Single-cell entropy network detects the activity of immune cells based on ribosomal protein genes

Qiqi Jin [a,b,c], Chunman Zuo [a], Haoyue Cui [a,b,c], Lin Li [a], Yiwen Yang [a,b,c], Hao Dai [a,*], Luonan Chen [a,b,d,e,*]

[a] Key Laboratory of Systems Biology, Shanghai Institute of Biochemistry and Cell Biology, Center for Excellence in Molecular Cell Science, Chinese Academy of Sciences, Shanghai 200031, China
[b] School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, China
[c] University of Chinese Academy of Sciences, Beijing 100049, China
[d] Guangdong Institute of Intelligence Science and Technology, Hengqin, Zhuhai, Guangdong 519031, China
[e] Key Laboratory of Systems Biology, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Hangzhou 310024, China

ARTICLE INFO

ABSTRACT

We developed a new computational method, Single-Cell Entropy Network (SCEN) to analyze single-cell RNA-seq data, which used the information of gene-gene associations to discover new heterogeneity of immune cells as well as identify existing cell types. Based on SCEN, we defined association-entropy (AE) for each cell and each gene through single-cell gene co-expression networks to measure the strength of association between each gene and all other genes at a single-cell resolution. Analyses of public datasets indicated that the AE of ribosomal protein genes (RP genes) varied greatly even in the same cell type of immune cells and the average AE of RP genes of immune cells in each person was significantly associated with the healthy/disease state of this person. Based on existing research and theory, we inferred that the AE of RP genes represented the heterogeneity of ribosomes and reflected the activity of immune cells. We believe SCEN can provide more biological insights into the heterogeneity and diversity of immune cells, especially the change of immune cells in the diseases.

1. Introduction

The advances of single-cell transcriptomics help researchers reveal the heterogeneity and functional diversity among cell populations and discover novel cell types [1–4]. However, existing methods usually focus on the differential expression of genes, but ignore the heterogeneity of gene-gene associations. Theoretically, single-cell RNA sequencing (scRNA-seq) data not only explicitly describe the gene expressions at a single-cell resolution, but also implicitly contain the information on the gene-gene associations [5].

In this work, we propose a new computational method *Single-Cell Entropy Network* (*SCEN*) to analyze scRNA-seq data, which constructs a gene association network for each single cell (i.e. one network for one cell), where nodes are genes and edges are gene-gene associations measured by mutual information. Based on SCEN, we can define *association-entropy* (*AE*) for each cell to measure the association strength of gene coexpression at single-cell resolution. The higher AE level indicates that the genes are more strongly-associated with each other in a single cell. From dynamical viewpoint, gene expressions are the variables of a nonlinear biological system which change dynamically with time and conditions due to various fluctuations or noises, but their gene associations/regulations are relatively stable [6,7]. Thus, gene associations can characterize the features of single cell more reliably, which may help us find new cell heterogeneity that is ignored by traditional gene expression analysis.

In this study, we found the AE of ribosomal protein genes (RP genes) varied greatly even in the same cell type of immune cells. Ribosomal proteins are the main components of ribosomes, which act as a translational machine in cells related to protein expression. Some studies found the heterogeneity of ribosomes [8], and different ribosomes have different mRNA selectivity and produce different proteins [9,10], e.g. Rps26-deficient ribosomes preferentially translate the mRNA from stress-response pathways [11]. Based on the definition of AE, we think AE of RP genes can reflect the

* Corresponding authors at: Center for Excellence in Molecular Cell Science, Chinese Academy of Sciences, Yueyang Road 320, Xuhui District, Shanghai 200031, China (Luonan Chen); Center for Excellence in Molecular Cell Science, Chinese Academy of Sciences, Yueyang Road 320, Xuhui District, Shanghai 200031, China (Hao Dai).

E-mail addresses: daihao@sibcb.ac.cn (H. Dai), lnchen@sibcb.ac.cn (L. Chen).

heterogeneity of ribosomes. High AE of RP genes indicates that these RP genes are strongly associated and represent normal ribosomes, while low AE of RP genes represent defective ribosomes.

Furthermore, we found that the average AE of RP genes of immune cells in each person was significantly associated with the healthy/disease state of this person, which validated the biological significance of AE of RP genes. Based on existing research and theory [12–14], we inferred that the AE of RP genes might reflect the activity of immune cells. Usually, immune cells will be activated by virus infection or other external stimuli to kill infected cells before progeny viruses are released, i.e. the activity of immune cells will rise after infection. This change is rapid and reversible, and is related to the heterogeneity of ribosomes. Thus, as a new method of single-cell sequencing data bioinformatics tool, we believe SCEN can provide more biological insights into the heterogeneity and diversity of immune cells, especially the change of immune cells in the diseases.

2. Methods

In this paper, we proposed a new method "Single-Cell Entropy Network" (SCEN) to analyze single-cell RNA-seq data (Fig. 1).

From the gene expression matrix comprised of m genes and n cells, we first make $m*(m-1)/2$ scatter diagrams, where each scatter diagram includes n points/cells, and the x-axis and y-axis represent the expression levels of two genes in the n cells. Then, from the scatter diagram of genes x and y, we could calculate the mutual information of the two genes based on Eqn. (2).

$$MI(x,y) = \sum_{X,Y \in \Omega} p(X,Y) \log \frac{p(X,Y)}{p(X)p(Y)} \tag{2}$$

where $p(X)$ and $p(Y)$ are marginal probability distributions of genes x and y respectively, $p(X,Y)$ is their joint probability distribution function, X and Y are expressions of genes x and y, and Ω is the domain of definition.

To estimate $MI(x, y)$, we substitute the probability by the frequency numerically. As shown in Fig. 1, we draw light orange box that covers the plot k (or cell k) to represent the neighborhood of x_k. The width of the box is large just enough to make the box cover $0.1n$ points ($0.05n$ points on the left of x_k and $0.05n$ points on the right of x_k), and thus the number of points in the neighborhood of x_k is fixed to $0.1n$ ($n_x^{(k)} = 0.1n$). Similarly, we draw dark orange box to represent the neighborhood of y_k, and $n_y^{(k)}$ is also equal to $0.1n$. Then, if the intersection of the previous two boxes cover $n_{xy}^{(k)}$ points, mutual information of genes x and y for all n cells could be approximately calculated by Eqn. (3).

$$MI(x,y) = \sum_{i=1}^{n} \frac{n_{xy}^{(i)}}{n} \log \frac{\frac{n_{xy}^{(i)}}{n}}{\frac{n_x^{(i)}}{n} \cdot \frac{n_y^{(i)}}{n}} \tag{3}$$

Then, we define the contribution of cell k in $MI(x, y)$ by decomposing Eqn. (3) into Eqn. (4).

$$MI^{(k)}(x,y) = \frac{n_{xy}^{(k)}}{n} \log \frac{\frac{n_{xy}^{(k)}}{n}}{\frac{n_x^{(k)}}{n} \cdot \frac{n_y^{(k)}}{n}} \tag{4}$$

If two variables are correlated in some points and uncorrelated in the others, the former points give positive contributions to $MI(x, y)$ and thus $MI^{(k)}(x, y) > 0$, while the latter points give zero contributions or noises to $MI(x, y)$ and thus $MI^{(k)}(x, y) \leq 0$ (Supplementary Note 1).

The parameter 0.1 could be substituted by other values, but we find that the result is not sensitive to this parameter in the range of 0.05 to 0.2 based on the test of several public datasets, and 0.1 seems to be the best (Supplementary Note 2).

