*Briefings in Bioinformatics*, 2025, **26(4)**, bbaf309

https://doi.org/10.1093/bib/bbaf309

**Problem Solving Protocol**

# Inference of gene coexpression networks from single-cell transcriptome data based on variance decomposition analysis

Bin Lian[1], Haohui Zhang[1], Tao Wang[1], Yongtian Wang[1], Xuequn Shang[1], N. Ahmad Aziz[2,3], Jialu Hu[1,2,*]

[1]School of Computer Science, Northwestern Polytechnical University, 710072 Xi'an, Shaanxi, China

[2]Population Health Sciences, German Center for Neurodegenerative Diseases (DZNE), Venusberg-Campus 1/99, 53127 Bonn, Germany

[3]Department of Neurology, Faculty of Medicine, University of Bonn, 53105 Bonn, Germany

[*]Corresponding author. E-mail: jhu@nwpu.edu.cn

## Abstract

Gene regulation varies across different cell types and developmental stages, leading to distinct cellular roles across cellular populations. Investigating cell type-specific gene coexpression is therefore crucial for understanding gene functions and disease pathology. However, reconstructing gene coexpression networks from single-cell transcriptome data is challenging due to artifacts, noise, and data sparsity. Here, we present an efficient method for inference of gene coexpression networks via variance decomposition analysis (GCNVDA) to explore the underlying gene regulatory mechanisms from single-cell transcriptome data. Our model incorporates multiple sources of variability, including a random effect term $G$ to capture gene-level variance and a random effect term $E$ to account for residual errors. We applied GCNVDA to three real-world single-cell datasets, demonstrating that our method outperforms existing state-of-the-art algorithms in both sensitivity and specificity for identifying tissue- or state-specific gene regulations. Furthermore, GCNVDA facilitates the discovery of functional modules that play critical roles in key biological processes such as embryonic development. These findings provide new insights into cell-specific regulatory mechanisms and have the potential to significantly advance research in developmental biology and disease pathology.

**Keywords:** gene coexpression networks; single-cell RNA sequencing; linear mixed model; gene functional modules

## Introduction

Genes play distinct regulatory roles across different tissues, developmental stages, and cell states within biological systems [1, 2]. Investigating the underlying gene regulatory mechanisms in normal and diseased tissues is crucial for understanding these roles. Gene coexpression networks (GCNs) have become a widely used tool for modeling complex regulatory systems, where nodes represent genes and edges represent the coexpression relationships between them. Accurately reconstructing GCNs can provide profound insights into the causes of diseases through downstream network-based analyses. Numerous studies on coexpression networks have demonstrated that transcriptional dysregulation can lead to abnormal cellular development, contributing to diseases such as various cancers [3], neurological disorders [4], and psychiatric conditions [5].

Many research studies [6] have focused on inferring gene coexpression networks from bulk RNA-seq data, which enables the identification of functional gene modules based on total gene expression read counts of cells in an entire tissue. However, these approaches are limited in their ability to detect regulatory changes across cell types. The advent of single-cell RNA sequencing (scRNA-seq) technologies [7] has revolutionized our ability to investigate cellular heterogeneity by enabling the inference of gene coexpression networks at single-cell resolution. These technologies facilitate the characterization of cell type-specific coexpression patterns and the analysis of dynamic gene regulatory programs during processes such as cellular differentiation and development. However, a major challenge in constructing accurate GCNs lies in the limited availability or absence of prior knowledge regarding cell type annotations in most scRNA-seq datasets. This lack of predefined labels complicates the identification of biologically meaningful coexpression modules that are specific to distinct cell populations. Therefore, many computational tools have been developed to capture cellular heterogeneity [8–11] and cell-type-specific gene coexpression networks [12].

These GCN inference tools can be grouped into four main categories: ordinary differential equation (ODE)-based models for dynamic system behavior, regression models for predictive relationships, correlation models for association inference, and Boolean network models for logic-based system representation. ODE-based models typically utilize a set of differential equations along with cell pseudotime ordering to model the time-delayed regulatory effects of one gene on another. Some of the algorithms

**Received:** April 27, 2025. **Revised:** May 16, 2025. **Accepted:** June 3, 2025

most widely used in this category include SCODE [13] and GRISLI [14]. Regression-based models employ dependent and independent variables to predict gene regulatory relationships through parametric or non-parametric functions. These models examine how the expression level of one gene (the dependent variable) changes in response to variations in the expression levels of other genes in other genes (the independent variables) while holding the other variables constant. Examples of regression-based models for inferring GCNs include SINCERITIES [15] and GENIE3 [16]. However, both ODE-based and regression-based models rely on strong assumptions. ODE-based models assume that the rate of expression change for each transcription factor (TF) linearly depends on the expression levels of other genes, while regression-based models assume linear or non-linear dependencies between the expression of a target gene and other driving genes. Correlation-based models, such as LEAP [17], scLink [18], and PIDC [19], calculate a correlation matrix based on statistical models to characterize gene coexpression relationships. However, inappropriate assumptions about gene expression distribution can lead to high false positive rates in gene regulation predictions [20, 21]. For example, LEAP and scLink focus on identifying relationships between genes without considering random effects between cells, which may reduce the impact of inter-cellular randomness on accurately inferring gene coexpression networks. Boolean network models [22] consist of nodes that represent genes in the regulatory system, with each node's state quantified as 0 (not expressed) or 1 (expressed). These models simply quantify the regulatory interactions between genes using Boolean discrete variables, predicting only the existence of potential edges between two genes.

