



Identifying the critical states of complex diseases by the dynamic change of multivariate distribution

Hao Peng [†], Jiayuan Zhong[†], Pei Chen and Rui Liu 

Corresponding authors: Pei Chen, School of Mathematics, South China University of Technology, Guangzhou 510640, China. Tel: +86-13016079219; E-mail: chenpei@scut.edu.cn; Rui Liu, School of Mathematics, South China University of Technology, Guangzhou 510640, China and Pazhou Lab., Guangzhou 510330, China. Tel: +86-13016098877; E-mail: scliurui@scut.edu.cn

[†]Hao Peng and Jiayuan Zhong are contributed equally to this work.

Abstract

The dynamics of complex diseases are not always smooth; they are occasionally abrupt, i.e. there is a critical state transition or tipping point at which the disease undergoes a sudden qualitative shift. There are generally a few significant differences in the critical state in terms of gene expressions or other static measurements, which may lead to the failure of traditional differential expression-based biomarkers to identify such a tipping point. In this study, we propose a computational method, the direct interaction network-based divergence, to detect the critical state of complex diseases by exploiting the dynamic changes in multivariable distributions inferred from observable samples and local biomolecular direct interaction networks. Such a method is model-free and applicable to both bulk and single-cell expression data. Our approach was validated by successfully identifying the tipping point just before the occurrence of a critical transition for both a simulated data set and seven real data sets, including those from The Cancer Genome Atlas and two single-cell RNA-sequencing data sets of cell differentiation. Functional and pathway enrichment analyses also validated the computational results from the perspectives of both molecules and networks.

Keywords: dynamic network biomarker, distribution divergence, tipping point, multivariable distributions, critical transition, direct interaction network (DIN)

Introduction

Abundant clinical and experimental evidences show that the progression of many complex diseases is not always smooth, but sometimes with abrupt deterioration at a tipping point [1–4]. For instance, some chronic diseases exist for years or even decades before a catastrophic deterioration such as metastasis or the sudden onset of stroke that may occur within a short period of time [1, 5, 6]. Besides, the irreversible critical transitions also appear in a variety of biological processes such as cell fate commitment [7, 8]. The identification of such tipping point is not only crucial for a better understanding of the underlying mechanisms during the disease progression but also provides the early warning signal of the upcoming deterioration for diagnosis reference. There are many outstanding studies being presented to investigate the mechanisms of complex diseases from a microscopic perspective [9–12]. Generally, the progression of a complex disease is modelled as a non-linear dynamic

system, while a critical transition can be regarded as the system state shift at a bifurcation point [2, 13]. With such settings, the progression process of a complex disease is roughly divided into three stages/states (Figure 1A): a before-transition state, a pre-disease/critical state and a disease/after-transition state. Specifically, for complex diseases, the before-transition state is a stable and relatively healthy stage with high resilience. The pre-disease state is an unstable critical state with low resilience just before the deterioration at a tipping point. The after-transition state is another stable stage after the irreversible disease deterioration [14, 15]. Clearly, detecting the pre-disease state may offer appropriate timing for effective medical interventions that prevent or delay an undesirable critical transition.

Based on the critical slowing down phenomenon [16], the dynamic network biomarker (DNB) was recently proposed to qualitatively describe the dynamics of a biological system when it is in a critical state [14, 17]. Specifically, when the system approaches a bifurcation point,

Hao Peng is a PhD candidate at the South China University of Technology, and he received her BS degree. His current research interests include deep neural networks and data mining for complex dynamic systems.

Jiayuan Zhong is a PhD candidate at the South China University of Technology. His research interest mainly focuses on developing computational approaches to detect the tipping points of nonlinear dynamic systems.

Pei Chen received her BS and MS degrees from Peking University, and PhD degree from the South University of Technology. Currently, she is an associate professor at the South China University of Technology. Her research interest includes deep learning, data mining and computational biology.

Rui Liu is a full professor at the School of Mathematics, South China University of Technology and also affiliated with Pazhou Lab., Guangzhou. He received the BS and PhD degrees in applied mathematics from Peking University. His research interest includes nonlinear dynamics, modelling and computational methods.

Received: March 2, 2022. **Revised:** April 10, 2022. **Accepted:** April 18, 2022

© The Author(s) 2022. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

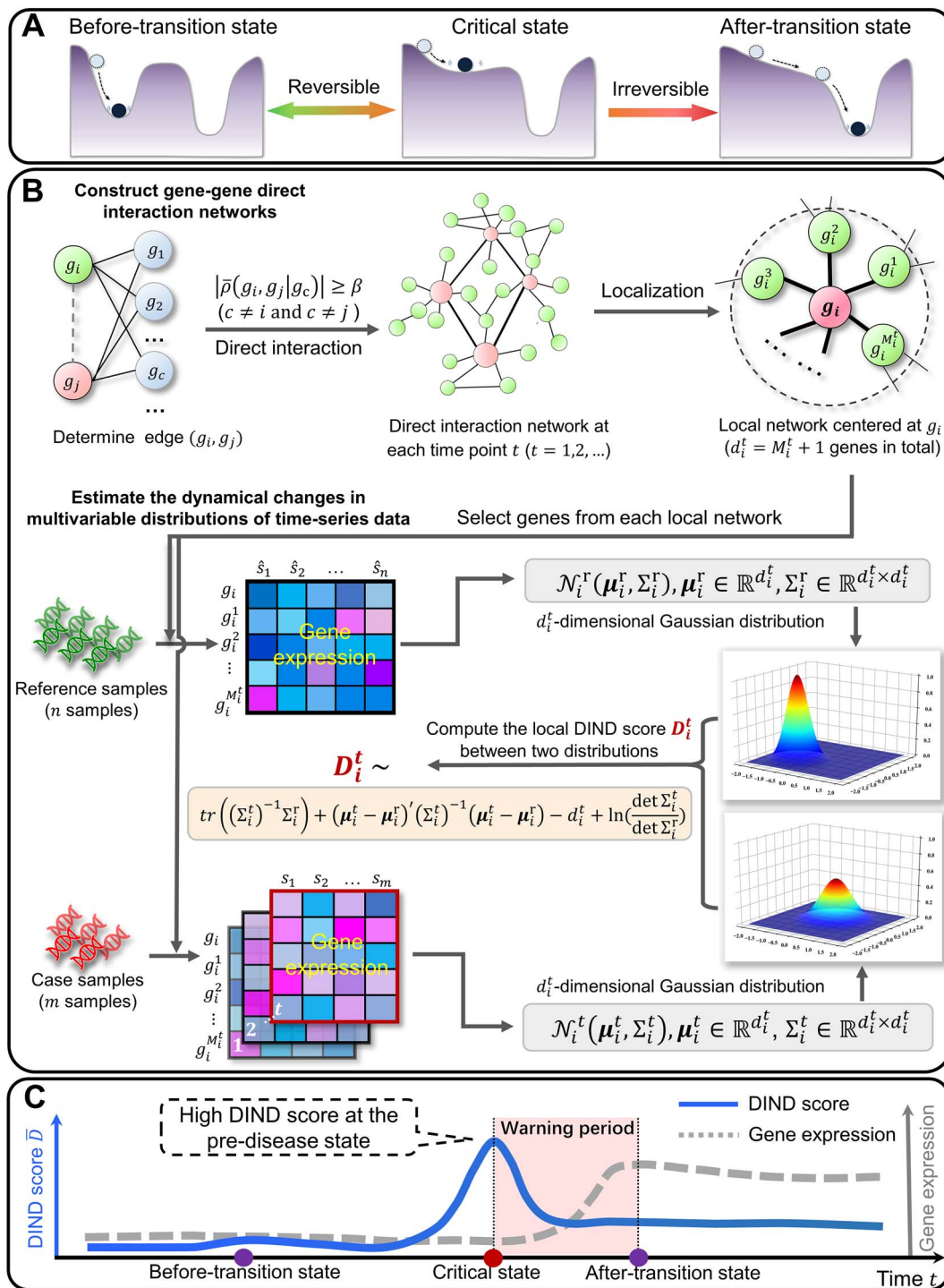


Figure 1. The schematic of the DIND method for measuring the differential distribution and detecting the critical transition of complex diseases. **(A)** The progression of a disease is roughly divided into three states: a before-transition state, a pre-disease/critical state and a disease/after-transition state. Generally, there are significant differences between the before- and after-transition states but few differences in terms of gene expression between the before-transition and critical states. **(B)** Given a group of control samples from relatively healthy individuals and case samples derived at time point t , a gene-gene DIN is constructed by excluding indirect interactions. Then, the DIN can be partitioned into a set of local networks. Thus, for each local network centered at gene g_i , the local DIND D_i^t [Equation (6)] is utilized to quantify the difference between two distributions of the local DIN with two sets of samples. **(C)** During the progression of a complex disease, the DIND index [Equation (7)] is relatively high in the pre-disease state and relatively low in the before-transition state. Such significant changes in DIND can indicate the critical state of a complex disease.

there is a dominating group of biomolecules, i.e. the DNB group, which behaves dynamically in a strongly collective manner and has a few generic properties that can be used to characterize the dynamic changes in molecular

associations rather than their expression patterns. The DNB offers the theoretical background for developing the computational methods for identifying the critical transition with early warning signals [2, 3, 18–20]. More

details of the DNB concept can be found in Supplementary Section A (see Supplementary Data available online at <http://bib.oxfordjournals.org/>).

