Quantitative Biology 2021, 9(4): 451–462
https://doi.org/10.15302/J-QB-021-0261

METHOD

Adaptive total variation constraint hypergraph regularized NMF for single-cell RNA-seq data analysis

Ya-Li Zhu, Xiao-Ning Zhang, Chuan-Yuan Wang, Jin-Xing Liu, Xiang-Zhen Kong*

School of Computer Science, Qufu Normal University, Rizhao 276826, China
* Correspondence: kongxzhen@163.com

Received October 28, 2020; Revised January 1, 2021; Accepted January 26, 2021

Background: **Single-cell RNA sequencing (scRNA-seq) data provides a whole new view to study disease and cell differentiation development. With the explosive increment of scRNA-seq data, effective models are demanded for mining the intrinsic biological information.**
Methods**: This paper proposes a novel non-negative matrix factorization (NMF) method for clustering and gene co-expression network analysis, termed Adaptive Total Variation Constraint Hypergraph Regularized NMF (ATV-HNMF). ATV-HNMF can adaptively select the different schemes to denoise the cluster or preserve the cluster boundary information between clusters based on the gradient information. Besides, ATV-HNMF incorporates hypergraph regularization, which can consider high-order relationships between cells to reserve the intrinsic structure of the space.**
Results: **Experiments show that the performances on clustering outperform other compared methods, and the network construction results are consistent with previous studies, which illustrate that our model is effective and useful.**
Conclusion: **From the clustering results, we can see that ATV-HNMF outperforms other methods, which can help us to understand the heterogeneity. We can discover many disease-related genes from the constructed network, and some are worthy of further clinical exploration.**

Keywords: adaptive total variation; single-cell RNA sequencing; network analysis; nonnegative matrix factorization; hypergraph

Author summary: Single-cell RNA sequencing techniques are helpful for researchers to study the development of disease and cell differentiation. We propose a novel non-negative matrix factorization (NMF) method called Adaptive Total Variation Constraint Hypergraph Regularized NMF (ATV-HNMF), which incorporates hypergraph regularization and adaptive total variation schemes. The results of clustering and gene co-expression network construction show that our model is effective and useful.

INTRODUCTION

With the advancement of scRNA-seq technology, researchers can separate individual cells from each other and sequence the transcriptome data at the individual cell level, which can provide a deeper insight into the biological process [1–3]. The technology has been used on diverse organs and cells to find out the important biological signals about diseases and cancers, such as islet cells [4], Lung Epithelial cells [5], neuronal cells [6], glioblastoma [7], and other use cases. It also can be used to identify the types in differentiating embryonic stem cells [8]. ScRNA-seq data profoundly influence us to understand the complexity, variety, and irregularity of cellular biological activity. While the technology has great potential to explore the gene expression of cells,

they also present new challenges that require advanced algorithm tools to extract potential biological information.

There are many methods to analyze single-cell sequencing data. T-distributed stochastic neighbor embedding (t-SNE) [9] is based on the local similarity that places the similar cells clustered together and separates different cells from each other, but the relative position used is not always meaningful. Principle component analysis (PCA) [10] is also popular to cluster the scRNA-seq data. Spectral clustering (SC) [11] uses the similarity graph to cluster single-cells. Its variant sparse spectral clustering (SSC) [11] adds the sparse regularization on low-dimensional matrix to obtain block diagonal, which can have better performance. Wang *et al*. [12] proposed the multi-kernel learning method for single-cell interpretation called SIMLR, which can learn from 55 Gaussian kernels to get the final results. Moreover, nonnegative matrix factorization (NMF) is also an efficient method that can be used to analyze single-cell data.

NMF has been widely used in the field of bioinformatics in recent years [13–16]. It is less influenced by initial conditions of the scRNA-seq data and can discover the patterns of a set of genes in differentiating individual cells [17]. Based on this, a lot of mature models have been proposed. In [18], a dimensionality reduction method based on manifold learning was proposed for cellular space digging on single-cell data. Cai *et al*. [19] proposed graph regularized NMF (GNMF) that considers the intrinsic manifold of the data space, which is practical to actual applications. Zeng *et al*. [20] proposed hypergraph regularized NMF (HNMF), extending the simple graph on two samples to higher-order relationships, considering multiple sample interactions. However, these methods do not deal with noise and outliers very well. Rudin *et al*. [21] proposed the total variation (TV), which effectively removes noise or keeps the detailed information of boundary. Though it has been widely used for denoising, it cannot preserve the feature structure and choose the proper total variation (TV) regularization parameter. Recently, Leng *et al*. [22] presented an adaptive total variation constrained non-negative matrix factorization on manifold (ATV-NMF), which considers the total variation and inner geometric structure simultaneously. However, the manifold regularization only considers the pairwise relationship between samples.

Based on the above problems, we consider introducing adaptive total variation (ATV) and hypergraph regularization into the NMF model, named ATV-HNMF. The ATV term can adaptively select different schemes to deal with data noise, making the model more robust. To retain the inherent high-order geometric structure, the hypergraph regularization term is incorporated to consider the complex relationship of more than two samples.

The main contributions of our work are listed as follows:

1. The incorporation of hypergraph regularization term retains the intrinsic geometric structure, which considers the high-order relationships among more than two samples. Unlike the simple graph with pairwise relationships, a hyperedge includes a set of samples, thus preserving the high-order geometry of the sample space, leading to a deeper understanding of the cells.

2. Adding adaptive total variation term to denoise or preserve the detailed information of the boundary. Based on the gradient information, different schemes are adaptively selected to denoise or retain the details. The adaptive total variation term makes the model more effective in processing single-cell data.

The rest of the paper is organized as follows: we present experimental results about clustering and network analysis in Section of Results; in Section of Discussion, we conclude the paper; a brief review of related work and the algorithm model of ATV-HNMF are presented in Section of Materials and methods.

RESULTS

Datasets

The seven detailed single-cell data are downloaded from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/), including Islet [4], Lung Epithelial [5], Darmanis [23], Goolam [24], Treutlein [25], Grover [26], Breton [27]. The statistics for datasets are summarized in Table 1.

Table 1 Data information about single-cell datasets

Datasets	The Number of genes	The Number of cells	Cell types	Species
Islet	39851	1600	4	*Homo sapiens*
Lung Epithelial	34816	540	2	*Homo sapiens*
Darmanis	22085	420	8	*Homo sapiens*
Goolam	40315	124	5	*Mus musculus*
Treutlein	959	80	5	*Mus musculus*
Grover	14739	135	2	*Mus musculus*
Breton	20689	957	4	*Homo sapiens*

In ATV-HNMF, we filter the gene and normalize the data matrix as preprocessing steps. We delete genes whose expression (gene expression value is non-zero) is less than 5% of the total number of cells for the gene filtering step. Since the expression values of different genes vary vastly, the L_2-norm is used to eliminate scale differences between samples in the normalize step.