Based on Eqns. (3) and (4), we could construct a network for each cell, where nodes = genes and edges = gene-gene associations. By the normalization, the weight of edge x-y in cell-k network with m genes/variables is defined as.

$$w^{(k)}(x,y) = \max\left\{ MI^{(k)}(x,y) \cdot \frac{\sum_{i=1}^{m} \sum_{j=i+1}^{m} MI(i,j)}{n \cdot \sum_{i=1}^{m} \sum_{j=i+1}^{m} MI^{(k)}(i,j)}, 0 \right\} \tag{5}$$

Specially, we require $w^{(k)}(x, y) \geq 0$ based on Eqn. (5) to make the weight of each edge non-negative. In addition, we set $w^{(k)}(x, y) = 0$ artificially when $x_k = 0$ or $y_k = 0$, because most zeros in the gene expression matrix come from the technique problems in single-cell RNA-seq, which is meaningless in biology and makes the wrong estimation in gene association analysis. Thus, we construct a single-cell entropy network for each cell k based on Eqn. (5), from scRNA-seq data.

At last, by Eqn. (5) we obtain n single-cell networks for the n cells. Based on these networks, we could find the network heterogeneity at a single-cell level. Furthermore, we could define the association-entropy (AE) of gene x in cell k based on Eqn. (6).

$$AE_{xk} = \sum_{\substack{y = 1 \\ y \neq x}}^{m} w^{(k)}(x,y) \tag{6}$$

AE_{xk} is the degree of gene x in cell-k network. Thus, we can construct an association-entropy matrix (AE matrix) with AE_{xk} as an element. Clearly, the AE matrix has the same number of rows and columns as the original gene expression matrix whose element is gene expression, and thus any traditional algorithm for gene expression matrix can be adopted for AE matrix, i.e. to analyze cell clustering and pseudotime trajectory.

It should be noted that AE is different from the traditional definition of entropy. In this paper, AE represents the summation of mutual information. Intuitively, the higher AE level means this gene in the cell is associated with other genes more strongly due to its higher MI in the single-cell network, so high AE level represents the state of high order and low entropy (traditional definition). As AE quantifies the importance of genes from a network viewpoint, the analyses of AE matrix might help to find more features that is ignored by traditional gene expression analysis. Note that We can similarly construct the direct gene association network or single-cell entropy direct-network in the form of conditional mutual information [1–4].

3. Results

3.1. SCEN shows better clustering performance

In this paper, we proposed a new method *Single-Cell Entropy Network* (*SCEN*) to analyze single-cell RNA-seq data (Fig. 1, Supplementary Note 1, 2). Based on SCEN, we can obtain n single-cell networks for the n cells, and further obtain an association-entropy (AE) matrix that has the same number of rows and columns as the original gene expression matrix. Thus, any traditional scRNA-seq algorithm for gene expression matrix can be adopted for our AE matrix to analyze cell clustering. To validate the performance of SCEN method, we collected seven high-quality scRNA-seq datasets with clear classification labels from literatures [15–20] and performed cell clustering analysis on both gene expression matrix and AE matrix. By comparing the clustering result with the known classification labels, we could evaluate the clustering performances. As shown in Table 1, obviously our AE matrix is superior to original gene expression matrix on various datasets by various clustering algorithms [21–23]. In fact, the best result of each dataset is all from our AE matrix, which demonstrates the advantage of
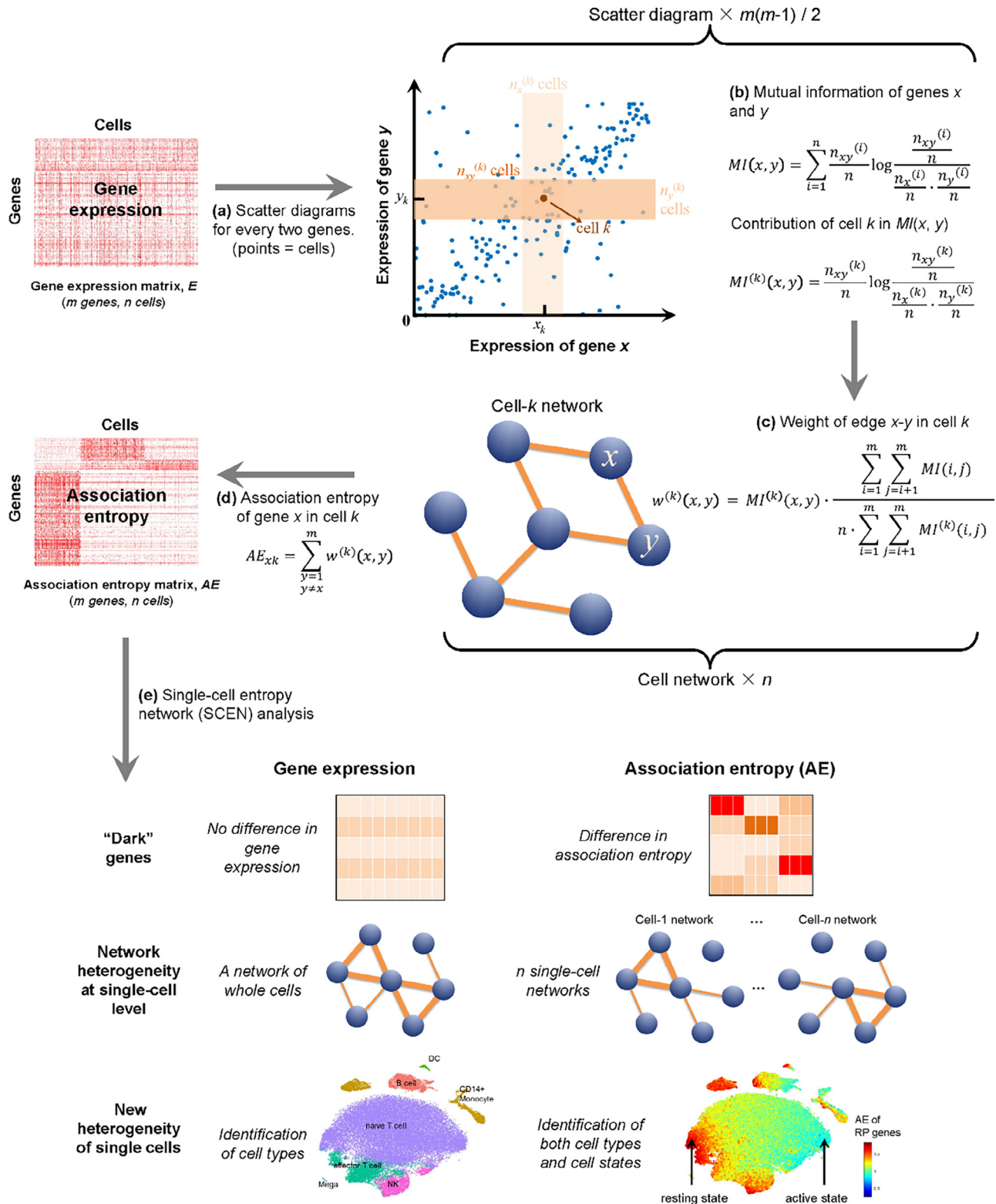
Fig. 1. Schematic illustration of SCEN method. **(a)** First, make a scatter diagram for every two genes based on the gene expression profile comprised of m genes and n cells, where points = cells, and then m genes can produce $m*(m-1)/2$ scatter diagrams. **(b)** In the scatter diagram of genes × and y, make a light orange box and a dark orange box near the cell k (red plot) to represent the neighborhood of x_k and y_k respectively. The number of plots/cells in the two boxes is $n_x^{(k)}$ and $n_y^{(k)}$ respectively, and the number of plots/cells in the intersection of the two boxes is $n_{xy}^{(k)}$. Then, calculate $MI(x, y)$ and $MI^{(k)}(x, y)$. **(c)** Calculate $w^{(k)}(x, y)$ that represents the weight of edge x-y in cell-k network, and construct n single-cell networks. **(d)** Calculate the association entropy (AE) of each gene in each cell, and get an AE matrix. **(e)** Based on SCEN method, we can discover "dark" genes ignored by original gene expression matrix, characterize network heterogeneity at a single-cell level, and distinguish the active or resting state of immune cells. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Clustering performance of gene expression matrix and AE matrix, evaluated by adjusted rand index (ARI). Hierarchical (t-SNE) and k-means (t-SNE) represent that the clustering is performed after dimension-reduction by t-SNE.

