OXFORD

Briefings in Bioinformatics, 2023, **24(2)**, 1–12
https://doi.org/10.1093/bib/bbad025
Advance access publication date 25 January 2023
Problem Solving Protocol

Subtype-DCC: decoupled contrastive clustering method for cancer subtype identification based on multi-omics data

Jing Zhao, Bowen Zhao, Xiaotong Song, Chujun Lyu, Weizhi Chen, Yi Xiong and Dong-Qing Wei

Corresponding authors. Yi Xiong and Dong-Qing Wei, State Key Laboratory of Microbial Metabolism, and School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China. Tel: +86 21-34204573; Email: xiongyi@sjtu.edu.cn, dqwei@sjtu.edu.cn

Abstract

Due to the high heterogeneity and complexity of cancers, patients with different cancer subtypes often have distinct groups of genomic and clinical characteristics. Therefore, the discovery and identification of cancer subtypes are crucial to cancer diagnosis, prognosis and treatment. Recent technological advances have accelerated the increasing availability of multi-omics data for cancer subtyping. To take advantage of the complementary information from multi-omics data, it is necessary to develop computational models that can represent and integrate different layers of data into a single framework. Here, we propose a decoupled contrastive clustering method (Subtype-DCC) based on multi-omics data integration for clustering to identify cancer subtypes. The idea of contrastive learning is introduced into deep clustering based on deep neural networks to learn clustering-friendly representations. Experimental results demonstrate the superior performance of the proposed Subtype-DCC model in identifying cancer subtypes over the currently available state-of-the-art clustering methods. The strength of Subtype-DCC is also supported by the survival and clinical analysis.

Keywords: cancer subtyping, multi-omics, deep clustering, contrastive learning

Introduction

Cancer is a disease with complex origins and accounts for one in six global deaths, according to the World Health Organization [1]. Gene alterations, epigenetic changes, the cellular biological context and patient-specific characteristics may all determine cancer formation and proliferation [2]. Since cancer is a heterogeneous disease with diverse pathogeneses and clinical features, morphologically similar tumors can have distinct pathogeneses belonging to various subtypes, which refer to the clusters of tumors that have shared characteristics within a cancer type [3]. The prognostic response and treatment outcome of different cancer subtypes vary considerably, so determining the cancer subtype is vital to cancer diagnosis, prognosis and treatment.

With the advancement of high-throughput sequencing and experimental techniques, omics and clinical data are increasingly accumulating from various cancer profiling projects, such as The Cancer Genome Atlas (TCGA) project [4], which can facilitate a more comprehensive understanding of the complex mechanisms of various cancers. Early cancer subtyping studies focus on single omic data (such as gene expression). However, the integration of multi-omics data associated with cancer occurrence and development can lead to a better understanding of the pathogenic mechanism of cancers, cancer subtyping and personalized treatment plans, which cannot be attained by utilizing only single omic data.

Integration, analysis and interpretation of large-scale multi-omics data have become an area of increasing interest in cancer research [5, 6]. The complex heterogeneity and high dimensionality of multi-omics data make the effective integration of them challenging. Over the last decade, considerable effort has been devoted to the development of numerous computational methods for multi-omics data integration [7, 8]. These approaches can be roughly categorized into three classes in terms of the major strategies for multi-omics data integration: early, intermediate and late integration [5]. Early integration methods perform a simple concatenation of features from the omic data into a single feature combination, increasing dimensionality and ignoring the unique data distribution in different omics levels. The early integration approaches include K-means, Spectral clustering

Jing Zhao is a master student at School of Life Sciences and Biotechnology, Shanghai Jiao Tong University. She works on multi-omics cancer subtyping through deep learning methods.
Bowen Zhao is a master student at School of Life Sciences and Biotechnology, Shanghai Jiao Tong University. He works on integrating single-cell omics data with machine learning methods.
Xiaotong Song is a PhD candidate at the School of Mathematical Sciences, Shanghai Jiao Tong University. Her research focuses on machine-learning-based multi-omics data analysis.
Chujun Lyu is a master student at the School of Life Sciences and Biotechnology, Shanghai Jiao Tong University. Her research focuses on bioinformatics, deep learning and computer-aided drug design.
Weizhi Chen is a master student at the School of Life Sciences and Biotechnology, Shanghai Jiao Tong University. He works on computational structural biology and computer-aided drug design through molecular dynamics simulations.
Yi Xiong is an associate professor at School of Life Sciences and Biotechnology, Shanghai Jiao Tong University. He is also affiliated with Shanghai Artificial Intelligence Laboratory.
Dong-Qing Wei is a tenured professor at the School of Life Sciences and Biotechnology, Shanghai Jiao Tong University. He is also affiliated with Zhongjing Research and Industrialization Institute of Chinese Medicine and Peng Cheng Laboratory.
Received: October 21, 2022. **Revised:** December 21, 2022. **Accepted:** January 8, 2023

and LRAcluster (low-rank-approximation-based multi-omics data clustering) [9] and so on. Late integration methods separately learn each omic layer and then merge the clustering results into a single solution in either a hierarchical, ensemble or linear way, in which weak signals from each omic layer may be lost during the integration phase. The late integration approaches include consensus clustering [10] and perturbation clustering for data integration and disease subtyping (PINS) [11]. Both early and late integration methods fail to model the interactions among the features in different omics data. Instead, intermediate integration methods have gradually become mainstream, which consolidate data by constructing a holistic model for joint dimensionality reduction and clustering without simply concatenating features or merging results. The intermediate integration methods could be further divided into at least three main classes, including statistical methods, deep learning and similarity-based methods.

Statistical methods (such as NMF [12], MCCA [13], iCluster [14], iCluster+ [15] and iClusterBayes [16]) model the distribution of each data type and then maximize the likelihood of multi-omics data based on joint latent variables. However, due to the complexity of multi-omics data, traditional statistical or mathematical models still face significant challenges in accurately modeling the high-dimensional multi-omics data. More recently, deep learning algorithms have been trained to model complex patterns in multi-omics data [17]. Most multi-omics clustering methods using deep learning algorithms are based on projection of heterogeneous omics data to a common latent subspace by Autoencoder (AE) [18–24], Variational Autoencoders [25], Generative Adversarial Network (GAN) [26], manifold optimization [27], subspace learning [28–31] and others [32–36]. The pioneering similarity-based method is similarity network fusion (SNF) method [37], which constructs a sample (e.g. patient) similarity network for the omic data and then integrates these networks into a unified similarity network that represents the full spectrum of underlying data, using a nonlinear combination method. NEMO [38] was inspired to calculate the similarity matrix by radial basis function kernels and perform spectral clustering, which is suitable for incomplete and intersecting omics datasets. Additionally, there are other kinds of similarity-based methods for cancer subtyping [39].

Although these methods provide data integration solutions for cancer subtyping, most traditional methods separate feature extraction from clustering tasks and offer different unsupervised classification algorithms for cancer subtype identification after dimensionality reduction of multi-omics data. No prior studies have focused on developing both tasks simultaneously, which extract the features at the dimensionality reduction stage without suitable clustering structure, and cannot obtain a competitive subtyping performance. For high-dimensional multi-omics data, it is challenging to extract features that are both individually different and cluster-friendly. At the same time, the small training datasets trouble the optimization problem of the model when training on different datasets. In order to improve the generalization ability of the model and achieve comparable performance on different cancer datasets, we introduce Subtype-DCC, a novel subtype model that combines deep clustering [40] and decoupled contrastive learning [41]. The deep clustering part ensures that this method is an end-to-end approach, which implies that the deep representation learning and clustering are jointly optimized, resulting in representations that are both individually diverse and more suitable for clustering. In addition, decoupled contrastive learning is suitable for optimizing with small batch size, which helps to solve the optimization dilemma of small sample datasets.

To evaluate the prediction performance of Subtype-DCC, we compared its performance with that of ten state-of-the-art multi-omics data clustering methods on nine datasets from TCGA. Additionally, we conduct a series of survival and clinical analysis to demonstrate the strength of Subtype-DCC.

In summary, our innovations are as follows:

(i) For the first time, we apply the deep clustering algorithm to cancer subtyping, expanding the breadth of cancer subtyping methods, and introduce self-supervised learning to clustering methods.
(ii) Subtype-DCC conducts contrastive learning at both the sample space and cluster space. Moreover, the model is jointly optimized to ensure the learned features are both individually different and cluster-friendly.
(iii) We propose an end-to-end model that can be applied to datasets with different sample sizes, eliminating the need for manual feature extraction and staged task execution, which is more user-friendly.

