OXFORD

Briefings in Bioinformatics, 2023, **24(6)**, 1–12
https://doi.org/10.1093/bib/bbad366
Problem Solving Protocol

SGAE: single-cell gene association entropy for revealing critical states of cell transitions during embryonic development

Jiayuan Zhong†, Chongyin Han†, Pei Chen and Rui Liu

Corresponding authors. R. Liu, School of Mathematics, South China University of Technology, Guangzhou 510640, China. E-mail: scliurui@scut.edu.cn; P. Chen, School of Mathematics, South China University of Technology, Guangzhou 510640, China. E-mail: chenpei@scut.edu.cn

†These authors contributed equally to this work.

Abstract

The critical point or pivotal threshold of cell transition occurs in early embryonic development when cell differentiation culminates in its transition to specific cell fates, at which the cell population undergoes an abrupt and qualitative shift. Revealing such critical points of cell transitions can track cellular heterogeneity and shed light on the molecular mechanisms of cell differentiation. However, precise detection of critical state transitions proves challenging when relying on single-cell RNA sequencing data due to their inherent sparsity, noise, and heterogeneity. In this study, diverging from conventional methods like differential gene analysis or static techniques that emphasize classification of cell types, an innovative computational approach, single-cell gene association entropy (SGAE), is designed for the analysis of single-cell RNA-seq data and utilizes gene association information to reveal critical states of cell transitions. More specifically, through the translation of gene expression data into local SGAE scores, the proposed SGAE can serve as an index to quantitatively assess the resilience and critical properties of genetic regulatory networks, consequently detecting the signal of cell transitions. Analyses of five single-cell datasets for embryonic development demonstrate that the SGAE method achieves better performance in facilitating the characterization of a critical phase transition compared with other existing methods. Moreover, the SGAE value can effectively discriminate cellular heterogeneity over time and performs well in the temporal clustering of cells. Besides, biological functional analysis also indicates the effectiveness of the proposed approach.

Keywords: critical transition; single-cell gene association entropy (SGAE); dynamic network biomarker (DNB); cell differentiation; cell fate transition

INTRODUCTION

Many early developmental processes involve a cell fate transition or critical state of cell transition, with a considerable and qualitative alteration occurring [1, 2]. From a dynamics perspective, early embryonic development is commonly interpreted as a dynamically evolving system over time, characterized by three distinct states: a stable and robust before-transition state, a highly sensitive and unstable critical states of cell transitions, and another stable after-transition state (Figure 1A) [3–5]. The detection of this critical state transition in embryonic development [6, 7] has garnered increasing attention, as it offers valuable insights into the biological mechanisms underlying potentially patient-specific tissue regeneration [8] and disease modeling [9]. However, it is considerable challenging to characterize bioprocess dynamics and effectively capture the signal that indicates cell fate transition from single-cell data because of the inherent sparsity, noise and heterogeneity characteristic of such data. Current approaches primarily concentrate on analyses of gene expression levels [6, 7], but single-cell expression data may provide deeper insights into gene–gene associations. Compared with the single-cell gene-expression pattern, gene–gene associations have demonstrated more consistent ability to characterize the biological processes or heterogeneity of cell populations [10, 11].

In this research, considering gene associations within the cell-specific network, we introduce a computational approach known as SGAE to discern the critical signals accountable for cell transitions in embryonic development. Specifically, there are three main procedures in the proposed method: the construction of cell-specific networks using a statistical dependency index, the calculation of a local SGAE for each localized network, and the detection of the critical states of cell transitions by employing the SGAE index (Figure 1B and C). Notably, such an approach enables the transformation of the inherently 'unstable' single-cell gene expression matrix into a comparatively 'stable' SGAE matrix.

Jiayuan Zhong is a research fellow in School of Mathematics and Big Data at Foshan University. His research interest mainly focuses on developing computational approaches to detect the tipping points of nonlinear dynamical systems.
Chongyin Han received the B.S. at South China University of Technology. His current research interest mainly focuses on data mining for complex dynamic systems.
Rui Liu is a full professor at the School of Mathematics, South China University of Technology. He received the B.S. and Ph.D. degrees in applied mathematics from Peking University. His research interest includes nonlinear dynamics, modeling and computational methods.
Pei Chen received her B.S. and M.S. degrees from Peking University, and Ph.D. degree from South China University of Technology. Currently she is an associate professor at the School of Mathematics, South China University of Technology. Her research interest includes deep learning, data mining and computational biology.
Received: June 27, 2023. **Revised:** August 29, 2023. **Accepted:** August 31, 2023
© The Author(s) 2023. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bib/article/24/6/bbad366/7313455 by Zhejiang University user on 22 October 2023

This SGAE matrix can be seamlessly substituted for the original expression matrix, allowing for the application of traditional scRNA-seq analysis like reducing dimensions and clustering cells. By capturing dynamic information from gene associations on the single-cell scale, SGAE can reveal a dynamic shift in cellular heterogeneity over time. In this study, five single-cell datasets for embryonic development are employed to validate our method, which indicates that the predicted critical states of cell transitions are in accordance with experimental observations. Thus, SGAE serves as a framework for analyzing scRNA-seq data and provides a quantitative approach to track the changes in biological systems over time using network entropy.

MATERIALS AND METHODS

Theoretical background

In the early phases of embryonic development, cell differentiation encompasses critical states of cell transitions [12]. This process is considered a complex dynamic system, where cell state transition is recognized a state shift occurring at the point of bifurcation [2]. The dynamic course of embryonic development can be partitioned into three phases or states (depicted in Figure 1A): a highly stable before-transition state, an unstable a critical phase denoting cell fate transition, and a subsequent after-transition state marked by heightened stability. In the vicinity of the dynamic system's critical state, a set of closely associated dynamical network biomarkers (DNBs) emerges, exhibiting significant fluctuations in collective behavior [3]. By leveraging the dynamic information of gene–gene associations within this set of variables, it is possible to predict the critical signal of a drastic or qualitative state.

To explore and measure the dynamic alterations in gene–gene associations, the mutual information defined as Equation (1) can be applied to quantify the statistical dependency between pairs of genes.

$$I(X,Y)=\sum_{x\in X,y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}, \tag{1}$$

where $p(x)$ and $p(y)$, denoted as a marginal probability distribution, corresponds to the probability distribution of genes X and Y respectively, whereas $p(x,y)$, labeled as a joint probability distribution, characterizes their combined probabilities. In this study, to determine gene–gene associations and construct cell-specific networks at the single-cell level, a statistical dependency index (as presented in Equation (3)) is developed using numerically estimated probability distributions based on frequency [11]. A positive statistical dependency value indicates a statistically significant correlation between gene pairs, implying the presence of an edge between them (Figure S1). The relationship between the entropy $H(X)$ of X, the conditional entropy $H(X|Y)$ of X given Y, and the mutual information $I(X,Y)$ of X and Y allows for the representation of $H(X)$ using $H(X|Y)$ and $I(X,Y)$, as shown in the following formula (refer to Supplementary Information A for the derivation process).

$$H(X)=I(X,Y)+H(X|Y). \tag{2}$$

For a certain local network/subnetwork, as seen easily from Equation (2), the entropy of the central node is dependent on the contribution from the strength of the association with its connecting or neighboring nodes.

An algorithm for capturing the critical state of cell transition utilizing SGAE

Based on the single-cell expression data with temporal information, the SGAE procedure is utilized to reveal the critical state of cell transition, and its specific implementation steps are as follows:

[Step 1] Normalization of single-cell expression data. For each specific time point, the general methodology **log**–transformation is carried out to normalize the initial gene expression matrix of size M (rows) × N (columns), where rows signify genes and columns depict cells.

