


Single-cell causal network inferred by cross-mapping entropy

Lin Li , Rui Xia, Wei Chen, Qi Zhao, Peng Tao and Luonan Chen 

Corresponding author. L. Chen, Key Laboratory of Systems Biology, Center for Excellence in Molecular Cell Science, Shanghai Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai 200031, China. Tel.: 021-54920100; E-mail: lnchen@sibcb.ac.cn

Lin Li, Rui Xia and Wei Chen contributed equally to this work.

Abstract

Gene regulatory networks (GRNs) reveal the complex molecular interactions that govern cell state. However, it is challenging for identifying causal relations among genes due to noisy data and molecular nonlinearity. Here, we propose a novel causal criterion, neighbor cross-mapping entropy (NME), for inferring GRNs from both steady data and time-series data. NME is designed to quantify ‘continuous causality’ or functional dependency from one variable to another based on their function continuity with varying neighbor sizes. NME shows superior performance on benchmark datasets, comparing with existing methods. By applying to scRNA-seq datasets, NME not only reliably inferred GRNs for cell types but also identified cell states. Based on the inferred GRNs and further their activity matrices, NME showed better performance in single-cell clustering and downstream analyses. In summary, based on continuous causality, NME provides a powerful tool in inferring causal regulations of GRNs between genes from scRNA-seq data, which is further exploited to identify novel cell types/states and predict cell type-specific network modules.

Keywords: GRN, cross-mapping entropy, causality, NEM, single cell

INTRODUCTION

Revealing gene regulatory networks (GRNs) or mechanisms from scRNA-seq data is critically important for understanding cellular processes. A GRN for a cell not only regulates most biological processes [1–3], such as developmental, homeostatic and disease processes, but also characterizes the transcriptional state of this cell. In addition, insight of the network topology in a cell can actually provide deep understanding of molecular mechanisms in these processes at a network level.

With rapid development of high-throughput sequencing technology, many methods have been proposed to infer GRNs. Generally, these GRN inference methods fall into three categories [4], i.e. model-based, machine learning-based and information-theoretic approaches. Model-based approaches such as Boolean network methods [5, 6] and Bayesian network methods [7–10] can describe gene association through fitting gene expression profiles based on various models, which showed high flexibility and scalability. Machine learning-based methods such as GENIE3 [11] and GRNBoost2 [12] can identify the direction of regulation and obtain directed networks. However, these methods incorporate GRN inference based on linear models or tree-based models, which is hard to be generalized to more comprehensive nonlinear frameworks. Information-theoretic methods are the most popular methods in inferring GRNs. However, GRNs inferred by these

methods are generally undirected networks, which inhibited the application to real biological networks.

To address these problems, we proposed a novel information-theoretic method, neighbor cross-mapping entropy (NME), to reliably estimate causality between variables and infer directed GRN for both steady-state data and time-series data, which overcomes the limitation of information-based GRN methods (Figure 1A and Supplementary Figures S1, S2). NME can quantify the ‘continuous causality’ as well as its causal strength between two variables by varying neighbor size based on a rigorous mathematical framework, i.e. continuity scaling law [13, 14], which is also logically consistent with natural interpretation as functional dependency. Furthermore, we also extend NME to conditional neighbor cross-mapping entropy (cNME) to obtain direct causality between variables. Moreover, by exploiting network information from scRNA-seq data, we develop scNME based on NME/cNME which can identify stable cell states and also accurately cluster cell types in a robust manner. scNME can also effectively group cells from different species by eliminating species heterogeneity through GRN matrix. Furthermore, scNME can be applied to identify novel cell states and evaluate GRN activity for scRNA-seq datasets. We demonstrated the power of NME in inferring causal networks, which can be further applied to downstream analysis on various biological processes at a network level. NME is freely available at <https://github.com/LinLi-0909/NME>.

Lin Li is a postdoctoral fellow in Center for Excellence in Molecular Cell Science. Her research interest is computational systems biology.

Rui Xia is a PhD student in Center for Excellence in Molecular Cell Science. Her research interest is deep learning.

Wei Chen is a master student in Center for Excellence in Molecular Cell Science. Her research interests include computational systems biology.

Qi Zhao is a master student in Center for Excellence in Molecular Cell Science. His research interest is deep learning.

Peng Tao is a research associate in Hangzhou Institute for Advanced Study. His research interests include system biology.

Luonan Chen is a professor and executive director at Key Laboratory of Systems Biology in Center for Excellence in Molecular Cell Science. His interests include systems biology, computational biology and applied mathematics.

Received: March 16, 2023. **Revised:** July 3, 2023. **Accepted:** July 19, 2023

© The Author(s) 2023. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

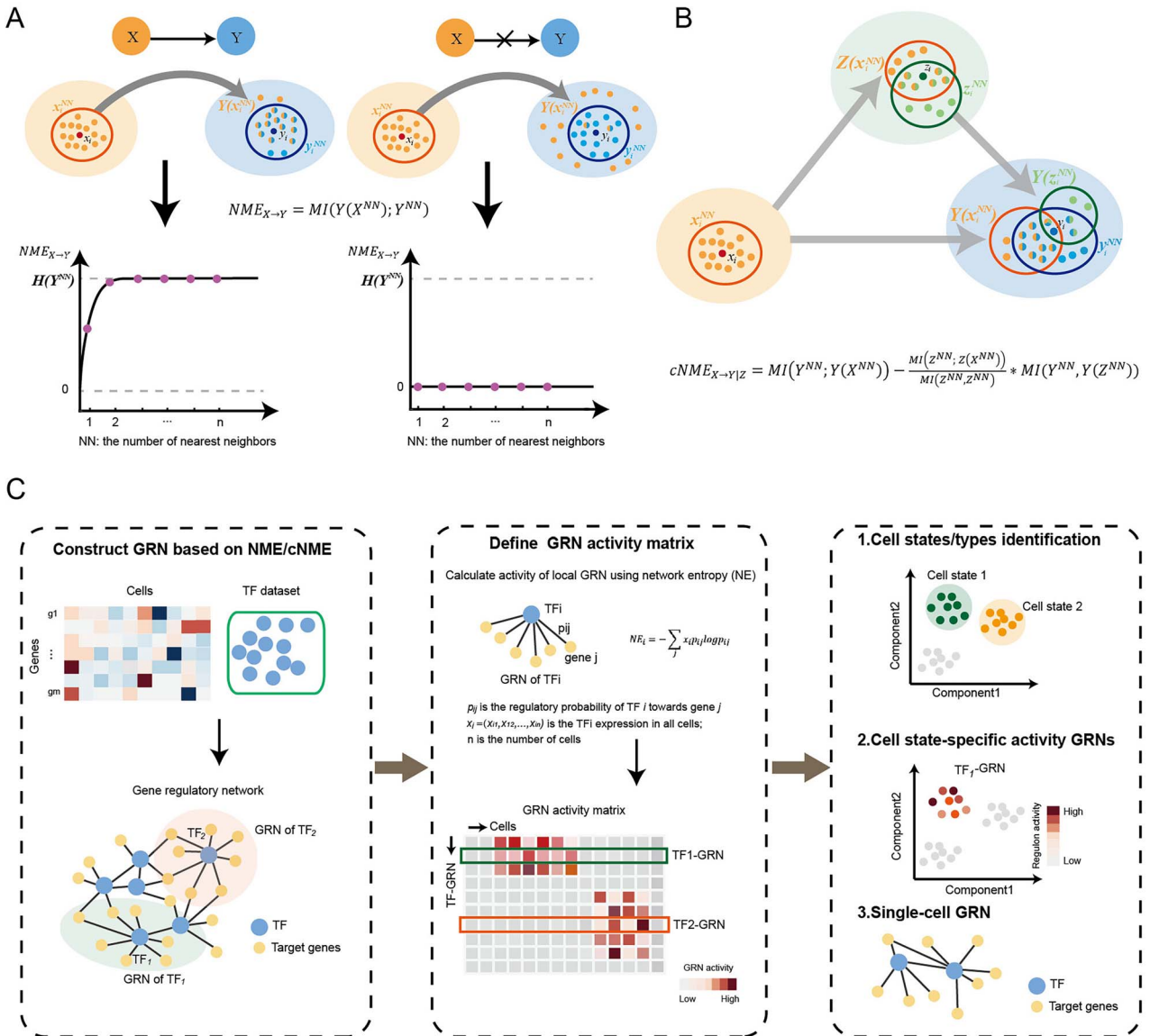


Figure 1. The framework of NME algorithm. **(A)** Illustration of causal relation between two variables quantified by neighbor cross-mapping entropy (NME) with varying NN. X and Y are two random variables with n samples. We define continuous causality using continuity of a function based on continuity scaling law or functional dependency, which is quantified by $NME_{X \rightarrow Y}$ to estimate the causal strength from X to Y. If there exists causality from X and Y, then mapping from the NN nearest neighbors x_i^{NN} of each x_i to Y, i.e. $Y(x_i^{NN})$, is expected to be a continuous function of x_i^{NN} , in other words $Y(x_i^{NN})$ are expected to be the NN nearest neighbors of y_i or have high overlap with y_i^{NN} . With varying NN, $NME_{X \rightarrow Y}$ keeps high, which can quantify such a continuity or causal strength. On the other hand, if there is no causality from X to Y, then mapping from the NN nearest neighbors x_i^{NN} of each x_i to Y, i.e. $Y(x_i^{NN})$, has no functional relation with x_i^{NN} , in other words $Y(x_i^{NN})$ are randomly distributed or have low overlap with y_i^{NN} . Thus, with varying NN, $NME_{X \rightarrow Y}$ always approaches 0. **(B)** Direct causal model between two variables based on conditional neighbor cross-mapping entropy (cNME). We consider the direct causality from X to Y in a third variable Z with n measured samples. **(C)** Schematic of scNME workflow. There are three main steps of scNME. 1) Inferring gene regulatory networks (GRNs) by cNME from scRNA-seq and TF datasets; 2) Defining the GRN activity in each cell by network entropy (Methods); 3) Clustering, dimensional reduction based on GRN activity matrix. Thus, cell state-specific activity GRNs and single-cell GRNs can be obtained.