Methods

Method overview

Deep clustering is a series of clustering methods that adopt deep neural networks to learn clustering-friendly representations [40]. The preliminary knowledge of deep clustering includes the relevant network architectures for feature representation and optimizing objectives. As a simple and effective paradigm of unsupervised learning, contrastive learning has achieved outstanding performance in the computer vision field [42, 43]. The essence of contrastive learning is to map the original data to a feature space wherein the similarities of positive pairs are maximized, while those of negative pairs are minimized [44]. As a member of the deep clustering family, contrastive clustering [45] simultaneously utilizes contrastive samples to facilitate clustering in both sample space and cluster space. Such a clustering-oriented contrastive learning paradigm helps the model minimize the intercluster similarities to separate different clusters. However, the contrastive clustering model may depend on a large batch size to achieve competitive performance. To address this dilemma, we employ a decoupled contrastive learning loss [41] to optimize contrastive clustering.

Subtype-DCC (shown in Figure 1) is a decoupled contrastive clustering method based on multi-omics datasets for cancer subtyping, which can extract suitable features through contrastive learning and help balance the optimization problems of the model adaptability when training on different datasets. The model inherits the framework of contrastive clustering [45], which consists of the pair construction backbone (PCB), the instance-level contrastive head (ICH) and the cluster-level contrastive head (CCH), and the three components are jointly learned. First, PCB takes the combined data of the four omics as input, constructs the data pair through data augmentation and then reduces the dimensionality of the data to extract the embeddings from the augmented samples. Then, ICH and CCH apply contrastive learning in the sample and cluster spaces of the embedding matrix, respectively. After training, the subtype clustering results of samples can be easily obtained through the soft labels predicted by CCH.

Benchmark datasets

To evaluate the performance of our proposed model and the comparison with the state-of-the-art methods, we utilized nine TCGA cancers with multi-omics data from four molecular

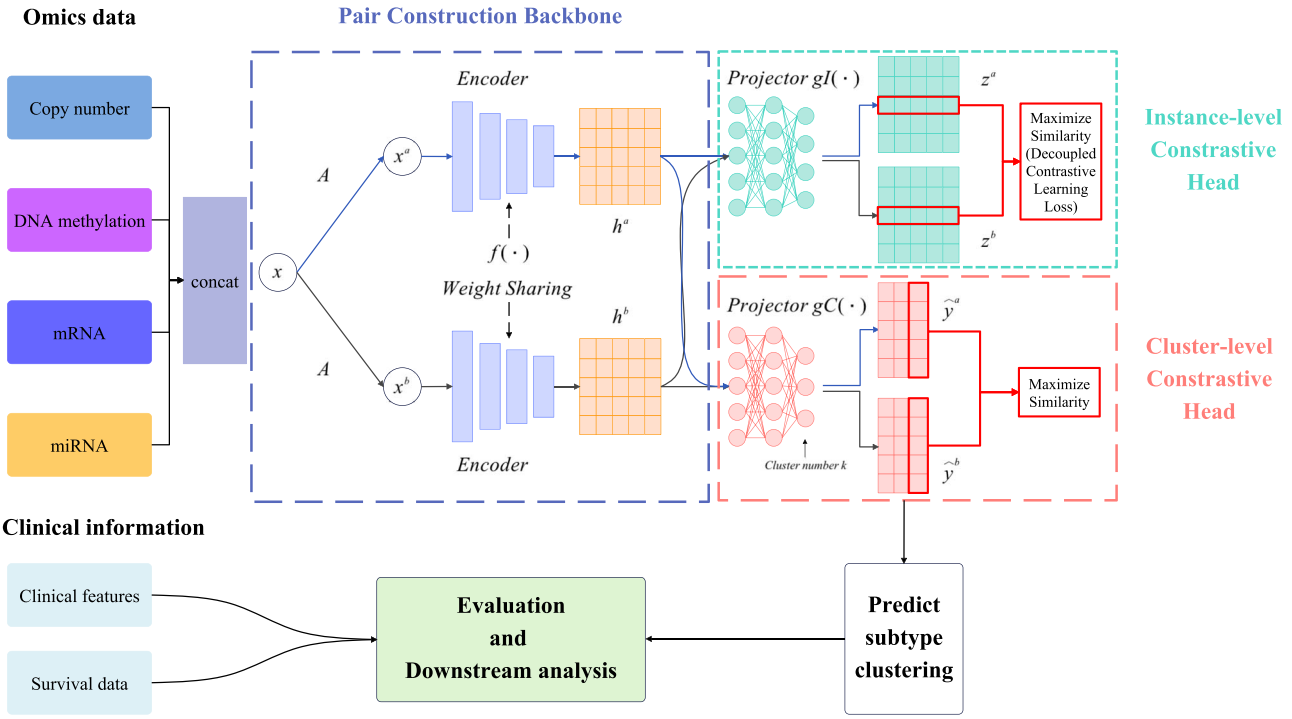

Figure 1. The framework of Subtype-DCC. First, the four omics data are concatenated as model inputs. Then we construct data pairs using data augmentation *A*. Given data pairs, one shared deep neural network is used to extract embeddings from data augmentations. Two separate projectors are used to project the embeddings into the row and column space, where the instance-level and cluster-level contrastive learning are conducted, respectively. To avoid the coupling between instance-level, decoupled contrastive learning loss is used. After training, the CCH is used to predict the subtype clustering. The final subtyping results were combined with clinical information for model evaluation and downstream analysis.

platforms (Copy Number, messenger RNA (mRNA), micro RNA (miRNA) and DNA methylation) by the previous study [26]. The cancer datasets include Breast Invasive Carcinoma (BRCA), Bladder Urothelial Carcinoma (BLCA), Kidney Renal Clear Cell Carcinoma (KIRC), Lung Adenocarcinoma (LUAD), Pancreatic Adenocarcinoma (PAAD), Skin Cutaneous Melanoma (SKCM), Stomach Adenocarcinoma (STAD), Uterine Corpus Endometrial Carcinoma (UCEC) and Uveal Melanoma (UVM). After performing the normalization provided by [26], we finally obtained 1031 samples in BRCA, 399 in BLCA, 488 in KIRC, 490 in LUAD, 176 in PAAD, 446 in SKCM, 407 in STAD, 510 in UCEC and 80 in UVM. There are a total of 4027 samples, which ensure a reliable number of samples for the stability of the analysis results. Finally, we used 3105 copy number features, 3217 mRNA features, 383 miRNA features and 3139 DNA methylation features for model training.

Evaluation criteria

We utilized two evaluation criteria (i.e. −log10 *P*-values and the number of significant clinical parameters) used by the previous study [46] to evaluate the clustering performance of our method. They are defined as below.

First, the differential survival −log10 *P*-values were measured between the obtained clusters using the log-rank test [47]. The metric assumes that subtypes of patients are biologically meaningful if they have significantly different survival. Second, the number of significant clinical parameters was tested for the enrichment of clinical labels in the clusters. A total of six clinical labels were chosen for which we tested enrichment: age at diagnosis, gender, pathologic stage, pathologic T, pathologic N and pathologic M. The four latter parameters are discrete pathological parameters, measuring the progression of the tumor (T), metastases (M), cancer in lymph nodes (N) and the total progression (pathologic stage). Enrichment for discrete parameters was calculated using the χ^2 test for independence and for numeric parameters using Kruskal–Wallis test [46].

Pair construction backbone

Subtype-DCC uses data augmentations to construct data pairs; specifically, applying data augmentation *A* to a given data instance x, resulting in two correlated samples denoted as x_i^a, x_i^b. The previous works have shown that an appropriate augmentation strategy is critical for good performance in downstream tasks [45]. In this work, three types of data augmentation methods are tried, including Noise, Mask and Dropout. For a given data matrix, each augmentation is applied independently with a certain probability. Specifically, Noise adds a standard Gaussian noise to the original matrix; Mask randomly masks the original matrix at a certain probability and Dropout randomly drops the original matrix at a certain probability.

One shared deep neural network $f(\cdot)$ is used to extract embedding from the augmented samples via $h_i^a = f(x_i^a)$ and $h_i^b = f(x_i^b)$. In terms of network architecture, our method adopts a four-layer deep neural network.

Instance-level contrastive head

Subtype-DCC follows the idea of contrastive learning and aims to maximize the similarities of positive pairs while minimizing those of negative ones. In the cancer subtyping task, since no prior labels are available, instance-level positive and negative sample pairs are constructed from pseudo-labels generated by data augmentations [45]. In detail, samples augmented from the same sample form positive pairs, while other samples form negative pairs.

Formally, given a mini-batch of size *N*, Subtype-DCC performs data augmentation on each sample x_i and results in 2*N* data

samples $\{x_1^a, \cdots, x_N^a, x_1^b, \cdots, x_N^b\}$. For a specific sample x_i^a, there are $2N-1$ pairs in total, of which the corresponding augmented sample x_i^b form a positive pair $\{x_i^a, x_i^b\}$ and leave other $2N-2$ pairs to be negative.