[Step 2] Construction of a statistical dependency index $I^{(k)}_{x,y}$. The detailed process is described as follows: (i) For each gene pair (g_x, g_y), scatter plots are constructed in a cartesian coordinate system where the g_x–axes and g_y–axes correspond to the expression values of these two genes across N cells. Notably, each point within the scatter plot represents a cell, with its g_x-coordinate denoted as $E^{(k)}_x$ (representing the expression of g_x in cell C_k), and its g_y-coordinate indicated as $E^{(k)}_y$ (reflecting the expression of g_y in cell C_k). A total of $M(M-1)/2$ scatter plots are generated by plotting a scatter plot for each gene pair (as illustrated in Figure 1B). (ii) Within the scatter plot of the gene pair (g_x, g_y), near the specific cell C_k, we designate the brown and blue boxes as representing the neighborhood of $E^{(k)}_x$ and $E^{(k)}_y$ based on the preset parameter $n^{(k)}(E_x) = 0.1N$ (the point/cell count in the green box) and $n^{(k)}(E_y) = 0.1N$ (the point/cell count in the blue box), respectively. We have conducted an analysis on the critical signal for cell transition when the parameter $n^{(k)}(E_x)$ varies. It can be seen from Figure S2 that the preset parameter $n^{(k)}(E_x)$within a certain range has negligible impact on the evolutionary trend of the signal curve. The red box signifies the overlap between two mentioned boxes, with the count of points/cells represented by $n^{(k)}(E_x, E_y)$. (iii) The construction of the statistical dependency index$I^{(k)}_{x,y}$ is achieved through the utilization of the three aforementioned statistics, namely $n^{(k)}(E_x)$, $n^{(k)}(E_y)$, and $n^{(k)}(E_x, E_y)$.

$$I^{(k)}_{x,y}=\frac{n^{(k)}(E_x,E_y)}{N}\log\frac{\frac{n^{(k)}(E_x,E_y)}{N}}{\frac{n^{(k)}(E_x)}{N}\cdot\frac{n^{(k)}(E_y)}{N}}, \tag{3}$$

[Step 3] Construction of a cell-specific network tailored to each individual cell. By employing the statistical dependency index $I^{(k)}_{x,y}$ as depicted in Equation (3), we construct a cell-specific network $N^{(k)}$ tailored to cell C_k, where genes are represented as nodes, and the connections between these nodes are defined by gene–gene associations indicated by the dependency indicator $I^{(k)}_{x,y}$. Specifically, the gene–gene association (an edge) exists between g_x and g_y in cell C_k if $I^{(k)}_{x,y}$ is a positive value; otherwise, there is not an edge between g_x and g_y. Then, the construction of the cell-specific network $N^{(k)}$ tailored to cellC_k follows this way: (see Supplementary Information B for the specific construction process).

[Step 4] Extraction of each local network from a cell-specific network. Specifically, focusing on cell C_k, the cell-specific network $N^{(k)}$ is composed of M local networks $LN^{(k)}_x$ ($x = 1, 2, 3, \ldots, M$), each of which is identified by a specific gene (i.e. one gene for one local network). Each local network$LN^{(k)}_x$ is represented as a subnetwork centered at a gene g_x, with Q edges representing its 1st-order neighbors $\left\{g^x_1, g^x_2, \ldots, g^x_Q\right\}$.

[Step 5] Calculation of the distinct local SGAE value $H^{(k)}_x$. Based on Equation (2) above, the calculation of the SGAE value $H^{(k)}_x$ for a

Downloaded from https://academic.oup.com/bib/article/24/6/bbad366/7313455 by Zhejiang University user on 22 October 2023

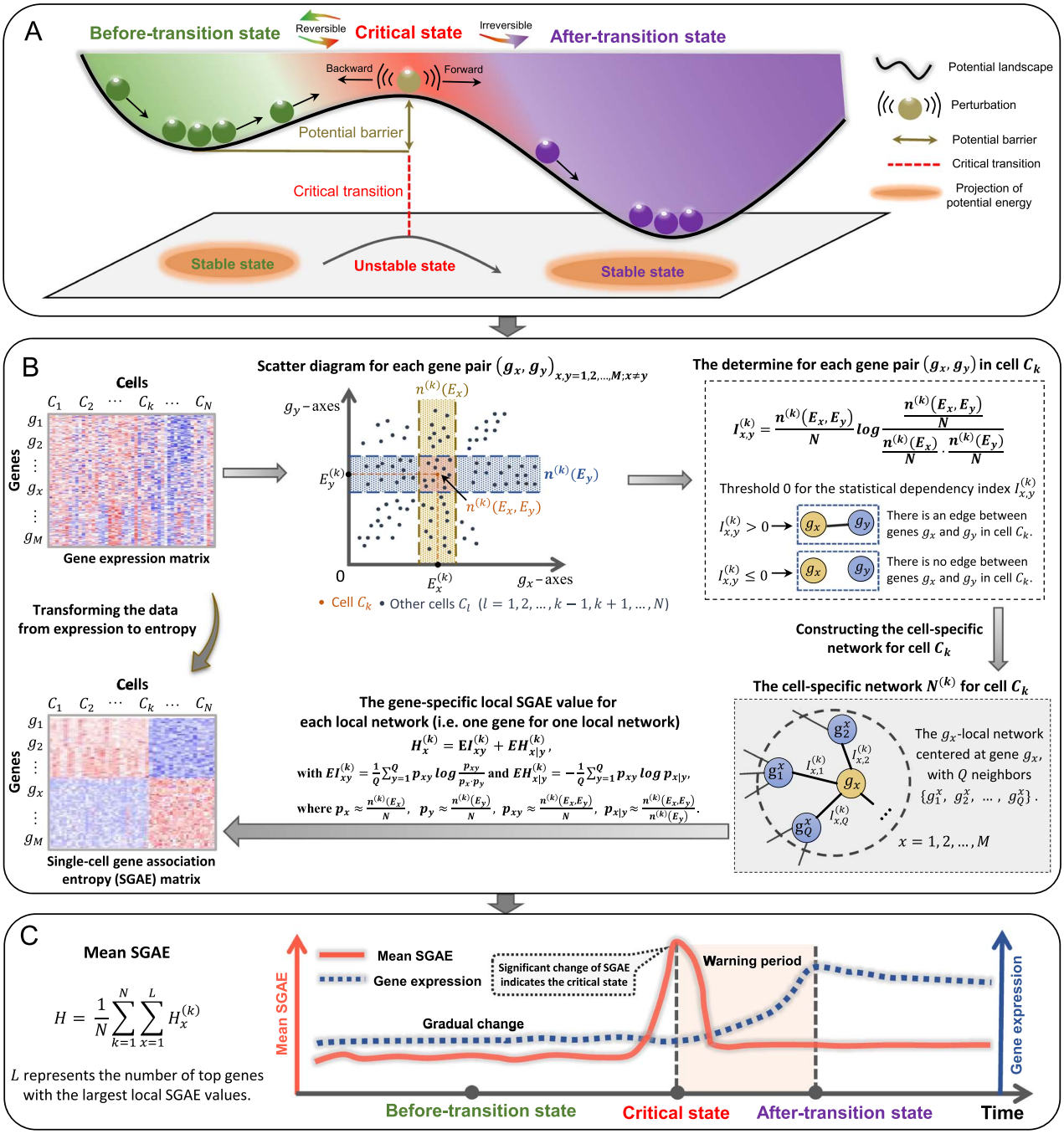

Figure 1. Schematic illustration of the proposed SGAE method for revealing critical states of cell transitions. (**A**) Complex biological systems can be broadly categorized into three states: a stable and robust before-transition state, a highly sensitive and unstable critical state at which an abrupt and qualitative shift occurs, and another stable after-transition state. (**B**) Infer gene-gene association (an edge) between gene pairs with the statistical dependency index $I_{x,y}^{(k)}$, and then construct the cell-specific network $N^{(k)}$ for cell C_k. The local SGAE value is calculated for each localized network extracted from the cell-specific network. (**C**) The critical state of cell transition can be indicated by the abrupt increase of SGAE. There is no obvious change in SGAE when the system is in a before-transition state; however, when the system is close to the critical state, SGAE increases significantly.

localized network $LN_x^{(k)}$ centered on gene g_x is performed as below.