		Buettner	Kolod ziejczyk	Darmanis	Chu-type	Chu-time	Kim	Trapnell
Hierarchical	Expression	0.48	0.49	0.63	0.75	0.67	0.66	0.08
	AE	**0.86**	**0.99**	**0.64**	**0.75**	**0.77**	**0.67**	**0.23**
k-means	Expression	0.31	0.53	0.58	0.73	0.59	0.60	0.14
	AE	**0.67**	**0.90**	**0.71**	**0.75**	**0.67**	**0.75**	**0.42**
Hierarchical (t-SNE)	Expression	0.32	0.99	0.67	0.98	0.68	0.66	0.16
	AE	**0.67**	**1.00**	**0.82**	**0.99**	**0.69**	**0.95**	**0.47**
k-means (t-SNE)	Expression	0.30	0.99	0.65	0.98	0.69	0.72	0.16
	AE	**0.70**	**1.00**	**0.82**	**0.99**	**0.69**	**0.89**	**0.47**
SIMLR	Expression	0.92	0.99	0.75	0.74	0.66	0.97	0.21
	AE	**0.98**	**1.00**	**0.88**	**0.75**	**0.70**	**0.95**	**0.28**
Seurat	Expression	0.27	0.73	0.86	0.95	0.72	0.77	0.17
	AE	**0.62**	**0.84**	**0.82**	**0.76**	**0.69**	**0.98**	**0.51**

SCEN method. More details of clustering analysis are listed in Supplementary Note 3.

3.2. SCEN discovers new heterogeneity of immune cells marked by ribosomal protein genes

In this work, we performed our SCEN method on a public dataset of 2,700 Peripheral Blood Mononuclear Cells (PBMCs) made by 10X Genomics, which was a sample dataset in Seurat tutorial (https://satijalab.org/seurat/v3.1/pbmc3k_tutorial.html). As shown in Fig. 2a, clustering analysis of our AE matrix identifies all known cell types that are also identified by gene expression matrix, but our method detects two new subgroups of cells, RP + Monocyte and RP + T cell by Seurat 3.1 [23]. RP + Monocyte is a subgroup of CD14 + Monocyte, and RP + T cells come from three types of T cells (CD8 + T cells, naïve CD4 + T cells and memory CD4 + T cells) (Supplementary Note 4).

Differential expression genes (DEGs) and differential AE genes (DAEGs) analyses indicate that the AE of ribosomal protein genes (RP genes) is much higher in RP + cells than other cells, and thus we marked the two new subgroups with "RP+" (Fig. 2b). As shown in Fig. 2c, the number of DAEGs is much larger than DEGs, and DAEGs cover most DEGs, which implies that the major differences between RP + T cells/Monocytes and other T cells/Monocytes are in AE instead of gene expression. We define the DAEGs that do not belong to DEGs as "dark" genes, i.e. non-differential expression genes but differential AE genes. There are 176 "dark" genes in RP + Monocyte and 228 "dark" genes in RP + T cell. GO analyses show that these "dark" genes are enriched to translation-related GOs including translation initiation, mRNA catabolic process and co-translational protein targeting to membrane (Supplementary Note 5), demonstrating the major differences on the regulation of translation level between RP + cells and other cells.

Furthermore, we constructed single-cell networks by our SCEN method. We selected 4 cells from 4 different sources and constructed the networks of RP genes. Fig. 2d clearly demonstrates that the associations of RP genes in RP + cells are much stronger than other cells, which implies that the ribosomes are different between RP + cells and other cells, i.e. the ribosomes in RP + cells are complete while the ribosomes in other cells may be defective comparatively.

Besides the dataset of 2,700 PBMCs, we also performed SCEN method on other public datasets including PBMC-68 k dataset [24] (containing about 68,000 PBMCs, NCBI SRA: SRP073767), UCB dataset [25] (containing about 19,000 human umbilical cord blood cells, NCBI BioProject: PRJNA524398) and Melanoma tumor dataset [26] (containing 2,887 immune cells from melanoma tumor, NCBI GEO: GSE72056). As shown in Fig. 3, although the

expressions of the RP genes do not show obvious difference, it is clear that the AE of RP genes varied greatly in the same cell type, and RP + cells are universal in the immune cells from various sources.

To confirm whether RP + cells represent a new cell type or not, we inspect the expression levels of all marker genes of classical cell types, and the results indicate that RP + cells do not express any known marker genes specifically (Supplementary Note 6). We also analyze the differential expression genes of RP + cells, and we cannot find any gene that is only expressed in RP + cells. In addition, we find the cells with low AE of RP genes highly express MHC class I molecules (T cell) and MHC class II molecules (Monocyte). In view of the key role of MHC molecules in immune response and antigen presentation, we do not consider RP + cells as a new cell type, but reflect the active or inactive state of immune cells.

3.3. RP + cells are in the initial stage of pseudo-trajectory

In this work, we performed pseudo-trajectory analyses on T cells and Monocytes from the human umbilical cord blood (UCB) dataset [25] (Fig. 4). As shown in Fig. 4a-4d, Monocytes and T cells are re-clustered separately, and trajectory analysis is performed using Monocle2 [27].