Performance comparisons

Parameter settings

This paper introduces the ATV regularization term to denoise the cluster or preserve the cluster boundary information between clusters. Fortunately, we do not have to set parameters on it. For all methods, we set the dimensionality reduction parameter k that equals to the number of cell types. For GNMF and ATV-HNMF, we use 0-1 weight scheme to construct K-NN graph, in which K is set to 5 as default. We vary the parameter λ from $\{0.01, 0.1, 1, 10, 100, 1000\}$, the results shown in Fig. 1. From Fig. 1, we can discover that when λ is too large, the accuracy is extremely low. Therefore, we do not recommend setting it to a larger value. Especially, the parameter λ is not sensitive to Lung Epithelial dataset, so our method can achieve robust performance. We choose the parameters corresponding to their best performance for all datasets.

Clustering results and analysis

Since t-SNE, Kmeans, SSC, SC, NMF, ATV-NMF, and our proposed method ATV-HNMF have unstable clustering results, we run all the algorithms 30 times and choose their average value as the final result. Using the average value can reduce the effects of random initialization. Tables 2 and 3 list the clustering results about eight comparative methods, and we can discover that:

1. In most cases, our method can achieve the best performance. Compared to NMF, we take the graph regularization and ATV scheme into consideration. T-SNE, Kmeans, SSC, and SC directly cluster the samples without dimension reduction, so their performance is not as good as NMF. SIMLR is an algorithm specifically designed to analyze single-cells, it considers the similarity between cells; however, it does not consider the inner geometric structure. ATV-HNMF considers the ATV scheme to adaptively chooses the different schemes to denoise the cluster or preserve the cluster boundary information between clusters, so our method has better clustering performance.

2. ATV-HNMF does not always have the best performance on NMI among the seven datasets; for dataset Islet and Lung Epithelial, NMF and t-SNE have the highest NMI performance. Taking all datasets as a whole, ATV-HNMF has the best clustering performance of the average value, and it outperforms the compared methods of ARI value 4.8% and NMI value 1.3%, respectively.

To observe the clustering results more intuitively, we further use the coefficient matrix **F** to visualize the results. T-SNE is a popular dimensionality reduction method, and it has become a mainstream visualization tool [12]. To show the clustering results in two-dimensional space, firstly, we use the ATV-HNMF model to reduce the original data matrix to two dimensions to obtain the coefficient matrix with size $2 \times n$. Then, Pearson Correlation Coefficient (PCC) is applied to compute the similarity between samples. Finally, t-SNE is used to

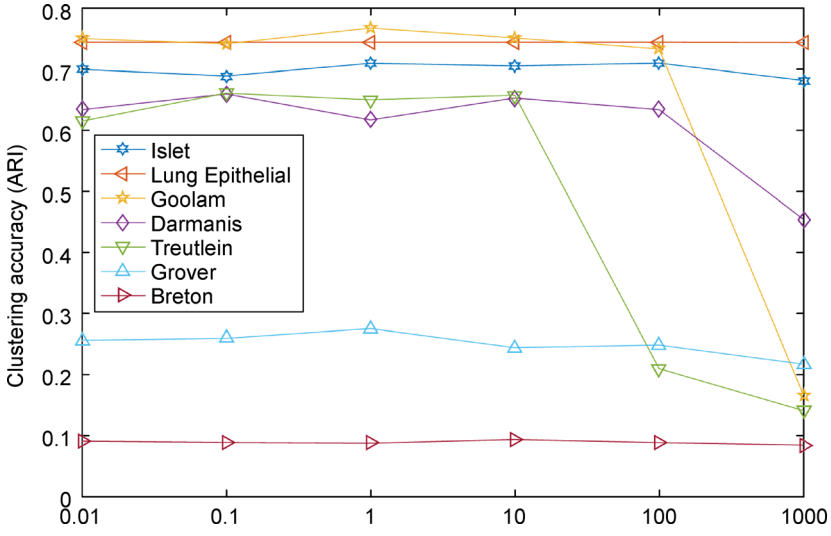

Figure 1. The impact of parameter λ on seven datasets.

Table 2 ARI results on single-cell datasets

Dataset	t-SNE	PCA	Kmeans	SSC	SC	SIMLR	NMF	ATV-HNMF
Islet	0.5975	0.6008	0.601	0.5115	0.7081	0.0534	0.68	0.7093
Lung Epithelial	0.6031	0.5822	0.5757	0.6125	0.5449	0.0714	0.7137	0.7437
Darmanis	0.5225	0.4594	0.4182	0.4441	0.4445	0.2991	0.7305	0.7617
Goolam	0.5725	0.3278	0.3453	0.5202	0.5258	0.3982	0.5904	0.6593
Treutlein	0.5473	0.5727	0.6172	0.5242	0.6191	0.5114	0.5297	0.6573
Grover	0.2712	0.2712	0.2712	0.1849	0.2261	0.0946	0.2336	0.2749
Breton	0.0902	0.0396	0.0411	0.0397	0.0686	0.0442	0.0801	0.0931
Average	0.4577	0.4076	0.4099	0.4053	0.4481	0.2103	0.5082	0.5570

Table 3 NMI results on single-cell datasets

Dataset	t-SNE	PCA	Kmeans	SSC	SC	SIMLR	NMF	ATV-HNMF
Islet	0.3843	0.3864	0.3870	0.3174	0.5592	0.0678	0.7685	0.7247
Lung Epithelial	0.7183	0.6741	0.6617	0.7115	0.6255	0.0567	0.5977	0.6320
Darmanis	0.7043	0.6253	0.5686	0.584	0.591	0.5693	0.7765	0.795
Goolam	0.6021	0.4445	0.4654	0.5871	0.5826	0.6071	0.7279	0.7403
Treutlein	0.7346	0.7530	0.7157	0.7088	0.8196	0.6881	0.7018	0.7558
Grover	0.2197	0.2125	0.2080	0.1382	0.1717	0.0711	0.1835	0.2108
Breton	0.2370	0.1467	0.1486	0.1672	0.1771	0.2097	0.2034	0.1939
Average	0.5143	0.4632	0.4507	0.4591	0.5038	0.3242	0.5656	0.5789

project the obtained similarity matrix into the two-dimensional space. We selected five involved datasets to show the results; Figure 2 shows the final results, and it was evident that our method can clearly separate different types of cells.