$$H_x^{(k)} = EI_{xy}^{(k)} + EH_{x|y}^{(k)}, \tag{4}$$

with

$$EI_{xy}^{(k)} = \frac{1}{Q}\sum_{y=1}^{Q} p_{xy} \log \frac{p_{xy}}{p_x \cdot p_y} \tag{5}$$

and

$$EH_{x|y}^{(k)} = -\frac{1}{Q}\sum_{y=1}^{Q} p_{xy} \log\ p_{x|y},$$

where the probability distributions can be estimated by statistical analysis as follows:

$$p_x \approx \frac{n^{(k)}(E_x)}{N}, \tag{6}$$

Downloaded from https://academic.oup.com/bib/article/24/6/bbad366/7313455 by Zhejiang University user on 22 October 2023

$$p_y \approx \frac{n^{(k)}(E_y)}{N},$$

$$p_{xy} \approx \frac{n^{(k)}(E_x, E_y)}{N},$$

$$p_{x|y} \approx \frac{n^{(k)}(E_x, E_y)}{n^{(k)}(E_y)}.$$

Clearly, Equation (4) has the capability to convert the inherent sparsity of the gene expression matrix of single-cell datasets into a non-sparsity entropy matrix in light of gene–gene associations (Figure 1B). Consequently, the local SGAE value $H_x^{(k)}$ is determined by both the central gene's expression within a local network and that of its neighboring genes.

[Step 6] Calculation of the cell-specific SGAE value $H^{(k)}$ for each cell. Given the cell C_k, the calculation of its cell-specific SGAE value $H^{(k)}$ relies on a gene cluster marked by the highest local SGAE values, and the specific computation formula is as follows:

$$H^{(k)} = \sum_{x=1}^{L} H_x^{(k)}, \tag{7}$$

where the constant L indicates the count of genes ranking in the top 5% with highest local SGAE values. The pattern of signal curve evolution remains consistent despite variations in the adjustable parameter L, typically set within a range from the top 3% to 10% (Figure S3). Moreover, the average of the cell-specific SGAE values H_t (as indicated in Equation (8)) for each time point t is utilized to capture the critical signal associated with cell transition.

$$H_t = \frac{1}{N}\sum_{k=1}^{N}\sum_{x=1}^{L} H_x^{(k)}, \tag{8}$$

Close to the critical point, there is a noticeable collective fluctuation behavior among the signaling molecules or DNBs. This leads to gene associations (dependent correlations) among DNBs in a critical state, exhibiting substantial distinctions from those observed in a before-transition state. Consequently, upon nearing the critical state, the SGAE index H_t shows a sudden increase.

[Step 7] Identification of the critical state based on the one-sample t test. For the purpose of assessing the capability of the SGAE H_t in capturing the critical signal, a one-sample t test [13] is implemented to ascertain whether the critical state significantly differs from the before-transition state. As represented in Equation (9), we employ the statistic TS to evaluate whether a significant distinction exists between the constant z and the average of the vector $Z = (z_1, z_2, \cdots, z_n)$.

$$TS = \sqrt{n}\frac{\mathbf{mean}(Z) - z}{\mathbf{SD}(Z)}, \tag{9}$$

where $\mathbf{mean}(Z)$ stands for the average of vector Z, while $\mathbf{SD}(Z)$ denotes its standard deviation. Subsequently, the calculation of the P value for statistic TS is performed to evaluate the statistical significance between $\mathbf{mean}(Z)$ and z. Statistical significance is attained when the P value < 0.05; otherwise, it implies a lack of significant difference. Thus, a time point t is deemed a critical state when the SGAE index H_t achieves these two specified requirements: (i) $H_t > H_{t-1}$ and (ii) H_t exhibits a statistical difference (P value< 0.05) compared to the prior values (refer to Supplementary Information C for a detailed description).

Data sources and functional analysis

The SGAE method approach was implemented in five distinct single-cell datasets related to embryonic developmental processes, encompassing the transition from mouse hepatoblast cell (MHC) to hepatocyte and cholangiocyte cell (HCC) (MHC-to-HCC; ID: GSE90047), epithelial basal cell (EBC) to mouse hair follicle (MHF) (EBC-to-MHF; ID: GSE147372), inner cell mass (ICM) to visceral endoderm cell (VEC) (ICM-to-VEC; ID: GSE100597), human prefrontal cortex (HPC) to neuron (HPC-to- neuron; ID: GSE104276) and neural progenitor cell (NPC) to neuron (NPC-to-neuron; ID: GSE102066). Comprehensive information regarding these datasets is provided in Supplementary Information D. Functional enrichment analysis was conducted based on the Metascape online tool [14] and ClusterProfiler package [15]. The pathway-related information can be found in the Kyoto Encyclopedia of Genes and Genomes (KEGG) (https://www.kegg.jp/).

RESULTS

Revealing the critical state of cell transition

To demonstrate how the proposed SGAE works, we implemented it on five distinct single-cell datasets associated with embryonic developmental processes, i.e. MHC-to-HCC data [16], EBC-to-MHF data [17], ICM-to-VEC data [18], HPC-to-neuron data [19] and NPC-to-neuron data [20]. The cell-specific SGAE value for each single cell was calculated using Equation (7) defined in the Methods section. The mean cell-specific SGAE value was utilized to detect any possible critical states of cell transitions at a specific time point. The SGAE approach effectively pinpointed cell fate transitions in embryonic development across all datasets, thereby validating the accuracy and effectiveness of our method. The algorithm's source code is freely available at https://github.com/zhongjiayuan-fs/SGAE-project.

For the MHC-to-HCC data, the red curve in Figure 2A clearly displays a notable upsurge in the mean SGAE value at embryonic day 12.5 (E12.5) ($P = 0.0229$), which serves as an indicator of an impending cell fate transition, aligning with the findings of original experiment that hepatoblasts differentiate into hepatocytes and cholangiocytes after E12.5 [16]. Furthermore, a box plot depicting the cell-specific SGAE values is provided for each time point, showcasing the resilience of our proposed approach. As shown by the red box plot of the SGAE value in Figure 2A, a clear critical transition signal can be seen in the median values, indicating that the SGAE value is highly robust to sample noise. For each cell, the average expression of the highest 5% of genes ranked by expression value was employed to analyze the dynamic changes. Unlike the SGAE value, the gene expression fails to provide early indication of cell fate transition (the green curve in Figure 2A). In the case of the EBC-to-MHF data, the mean SGAE value (the red curve in Figure 2B) abruptly increases from embryonic day 13 (E13) to embryonic day 13.5 (E13.5) ($P = 3.1741E-12$), after which epithelial basal cells were guided to differentiate into hair follicle stem cells [17]. Additionally, the robustness of the SGAE method in pinpointing the tipping point of cell transitions is highlighted by the median values shown in the red box plot of the SGAE value in Figure 2B. However, in the context of gene expression, the green curve in Figure 2B demonstrates no significant variation among the six time points. When applied to the ICM-to-VEC data, the statistical significance ($P = 0.0436$) is indicated by the red curve in Figure 2C, revealing a commitment to a visceral endoderm fate after embryonic day 4.5 (E4.5) [18]. Moreover, the median values depicted by the red box plot of the SGAE value in Figure 2C

Downloaded from https://academic.oup.com/bib/article/24/6/bbad366/7313455 by Zhejiang University user on 22 October 2023