In this study, we propose a computational approach, direct interaction network-based divergence (DIND), based on the combination of direct interaction network (DIN) inference and the DNB concept, to characterize the differential distribution of samples, thus detecting the critical state in a robust manner during a complex disease process (Figure 1). Specifically, based on a network inference procedure by determining whether an edge or a direct interaction relationship exists between two biomolecules, a local network-based Kullback–Leibler (KL) divergence is utilized to quantify the dynamic difference between two multivariate distributions that are estimated from the temporally adjacent or stage-wise samples (Figure 1B). The proposed method provides an effective computational tool that facilitates analysis in two aspects. On the one hand, our method reconstructs a set of stage-specific networks by eliminating all indirect interactions among biomolecules and captures the significant dynamic changes in gene associations during the progression of a complex disease. On the other hand, DIND provides an applicable way to quantitatively detect the pre-disease state or a tipping point of complex disease (Figure 1C). Furthermore, based on the DIND score, one can identify a group of genes that contributed the most to the differential distribution as the signalling biomolecules for further functional analysis and selection of potential drug targets. The DIND method was applied to one numerical simulation data set and seven real-world data sets, including those with bulk sequencing and single-cell RNA sequencing (scRNA-seq) data. We successfully identified the pre-disease states for acute lung injury, colon adenocarcinoma (COAD), stomach adenocarcinoma (STAD), thyroid carcinoma (THCA) and lung adenocarcinoma (LUAD). In addition, signals involved in cell fate commitment were detected at the single-cell level in cell differentiation data sets, including those on human embryonic stem cells (hESCs) to definitive endoderm cells (DECs) and mouse embryonic fibroblasts (MEFs) to neurons. All the results are consistent with original clinical or experimental observations, supporting the effectiveness and robustness of the proposed method. The corresponding signalling genes were analyzed in terms of their functions and potential roles in the critical transitions.

Materials and methods

Theoretical background

An abrupt and catastrophic deterioration of a complex disease is usually mathematically described as a state shift or phase transition through a bifurcation at a tipping point. Thus, the progression of a complex disease is correspondingly divided into three states or stages (Figure 1A) [14, 21]: a before-transition state in which a

biological system is far away from the bifurcation and thus with high stability and resilience, a critical state in which the system is in the vicinity of the bifurcation point, unstable and sensitive to external perturbations and an after-transition state which is another stable state when the system cross the bifurcation point. According to the DNB theory [14], when a complex system approaches the tipping point, a group of variables/biomolecules, i.e. the DNB group, arises with the following critical behaviours:

- The correlation between each pair of members in the DNB group rapidly increases;
- The correlation between a DNB member and any other non-DNB molecule rapidly decreases;
- The variation of each member in the DNB group drastically increases.

Clearly, it is the dynamic change in the molecular association and fluctuation rather than differential gene expression that makes a difference. The complete proof of these properties was presented in our previous work [14]. From the above description, there emerged a generic property that there are differential multivariate distributions of samples from the critical state. By exploiting such differential distributions, it is possible to detect the upcoming critical transition in a robust way. Therefore, the dynamic multivariate distributions are to be inferred from samples. To obtain the accurate distributions, we first constructed a set of DINs, so that dynamically changed molecular interactions can be employed in the distribution estimation. Then, the difference between the estimated multivariate distributions can be measured by the KL divergence.

DIN construction

The interaction of two genes determined by the Pearson correlation coefficient is widely used. However, there is an overestimation of gene–gene direct interactions due to common neighbours. In order to accurately depict the biomolecular associative relationship, we construct a gene–gene DIN by excluding indirect interactions. Given a microarray data set with M genes and n samples, denote $\mathbf{X} = [x_1, x_2, \dots, x_n]'$ and $\mathbf{Y} = [y_1, y_2, \dots, y_n]'$ as the expressions of two genes g_X and g_Y whose direct interaction is to be determined and $\mathbf{Z} = [z_1, z_2, \dots, z_n]'$ as the expressions of a common neighbour g_Z of g_X and g_Y , where the symbol $'$ represents the transpose of a vector. We decide whether there is a direct interaction between genes g_X and g_Y in the following two steps.

- (1) Calculate two coefficient parameters, w_x and w_y , as in Equation (1):

$$\begin{aligned} w_x &= \arg \min_w \sum_{i=1}^n (x_i - wz_i)^2 = \frac{\sum_{i=1}^n x_i z_i}{\sum_{i=1}^n z_i^2} \\ w_y &= \arg \min_w \sum_{i=1}^n (y_i - wz_i)^2 = \frac{\sum_{i=1}^n y_i z_i}{\sum_{i=1}^n z_i^2} \end{aligned} \quad (1)$$

The detailed solution process is provided in Supplementary Section H (see Supplementary Data available online at <http://bib.oxfordjournals.org/>). Then, the residuals $e_{X,i}$ and $e_{Y,i}$ are presented as follows:

$$\begin{aligned} e_{X,i} &= x_i - w_{XZ_i} \\ e_{Y,i} &= y_i - w_{YZ_i} \end{aligned} \quad (2)$$

- (2) Calculate the direct-interaction index $\rho(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$ between genes g_X and g_Y with a common neighbour g_Z as follows:

$$\rho(\mathbf{X}, \mathbf{Y}|\mathbf{Z}) = \frac{n \sum_{i=1}^n e_{X,i} e_{Y,i} - \sum_{i=1}^n e_{X,i} \sum_{i=1}^n e_{Y,i}}{\sqrt{n \sum_{i=1}^n e_{X,i}^2 - \left(\sum_{i=1}^n e_{X,i} \right)^2} \sqrt{n \sum_{i=1}^n e_{Y,i}^2 - \left(\sum_{i=1}^n e_{Y,i} \right)^2}} \quad (3)$$

[[DmEquation3]]

Then, the average score $\bar{\rho}(\mathbf{X}, \mathbf{Y}|\mathbf{Z}_{\text{all}})$ across all possible common neighbouring genes was used to build the gene-gene DIN at each time point, that is, if $|\bar{\rho}(\mathbf{X}, \mathbf{Y}|\mathbf{Z}_{\text{all}})| \geq \beta$, then there is a direct link between g_X and g_Y , where β is a data set-specific constant. Thus, a time-point/stage-specific DIN with average score $\bar{\rho}(\mathbf{X}, \mathbf{Y}|\mathbf{Z}_{\text{all}})$ as the edge weight is constructed by excluding indirect associations influenced by the common neighbouring genes. More details of the DIN were presented in Supplementary Section B (see Supplementary Data available online at <http://bib.oxfordjournals.org/>).