To mitigate the information loss caused by contrastive loss, we do not directly perform contrastive learning on the embedding matrix. Instead, a two-layer nonlinear Multilayer Perceptron (MLP) $g_I(\cdot)$ is stacked to map the embedding matrix to a subspace via $z_i^a = g_I(h_i^a)$ where the decoupled instance-level contrastive loss is applied. The similarity of paired samples is measured by cosine distance, i.e.

$$s\left(z_i^{k1}, z_j^{k2}\right) = \frac{\left(z_i^{k1}\right)\left(z_j^{k2}\right)^T}{\left\|z_i^{k1}\right\|\left\|z_j^{k2}\right\|} \tag{1}$$

where $k1, k2 \in \{a, b\}$and $i, j \in [1, N]$. In contrastive clustering [45], to optimize pairwise similarities, the loss for a given sample $x_i^a$is in the form of

$$l_i^a = -\log \frac{\exp\left(s\left(z_i^a, z_i^b\right)/\tau_I\right)}{\sum_{j=1}^{N}\left[\exp\left(s\left(z_i^a, z_j^a\right)/\tau_I\right)+\exp\left(s\left(z_i^a, z_j^b\right)/\tau_I\right)\right]} \tag{2}$$

where τ_I is the instance-level temperature parameter to control the softness. However, there is a significant negative–positive-coupling effect in this cross-entropy loss (InfoNCE), resulting in unsuitable learning efficiency relative to batch size [41]. We refer to Decoupled Contrastive Learning (DCL) [41] objective to address this coupling phenomenon. A decoupled instance-level contrastive loss is achieved by removing positive pairs from the denominator of Equation 2.

$$l_{DCi}^a = -\log \frac{\exp\left(s\left(z_i^a, z_i^b\right)/\tau_I\right)}{\sum_{j=1, j\neq i}^{N}\left[\exp\left(s\left(z_i^a, z_j^a\right)/\tau_I\right)+\exp\left(s\left(z_i^a, z_j^b\right)/\tau_I\right)\right]} \tag{3}$$

$$= -s\left(z_i^a, z_i^b\right)/\tau_I + \log\sum_{j=1, j\neq i}^{N}\left[\exp\left(s\left(z_i^a, z_j^a\right)/\tau_I\right) + \exp\left(s\left(z_i^a, z_j^b\right)/\tau_I\right)\right] \tag{4}$$

The model identifies all positive pairs in the entire dataset by computing a decoupled instance-level contrastive loss on each augmented sample, namely,

$$L_{ins} = \frac{1}{2N}\sum_{i=1}^{N}\left(l_{DCi}^a + l_{DCi}^b\right) \tag{5}$$

Cluster-level contrastive head

According to the concept of 'label as representation' in contrastive clustering [45], when projecting a data sample into a space of the same dimensionality as subtype clusters, the i-th element of its feature indicates its probability that it belongs to the i-th cluster, and the feature vector denotes its soft label accordingly.

Formally, if we define $Y^a \in R^{N\times M}$as the output of CCH for a mini-batch under one of the augmented views (and Y^b for the other augmented view), where N represents the batch size and M is equal to the number of subtype clusters, then $Y_{n,m}^a$represents the probability of a sample n being assigned to cluster m. Since each patient belongs to only one cancer subtype, the rows of Y^a should ideally be one-hot vectors. In this sense, the i-th column of Y^a can be viewed as representing the i-th subtype cluster and all columns should be different from each other.

Similar to the projection layer structure used in the ICH, the embedding matrix is projected into an M-dimensional space using another two-layer MLP $g_C(\cdot)$ via $y_i^a = g_C(h_i^a)$, where y_i^a indicates the soft label of sample x_i^a (the i-th row of Y^a). We can consider $\hat{y}_i^a$ to be the i-th column of Y^a, which is the representation of subtype cluster i under the corresponding data augmentation view, and we combine it with $\hat{y}_i^b$ to form a positive cluster pair $\{\hat{y}_i^a, \hat{y}_i^b\}$, while considering the remaining $2M-2$ pairs as negative, where $\hat{y}_i^b$ denotes another augmented view representation of cluster i. The cosine distance is applied to measure the similarity between subtype cluster pairs, that is,

$$s\left(\hat{y}_i^{k1}, \hat{y}_j^{k2}\right) = \frac{\left(\hat{y}_i^{k1}\right)^T\left(\hat{y}_j^{k2}\right)}{\left\|\hat{y}_i^{k1}\right\|\left\|\hat{y}_j^{k2}\right\|} \tag{6}$$

where $k1, k2 \in \{a, b\}$ and $i, j \in [1, M]$. The following loss function is utilized to distinguish cluster $\hat{y}_i^a$ from all other clusters except $\hat{y}_i^b$, i.e.

$$\hat{l}_i^a = -\log \frac{\exp\left(s\left(\hat{y}_i^a, \hat{y}_i^b\right)/\tau_C\right)}{\sum_{j=1}^{M}\left[\exp\left(s\left(\hat{y}_i^a, \hat{y}_j^a\right)/\tau_C\right)+\exp\left(s\left(\hat{y}_i^a, \hat{y}_j^b\right)/\tau_C\right)\right]} \tag{7}$$

where τ_C is the temperature parameter that controls the softness at cluster-level. By traversing all clusters, the cluster-level contrastive loss is finally computed as follows:

$$L_{clu} = \frac{1}{2M}\sum_{i=1}^{M}\left(\hat{l}_i^a + \hat{l}_i^b\right) - H(Y) \tag{8}$$

where $H(Y) = -\sum_{i=1}^{M}\left[P\left(\hat{y}_i^a\right)\log P\left(\hat{y}_i^a\right) + P\left(\hat{y}_i^b\right)\log P\left(\hat{y}_i^b\right)\right]$ is the entropy of subtype cluster assignment probabilities $P\left(\hat{y}_i^k\right) = \sum_{t=1}^{N} Y_{ti}^k / \left\|Y^k\right\|_1$, $k \in \{a, b\}$ within a mini-batch under each data augmentation. This term is useful for avoiding the trivial solution that most instances are assigned to the same subtype cluster [48].

Objective function

ICH and CCH optimization is an end-to-end process involving two heads that are optimized simultaneously in one stage. The overall objective function is calculated based on the decoupled instance-level and cluster-level contrastive loss, i.e.

$$L = L_{ins} + \lambda L_{clu} \tag{9}$$

As a general rule, a dynamic weight parameter λ is set to balance the two losses during the training [42]. In practice, we set λ to 1 by default, i.e. follow the simple addition of the two losses in the original contrastive clustering setting [45].

Experimental settings

The model is developed with Python 3.10.6 and Pytorch 1.12.1. For optimal performance of the model, we tuned six main hyper-parameters: instance-level temperature, cluster-level temperature, batch size, feature dimension, learning rate and training epoch. These hyper-parameters may have a huge impact on model performance. The temperature of instance-level and cluster-level controls the softness of the model. Batch size affects decoupled contrastive learning performance. The feature dimension determines the size of the feature space for keeping data information. The learning rate determines how quickly the model converges. Training epochs can set the appropriate training time for the model. Due to the large number of parameter combinations, we tune six hyperparameters in the order of instance-level temperature, cluster-level temperature, batch size, feature dimension, learning rate and training epoch. While tuning one of the hyper-parameters, the other five hyper-parameters were kept constant. The parameter selection is shown in Table S1.

Table 1. Performance comparison of Subtype-DCC and other methods on nine cancer datasets (−log10 *P*-values/number of significant clinical parameters, bold indicates that this method performs best on the corresponding cancer dataset)

Method/Dataset	BRCA	BLCA	KIRC	LUAD	PAAD	SKCM	STAD	UCEC	UVM
Subtype-DCC	1.11/5	2.33/**6**	**8.79/6**	1.69/**4**	**3.75**/1	**5.94/4**	1.48/2	5.46/**1**	2.77/0
Subtype-GAN	**1.28/6**	1.45/4	7.77/**6**	**2.83**/3	1.65/1	0.1/2	0.39/2	**7.4**/1	2.62/0
NEMO	1.21/**6**	**2.8**/5	5.72/5	2.63/**4**	3.04/1	5.01/**4**	**1.8**/2	5.96/1	2.38/0
SNF	0.93/5	1.31/**6**	8.19/**6**	2.23/**4**	3.24/3	5.27/**4**	0.72/2	5/1	2.77/0
PINS	1.42/2	1.61/3	4.44/**6**	2.46/**4**	3.41/4	2.32/1	1.26/2	5.04/1	**3.63**/0
NMF	0.4/4	0.24/1	5.63/5	0.42/1	1.49/0	3.54/3	0.1/1	5.1/1	1.39/0
MCCA	1.73/3	1.03/3	7.91/5	0.49/3	2.15/4	0.89/3	0.18/1	3.75/1	1.1/**1**
ICluster	0.53/4	0.21/3	2.95/4	0.23/3	0.54/0	0.98/1	0.06/1	2.13/1	1.36/**1**
Spectral	0.08/4	1.67/3	5.46/**6**	0.6/1	2.39/0	1.77/2	0.19/2	0.81/1	1.82/0
K-Means	0.12/5	0.66/3	4.77/**6**	1.01/1	2.38/0	1.56/1	0.01/**3**	7.03/1	1.67/0
LRAclutser	0.27/5	0.63/1	6.83/**6**	0.19/1	2.03/1	2.05/1	0.14/1	4.58/1	2.52/0