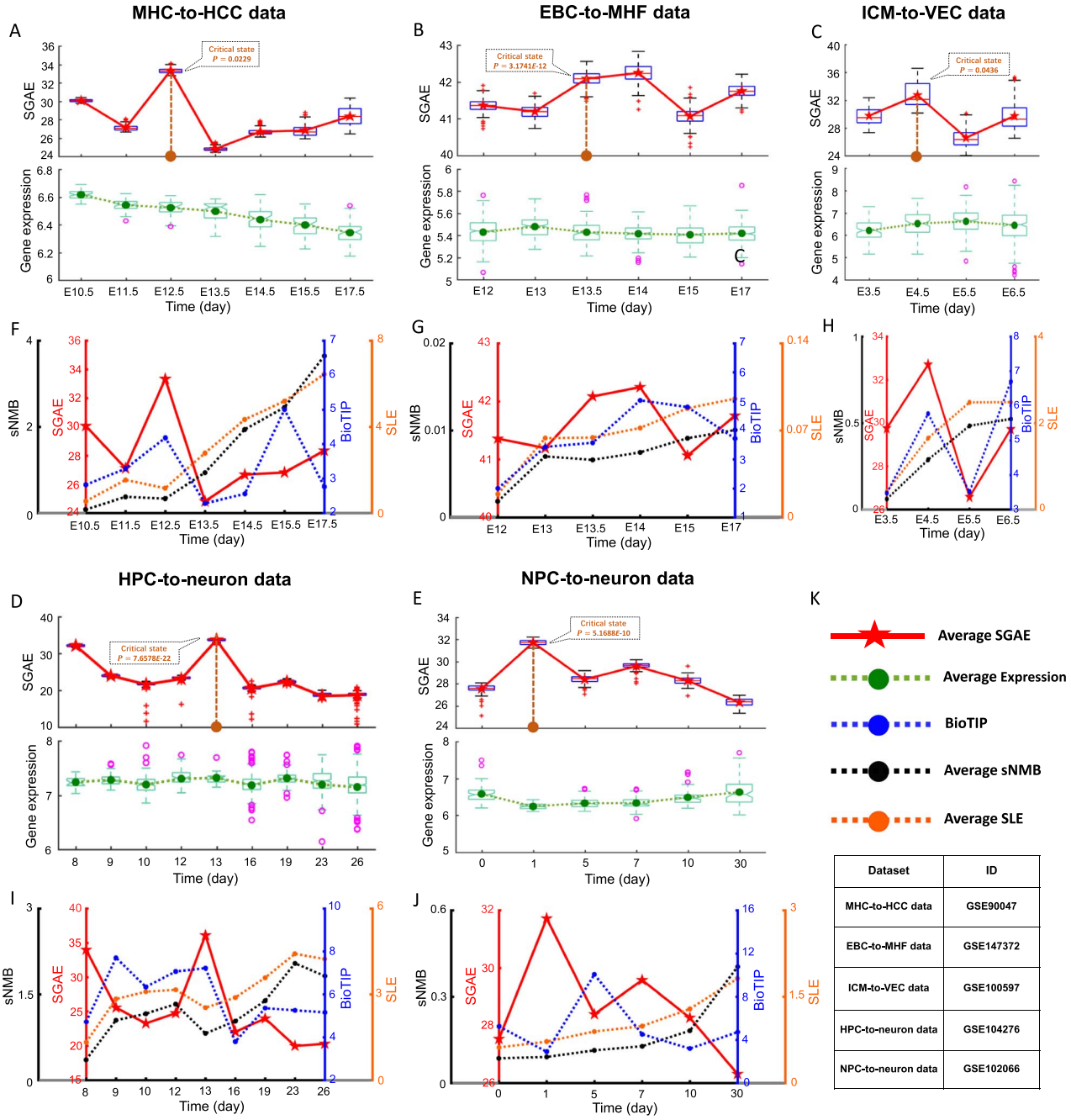

Figure 2. Revealing the critical state of cell transition. The performance of dynamic changes between the SGAE and gene expression for five single-cell datasets of embryonic differentiation: (**A**) MHC-to-HCC data, (**B**) EBC-to-MHF data, (**C**) ICM-to-VEC data, (**D**) HPC-to-neuron data, and (**E**) NPC-to-neuron data. The dynamic change performance between SGAE and other existing critical transition detection methods across five datasets: (**F**) MHC-to-HCC data, (**G**) EBC-to-MHF data, (**H**) ICM-to-VEC data, (**I**) HPC-to-neuron data, and (**J**) NPC-to-neuron data.

obviously show a critical signal at E4.5 as well. However, when considering gene expression, there is no significant difference (the green curve in Figure 2C). For the HPC-to-neuron data, as indicated by the red curve in Figure 2D, mean SGAE value experiences a dramatic increase from 12 weeks to 13 weeks ($P = 7.6578E-22$), after which neurons from the prefrontal cortex have been shown to increase expression of genes for cell fate commitment [19]. Additionally, it is seen from the red box plot of SGAE values in Figure 2D that the median values stably exhibit a clear signal at the critical state (13 weeks). Unlike the SGAE value, as demonstrated by the green curve in Figure 2D, there is no indication of the cell fate transition based on gene expression. Applying the SGAE to the NPC-to-neuron data, the red curve in Figure 2E reveals significant differences in the mean SGAE value at day 1 ($P = 5.1688E-10$), which supports the original experimental observation that the heterogeneity of transcriptional state increased after day 1 and reached the highest level of neuronal heterogeneity at day 30 [20]. In addition, as depicted in the red box plot of the SGAE value in Figure 2E, the median values also pinpoint day 1 as a critical point. However, as observed from the green curve in Figure 2E, there is no significant difference from the perspective of gene expression. Figure 2F-J demonstrates that our proposed method exhibits satisfied performance in capturing critical signals of

Downloaded from https://academic.oup.com/bib/article/24/6/bbad366/7313455 by Zhejiang University user on 22 October 2023

cell transitions compared to other existing critical transition detection methods, such as biological tipping-point (BioTIP) [21], single-sample landscape entropy [22], and single-sample network module biomarkers (sNMB) [23]. The above results indicate that our SGAE approach is capable of delivering more effective the critical signal for cell transition during embryonic development.

Performing analyses for temporal clustering of cells and dynamic evolution of the network

Beyond its role in pinpointing the critical transitions during embryonic development, SGAE also enables the transformation of the gene expression matrix into the SGAE matrix, which provides a SGAE-based way to perform cell clustering analysis for biological processes. At the identified critical state, a cluster of genes, consisting of signaling genes (molecules ranking in the top 5% for their highest local SGAE values) and low-entropy genes (molecules ranking in the top 5% for their lowest local SGAE values), was selected to carry out cell clustering analysis using local SGAE values. For the MHC-to-HCC, NPC-to-neuron, and ICM-to-VEC data, as presented in Figure 3A–C, SGAE-based clustering analysis effectively distinguishes the cellular states at different time points using t-distributed stochastic neighbor embedding [24]. Uniform manifold approximation and projection is also applied to carry out for the SGAE-based clustering analysis. The results of the cluster analysis and an evaluation of their clustering performance are given in Figure S4 of Supplementary Information E. In addition, we utilize the landscape of the local SGAE for signaling and non-signaling genes to depict the overall dynamic shifts in network entropy. For the MHC-to-HCC data, as illustrated in Figure 3D, an abrupt surge in the local SGAE of signaling genes occurs at the critical state (E12.5). Similarly, in the case of the NPC-to-neuron data, Figure 3E indicates a notable peak in local SGAE for signaling genes at the critical state (E4.5). For the ICM-to-VEC data, as depicted in Figure 3F, a substantial increase is evident in the local SGAE values of signaling genes at 13 weeks, indicating an upcoming critical point of cell transition. Moreover, we map the signaling genes onto the protein–protein interaction (PPI) network and extract the largest connected subgraph to study how it dynamically evolves at the network level. For the three mentioned embryonic development datasets, as illustrated in Figure 3G–I, a noticeable shift in the network structure is observable at the critical point. The complete temporal evolution of the signaling-gene network is presented in Figure S5 of Supplementary Information F. Therefore, by exploiting the dynamic changes in gene–gene associations, SGAE offers valuable insights into critical transition during early embryonic development in terms of network dynamics.