Network construction and mining

To illustrate the effectiveness of the ATV-HNMF method, we use the selected 1000 genes to construct the network and mine the genetic information. We use PCC to

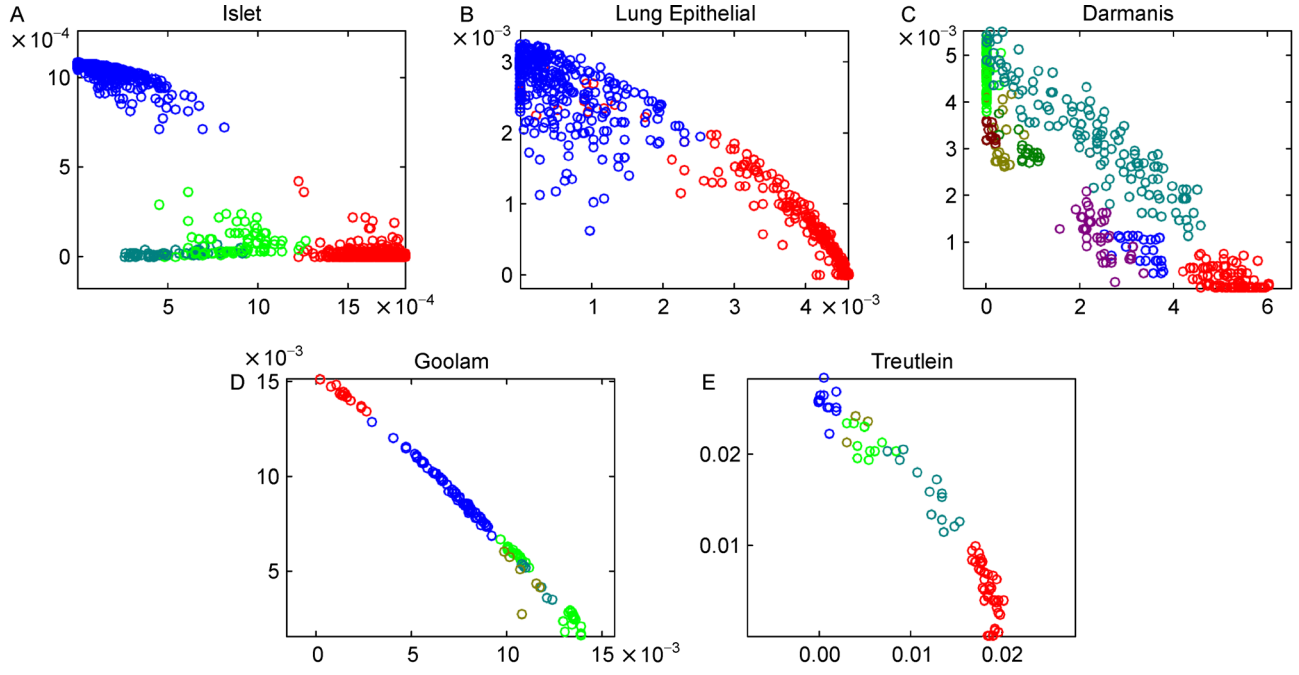

Figure 2. The clustering result graph of the five single-cell datasets. The different color represents the different cell types.

calculate the relationship between two nodes to obtain the adjacency matrix, and then sort the absolute PCC values to perform curve fitting, finally choose the inflection point as the threshold to construct the network. Betweenness indicates the role of the node in the interconnection of others. The higher the betweenness value, the more significant node that maintains the tightness of the network; therefore, the betweenness is selected as the metric to evaluate the importance of gene nodes.

Due to space limitations, we select Islet and Lung Epithelial datasets as instances to show the network construction results. If there are too few nodes in the network module, less information can be mined. Therefore, we set 20 nodes as the baseline to reserve modules with more gene nodes. Figures 3 and 4 (visualized by Cytoscape [28]) show the network constructed by ATV-HNMF. Larger size indicates a greater degree of node, while a darker node indicates a greater betweenness. The nodes with higher scores are more important genes and worth mining.

The genes with higher scores can be considered suspicious genes; we select the top 10 nodes and make annotations on Genecards (http://www.genecards.org/); the results are shown in Table 4. Many disease-related genes were found and shown in bold in Table 4. HIF1A is related to type 2 diabetes (T2D), and it keeps a consistently high expression level in transcriptional activity [29]. GPX1 plays a dual role in insulin synthesis, secretion, and signal transduction by regulating redox homeostasis. The overexpression of GPX1 is associated with the elevated protein level of SELENOT, which may partially affect the T2D phenotype [30]. The expression of EIF6 regulates the amount of histone acetylation and fatty acid synthase mRNA, so EIF6 may be a therapeutic target to the fasn-driven lipogenesis in T2D [31]. ATP6V1H is down-regulated in islet of T2D [32]; it plays a significant role in the regulation of vacuolar-ATPase activity and may be involved in the development of an important molecular mechanism of T2D [33]. The non-coding region of RPS19 may conduce to the pathogenesis of Diamond-Blackfan anemia (DBA) by regulating the expression level of RPS19 protein [34]. Liu *et al*. reveal the role of the RPL11-MDM2-p53 pathway in fat storage during periods of overnutrition, and targeting this pathway might be a potential obesity treatment [35].

The expression of SDR16C5 is significantly increased in six subtypes in idiopathic interstitial pneumonias (IIPs), and the researchers speculate that they play a vital role in all subgroups of IIP and worth further research [36]. ARPP19 is highly expressed in embryonic tissues. And the overexpression of ARPP19 may be related to cellular malignancies because it regulates mitosis by inhibiting protein phosphatase-2A [37,38]. The existence of CXCL5 contributes to the proliferation and invasion of prostate cancer and squamous cell carcinoma [39,40]. Patients with idiopathic pulmonary fibrosis have an increased risk of developing cancer. BANF1 is a marker of lung cancer cells [41]. USP22 induces epithelial-mesenchymal transformation of lung adenocarcinoma, and it promotes tumor progression [42]. There are still high-ranking genes such as TMEM14B, RPS16, ARPP19 that have not been clinically verified, which need further exploration.

DISCUSSION

This paper proposes a non-negative matrix factorization model for clustering and network analysis, called ATV-HNMF. On the one hand, this model introduces the high-order relationships between cells that can keep the intrinsic structure high-dimensional. On the other hand, adaptive total variation is used for reducing noise from interference. From the clustering results, we can see that ATV-HNMF outperforms other methods. We can discover many disease-related genes from the constructed network, and some are worthy of further clinical exploration. In the future, we will consider reducing the loss of error terms to make the model more robust.

MATERIALS AND METHODS

Related work

Adaptive total variation

ATV-HNMF is based on the idea of ATV, which is first proposed by Stacey Levine *et al*. [43]. The data can adaptively select an anisotropic scheme to preserve the cluster boundary information or denoise based on the gradient information. The term is defined as follows:

$$E(\mathbf{F}) = \|\mathbf{F}\|_{ATV}, \tag{1}$$

where E is the function of energy, and the size of $\mathbf{F}$ is $k \times n$. The ATV regularization scheme $\|\mathbf{F}\|_{ATV}$ is defined as $\int_{\Omega}\left(1/p(x,y)\right)|\nabla\mathbf{F}|^{p(x,y)}dxdy$, $p(x,y) = 1 + 1/(1+|\nabla\mathbf{F}|^2)$, $1 < p(x,y) < 2$, where $\Omega \subset \mathbb{R}^n$ denotes the data space. The discrete gradient form of $(\nabla\mathbf{F})(i,j) = ((\partial_x\mathbf{F}) \times (i,j), (\partial_y\mathbf{F})(i,j))$ is given as follows:

$$(\partial_x\mathbf{F})(i,j) = \begin{cases} \mathbf{F}(i+1,j) - \mathbf{F}(i,j) & \text{if } i < k \\ \mathbf{F}(1,j) - \mathbf{F}(k,j) & \text{if } i = k \end{cases}, \tag{2}$$

$$(\partial_y\mathbf{F})(i,j) = \begin{cases} \mathbf{F}(i,j+1) - \mathbf{F}(i,j) & \text{if } i < n \\ \mathbf{F}(i,1) - \mathbf{F}(i,n) & \text{if } i = n \end{cases}. \tag{3}$$