MATERIALS AND METHODS

Continuous causality and NME framework

Here, we define continuous causality based on functional continuity from one variable to another with the scaling law and propose a new causality criterion, i.e. NME, which can be used to infer direct gene regulatory network from omics data (Figure 1A). That is, based on the transformed (neighbor) data rather than the original data, NME estimates continuous causality or functional dependency from cause variable to effect one, which is able to infer directed GRN for both steady-state and time-series data. In other words, the consistent continuity of a dependent

function from one variable to another across all measured samples implies their causal relation, which is quantified by their NME with varying neighbor size. Here, the consistent continuity of a function means that the functional relation or NME (cross-map mutual information) from one variable to another is persistent from a small (local) neighbor size to a large (global) neighbor size.

Specifically, suppose that there are two random variables X and Y with n original data/samples, i.e.

$$X = x_i \text{ and } Y = y_i, i = 1, \dots, n, \quad (1)$$

which will be used to infer the causal relation from X to Y . By the same definition of Equation (1), we can select the top p nearest neighbors of x_i and y_i , respectively. We define x_i^{NN} as the top p nearest neighbors of x_i , $x_i^{NN} = (x_{i_1}, \dots, x_{i_p})$ and y_i^{NN} as the top p nearest neighbors of y_i , $y_i^{NN} = (y_{i_1}, \dots, y_{i_p})$. x_i^{NN} and y_i^{NN} are p -dimension vectors and $i_1, i_2, \dots, i_p, i'_1, i'_2, \dots, i'_p \in \{1, 2, \dots, n\}$.

$$x_i^{NN} = (x_{i_1}, \dots, x_{i_p}), y_i^{NN} = (y_{i'_1}, \dots, y_{i'_p}) \quad (2)$$

where x_{i_j} or $y_{i'_j}$ is the j -th nearest neighbor/sample of x_i or y_i among all n samples, respectively, in terms of the absolute value or Euclidian distance (Figure 1A).

For each sample i , we can obtain the nearest neighbors of x_i and y_i , $x_i^{NN} = (x_{i_1}, \dots, x_{i_p})$ and $y_i^{NN} = (y_{i'_1}, \dots, y_{i'_p})$. Also, we defined $X^{NN} = (x_1^{NN}, x_2^{NN}, \dots, x_n^{NN})'$ and $Y^{NN} = (y_1^{NN}, y_2^{NN}, \dots, y_n^{NN})'$, where x_i^{NN} or y_i^{NN} is a random variable/vector with p elements. Thus, we can quantify the causality or continuity of $Y=f(X)$ from X to Y by designing a new measure or mutual information (MI) called neighbor cross-mapping entropy (NME) or Equation (3):

$$\begin{aligned} NME(X, Y) &= NME_{X \rightarrow Y} = MI(Y(X^{NN}); Y^{NN}) \\ &= \sum_{i=1}^n P(Y(x_i^{NN}), y_i^{NN}) \ln \frac{P(Y(x_i^{NN}), y_i^{NN})}{P(Y(x_i^{NN}))P(y_i^{NN})} \end{aligned} \quad (3)$$

where $P(\cdot)$ is the probability distribution of the corresponding random variable/vector. To simplify the notation, we denote $Y(x_i^{NN})$ as the corresponding p points of x_i^{NN} in Y , or the cross-maps of x_i^{NN} to Y , whereas y_i^{NN} is the p nearest neighbors of y_i .

In this work, we define continuous causality based on the continuity of a function by extending the dynamical causality in dynamical systems to steady data [13, 14].

Definition 1. $X=x_i$ and $Y=y_i$ are two random variables. If there exists a continuous function f such that $y_i^{NN} = f(x_i^{NN})$ holds for all points $i=1, \dots, n$ with varying neighbor sizes $p=1, \dots, n$, then there is continuous causality from X to Y .

Considering the continuity of a function f at any observed point (x^*, y^*) from X to Y , i.e. $Y=f(X)$, we interpret the continuous causality stipulated by the following $\epsilon - \delta$ definition. If for every $\epsilon > 0$ there exists a δ such that $|f(x) - f(x^*)| < \epsilon$ holds for any x satisfying $|x - x^*| < \delta$, then f is continuous at point (x^*, y^*) from X to Y . In our NME method with NN as ϵ and δ , we use neighbor cross-mapping to approximately quantify the continuity. Clearly, such a causality is characterized by a functional continuity from X to Y across all samples, perturbed by neighbor size. Actually, f is a function series in Definition 1 as the dimension of variables (x_i^{NN}, y_i^{NN}) changes with p .

Definition 2. X and Y are two random variables. Then, the continuous causality strength from variables X to Y is defined as $NME(X, Y) = NME_{X \rightarrow Y} = MI(Y(X^{NN}); Y^{NN})$ of Equation (3), and the normalized strength is defined as $NME_{X \rightarrow Y}/H(Y^{NN}) = MI(Y(X^{NN}); Y^{NN})/H(Y^{NN})$, where $H(Y^{NN})$ is the entropy

$$H(Y^{NN}) = -\sum_{i=1}^n P(y_i^{NN}) \ln P(y_i^{NN}) \quad (4)$$

Clearly, $NME(X, Y) = NME_{X \rightarrow Y}$ is asymmetric, that is, $NME_{X \rightarrow Y} \neq NME_{Y \rightarrow X}$ or $MI(Y(X^{NN}); Y^{NN}) \neq MI(X(Y^{NN}); X^{NN})$. $NME_{X \rightarrow Y}$ or

$NME(X, Y)$ represents the strength of X 's influence on Y (causality from X to Y) and $NME_{Y \rightarrow X}$ or $NME(Y, X)$ represents the strength of Y 's influence on X . So we can determine whether NME have symmetry by comparing whether $NME(X, Y)$ are equal to $NME(Y, X)$. From Equation (3), we can obtain that $NME(X, Y) = MI(Y(X^{NN}); Y^{NN})$ and $NME(Y, X) = MI(X(Y^{NN}); X^{NN})$

$$NME(X, Y) - NME(Y, X)$$

$$= MI(Y(X^{NN}); Y^{NN}) - MI(X(Y^{NN}); X^{NN})$$

So, the NME statistics are asymmetrical.

We can also show the following relation:

$$MI(Y(X^{NN}); Y^{NN}) = H(Y^{NN}) - H(Y^{NN}|Y(X^{NN})) \quad (5)$$

Thus, the continuous causality criteria can be proven⁷³ from Equations (4) and (5).

- (i) If $NME_{X \rightarrow Y} = MI(Y(X^{NN}); Y^{NN}) = H(Y^{NN})$ or $H(Y^{NN}|Y(X^{NN})) = 0$, then there is a continuous function $y_i^{NN} = f(x_i^{NN})$ or continuous causality from X to Y ;
- (ii) If $NME_{X \rightarrow Y} = MI(Y(X^{NN}); Y^{NN}) = 0$ or $H(Y^{NN}|Y(X^{NN})) = H(Y^{NN})$, then x_i^{NN} is independent of y_i^{NN} or there is no continuous causality from X to Y ;
- (iii) If $y_i^{NN} = f(x_i^{NN})$, then $H(Y(X^{NN})) = H(X^{NN}) \geq H(Y^{NN})$.

Hence, for noisy data, those results imply $NME_{X \rightarrow Y} \approx H(Y^{NN})$ is high for causal relation while $NME_{X \rightarrow Y} \approx 0$ is low for non-causal relation. Based on those results, we can use $NME_{X \rightarrow Y}$ of Equation (3) to infer causal relationship from X to Y .

Direct causality and cNME framework

Considering the complexity in real networks, there are many variables, which implies that there may exist indirect causality due to the transitivity of causality (Figure 1B and Supplementary Figure S3). To quantify direct causality between variables, we further propose the cNME. Here, we define the direct continuous causality from X to Y in a third variable/vector $Z = z_i$ and its p nearest neighbors of z_i as $Z^{NN} = (z_1^{NN}, z_2^{NN}, \dots, z_n^{NN})$.