Revealing the molecular signaling mechanism underlying developmental process

Differential expression analysis is a valuable approach for identifying new drug targets and biomarkers. However, it often overlooks non-differentially expressed genes (non-DEGs) that play crucial roles in essential biological processes. These genes are recognized for their significant roles in immune cell activity [25] and are enriched in development-related functional pathways [26]. In our study, genes are considered 'dark genes' when they satisfy the following two conditions: (i) they show non-differential expression based on expression levels and (ii) a notable difference can be observed in the SGAE value levels between the critical state and the non-critical state. By focusing on signaling genes (top 5% molecules with the highest local SGAE values), we compare the changes in SGAE values and gene expression to discover 'dark genes'. Figure 4 demonstrates some of the 'dark genes' from the MHC-to-HCC, NPC-to-neuron, and ICM-to-VEC data, which exhibit insignificant variations in gene expression but show significant alterations in SGAE values. Other dark genes for these three datasets, as well as for the EBC-to-MHF and HPC-to-neuron data, can be found in Supplementary Information G. It has been revealed that some dark genes are implicated in developmental processes, suggesting their potential importance in embryonic development. For the MHC-to-HCC, NPC-to-neuron, and ICM-to-VEC data, the 'dark genes' related to embryonic development are outlined in Tables 1, 2 and 3, respectively.

An analysis for KEGG pathway enrichment was conducted to enhance our understanding of the biological roles of the dark genes. For the NPC-to-neuron data and HPC-to-neuron data, as shown in Figure 5A-B, we found that the enriched pathways associated with the dark genes were closely related to embryonic developmental processes, such as the FoxO signaling pathway [27], Wnt signaling pathway [28], cell cycle [29], gap junction [30] and oocyte meiosis [31]. There were seven common dark genes shared between the above two human embryonic development datasets (Figure S6). To investigate the potential regulatory mechanism involved in embryonic development on a network level, we performed functional analysis on the protein–protein interaction (PPI) subnetwork composed of these common dark genes (CDGs) and their 1st-order differentially expressed gene (DEG) neighbors within the PPI network, which can be viewed as a CDG-related network. The 1st-order DEG neighbors are characterized by meeting two requirements: (i) they serve as the 1st-order neighbors of common signaling genes within the PPI network and (ii) they indicate statistically significant variations ($P < 0.05$) in gene expression levels before and after the identified critical state. From Figure 5C, it is evident that the CDG-related network based on the NPC-to-neuron data comprises 7 CDGs and their 180 1st-order DEG neighbors. Notably, a significant change in gene expression, transitioning from low to high or vice versa, occurs after a critical transition. Furthermore, pathway enrichment analysis highlights that the 1st-order DEG neighbors exhibit enrichment in signaling pathways closely associated with embryonic development (Figure 5D-E). For the NPC-to-neuron data, the functional analysis of signaling genes and 1st-order DEG neighbors (Figure 5F) reveals insights into the underlying signaling mechanism. The WNT pathway is known to have a significant role in brain development [32]. Our method identifies a further potential molecular driving axis for neuron development. Our findings indicate that the long-range axis of action based on *WNT5A* and *FZD3* may exhibit spatiotemporal dynamics during the transition from NPCs to neurons. Specifically, *WNT5A* is a 1st-order DEG, and its expression is markedly downregulated in the vicinity of the crucial point, suggesting that the signal is initiated during this time. After *WNT5A* activates a variety of regulatory and downstream receptors, the transcription factor *TCF7L1* (a signaling gene) is released by inhibiting *CTNNB1* expression. It is generally accepted that cell cycle inhibition is related to elevated *TCF7L1* expression [33–35]. In this particular dataset, cell cycle inhibition and neural differentiation begin simultaneously [20]. These outcomes further substantiate the accuracy of the pathway axis of action identified by our method.

DISCUSSION

The identification of cell fate decisions or critical states of cell transitions holds significant biological and clinical importance. It can contribute to the design of therapeutic strategies intended

Downloaded from https://academic.oup.com/bib/article/24/6/bbad366/7313455 by Zhejiang University user on 22 October 2023

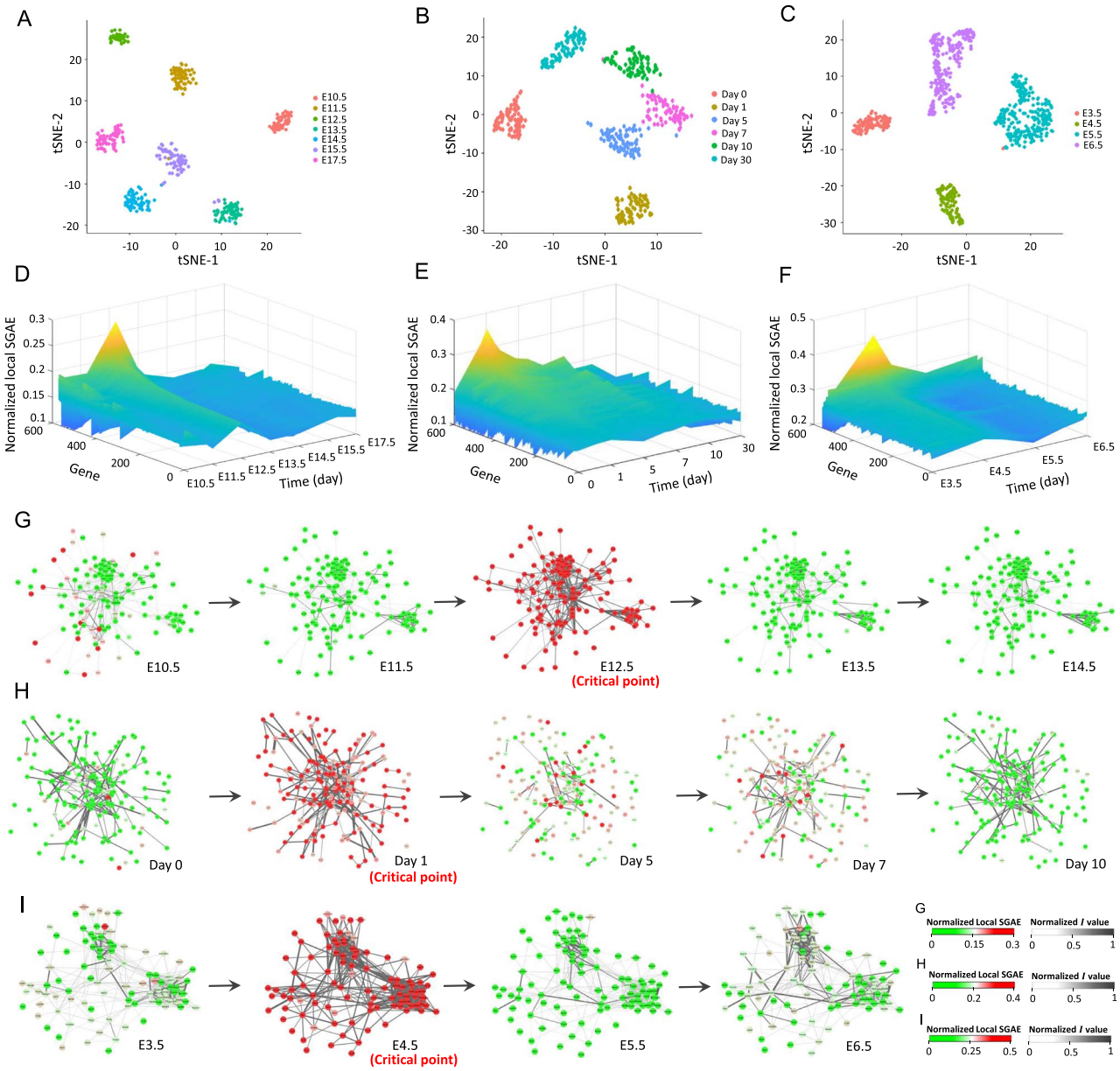

Figure 3. Temporal clustering of cells based on SGAE and dynamical evolution of the signaling-gene network. Based on the local SGAE values of the top high- and low-entropy genes, the temporal clustering analysis is carried out for (**A**) MHC-to-HCC data, (**B**) NPC-to-neuron data, and (**C**) ICM-to-VEC data, respectively. The landscape of the local SGAE is employed to demonstrate the dynamic changes of the network entropy from a global perspective for (**D**) MHC-to-HCC data, (**E**) NPC-to-neuron data, and (**F**) ICM-to-VEC data, respectively. The dynamical evolution of the signaling-gene network is analyzed for (**G**) MHC-to-HCC data, (**H**) NPC-to-neuron data, and (**I**) ICM-to-VEC data, respectively.

for personalized tissue regeneration and the construction of differentiation-related disease models [36]. Most existing computational methods for analyzing scRNA-seq data rely on statistical measures (e.g. mean and variance) guided by gene expression. However, characterizing the dynamic evolution of complex biological systems using scRNA-seq data often presents challenges due to the inherent sparsity and noise in single-cell gene expression data. This study introduces a model-free approach named SGAE to quantitatively leverage gene–gene associations information and their dynamic changes at the single-cell level, consequently revealing a critical state of cell differentiation. For five single-cell datasets of embryonic development, the proposed approach successfully identified their corresponding critical states or cell fate decisions. Moreover, SGAE transforms a sparse matrix of single-cell gene expressions into a non-sparse matrix of gene–gene association entropy values. This enables temporal clustering analysis based on SGAE values, accurately distinguishing cellular heterogeneity over time. Besides, our approach can detect some SGAE-sensitive 'dark genes' that hold significance in the biological processes of embryonic development.