Divergence evaluation

The KL divergence between two continuous random variable distributions \mathcal{P} and \mathcal{Q} is defined as:

$$D(\mathcal{P} \parallel \mathcal{Q}) = \int_{-\infty}^{\infty} p(x) \log \left(\frac{p(x)}{q(x)} \right) dx, \quad (4)$$

where p and q denote the probability densities of \mathcal{P} and \mathcal{Q} , respectively. Generally, we assume that the variables (the expression levels of genes) in a local gene network (Figure 1B) conform to a multivariate normal distribution; thus, the divergence between two multivariate normal distributions \mathcal{N}_1 and \mathcal{N}_2 with means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and (non-singular) covariance matrices $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$ is as follows:

$$\begin{aligned} D(\mathcal{N}_1 \parallel \mathcal{N}_2) &= \frac{1}{2} (\text{tr}(\boldsymbol{\Sigma}_2^{-1} \boldsymbol{\Sigma}_1) \\ &+ (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)' \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + \ln \left(\frac{\det \boldsymbol{\Sigma}_2}{\det \boldsymbol{\Sigma}_1} \right) - d), \end{aligned} \quad (5)$$

where d denotes the dimension of two multivariate normal distributions. In general, $D(\mathcal{N}_1 \parallel \mathcal{N}_2) \neq D(\mathcal{N}_2 \parallel \mathcal{N}_1)$. Thus, a symmetric divergence is adopted in our method and is defined as

$$D(\mathcal{N}_1, \mathcal{N}_2) = \frac{D(\mathcal{N}_1 \parallel \mathcal{N}_2) + D(\mathcal{N}_2 \parallel \mathcal{N}_1)}{2}. \quad (6)$$

Algorithm for identifying the tipping point

Given a series of reference/control samples (samples from a relatively healthy normal cohort which represents the healthy or relatively healthy individuals) and a number of case samples, we identify the tipping point/critical state in the following steps:

- (1) At each time point t ($t = 1, 2, \dots$), construct a DIN by using the method proposed in Section 2.2.
- (2) Localize the DIN, such that each local DIN contains a centre gene g_i and its 1st-order neighbours $\{g_i^1, g_i^2, \dots, g_i^{M_i^t}\}$, ($i = 1, 2, \dots, M$, M is the total number of genes and $d_i^t = M_i^t + 1$ is the number of genes in the g_i -centered local network at time point t).
- (3) Fit a multivariate normal distribution for each local DIN. Specifically, for d_i^t genes in a local network centered at gene g_i , two multivariate normal distributions \mathcal{N}_i^r and \mathcal{N}_i^t were obtained from the reference and case samples at time point t (Figure 1B), that is, the d_i^t -dimensional vectors $\boldsymbol{\mu}_{g_i}^r, \boldsymbol{\mu}_{g_i}^t$ and $d_i^t \times d_i^t$ matrices $\boldsymbol{\Sigma}_i^r, \boldsymbol{\Sigma}_i^t$ were obtained.
- (4) Calculate the local DIND score D_i^t between $\mathcal{N}_i^r, \mathcal{N}_i^t$ based on Equation (6). Then, the DIND score \bar{D}^t is calculated as:

$$\bar{D}^t = \frac{1}{\hat{M}} \sum_{i=1}^{\hat{M}} D_i^t, \quad (7)$$

where \hat{M} is the number of top 5% genes with the largest local DIND scores and D_i^t denotes the corresponding local DIND scores.

According to the DNB theory, DNB biomolecules show significant collective behaviours with strong fluctuations, when a complex system approaches the critical transition [18]. The distributions of local networks containing DNB biomolecules in the critical state exhibit significant differences from those in the before-transition state, which may lead to an abrupt increase in the DIND score (Equation (7)).

Data processing and functional analysis

We eliminated the probes which did not include corresponding NCBI Entrez gene symbols. The average value for each gene mapped by multiple probes was recorded as its expression. Then, gene expression was normalized using Z-score.

The functional annotations were performed with the NCBI Gene database (<http://www.ncbi.nlm.nih.gov/gene>). The enrichment analyses were using web service tools from the Gene Ontology Consortium (<http://geneontology.org>) and client software from Ingenuity Pathway Analysis (IPA, <http://www.ingenuity.com/products/ipa>).

Results

To demonstrate the performance of DIND in identifying the critical state, we applied it to a numerical simulation

data set and seven real-world high-throughput omics data sets, including those with bulk sequencing data including acute lung injury (GSE2565), LUAD, STAD, THCA and COAD from The Cancer Genome Atlas (TCGA) database (<http://cancergenome.nih.gov>) and scRNA-seq data (embryonic differentiation of hESCs to DEC cells (ID: GSE75748) [22] and MEFs to neurons (ID: GSE67310) [23] from the NCBI Gene Expression Omnibus database <http://www.ncbi.nlm.nih.gov/geo>). For each application, the local DIND score D_i (for each local network) was calculated and sorted. Then, the top 5% genes with the largest DIND scores were taken as the DIND signalling genes.

Performance on numerical simulation

As shown in Figure 2A, a theoretical network model is used to illustrate how DIND detects the early warning signal when a system approaches a tipping point. Such a regulatory network with eight nodes is constructed by a set of stochastic differential equations in Michaelis–Menten form Equation (S4), which is frequently employed to study the biomolecular regulatory activities such as transcription and translation processes [24–26] and other multi-stable non-linear biological processes [27, 28]. A data set was generated for numerical simulation from the network by Equation (S4), with parameter q varying from -0.1 to 0.1 and $q = 0$ as the bifurcation point. Details of the network are shown in Supplementary Section C (see Supplementary Data available online at <http://bib.oxfordjournals.org/>).

We randomly set 50 initial values and applied DIND to the numerical simulation with each initial value independently. As shown in Figure 2B, when the network system approaches the bifurcation point at $q = 0$, the DIND \bar{D} score increases significantly, which indicates the upcoming critical transition. The DIND scores and their distributions shown in Figure 2B provide a global view of how DIND changes in the whole process, during which the system undergoes a stability reversal at $q = 0$. As shown in the first two rows of Figure 2C, dramatic changes of the distributions of the DIND signalling nodes occur when the system approaches the bifurcation point. The last row of Figure 2C shows the almost invariant distribution of non-signalling nodes which are insensitive to the approaching of critical transition. To better illustrate the characteristics of different local D_i scores at the tipping point, we visualized a landscape of the evolution of the local scores D_i in Supplementary Figure S2 (see Supplementary Data available online at <http://bib.oxfordjournals.org/>). It is seen that some of the local D_i scores (D_1, D_2, D_3, D_4, D_5) exhibit an abrupt increase before the tipping point ($q = -0.005$), that is, the expression levels of these nodes fluctuate significantly, resulting in a distinct multivariate distribution when the regulatory network approaches the tipping point. Moreover, the effectiveness of DIND under different noise situations is shown in Supplementary Section K (see Supplementary Data available

online at <http://bib.oxfordjournals.org/>). The details of this dynamic system are provided in Supplementary Section C (see Supplementary Data available online at <http://bib.oxfordjournals.org/>).

Identifying cell fate commitment during embryonic differentiation

Pluripotent stem cells play an important role in *in vitro*/biliary disease modelling and drug discovery [29–31], and the study of differentiation of nonneural cells to functional neurons has great promise in neurological disease modelling [32]. To reveal the underlying mechanism of cell differentiation which has a close relationship with a series of diseases, we applied the DIND method to two sets of cell differentiation scRNA-seq data sets: hESCs to DEC cells and MEFs to neurons. Details of these cell differentiation processes are provided in Supplementary Section G (see Supplementary Data available online at <http://bib.oxfordjournals.org/>).

In both data sets, the DIND scores \bar{D} were calculated to quantify the criticality of the cell population. The dynamic evolutions of the direct gene–gene networks are shown in Supplementary Figures S11 and S12 (see Supplementary Data available online at <http://bib.oxfordjournals.org/>). As shown in Figure 3A, the DIND score from hESC-to-DEC process increases significantly at 36 h, which is prior to the differentiation induction into definitive endoderm (DE) at 72 h [22, 33]. As shown in Figure 3B, for MEF-to-neuron process, the DIND score rises sharply from day 5 to day 20, which provides an early warning signal of the upcoming differentiation of mouse embryonic intermediate cells into induced neuron cells at day 22 [29]. Dynamic changes of the distributions of signalling genes for hESC-to-DEC and MEF-to-neuron processes are shown in Supplementary Figures S13 and S14 (see Supplementary Data available online at <http://bib.oxfordjournals.org/>), respectively. Besides, it is seen that the clustering results of the cells clearly distinguish between stages before and after the identified tipping points around 36 h for hESC-to-DEC and day 20 for MEF-to-neuron (Figure 3C and D).