Definition 3. X and Y are two variables and Z is the third variable/vector. Then, the direct continuous causality strength from variables X to Y in conditional variable/vector Z is defined as

$$\begin{aligned} cNME_{X \rightarrow Y|Z} &= MI(Y^{NN}; Y(X^{NN})) \\ &\quad - \frac{MI(Z^{NN}; Z(X^{NN}))}{H(Z^{NN})} * MI(Y^{NN}, Y(Z^{NN})) \end{aligned} \quad (6)$$

where X^{NN} , Y^{NN} and Z^{NN} represent the p nearest neighbors of X , Y and Z ; $Y(X^{NN})$ is corresponding points of X^{NN} in Y ; $Z(X^{NN})$ is corresponding points of X^{NN} in Z ; $Y(Z^{NN})$ is corresponding points of Z^{NN} in Y .

By Equation (6), we can remove the influence of Z to obtain the direct causality from X to Y . Actually, the continuous causality in this work is functional dependency from one variable to another, which is measured by NME not from the original data but the transformed (neighbor vector) data, thus ensured mathematically by the function continuity and scaling law [13, 14].

Numerical estimation of NME and cNME

NEM and cNME are calculated based on mutual information of multi-variables, which can be numerically estimated from data. For given variables X and Y , the mutual information can be described as follows:

$$MI(X; Y) = H(X) + H(Y) - H(X, Y)$$

where $H(X)$ is entropy of X ; $H(Y)$ is entropy of Y ; $H(X, Y)$ is joint entropy of X and Y .

Many algorithms [15–21] have been developed to estimate the entropy from data. And kNN estimation method [19] is a suitable nonparametric approach for entropy estimation of high dimensional data. Mutual information $MI(X; Y)$ from Equation (6) can be estimated based on kNN estimation. Equations (7)–(11) describe the details of the kNN method [19].

Suppose that x_1, x_2, \dots, x_n are assumed to be iid (independent identically distributed) realizations of a random variable X , the k -th nearest neighbor of x_i is x_{ik} and $\epsilon(i)/2 = |x_i - x_{ik}|$ represents distance from x_i to its k -th neighbor x_{ik} . $H(X)$ can be estimated as

$$\hat{H}(X) = -\psi(k) + \psi(n) + \frac{d}{n} \sum_{i=1}^n \ln \epsilon(i) \quad (7)$$

where $\epsilon(i)$ is twice the distance from x_i to its k -th nearest neighbor; n is the number of samples; d is the dimension of X .

We further calculate estimation of $H(X, Y)$. Suppose that $z_i = (x_i, y_i)$, $i = 1, 2, \dots, n$ are assumed to be iid (independent identically distributed) realizations of a random variable $Z = (X, Y)$, the k -th nearest neighbor of z_i is z_{ik} and $\epsilon(i)/2 = |z_i - z_{ik}|$ represents distance from z_i to its k -th neighbor z_{ik} . The estimation of $H(X, Y)$ can be described as follows based on kNN estimation:

$$\hat{H}(X, Y) = -\psi(k) + \psi(n) + \frac{d_X + d_Y}{n} \sum_{i=1}^n \ln \epsilon(i) \quad (8)$$

where $\epsilon(i)$ is twice the distance from (x_i, y_i) to its k -th nearest neighbor; n is the number of samples; d_X is the dimension of X ; d_Y is the dimension of Y .

We denote by $\epsilon_x(i)/2$ and $\epsilon_y(i)/2$ the distances between the same points projected into the X and Y subspaces. $n_x(i)$ is the number of points within the vertical $x = x_i \pm \epsilon(i)/2$ then $\epsilon(i)/2$ is the distance to the $(n_x(i)+1)$ -st neighbor of x_i . So based on Equation (7), we have

$$\hat{H}(X) = -\langle \psi(n_x + 1) \rangle + \psi(n) + \frac{d_X}{n} \sum_{i=1}^n \log \epsilon(i) \quad (9)$$

Similarly, we have

$$\hat{H}(Y) = -\langle \psi(n_y + 1) \rangle + \psi(n) + \frac{d_Y}{n} \sum_{i=1}^n \log \epsilon(i) \quad (10)$$

Thus, the estimation of mutual information is

$$\hat{MI}(X; Y) = \hat{H}(X) + \hat{H}(Y) - \hat{H}(X, Y)$$

$$\begin{aligned} &= -\langle \psi(n_x + 1) \rangle + \psi(n) + \frac{d_X}{n} \sum_{i=1}^n \log \epsilon(i) \\ &\quad - \langle \psi(n_y + 1) \rangle + \psi(n) + \frac{d_Y}{n} \sum_{i=1}^n \log \epsilon(i) \\ &\quad - \psi(k) + \psi(n) + \frac{d_X + d_Y}{N} \sum_{i=1}^N \ln \epsilon(i) \\ &= \psi(k) - \langle \psi(n_x + 1) + \psi(n_y + 1) \rangle + \psi(n) \end{aligned} \quad (11)$$

where $\langle \cdot \rangle$ represents mean; n is the number of samples. Thus, we can also obtain the estimation of NME and cNME through computing $\hat{MI}(Y(X^{NN}); Y^{NN})$, $\hat{MI}(Z^{NN}; Z(X^{NN}))$, $\hat{H}(Z^{NN})$ and $\hat{MI}(Y^{NN}, Y(Z^{NN}))$.

scNME framework

We proposed scNME framework specifically for scRNA-seq data based on NME/cNME (Figure 1C), which consists of three main steps:

- (i) Inferring gene regulatory networks (GRNs) using NME or cNME framework;
- (ii) Defining the activity of GRNs in each cell using network entropy;
- (iii) Clustering and dimensional reduction based on GRN activity matrix.

In this scNME framework, we first infer GRN using cNME (or NME). We consider TF datasets, e.g. TRRUST (<http://www.grnpedia.org/trrust/>) and hTFtarget (<http://bioinfo.life.hust.edu.cn/hTFtarget#!/>). Using cNME or NME, we can infer the GRN of scRNA-seq data which consist of TF GRNs. For threshold selection of each TF's GRN, we approximatively used the mean and variance of cNME values from the given TF toward all target genes to replace the mean and variance in Supplementary Equation (4) and (5). Second, we designed a new method, network entropy (NE), which can be used to calculate the activity of TF GRNs for each cell in scRNA-seq. TF GRNs consist of TFs and their target genes. The regulatory probability for TF_i and target gene j is $A_{ij} = 0$ or 1 . NE of a GRN of TF_i can be described as follows:

$$NE_i = - \sum_j x_i p_{ij} \log p_{ij} \quad (12)$$

where p_{ij} is the normalized regulatory probability distribution of TF_i toward gene j , defined as $p_{ij} = \frac{A_{ij}}{\sum_j A_{ij}}$; $x_i = (x_{i1}, x_{i2}, \dots, x_{in})$ is the TF_i expression in all cells; n is the number of cells.

The GRN activity matrix can be obtained after computing NE of all TFs. The GRN activity matrix reflects the activity of each GRN in all cells and unravels relevant cell states. Finally, the GRN activity matrix can be used to identify cell states and cell state-specific GRNs.

Benchmark datasets

We simulated five causal scenarios between three variables, sequential, fan-in, fan-out, cascade and loop, which can represent all one-way causal scenarios among the three variables (Figure 2A, Supplementary Figure S4).

DREAM3 and DREAM4 *Insilico* Network Challenges data, which contain both steady-state and time-series data with different gene scales, were also applied for benchmarking. The real-world dataset SOS DNA repair network which is a small-scale and