The adaptive total variation regularization term contains a parameter expressed as $1/|\nabla\mathbf{F}|^{2-p}$ in Eq. (13) to control

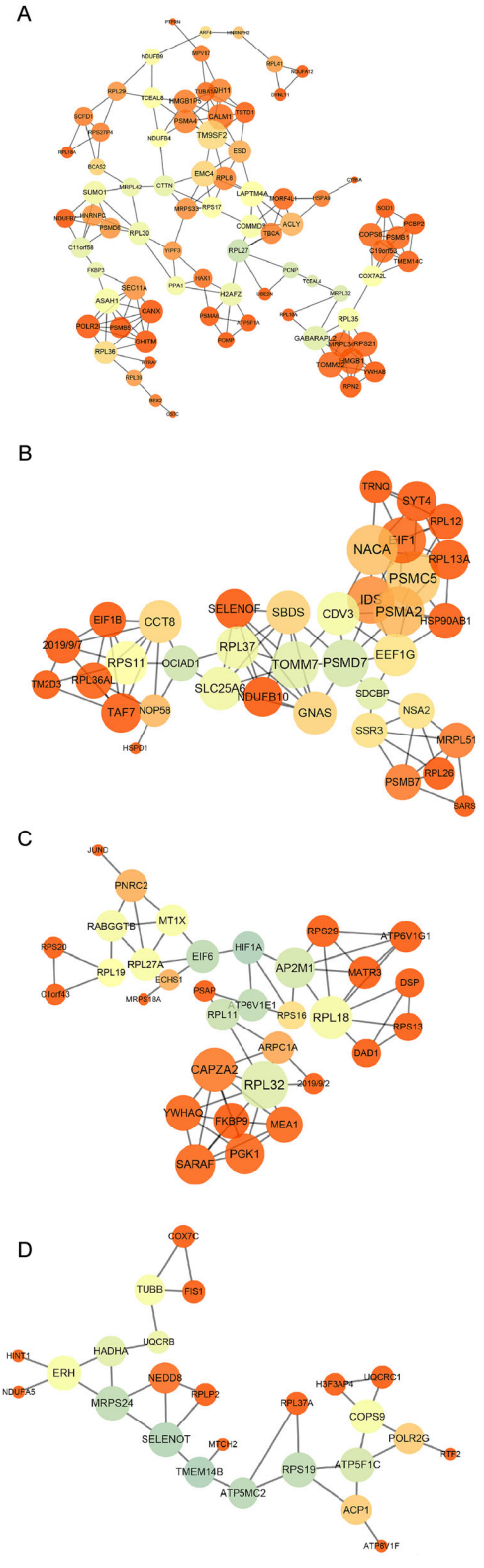

Figure 3. Network construction based on human islet cells. Larger size indicates a greater degree of node, while a darker node indicates a greater betweenness.

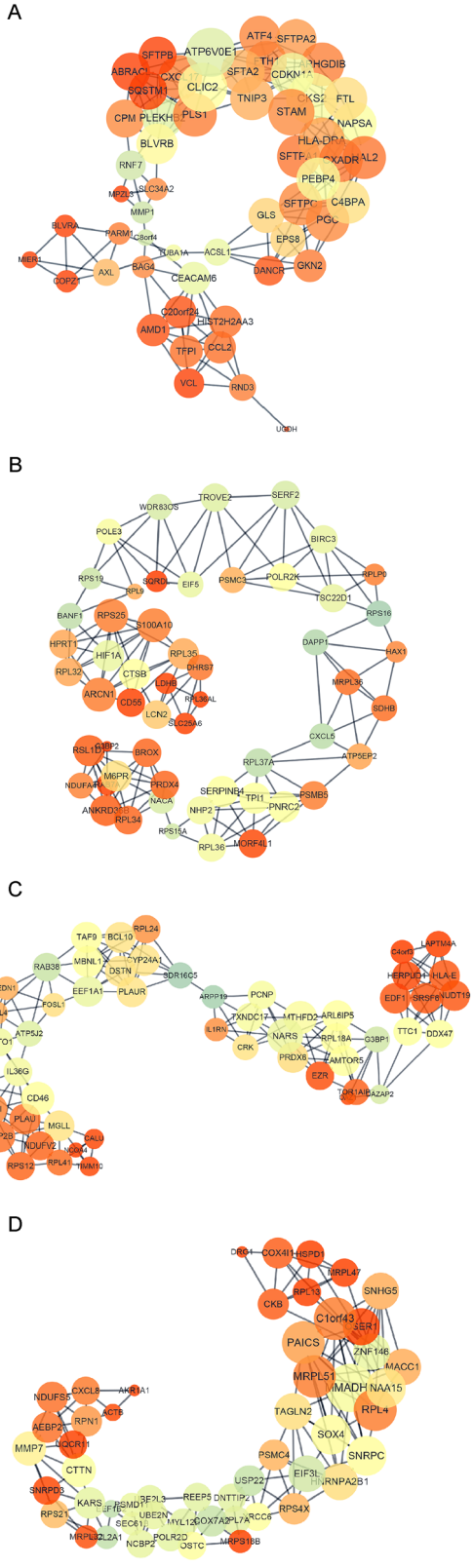

Figure 4. Network construction based on Lung Epithelial cells. Larger size indicates a greater degree of node, while a darker node indicates a greater betweenness.

Table 4 Detailed information on the top ten selected genes

Islet		Lung Epithelial	
Gene	Annotations	Gene	Annotations
HIF1A	This gene encodes the alpha subunit of transcription factor hypoxia-inducible factor-1 (HIF-1)	**SDR16C5**	Diseases associated with SDR16C5 include psoriasis
TMEM14B	Protein coding gene	RPS16	Protein coding gene
SELENOT	This gene encodes a seleno protein, containing a seleno cysteine (Sec) residue at the active site	**ARPP19**	Among its related pathways are cell cycle, mitotic and DNA damage
ATP5MC2	Among its related pathways are respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins	DAPP1	Protein coding gene
EIF6	Diseases associated with EIF6 include pyloric atresiaand shwachman-diamond syndrome 1	**CXCL5**	Diseases associated with CXCL5 include pediatric ulcerative colitis and pulmonary sarcoidosis
ATP6V1E1	This gene encodes a component of vacuolar ATPase (V-ATPase)	RPL37A	Among its related pathways are Viral mRNA translation and influenza viral RNA transcription and replication
MRPS24	Among its related pathways are mitochondrial translation and organelle biogenesis and maintenance	EEF1B2	This gene encodes a translation elongation factor
RPS19	Protein coding gene	**BANF1**	Among its related pathways are cell cycle, mitotic and HIV life cycle
RPL11	Diseases associated with RPL11 include diamond-blackfan anemia 7 and diamond-blackfan anemia	BCL2A1	This gene encodes a member of the BCL-2 protein family
RPL27	Among its related pathways are viral mRNA translation and influenza viral RNA transcription and replication	**USP22**	Protein coding gene

the diffusion speed of the different directions. In the boundary part of the cluster, the value of $|\nabla \mathbf{F}|^{2-p}$ is large, and the diffusion coefficient is small, so the diffusion speed is slow, which can keep the difference from cluster to cluster. Between clusters, based on the small changes of gradient information, the value of $1/|\nabla \mathbf{F}|^{2-p}$ is large so that the diffusion is strong, which helps denoise and keeps samples between clusters tighter.

