Genomics Proteomics Bioinformatics 19 (2021) 319–329

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

METHOD

c-CSN: Single-cell RNA Sequencing Data Analysis by Conditional Cell-specific Network

Lin Li [1,2,#], Hao Dai [1,2,3,#], Zhaoyuan Fang [1,2,*], Luonan Chen [1,2,3,4,5,*]

[1] *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*
[2] *University of Chinese Academy of Sciences, Beijing 100049, China*
[3] *Shanghai Research Center for Brain Science and Brain-Inspired Intelligence, Shanghai 201210, China*
[4] *CAS Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China*
[5] *Hangzhou Institute for Advanced Study, Chinese Academy of Sciences, Hangzhou 310024, China*

Received 15 October 2019; revised 13 April 2020; accepted 8 July 2020
Available online 5 March 2021

KEYWORDS

Network flow entropy;
Cell-specific network;
Single-cell network;
Direct association;
Conditional independence

Abstract The rapid advancement of single-cell technologies has shed new light on the complex mechanisms of cellular heterogeneity. However, compared to bulk RNA sequencing (RNA-seq), single-cell RNA-seq (scRNA-seq) suffers from higher noise and lower coverage, which brings new computational difficulties. Based on statistical independence, **cell-specific network** (CSN) is able to quantify the overall associations between genes for each cell, yet suffering from a problem of overestimation related to indirect effects. To overcome this problem, we propose the c-CSN method, which can construct the conditional cell-specific network (CCSN) for each cell. c-CSN method can measure the **direct associations** between genes by eliminating the indirect associations. c-CSN can be used for cell clustering and dimension reduction on a network basis of single cells. Intuitively, each CCSN can be viewed as the transformation from less "reliable" gene expression to more "reliable" gene–gene associations in a cell. Based on CCSN, we further design **network flow entropy** (NFE) to estimate the differentiation potency of a single cell. A number of scRNA-seq datasets were used to demonstrate the advantages of our approach. 1) One direct association network is generated for one cell. 2) Most existing scRNA-seq methods designed for gene expression matrices are also applicable to c-CSN-transformed degree matrices. 3) CCSN-based NFE helps resolving the direction of differentiation trajectories by quantifying the potency of each cell. c-CSN is publicly available at https://github.com/LinLi-0909/c-CSN.

Introduction

With the development of high-throughput single-cell RNA sequencing (scRNA-seq), novel cell populations in complex tissues [1–5] can be identified and the differentiation trajectory of cell states [6–8] can be obtained, which opens a new way to understand the heterogeneity and transition of cells [9–11].

* Corresponding authors.
E-mail: fangzhaoyuan@sibs.ac.cn (Fang Z), lnchen@sibs.ac.cn (Chen L).
# Equal contribution.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2020.05.005

However, compared to traditional bulk RNA-seq data, the prevalence of high technical noise and dropout events is a major problem in scRNA-seq [12–17], which raises substantial challenges for data analysis. To analyze high-dimensional scRNA-seq data, principal component analysis (PCA), non-negative matrix factorization (NMF), and t-distributed Stochastic Neighbor Embedding (t-SNE) are widely used for dimension reduction. Subsequently, clustering methods such as hierarchical clustering, K-means, SNN-Cliq [18], Corr [19], SC3 [20], and SIMLR [21] could be applied to identify potential cell types, further corroborated with known marker genes. For developmental or differentiation studies, trajectory inference methods such as Monocle [22], TSCAN [23], and DPT [24] can be used to order cells along a pseudo temporal trajectory. Besides these approaches, several methods have been developed to offer special treatments of the dropouts in scRNA-seq data. One way is to explicitly model the dropout events during dimension reduction, *e.g.*, the zero-inflated factor analysis model developed in ZIFA [25]. Another way is to incorporate biological information, especially functional gene–gene association networks. In this direction, SCRL [26] takes another step forward by leveraging gene–gene interactions, learning a more meaningful low-dimensional projection. A recent method, netNMF-sc [27] derives a robust factorization or clustering against dropouts, by regularizing the original NMF model with a given gene correlation network. Furthermore, gene–gene correlations could also be employed to directly estimate the 'true' expression values for those observed zero counts, which is known as the data imputation approach, as exemplified by several well-known methods including SAVER [28], MAGIC [29], and scImpute [30]. However, data imputation is with some limitations, such as over-imputation of genes unexpressed in certain cell types and inducing artificial effects that may confound downstream analyses [31].

Several network inference algorithms were also developed for scRNA-seq. MTGO-SC [32] can detect the network modules of genes for each cell cluster though combing the information of network structures and annotations of genes. SCODE [33] can construct regulatory networks and expression dynamics through linear ordinary differential equations (ODEs). These methods only infer the network of a group or cluster of cells, and do not construct networks for individual cells. Recently, cell-specific network (CSN) has been proposed to infer CSNs based on scRNA-seq data [34], which elegantly infers a network for each cell. Moreover, unlike imputation methods, CSN employs a data transformation strategy, and successfully transforms the noisy and "unreliable" gene expression data to the more "reliable" gene association data, thereby alleviating the dropout problem to a certain extent. The network degree matrix (NDM) derived from CSN can be further applied in downstream single-cell analyses, which performs better than traditional expression-based methods in terms of robustness and accuracy. CSN is able to identify the dependency between two genes from single-cell data based on statistical independence. However, CSN suffers from a problem of overestimation on gene–gene associations and includes both direct and indirect associations due to interactive effects from other genes in a network. In other words, a gene pair without direct association can be falsely identified to have a link just because they both have true associations with some other genes. Thus, the gene–gene network of a cell constructed by CSN may be much denser than the real molecular network in this cell, in particular when there are many complex associations among genes.

To overcome these shortcomings of CSN, we introduce a novel computational method c-CSN, which can construct a conditional cell-specific network (CCSN) from scRNA-seq data. Specifically, c-CSN identifies direct associations between genes by filtering out indirect associations in the gene–gene network based on conditional independence. Thus, c-CSN can transform the original gene expression data of each cell to the direct and robust gene–gene association data (or network data) of the same cell. In this study, we first demonstrate that the transformed gene–gene association data not only are fully compatible with traditional analyses such as dimension reduction and clustering, but also enable us to delineate the CSN topology and its dynamics along developmental trajectories. Then, by defining the network flow entropy (NFE) on the gene–gene association data of each cell based on c-CSN, we estimate the differentiation potency of individual cells. We show that NFE can illustrate the lineage dynamics of cell differentiation by quantifying the differentiation potency of cells, which is also one of the most challenging tasks in developmental biology.

Method

Assume that x and y are two random variables, and z is the third random variable. If x and y are independent, then

$$p(x)p(y) = p(x, y) \tag{1}$$

where $p(x, y)$ is the joint probability distribution of x and y; $p(x)$ and $p(y)$ are the marginal probability distributions of x and y, respectively.