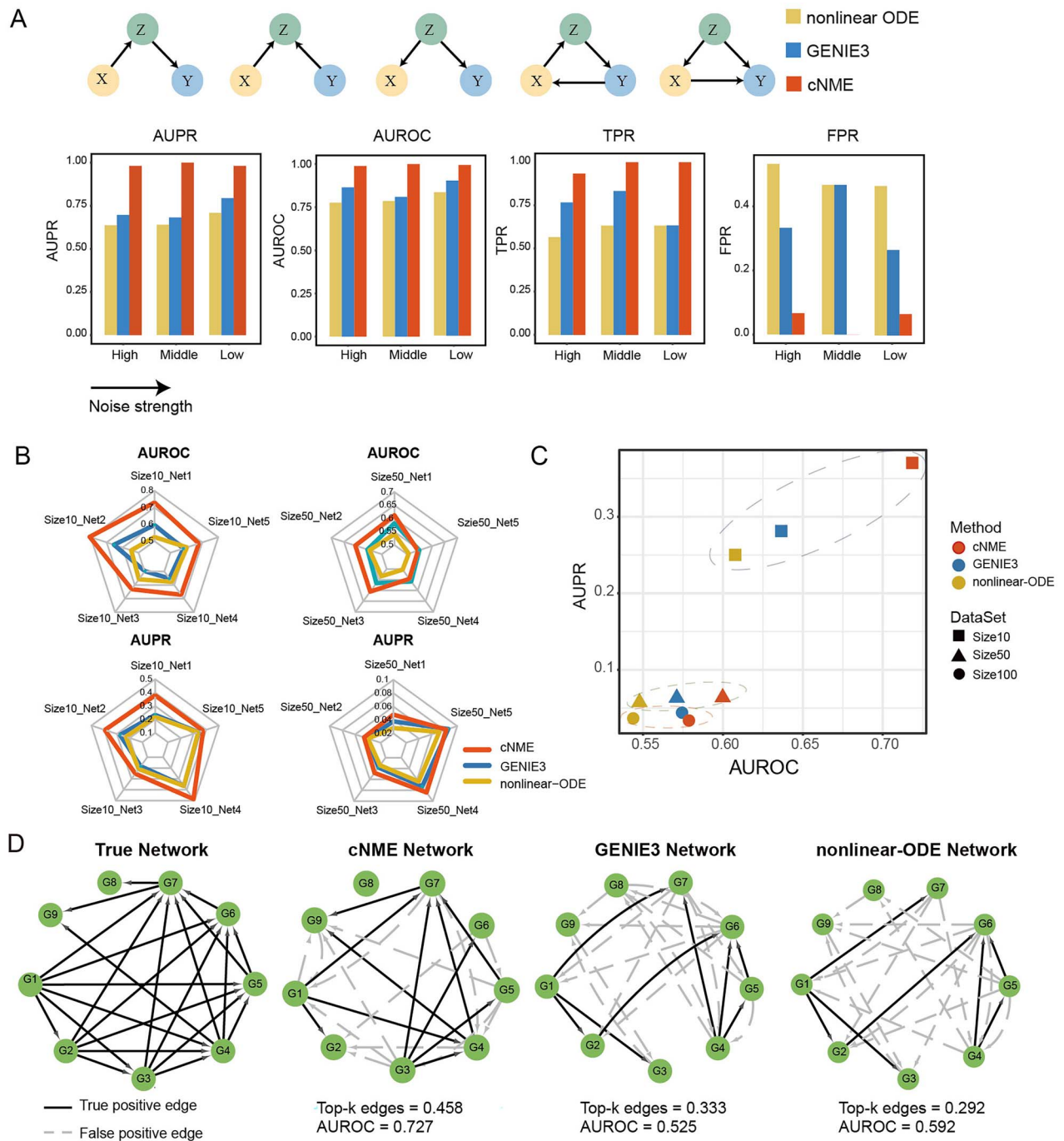


Figure 2. The comparison performances of cNME on simulation and benchmark datasets. **(A)** Multivariable network structure for simulation and evaluation metrics scores of different methods with increasing noise strength. **(B)** Radar chart showing the performances of different methods on DREAM3 dataset. Left column shows the AUROC and AUPR values of in silico size_10 timeseries data, where right column represents in silico size_50 steady data. **(C)** Average performances of DREAM3 and DREAM4 dataset with different sizes. **(D)** Top n predicted edges and corresponding AUROC values of different methods. The black solid line represents true positive edges, and the grey dotted line represents false positive predictions.

directed network composed of 9 genes and 24 regulations [22, 23] was also used for comparison.

We selected AUROC, AUPR and top-k edges to evaluate the performance on the simulated data and benchmark datasets. AUROC and AUPR are common metrics used in classification and prediction models, which select different thresholds to comprehensively assess the experimental results. The *sklearn.metrics* python package was used to calculate the area under ROC and PR curve. In the simulated data, we set the average weight of all edges as cutoff for each network and computed its TPR and FPR values.

In DNA SOS repair network data, we plotted the top-k predicted edges for each method, where k is the number of true edges in the ground-truth network.

cNME analyses between microbes and skin parameters

For all skin parameters, we set $p = 15$, $k = 3$. We considered a significant causal relationship between microbes and skin parameters as the P-value of cNME less than 0.1.

Analysis of Chu-time dataset

In Chu-time dataset, genes whose average expression less than 0.5 were removed (11 077 genes remained) from following analysis. We downloaded well-known human TF genes from TRRUST database and obtained 529 TFs in this dataset. The weight of potential regulations was inferred by cNME using gene expression matrix with parameters ($nz=2, p=15, k=5$). We obtain the weight matrix with a P-value cutoff set as 0.05 and calculated GRN activity for each cell. The NE matrix was used for downstream clustering and differential GRN activity analysis.

Joint analysis of human and mouse ESC dataset

The mESC dataset was from mESC cell lines to NPC differentiation. Cells with fewer than 10 genes and genes expressed in fewer than 3 cells were removed. For hESC dataset, we selected those cells corresponding to mESC cell types. Then, mouse gene symbols were transformed to homologous genes of human by *homologene* packages. In total, 1006 cells and 6735 genes remained in further analysis. The regulatory relationships of 388 TFs among them were inferred by cNME analysis with parameters set as below ($nz=1, p=20, k=10$). GRN activity matrix was calculated with P-value cutoff of 0.1. Hierarchical clustering with Spearman's distance was conducted based on the GRN activity matrix.

Comparison of clustering methods

We compared scNME performance on different clustering methods, such as *pcaReduce*, SC3, Leiden, Louvain, SLM, SIMLR and SNN-Cliq. For each clustering method, we used the default parameters except that the number of clusters was set to 6 corresponding to true labels. Methods involving random steps were repeated 20 times, and the average ARI was plotted.

Target enrichment

For each TF gene, cNME inferred GRN with different weights of targets. The targets were further submitted to Metascape for enrichment analysis with default parameters, and we only focused on GO Biological Progress Pathways.

Calculating differential GRN

We used Wilcoxon rank-sum test to identify differentially activated GRNs for each group with a minimum logfc of 1 and minimum difference fraction of 0.1 (Supplementary Table S1).

Analysis of fibroblast dataset

To demonstrate the application of scNME in identifying cell state and cell state-specific GRNs, we performed reanalysis on fibroblasts from human lung cancer dataset [24] based on the GRN activity matrix. The dataset consists of cells from human NSCLC samples. These patients were from three stages, before initiating systemic targeted therapy (TN), at the residual disease (PR) state, and the tumors showed acquired drug resistance (PD). We performed clustering on 2991 fibroblasts using Seurat. We further applied scNME on fibroblasts to obtain GRN activity matrix and identify the cell state. To reduce the computational complexity of scNME, we selected 6000 highly variable genes in construction of GRN. For each TF's GRN, we applied 0.1 as significance level to identify causal relationship. Based on the GRN activity matrix, we identify eight fibroblast subtypes (resolution = 0.3).

To investigate the association of fibroblasts' cell state with drug resistance in tumors, we further performed clustering in tumor cells [24]. Cellchat [25] was applied to identify the cell-cell

communication between fibroblasts and tumor cells. GEPIA2 [26] was further used to perform survival analysis.

RESULTS

Overview of causality inference algorithm

Information-based method is one of the most popular methods in inferring GRN due to low computational complexity and sample (size) demand. However, the inferred GRN by such a method, e.g. mutual information or correlation-based method, is undirected or bidirectional not causal regulations. Here, instead of the original data, we proposed a novel information-theoretic method for the transformed (neighbor) data based on 'continuous causality' concept, i.e. NME, with varying neighbor sizes to estimate causality or functional dependency between variables, which is able to infer directed GRN for both steady-state and time-series data, thus overcoming such a limitation (Figure 1A, Methods and Supplementary Figures S1, S2).

Given that direct causality in a molecular network is vital for understanding regulatory mechanisms and biological functions, such as gene regulations, signaling processes and metabolic pathways, we further extend NME to cNME to identify direct causality between variables (Figure 1B). cNME can identify direct causality between variables effectively through eliminating indirect causality. We also developed scNME framework (Figure 1C) by NME to construct GRN specifically from scRNA-seq data, evaluate GRNs' activity for each cell and identify stable cell states.

Performance on benchmark datasets

We benchmarked cNME method against two well-known network reconstruction models (GENIE3 [11] and nonlinear-ODE [27]) using simulated toy networks. The average performance of different methods on the stimulated data was illustrated (Figure 2A). cNME performed better than other methods and maintained robustness with increasing noise levels. Furthermore, we observed that both GENIE3 and nonlinear-ODE failed to discern the true direction of one-way causality in the first three simulated networks (Figure 2A, Supplementary Figure S5), which may be caused by nonlinearity. For more complex models, where direct and indirect causality co-exist, GENIE3 failed to recognize the true direct link from z to y , which may be regressed out by x . Due to the interference of the variable z on both x and y , it was difficult for nonlinear-ODE to discern the true direction between x and y . It was also difficult for nonlinear-ODE to deal with the case containing loop structures. Again, only the cNME method can correctly predict both causal networks.