Overall, the proposed SGAE method offers several key advantages as follows. First, in terms of analyzing dynamic changes, SGAE represents a valuable tool for uncovering critical signals during embryonic development at the single-cell level. It is capable of capturing critical signals of cell transitions and performs better than other critical-transition detection methods. Additionally, in conjunction with dynamic predictive techniques [37–39], it holds the potential to forecast future critical states using omics data. Second, SGAE performs well in cell clustering. It accurately

Downloaded from https://academic.oup.com/bib/article/24/6/bbad366/7313455 by Zhejiang University user on 22 October 2023

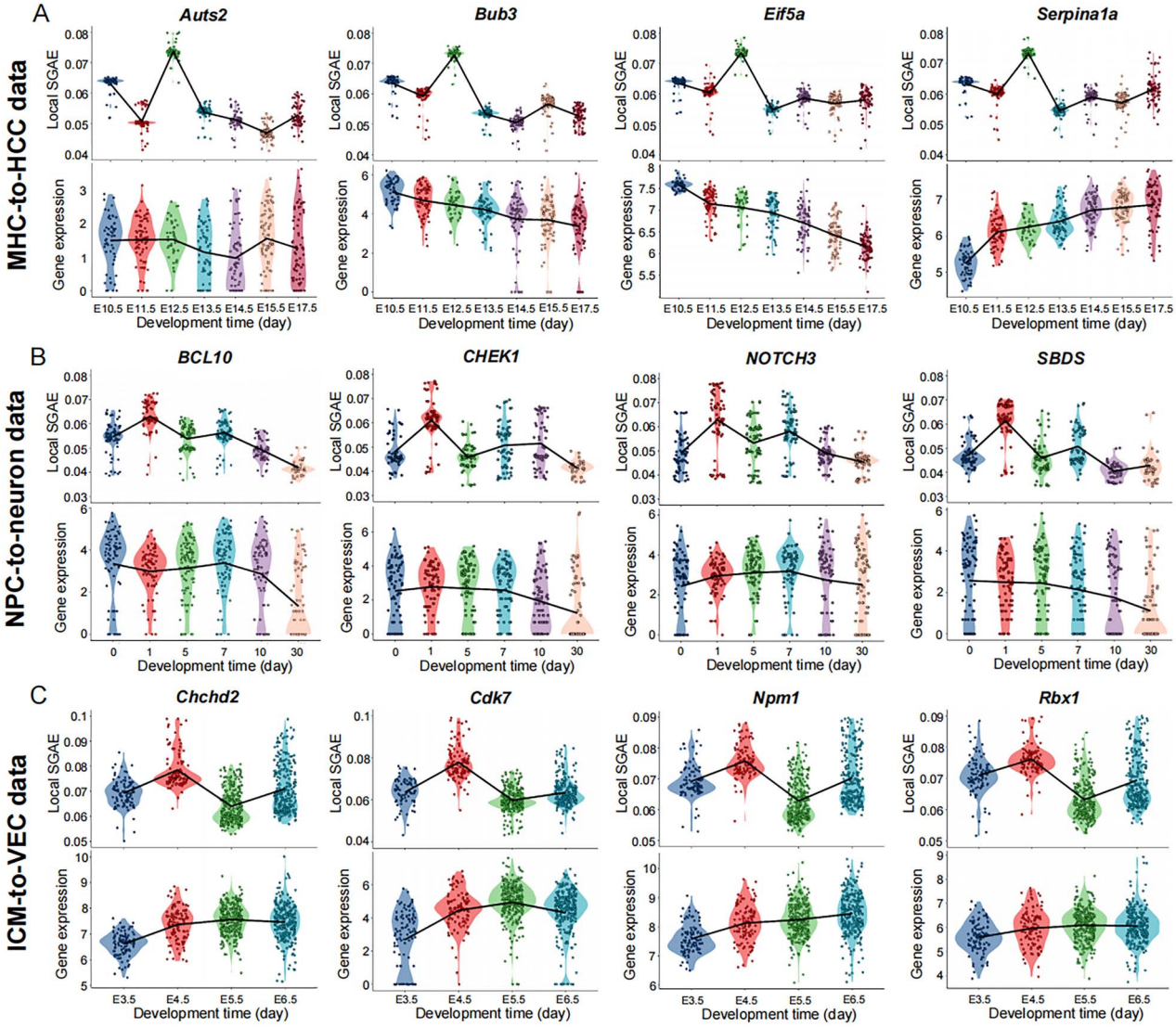

Figure 4. The SGAE-sensitive "dark genes". The local SGAE (top) and gene expression (bottom) of dark genes for (**A**) MHC-to-HCC data, (**B**) NPC-to-neuron data, and (**C**) ICM-to-VEC data are provided.

Table 1. The information of important 'dark genes' in MHC-to-HCC data

Gene	Location	Family	Relation with embryonic development	PMID
Auts2	Nucleus	Translation regulator	*Auts2* isoforms play an important role in regulating transcription and neuronal differentiation	30 953 002
Ambp	Cytoplasm	Other	Expression of the *Ambp* can affect mouse embryogenesis	12 204 273
Bub3	Cytoplasm	Enzyme	*Bub3* disruption reveals essential mitotic spindle checkpoint function during early embryogenesis	10 995 385
Dkc1	Nucleus	Enzyme	Targeted disruption of *Dkc1* can cause embryonic lethality in mice	12 400 016
Eif5a	Cytoplasm	Other	*Eif5a* is related to the embryogenesis and cell differentiation	20 458 750
Gdi2	Cytoplasm	Regulatory	Targeted disruption of *Gdi2* causes early embryonic lethality	35 689 892
Prmt7	Nucleus	Enzyme	*PRMT7* is involved in regulation of germ cell proliferation during embryonic stage	33 008 598
Serpina1a	Cytoplasm	Other	Deletion of *serpina1a* can result in embryonic lethality	21 574 874

distinguishes cellular heterogeneity over time at the single-cell level through clustering analysis based on SGAE values. Third, SGAE helps uncover 'dark genes' that are sensitive to SGAE but often overlooked by traditional differential expression analysis. These genes are likely to play significant roles in key biological processes [40]. Final, the proposed SGAE is a data-driven method, and thus can view as a model-free process, eliminating the need for selecting features or training the model/parameter.