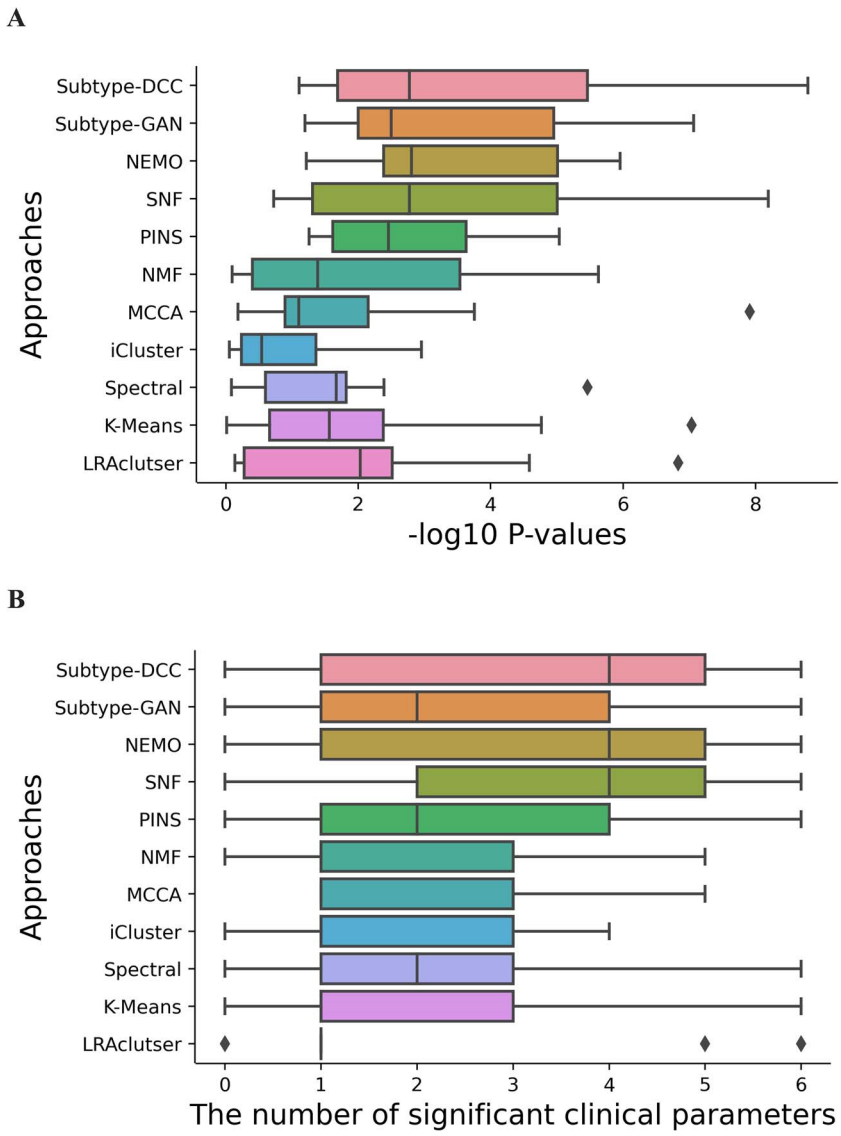

Figure 2. Performances of Subtype-DCC and other approaches. Subtype-DCC is compared with ten state-of-the-art multi-omics data clustering methods, including Subtype-GAN, NEMO, SNF, PINS, NMF, MCCA, iCluster, Spectral, Kmeans and LRAcluster. (**A**) the −log10 *P*-values, (**B**) the number of significant clinical parameters (age at diagnosis, gender, pathologic stage, pathologic T, pathologic N and pathologic M).

Subtype-DCC uses the addition of Gaussian noise for data augmentation. Our encoder uses MLPs with 5000, 2000, 1000 and 256 neurons, respectively. The activation function uses RELU [49]. The instance-level projection head and cluster-level projection head both use two layers of MLP and also use RELU for nonlinear activation. The number of neurons projected at the instance level

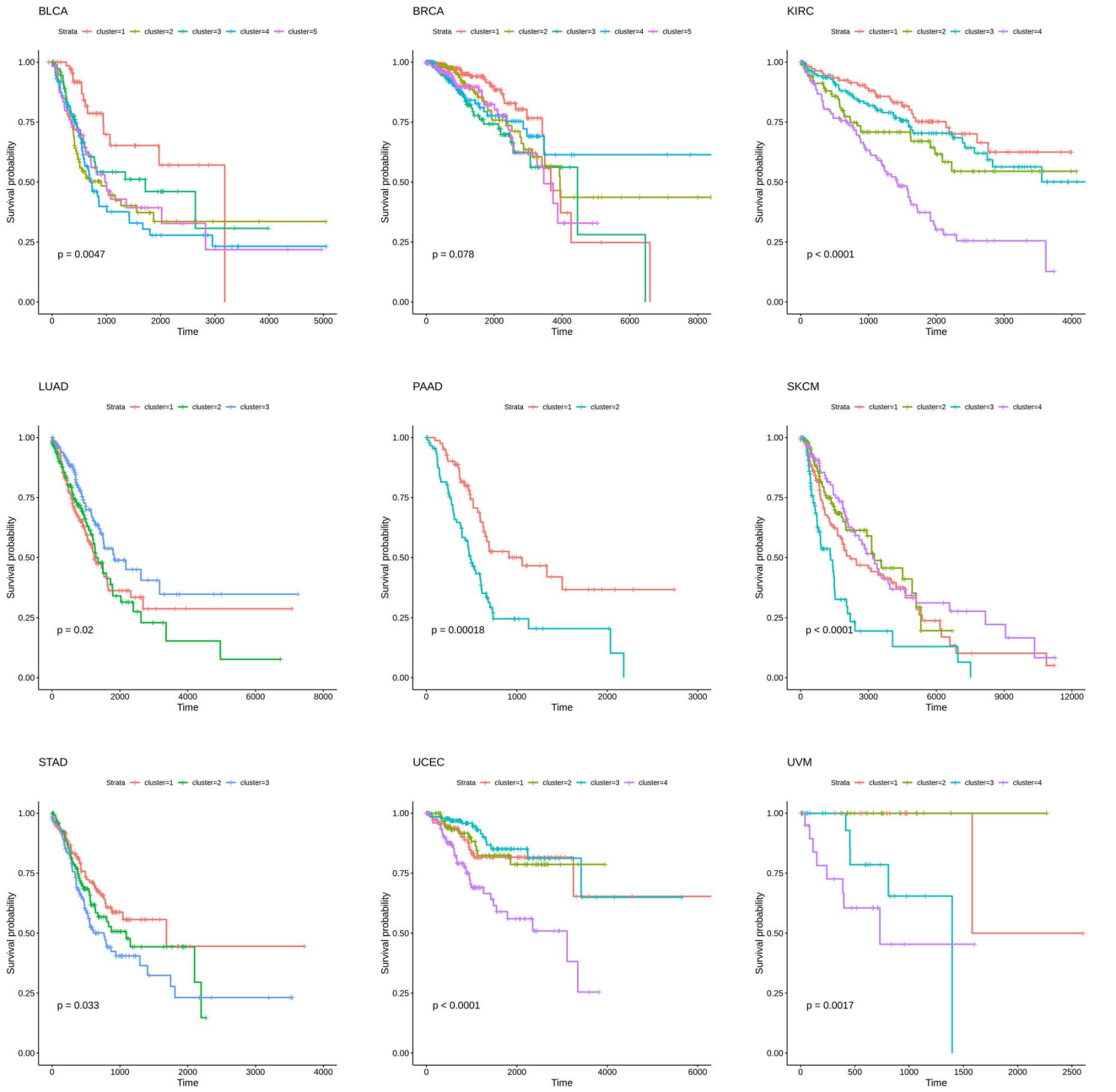

Figure 3. Survival analysis curves of Subtype-DCC on nine datasets. The different colors represent the grouping of samples according to the cluster labels output by Subtype-DCC.

is 256 and 128, respectively. The 128-dimensional vector output by instance-level projection head is used to calculate the contrastive loss between samples, and the temperature parameter is set to 0.5. The number of neurons projected at the clustering level is 256 and M, respectively. M corresponds to the pre-set number of clusters for each cancer dataset. The cluster-level temperature parameter is set to 1. The optimal value of batch size is 64, and the learning rate is set to 3e-4.