Figure 3E depicts the underlying mechanism revealed by the functional analysis of the signalling genes for the hESC-to-DEC data set. The upstream regulator collagen type I alpha 2 chain (COL1A2) is a component of the extracellular matrix and may drive cells into a developmental critical transition during cell differentiation. Specifically, this regulator play crucial roles in the upregulation of integrin beta 1 (ITGB1), which together with the upregulation of Erb-B2 receptor tyrosine kinase 4 (ERBB4), GNG11 and protein phosphatase 2 regulatory subunit Bbeta (PPP2R2B) expression, activates the phosphatidylinositol 3-kinase/protein kinase B (PI3K/AKT) pathway and then downregulates the expression of the downstream cyclin D2 (CCND2) gene and promotes cell proliferation and differentiation. According to the literature [22], the detected tipping point

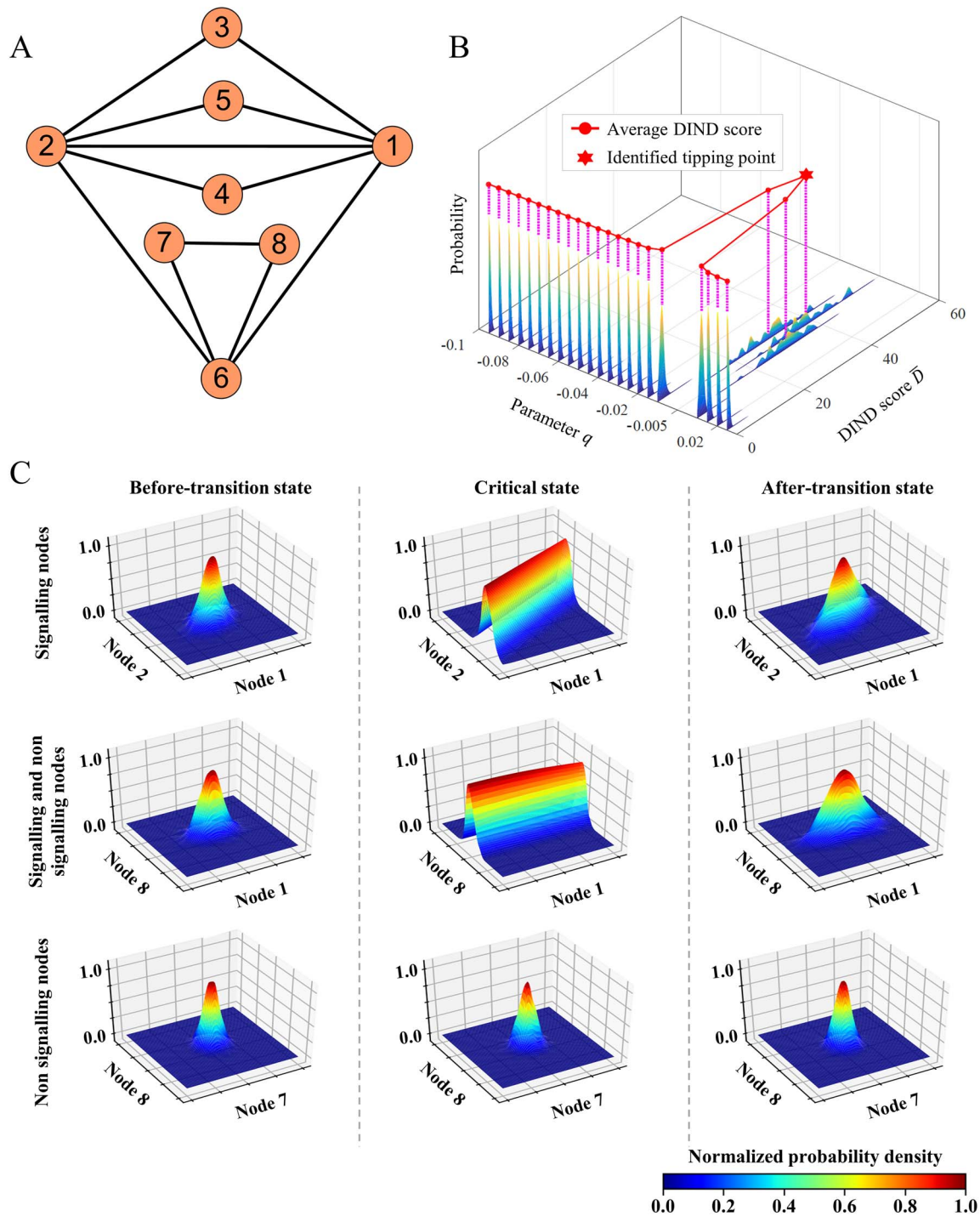


Figure 2. The illustration of the DIND performance through a representative network model. (A) The simulation data sets were generated from an eight-node network, which is governed by a set of differential equations with Michaelis–Menten form. (B) The curve of the DIND \bar{D} [defined in Equation (7)] for numerical simulation with 50 rounds is presented. The sudden increase in the DIND score around $q = -0.005$ signals an upcoming state transition, which is coincident with the fact that the stability of the system shifts at the bifurcation parameter value $q = 0$. (C) Illustration of the distributions constructed by gene pairs from DIND signalling variables or other variables. Notably, when the system approaches the bifurcation point $q = 0$, the distributions related to the DIND signalling nodes (the first two rows) exhibit great differences at the critical state while those of non-signalling nodes remain almost unchanged.

may be a key time point of guiding the differentiation of pluripotent stem cells to DE since there was a significant overturn in the expression of related regulatory genes between 24 and 72 h. Moreover, among the DIND signalling genes, there are some non-differentially

expressed genes which are closely related to important biological functions during the cell differentiation and shown in Supplementary Table S2 (see Supplementary Data available online at <http://bib.oxfordjournals.org/>).