Downloaded from https://academic.oup.com/bib/article/24/6/bbad366/7313455 by Zhejiang University user on 22 October 2023

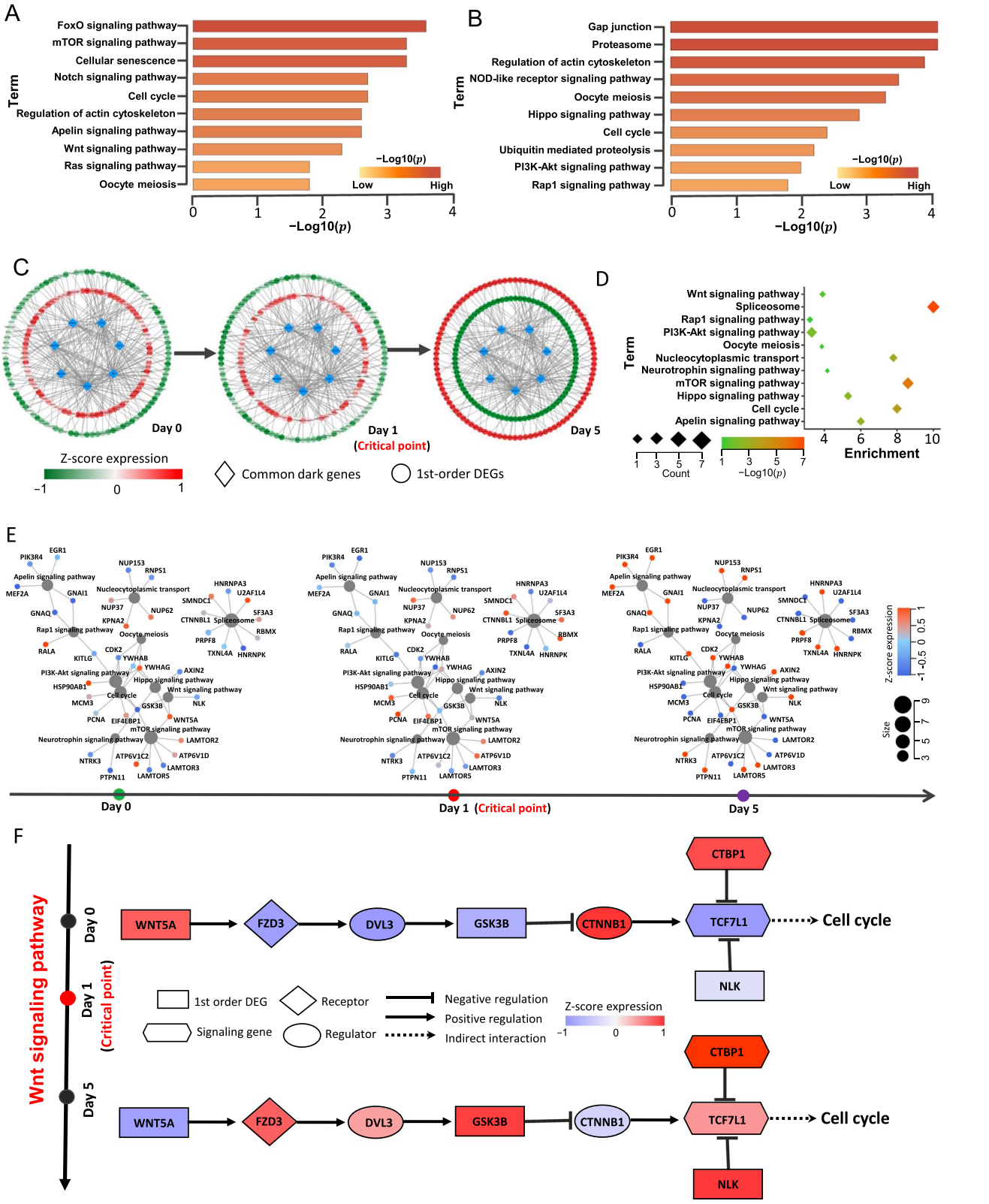

Figure 5. The molecular signaling mechanism underlying the developmental process. (**A**) KEGG pathway enrichment analysis for NPC-to-neuron data. (**B**) KEGG pathway enrichment analysis for HPC-to-neuron data. (**C**) Dynamic changes of the CDG-related network based on NPC-to-neuron data, which is composed of 7 CSGs and their 180 1st-order DEG neighbors. (**D**) KEGG pathway enrichment analysis for the above-mentioned 1st-order DEG neighbors. (**E**) The expression patterns of these 1st-order DEG neighbors involved in different development-related pathways change significantly before and after the critical transition. (**F**) The Wnt signaling pathway presents different regulatory patterns before and after the critical point, which plays an essential role in the developmental process.

Downloaded from https://academic.oup.com/bib/article/24/6/bbad366/7313455 by Zhejiang University user on 22 October 2023

Table 2. The information of important 'dark genes' in NPC-to-neuron data

Gene	Location	Family	Relation with embryonic development	PMID
BANF1	Cytoplasm	Other	*BANF1* is required to maintain the self-renewal of human embryonic stem cells	21 750 191
BCL10	Cytoplasm	Regulatory	Deficiency of *BCL10* results in partial embryonic lethality caused by a neural tube closure defect	11 163 238
CFDP1	Nucleus	Other	*CFDP1* can control neural differentiation by regulating the cell cycle	33 987 914
CHEK1	Cytoplasm	Regulatory	*CHEK1* promotes DNA repair and is needed for embryo development	34 349 269
CSNK2B	Cytoplasm	Enzyme	Loss of *CSNK2B* compromises proliferation and differentiation of embryonic neural progenitor cells	35 571 680
DGCR8	Nucleus	Regulatory	*DGCR8* is essential for microRNA biogenesis and silencing of embryonic stem cell self-renewal	17 259 983
NOTCH3	Cytoplasm	Regulatory	*NOTCH3* is necessary for neuronal differentiation and maturation in the adult spinal cord	25 164 209
SBDS	Nucleus	Other	*SBDS* is an essential gene for early mammalian development and plays a critical role in cell proliferation.	16 914 746

Table 3. The information of important 'dark genes' in ICM-to-VEC data

Gene	Location	Family	Relation with embryonic development	PMID
Banf1	Nucleus	Other	The knockdown of *Banf1* promotes the differentiation of mouse embryonic stem cells	21 750 191
Cdk7	Cytoplasm	Enzyme	*Cdk7* plays an important role in embryonic stem pluripotency	20 231 280
Chchd2	Nucleus	Regulatory	*Chchd2* can regulate pluripotent stem cell differentiation	27 810 911
Eif3h	Cytoplasm	Other	*Eif3h* is required for normal embryonic development in the mouse	23 173 090
Npm1	Nucleus	Other	Suppression of *Npm1* expression in mouse embryonic stem cells resulted in reduced cell proliferation	27 939 217
Rbx1	Cytoplasm	Enzyme	Physiological function of *Rbx1* is to ensure cell proliferation during the early embryonic development	19 325 126
Hmgn1	Nucleus	Other	*Hmgn1* can affect the transcriptional profile of mouse embryonic stem cells and neural progenitors	23 775 126
Sec61b	Cytoplasm	Transporter	*Sec61b* is an essential factor for the embryonic development	19 226 464

However, the SGAE method has certain limitations. For example, the identified gene–gene associations do not necessarily imply causal relationships between the two molecules, which will be a potential focus of our future research topics.

Key Points

- In contrast to conventional differential-gene or static approaches that emphasize the classification of cell types, we presented a novel computational approach, single-cell gene association entropy (SGAE), to reveal critical states of cell transitions among cell populations, which supports the monitoring of biological system dynamics at the single-cell level.
- SGAE has a better performance in capturing critical signals of cell transitions compared with other existing critical transition detection methods.
- Temporal changes in cellular heterogeneity at the resolution of single cells can be accurately distinguished by the clustering analysis based on SGAE values, suggesting SGAE has good performance in cell clustering.
- Our method can uncover SGAE-sensitive 'dark genes', which are often neglected by traditional differential expression analysis but are likely to be implicated in essential biological processes of embryonic development.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

FUNDING

National Natural Science Foundation of China (Nos. 12322119, 62172164 and 12271180), Guangdong Provincial Key Laboratory of Human Digital Twin (2022B1212010004), Educational Commission of Guangdong Province of China (2023KQNCX073) and the Natural Science Foundation of Guangdong Province of China (2022A1515110759).

Downloaded from https://academic.oup.com/bib/article/24/6/bbad366/7313455 by Zhejiang University user on 22 October 2023

AUTHOR CONTRIBUTIONS

The research was conceptualized by R.L. and P.C., and the practical data analysis was carried out by J.Y.Z. and C.Y.H. All authors contributed to the writing of the paper and participated in reviewing and approving the final manuscript.