DREAM datasets are the most popular benchmark datasets used in GRN reconstruction algorithms [28–30]. We also investigated the performance of these methods in different sample sizes and data types of DREAM dataset by common evaluation metrics. cNME obtained higher AUROC and AUPR values than other methods on all five networks at size_10 and size_50 datasets (Figure 2B). A comprehensive comparison of these methods at different gene scales was conducted on DREAM3 and DREAM4 data (Figure 2C, Supplementary Table S2). With the increase of network sizes, cNME could maintain a relatively better average performance, which indicates the robustness and superiority of cNME for gene regulation inference compared with other methods.

Real gene expression dataset was also considered to test the effectiveness of algorithms. SOS DNA repair network is a small-scale and directed network composed of 9 genes and 24 regulations [22, 23]. The top 24 predictions of each method were

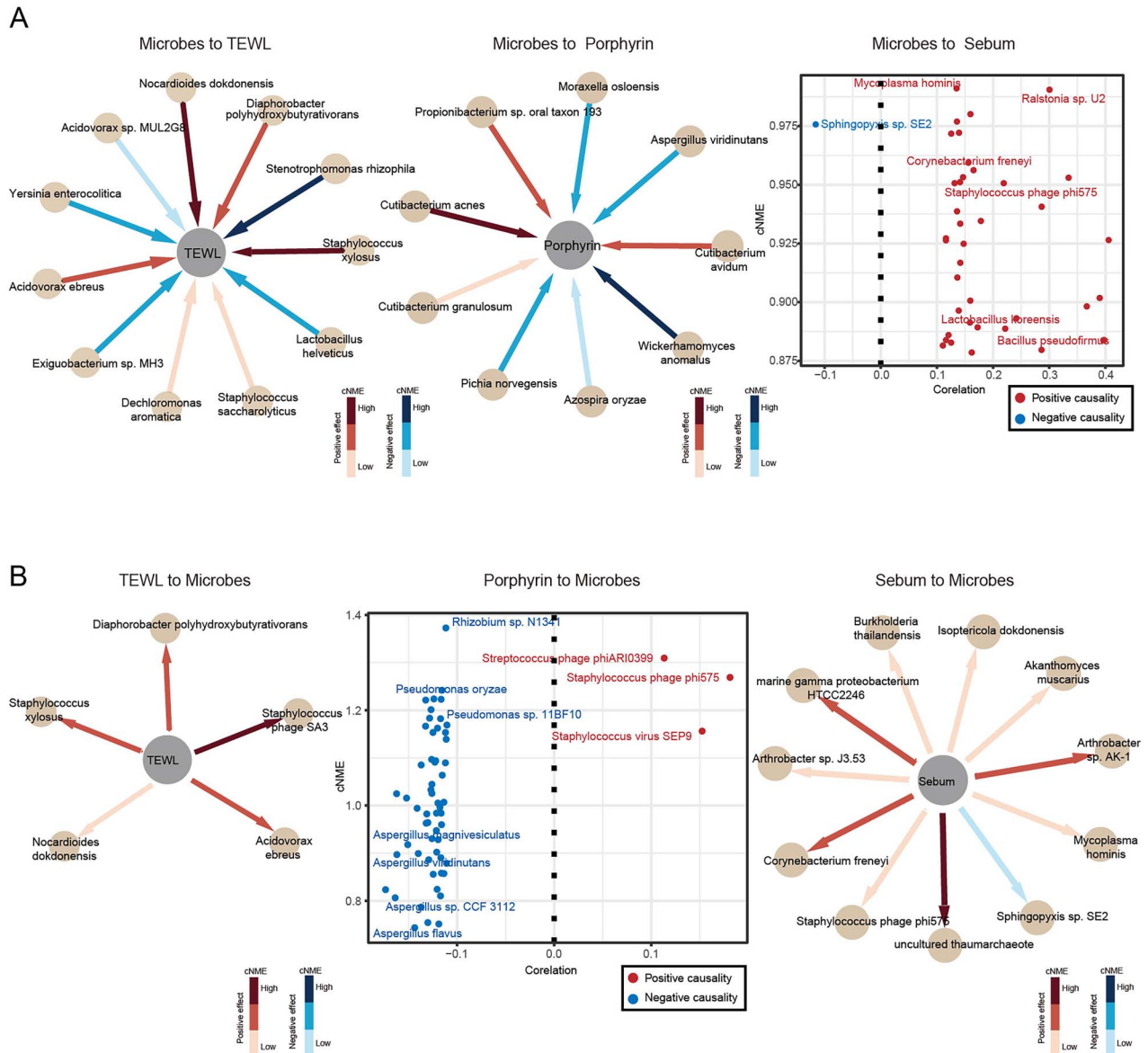


Figure 3. cNME analyses between microbes and skin parameters. (A) Results of cNME analyses of microbes to phenotypes, including TEWL(left), Porphyrin(middle) and Sebum (right). (B) Results of cNME analyses of phenotypes (TEWL, Porphyrin and Sebum) to microbes. The causal effects from low to high are indicated by the color shade of arrows.

extracted. GENIE3 and nonlinear-ODE identified 8 and 6 true positive regulatory links, respectively. It is obvious that cNME assigned higher weights to edges in the ground-truth network compared with GENIE3 and nonlinear-ODE (Figure 2D), while the other two methods tended to obtain more indirect links, such as gray edges to G8 and they had difficulty recognizing the true positive cascade pattern between triples. In conclusion, cNME could identify the nonlinear relationships and be more sensitive to directional edges compared to these methods.

Validation of cNME using metagenomic data

The skin microbiome plays important roles in maintaining skin homeostasis and microbial dysbiosis, which is associated with the development and progression of many common skin diseases [31–33]. Here, we applied cNME to infer causal relationships between microbial species levels and skin physiological indicators (Figure 3 and Supplementary Figure S6). Positive or negative causality

is judged by Pearson correlation coefficients. Our results show that *Staphylococcus* spp. has strong positive effects on TEWL, and *Staphylococcus xylosum* is particularly prominent, which is consistent with a previous study [34]. In addition, *Cutibacterium acnes* and *Propionibacterium* species were found to be porphyrin-producing strains [35, 36]. Also, they actually show high cNME scores in causality from microbes to porphyrin. Besides, *Corynebacterium freneyi* has a high positive causal relationship to the sebum level, which was also reported for releasing free fatty acids [37].

We further investigated causality from those skin statuses to microbes based on cNME (Figure 3B) since microorganisms are easily influenced by the host environment. *Staphylococcus xylosum* and TEWL showed bidirectional positive causality. Porphyrin may have inhibitory effects on *Aspergillus* species and *Pseudomonas* species. Sebum showed the ability to promote the growth of *C. freneyi*. These results suggest that most microbes can decrease the hydration of skin, which can be further investigated.