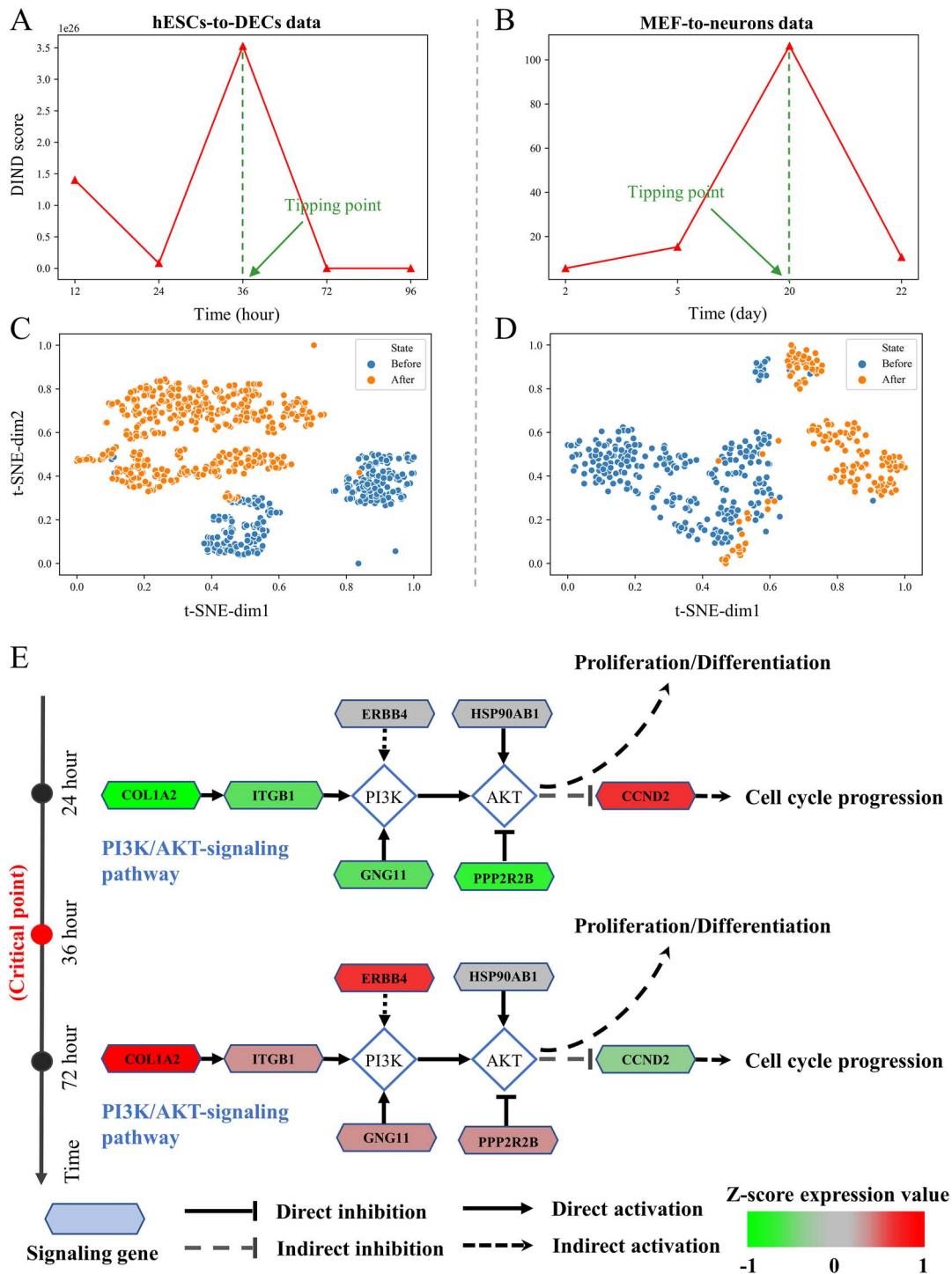


Figure 3. Identification of the critical states of two cell differentiation scRNA-seq data sets. The DIND curves of cell differentiation processes for (A) hESC-to-DEC and (B) MEF-to-neuron. The temporal clustering analysis t-distributed stochastic neighbor embedding (t-SNE) of the DIND signalling genes of (C) hESC-to-DEC and (D) MEF-to-neuron. The clustering based on the DIND signalling genes can distinguish cells before and after the transition state (represented in different colours). (E) Functional analysis shows that the DIND signalling genes were associated with cell proliferation and differentiation for hESC-to-DEC differentiation.

Identifying the critical state during cancer progression

By using DIND, we identified the critical states for four cancers COAD, THCA, STAD and LUAD based on their TCGA data sets. The number of samples within each stage in four tumour data sets is presented

in Supplementary Table S1 (see Supplementary Data available online at <http://bib.oxfordjournals.org/>). The tumour-adjacent samples were used as the reference samples, while the tumour samples were regarded as the stage-wise case samples based on their clinical information.

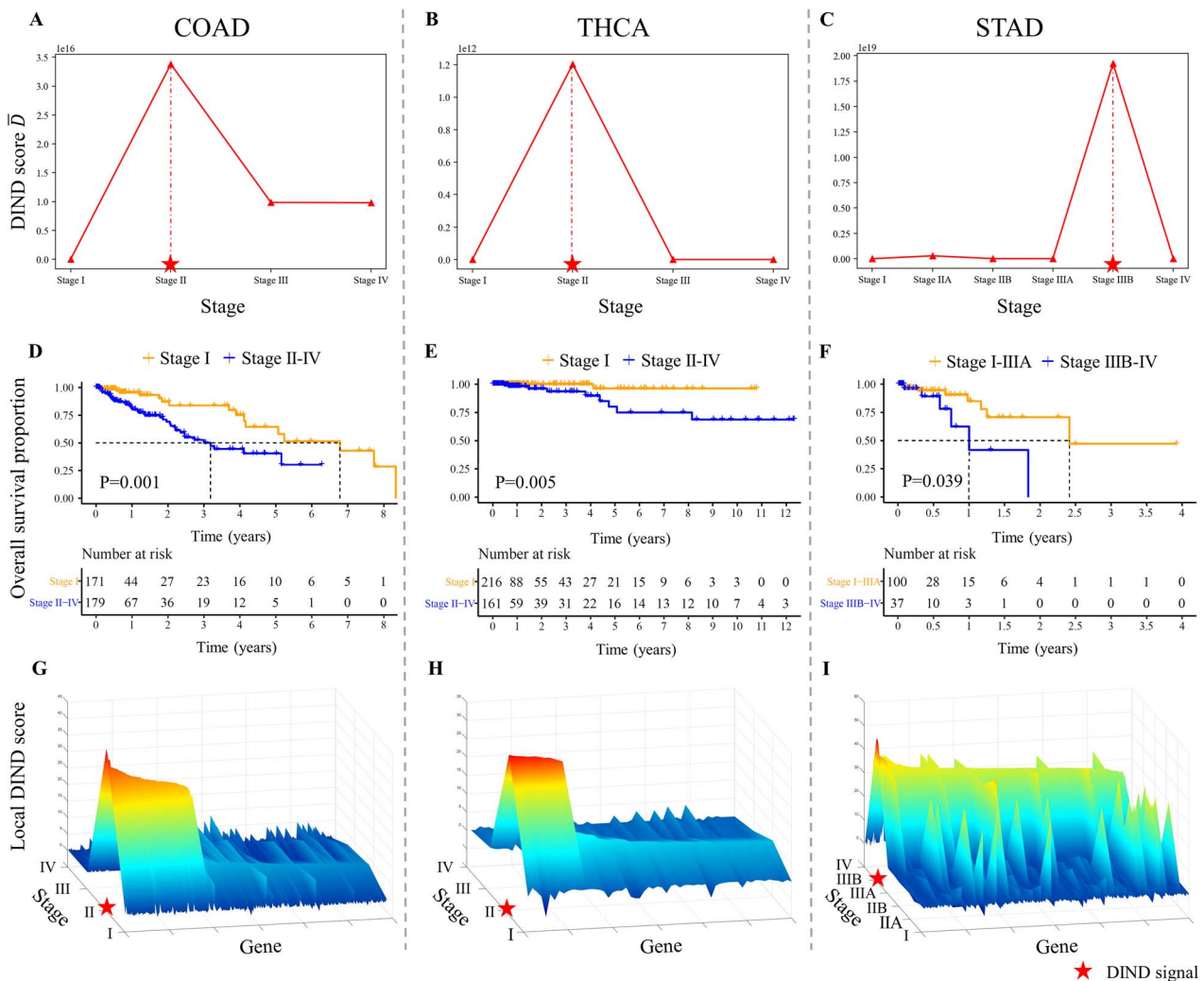


Figure 4. Detection of the critical states for cancers metastasis. The DIND score for (A) COAD, (B) THCA and (C) STAD. Survival analysis before and after the identified critical states for (D) COAD, (E) THCA and (F) STAD. Landscapes of the local DIND score for (G) COAD, (H) THCA and (I) STAD, which show the dynamic changes of multivariate distributions in a global view.

The critical states for both COAD and THCA were detected at Stage II (Figure 4A and B), and those for STAD and LUAD were detected at Stage IIIB (Figures 4C and 5A); the landscapes (Figures 4G-I and 5C), which show the dynamic changes of local multivariate distributions in a global view, also demonstrate the systematic abnormality. Most mortality of cancer patients is metastasis: the process of which tumor cells migrate from the primary focus and colonize distant areas [34]. Thus, it is important to detect the critical state/tipping point before the occurrence of distant cancer metastasis or lymph node metastasis, so that chemotherapy, radiotherapy and other strategies can be carried out in a timely manner to slow cancer progression or prevent serious deterioration [35–38]. Therefore, the early warning signals provided by DIND in the progression of tumour may contribute to appropriate and timely clinical interventions.

Specifically, for the COAD data set, the abrupt increase of the DIND score from Stages I to II, illustrated in Figure 4A, indicates that a critical deterioration event would occur after Stage II. In fact, lymph node metastasis

and direct metastasis to another organ or structure usually occur at Stage III [39]. The peak of the DIND at Stage II for THCA data reveals the upcoming critical transition, which is in accordance with the fact that critical deterioration, including extension into sternothyroid muscle or parathyroid soft tissues and regional lymph node metastasis, appears at Stage III [40, 41].

As shown in Figures 4D–F and 5B, the survival curves before and after the critical stage are easily distinguishable with significant P-values $P = 0.001$, $P = 0.005$, $P = 0.039$ and $P = 0.012$ for COAD, THCA, STAD and LUAD respectively, which show that the patients diagnosed before the identified stages have a substantially better prognosis than those after the identified stages, suggesting that the detected early warning signals are able to predict upcoming serious deterioration. More details of the survival analysis of cancers are presented in Supplementary Section E and Figure S6 (see Supplementary Data available online at <http://bib.oxfordjournals.org/>).