DATA AVAILABILITY

MHC-to-HCC data (ID: GSE90047), EBC-to-MHF data (ID: GSE 147372), ICM-to-VEC data (ID: GSE100597), HPC-to-neuron data (ID: GSE104276), and NPC-to-neuron data (ID: GSE102066) were downloaded from the NCBI Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo). The algorithm's source code and related data can be accessed at https://github.com/zhongjiayuan-fs/SGAE-project.

REFERENCES

1. Scheffer M, Bascompte J, Brock WA, *et al.* Early-warning signals for critical transitions. *Nature* 2009;**461**(7260):53–9.
2. Bargaje R, Trachana K, Shelton MN, *et al.* Cell population structure prior to bifurcation predicts efficiency of directed differentiation in human induced pluripotent cells. *Proc Natl Acad Sci U S A* 2017;**114**(9):2271–6.
3. Chen L, Liu R, Liu ZP, *et al.* Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. *Sci Rep* 2012;**2**:342.
4. Peng H, Zhong J, Chen P, Liu R. Identifying the critical states of complex diseases by the dynamic change of multivariate distribution. *Brief Bioinform* 2022;**23**.
5. Zhong J, Ding D, Liu J, *et al.* SPNE: sample-perturbed network entropy for revealing critical states of complex biological systems. *Brief Bioinform* 2023;**24**.
6. Mojtahedi M, Skupin A, Zhou J, *et al.* Cell fate decision as high-dimensional critical state transition. *PLoS Biol* 2016;**14**(12):e2000640.
7. Wang Z, Zhong Y, Ye Z, *et al.* MarkovHC: Markov hierarchical clustering for the topological structure of high-dimensional single-cell omics data with transition pathway and critical point detection. *Nucleic Acids Res* 2022;**50**(1):46–56.
8. Duan X, Tu Q, Zhang J, *et al.* Application of induced pluripotent stem (iPS) cells in periodontal tissue regeneration. *J Cell Physiol* 2011;**226**(1):150–7.
9. Stepniewski J, Kachamakova-Trojanowska N, Ogrocki D, *et al.* Induced pluripotent stem cells as a model for diabetes investigation. *Sci Rep* 2015;**5**:8597.
10. 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.
11. Li L, Dai H, Fang Z, Chen L. C-CSN: single-cell RNA sequencing data analysis by conditional cell-specific network. *Genomics Proteomics Bioinformatics* 2021;**19**(2):319–29.
12. Gilbert DM. Cell fate transitions and the replication timing decision point. *J Cell Biol* 2010;**191**(5):899–903.
13. Rochon J, Kieser M. A closer look at the effect of preliminary goodness-of-fit testing for normality for the one-sample t-test. *Br J Math Stat Psychol* 2011;**64**(3):410–26.
14. Zhou Y, Zhou B, Pache L, *et al.* Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun* 2019;**10**(1):1523.
15. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 2012;**16**(5):284–7.
16. Yang L, Wang WH, Qiu WL, *et al.* A single-cell transcriptomic analysis reveals precise pathways and regulatory mechanisms underlying hepatoblast differentiation. *Hepatology* 2017;**66**(5):1387–401.
17. Morita R, Sanzen N, Sasaki H, *et al.* Tracing the origin of hair follicle stem cells. *Nature* 2021;**594**(7864):547–52.
18. Mohammed H, Hernando-Herraez I, Savino A, *et al.* Single-cell landscape of transcriptional heterogeneity and cell fate decisions during mouse early gastrulation. *Cell Rep* 2017;**20**(5):1215–28.
19. Zhong S, Zhang S, Fan X, *et al.* A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature* 2018;**555**(7697):524–8.
20. Wang J, Jenjaroenpun P, Bhinge A, *et al.* Single-cell gene expression analysis reveals regulators of distinct cell subpopulations among developing human neurons. *Genome Res* 2017;**27**(11):1783–94.
21. Yang XH, Goldstein A, Sun Y, *et al.* Detecting critical transition signals from single-cell transcriptomes to infer lineage-determining transcription factors. *Nucleic Acids Res* 2022;**50**(16):e91.
22. Liu R, Chen P, Chen L. Single-sample landscape entropy reveals the imminent phase transition during disease progression. *Bioinformatics* 2020;**36**(5):1522–32.
23. Zhong J, Liu H, Chen P. The single-sample network module biomarkers (sNMB) method reveals the pre-deterioration stage of disease progression. *J Mol Cell Biol* 2022;**14**.
24. Zhou B, Jin W. Visualization of single cell RNA-Seq data using t-SNE in R. *Methods Mol Biol* 2020;**2117**:159–67.
25. Jin Q, Zuo C, Cui H, *et al.* Single-cell entropy network detects the activity of immune cells based on ribosomal protein genes. *Comput Struct Biotechnol J* 2022;**20**:3556–66.
26. Zhong J, Han C, Zhang X, *et al.* scGET: predicting cell fate transition during early embryonic development by single-cell graph entropy. *Genomics Proteomics Bioinformatics* 2021;**19**(3):461–74.
27. Zhang X, Yalcin S, Lee DF, *et al.* FOXO1 is an essential regulator of pluripotency in human embryonic stem cells. *Nat Cell Biol* 2011;**13**(9):1092–9.
28. van Amerongen R, Nusse R. Towards an integrated view of Wnt signaling in development. *Development* 2009;**136**(19):3205–14.
29. White J, Dalton S. Cell cycle control of embryonic stem cells. *Stem Cell Rev* 2005;**1**(2):131–8.
30. Houghton FD. Role of gap junctions during early embryo development. *Reproduction* 2005;**129**(2):129–35.
31. Qi ST, Ma JY, Wang ZB, *et al.* N6-Methyladenosine sequencing highlights the involvement of mRNA methylation in oocyte meiotic maturation and embryo development by regulating translation in xenopus laevis. *J Biol Chem* 2016;**291**(44):23020–6.
32. Kim JY, Lee JS, Hwang HS, *et al.* Wnt signal activation induces midbrain specification through direct binding of the beta-catenin/TCF4 complex to the EN1 promoter in human pluripotent stem cells. *Exp Mol Med* 2018;**50**(4):1–17.
33. Beck B, Blanpain C. Unravelling cancer stem cell potential. *Nat Rev Cancer* 2013;**13**(10):727–38.
34. Ben-Porath I, Thomson MW, Carey VJ, *et al.* An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. *Nat Genet* 2008;**40**(5):499–507.
35. Eshelman MA, Shah M, Raup-Konsavage WM, *et al.* TCF7L1 recruits CtBP and HDAC1 to repress DICKKOPF4 gene expression

Downloaded from https://academic.oup.com/bib/article/24/6/bbad366/7313455 by Zhejiang University user on 22 October 2023

in human colorectal cancer cells. *Biochem Biophys Res Commun* 2017;**487**(3):716–22.
36. Jiang B, Jen M, Perrin L, *et al.* SIRT1 overexpression maintains cell phenotype and function of endothelial cells derived from induced pluripotent stem cells. *Stem Cells Dev* 2015;**24**(23):2740–5.
37. Chen P, Liu R, Aihara K, Chen L. Autoreservoir computing for multi-step ahead prediction based on the spatiotemporal information transformation. *Nat Commun* 2020;**11**:4568.
38. Peng H, Chen P, Liu R, Chen L. Spatiotemporal information conversion machine for time-series forecasting. *Fundamental Research* 2022. https://doi.org/10.1016/j.fmre.2022.12.009.
39. Chen P, Zhong J, Yang K, *et al.* TPD: a web tool for tipping-point detection based on dynamic network biomarker. *Brief Bioinform* 2022;**23**.
40. Zhong J, Han C, Wang Y, *et al.* Identifying the critical state of complex biological systems by the directed-network rank score method. *Bioinformatics* 2022;**38**:5398–405.

Downloaded from https://academic.oup.com/bib/article/24/6/bbad366/7313455 by Zhejiang University user on 22 October 2023