We find that the RP + cells are located in the beginning of the trajectory where the AE of RP genes presents a downward trend along the trajectory. In Fig. 4a, for Monocytes, the trajectory starts from RP + Monocytes and ends up with Monocytes highly expressing CD69, also known as a marker of Monocyte activation [28]. For T cells, the trajectory starts from RP + T cells, follows by CD4 T_{CM} and ends up with T_{EM} (Fig. 4d). It has been reported that in naïve T cells, RP genes are highly expressed and the ribosomal protein are preferentially translated for the preparation of the upcoming translation activities in cell proliferation [29]. Hence, we suppose that the RP + cells in our work may be also related to the translation activities. To demonstrate this, we performed enrichment analysis based on the differential AE genes (DAEGs) between the cells on both ends of the trajectory. Besides the translation-related GOs, we find some cytokine-mediated signaling pathways, such as IL-17 signaling pathway and TNF-alpha / NF-kappa B signaling pathway (Supplementary Note 7). This result implies that the RP level of cells may be related to the external stimulus. We also selected five representative pathways and analyze their associations. We find the associations are the strongest in RP + cells, which indicates the relationship between ribosome biogenesis / assembly and T cell activation / differentiation / proliferation (Fig. 4e). In addition, we also performed pseudo-trajectory analyses and GO analyses on PBMC-68 k dataset [24] (Supplementary Note
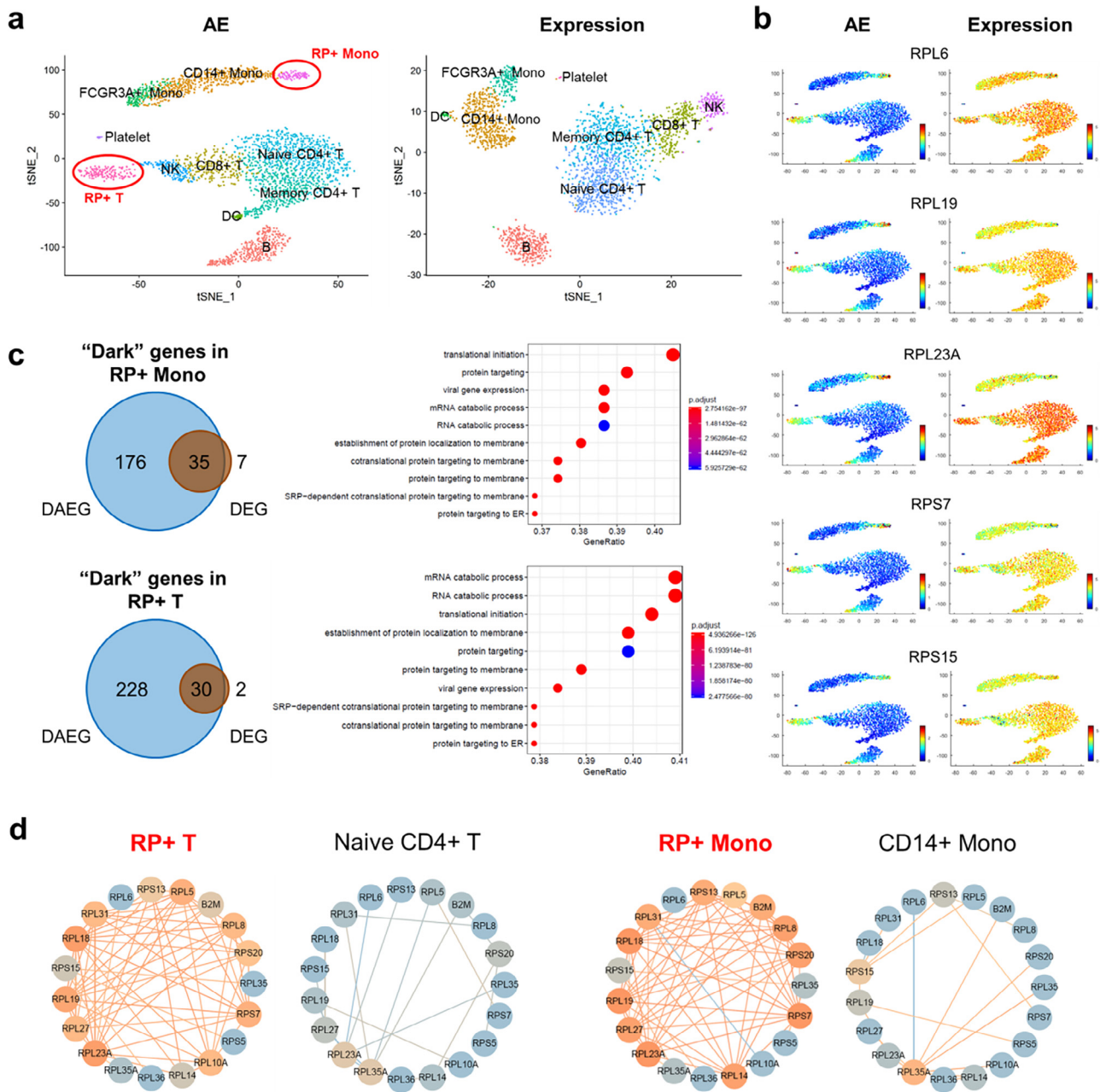
Fig. 2. Performance of SCEN on the PBMC dataset with 2700 cells. **(a)** Clustering based on AE matrix can identify two new groups of cells, RP + Monocyte and RP + T cell. **(b)** AE and gene expression of some RP genes. Plots are colored by log(1 + *AE*) or log(1 + *normalized counts*). **(c)** Differential AE genes (DAEGs) and differential expression genes (DEGs) analyses identify 176 "dark" genes (DAEGs but not DEGs) between RP + Monocytes and other CD14 + Monocytes, and 228 "dark" genes between RP + T cells and other T cells. Some translation-related GOs are enriched from these "dark" genes. **(d)** 4 single-cell entropy networks from 4 single cells with different sources, where the nodes are some RP genes.

8) and melanoma tumor dataset [26] (Supplementary Note 9), and similar conclusion can be obtained.

3.4. AE of RP genes is associated with diseases and reflects the activity of immune cells

In this work, we collected four public datasets related with diseases, and each dataset contains multiple samples that were taken from healthy donors or patients. For each sample/individual, scRNA-seq produced a gene expression matrix, and we performed SCEN method based on this matrix to produce an AE matrix. Then we calculated RP level (average AE of RP genes) for each

sample/individual based on the definition of Eqn. (1). We found RP level is significantly associated with diseases, while AE of RP genes is able to represent the activity of immune cells.

$$RP \; level = \log_{10}\left(1 + \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{m} AE_{RP_i,j}\right) \qquad (1)$$

where $AE_{RP_i,j}$ represents the AE of RP gene i in cell j, m is the number of RP genes and n is the number of cells.

We first analyzed a COVID-19 dataset that contains 135,600 cells derived from 24 donors (32 samples) (FigShare: https://doi.org/10.6084/m9.figshare.12436517) [30], including 5 samples
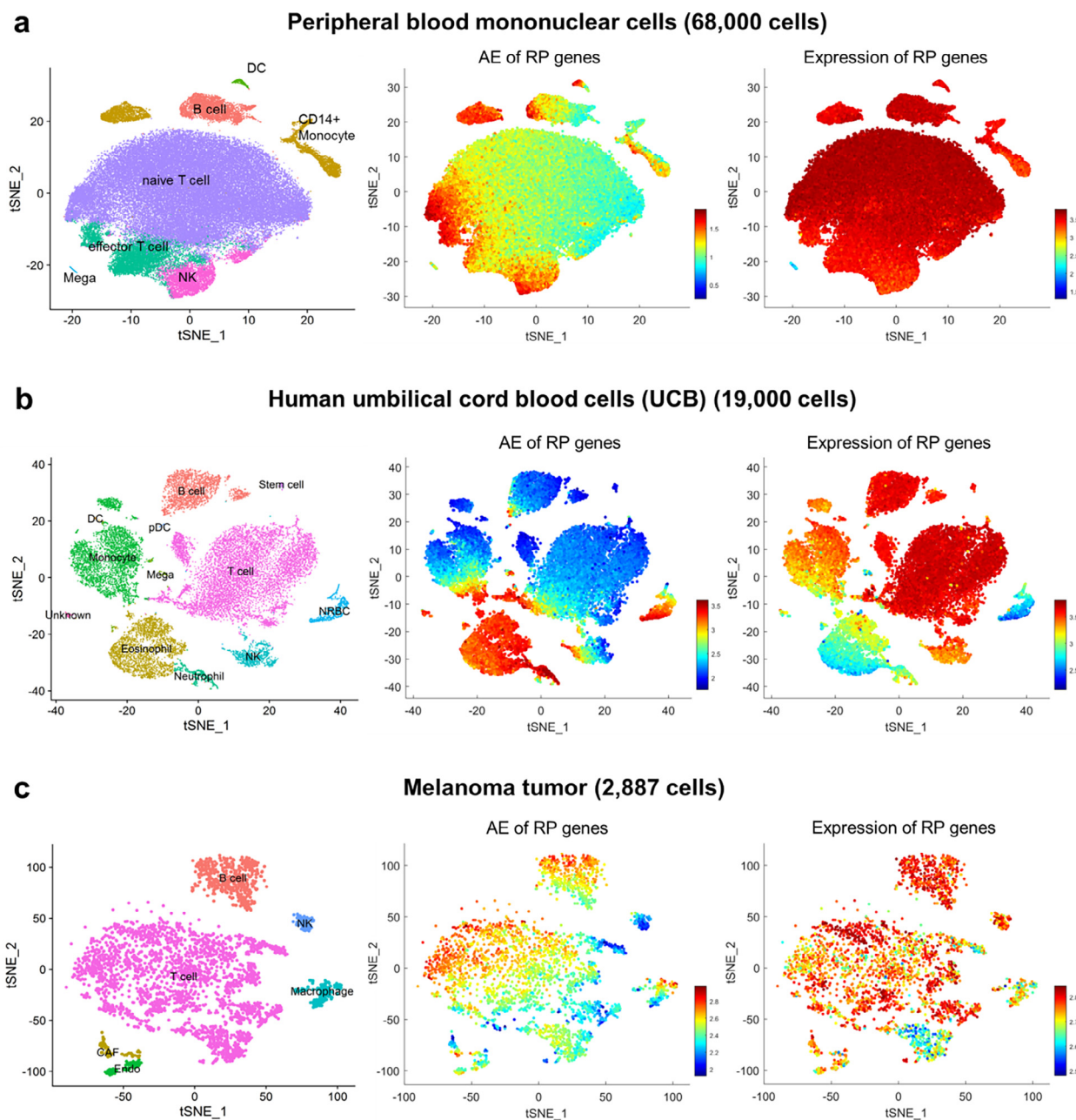
Fig. 3. The wide existence of RP + cells in (**a**) Peripheral blood mononuclear cells, (**b**) Human umbilical cord blood cells, and (**c**) Immune cells from melanoma tumor tissue. Clustering performance, AE of all RP genes and expression of all RP genes are illustrated.