While recent approaches have made progress in addressing noise and sparsity in scRNA-seq data, they often suffer from key limitations. For example, methods based on correlation or regression typically assume independence among cells or treat all cells as equally related, ignoring hierarchical or continuous structures such as developmental trajectories. Others rely on strong parametric assumptions that may not hold in heterogeneous tissues, leading to biased or unstable coexpression estimates. Moreover, some techniques require predefined cell type labels or network topologies, which may not be available or reliable in practice.

To address the challenges inherent in constructing gene coexpression networks from single-cell RNA sequencing data, we propose GCNVDA, a novel variance decomposition-based algorithm designed to explicitly account for cellular heterogeneity. Unlike many existing methods that rely on restrictive modeling assumptions—such as linear dependencies, pre-defined gene modules, or the assumption of homogeneity within cell populations—GCNVDA integrates a random effects model with a structured cell–cell covariance matrix to better reflect the underlying biological complexity. This design enables more accurate and robust inference of gene coexpression patterns across diverse cellular contexts. We applied GCNVDA to three biologically diverse single-cell datasets and demonstrated its superior performance in constructing biologically meaningful coexpression networks. Our results show that GCNVDA not only improves the accuracy of gene network inference but also enhances downstream analyses, including clustering, functional module detection, and identification of condition- or state-specific regulatory programs. These improvements highlight GCNVDA's potential as a general-purpose tool for single-cell systems biology and reinforce its advantage over existing state-of-the-art algorithms.

# Materials and Methods

## The GCNVDA model

Here, our aim is to reconstruct a gene coexpression network from single-cell transcriptome data. Suppose that an expression matrix $Y_{p\times n}$ of $n$ cells measured in $p$ genes is randomly sampled from a matrix normal distribution. We fit a model with two covariance components to these samples,

$$\mathbf{Y} = \mathbf{M} + \mathbf{G} + \mathbf{E}; \quad \mathbf{G} \sim MN_{p\times n}(\mathbf{0}, \mathbf{V}_g, \mathbf{K}), \quad \mathbf{E} \sim MN_{p\times n}(\mathbf{0}, \mathbf{V}_e, \mathbf{I}), \qquad (1)$$

where $\mathbf{Y}$ is a random variable of a $p$ by $n$ expression matrix, $\mathbf{M}$ is a $p$ by $n$ matrix representing the mean of gene expression, $\mathbf{G}$ is a $p$ by $n$ matrix of random effects, and $\mathbf{E}$ is a $p$ by $n$ matrix of residual errors. The key point is that the variance of the outcome variable $\mathbf{Y}$ is partitioned into two components represented by $\mathbf{G}$ and $\mathbf{E}$. In our model, we assume that the latent matrix $\mathbf{G}$ follows a matrix normal distribution with mean $\mathbf{0}$, a $p \times p$ row covariance matrix $\mathbf{V}_g$, and a known $n \times n$ column covariance matrix $\mathbf{K}$, which encodes the pairwise relationships among cells. Similarly, the noise matrix $\mathbf{E}$ is assumed to follow a matrix normal distribution with mean $\mathbf{0}$, a row covariance matrix $\mathbf{V}_e$, and a column covariance matrix $\mathbf{I}_{n\times n}$, reflecting independent measurement noise across cells. The use of matrix normal distributions enables us to explicitly model structured dependencies across both genes and cells. The inclusion of $\mathbf{K}$ as the column covariance of $\mathbf{G}$ is motivated by the need to account for structured variation arising from latent cell states, such as cell type, cell cycle phase, or differentiation stage. These factors introduce correlations among cells that, if unaccounted for, may confound the estimation of gene coexpression patterns. By incorporating $\mathbf{K}$, GCNVDA borrows strength from prior knowledge or learned similarity between cells—e.g. derived from diffusion maps, graph-based distances, or principal components—to better disentangle biologically relevant signal from noise. Since $\mathbf{K}$ must be a valid covariance matrix, it is required to be symmetric and positive definite. Therefore, there exists an eigendecomposition $\mathbf{K} = \mathbf{UDU}^{\top}$, where $\mathbf{U}$ is an orthogonal matrix and $\mathbf{D} = \mathrm{diag}(\delta_1, \delta_2, \ldots, \delta_n)$ contains the eigenvalues of $\mathbf{K}$. This decomposition facilitates a transformation of the matrix normal model (Equation 1) into a multivariate normal model, simplifying inference and computation. While the assumptions regarding $\mathbf{G}$ and $\mathbf{E}$ are rooted in established statistical models, their biological plausibility is supported by empirical observations: gene expression variation across cells often reflects both coordinated biological programs (modeled via $\mathbf{G}$) and technical or stochastic noise (modeled via $\mathbf{E}$). Real data analyses in the section of results further confirm that incorporating structured cell–cell covariance significantly improves the accuracy and interpretability of inferred gene coexpression networks. Then, we can transform the matrix normal model in Equation 1 into a multivariate normal model as follows