If x and y with the condition z are conditionally independent, then

$$p(x|z)p(y|z) = p(x, y|z) \tag{2}$$

where $p(x, y|z)$ is the joint probability distribution of x and y with the condition z, $p(x|z)$ and $p(y|z)$ are conditionally marginal probability distributions. Note that Equations (1) and (2) are both necessary and sufficient conditions on mutual independence and conditional independence, respectively. Here, we define

$$\rho_{xy} = p(x, y) - p(x)p(y) \tag{3}$$

$$\rho_{xy|z} = p(x, y|z) - p(x|z)p(y|z) \tag{4}$$

The original CSN method uses ρ_{xy} to distinguish the independency and association between x and y (File S1 Note 1). However, if two independent variables x and y are both associated with a third random variable z, ρ_{xy} cannot measure the direct independency because there is an indirect association between x and y. In other words, the associations defined by CSN or Equation (3) include both direct and indirect dependencies, thus resulting in the overestimation on gene–gene associations. To overcome this problem of CSN, we develop a novel method, c-CSN, which measures the direct gene–gene associations based on the conditional independency $\rho_{xy|z}$, *i.e.*, Equation (4), by filtering out the indirect associations in the reconstructed network. The computational framework of c-CSN is shown in **Figure 1**, and the method is described in next sections.

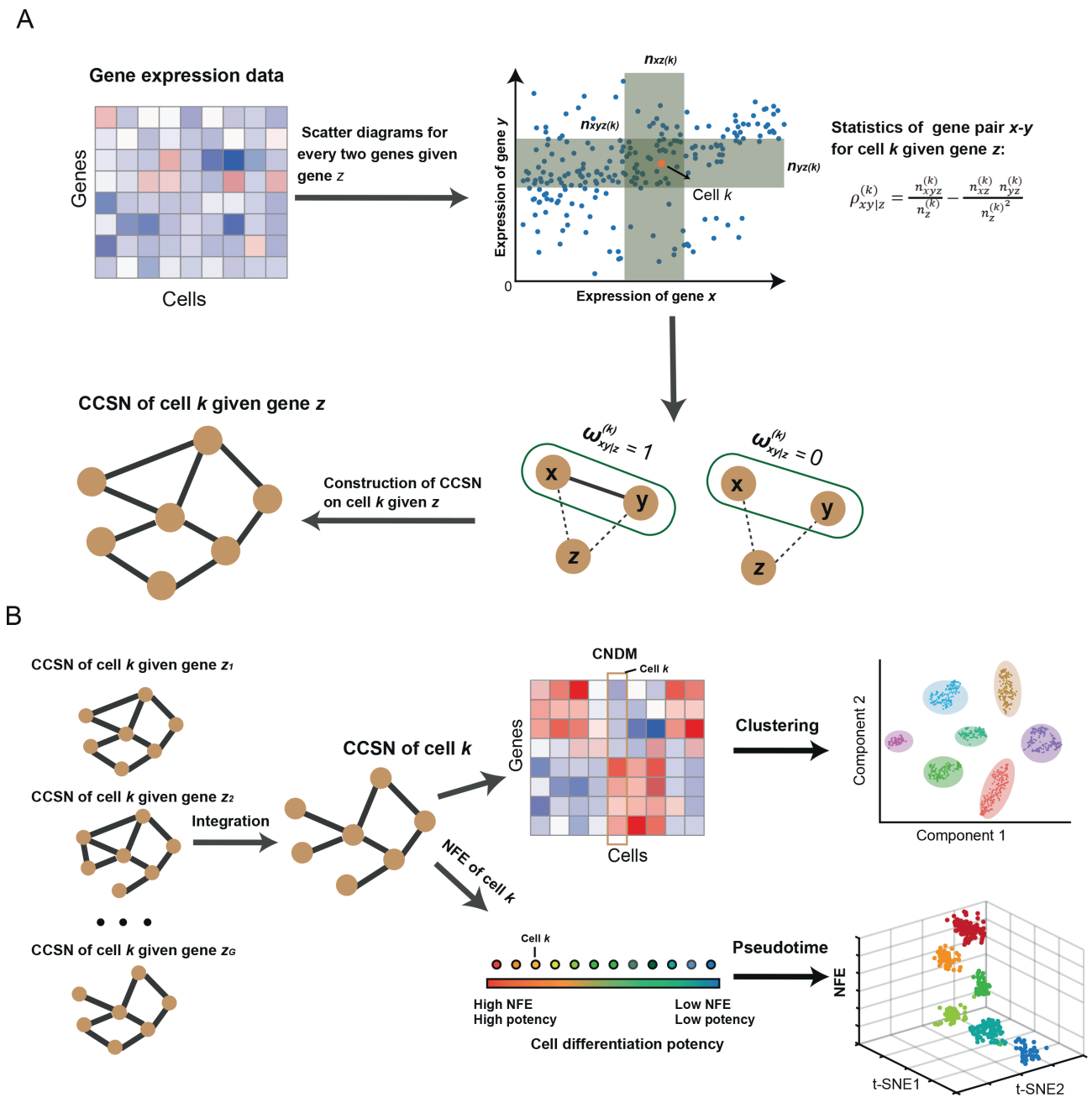

Figure 1 Overview of c-CSN method

A. The CCSN of each cell, *e.g.*, cell k, is constructed given a conditional gene z using gene expression data. For every two genes, *e.g.*, gene x and gene y, we use the statistics $\rho_{xy|z}^{(k)}$ to measure whether gene x and gene y are conditionally independent given gene z. If gene x and gene y are directly dependent ($\omega_{xy|z}^{(k)} = 1$), there is an edge between gene x and gene y in CCSN of cell k given gene z. Otherwise, there is no edge between gene x and gene y in CCSN of cell k given gene z. **B.** Construction of CCSNs of cell k given conditional gene z_i ($i = 1, 2, 3, \cdots, G$). These CCSNs of cell k given conditional gene z_i ($i = 1, 2, 3, \cdots, G$) can be integrated into a CCSN of cell k. On the one hand, CCSN of cell k can be applied to compute NFE of cell k. The pseudotime of cells can be obtained based on NFE of cells. On the other hand, CNDM can be calculated by obtaining the CCSN of each cell. The CNDM can be further applied to clustering analysis. CCSN, conditional cell-specific network; CNDM, conditional network degree matrix; NFE, network flow entropy; t-SNE, t-distributed Stochastic Neighbor Embedding.

Probability distribution estimation

We numerically estimate the value of $\rho_{xy|z}$ by making a scatter diagram based on gene expression data. Suppose there are m genes and n cells in the data. We depict the expression values of gene x, gene y, and the conditional gene z in a three-dimensional space (Figure S1A–G), where each dot represents one cell. First, we draw two parallel planes which are orthogonal with z axis near the dot k to represent the upper and lower bounds of the neighborhoods of z_k. And the number of dots in

the space between the two parallel planes (*i.e.*, the neighborhood of z_k) is $n_z^{(k)}$ (Figure S1D). Now we get a subspace on condition of gene z. Then, we draw other four planes near the dot k, where two planes are orthogonal with x axis and the other two planes are orthogonal with y axis. We can get the neighborhoods of (x_k, z_k), (y_k, z_k), and (x_k, y_k, z_k) according to the intersection space of six planes (Figure S1E–G), where the numbers of dots are $n_{xz}^{(k)}$, $n_{yz}^{(k)}$, and $n_{xyz}^{(k)}$, respectively. Then, we can get the estimation of probability distributions:

$$p^{(k)}(x, y|z) \approx \frac{n_{xyz}^{(k)}}{n_z^{(k)}},\ p^{(k)}(x|z) \approx \frac{n_{xz}^{(k)}}{n_z^{(k)}},\ p^{(k)}(y|z) \approx \frac{n_{yz}^{(k)}}{n_z^{(k)}}$$