from healthy people, 14 samples from moderate patients infected by COVID-19 and 13 samples from critical patients. Each sample includes various cell types that can be classified into immune cells and epithelial cells. As shown in Fig. 5a, RP levels of immune cells have the largest values in healthy people, the moderate values in moderate patients, and the smallest values in critical patients. As there are more active immune cells in patients than in healthy people, there seems to be negative correlation between AE of RP genes and the activity of immune cells. In addition, we can find the significant difference of RP level in lung epithelial cells between healthy people and COVID-19 patients. This result is of course as lung epithelial cells are the target of COVID-19.

We next analyzed a public scRNA-seq dataset (GSE131907) that contains 208,506 cells derived from 44 patients pathologically diagnosed with lung adenocarcinoma (LUAD) (58 samples) [31],

which cover 7 metastatic lymph node samples (mLN), 10 normal lymph node samples (nLN), 4 lung tumor tissue samples from advanced stage LUAD patients (tL/B), 11 primary tumor tissue samples (tLung), 11 distant normal lung tissue samples (nLung), 5 pleural fluid samples (PE) from LUAD patients with malignant pleural effusion, and 10 metastatic brain tissue samples (mBrain). The extensive single cell profiles include cancer, stromal, and immune cells in the surrounding tumor microenvironments. As shown in Fig. 5b, in T cells, B cells and myeloid cells, RP level in normal lymph nodes is much lower than that in metastatic lymph nodes, suggesting the great influences of cancer cells on immune cells, which is consistent with the existing studies [32]. In cancer patients, immune evasion is a common phenomenon [33–35], i.e. the immune cells cannot attack tumor cells. Thus, there are less active immune cells in metastatic lymph nodes than normal lymph
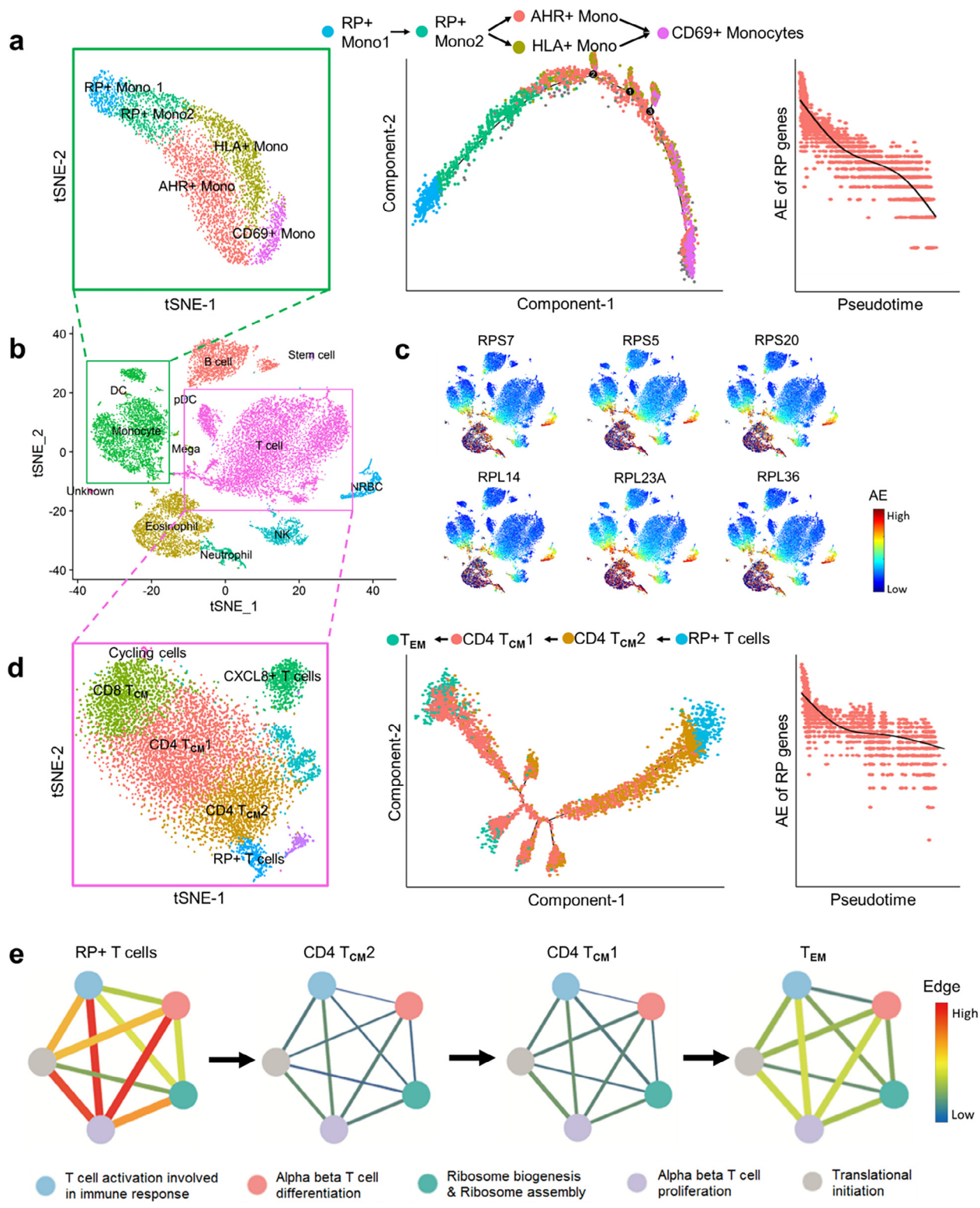
Fig. 4. The characteristics of RP + cells in UCB dataset. (**a, d**) Clustering and pseudo-trajectory results of Monocytes or T cell based on AE matrix, as well as the dynamic change of RP genes along the pseudotime. (**b**) Clustering of all UCB cells based on AE matrix. (**c**) AE levels of RP genes. (**e**) The network among 5 KEGG pathways in 4 different T cell subtypes.

nodes. And we can still find the negative correlation between AE of RP genes and the activity of immune cells, which is consistent with the previous conclusion in COVID-19 dataset. In addition, in T cells