Hypergraph theory

The establishment of hypergraph theory is based on the simple graph concept. In a simple graph, the relationships are between two vertices, the edge is connected by the two vertices. In a hypergraph, the relationship is between two or more vertices, so the hypergraph is composed of many hyperedges and one hyperedge is composed of many vertices [44,45]. In reality, interactions between more than two samples are more critical to preserve the geometrical structure.

The hypergraph is denoted as $G=(V,E,\mathbf{W})$, where V denotes all the samples, E is the family of e that $\cup_{e\in E}=V$, $\mathbf{W}$ is a diagonal matrix with hyperedge weights elements $w(e)$. The incidence matrix is defined as follows:

$$\mathbf{H}(v,e)=\begin{cases}1, & \text{if } v\in e\\ 0, & \text{if } v\notin e\end{cases}. \tag{4}$$

Based on $\mathbf{H}$, the degree of each vertex is defined as

$$d(v)=\sum_{e\in E} w(e)\mathbf{H}(v,e). \tag{5}$$

And the degree of each hyperedge is given by

$$\delta(e)=\sum_{v\in V}\mathbf{H}(v,e). \tag{6}$$

We use the diagonal matrices $\mathbf{D}_e$ and $\mathbf{D}_v$ to denote the vertices degree and hyperedges degree respectively. As for the hypergraph Laplacian matrix, it is denoted as

$$\mathbf{L}=\mathbf{D}_v-\mathbf{A}, \tag{7}$$

where $\mathbf{A}=\mathbf{H}\mathbf{W}\mathbf{D}_e^{-1}\mathbf{H}^T$.

The affinity matrix of vertices is defined as

$$\mathbf{S}_{ij}=\exp\left(-\frac{\|\mathbf{v}_i-\mathbf{v}_j\|^2}{\sigma}\right), \tag{8}$$

and the weight of each hyperedge is given by

$$w(e_j)=\sum_{v_i\in e_j}\mathbf{S}_{ij}. \tag{9}$$

Method

The methodology of ATV-HNMF

Assuming that the input matrix $\mathbf{X}=(\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_n)\in\mathbb{R}^{m\times n}$,

each row of $\mathbf{X}$ indicates the expression level of a gene among all single cells, each column of $\mathbf{X}$ indicates all gene expression levels in a single cell. The task of ATV-HNMF aims to factorize the original data into two matrixes $\mathbf{U}=(\mathbf{u}_1,\mathbf{u}_2,...,\mathbf{u}_k)\in\mathbb{R}^{m\times k}$ and $\mathbf{F}=(\mathbf{f}_1,\mathbf{f}_2,...,\mathbf{f}_n)\in\mathbb{R}^{k\times n}$, so as to get the approximation $\mathbf{X}\simeq\mathbf{UF}$. The objective function is designed as follows:

$$\min_{\mathbf{U},\mathbf{F}} O=\|\mathbf{X}-\mathbf{UF}\|_F^2+\lambda Tr(\mathbf{FLF}^T)+2\|\mathbf{F}\|_{ATV}$$
$$s.t.\ \ \mathbf{U}\geqslant 0,\mathbf{F}\geqslant 0, \tag{10}$$

where $\|\cdot\|_{\mathbf{F}}$ denotes the Frobenius norm to indicate the error term, $Tr(\cdot)$ represents the trace of the matrix, λ is the parameter of hypergraph regularization term $(\lambda>0)$.

The framework of our method is clearly shown in Fig. 5. We decompose the data matrix into two low-dimensional matrices, then we use the basis matrix $\mathbf{U}$ to mine the related genes by network construction and use the coefficient matrix $\mathbf{F}$ to process sample clustering.

Note that $\|\mathbf{A}\|_F^2=Tr(\mathbf{AA}^T)$. Eq. (10) can be rewritten as follows:

$$\min_{\mathbf{U},\mathbf{F}} O=Tr((\mathbf{X}-\mathbf{UF})(\mathbf{X}-\mathbf{UF})^T)+\lambda Tr(\mathbf{FLF}^T)+2\|\mathbf{F}\|_{ATV}$$
$$s.t.\ \ \mathbf{U}\geqslant 0,\mathbf{F}\geqslant 0. \tag{11}$$

The multiplicative iterative updating rules of $\mathbf{U}$ and $\mathbf{F}$ are obtained as follows:

$$\mathbf{U}_{i,r}\leftarrow\mathbf{U}_{i,r}\frac{(\mathbf{XF}^T)_{i,r}}{(\mathbf{UFF}^T)_{i,r}}, \tag{12}$$

$$\mathbf{F}_{r,j}\leftarrow\mathbf{F}_{r,j}\frac{\left(\mathbf{U}^T\mathbf{X}+\lambda\mathbf{FA}+div\left(\frac{\nabla\mathbf{F}}{|\nabla\mathbf{F}|^{2-p}}\right)\right)_{r,j}}{(\mathbf{U}^T\mathbf{UF}+\lambda\mathbf{FL})_{r,j}}, \tag{13}$$

where $div=\left(\frac{\partial}{\partial_x},\frac{\partial}{\partial_y}\right)$ denotes divergence. $\nabla\mathbf{F}=(\partial_x\mathbf{F},\partial_y\mathbf{F})$ indicates the gradient information, and $|\nabla\mathbf{F}|=\sqrt{(\partial_x\mathbf{F})^2+(\partial_y\mathbf{F})^2}$ is the gradient norm.

In summary, the algorithm of ATV-HNMF is shown as follows:

Algorithm: ATV-HNMF

Input: $\mathbf{X}$, parameter λ

Output: $\mathbf{U}\in\mathbb{R}^{m\times k}$, $\mathbf{F}\in\mathbb{R}^{k\times n}$

Construct hypergraph Laplacian matrix $\mathbf{L}\in\mathbb{R}^{n\times n}$

Initialize: $\mathbf{U}\geqslant 0$, $\mathbf{F}\geqslant 0$

Repeat

Update $\mathbf{U}$ by Eq. (12)

Update $\mathbf{F}$ by Eq. (13)

Until convergence

Convergence analysis

Definition 1: If a function $A(x,x')$ satisfies $J(x)\leqslant A(x,x')$ and $J(x)=A(x,x)$, we suppose $A(x,x')$ is an auxiliary function of $J(x)$.