Based on Equation (4), we construct a statistic

$$\rho_{xy|z}^{(k)} = \frac{n_{xyz}^{(k)}}{n_z^{(k)}} - \frac{n_{xz}^{(k)} n_{yz}^{(k)}}{{n_z^{(k)}}^2} \tag{5}$$

to measure the conditional independence between gene x and gene y on the condition of gene z in cell k. And when gene x and gene y given gene z are conditionally independent, the expectation $\mu_{xy|z}^{(k)}$ and standard deviation $\sigma_{xy|z}^{(k)}$ (File S1) of the statistic $\rho_{xy|z}^{(k)}$ can be obtained:

$$\mu_{xy|z}^{(k)} = 0$$

$$\sigma_{xy|z}^{(k)} = \sqrt{\frac{n_{xz}^{(k)} n_{yz}^{(k)} \cdot \left(n_z^{(k)} - n_{xz}^{(k)}\right) \cdot \left(n_z^{(k)} - n_{yz}^{(k)}\right)}{{n_z^{(k)}}^4 \left(n_z^{(k)} - 1\right)}}$$

Then, we normalize the statistic as

$$\widehat{\rho}_{xy|z}^{(k)} = \frac{\rho_{xy|z}^{(k)} - \mu_{xy|z}^{(k)}}{\sigma_{xy|z}^{(k)}} \tag{6}$$

If gene x and y are conditionally independent on the condition of gene z, it can be proved that the normalized statistic follows the standard normal distribution (File S1 Note 1; Figure S2), and it is less than or equal to 0 when gene x and y are conditionally independent (File S1 Note 2).

Construction of CCSN

To estimate the conditional independency of gene x and gene y given the conditional gene z in cell k, we first use CSN or Equation (3) to distinguish the independence of gene x and gene y and we then use the following hypothesis test. H_0(null hypothesis): gene x and gene y are conditionally independent given gene z in cell k. H_1(alternative hypothesis): gene x and gene y are conditionally dependent given gene z in cell k. If $\hat{\rho}_{xy|z}^{(k)}$, the normalized statistic, is larger than $\mathcal{N}_\alpha$ (significance level α, $\mathcal{N}_\alpha$ is the alpha quantile of the standard normal distribution), the null hypothesis will be rejected and then $\omega_{xy|z}^{(k)} = 1$ ($\omega_{xy|z}^{(k)}$ is the edge weight of genes x and y on condition of gene z).

$$\omega_{xy|z}^{(k)} = \begin{cases} 1 & \text{genes } x \text{ and } y \text{ are directly dependent given gene } z \\ 0 & \text{genes } x \text{ and } y \text{ are conditionally independent given gene } z \end{cases} \tag{7}$$

All gene pairs can be tested if they are conditionally independent given gene z in cell k. And the CCSN $C_z^{(k)}$ given conditional gene z is obtained for cell k.

Then, to estimate the direct association between a pair of genes in a cell, theoretically we should use all the remaining $m - 2$ genes as conditional genes, which is computationally intensive. Suppose there are m genes in our analysis, then $m \times (m - 1)/2$ gene pairs should be tested. Fortunately, a molecular network is generally sparse, which means that a pair of genes (*i.e.*, genes x and y) are expected to have a very small number of commonly interactive genes (as conditional genes z). In other words, numerically we can use a small number of conditional genes to identify the direct association between a pair of genes in a cell, which can significantly reduce the computational cost (File S1 Note 3; Table S1). For each gene pair in a cell, we choose G ($1 \leq G \leq m - 2$) genes as the conditional genes to test if the gene pair is conditionally independent or not. Generally, the conditional genes may be the key regulatory genes in a biological process, such as transcription factor genes and kinase genes. From a network viewpoint, these genes are usually hub genes in the gene–gene network, and the network degrees of these genes would be higher.

Practically, the conditional genes could be obtained from many available methods, such as highly expressed genes, highly variable genes, key transcription factor genes, and the hub genes in the CSN. For the c-CSN method, the conditional gene sets were defined by CSN. Two steps were used to obtain the conditional genes although other appropriate schemes can also be used.

1) For a given cell, we first construct a CSN without the consideration of conditional genes, where the edge between gene x and gene y in cell k is determined by the following hypothesis test:
H_0(null hypothesis): gene x and gene y are independent in cell k.
H_1(alternative hypothesis): gene x and gene y are dependent in cell k.
The statistic ρ_{xy} can be used to measure the independency of genes x and y (File S1 Note 1). If ρ_{xy} is larger than a significant level, we will reject the null hypothesis and $edge_{xy}^{(k)} = 1$, otherwise $edge_{xy}^{(k)} = 0$.

$$edge_{xy}^{(k)} = \begin{cases} 1 & \text{genes } x \text{ and } y \text{ are dependent} \\ 0 & \text{genes } x \text{ and } y \text{ are independent} \end{cases}$$

Then we use $D_z^{(k)}$ to measure the importance of conditional gene z in cell k:

$$D_z^{(k)} = \sum_{y=1, y \neq z}^{M} edge_{zy}^{(k)} \tag{8}$$

Equation (8) means that if a gene is connected to more other genes, this gene is more important.

2) For a given cell k, we choose the top G ($G \geq 1$) largest 'importance' genes as the conditional genes. We assume that the conditional gene set is $\{z_g,\ g = 1, 2, 3, \cdots, G\}$, and CCSN $C_{z_g}^{(k)}$ is obtained for cell k given conditional gene z_g. The CCSNs of the cell k on the condition of gene set $\{z_g,\ g = 1, 2, 3, \cdots, G\}$ are $\{C_{z_1}^{(k)}, C_{z_2}^{(k)}, \cdots, C_{z_G}^{(k)}\}$. Then, we use

$$\overline{C}_k = \frac{1}{G}\sum_{g=1}^{G} C_{z_g}^{(k)} = \left(c_{ij}^{(k)}\right) \tag{9}$$

to represent the degrees of gene–gene interaction network of cell k, where $c_{ij}^{(k)}$ for $i,j = 1,\cdots,m$ is the (i, j) element of the matrix $\overline{C}_k$.

For scRNA-seq data with all n cells, we can construct n CCSNs, which can be used for further dimension reduction and clustering. In other words, instead of the originally measured gene expression data with n cells, we use the n transformed CCSNs for further analysis.

Network degree matrix from CCSN

CCSN could be used for various biological studies by exploiting the gene–gene conditional association network from a network viewpoint. We transform Equation (9) to a conditional network degree vector based on the following transformation

$$v_{ik} = \sum_{j=1}^{m} c_{ij}^{(k)} \tag{10}$$

Then, for $\left\{\overline{C}_1, \overline{C}_2, \cdots, \overline{C}_n\right\}$, an $m \times n$ matrix conditional network degree matrix (CNDM) is obtained.

$$CNDM = (v_{ik}) \text{ with } i = 1,\cdots,m;\ k = 1,\cdots,n \tag{11}$$

The matrix has the same dimension with the gene expression matrix (GEM), *i.e.*, $GEM = (x_{ik})$ (with $i = 1,\cdots,m$; $k = 1,\cdots,n$), but CNDM can reflect the gene–gene direct association in terms of interaction degrees. Moreover, this CNDM matrix after normalization could be further analyzed by most traditional scRNA-seq methods for dimension reduction and clustering analysis. The input/output settings as well as application fields of our c-CSN method are listed in File S1 Note 4.

Network analysis of c-CSN

The relationship between gene pairs can be obtained by c-CSN at a single-cell level. c-CSN also provides a new way to build gene–gene interaction network for each cell. And the CNDM derived from CCSNs can be further used in dimension reduction, clustering and NFE analysis by many existing methods.