In Figure 5D, we illustrated the evolution of DINs constructed by signalling genes for LUAD data. It

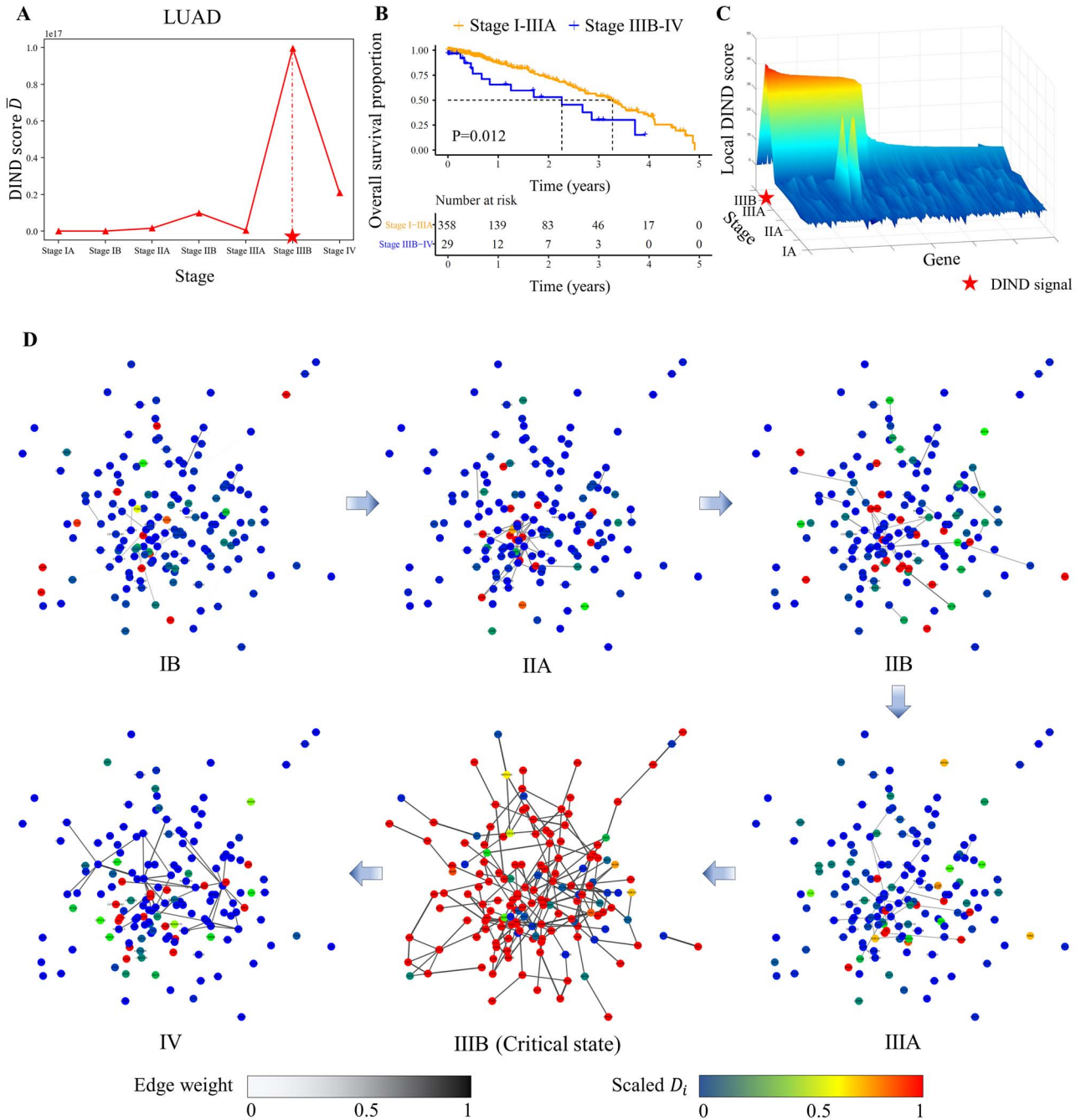


Figure 5. Identification of the critical state for LUAD. (A) The curve of average DIND score for LUAD. (B) Survival analysis before and after the critical stages in LUAD. (C) Landscape of the local DIND score for LUAD. (D) The dynamic evolution of DINDs constructed by signalling genes of LUAD.

is found that there is a significant change in the structure of DINDs at Stage IIIB, indicating the critical transition later into distant metastasis at Stage IV. All these results demonstrate that our proposed method is capable of effectively detecting the early warning signals of serious deterioration during cancer progression to identify the critical stage, which can be regarded as an important indicator for patient survival. In [Supplementary Figure S7](#) (see [Supplementary Data available online at http://bib.oxfordjournals.org/](http://bib.oxfordjournals.org/)), the details of dynamic changes in terms of distributions

containing signalling and non-signalling genes are illustrated, reflecting the increase in DIND score. The signalling genes collagen type IV alpha 2 chain (COL4A2), collagen type IV alpha 4 chain (COL4A4) are enriched in the PI3K-Akt signalling pathway related to LUAD [41]. Dynamic changes of the distributions of signalling genes for COAD, THCA and STAD were shown in [Supplementary Figures S8–S10](#) (see [Supplementary Data available online at http://bib.oxfordjournals.org/](http://bib.oxfordjournals.org/)), respectively. Moreover, the dynamic evolutions of DINDs constructed by signalling genes for the above three can-

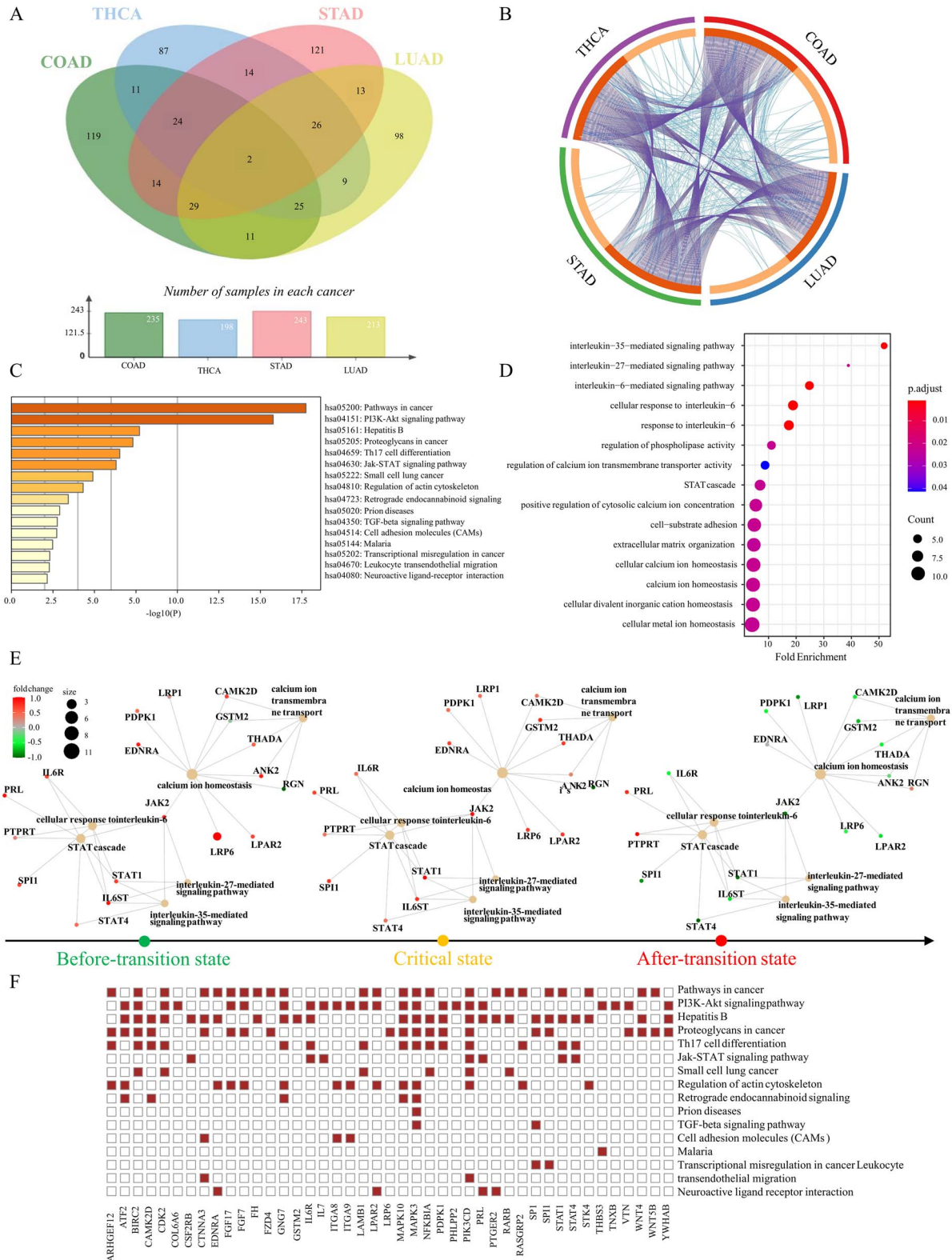


Figure 6. Functional analysis of common DIND signalling genes in different cancer cohorts. **(A)** Common DIND signalling genes among COAD, THCA, STAD and LUAD. **(B)** Plenty of overlap appears not only in the identical signalling genes among the four cancers but also in their biological functions. The outer ring represents different groups of cancer signalling genes, and the inner ring represents their identical genes and functions. The identical genes are linked with each other are depicted in purple lines, and functions are linked are depicted in blue lines. **(C)** These common genes were found enriched in cancer-related functional annotations. **(D)** Functional enrichment through Gene Ontology (GO) analysis showed that the common DIND signalling genes are enriched in multiple cancer-related biological processes. **(E)** The activity of some biological processes changed from a high level to a low level after reaching the critical state. **(F)** The specific DIND signalling genes are involved in multiple cancer-related pathways.

cers were shown in [Supplementary Figures S15–S17](#) (see [Supplementary Data](#) available online at <http://bib.oxfordjournals.org/>), respectively.