and B cells, the RP levels of tumor lung tissues and distant normal lung tissues are almost the same, and are both significantly higher than those in normal lymph nodes, but similar to those in
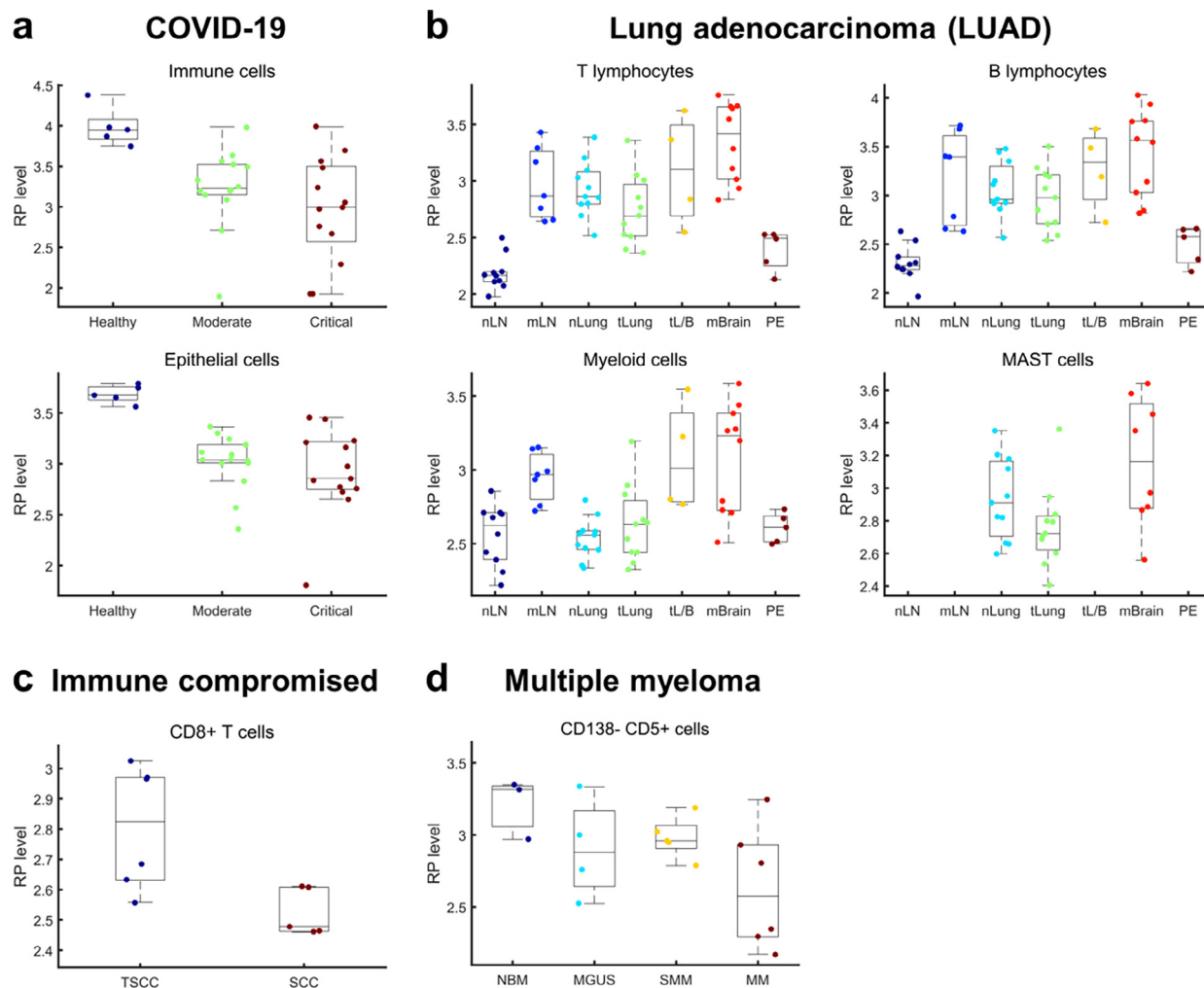
Fig. 5. The associations between RP levels and diseases: **(a)** COVID-19, **(b)** Lung adenocarcinoma (LUAD) (nLN = normal lymph node; mLN = metastatic lymph node; nLung = distant normal lung tissue; tLung = primary tumor tissue; tL/B = tumor tissue from advanced stage LUAD; mBrain = metastatic brain tissue; PE = pleural fluid), **(c)** Immune compromised (TSCC = immune compromised organ transplant recipients; SCC = immune competent patients with cutaneous squamous cell carcinoma tumors), **(d)** multiple myeloma (NBM = healthy donors; MGUS = monoclonal gammopathy of undetermined significance; SMM = smoldering myeloma; MM = full-blown multiple myeloma). RP level of each sample was illustrated in box plots (center line = median; box limits = upper (3/4) and lower (1/4) quartiles; whiskers = 1.5 × interquartile range).

metastatic lymph nodes. This result suggests that even if the tumor cells have not metastasized, perhaps they have inhibited the activity of immune cells in a wide range.

The third dataset (GSE145328) is from a study that reveals T-cell landscape differences between cutaneous squamous cell carcinoma tumors in immune competent (SCC in IC) and immune-compromised organ transplant recipients (TSCC in OTR) [36]. This dataset contained 32,738 cells derived from 5 SCC samples and 6 TSCC samples. As shown in Fig. 5c, RP levels in TSCC samples from immune-compromised patients are significantly higher than those in SCC samples from immune competent patients. This result is also consistent with the previous conclusion that there is negative correlation between AE of RP genes and the activity of immune cells.

Fig. 5d illustrates the change of RP level during the progression of multiple myeloma (MM) [37], which is a cancer derived from B cells. The dataset (GSE124310) contains 11,450 immune cells from varying stages of MM progression: from healthy stage (NBM, 3 samples), to precursor stages: monoclonal gammopathy of undetermined significance (MGUS, 4 samples) and smoldering myeloma (SMM, 5 samples), to full-blown MM (6 samples). We can see that the RP level is associated with the progress of MM. RP levels are the highest from the cells in healthy donors, is moderate from the pre-

cursor stages MGUS and SMM, and is the lowest from full-blown MM. This result demonstrates the association between disease progression and AE of RP genes again, and gives an insight into the early immune changes of multiple myeloma.

In summary, the average AE of RP genes of immune cells in each person is significantly associated with the healthy/disease state of this person, which validated the biological significance of AE of RP genes. Especially, we find the negative correlation between AE of RP genes and the activity of immune cells. It seems that high AE of RP genes (represent normal ribosomes) is associated with the inactive or resting state of immune cells, and low AE of RP genes (represent defective ribosomes) is associated with the active state. This inference is consistent with DRiPs hypothesis [13,14]. More statistical details of these results are provided in Supplementary Note 10.

4. Discussion

Traditional scRNA-seq analyses are based on gene expression profiles, which aim to find new cell types with unique functions. In this paper, we present a new method SCEN to analyze scRNA-seq data based on gene-gene associations. From a dynamical viewpoint, gene associations can more reliably and stably characterize the feature/state of each cell, which may help us find the cell diver-

sity that is ignored by traditional gene expression analysis. Based on SCEN, we define association-entropy (AE) for each cell to measure the association strength of gene coexpression at single-cell resolution, and we find the AE of ribosomal protein genes (RP genes) varied greatly even in the same cell type of immune cells.

In order to rule out the possibility that the changes of RP genes are randomly arisen from the use of particular technological platforms, we further explored the relationship between the AE of RP genes and diseases. We found the average AE of RP genes of immune cells in each person is significantly associated with the healthy/disease state of this person, while AE of RP genes is not associated with the sex or age of donors. Thus, the possibility that technical artifacts lead to the changes of RP genes is considered to be quite low.

In traditional opinion, ribosomes act as a translation machine in cells related to protein expression, in which ribosomal proteins (RP) and ribosomal RNA (rRNA) are the main components. However, more and more studies have found that ribosomes are also involved in DNA repair, regulation of cell development and stress response [38–40]. Moreover, the heterogeneity of ribosomes has been confirmed. Genuth et al. [8] used the state-of-the-art quantitative mass spectrometry measurements to measure the absolute abundance of 15 of the 80 core RPs in mouse embryonic stem cells, and found that 4 of the 15 RPs existed in only 60–70% ribosomes. This result indicates that ribosomes are different depending on the composition of RPs, and some ribosomes are not complete but deficient. In fact, even small changes of RPs could have large effects on translation given that millions of ribosomes are present in a single cell. Further studies found that different ribosomes have different physiological functions, depending on mRNA selectivity [9,10]. For example, Ferretti et al. [11] found that Rps26 contributes to mRNA-specific translation by recognition of the Kozak sequence in well-translated mRNAs, and Rps26-deficient ribosomes preferentially translate the mRNA from stress-response pathways. This result confirms the relationship between RP composition and protein translation. Therefore, we infer that AE of RP genes represents the heterogeneity of ribosomes based on existing research and theory.