Dimension reduction

We used PCA [35] and t-SNE [36] which respectively represent linear and nonlinear methods, to perform dimension reduction on public scRNA-seq datasets with known cell types.

Clustering

To validate the good performance of c-CSN in clustering analysis, several traditional clustering methods such as K-means, Hierarchical clustering analysis, and K-medoids were applied to clustering analysis. Furthermore, state-of-the-art scRNA-seq data clustering methods such as SC3, SIMLR, and Seurat [20,21,37] were also used for comparison.

NFE analysis

Quantifying the differentiation potency of a single cell is one of the important tasks in scRNA-seq studies [15,38,39]. A recent study developed SCENT [40], which uses protein–protein interaction (PPI) network and gene expression data as input to obtain the potency of cells. However, SCENT depends on the PPI network, which may ignore many important relationships between genes in specific cells. In this study, we developed NFE to estimate the differentiation potency of a cell from its CSN or CCSN, which is constructed for each cell. The normalized gene expression profile and CSN/CCSN are used when we compute the NFE. The value of NFE is expected to be lower for differentiated cells, since differentiation is accompanied by activation of a specific subnetwork, which actually diverts the signaling flux from other parts of the network.

Estimating NFE requires a background network, which could be provided by CSN or CCSN. Based on CSN or CCSN, we could know whether or not there is an edge between gene i and gene j. We assume that the weight of an edge between gene i and gene j, p_{ij}, is proportional to the normalized expression levels of gene i and gene j, that is $p_{ij} \propto x_i x_j$ with $\sum_{j=1}^{m} p_{ij} = 1$. These weights are interpreted as interaction probabilities. Then, we normalize the weighted network as a stochastic matrix, $P = \left(p_{ij}\right)$ with

$$p_{ij} = \frac{x_j}{\sum_{k\in E(i)} x_k} = \frac{x_j}{(Ax)_i} \text{ for } i,j = 1,\cdots,m$$

where $E(i)$ contains the neighbors of gene i, and A is the CSN or CCSN ($A_{ij} = 1$ if i and j are connected, otherwise $A_{ij} = 0$).

And then, we define the NFE as:

$$NFE = -\sum_{i,j} x_i p_{ij} \log\left(x_i p_{ij}\right) \tag{12}$$

where x_i is the normalized gene expression of gene i. From the definition, NFE is clearly different from network entropy.

Datasets used

Twelve scRNA-seq datasets and one bulk RNA-seq dataset [15,40–47] were used to validate our c-CSN method. The number of cells in these datasets ranges from 100 to 20,000. Table S2 gives a brief introduction of these datasets.

Results and discussion

Visualization and clustering of scRNA-seq datasets with CNDM

Characterizing cell heterogeneity is one of the important tasks for scRNA-seq data analysis. To test whether CCSN-transformed network data can help segregate cell types, we performed dimension reduction and clustering on the CNDMs of gold-standard scRNA-seq datasets, using algorithms widely employed in scRNA-seq studies. The numbers of conditional genes used in CCSN construction are listed in Table S2.

For visualizing the structure of these datasets in a two-dimensional space, we used the representative linear and nonlinear dimension reduction methods, PCA [48] and t-SNE [36], respectively. As shown in **Figure 2** and Figure S3, CNDMs can separate different cell types clearly in the low-dimensional space by both PCA and t-SNE. Notably, they generally perform even better than GEM (Figure 2, Figure S3). Hence,

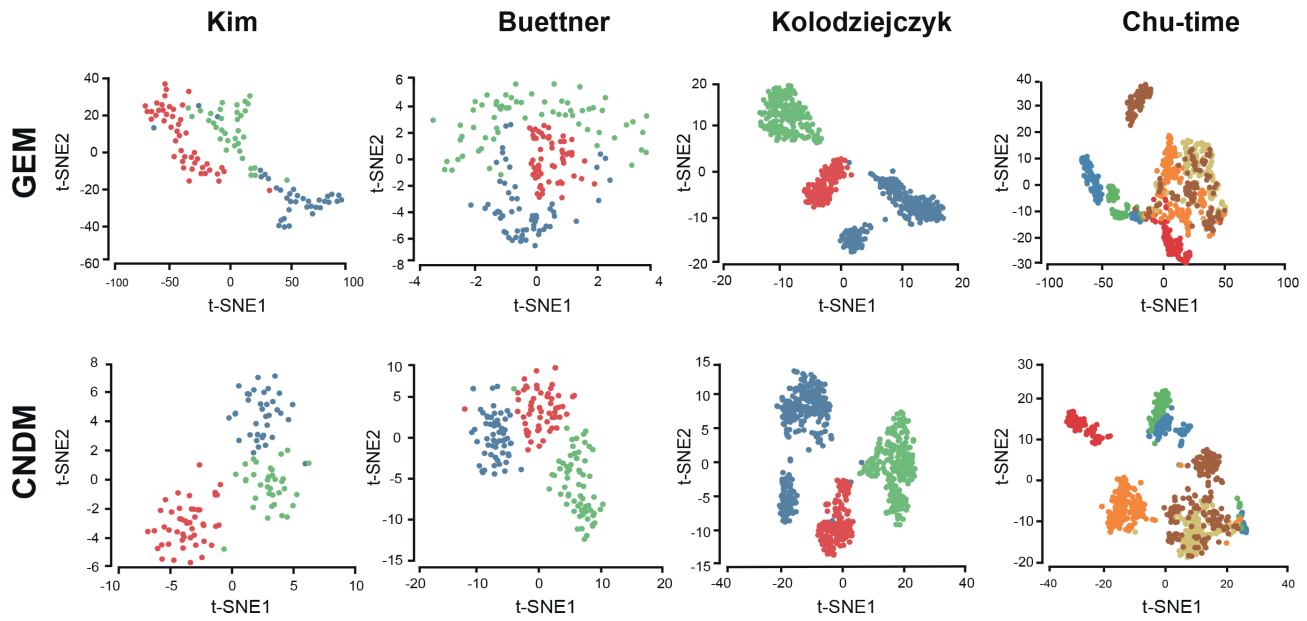

Figure 2 Comparison of traditional GEM and CNDM on visualization

GEM (top panel) and CNDM (bottom panel) are benchmarked for visualizing four scRNA-seq datasets (Kim [43], Buettner [15], Kolodziejczyk [41], and Chu-time [42]) with t-SNE. Different colors represent different cell types. GEM, gene express matrix.

the network data of CNDMs contain sufficient information for separating cell types in scRNA-seq datasets.

To quantitatively evaluate the power of CNDMs in cell type identification, we performed clustering on CNDMs and computed the adjusted Rand index (ARI) for each dataset based on the background truth (File S1 Note 5; Figure S4). As shown in **Table 1** and Figure S5, CNDM performs obviously better than GEM on all datasets. These provide a strong support of the notion that the CCSN-transformed network data are highly informative for characterizing single-cell populations. Interestingly, when further compared to NDM, CNDM also shows a good performance (**Table 2**; Figure S6).

We further evaluated the performance of c-CSN in larger datasets. The Tabula Muris droplet1 dataset [47] comprising more than 20,000 cells from three tissues (bladder, trachea, and spleen) were tested. The Seurat package was used to perform dimension reduction and clustering analysis on the CNDM [37]. The cells were clearly segregated into three dominant groups in the t-SNE map, which were largely defined by their cell origins (ARI = 0.73 and Figure S7). This indicates that CCSN can be effectively extended to larger datasets in addition to the relatively small gold-standard datasets benchmarked above.