Results

Comparison of Subtype-DCC with the state-of-the-art methods on nine cancer datasets

In this section, we conduct the clustering performance comparison of Subtype-DCC against ten state-of-the-art methods for multi-omics data integration (i.e. Subtype-GAN [26], NEMO [38], SNF [37], PINS [11], NMF [12], MCCA [13], iCluster [16], Spectral, K-means and LRAcluster [9]). Here, we present the detailed results for the nine cancer datasets (as shown in Table 1), in which previous studies have obtained reasonable numbers of subtypes of these datasets, ensuring the fairness of comparison among all the approaches. We used the survival $-\log10$ *P*-values and the number of significant clinical parameters to evaluate the clustering performances. For methods with fluctuating results, we train the models five times and take their average as the final evaluation performance. Overall, Subtype-DCC achieved the best results on at least one metric over six datasets. In particular, the top-performing results for all metrics were achieved in two datasets. In the KIRC dataset, the survival $-\log10$ *P*-values is 8.79, and the number of significant clinical parameters is 6, which

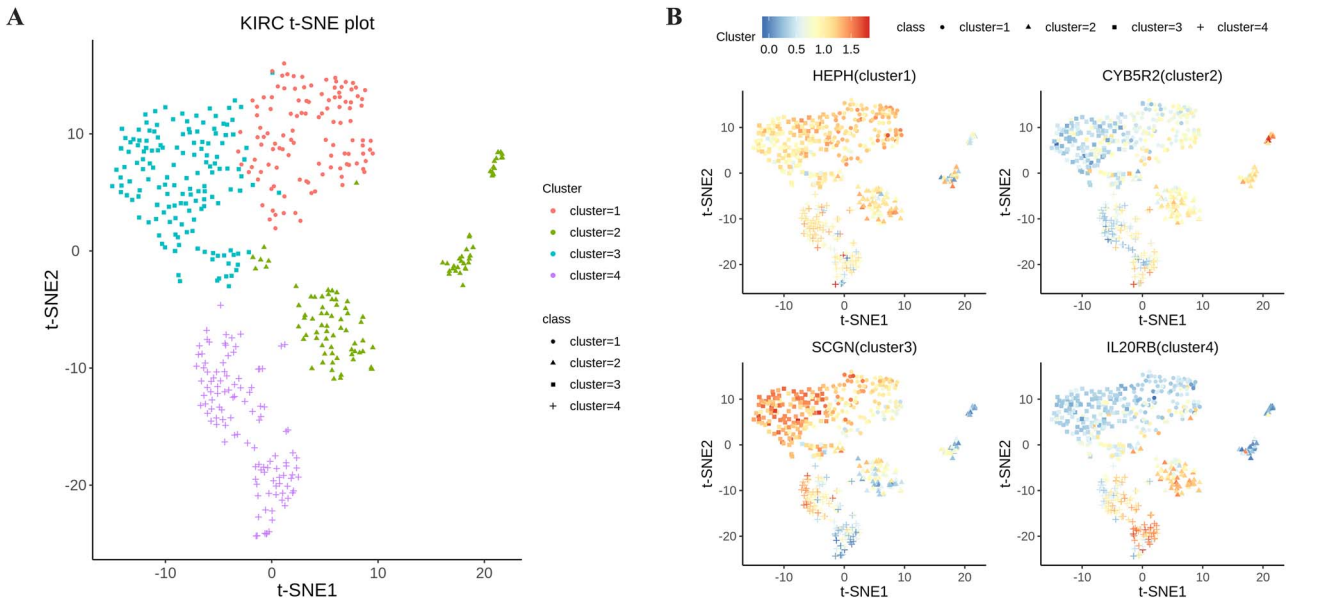

Figure 4. Subtypes identified in KIRC and biomarkers screened from subtyping results. (**A**) the t-SNE visualization of latent embedding generated by Subtype-DCC on the KIRC dataset; (**B**) the expression of biomarker mRNAs screened in four different subtype clusters across all samples; red indicates high expression, and blue indicates low expression.

suggests that Kidney renal clear cell carcinoma patients are well subtyped, including KIRC-T1, KIRC-T2A, KIRC-T2B and KIRC-T2C [50]. In the SKCM dataset, the survival −log10 *P*-values is 5.94 and the number of significant clinical parameters is 4, which suggests that Skin Cutaneous Melanoma patients are well subtyped.

The −log10 *P*-values shown in Figure 2A indicate that Subtype-DCC outperformed the other ten methods over nine cancer datasets, and especially achieved the most significant results on three datasets (KIRC, PAAD and SKCM). For the clinical parameters enrichment analysis, Subtype-DCC delivered the most substantial results on five datasets (BLCA, KIRC, LUAD, SKCM and UCEC). These findings indicate that the performance of Subtype-DCC was better than or competitive with those of the ten state-of-the-art clustering methods (Figure 2B). It implies that Subtype-DCC effectively captured and integrated the dominant part of each omic dataset.

To verify the effects of the prognosis predictions of different cancer subtypes, we plotted survival curves of Subtype-DCC on nine cancer datasets. According to Figure 3, other than BRCA, the cancer subtypes identified by our method on the other eight datasets all show significantly different survival curves. A significant difference in survival curves was observed between the subtypes, and this difference increased over time, indicating different subtypes have varying survival probabilities. Specifically, in the case of KRIC, cluster 4 had a lower survival probability compared with the other subtypes when the time was above 2000. This implies that our method might help identify groups of patients with different prognoses and aid in precision treatment planning.

Ablation experiment on multi-omics data

To demonstrate the advantages of multi-omics data for cancer subtyping tasks, we performed ablation experiments on nine datasets based on different omics. (Figure S1) The single omic data from copy number features, DNA methylation features, mRNA features and miRNA features were removed in turn. The results show that after removing each omic data, the evaluation metrics of the model clustering have decreased to varying degrees. The advantages of multi-omics data fusion for cancer subtyping are demonstrated. At the same time, we found that after removing the mRNA omic data, all evaluation metrics dropped significantly, which proved the importance of mRNA characteristics and provided guidance for our subsequent screening of biomarkers.

Subtypes identified in KIRC

Based on the embedding learned by our model, we reduced the dimensionality of latent layer factors and visualized corresponding clusters to study the subtypes identified by Subtype-DCC. For KIRC, one of the top-performing datasets, it can be observed that different subtype clusters are well separated, proving that the model learned a meaningful latent representation (Figure 4A). We further performed differential expression analysis to discover biomarkers by calculating the t-test of each feature in different omics and sorting them. (A t-test calculation was performed on each gene, and then the top n most significant genes were selected by sorting according to the *P*-value of each gene.)

Each omic was screened for biomarkers for each cluster type. Differentially expressed mRNAs were found in the profile named KIRC-differential-genes of each KIRC subtype. We visualized one of the most prominent biomarkers for each subtype and observed that differentially expressed mRNAs could provide an intuitive distinction between these subtypes (Figure 4B). That is, differentially expressed mRNA biomarkers are highly expressed in their own cluster (indicated in red) and have low expression in other clusters (indicated in blue), which indicates a robust and interpretable relationship between biomarkers and the identified subtypes.

To understand the biological roles and potential functions of the biomarker mRNAs, Gene Ontology (GO) [51] enrichment analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) [52] signal pathway enrichment analysis were performed on each set of differentially expressed mRNAs for KIRC through R packages: 'clusterProfiler' (version:3.14.0) [53]. Figure 5 shows each set of differentially expressed mRNAs enriched in GO pathways. Figure 6