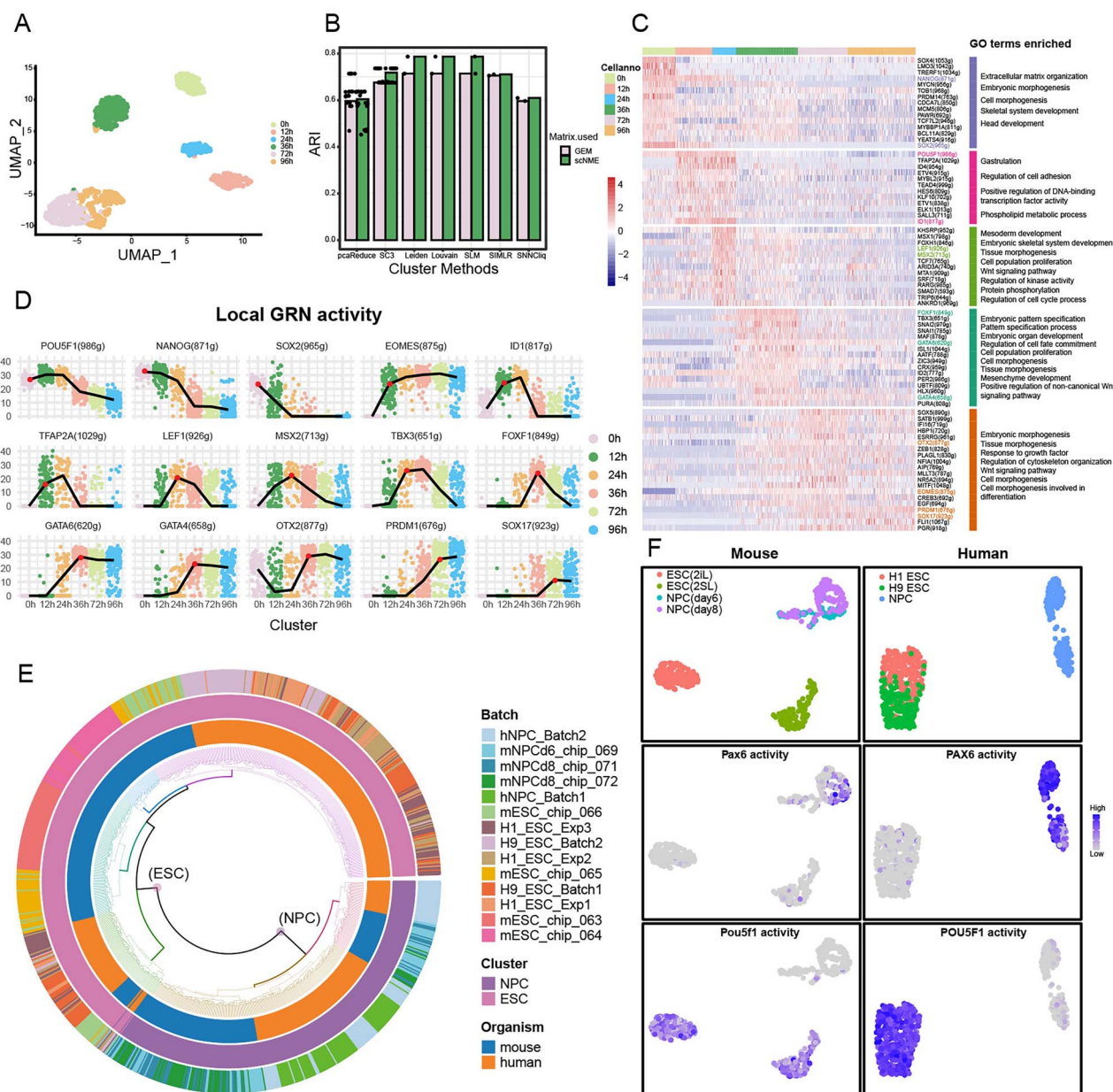


Figure 4. scNME results on human embryo dataset and cross-species integration. (A) UMAP plot based on the GRN activity matrix. Cells are colored by their differentiation time. (B) Accuracy of different clustering methods on expression matrix and scNME activity matrix. The random methods are repeated several times. (C) Heatmap of scNME activity matrix for chu-tomo dataset. Differential GRNs are shown in rows, and master TFs corresponding to each cell state are highlighted. GRNs for master TFs are enriched and the associated GO terms are listed, respectively. (D) GRN activity over time based on scNME activity matrix. The black line indicates mean activity score at each time point and the highest mean activity scores are marked by red dots. (E) Circus plot showing the joint clustering of human and mouse embryo dataset based on GRN activity matrix. (F) The activity of master TFs of ESC development across two datasets.

scNME identifies regulators for human embryonic stem cell differentiation

In order to evaluate the performance of scNME on scRNA-seq data, we applied scNME to a human embryonic stem cell dataset [38], which contains cells developing from human embryonic stem (ES) cells to induced endoderm (DE) cells at 0, 12, 24, 36, 72 and 96 h of differentiation [39]. Visualization based on GRN activity matrix showed the distribution of cells (Figure 4A). Overlapping domains between cells from 72 to 96 h of differentiation indicated a similar stable state of transcriptome. To determine whether GRN activity matrix is beneficial for different clustering methods, we compared the performance on both GRN activity

matrix and gene expression matrix (GEM) using seven clustering methods (Figure 4B). Overall, scNME obtained a higher score of adjusted Rand index (ARI), especially on community-based clustering methods, which demonstrated that scNME based on GRN activity could identify different cell states more accurately than GEM.

We further characterized GRNs associated with stage of ES cell differentiation by performing differential GRN analysis. At the beginning of differentiation of human ES cells, the GRN of pluripotency transcription factor (TF) NANOG, POU5F1 and SOX2 maintained high activity, which corresponded to characteristics of stem cells (Figure 4C, Supplementary Table S1). At 12 h of

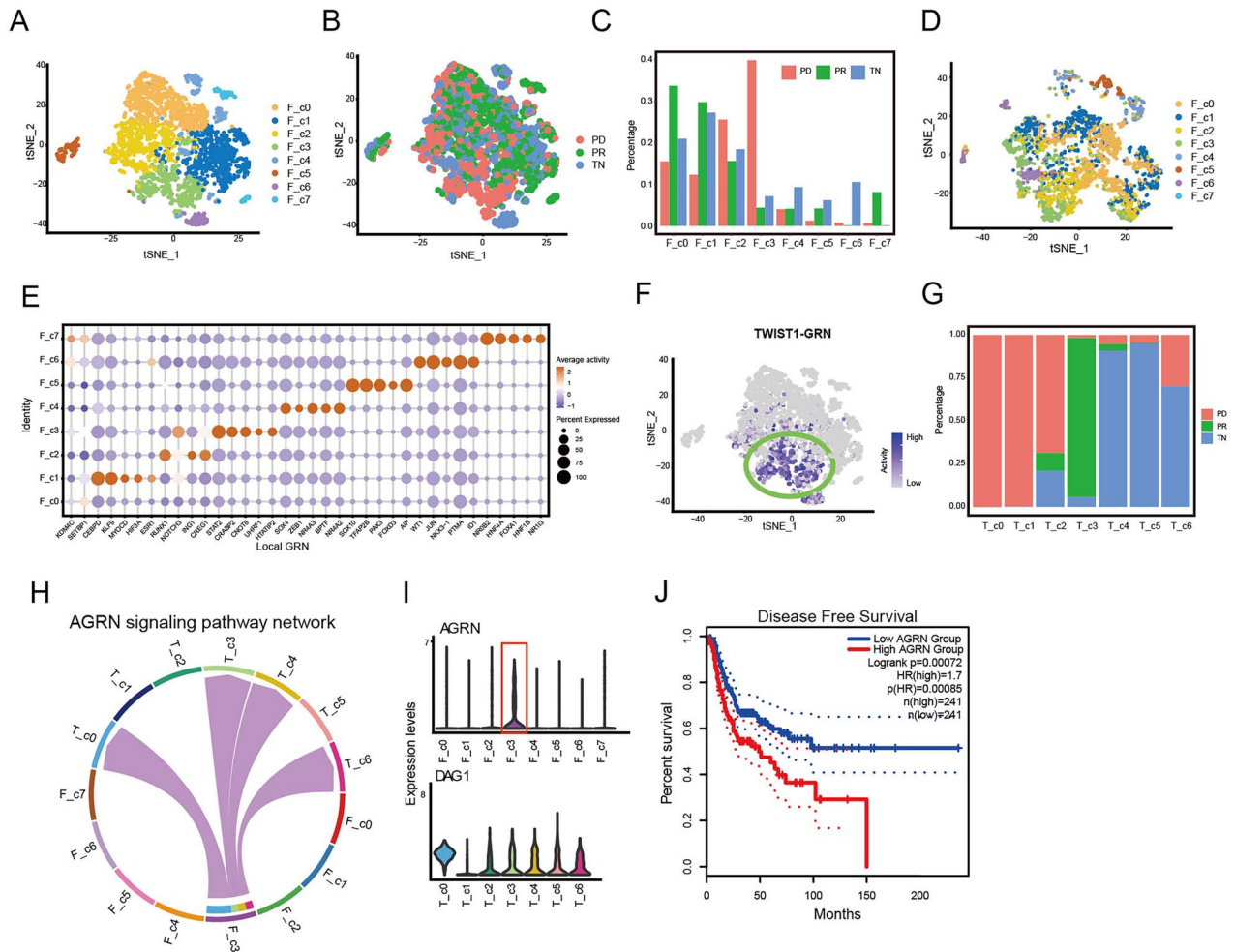


Figure 5. scNME-guided analysis identifies a novel cancer-associated fibroblast related with drug resistance. t-SNE plot of fibroblasts based on GRN activity matrix, and clusters were obtained using GRN activity matrix. Different colors represent different cell types (A), and states of patients (B). (C) The distribution of fibroblast cell types in three patients' states. (D) t-SNE plot of fibroblasts based on gene expression matrix. Colors represent cell types from (A). (E) Activity of cell-state specific GRNs of 8 fibroblast clusters. (F) t-SNE plot shows the activity levels of TWIST1-GRN. (G) The distribution of tumor cells in three patients' states. (H) Circle plot showing interaction strength of AGRN signaling pathway network from F_c3 to all tumor cell types. (I) Violin plots showing the RNA expression levels of ligand (AGRN) in fibroblast cell types and receptor (DAG1) in tumor cell types. (J) The Kaplan-Meier curve showing that patients with high expression of AGRN had significantly worse overall survival based on TCGA LUAD and LUSD datasets.

differentiation, a second wave of GRNs exhibited high levels of activity, such as ID1 and EOMES whose target genes are enriched at gastrulation pathway (Figure 4D, Supplementary Figure S7), indicating a transition of embryo stem cells toward three primary germ layers. The GRN activities of MSX2, LEF1 and TBX3 were upregulated at 24 h of differentiation. Targets of these GRNs were enriched at mesoderm development, cell population proliferation etc., which suggested that cells were in proliferation and morphogenesis. At 36 h of differentiation, the GRN activities of early DE-specific TFs, such as GATA4, GATA6 and FOXF1, were upregulated. Enrichment analysis focused on embryonic pattern specification and regulation of cell fate commitment. At 72 h of differentiation, most cells activated GRNs of DE markers like OTX2, EOMES, PRDM1 and SOX17, which remained at high levels among cells at 96 h. It marked that the cells have differentiated into mature DE cells. In scNME analysis, we identified the key regulation point around 12 h due to high GRN activity of gastrulation, which was revealed as a tipping point by dynamic network biomarker analysis [39]. Besides, it was noticed that GRN of TFAP2A was activated at 12 h, which was reported to encode a transcription factor binding to consensus sequence and involve in

neural crests and epidermis development [40, 41]. Its high activity of GRN suggested that TFAP2A might play a significant role at early embryonic development.