CCSN reveals network structure and dynamics on a single-cell basis

In this study, we applied c-CSN to Wang dataset [45], which comes from a study of neural progenitor cells (NPCs) that dif-

Table 1 Performance comparison of CNDM and GEM in clustering of scRNA-seq data

Method	Input	Buettner [15]	Kolodziejczyk [41]	Gokce [46]	Chu-time [42]	Chu-type [42]	Kim [43]
K-means	GEM	0.29	0.54	0.42	0.17	0.22	0.20
	CNDM	**0.87**	**0.85**	**0.75**	**0.45**	**0.57**	**0.81**
Hierarchical	GEM	0.32	0.49	0.47	0.22	0.22	0.12
	CNDM	**0.73**	**0.65**	**0.92**	**0.47**	**0.61**	**0.77**
K-means (t-SNE)	GEM	0.41	0.87	0.43	0.33	0.55	0.53
	CNDM	**0.95**	**0.91**	0.36	0.56	**0.70**	**0.93**
Hierarchical (t-SNE)	GEM	0.55	0.99	0.50	0.39	0.67	0.73
	CNDM	**0.95**	0.99	0.39	**0.61**	**0.80**	**0.95**
K-medoids	GEM	0.23	0.29	0.40	0.33	0.33	0.79
	CNDM	**0.53**	**0.63**	**0.81**	0.17	**0.38**	**0.61**
SC3	GEM	0.89	1	0.56	0.66	0.78	0.89
	CNDM	**0.98**	0.72	**0.72**	0.63	**0.98**	**0.96**
SIMLR	GEM	0.89	0.49	0.43	0.30	0.48	0.38
	CNDM	0.63	**0.52**	**0.85**	**0.58**	**0.54**	**0.95**
Seurat	GEM	0.67	0.43	0.35	0.52	0.52	0.41
	CNDM	**0.90**	**0.56**	0.32	**0.56**	**0.69**	**0.84**

Note: The performance of clustering is evaluated by ARI. Hierarchical (t-SNE) and K-means (t-SNE) indicate clustering after t-SNE. CNDM, conditional network degree matrix; GEM, gene expression matrix; ARI, adjusted Rand index; t-SNE, t-distributed Stochastic Neighbor Embedding. Bold font (ARI) indicates that CNDM performs better.

Table 2 Performance comparison of CNDM and NDM in clustering of scRNA-seq data

Method	Input	Buettner [15]	Kim [43]	Wang [45]	Gokce [46]	Tabula Muris [47] (aorta)	Tabula Muris [47] (limb muscle)
K-means	NDM	0.50	0.50	0.30	0.79	0.21	0.58
	CNDM	**0.87**	**0.81**	**0.45**	**0.75**	**0.63**	**0.66**
Hierarchical	NDM	0.69	0.59	0.38	0.95	0.12	0.65
	CNDM	**0.73**	**0.77**	**0.45**	**0.92**	**0.75**	**0.76**
K-means (t-SNE)	NDM	0.83	0.84	0.61	0.38	0.46	0.62
	CNDM	**0.95**	**0.93**	**0.67**	0.36	**0.61**	**0.65**
Hierarchical (t-SNE)	NDM	0.89	**0.98**	0.58	0.47	0.50	0.66
	CNDM	**0.95**	**0.95**	**0.72**	0.39	**0.50**	0.66
K-medoids	NDM	0.26	0.49	0.31	0.60	0.35	0.14
	CNDM	**0.53**	**0.61**	0.21	**0.81**	**0.53**	**0.39**
SC3	NDM	0.67	1	0.70	0.45	0.29	0.66
	CNDM	**0.98**	0.96	**0.86**	**0.72**	**0.73**	**0.76**
SIMLR	NDM	0.64	0.75	0.29	0.74	0.40	0.60
	CNDM	0.63	**0.95**	**0.60**	**0.85**	**0.70**	**0.71**
Seurat	NDM	0.82	0.97	0.59	0.44	0.45	0.66
	CNDM	**0.90**	0.84	0.59	0.32	**0.76**	**0.75**

Note: The performance of clustering is evaluated by ARI. Hierarchical (t-SNE) and K-means (t-SNE) indicate clustering after t-SNE. NDM, network degree matrix. Bold font (ARI) indicates that CNDM performs better.

ferentiate into mature neurons. The dataset contains six time points over a 30-day period.

The CSN and c-CSN were performed on a single cell (Day 0, RHB1742_d0) using 195 transcription factors that are differentially expressed across all the cell subpopulations and all time points. In CCSN, two genes (*HMGB1* and *SOX11*) of high coefficients of variation (CV) were chosen as the conditional genes. The results (**Figure 3**A) illustrate that the network of CCSN is much sparser than the network of CSN. There are three modules in the CCSN, while there is only one dense network in the CSN. Furthermore, three hub genes were obtained in three modules in the CCSN. One of the hub genes is *ASCL1*, which plays an important role in neural development [13,49]. Thus, by removing indirect associations, c-CSN can extract a more informative network structure than CSN, which could improve the characterization of key regulatory factors in individual cells.

c-CSN also reveals the network dynamics over the differentiation trajectory. As illustrated in Figure 3B, a core neural differentiation network composed of eight regulatory genes was dynamically modulated through the temporal progression of NPC differentiation. At Day 0, the associations among these genes were the strongest, consistent with the high potency of progenitor cells. As NPC differentiates, the network became much sparser, suggesting more specified cell fates. In addition, when constructing CCSN from all genes, the degrees of *MEIS2*, *PBX1* and *POU3F2* were also larger at Day 0 and quickly decreases afterward (Figure 3C). These indicate that these genes are highly connected with other genes in NPCs, which is consistent with their known important roles in early differentiation of NPCs [45].

Both theoretically and computationally, c-CSN can also construct a gene–gene network for a single bulk RNA-seq sample, in addition to a single cell. To validate this biologically, we applied c-CSN to the TCGA lung adenocarcinoma (LUAD) RNA-seq dataset. The t-SNE plot based on CNDM reveals two obvious clusters, which respectively corresponding to normal adjacent lung tissues and lung tumors (Figure S8A), supporting the effective application of c-CSN to bulk RNA-seq data as well. Moreover, the EGFR pathway, a well-known oncogenic driver pathway for LUAD [50–52], was densely connected in tumor samples but not in benign tissues, as illustrated in the representative single-sample EGFR networks (Figure S8B), and the CCSN degrees of EGF and EGFR in each normal and tumor samples (Figure S8C). These data demonstrate that c-CSN well extends to single sample bulk RNA-seq data analysis and uncovers important biological connections related to disease states.

CCSN-based NFE analysis

To quantify the differentiation state of cells, we further develop a new method, NFE, to estimate the differentiation potency of cells by exploiting the gene–gene network constructed by c-CSN.

To assess the performance of NFE, we applied it to two datasets. In Wang dataset [45], there were 483 cells with 6 stages (Day 0, Day 1, Day 5, Day 7, Day 10, Day 30) and the CCSNs with one conditional gene were used to compute the NFE. We compared NPCs at Day 0 and Day 1 with mature neurons at Day 30 (**Figure 4**A). In Yang dataset [44], we compared the cells at embryonic day 10 (E10) with those at embryonic day 17 (E17) in differentiation of mouse hepatoblasts (Figure 4B) and the CSN was used to compute the NFE. In both datasets, NFE assigned significantly higher scores to the progenitors than the differentiated cells (one-sided Wilcox rank sum test, $P = 2.062\text{E}{-12}$ in Wang dataset, $P = 3.756\text{E}{-19}$ in Yang dataset).