The relationship between immune response and ribosomes has been concerned for twenty years [13]. Kinetic assays indicated that CD8 + T cells are activated only one hour after adding influenza viruses to cells [41,42]. This rapid response is essential to kill infected cells before progeny viruses are released. However, from the kinetics of protein translation, cells are impossible to produce sufficient level of viral peptides and present on the cell surface to activate T cells [43,44], so there must be a special mechanism for rapid immune response. Many studies have confirmed a unique translation mechanism, including CUG and other non AUG codes can be used to initiate translation of antigenic peptides [45]; CUG initiation of antigenic peptides uses eIF2A in place of eIF2 as translation initial factor [46]; the CUG initiation pathway is activated by pro-inflammatory stimuli, including influenza A virus (IAV) infection [47]. Especially, Lee et al. [12] demonstrated special relevance of ribosome heterogeneity to viral infections by knocking down each RP in cultured cells: 61 RPs were required for cell viability, while 8 RPs were required for vesicular stomatitis virus infection but not cell viability. Yewdell [13,14] proposed a hypothesis of immunoribosomes, which suggested that antigenic peptides derived from defective ribosomal products (DRiPs), i.e. there is a special ribosome that is particularly good at synthesizing antigens.

Recently, with the advent of ribosome analysis data, DRiPs has been recognized as an important source of tumor-specific antigens [48], and plays a critical role in immunesurveillance of viruses [49]. The non-canonical translation of mRNA by ribosomes is the major cause of DRiPs. Wei et al. knocked down 80 ribosomal proteins separately and found that 14 ribosomal proteins can regulate specific peptides or alter MHC class I expression. Knockdown of RPS28 increased the total number of peptides in cells by promoting DRiPs synthesis from non-canonical translations of "untranslated" regions (UTRs) and non-AUG start codons in mRNA, and enhanced the ability of CD8 + T cells to kill melanoma cells [50]. These results demonstrated that ribosomal protein mutations are closely related to DRiPs synthesis. Patel et al. found that RPL23-pertubed cells can resist T-cell mediated lysis, demonstrating that RPL23 negatively regulates CD8 + T cells to kill melanoma cells [51]. These previous findings support the relevance of ribosomal proteins to immunesurveillance.

In this study, we clearly illustrate the negative correlation between AE of RP genes and the activity of various immune cells, which implies that the deficient ribosomes (low AE of RP genes) are more enriched in active immune cells. This result is consistent with the DRiPs hypothesis and provides new evidence for the relationship between ribosome heterogeneity and immune response. In addition, our analysis also found that there was no significant difference on transcriptomic level between the immune cells of high AE of RP genes and low AE of RP genes. This result is also consistent with the experimental evidence that knockdown of several RPs had minor effects on the transcriptome, which suggests that the effects of RPs on antigen presentation may arise directly from the affected ribosomes rather than the downstream "ribosomal stress" effects [52].

Generally, it is hard to measure gene expression and detect the activity of immune cells simultaneously at a single-cell level by current experimental technology, but our network-based method opens a new way, which is only based on scRNA-seq data. As the activity of immune cells is strongly associated with many diseases, our method provides a new tool to discover the change of immune cells during the disease progression. Clinically, human immunotherapy urgently requires more detailed characterization of immune cell response at the level of target peptides. As a new bioinformatics tool for single-cell RNA-seq data, SCEN method can predict tumor-specific ribosomal protein alterations for individual patients, which provides a new perspective to the improvement of immunotherapy.

5. Conclusions

In this paper, we present a new network-based method SCEN for scRNA-seq analysis. SCEN can identify all known cell types that have been discovered by traditional gene-expression-based analysis. Furthermore, SCEN can detect the activity of immune cells based on AE of RP genes that characterizes the heterogeneity of ribosomes and implicitly reflects the differences at the protein level. Analyses of several public datasets on immune cells indicate that the average AE of RP genes for each person is significantly associated with the healthy/disease state of this person, and AE of RP genes is negatively correlated with the activity of immune cells. We believe that SCEN method provides a new bioinformatics tool to measure gene expression and detect the activity of immune cells simultaneously at single-cell level. In addition to single-cell data, clearly SCEN can be directly applied to bulk tissue data to construct single-sample gene entropy/association network for network biomarker analysis [53,54] and dynamic network biomarker (DNB) analysis [55–59]. And SCEN method may provide more biological insights into the heterogeneity of single cells and discover further other biological processes beyond gene expression.

6. Availability

More details of our SCEN method are illustrated in the Supplementary Information. The source code in this work is available in

the Hao Dai's repository: https://github.com/wys8c764/SCEN, which can be only used for learning and communication.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.06.056.

References

[1] Grün D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature 2015;525(7568):251–5.

[2] Guo W, Li L, He J, Liu Z, Han M, Li F, et al. Single-cell transcriptomics identifies a distinct luminal progenitor cell type in distal prostate invagination tips. Nat Genet 2020;52(9):908–18.

[3] Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science 2014;343(6172):776–9.

[4] Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat Rev Genet 2013;14(9):618–30.

[5] Dai H, Jin Q-Q, Li L, Chen L-N. Reconstructing gene regulatory networks in single-cell transcriptomic data analysis. Zool Res 2020;41(6):599–604.

[6] Dai H, Li L, Zeng T, Chen L. Cell-specific network constructed by single-cell RNA sequencing data. Nucleic Acids Res 2019;47(11):e62.

[7] Liu X, Wang Y, Ji H, Aihara K, Chen L. Personalized characterization of diseases using sample-specific networks. Nucleic Acids Res 2016;44(22):e164.

[8] Genuth NR, Barna M. The Discovery of Ribosome Heterogeneity and Its Implications for Gene Regulation and Organismal Life. Mol Cell 2018;71 (3):364–74.

[9] Shi Z, Fujii K, Kovary KM, Genuth NR, Röst HL, Teruel MN, et al. Heterogeneous Ribosomes Preferentially Translate Distinct Subpools of mRNAs Genome-wide. Mol Cell 2017;67(1):71–83.e07.

[10] Slavov N, Semrau S, Airoldi E, Budnik B, Av O. Differential Stoichiometry among Core Ribosomal Proteins. Cell Rep 2015;13(5):865–73.

[11] Ferretti MB, Ghalei H, Ward EA, Potts EL, Karbstein K. Rps26 directs mRNA-specific translation by recognition of Kozak sequence elements. Nat Struct Mol Biol 2017;24(9):700–7.

[12] Lee A-S-Y, Burdeinick-Kerr R, Whelan SPJ. A ribosome-specialized translation initiation pathway is required for cap-dependent translation of vesicular stomatitis virus mRNAs. Proc Natl Acad Sci U S A 2013;110(1):324–9.

[13] Yewdell JW, Antón LC, Bennink JR. Defective ribosomal products (DRiPs): a major source of antigenic peptides for MHC class I molecules? J Immunol 1996;157(5):1823–6.

[14] Yewdell JW, Nicchitta CV. The DRiP hypothesis decennial: support, controversy, refinement and extension. Trends Immunol 2006;27(8):368–73.

[15] Kim KT, Lee HW, Lee HO, Song HJ, Jeong DE, Shin S, et al. Application of single-cell RNA sequencing in optimizing a combinatorial therapeutic strategy in metastatic renal cell carcinoma. Genome Biol 2016;17:80.

[16] Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat Biotechnol 2015;33(2):155–60.

[17] Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol 2014;32(4):381–6.

[18] Kolodziejczyk AA, Kim JK, Tsang JCH, Ilicic T, Henriksson J, Natarajan KN, et al. Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. Cell Stem Cell 2015;17(4):471–85.

[19] Chu LF, Leng N, Zhang J, Hou Z, Mamott D, Vereide DT, et al. Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. Genome Biol 2016;17(1):173.

[20] Darmanis S, Sloan SA, Zhang Y, Enge M, Caneda C, Shuer LM, et al. A survey of human brain transcriptome diversity at the single cell level. Proc Natl Acad Sci U S A 2015;112(23):7285–90.