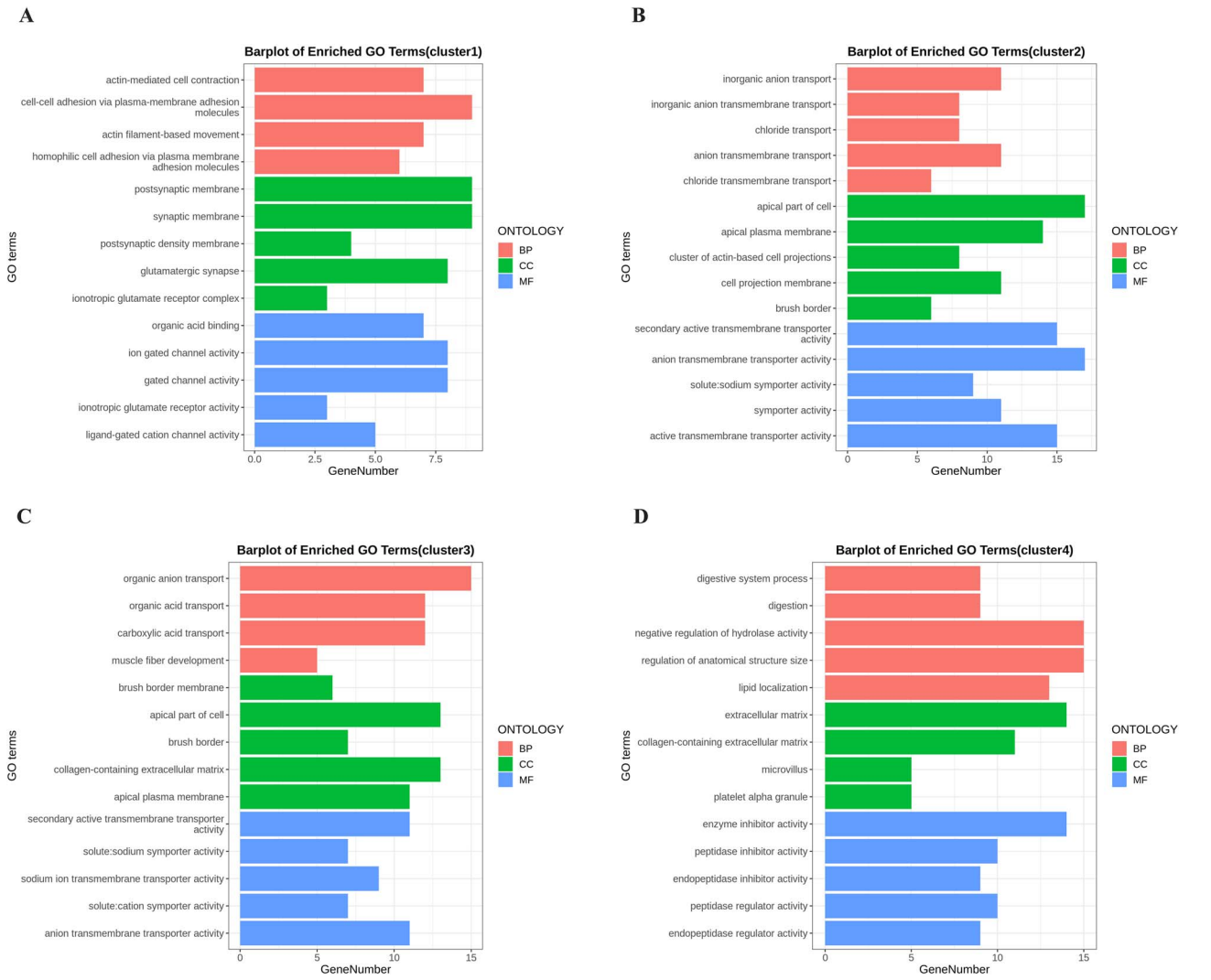

Figure 5. GO enrichment analysis of each set of differentially expressed mRNAs on KIRC. The y-axis represents GO-enriched terms. The x-axis represents the number of genes. The size of the bar represents the number of genes under a specific GO term. The BP (biological processes), CC (cellular component) and MF (molecular function) GO terms are colored by the adjusted *P*-values. (**A**) Differentially expressed gene enrichment analysis results in cluster 1. (**B**) Differentially expressed gene enrichment analysis results in cluster 2. (**C**) Differentially expressed gene enrichment analysis results in cluster 3. (**D**) Differentially expressed gene enrichment analysis results in cluster 4.

shows each set of differentially expressed mRNAs enriched in KEGG pathways.

Differentially expressed genes in cluster 1 are concentrated in the biological processes of actin-mediated cell contraction and cell–cell adhesion via plasma-membrane adhesion molecules (Figure 5A). For cluster 2, the differentially expressed genes are mainly involved in the process of inorganic anion transport (Figure 5B). Similarly, the differentially expressed genes in cluster 3 are related to organic anion transport (Figure 5C). The differential genes set of cluster4 are mainly enriched in processes such as the digestive system process and regulation of hydrolase activity (Figure 5D). These biological processes are closely associated with kidney renal clear cell carcinoma, further increasing the interpretability of mRNA biomarkers. The renal digestion and absorption related pathways enriched by KEGG will help elucidate the mechanism of tumor progression and metastasis of kidney renal clear cell carcinoma and the research of related targeted drugs (Figure 6).

We also performed similar differential expression and enrichment analysis on DNA methylation omic data by 'methylGSA' package [54]. According to the different methylation data platforms analyzed, the corresponding relationship between CpG sites and genes was obtained. The number of CpGs was included as a covariate for logistic regression analysis using the 'methylglm' function, and the number of CpGs in the DNA methylation data was corrected for enrichment analysis. The results showed that many differential genes were involved in the trans-Golgi network, pattern recognition receptor signaling pathway and cholesterol biosynthesis (Supplementary Figures S2–S5). The results of KEGG enrichment analysis showed that these DNA methylation changes were associated with pathways in cancer and MAPK signaling pathways, further confirming the contribution of epigenetics to cancer subtyping (Supplementary Figures S6–S9).

We employed DIANA-miRPath [55], which renders possible the functional annotation of one or more miRNAs provided by experimentally validated miRNA interactions derived from DIANA-TarBase to probe the signaling pathways that may involve differentially expressed miRNAs. As shown in Figure 7, these miRNAs are involved in some pathways related to the development of cancer, such as the transcriptional misregulation in cancer,

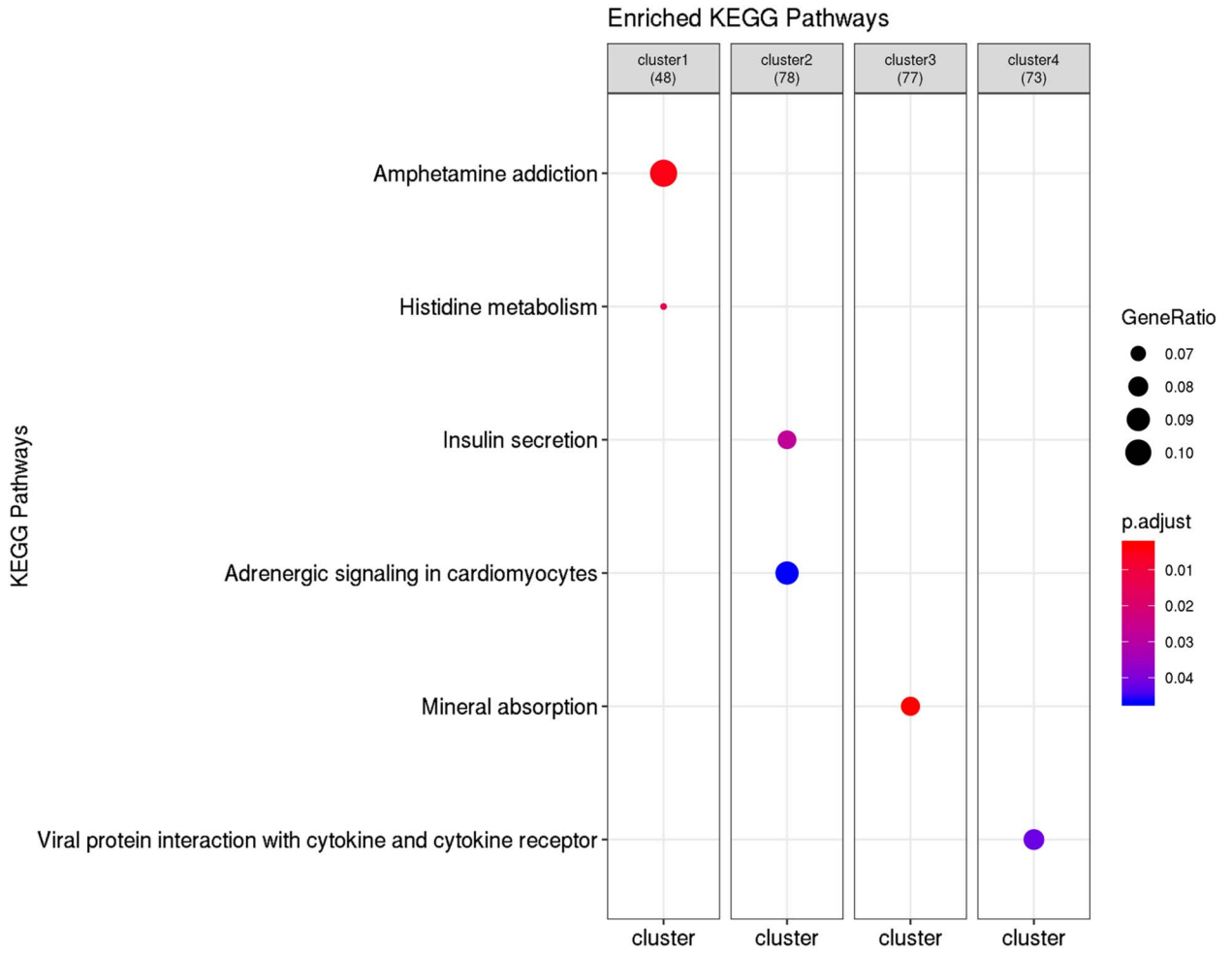

Figure 6. KEGG signal pathway enrichment analysis of each set of differentially expressed mRNAs on KIRC. The *x*-axis represents the different clusters, which indicates subtypes in KIRC. The *y*-axis represents the KEGG pathways, and the path name is shown on the left vertical axis. The size of the dots represents the ratio of the proportion of differentially expressed proteins annotated to this pathway to the proportion of proteins annotated to a certain pathway in the species. The larger the GeneRatio, the more reliable the enrichment significance of differential proteins in this pathway. The color of a dot represents the level of the adjusted *P*-value.

proteoglycans in cancer, microRNAs in cancer and pathways in cancer. Moreover, miR-200c [56] has been reported to play an important role in promoting kidney tumor growth and metastasis. MiRNAs such as miR-193b-3p [57, 58], miR-92b-3p [59] and Hsa-let-7a [60] can function as tumor suppressors in renal cell carcinoma.