We further applied scNME to scRNA-seq data during ES cell differentiation into neural progenitor cells in both human [38] and mouse [42] to investigate cross-species GRNs. scNME could effectively group cells by cell types, regardless of their experiment batch and species (Figure 4E, Supplementary Figure S8). The GRN activity of master TFs was examined on two datasets (Figure 4F). After integration, ESC-derived cells have comparable activity level of TF associated with pluripotency (POU5F1) and neural development (PAX6).

scNME identified a novel cancer-associated fibroblast related with drug resistance

The tumor microenvironment (TME) plays a vital role in the occurrence, development, metastasis and therapeutic resistance of tumor [43–47]. Cancer-associated fibroblasts (CAFs) are the main player in the TME and participated in tumor cell proliferation, treatment resistance and metastases [48, 49]. The diversity of CAFs has led to exploiting CAFs to improve personalized

cancer treatment. However, identification of CAF subpopulation associated with tumor development or drug resistance is also a huge challenge due to characterization of CAF heterogeneity. Here, we applied scNME to identify novel CAF subtypes in scRNA-seq dataset of metastatic lung cancer, which were obtained from patients before and during targeted therapy [24]. These patients are divided into three states, before targeted therapy (TN), at the residual disease state (RD), showing the acquired drug resistance (PD). In order to ensure that each cluster has heterogeneity without over-clustering, we determined the number of clusters based on biological interpretation and dimensionality reduction analysis. Also, eight subtypes of fibroblast were identified based on the GRN activity matrix (Figure 5A and B). We observed cells from F_c3 highly enriched in PD state, suggesting that F_c3 may be a CAF subtype associated with drug resistance (Figure 5C and D). Gene expression-based clustering result shows that cells from F_c3 did not fall into a group (Figure 5B). The F_c3 shows high activity of TWIST1-GRN (Figure 5E and F). Previous studies reported that Twist1 was a vital inducer of epithelial-mesenchymal transition and associated with poor prognosis in a variety of epithelial cancer cells [50–52]. Several publications have reported that Twist1 was implicated in drug resistance in various cancers [53–59].

Previous reports have demonstrated that CAFs communicated with tumor cells by cytokines, chemokines, growth factors and exosomes to inhibit immune cell function and promote tumor development [60–62]. However, it is still unclear for the detailed interaction mechanisms of fibroblasts and tumor cells in PD state of lung cancer. We observed that cells from T_c0 were all derived from PD patients who have shown drug resistance (Figure 5G and Supplementary Figure S9A and B). To dissect the mechanisms underlying the fibroblasts in PD state, we applied CellChat [25] inference and analysis of ligand–receptor interactions. AGRN and DAG1 in AGRN signaling pathway network were identified as major sources of signaling ligands involved in fibroblasts and tumor interaction in PD stage (Figure 5H and I). As previously reported, AGRN (AGRN) is a component of extracellular matrix and is involved in tumorigenesis [63, 64]. AGRN was also reported to exert a vital effect on resistance to radio-chemotherapy [65]. We also noted that patients with high expression of AGRN had significantly worse overall survival (Figure 5J) based on lung adenocarcinoma cohort dataset and lung squamous carcinoma cohort dataset. The clinical implication demonstrated that AGRN is a key factor in the interaction between cancer cells and CAFs in drug resistance (PD state). Thus, targeting F_C3 may be translational advances and potential therapeutic for lung cancer treatment in drug resistance.

DISCUSSION

In our study, we present an information-theoretic causal network inference method that can be applied to infer gene causal network and analyze single-cell data. Based on the continuous causality, we first proposed a novel estimation method, i.e. NME designed by neighbor cross-mapping, which overcomes the limitation of information-based method. NME quantifies continuous causality from one variable to another based on their function continuity measured across samples and neighbor sizes, i.e. a rigorous mathematical framework (continuity scaling law) [13, 14], which is also logically consistent with natural interpretation as functional dependency. NME can estimate direct causality of variables in an accurate and robust manner, which achieves better performance compared with other methods. In addition, we extended scNME for scRNA-seq which can build GRNs and evaluate GRNs' activity,

which further leads to the discovery of cell-state types. Of course, there are still some shortcomings in this paper. We observed that with the increase of network scale, our method has some limitations, such as increased computational complexity and decreased inference accuracy. In addition, the estimation mutual information method also affects the accuracy of NME/cNME. In the future, we hope to improve the accuracy of the method by integrating multiomics and improving the estimation methods.

Theoretical and numerical results demonstrate that NME will have various potential applications in omics data analysis. For instance, for scRNA-seq analysis, we can identify driver cells using the causal relation between cell types, and infer cell–cell communications by combining the pathway information and ligand–receptor datasets.

Key Points

- We propose a novel causal concept of 'continuous causality' with its quantitative criterion, NME, for inferring GRNs from both steady-state data and time-series data.
- NME shows superior performance on benchmark datasets, comparing with existing methods.
- We propose scNME for single-cell data based on NME, which not only reliably infer GRNs for cell types but also identify cell states.

SUPPLEMENTARY DATA

Supplementary data are available online at <https://academic.oup.com/bib/article/24/5/bbad281/7237942>.

ACKNOWLEDGEMENTS

We thank all the members of Chen laboratory for technical assistance.

AUTHOR CONTRIBUTIONS

L.L. and L.C. conceived and designed this work. L.L., R.X., W.C., Q.Z. and P.T. acquired and analyzed the data. L.L. and R.X. wrote the source code. L.L. and R.X. drafted the manuscript. L.C. revised the manuscript.

FUNDING

The National Key R&D Program of China (No. 2022YFA1004800); the Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDB38040400); the National Natural Science Foundation of China (Nos. 32200520, 12131020, 31930022, T2341007 T2350003 and 62132015); the Special Fund for Science and Technology Innovation Strategy of Guangdong Province (Nos. 2021B0909050004 and 2021B0909060002); and JST Moonshot R&D (No. JPMJMS2021).

DATA AVAILABILITY

All related codes are available at <https://github.com/LinLi-0909/NME>. All datasets used are listed in Supplementary Table S3.