Lemma 1. If A is an auxiliary function of J, the J is

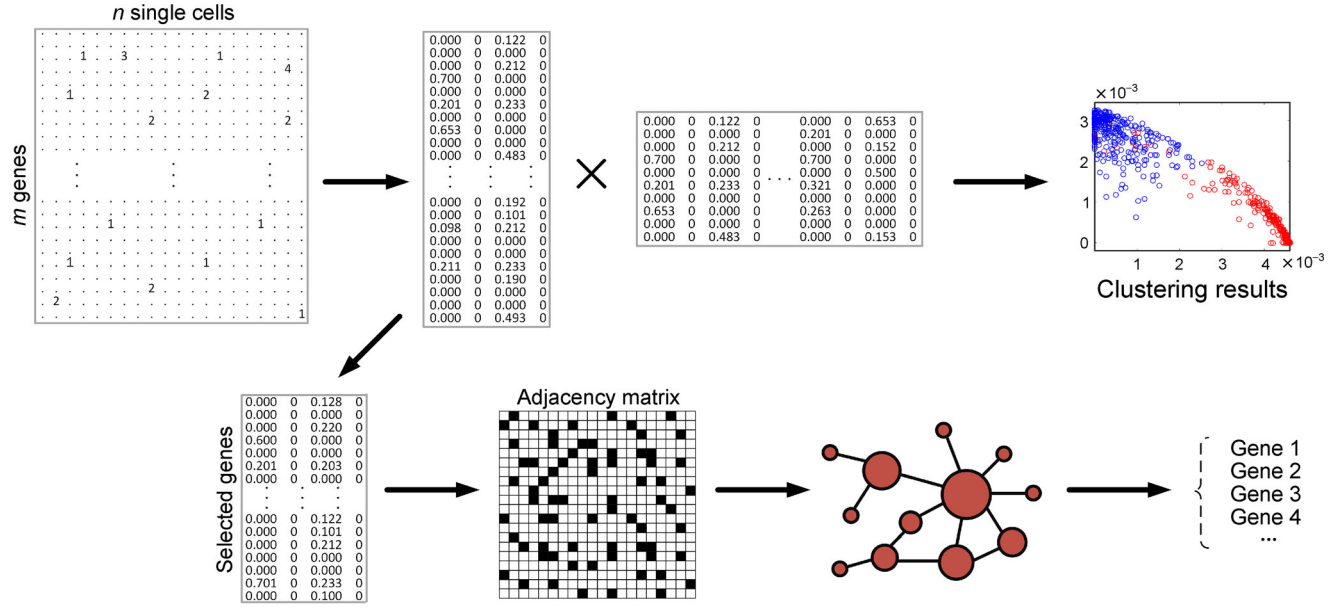

Figure 5. The framework of the method ATV-HNMF. The matrix $\mathbf{X}$ is decomposed into a basis matrix $\mathbf{U}$ and a coefficient matrix $\mathbf{F}$. The matrix $\mathbf{U}$ is selected for gene co-expression network construction and analysis; the matrix $\mathbf{F}$ is used for cell clustering.

non-increasing according to the below updating rules:

$$x^{t+1} = \arg\ \min_x A(x,x^t). \tag{14}$$

Proof.

$$J(x^{t+1}) = A(x^{t+1},x^{t+1}) \leqslant A(x^{t+1},x^t) \leqslant J(x^t). \tag{15}$$

Let the element of $\mathbf{U}$ be u_{ij}, the auxiliary function only relevant to $\mathbf{U}$ denoted as $J_{u_{ij}}$. The partial derivative of $\mathbf{U}$ is shown as follows:

$$J'_{u_{ij}} = \left(\frac{\partial O_{ATV-HNMF}}{\partial \mathbf{U}}\right)_{ij} = -2(\mathbf{X}\mathbf{F}^T - \mathbf{U}\mathbf{F}\mathbf{F}^T)_{ij}, \tag{16}$$

and

$$J''_{u_{ij}} = \left(\frac{\partial^2 O_{ATV-HNMF}}{\partial \mathbf{U}^2}\right)_{ij} = (2\mathbf{F}\mathbf{F}^T)_{jj}. \tag{17}$$

Lemma 2. If $A(u,u_{ij}^t)$ satisfies

$$A(u,u_{ij}^t) = J_{u_{ij}}(u_{ij}^t) + J'_{u_{ij}}(u_{ij}^t)(u-u_{ij}^t) + \frac{(\mathbf{U}\mathbf{F}\mathbf{F}^T)_{ij}}{u_{ij}^t}(u-u_{ij}^t)^2, \tag{18}$$

then $A(u,u_{ij}^t)$ is one of the auxiliary functions of $J_{u_{ij}}$.

Proof.

It is evident that $J(x^t) = A(x^t,x^t)$, to prove that $A(x,x_{ij}^t) \geqslant J(x)$, we rewrite $J(x)$ to the form of Taylor expansion, it is shown as follows:

$$J_{u_{ij}}(u) = J_{u_{ij}}(u_{ij}^t) + J'_{u_{ij}}(u_{ij}^t)(u-u_{ij}^t) + ((\mathbf{F}\mathbf{F}^T)_{jj})(u-u_{ij}^t)^2, \tag{19}$$

We can see that proving $A(u,u_{ij}^t) \geqslant J(u)$ is equivalent to prove that:

$$\frac{(\mathbf{U}\mathbf{F}\mathbf{F}^T)_{ij}}{u_{ij}^t} \geqslant (\mathbf{F}\mathbf{F}^T)_{jj}. \tag{20}$$

$$(\mathbf{U}\mathbf{F}\mathbf{F}^T)_{ij} = \sum_{r=1}^{k} u_{ik}^t (\mathbf{F}\mathbf{F}^T)_{kj} \geqslant u_{ij}^t (\mathbf{F}\mathbf{F}^T)_{jj}. \tag{21}$$

Thus, we can prove the correctness of Eq. (19), and the detailed information can be found in [22,46].

Theorem 1. In each iteration, under the updating rule of Eqs. (12) and (13), the value of the objective function is non-increasing.

Proof.

According to Lemma 1 and 2, $A(u,u_{ij}^t)$ is the auxiliary function of $J_{u_{ij}}$ so that under the updating rule $u_{ij}^{t+1} = \arg\min_u A(u,u_{ij}^t)$, $J_{u_{ij}}$ is non-increasing.

Taking partial derivative of u and set it to 0, we have:

$$\frac{\partial A(u,u_{ij}^t)}{\partial u} = -2(\mathbf{X}\mathbf{F}^T - \mathbf{U}\mathbf{F}\mathbf{F}^T)_{ij} + \frac{2(\mathbf{U}\mathbf{F}\mathbf{F}^T)_{ij}}{u_{ij}^t}(u-u_{ij}^t) = 0. \tag{22}$$

Then, we have:

$$u = u_{ij}^t \frac{(\mathbf{X}\mathbf{F}^T)_{ij}}{(\mathbf{U}\mathbf{F}\mathbf{F}^T)_{ij}}. \tag{23}$$

Considering Eq. (14), we can conclude that

$$u_{ij}^{t+1} = \arg\min_u A(u,u_{ij}^t) = u_{ij}^t \frac{(\mathbf{X}\mathbf{F}^T)_{ij}}{(\mathbf{U}\mathbf{F}\mathbf{F}^T)_{ij}}. \tag{24}$$

Similarly, we can get:

$$f_{ij}^{t+1} = \arg\min_f A(f,f_{ij}^t) = f_{ij}^t \frac{\left(\mathbf{U}^T\mathbf{X} + \lambda\mathbf{F}\mathbf{A} + div\left(\frac{\nabla\mathbf{F}}{|\nabla\mathbf{F}|^{2-P}}\right)\right)_{ij}}{(\mathbf{U}^T\mathbf{U}\mathbf{F} + \lambda\mathbf{F}\mathbf{L})_{ij}}. \tag{25}$$

Thus, we can conclude that the objective function can get the local optimal solution under the iterative updating rules. Moreover, the related theory can also refer to [47].