Discussion and conclusion

Recent technological advances have accelerated the increasing availability of multi-omics biological data, which can represent data from different views. To take advantage of the complementary information contained in multi-omics data, there is a need to develop models that can represent and integrate different layers of data into a single framework.

Inspired by contrastive learning, we proposed Subtype-DCC method to identify cancer subtypes by integrating multi-omics data. Subtype-DCC combined contrastive learning representation and clustering at one stage, which implies that the clustering assignments and network parameters are jointly optimized. Benefiting from suitable feature representations learned by Subtype-DCC, our model shows better performance than several state-of-the-art methods on nine TCGA datasets. The best results were achieved in at least one of −log10 *P*-values or the number of significant clinical parameters over six datasets. Experimental results showed that Subtype-DCC has superiority over other methods on eight datasets in terms of *P*-value and survival curves. We also analyzed the omics data on KIRC to screen biomarkers of different subtypes for clinical application. Functional enrichment analysis of biomarker mRNAs demonstrated that the identified subtypes for cancer were related to renal digestion and absorption pathways. Visualization of biomarker mRNA expression levels intuitively demonstrated that the identified subtypes of cancer have some biological significance. Signal pathway enrichment analysis of differential methylation expressions and heatmaps of miRNA expressions provided insight into the elucidation of mechanisms of tumor occurrence, progression and metastasis as well as for research of related targeted drugs by studying cancer-related pathways. Although Subtype-DCC introduced contrastive learning to feature representation, it also has some limitations that influence its performance. In our model, the contrastive

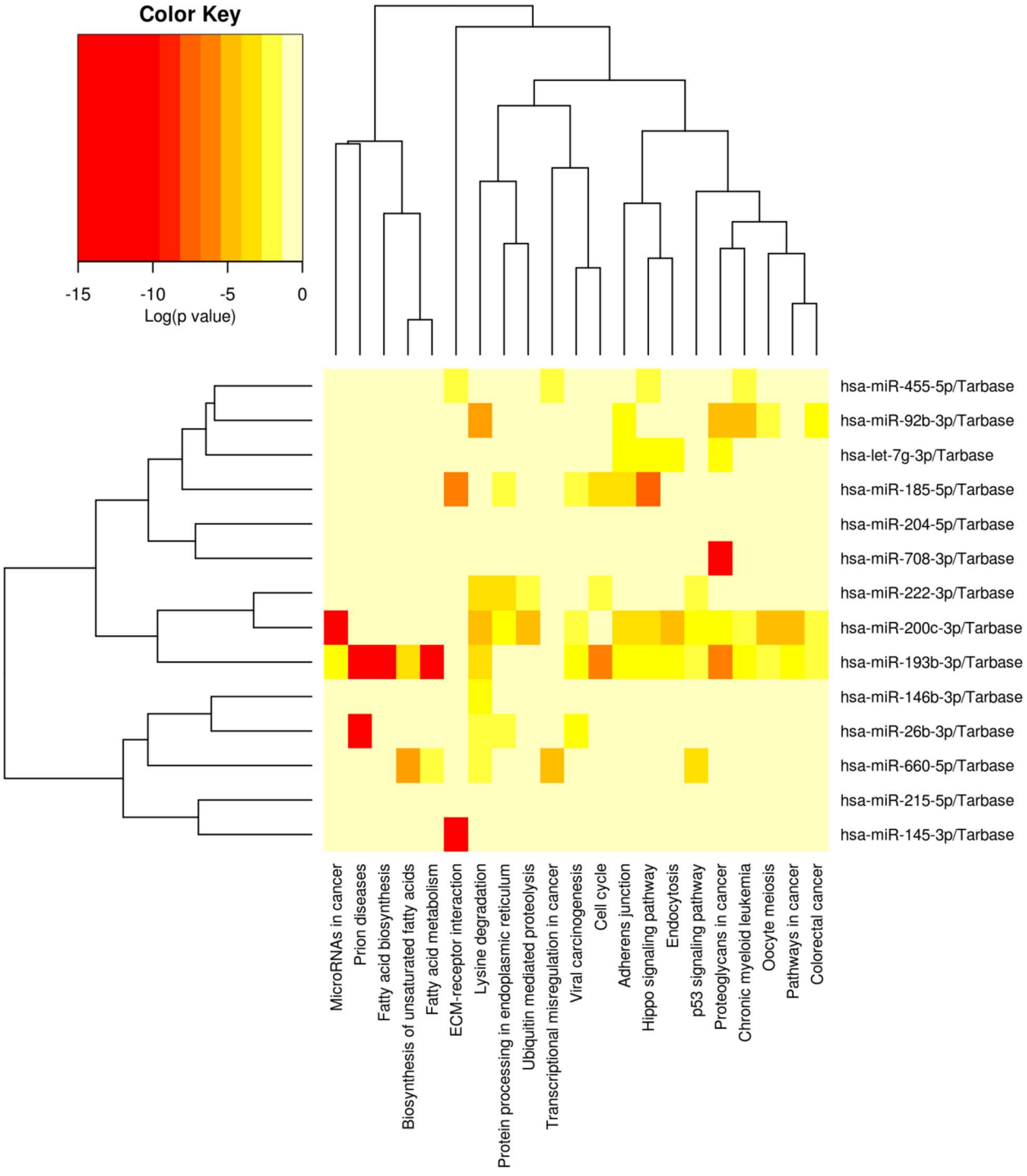

Figure 7. Heatmaps of significantly differentially expressed miRNAs among KIRC subtypes. The x-axis represents the significant signaling pathways that involve the differentially expressed miRNAs by utilizing the DIANA-miRPath. The y-axis represents the significantly differentially expressed miRNAs among KIRC subtypes.

learning only focuses on the matching of self-augmented positive pairs, which may lead to information loss at the embedding level compared with AE-based methods. The discovery of cancer subtypes is also inseparable from the role of drug treatment. How to integrate the known information on the effects of these known useful drugs has not been proposed. In future work, we will consider optimizing intra-cluster information for better subtypes and adding omics data, such as pharmacogenomics, to network integration to discover relationships among omics data.

Key Points

- This work introduces contrastive learning on cancer subtype identification based on multi-omics data. The self-supervised learning paradigm jointly optimized deep representation learning and the clustering parameters, enabling Subtype-DCC to learn suitable feature representations.

- Subtypes obtained by modeling on multi-omics data have certain guiding significance for subtype discovery. Important biomarkers of different subtypes for KIRC are identified, and the biological role and potential functions are determined by effectively utilizing functional enrichment analysis.
- Experimental results demonstrate that Subtype-DCC achieves excellent predictive ability in cancer subtyping, survival and clinical analysis, proving the superiority of Subtype-DCC over competing methods.

Data Availability

The source codes are available at https://github.com/zhaojingo/Subtype-DCC.

Funding

This work is supported by grants from the National Science Foundation of China (Grant Nos. 32070662, 62172274, 61832019, 32030063), the Science and Technology Commission of Shanghai Municipality (Grant No. 19430750600) and the Joint Research Fund for Medical and Engineering and Scientific Research at Shanghai Jiao Tong University (Grant No. YG2021ZD02, YG2019ZDA12, YG2019GD01).