REFERENCES

- Fang L, Li Y, Ma L, et al. GRNdb: decoding the gene regulatory networks in diverse human and mouse conditions. *Nucleic Acids Res* 2021;**49**:D97–103.
- Chen X, Li M, Zheng R, et al. D3GRN: a data driven dynamic network construction method to infer gene regulatory networks. *BMC Genomics* 2019;**20**:929.
- Liao J, Huang Y, Wang Q, et al. Gene regulatory network from cranial neural crest cells to osteoblast differentiation and calvarial bone development. *Cell Mol Life Sci* 2022;**79**:158.
- Zhao M, He W, Tang J, et al. A comprehensive overview and critical evaluation of gene regulatory network inference technologies. *Brief Bioinform* 2021;**22**:bbab009.
- Thomas R. Boolean formalization of genetic control circuits. *J Theor Biol* 1973;**42**:563–85.
- Xiao Y. A tutorial on analysis and simulation of Boolean gene regulatory network models. *Curr Genomics* 2009;**10**:511–25.
- Perrin BE, Ralaivola L, Mazurie A, et al. Gene networks inference using dynamic Bayesian networks. *Bioinformatics* 2003;**19**:ii138–48.
- Kim SY, Imoto S, Miyano S. Inferring gene networks from time series microarray data using dynamic Bayesian networks. *Brief Bioinform* 2003;**4**:228–35.
- Sanchez-Castillo M, Blanco D, Tienda-Luna IM, et al. A Bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data. *Bioinformatics* 2018;**34**:964–70.
- Liu F, Zhang SW, Guo WF, et al. Inference of gene regulatory network based on local Bayesian networks. *PLoS Comput Biol* 2016;**12**:e1005024.
- Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring regulatory networks from expression data using tree-based methods. *PLoS One* 2010;**5**:5.
- Moerman T, Aibar Santos S, Bravo Gonzalez-Blas C, et al. GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics* 2019;**35**:2159–61.
- Pearl J. Statistics and causality: separated to reunite-commentary on Bryan Dowd's "separated at birth". *Health Serv Res* 2011;**46**:421–9.
- Ying X, Leng SY, Ma HF, et al. Continuity scaling: a rigorous framework for detecting and quantifying causality accurately. *Research (Wash D C)* 2022;**2022**:9870149.
- Fraser AM, Swinney HL. Independent coordinates for strange attractors from mutual information. *Phys Rev A Gen Phys* 1986;**33**:1134–40.
- Cellucci CJ, Albano AM, Rapp PE. Statistical validation of mutual information calculations: comparison of alternative numerical algorithms. *Phys Rev E Stat Nonlin Soft Matter Phys* 2005;**71**:066208.
- Moon YI, Rajagopalan B, Lall U. Estimation of mutual information using kernel density estimators. *Physical Review E* 1995;**52**:2318–21.
- Victor JD. Binless strategies for estimation of information from neural data. *Phys Rev E Stat Nonlin Soft Matter Phys* 2002;**66**:051903.
- Kraskov A, Stogbauer H, Grassberger P. Estimating mutual information. *Phys Rev E Stat Nonlin Soft Matter Phys* 2004;**69**:066138.
- Van Hulle MM. Edgeworth approximation of multivariate differential entropy. *Neural Comput* 2005;**17**:1903–10.
- Van Hulle MM. Multivariate Edgeworth-based entropy estimation. 2005 *IEEE Workshop on Machine Learning for Signal Processing (MLSP)* 2005, pp. 311–6. IEEE.
- Ronen M, Rosenberg R, Shraiman BI, et al. Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proc Natl Acad Sci U S A* 2002;**99**:10555–60.
- Shen-Orr SS, Milo R, Mangan S, et al. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat Genet* 2002;**31**:64–8.
- Maynard A, McCoach CE, Rotow JK, et al. Therapy-induced evolution of human lung cancer revealed by single-cell RNA sequencing. *Cell* 2020;**182**:1232–1251 e1222.
- Jin S, Guerrero-Juarez CF, Zhang L, et al. Inference and analysis of cell-cell communication using CellChat. *Nat Commun* 2021;**12**:1088.
- Tang Z, Kang B, Li C, et al. GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic Acids Res* 2019;**47**:W556–60.
- Ma B, Fang M, Jiao X. Inference of gene regulatory networks based on nonlinear ordinary differential equations. *Bioinformatics* 2020;**36**:4885–93.
- Stolovitzky G, Prill RJ, Califano A. Lessons from the DREAM2 challenges. *Ann N Y Acad Sci* 2009;**1158**:159–95.
- Stolovitzky G, Monroe D, Califano A. Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference. *Ann N Y Acad Sci* 2007;**1115**:1–22.
- Marbach D, Schaffter T, Mattiussi C, et al. Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *J Comput Biol* 2009;**16**:229–39.
- Fyhrquist N, Muirhead G, Prast-Nielsen S, et al. Microbe-host interplay in atopic dermatitis and psoriasis. *Nat Commun* 2019;**10**:4703.
- Chen YE, Fischbach MA, Belkaid Y. Skin microbiota-host interactions. *Nature* 2018;**553**:427–36.
- Dethlefsen L, McFall-Ngai M, Relman DA. An ecological and evolutionary perspective on human-microbe mutualism and disease. *Nature* 2007;**449**:811–8.
- Kim Y, Lee Y-S, Yang J-Y, et al. The resident pathobiont *Staphylococcus xylosum* in Nfkbiz-deficient skin accelerates spontaneous skin inflammation. *Sci Rep* 2017;**7**:6348.
- Spittaels KJ, van Uytanghe K, Zouboulis CC, et al. Porphyrins produced by acneic *Cutibacterium acnes* strains activate the inflammasome by inducing K(+) leakage. *iScience* 2021;**24**:102575.
- Barnard E, Johnson T, Ngo T, et al. Porphyrin production and regulation in cutaneous *Propionibacteria*. *mSphere* 2020;**5**:e00793–19.
- Bomar L, Brugger SD, Yost BH, et al. *Corynebacterium accolens* releases antipneumococcal free fatty acids from human nostril and skin surface triacylglycerols. *MBio* 2016;**7**:e01725–15.
- Chu LF, 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**:173.
- Lin L, Yilin X, Lili Y, et al. Dynamic network biomarker factors orchestrate cell-fate determination at tipping points during hESC differentiation. *Innovation* 2022;**4**:100364.
- Wenke AK, Bosserhoff AK. Roles of AP-2 transcription factors in the regulation of cartilage and skeletal development. *FEBS J* 2010;**277**:894–902.
- Eckert D, Buhl S, Weber S, et al. The AP-2 family of transcription factors. *Genome Biol* 2005;**6**:246.
- Tuck AC, Natarajan KN, Rice GM, et al. Distinctive features of lincRNA gene expression suggest widespread RNA-independent functions. *Life Sci Alliance* 2018;**1**:e201800124.

43. Vitale I, Manic G, Coussens LM, et al. Macrophages and metabolism in the tumor microenvironment. *Cell Metab* 2019;**30**: 36–50.
44. Meurette O, Mehlen P. Notch signaling in the tumor microenvironment. *Cancer Cell* 2018;**34**:536–48.
45. Vuong L, Kotecha RR, Voss MH, et al. Tumor microenvironment dynamics in clear-cell renal cell carcinoma. *Cancer Discov* 2019;**9**: 1349–57.
46. Zhou Z, Lu ZR. Molecular imaging of the tumor microenvironment. *Adv Drug Deliv Rev* 2017;**113**:24–48.
47. Roswall P, Bocci M, Bartoschek M, et al. Microenvironmental control of breast cancer subtype elicited through paracrine platelet-derived growth factor-CC signaling. *Nat Med* 2018;**24**:463–73.
48. Feng B, Wu J, Shen B, et al. Cancer-associated fibroblasts and resistance to anticancer therapies: status, mechanisms, and countermeasures. *Cancer Cell Int* 2022;**22**:166.
49. Cirri P, Chiarugi P. Cancer-associated-fibroblasts and tumour cells: a diabolic liaison driving cancer progression. *Cancer Metastasis Rev* 2012;**31**:195–208.
50. Ansieau S, Morel AP, Hinkal G, et al. TWISTing an embryonic transcription factor into an oncoprotein. *Oncogene* 2010;**29**: 3173–84.
51. Puisieux A, Valsesia-Wittmann S, Ansieau S. A twist for survival and cancer progression. *Br J Cancer* 2006;**94**:13–7.
52. Yang J, Mani SA, Donaher JL, et al. Twist, a master regulator of morphogenesis, plays an essential role in tumor metastasis. *Cell* 2004;**117**:927–39.
53. Khan MA, Chen HC, Zhang D, et al. Twist: a molecular target in cancer therapeutics. *Tumour Biol* 2013;**34**:2497–506.
54. Saxena M, Stephens MA, Pathak H, et al. Transcription factors that mediate epithelial-mesenchymal transition lead to multidrug resistance by upregulating ABC transporters. *Cell Death Dis* 2011;**2**:e179.
55. Li QQ, Xu JD, Wang WJ, et al. Twist1-mediated adriamycin-induced epithelial-mesenchymal transition relates to multidrug resistance and invasive potential in breast cancer cells. *Clin Cancer Res* 2009;**15**:2657–65.
56. Deng JJ, Zhang W, Xu XM, et al. Twist mediates an aggressive phenotype in human colorectal cancer cells. *Int J Oncol* 2016;**48**: 1117–24.
57. Zhang X, Wang Q, Ling MT, et al. Anti-apoptotic role of TWIST and its association with Akt pathway in mediating taxol resistance in nasopharyngeal carcinoma cells. *Int J Cancer* 2007;**120**: 1891–8.
58. Jin HO, Hong SE, Woo SH, et al. Silencing of Twist1 sensitizes NSCLC cells to cisplatin via AMPK-activated mTOR inhibition. *Cell Death Dis* 2012;**3**:e319.
59. Zhao ZX, Rahman MA, Chen ZG, et al. Multiple biological functions of Twist1 in various cancers. *Oncotarget* 2017;**8**:20380–93.
60. Farhood B, Najafi M, Mortezaee K. Cancer-associated fibroblasts: secretions, interactions, and therapy. *J Cell Biochem* 2019;**120**: 2791–800.
61. Kobayashi H, Enomoto A, Woods SL, et al. Cancer-associated fibroblasts in gastrointestinal cancer. *Nat Rev Gastroenterol Hepatol* 2019;**16**:282–95.
62. Martinez-Outschoorn UE, Lisanti MP, Sotgia F. Catabolic cancer-associated fibroblasts transfer energy and biomass to anabolic cancer cells, fueling tumor growth. *Semin Cancer Biol* 2014;**25**: 47–60.
63. Mao XQ, Xu J, Wang W, et al. Crosstalk between cancer-associated fibroblasts and immune cells in the tumor microenvironment: new findings and future perspectives. *Mol Cancer* 2021;**20**:131.
64. Wang ZQ, Sun XL, Wang YL, et al. Agrin promotes the proliferation, invasion and migration of rectal cancer cells via the WNT signaling pathway to contribute to rectal cancer progression. *J Recept Signal Transduct Res* 2021;**41**:363–70.
65. Kawahara R, Granato DC, Carnielli CM, et al. Agrin and perlecan mediate tumorigenic processes in oral squamous cell carcinoma. *PLoS One* 2014;**9**:e115004.