To further validate the accuracy of NFE, we generated a three-dimensional representation of the cell-lineage trajectory for the Wang dataset [45]. In the time-course differentiation experiment of NPCs into neurons [45], NFE correctly predicted a gradual decrease in differentiation potency (**Figure 5**). Therefore, NFE is effectively applicable to single-cell differentiation studies and highly predictive of developmental states and directions.

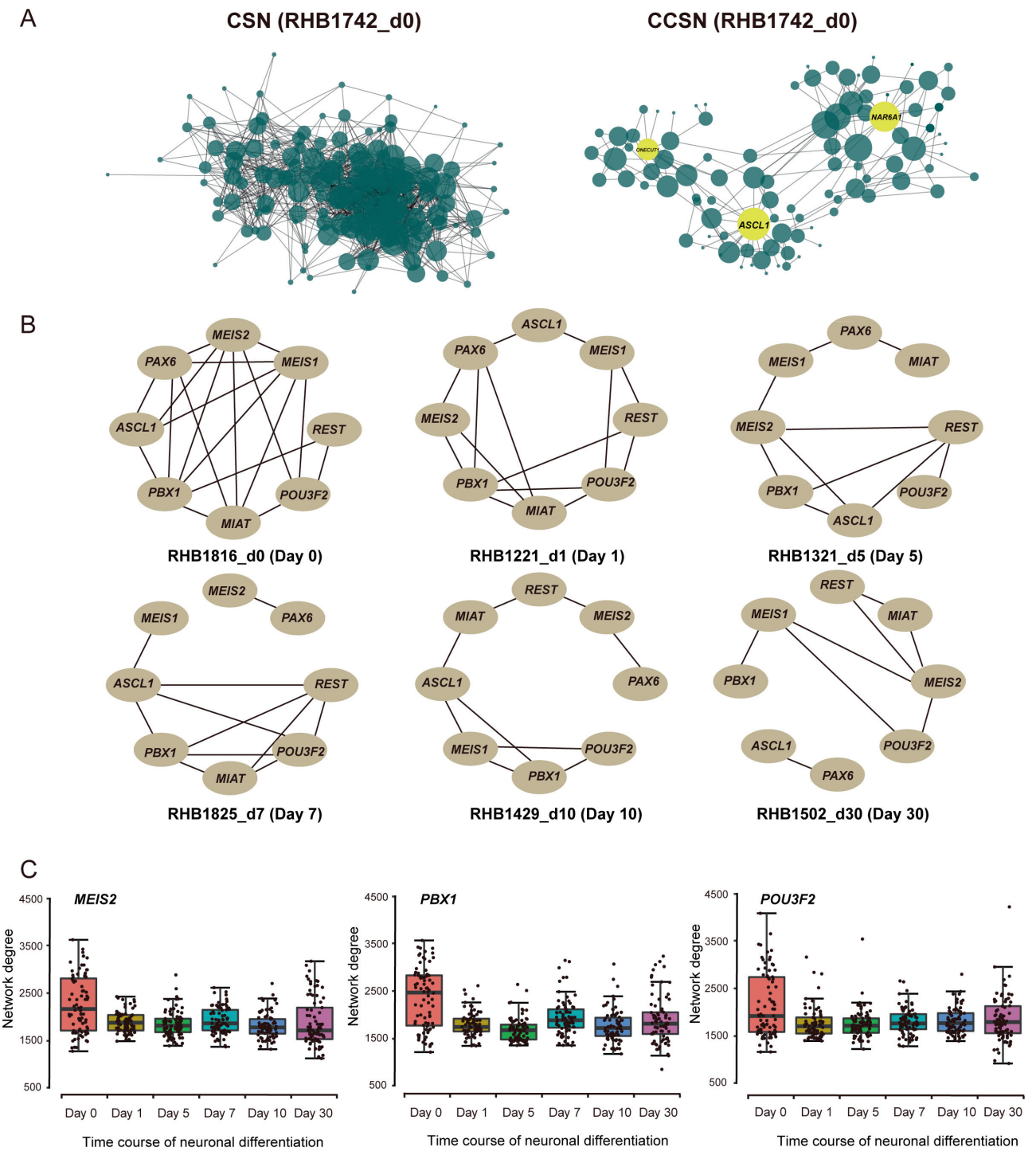

Figure 3 The network analyses for single cells based on CCSN

A. CSN and CCSN of the same single cell (RHB1742_d0) from Wang dataset [45]. The same genes are used in network construction. Three hub genes *ONECUT1*, *ASCL1*, and *NAR6A1* are highlighted in yellow. **B.** CCSNs of six cells from six time points (from Day 0 to Day 30) with eight genes that are involved in neuronal differentiation. The edge between two genes means the direct dependency of genes. **C.** The network degrees of *MEIS2*, *PBX1*, and *POU3F2* along six time points of the neuronal differentiation. Each point indicates a single cell, colored according to the time point of sampling (from Day 0 to Day 30). The midlines indicate the median levels of network degree at six time points.

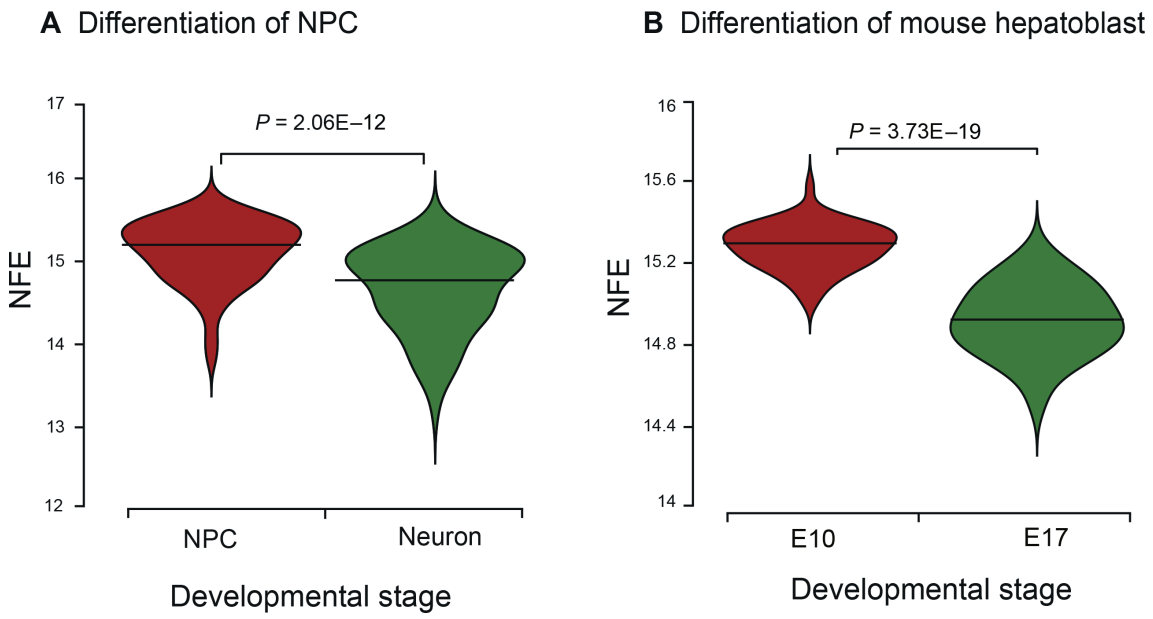

Figure 4 NFE analyses for differentiated cells and progenitors

A. Violin plot comparing NFE of NPCs at Day 0 and Day 1 and mature neurons at Day 30 during the differentiation of NPCs. **B.** Violin plot of NFE values for cells at embryonic day 10 (E10) and embryonic day 17 (E17) during differentiation of mouse hepatoblasts. Cells of various differentiation states are compared for their differentiation potency by NFE. Red represents more pluripotent cells; green represents more differentiated cells. The midlines indicate the median levels of NFE in each cell type. *P* values are from one-sided Wilcoxon rank-sum test. NPC, neural progenitor cell.