References

1. Ferlay J EM, Lam F, Colombet M, *et al. Global Cancer Observatory: Cancer Today*. https://gco.iarc.fr/today (February 2021, date last accessed).
2. Kristensen VN, Lingjærde OC, Russnes HG, *et al.* Principles and methods of integrative genomic analyses in cancer. *Nat Rev Cancer* 2014;**14**:299–313.
3. Hoadley KA, Yau C, Wolf DM, *et al.* Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* 2014;**158**:929–44.
4. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, *et al.* The cancer genome atlas pan-cancer analysis project. *Nat Genet* 2013;**45**:1113–20.
5. Lipkova J, Chen RJ, Chen B, *et al.* Artificial intelligence for multimodal data integration in oncology. *Cancer Cell* 2022;**40**:1095–110.
6. Leng D, Zheng L, Wen Y, *et al.* A benchmark study of deep learning-based multi-omics data fusion methods for cancer. *Genome Biol* 2022;**23**:171.
7. Zhong Y, Lin Y, Chen D, *et al.* Review on integration analysis and application of multi-omics data. *Comput Eng Appl* 2021;**57**:1–17.
8. Akhoundova D, Rubin MA. Clinical application of advanced multi-omics tumor profiling: shaping precision oncology of the future. *Cancer Cell* 2022;**40**:920–38.
9. Wu D, Wang D, Zhang MQ, *et al.* Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. *BMC Genomics* 2015;**16**:1022–2.
10. Monti S, Tamayo P, Mesirov J, *et al.* Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning* 2003;**52**:91–118.
11. Nguyen T, Tagett R, Diaz D, *et al.* A novel approach for data integration and disease subtyping. *Genome Res* 2017;**27**:2025–39.
12. Brunet JP, Tamayo P, Golub TR, *et al.* Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A* 2004;**101**:4164–9.
13. Witten DM, Tibshirani RJ. Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat Appl Genet Mol Biol* 2009;**8**:Article28.
14. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* 2009;**25**:2906–12.
15. Mo Q, Wang S, Seshan VE, *et al.* Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc Natl Acad Sci U S A* 2013;**110**:4245–50.
16. Mo Q, Shen R, Guo C, *et al.* A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics (Oxford, England)* 2018;**19**:71–86.
17. Cai Z, Poulos RC, Liu J, *et al.* Machine learning for multi-omics data integration in cancer. *iScience* 2022;**25**:103798.
18. Chaudhary K, Poirion OB, Lu L, *et al.* Deep learning-based multi-omics integration robustly predicts survival in liver cancer. *Clin Cancer Res* 2018;**24**:1248–59.
19. Ronen J, Hayat S, Akalin A. Evaluation of colorectal cancer subtypes and cell lines using deep learning. *Life Science Alliance* 2019;**2**:e201900517.
20. Guo LY, Wu AH, Wang YX, *et al.* Deep learning-based ovarian cancer subtypes identification using multi-omics data. *BioData Mining* 2020;**13**:10.
21. Zhang L, Lv C, Jin Y, *et al.* Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk Neuroblastoma. *Front Genet* 2018;**9**:477.
22. Zhao Z, Li Y, Wu Y, *et al.* Deep learning-based model for predicting progression in patients with head and neck squamous cell carcinoma. *Cancer Biomark* 2020;**27**:19–28.
23. Xu J, Wu P, Chen Y, *et al.* A hierarchical integration deep flexible neural forest framework for cancer subtype classification by integrating multi-omics data. *BMC bioinformatics* 2019;**20**:527–7.
24. Zhang C, Chen Y, Zeng T, *et al.* Deep latent space fusion for adaptive representation of heterogeneous multi-omics data. *Brief Bioinform* 2022;**23**:bbab600.
25. Rong Z, Liu Z, Song J, *et al.* MCluster-VAEs: an end-to-end variational deep learning-based clustering method for subtype discovery using multi-omics data. *Comput Biol Med* 2022;**150**:106085.
26. Yang H, Chen R, Li D, *et al.* Subtype-GAN: a deep learning approach for integrative cancer subtyping of multi-omics data. *Bioinformatics* 2021;**37**:2231–7.
27. Zhang Y, Kiryu H. MODEC: an unsupervised clustering method integrating omics data for identifying cancer subtypes. *Brief Bioinform* 2022;**23**:bbac372.
28. Song W, Wang W, Dai D-Q. Subtype-WESLR: identifying cancer subtype with weighted ensemble sparse latent representation of multi-view data. *Brief Bioinform* 2021;**23**:bbab398.
29. Yang Y, Tian S, Qiu Y, *et al.* MDICC: novel method for multi-omics data integration and cancer subtype identification. *Brief Bioinform* 2022;**23**:bbac132.
30. Yang B, Yang Y, Su XP. Deep structure integrative representation of multi-omics data for cancer subtyping. *Bioinformatics* 2022;**38**:3337–42.
31. Yang B, Xin TT, Pang SM, *et al.* Deep subspace mutual learning for cancer subtypes prediction. *Bioinformatics* 2021;**37**:3715–22.
32. Chen R, Yang L, Goodison S, *et al.* Deep-learning approach to identifying cancer subtypes using high-dimensional genomic data. *Bioinformatics* 2020;**36**:1476–83.

33. Moon S, Lee H. MOMA: a multi-task attention learning algorithm for multi-omics data interpretation and classification. *Bioinformatics* 2022;**38**:2287–96.
34. Poirion OB, Jing Z, Chaudhary K, *et al.* DeepProg: an ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data. *Genome Med* 2021;**13**:112.
35. Pfeifer B, Schimek MG. A hierarchical clustering and data fusion approach for disease subtype discovery. *J Biomed Inform* 2021;**113**:103636.
36. Liang C, Shang MC, Luo JW. Cancer subtype identification by consensus guided graph autoencoders. *Bioinformatics* 2021;**37**: 4779–86.
37. Wang B, Mezlini AM, Demir F, *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods* 2014;**11**: 333–7.
38. Rappoport N, Shamir R. NEMO: cancer subtyping by integration of partial multi-omic data. *Bioinformatics* 2019;**35**:3348–56.
39. Xu H, Gao L, Huang M, *et al.* A network embedding based method for partial multi-omics integration in cancer subtyping. *Methods* 2021;**192**:67–76.
40. Min E, Guo X, Liu Q, *et al.* A survey of clustering with deep learning: from the perspective of network architecture. *IEEE Access* 2018;**6**:39501–14.
41. Yeh C-H, Hong C-Y, Hsu Y-C, *et al.* Decoupled contrastive learning. 2021; arXiv:2110.06848.
42. Grill J-B, Strub F, Altché F, *et al.* Bootstrap your own latent: a new approach to self-supervised learning. 2020; arXiv:2006.07733.
43. Li J, Zhou P, Xiong C, *et al.* Prototypical contrastive learning of unsupervised representations. 2020; arXiv:2005.04966.
44. Hadsell R, Chopra S, LeCun Y. Dimensionality Reduction by Learning an Invariant Mapping. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. 2006, p. 1735-42.
45. Li Y, Hu P, Liu Z, *et al.* Contrastive clustering, proceedings of the AAAI conference on. *Artificial Intelligence* 2021;**35**:8547–55.
46. Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Res* 2018;**46**:10546–62.
47. Mukhopadhyay P, Ye JB, Anderson KM, *et al.* Log-rank test vs MaxCombo and difference in restricted mean survival time tests for comparing survival under nonproportional hazards in Immuno-oncology trials a systematic review and meta-analysis. *JAMA Oncol* 2022;**8**:1294–300.
48. Hu W, Miyato T, Tokui S, *et al.* Learning discrete representations via information maximizing self-augmented training. 2017; arXiv:1702.08720.
49. Glorot X, Bordes A, Bengio Y. Deep Sparse Rectifier Neural Networks. In: Geoffrey G, David D, Miroslav D (eds). *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research: PMLR*, 2011, 315–23.
50. Xu H, Zheng XN, Zhang SY, *et al.* Tumor antigens and immune subtypes guided mRNA vaccine development for kidney renal clear cell carcinoma. *Mol Cancer* 2021;**20**:20.
51. Berardini TZ, Li DH, Huala E, *et al.* The gene ontology in 2010: extensions and refinements the gene ontology consortium. *Nucleic Acids Res* 2010;**38**:D331–5.
52. Ogata H, Goto S, Sato K, *et al.* KEGG: Kyoto Encyclopedia of genes and genomes. *Nucleic Acids Res* 1999;**27**:29–34.
53. Yu G, Wang LG, Han Y, *et al.* clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics* 2012;**16**:284–7.
54. Ren X, Kuan PF. methylGSA: a Bioconductor package and shiny app for DNA methylation data length bias adjustment in gene set testing. *Bioinformatics* 2018;**35**:1958–9.
55. Vlachos IS, Zagganas K, Paraskevopoulou MD, *et al.* DIANA-miRPath v3.0: deciphering microRNA function with experimental support. *Nucleic Acids Res* 2015;**43**:W460–6.
56. Sellitti DF, Doi SQ. MicroRNAs in renal cell carcinoma. *Microrna* 2015;**4**:26–35.
57. Khordadmehr M, Shahbazi R, Sadreddini S, *et al.* miR-193: a new weapon against cancer. *J Cell Physiol* 2019;**234**:16861–72.
58. Trevisani F, Ghidini M, Larcher A, *et al.* MicroRNA 193b-3p as a predictive biomarker of chronic kidney disease in patients undergoing radical nephrectomy for renal cell carcinoma. *Br J Cancer* 2016;**115**:1343–50.
59. Wang C, Uemura M, Tomiyama E, *et al.* MicroRNA-92b-3p is a prognostic oncomiR that targets TSC1 in clear cell renal cell carcinoma. *Cancer Sci* 2020;**111**:1146–55.
60. Liu Y, Yin B, Zhang C, *et al.* Hsa-let-7a functions as a tumor suppressor in renal cell carcinoma cell lines by targeting c-myc. *Biochem Biophys Res Commun* 2012;**417**:371–5.