Functional analysis of the common DIND signalling genes among four cancers

Functional analysis was carried out to the common DIND signalling genes across the four cancers ([Figure 6A](#)). There are not only rich intersections across the identities of the signalling genes in different cancers but also close relationships of their functions ([Figure 6B](#)). Based on the functional enrichment analysis (IPA), these common DIND signalling genes are enriched in the PI3K-Akt signalling pathway [42], the Jak-STAT signalling pathway [43] and other pathways related to cancer ([Figure 6C](#)). Furthermore, these common genes play a crucial role in the biological processes associated with the progression of cancer from GO analysis ([Figure 6D](#)), such as the interleukin-35-mediated signalling pathway [44], the interleukin-6-mediated signalling pathway [45], the STAT cascade [46] and cellular calcium ion homeostasis [47]. We also found that the expression of common genes was associated with some biological processes, such as the regulation of calcium ion transmembrane transporter activity, calcium ion homeostasis, the interleukin-27-mediated signalling pathway, the interleukin-35-mediated signalling pathway and the STAT cascade, which changed from high levels to low levels ([Figure 6E](#)). Here, we focus on the role of those common signalling genes in both cancer-related GO biological processes, Kyoto Encyclopedia of Genes and Genomes pathways ([Figure 6E and F](#)). For example, calcium/calmodulin dependent protein kinase II delta (CAMK2D) has been reported to play an important role in the regulation of proliferation, differentiation, metastasis and survival of various cancer cells [48]. The previous studies demonstrated that activated signal transducer and activator of transcription 1 (STAT1) exhibited pro-apoptotic and anti-proliferative effects [49, 50]. In gastric cancer, high signal transducer and activator of transcription 4 (STAT4) expression is vital for a good patient prognosis [51]. low-density lipoprotein receptor-related protein 6 (LRP6) is a cancer-related gene whose expression promotes cancer cell proliferation and tumorigenesis by altering the subcellular distribution of β -catenin [52]. Immunohistochemical analysis has demonstrated that lysophosphatidic acid receptor 2 (LPAR2) plays a significant role in the gastric cancer progression [53]. The pyruvate dehydrogenase lipoamide kinase isozyme 1 (PDK1) protein, a specific locus amplification of 3-phosphoinositide-dependent protein kinase 1 (PDPK1), has been shown to be correlated with poor survival in breast cancer patients and has also been detected in lung and prostate cancers [54, 55]. The interleukin 6 receptor (IL6R) gene is also reported to play a critical role in the progression of tumour cell growth and patient survival and the tumour microenvironment [56]. These common genes may be crucial keys in the study of the university

of cancers and could potentially contribute to clinical interventions.

For all data sets, DIND successfully detected the tipping points or early warning signals before the critical transition and irreversible disease states, which shows the effectiveness of our method. For the lung injury data set, the obvious change in the DIND score at the 3rd time point (4 h) signalled the upcoming critical transition of lung injury, which agrees with the observation in the original experiment [57]. The analysis results on the lung injury data were provided in [Supplementary Section D](#) (see [Supplementary Data](#) available online at <http://bib.oxfordjournals.org/>).

Discussion

Early diagnosis offers patients the best chance of mitigating the risks associated with severe diseases. Therefore, the detection of the early warning signal of the catastrophic deterioration is of great importance. However, because generally there is little state change before the critical transition, it is difficult to distinguish the pre-disease samples based on the traditional biomarkers. In this study, aiming at the identification of the critical states of complex diseases, the DIND method was proposed to explore the dynamic changes in both the biomolecular interaction and the multivariate distribution of gene groups during the disease progression and thus signalling the abnormalities. With the support of DIND, we identified the critical states of five complex diseases and two biological processes of cell differentiation. The analysis results agree well with the clinical or experimental observations. Besides, further study on the DIND signalling genes, which are a group of biomolecules sensitive to the change of local network structure, shows that some of these genes may play important roles in the disease progression.

As a model-free computational method, DIND is applicable in both bulk and single-cell genomics data. However, there is a limitation of DIND that it may have a poor performance if there are too few samples to fit a proper distribution. Furthermore, by combining with dynamics prediction method [58], it may not only help to identify the critical states based on omics data but also reveal the dynamically differential information that provides us with new insights of how the biological system behaves in the vicinity of its tipping point.

Key Points

- We proposed a new model-free method, the direct interaction network-based divergence (DIND), to detect the early warning signal of the critical transition during disease progressions.
- DIND explores the dynamic changes in both the biomolecular interaction and the multivariate distribution of gene groups during the disease progression and thus signalling the abnormalities.

- With the support of DIND, we identified the critical states of five complex diseases and two biological processes of cell differentiation, which agree well with the clinical or experimental observations.

Supplementary data

Supplementary data are available online at <https://academic.oup.com/bib>.

Funding

National Natural Science Foundation of China (grant nos. 62172164, 12026608); Guangdong Basic and Applied Basic Research Foundation (grant nos. 2019B151502062, 2021A1515012317).

Data and code availability

All data needed to evaluate the conclusions are present in the paper and/or the Supplementary Materials. All data and the codes used in this study are available at <https://github.com/Peng154/DIND>.