Conclusion

Estimating functional gene networks from noisy single-cell data has been a challenging task. Motivated by network-based data transformation, we have previously developed CSN to uncover CSNs and successfully applied it to extract biologically important gene interactions. However, CSN does not distinguish direct and indirect associations and thus suffers from the so-called overestimation problem. In this study, we propose a more sophisticated approach termed c-CSN, which constructs direct gene–gene associations (network) of each cell by eliminating false connections introduced by indirect effects.

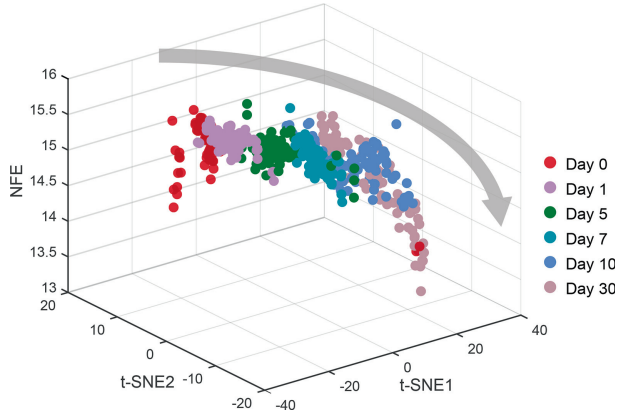

Figure 5 Differentiation landscape of single-cell data according to NFE

The three-dimensional plot shows that NFE of single cells gradually decreases along the differentiation time course of NPCs (at Day 0 and Day 1) into mature neurons (at Day 30). The grey arrow depicts the overall differentiation trajectory from stem cells to differentiated cells, consistent with the trend of decrease of NFE. Cells from the same time point are assigned the same color. The *z* axis represents the NFE. The *x* axis and *y* axis represent respectively the two components of t-SNE (t-SNE1 and t-SNE2).

c-CSN can transform GEM to CNDM for downstream dimension reduction and clustering analysis. These allow us to identify cell populations, generally better than GEM in the datasets tested above. In addition, c-CSN also shows good performance when compared to CSN. Moreover, we can construct one direct gene–gene association network by one cell based on c-CSN. From the networks of the individual cells, we can obtain the dynamically changed networks. As shown in Figure 3B, the CCSNs of these cells dynamically changed at different time points, and the network at Day 0 shows the strongest associations. Moreover, the hub genes of the networks constructed by c-CSN method may play an important role in biological processes. As shown in Figure 3A, the hub genes of three modules in the network constructed by c-CSN play a vital role in neural development. These clearly demonstrate the advantages of CCSN. In addition, individual networks of cells constructed by c-CSN can also be applied to construct network biomarkers [53,54] for accurate disease diagnosis/prognosis, or dynamic network biomarkers [55–59] for reliable disease prediction.

According to the Waddington's landscape model of cellular differentiation, cellular differentiation potency is decreased as a pluripotent cell "rolls" down from a "hill" to nearby "valleys", and cell fate transitions could be modeled as "canalization" events [60–62]. The differentiation potency quantifies the

relative number of fate choices that a cell may have and provides a useful indicator of cellular "stemness". Recently, SCENT [40] and MCE [63] use PPI network and gene expression data as input to obtain the potency of cells. However, these methods estimate the entropy of cells based on the available PPI network across various tissues, which may ignore many important relationships between genes in specific cells. Here, we develop the NFE to integrate the scRNA-seq profile of a cell with its gene–gene association network, and the results show that NFE performs well in distinguishing various cells of differential potency.

Nonetheless, the computational cost of c-CSN generally increases by G times comparing with the original CSN due to G conditional genes. Thus, a parallel computation scheme is desired to reduce the computation time. Also, c-CSN is not designed to construct the causal gene association networks, and the directions of the gene associations cannot be obtained. These could be our future research topics.

Code availability

CCSN is available at https://github.com/LinLi-0909/c-CSN.

CRediT author statement

Lin Li: Conceptualization, Methodology, Writing - review & editing. **Hao Dai:** Methodology, Writing - review & editing. **Zhaoyuan Fang:** Conceptualization, Methodology, Writing - review & editing. **Luonan Chen:** Conceptualization, Writing - review & editing. All authors read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgments

This work was supported by the National Key R&D Program of China (Grant No. 2017YFA0505500), the National Natural Science Foundation of China (Grant Nos. 31771476 and 31930022), and the Shanghai Municipal Science and Technology Major Project, China (Grant No. 2017SHZDZX01).

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gpb.2020.05.005.

ORCID

0000-0002-3558-323X (Lin Li)
0000-0003-0226-9292 (Hao Dai)
0000-0002-0393-2052 (Zhaoyuan Fang)
0000-0002-3960-0068 (Luonan Chen)

References

[1] Yan L, Yang M, Guo H, Yang L, Wu J, Li R, et al. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat Struct Mol Biol 2013;20:1131–9.
[2] Treutlein B, Brownfield DG, Wu AR, Neff NF, Mantalas GL, Espinoza FH, et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature 2014;509:371–5.
[3] Zeisel A, Munoz-Manchado AB, Codeluppi S, Lonnerberg P, La Manno G, Jureus A, et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science 2015;347:1138–42.
[4] Fuzik J, Zeisel A, Mate Z, Calvigioni D, Yanagawa Y, Szabo G, et al. Integration of electrophysiological recordings with single-cell RNA-seq data identifies neuronal subtypes. Nat Biotechnol 2016;34:175–83.
[5] Scialdone A, Tanaka Y, Jawaid W, Moignard V, Wilson NK, Macaulay IC, et al. Resolving early mesoderm diversification through single-cell expression profiling. Nature 2016;535:289–93.
[6] Bendall SC, Davis KL, Amir el AD, Tadmor MD, Simonds EF, Chen TJ, et al. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. Cell 2014;157:714–25.
[7] Nestorowa S, Hamey FK, Pijuan Sala B, Diamanti E, Shepherd M, Laurenti E, et al. A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. Blood 2016;128:e20–31.
[8] Woyke T, Doud DFR, Schulz F. The trajectory of microbial single-cell sequencing. Nat Methods 2017;14:1045–54.
[9] 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;343:776–9.
[10] Shapiro E, Biezuner T, Linnarsson S. Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat Rev Genet 2013;14:618–30.
[11] Grun 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;525:251–5.
[12] Kuznetsov VA, Knott GD, Bonner RF. General statistics of stochastic process of gene expression in eukaryotic cells. Genetics 2002;161:1321–32.
[13] Kim JK, Marioni JC. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. Genome Biol 2013;14:R7.
[14] Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods 2014;11:740–2.
[15] 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;33:155–60.
[16] Daigle Jr BJ, Soltani M, Petzold LR, Singh A. Inferring single-cell gene expression mechanisms using stochastic simulation. Bioinformatics 2015;31:1428–35.
[17] Vu TN, Wills QF, Kalari KR, Niu N, Wang L, Rantalainen M, et al. Beta-Poisson model for single-cell RNA-seq data analyses. Bioinformatics 2016;32:2128–35.
[18] Xu C, Su Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. Bioinformatics 2015;31:1974–80.
[19] Jiang H, Sohn LL, Huang H, Chen L. Single cell clustering based on cell-pair differentiability correlation and variance analysis. Bioinformatics 2018;34:3684–94.