Evaluation metrics

The introduction of ATV and hypergraph makes the similar cells closer and different types of cells more separated. So, we select sample clustering to evaluate the clustering performance. To prove the effectiveness of our method, we introduce two evaluation metrics to make a fair comparison.

Adjusted rand index (ARI) is one of the most popular metrics to reflect the clustering results, which is defined as

$$\mathrm{ARI}(L_p,L_t) = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_{ij}\binom{n_{ij}}{2}\sum_{ij}\binom{n_{ij}}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i\binom{e_i}{2} + \sum_j\binom{t_j}{2}\right] - \left[\sum_i\binom{e_i}{2}\sum_j\binom{t_j}{2}\right]/\binom{n}{2}}, \tag{26}$$

where L_p denotes the predicted labels obtained from the method, and L_t are the true labels from the original data, respectively. n denotes the number of single-cells, $\binom{n}{2} = n(n-1)/2$. e_i and e_j are the numbers of single-cell in the predicted cluster i and true cluster j, respectively; n_{ij} is the number both in the predicted cluster and the true cluster.

When the number of cells with different labels is unbalanced, normalized mutual information (NMI) is

more widely used, which is defined as follows:

$$\mathrm{NMI}(L_p,L_t)=\frac{M(L_p,L_t)}{[H(L_p)+H(L_t)]/2}, \tag{27}$$

where H denotes the entropy and $M(L_p,L_t)$ denotes the mutual information between L_p and L_t. The value range of ARI is $[-1,1]$, the value range of NMI is [0, 1]. The larger the value is, the more consistent the clustering result is with the actual situation.

ACKNOWLEDGEMENTS

This work was supported in part by the grants provided by the National Natural Science Foundation of China (No. 61872220).

COMPLIANCE WITH ETHICS GUIDELINES

The authors Ya-Li Zhu, Xiao-Ning Zhang, Chuan-Yuan Wang, Jin-Xing Liu and Xiang-Zhen Kong declare that they have no conflict of interest.

This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Villani, A. C., Satija, R., Reynolds, G., Sarkizova, S., Shekhar, K., Fletcher, J., Griesbeck, M., Butler, A., Zheng, S., Lazo, S., *et al.* (2017) Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science, 356, eaah4573
2. Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., *et al.* (2009) mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods, 6, 377–382
3. Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J. B., Lönnerberg, P. and Linnarsson, S. (2011) Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res., 21, 1160–1167
4. Xin, Y., Kim, J., Okamoto, H., Ni, M., Wei, Y., Adler, C., Murphy, A. J., Yancopoulos, G. D., Lin, C. and Gromada, J. (2016) RNA sequencing of single human islet cells reveals type 2 diabetes genes. Cell Metab., 24, 608–615
5. Xu, Y., Mizuno, T., Sridharan, A., Du, Y., Guo, M., Tang, J., Wikenheiser-Brokamp, K. A., Perl, A. T., Funari, V. A., Gokey, J. J., *et al.* (2016) Single-cell RNA sequencing identifies diverse roles of epithelial cells in idiopathic pulmonary fibrosis. JCI Insight, 1, e90558
6. Usoskin, D., Furlan, A., Islam, S., Abdo, H., Lönnerberg, P., Lou, D., Hjerling-Leffler, J., Haeggström, J., Kharchenko, O., Kharchenko, P. V., *et al.* (2015) Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. Nat. Neurosci., 18, 145–153
7. Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., Cahill, D. P., Nahed, B. V., Curry, W. T., Martuza, R. L., *et al.* (2014) Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science, 344, 1396–1401
8. Treutlein, B., Brownfield, D. G., Wu, A. R., Neff, N. F., Mantalas, G. L., Espinoza, F. H., Desai, T. J., Krasnow, M. A. and Quake, S. R. (2014) Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature, 509, 371–375
9. Maaten, L. d. and Hinton, G. (2008) Visualizing data using t-SNE. J. Mach. Learn. Res., 9, 2579–2605
10. Wold, S., Esbensen, K. and Geladi, P. (1987) Principal component analysis. Chemom. Intell. Lab. Syst., 2, 37–52
11. von Luxburg, U. (2007) A tutorial on spectral clustering. Stat. Comput., 17, 395–416
12. Wang, B., Zhu, J., Pierson, E., Ramazzotti, D. and Batzoglou, S. (2017) Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. Nat. Methods, 14, 414–416
13. Jiao, C.-N., Gao, Y.-L., Yu, N., Liu, J.-X. and Qi, L.-Y. (2020) Hyper-graph regularized constrained NMF for selecting differentially expressed genes and tumor classification. IEEE J. Biomed. Health Inform., 24, 3002–3011
14. Lin, X. and Boutros, P. C. (2020) Optimization and expansion of non-negative matrix factorization. BMC Bioinformatics, 21, 7
15. Yu, N., Wu, M. J., Liu, J. X., Zheng, C. H. and Xu, Y. (2020) Correntropy-based hypergraph regularized NMF for clustering and feature selection on multi-cancer integrated data. IEEE Trans. Cybern
16. Gao, Z., Wang, Y.-T., Wu, Q.-W., Ni, J.-C. and Zheng, C.-H. (2020) Graph regularized $L_{2,1}$-nonnegative matrix factorization for miRNA-disease association prediction. BMC Bioinformatics, 21, 61
17. Zhu, X., Ching, T., Pan, X., Weissman, S. M. and Garmire, L. (2017) Detecting heterogeneity in single-cell RNA-Seq data by non-negative matrix factorization. PeerJ, 5, e2888
18. Moon, K. R., Stanley, J. S. III, Burkhardt, D., van Dijk, D., Wolf, G. and Krishnaswamy, S. (2018) Manifold learning-based methods for analyzing single-cell RNA-sequencing data. Curr. Opin. Syst. Biol., 7, 36–46
19. Cai, D., He, X., Han, J. and Huang, T. S. (2011) Graph regularized nonnegative matrix factorization for data representation. IEEE Trans. Pattern Anal. Mach. Intell., 33, 1548–1560
20. Zeng, K., Yu, J., Li, C., You, J. and Jin, T. (2014) Image clustering