References

- Hu S, Aisner SC, Lubitz SE. Cinacalcet for management of tertiary hyperparathyroidism associated with chronic treatment of hypophosphatemia in an adult with tumor-induced osteomalacia. *AACE Clin Case Rep* 2015;**1**:e225–9.
- Liu R, Chen P, Chen L. Single-sample landscape entropy reveals the imminent phase transition during disease progression. *Bioinformatics* 2020;**36**:1522–32.
- Chen P, Li Y, Liu X, et al. Detecting the tipping points in a three-state model of complex diseases by temporal differential networks. *J Transl Med* 2017;**15**:217.
- Yang B, Li M, Tang W, et al. Dynamic network biomarker indicates pulmonary metastasis at the tipping point of hepatocellular carcinoma. *Nat Commun* 2018;**9**:1–14.
- Penney ME, Parfrey PS, Savas S, et al. A genome-wide association study identifies single nucleotide polymorphisms associated with time-to-metastasis in colorectal cancer. *BMC Cancer* 2019;**19**:1–12.
- Jayadevappa G, Ravishankar S. Risk factors and clinical profile of ischemic stroke patients attending emergency Care Facility in Bangalore City. *Sch J Appl Med Sci* 2021;**9**:572–7.
- Richard A, Boullu L, Herbach U, et al. Single-cell-based analysis highlights a surge in cell-to-cell molecular variability preceding irreversible commitment in a differentiation process. *PLoS Biol* 2016;**14**:e1002585.
- 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.
- Chen X, Sun L-G, Zhao Y. NCMCMDA: miRNA–disease association prediction through neighborhood constraint matrix completion. *Brief Bioinform* 2021;**22**:485–96.
- Zhao Y, Wang C-C, Chen X. Microbes and complex diseases: from experimental results to computational models. *Brief Bioinform* 2021;**22**:bbaa158.
- Chen X, Guan N-N, Sun Y-Z, et al. MicroRNA-small molecule association identification: from experimental results to computational models. *Brief Bioinform* 2020;**21**:47–61.
- Chu Y, Zhang Y, Wang Q, et al. A transformer-based model to predict peptide–HLA class I binding and optimize mutated peptides for vaccine design. *Nat Mach Intell* 2022;**4**:300–11.
- Liu R, Aihara K, Chen L. Collective fluctuation implies imminent state transition: comment on “Dynamic and thermodynamic models of adaptation” by AN Gorban et al. *Phys Life Rev* 2021;**37**:103–7.
- Chen L, Liu R, Liu Z-P, et al. Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. *Sci Rep* 2012;**2**:1–8.
- Liu R, Wang X, Aihara K, et al. Early diagnosis of complex diseases by molecular biomarkers, network biomarkers, and dynamical network biomarkers. *Med Res Rev* 2014;**34**:455–78.
- Steven HS. *Nonlinear Dynamics and Chaos: with Applications to Physics, Biology, Chemistry, and Engineering*. Boston, Massachusetts: Addison-Wesley, 1994.
- Liu R, Zhong J, Hong R, et al. Predicting local COVID-19 outbreaks and infectious disease epidemics based on landscape network entropy. *Sci Bull* 2021;**66**(22):2265–70.
- Liu R, Yu X, Liu X, et al. Identifying critical transitions of complex diseases based on a single sample. *Bioinformatics* 2014;**30**:1579–86.
- Chen P, Chen E, Chen L, et al. Detecting early-warning signals of influenza outbreak based on dynamic network marker. *J Cell Mol Med* 2019;**23**:395–404.
- Chen P, Liu R, Li Y, et al. Detecting critical state before phase transition of complex biological systems by hidden Markov model. *Bioinformatics* 2016;**32**:2143–50.
- Liu X, Chang X, Leng S, et al. Detection for disease tipping points by landscape dynamic network biomarkers. *Natl Sci Rev* 2019;**6**:775–85.
- Chu L-F, Leng N, Zhang J, et al. Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol* 2016;**17**:1–20.
- Treutlein B, Lee QY, Camp JG, et al. Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq. *Nature* 2016;**534**:391–5.
- De Jong H. Modeling and simulation of genetic regulatory systems: a literature review. *J Comput Biol* 2002;**9**:67–103.
- Sherman MS, Cohen BA. Thermodynamic state ensemble models of cis-regulation. *PLoS Comput Biol* 2012;**8**:e1002407.
- Cantone I, Marucci L, Iorio F, et al. A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. *Cell* 2009;**137**:172–81.
- Chen Y, Kim JK, Hirning AJ, et al. Emergent genetic oscillations in a synthetic microbial consortium. *Science* 2015;**349**:986–9.
- Li C, Chen L, Aihara K. Stability of genetic networks with SUM regulatory logic: Lur’e system and LMI approach. *IEEE Trans Circuits Syst I Fundam Theory Appl* 2006;**53**:2451–8.
- Musunuru K. Genome editing of human pluripotent stem cells to generate human cellular disease models. *Dis Model Mech* 2013;**6**:896–904.
- Ogawa M, Ogawa S, Bear CE, et al. Directed differentiation of cholangiocytes from human pluripotent stem cells. *Nat Biotechnol* 2015;**33**:853–61.
- Patsch C, Challet-Meylan L, Thoma EC, et al. Generation of vascular endothelial and smooth muscle cells from human pluripotent stem cells. *Nat Cell Biol* 2015;**17**:994–1003.

32. Chanda S, Ang CE, Davila J, et al. Generation of induced neuronal cells by the single reprogramming factor ASCL1. *Stem Cell Rep* 2014;**3**:282–96.
33. Teschendorff AE, Enver T. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. *Nat Commun* 2017;**8**:1–15.
34. Steeg PS, Bevilacqua G, Kopper L, et al. Evidence for a novel gene associated with low tumor metastatic potential. *J Natl Cancer Inst Monogr* 1988;**80**:200–4.
35. Bareschino MA, Schettino C, Rossi A, et al. Treatment of advanced non small cell lung cancer. *J Thorac Dis* 2011;**3**:122.
36. Twelves C, Wong A, Nowacki MP, et al. Capecitabine as adjuvant treatment for stage III colon cancer. *N Engl J Med* 2005;**352**:2696–704.
37. Gasent Blesa JM, Grande Pulido E, Provencio Pulla M, et al. Old and new insights in the treatment of thyroid carcinoma. *J Thyroid Res* 2010;**2010**:1–16.
38. Cheung DY, Kim JK. Perspectives of the stomach cancer treatment: the introduction of molecular targeted therapy and the hope for cure. *Korean J Gastroenterol* 2013;**61**:117–27.
39. Hari DM, Leung AM, Lee J-H, et al. AJCC Cancer staging manual 7th edition criteria for colon cancer: do the complex modifications improve prognostic assessment? *J Am Coll Surg* 2013;**217**:181–90.
40. Shaha AR. TNM classification of thyroid carcinoma. *World J Surg* 2007;**31**:879–87.
41. Li J, Wang J, Xie D, et al. Characteristics of the PI3K/AKT and MAPK/ERK pathways involved in the maintenance of self-renewal in lung cancer stem-like cells. *Int J Biol Sci* 2021;**17**:1191.
42. Luo J, Manning BD, Cantley LC. Targeting the PI3K-Akt pathway in human cancer: rationale and promise. *Cancer Cell* 2003;**4**:257–62.
43. Constantinescu SN, Girardot M, Pecquet C. Mining for JAK-STAT mutations in cancer. *Trends Biochem Sci* 2008;**33**:122–31.
44. Lee C-C, Lin J-C, Hwang W-L, et al. Macrophage-secreted interleukin-35 regulates cancer cell plasticity to facilitate metastatic colonization. *Nat Commun* 2018;**9**:1–18.
45. Kumari N, Dwarakanath B, Das A, et al. Role of interleukin-6 in cancer progression and therapeutic resistance. *Tumor Biol* 2016;**37**:11553–72.
46. Verhoeven Y, Tilborghs S, Jacobs J, et al. The potential and controversy of targeting STAT family members in cancer. *Semin Cancer Biol* 2020;**60**:41–56.
47. Durham AC, Walton JM. Calcium ions and the control of proliferation in normal and cancer cells. *Biosci Rep* 1982;**2**:15–30.
48. Wang Y, Zhao R, Zhe H. The emerging role of CaMKII in cancer. *Oncotarget* 2015;**6**:11725.
49. Bromberg JF, Horvath CM, Wen Z, et al. Transcriptionally active Stat1 is required for the antiproliferative effects of both interferon alpha and interferon gamma. *Proc Natl Acad Sci* 1996;**93**:7673–8.
50. Shankaran V, Ikeda H, Bruce AT, et al. IFN γ and lymphocytes prevent primary tumour development and shape tumour immunogenicity. *Nature* 2001;**410**:1107–11.
51. Nishi M, Batsaikhan B-E, Yoshikawa K, et al. High STAT4 expression indicates better disease-free survival in patients with gastric cancer. *Anticancer Res* 2017;**37**:6723–9.
52. Li Y, Lu W, He X, et al. LRP6 expression promotes cancer cell proliferation and tumorigenesis by altering β -catenin subcellular distribution. *Oncogene* 2004;**23**:9129–35.
53. Yamashita H, Kitayama J, Shida D, et al. Differential expression of lysophosphatidic acid receptor-2 in intestinal and diffuse type gastric cancer. *J Surg Oncol* 2006;**93**:30–5.
54. Muranen TA, Greco D, Fagerholm R, et al. Breast tumors from CHEK2 1100delC-mutation carriers: genomic landscape and clinical implications. *Breast Cancer Res* 2011;**13**:R90.
55. Choucair KA, Guérard K-P, Ejdelman J, et al. The 16p13.3 (PDPK1) genomic gain in prostate cancer: a potential role in disease progression. *Transl Oncol* 2012;**5**:453–60.
56. Taher MY, Davies DM, Maher J. The role of the interleukin (IL)-6/IL-6 receptor axis in cancer. *Biochem Soc Trans* 2018;**46**:1449–62.
57. Sciuto AM, Phillips CS, Orzolek LD, et al. Genomic analysis of murine pulmonary tissue following carbonyl chloride inhalation. *Chem Res Toxicol* 2005;**18**:1654–60.
58. Chen P, Liu R, Aihara K, et al. Autoreservoir computing for multi-step ahead prediction based on the spatiotemporal information transformation. *Nat Commun* 2020;**11**:4568.