[20] Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, et al. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods 2017;14:483–6.
[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;14:414–6.
[22] 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;32:381–6.
[23] Ji Z, Ji H. TSCAN: pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. Nucleic Acids Res 2016;44:e117.
[24] Angerer P, Haghverdi L, Buttner M, Theis FJ, Marr C, Buettner F. Destiny: diffusion maps for large-scale single-cell data in R. Bioinformatics 2016;32:1241–3.
[25] Pierson E, Yau C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol 2015;16:241.
[26] Li X, Chen W, Chen Y, Zhang X, Gu J, Zhang MQ. Network embedding-based representation learning for single cell RNA-seq data. Nucleic Acids Res 2017;45:e166.
[27] Elyanow R, Dumitrascu B, Engelhardt BE, Raphael BJ. netNMF-sc: leveraging gene-gene interactions for imputation and dimensionality reduction in single-cell expression analysis. Genome Res 2020;30:195–204.
[28] Huang M, Wang J, Torre E, Dueck H, Shaffer S, Bonasio R, et al. SAVER: gene expression recovery for single-cell RNA sequencing. Nat Methods 2018;15:539–42.
[29] van Dijk D, Sharma R, Nainys J, Yim K, Kathail P, Carr AJ, et al. Recovering gene interactions from single-cell data using data diffusion. Cell 2018;174:716–29.e27.
[30] Li WV, Li JJ. An accurate and robust imputation method scImpute for single-cell RNA-seq data. Nat Commun 2018;9:997.
[31] Andrews TS, Hemberg M. False signals induced by single-cell imputation. F1000Res 2018;7:1740.
[32] Nazzicari N, Vella D, Coronnello C, Di Silvestre D, Bellazzi R, Marini S. MTGO-SC, a tool to explore gene modules in single-cell RNA sequencing data. Front Genet 2019;10:953.
[33] Matsumoto H, Kiryu H, Furusawa C, Ko MSH, Ko SBH, Gouda N, et al. SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. Bioinformatics 2017;33:2314–21.
[34] Dai H, Li L, Zeng T, Chen L. Cell-specific network constructed by single-cell RNA sequencing data. Nucleic Acids Res 2019;47:e62.
[35] Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philos Trans A Math Phys Eng Sci 2016;374:20150202.
[36] Der Maaten LV, Hinton GE. Visualizing data using t-SNE. J Mach Learn Res 2008;9:2579–605.
[37] Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck 3rd WM, et al. Comprehensive integration of single-cell data. Cell 2019;177:1888–902.e21.
[38] MacArthur BD, Lemischka IR. Statistical mechanics of pluripotency. Cell 2013;154:484–9.
[39] Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet 2015;16:133–45.
[40] Teschendorff AE, Enver T. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. Nat Commun 2017;8:15599.
[41] Kolodziejczyk AA, Kim JK, Tsang JC, Ilicic T, Henriksson J, Natarajan KN, et al. Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. Cell Stem Cell 2015;17:471–85.
[42] 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;17:173.
[43] Kim KT, Lee HW, Lee HO, Song HJ, Jeong da E, 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.
[44] Yang L, Wang WH, Qiu WL, Guo Z, Bi E, Xu CR. A single-cell transcriptomic analysis reveals precise pathways and regulatory mechanisms underlying hepatoblast differentiation. Hepatology 2017;66:1387–401.
[45] Wang J, Jenjaroenpun P, Bhinge A, Angarica VE, Del Sol A, Nookaew I, et al. Single-cell gene expression analysis reveals regulators of distinct cell subpopulations among developing human neurons. Genome Res 2017;27:1783–94.
[46] Gokce O, Stanley GM, Treutlein B, Neff NF, Camp JG, Malenka RC, et al. Cellular taxonomy of the mouse striatum as revealed by single-cell RNA-Seq. Cell Rep 2016;16:1126–37.
[47] Tabula Muris Consortium, Overall coordination, Logistical coordination, Organ collection and processing, Library preparation and sequencing, Computational data analysis, et al. Single-cell transcriptomics of 20 mouse organs creates a *Tabula Muris*. Nature 2018;562:367–72.
[48] Baglama J, Reichel L. Augmented implicitly restarted lanczos bidiagonalization methods. SIAM J Sci Comput 2005;27:19–42.
[49] Ming GL, Song H. Adult neurogenesis in the mammalian brain: significant answers and significant questions. Neuron 2011;70:687–702.
[50] Ohsaki Y, Tanno S, Fujita Y, Toyoshima E, Fujiuchi S, Nishigaki Y, et al. Epidermal growth factor receptor expression correlates with poor prognosis in non-small cell lung cancer patients with p53 overexpression. Oncol Rep 2000;7:603–7.
[51] Nicholson RI, Gee JM, Harper ME. EGFR and cancer prognosis. Eur J Cancer 2001;37:S9–15.
[52] Sharma SV, Bell DW, Settleman J, Haber DA. Epidermal growth factor receptor mutations in lung cancer. Nat Rev Cancer 2007;7:169–81.
[53] Zhang W, Zeng T, Liu X, Chen L. Diagnosing phenotypes of single-sample individuals by edge biomarkers. J Mol Cell Biol 2015;7:231–41.
[54] Liu X, Wang Y, Ji H, Aihara K, Chen L. Personalized characterization of diseases using sample-specific networks. Nucleic Acids Res 2016;44:e164.
[55] Yang B, Li M, Tang W, Liu W, Zhang S, Chen L, et al. Dynamic network biomarker indicates pulmonary metastasis at the tipping point of hepatocellular carcinoma. Nat Commun 2018;9:678.
[56] Liu R, Chen P, Chen L. Single-sample landscape entropy reveals the imminent phase transition during disease progression. Bioinformatics 2020;36:1522–32.
[57] 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;11:649–64.
[58] 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 2018;6:775–85.
[59] Chen C, Li R, Shu L, He Z, Wang J, Zhang C, et al. Predicting future dynamics from short-term time series using an Anticipated Learning Machine. Natl Sci Rev 2020;7:1079–91.
[60] Moris N, Pina C, Arias AM. Transition states and cell fate decisions in epigenetic landscapes. Nat Rev Genet 2016;17:693–703.
[61] Laurenti E, Gottgens B. From haematopoietic stem cells to complex differentiation landscapes. Nature 2018;553:418–26.
[62] Lang AH, Li H, Collins JJ, Mehta P. Epigenetic landscapes explain partially reprogrammed cells and identify key reprogramming genes. PLoS Comput Biol 2014;10:e1003734.
[63] Shi J, Teschendorff AE, Chen W, Chen L, Li T. Quantifying Waddington's epigenetic landscape: a comparison of single-cell potency measures. Brief Bioinform 2018;21:248–61.