[21] Wang B, Zhu J, Pierson E, Ramazzotti D, Batzoglou S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. Nat Methods 2017;14(4):414–6.

[22] Lvd M, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res 2008;9:2579–605.

[23] Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 2018;36(5):411–20.

[24] Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun 2017;8:14049.

[25] Zhao Y, Li X, Zhao W, Wang J, Yu J, Wan Z, et al. Single-cell transcriptomic landscape of nucleated cells in umbilical cord blood. GigaScience 2019;8 (5):giz047..

[26] Tirosh I, Izar B, Prakadan 2nd SM, MHW, Treacy D, Trombetta JJ,, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science 2016;352(6282):189–96.

[27] Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, et al. Reversed graph embedding resolves complex single-cell trajectories. Nat Methods 2017;14 (10):979–82.

[28] Homsy E, Das S, Consiglio P, McAtee C, Zachman A, Nagaraja H, et al. Circulating Gasdermin-D in Critically Ill Patients. Crit Care Explor 2019;1(9): e0039.

[29] Araki K, Morita M, Bederman AG, Konieczny BT, Kissick HT, Sonenberg N, et al. Translation is actively regulated during the differentiation of CD8 + effector T cells. Nat Immunol 2017;18(9):1046–57.

[30] Chua RL, Lukassen S, Trump S, Hennig BP, Wendisch D, Pott F, et al. COVID-19 severity correlates with airway epithelium-immune cell interactions identified by single-cell analysis. Nat Biotechnol 2020;38(8):970–9.

[31] Kim N, Kim HK, Lee K, Hong Y, Cho JH, Choi JW, et al. Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. Nat Commun 2020;11(1):2285.

[32] Wang Z, Geng Z, Shao W, Liu E, Zhang J, Tang J, et al. Cancer-derived sialylated IgG promotes tumor immune escape by binding to Siglecs on effector T cells. Cell Mol Immunol 2020;17(11):1148–62.

[33] Montes P, Bernal M, Campo LN, González-Ramírez AR, Jiménez P, Garrido P, et al. Tumor genetic alterations and features of the immune microenvironment drive myelodysplastic syndrome escape and progression. Cancer Immunol Immunother 2019;68(12):2015–27.

[34] Garrido-Navas C, Miguel-Perez Dd, Exposito-Hernandez J, Bayarri C, Amezcua V, Ortigosa A, et al. Cooperative and Escaping Mechanisms between Circulating Tumor Cells and Blood Constituents. Cells 2019;8(11):1382..

[35] Jaeghere EAD, Denys HG, Wever OD. Fibroblasts Fuel Immune Escape in the Tumor Microenvironment. Trends Cancer 2019;5(11):704–23.

[36] Frazzette N, Khodadadi-Jamayran A, Doudican N, Santana A, Felsen D, Pavlick AC, et al. Decreased cytotoxic T cells and TCR clonality in organ transplant recipients with squamous cell carcinoma. NPJ Precis Oncol 2020;4:13.

[37] Zavidij O, Haradhvala NJ, Mouhieddine TH, Sklavenitis-Pistofidis R, Cai S, Reidy M, et al. Single-cell RNA sequencing reveals compromised immune microenvironment in precursor stages of multiple myeloma. Nat Cancer 2020;1:493–506.

[38] Lee SB, Kwon I-S, Park J, Lee K-H, Ahn Y, Lee C, et al. Ribosomal protein S3, a new substrate of Akt, serves as a signal mediator between neuronal apoptosis and DNA repair. J Biol Chem 2010;285(38):29457–68.

[39] Prakash V, Carson BB, Feenstra JM, Dass RA, Sekyrova P, Hoshino A, et al. Ribosome biogenesis during cell cycle arrest fuels EMT in development and disease. Nat Commun 2019;10(1):2110.

[40] Yang C, Zang W, Ji Y, Li T, Yang Y, Zheng X. Ribosomal protein L6 (RPL6) is recruited to DNA damage sites in a poly(ADP-ribose) polymerase-dependent manner and regulates the DNA damage response. J Biol Chem 2019;294 (8):2827–38.

[41] Esquivel F, Yewdell J, Bennink J. RMA/S cells present endogenously synthesized cytosolic proteins to class I-restricted cytotoxic T lymphocytes. J Exp Med 1992;175(1):163–8.

[42] Chen W, Pang K, Masterman K-A, Kennedy G, Basta S, Dimopoulos N, et al. Reversal in the immunodominance hierarchy in secondary CD8+ T cell responses to influenza A virus: roles for cross-presentation and lysis-independent immunodomination. J Immunol 2004;173(8):5021–7.

[43] Goldberg AL, Dice JF. Intracellular protein degradation in mammalian and bacterial cells. Annu Rev Biochem 1974;43:835–69.

[44] Goldberg AL, John ACS. Intracellular protein degradation in mammalian and bacterial cells: Part 2. Annu Rev Biochem 1976;45:747–803.

[45] Shastri N, Nguyen V, Gonzalez F. Major histocompatibility class I molecules can present cryptic translation products to T-cells. J Biol Chem 1995;270 (3):1088–91.

[46] Starck SR, Jiang V, Pavon-Eternod M, Prasad S, McCarthy B, Pan T, et al. Leucine-tRNA initiates at CUG start codons for protein synthesis and presentation by MHC class I. Science 2012;336(6089):1719–23.

[47] Prasad S, Starck SR, Shastri N. Presentation of Cryptic Peptides by MHC Class I Is Enhanced by Inflammatory Stimuli. J Immunol 2016;197(8):2981–91.

[48] Dersh D, Hollý J, Yewdell JW. A few good peptides: MHC class I-based cancer immunosurveillance and immunoevasion. Nat Rev Immunol 2021;21(2):116–28.

[49] Croft NP, Smith SA, Wong YC, Tan CT, Dudek NL, Flesch IEA, et al. Kinetics of antigen expression and epitope presentation during virus infection. PLoS Pathog 2013;9(1):e1003129.

[50] Wei J, Kishton RJ, Angel M, Conn CS, Dalla-Venezia N, Marcel V, et al. Ribosomal Proteins Regulate MHC Class I Peptide Generation for Immunosurveillance. Mol Cell. 2019;73(6):1162-73.e5.

[51] Patel SJ, Sanjana NE, Kishton RJ, Eidizadeh A, Vodnala SK, Cam M, et al. Identification of essential genes for cancer immunotherapy. Nature 2017;548(7669):537–42.

[52] Yewdell JW, Hollý J. DRiPs get molecular. Curr Opin Immunol 2020;64:130–6.

[53] Liu XP, Wang YT, Ji HB, Aihara K, Chen LN. Personalized characterization of diseases using sample-specific networks. Nucleic Acids Res 2016;44(22).

[54] Liu R, Wang J, Ukai M, Sewon K, Chen P, Suzuki Y, et al. Hunt for the tipping point during endocrine resistance process in breast cancer by dynamic network biomarkers. J Mol Cell Biol 2019;11(8):649–64.

[55] Chen L, Liu R, Liu ZP, Li M, Aihara K. Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. Sci Rep 2012;2:342.

[56] Yang BW, Li MY, Tang WQ, Liu WX, Zhang S, Chen LN, et al. Dynamic network biomarker indicates pulmonary metastasis at the tipping point of hepatocellular carcinoma. Nature. Communications 2018;9.

[57] Liu X, Chang X, Leng S, Tang H, Aihara K, Chen L. Detection for disease tipping points by landscape dynamic network biomarkers. Natl Sci Rev 2019;6(4):775–85.

[58] Gao R, Yan J, Li P, Chen L. Detecting the critical states during disease development based on temporal network flow entropy. Brief Bioinform 2022.

[59] Zhang CM, Zhang H, Ge J, Mi TY, Cui X, Tu FJ, et al. Landscape dynamic network biomarker analysis reveals the tipping point of transcriptome reprogramming to prevent skin photodamage (vol 13, pg 822, 2021). J Mol Cell Biol 2022.