by hyper-graph regularized non-negative matrix factorization. Neurocomputing, 138, 209–217
21. Rudin, L. I., Osher, S. and Fatemi, E. (1992) Nonlinear total variation based noise removal algorithms. Physica. D, 60, 259–268
22. Leng, C., Cai, G., Yu, D. and Wang, Z. (2017) Adaptive total-variation for non-negative matrix factorization on manifold. Pattern Recognit. Lett., 98, 68–74
23. Darmanis, S., Sloan, S. A., Zhang, Y., Enge, M., Caneda, C., Shuer, L. M., Hayden Gephart, M. G., Barres, B. A. and Quake, S. R. (2015) A survey of human brain transcriptome diversity at the single cell level. Proc. Natl. Acad. Sci. USA, 112, 7285–7290
24. Goolam, M., Scialdone, A., Graham, S. J. L., Macaulay, I. C., Jedrusik, A., Hupalowska, A., Voet, T., Marioni, J. C. and Zernicka-Goetz, M. (2016) Heterogeneity in oct4 and sox2 targets biases cell fate in 4-cell mouse embryos. Cell, 165, 61–74
25. Treutlein, B., Brownfield, D. G., Wu, A. R., Neff, N. F., Mantalas, G. L., Espinoza, F. H., Desai, T. J., Krasnow, M. A. and Quake, S. R. (2014) Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature, 509, 371–375
26. Grover, A., Sanjuan-Pla, A., Thongjuea, S., Carrelha, J., Giustacchini, A., Gambardella, A., Macaulay, I., Mancini, E., Luis, T. C., Mead, A., *et al.* (2016) Single-cell RNA sequencing reveals molecular and functional platelet bias of aged haematopoietic stem cells. Nat. Commun., 7, 11075
27. Breton, G., Zheng, S., Valieris, R., Tojal da Silva, I., Satija, R. and Nussenzweig, M. C. (2016) Human dendritic cells (DCs) are derived from distinct circulating precursors that are precommitted to become $CD1c^+$ or $CD141^+$ DCs. J. Exp. Med., 213, 2861–2870
28. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res., 13, 2498–2504
29. Yamada, N., Horikawa, Y., Oda, N., Iizuka, K., Shihara, N., Kishi, S. and Takeda, J. (2005) Genetic variation in the hypoxia-inducible factor-1α gene is associated with type 2 diabetes in Japanese. J. Clin. Endocrinol. Metab., 90, 5841–5847
30. Zhou, J.-C., Zhou, J., Su, L., Huang, K. and Lei, X. G. (2018) Selenium and Diabetes. In: Selenium. MICHALKE, B, 317–344. Cham: Springer International Publishing
31. Brina, D., Miluzio, A., Ricciardi, S., Clarke, K., Davidsen, P. K., Viero, G., Tebaldi, T., Offenhäuser, N., Rozman, J., Rathkolb, B., *et al.* (2015) eIF6 coordinates insulin sensitivity and lipid metabolism by coupling translation to transcription. Nat. Commun., 6, 8261
32. Olsson, A. H., Yang, B. T., Hall, E., Taneera, J., Salehi, A., Dekker Nitert, M. and Ling, C.. (2011) Decreased expression of genes involved in oxidative phosphorylation in human pancreatic islets from patients with type 2 diabetes. Eur. J. Endocrinol., 165, 589–595
33. Molina, M. F., Qu, H.-Q., Rentfro, A. R., Nair, S., Lu, Y., Hanis, C. L., McCormick, J. B. and Fisher-Hoch, S. P. (2011) Decreased expression of ATP6V1H in type 2 diabetes: a pilot report on the diabetes risk study in Mexican Americans. Biochem. Biophys. Res. Commun., 412, 728–731
34. Crétien, A., Proust, A., Delaunay, J., Rincé, P., Leblanc, T., Ducrocq, R., Simansour, M., Marie, I., Tamary, H., Meerpohl, J., *et al.* (2010) Genetic variants in the noncoding region of *RPS19* gene in Diamond-Blackfan anemia: potential implications for phenotypic heterogeneity. Am J Hematol, 85, 111–116
35. Liu, S., Kim, T.-H., Franklin, D. A. and Zhang, Y. (2017) Protection against high-fat-diet-induced obesity in mdm2c305f mice due to reduced p53 activity and enhanced energy expenditure. Cell Rep., 18, 1005–1018
36. Chen, H., Fang, X., Zhu, H., Li, S., He, J., Gu, P., Fan, D., Han, F., Zeng, Y., Yu, X., *et al.* (2014) Gene expression profile analysis for different idiopathic interstitial pneumonias subtypes. Exp. Lung Res., 40, 367–379
37. Gharbi-Ayachi, A., Labbé, J. C., Burgess, A., Vigneron, S., Strub, J. M., Brioudes, E., Van-Dorsselaer, A., Castro, A. and Lorca, T. (2010) The substrate of Greatwall kinase, Arpp19, controls mitosis by inhibiting protein phosphatase 2A. Science, 330, 1673–1677
38. Gong, Y., Wu, W., Zou, X., Liu, F., Wei, T. and Zhu, J. (2018) MiR-26a inhibits thyroid cancer cell proliferation by targeting ARPP19. Am J Cancer Res, 8, 1030–1039
39. Miyazaki, H., Patel, V., Wang, H., Edmunds, R. K., Gutkind, J. S. and Yeudall, W. A. (2006) Down-regulation of CXCL5 inhibits squamous carcinogenesis. Cancer Res., 66, 4279–4284
40. Begley, L. A., Kasina, S., Mehra, R., Adsule, S., Admon, A. J., Lonigro, R. J., Chinnaiyan, A. M. and Macoska, J. A. (2008) CXCL5 promotes prostate cancer progression. Neoplasia, 10, 244–254
41. Plowman, J., Bolderson, E., Burgess, J., Richard, D. and O'Byrne, K. (2019) Banf1 as a marker of lung cancer cell sensitivity to cisplatin. Lung Cancer, 127, S3
42. Hu, J., Yang, D., Zhang, H., Liu, W., Zhao, Y., Lu, H., Meng, Q., Pang, H., Chen, X., Liu, Y., *et al.* (2015) USP22 promotes tumor progression and induces epithelial-mesenchymal transition in lung adenocarcinoma. Lung Cancer, 88, 239–245
43. Levine, S., Chen, Y., and Stanich, J. (2004) Image restoration via nonstandard diffusion. Duquesne University, Department of Mathematics and Computer Science Technical Report. 04-01
44. Huang, S., Wang, H., Ge, Y., Huangfu, L., Zhang, X. and Yang, D. (2018) Improved hypergraph regularized nonnegative matrix factorization with sparse representation. Pattern Recognit. Lett., 102, 8–14
45. Jin, T., Yu, Z., Gao, Y., Gao, S., Sun, X. and Li, C. (2019) Robust $\ell 2-$ hypergraph and its applications. Inf. Sci., 501, 708–723
46. Yin, H. and Liu, H. (2010) Nonnegative matrix factorization with bounded total variational regularization for face recognition. Pattern Recognit. Lett., 31, 2468–2473
47. Hong, M., Razaviyayn, M., Luo, Z.-Q. and Pang, J.-S. (2016) A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing. IEEE Signal Proc. Mag., 33, 57–77