$$\mathbf{y} = \mathbf{g} + \mathbf{e}; \quad \mathbf{g} \sim MVN(0, \mathbf{D} \otimes \mathbf{V}_g), \quad \mathbf{e} \sim MVN(0, \mathbf{I} \otimes \mathbf{V}_e), \qquad (2)$$

where $\mathbf{y} = \mathrm{vec}(\mathbf{Y} - \mathbf{M})$, $\mathbf{g} = \mathrm{vec}(\mathbf{G} \cdot \mathbf{U})$, $\mathbf{e} = \mathrm{vec}(\mathbf{E} \cdot \mathbf{U})$, $\mathrm{vec}(\mathbf{X})$ denotes the vectorization of $X$ (i.e. stacking columns), $\mathbf{g}$ and $\mathbf{e}$ follow a multivariate normal distribution, and $\otimes$ denotes the kronecker product. Further, we can partition $\mathbf{y}$ in Equation 2 into $n$ independent (but not identical) multivariate normal variables as follows

$$\mathbf{y}_i = \mathbf{g}_i + \mathbf{e}_i; \quad g_i \sim MVN(0, \delta_i \mathbf{V}_g), \quad e_i \sim MVN(0, \mathbf{V}_e), \quad i = \{1, 2, \cdots, n\}, \qquad (3)$$

where $y_i$, $g_i$, and $e_i$ are the $ith$ column of $\mathbf{Y} - \mathbf{M}$, $\mathbf{G} \cdot \mathbf{U}$, and $\mathbf{E} \cdot \mathbf{U}$, respectively.

## The maximum likelihood estimation

To estimate the unknown parameters $\mathbf{V}_g$ and $\mathbf{V}_e$ in Equation 3, we maximize the likelihood probability of the observed gene expression data $\mathbf{Y}$, $P(\mathbf{Y}|\mathbf{V}_g, \mathbf{V}_e) = \Sigma_{\mathbf{G}} P(\mathbf{Y}, \mathbf{G}|\mathbf{V}_g, \mathbf{V}_e)$. The joint log likelihood of $\mathbf{Y}$ and $\mathbf{G}$ is the sum of the log likelihood of all cells, which is

$$\log l(\mathbf{Y}, \mathbf{G}|\mathbf{V}_g, \mathbf{V}_e) = \sum_{i=1}^{n} \log P(y_i, g_i|\mathbf{V}_g, \mathbf{V}_e). \tag{4}$$

It should be noted that $g_i$ is a hidden variable. The joint log likelihood probability of $y_i$ and $g_i$ can be further written as

$$\begin{aligned} \log P(y_i, g_i|\mathbf{V}_g, \mathbf{V}_e) &= \log P(y_i|g_i, \mathbf{V}_g, \mathbf{V}_e) + \log P(g_i|\mathbf{V}_g, \mathbf{V}_e) \\ &= -p\log(2\pi) - \frac{1}{2}\log|\mathbf{V}_e| - \frac{1}{2}e_i^T \mathbf{V}_e^{-1} e_i \\ &\quad - \frac{1}{2}\log|\delta_i \mathbf{V}_g| - \frac{1}{2} g_i^T (\delta_i \mathbf{V}_g)^{-1} g_i. \end{aligned} \tag{5}$$

However, as $g_i$ is a missing value, the maximum likelihood estimation task [23] of estimating parameters $V_g$ and $V_e$ in Equation 5 is a computational challenge. Methods for solving this optimization problem generally fall into two categories based on their use of gradient information: ‘derivative-based’ methods, which rely on analytical or numerical gradients, and ‘derivative-free’ methods, which optimize the objective function without requiring gradient computation. The derivative-free methods search for a combination of parameters along a searching path [24]. These methods are sometimes employed for convenience rather than by necessity. Although they are usually easy to implement, the decision to use a derivative-free method is typically limited by performance in terms of accuracy, expense, or problem size. The computational cost usually grows exponentially with the increasing number of genes and cells. The derivative-based methods determine search direction according to an objective function’s derivative information, which include the expectation maximization (EM) algorithm [25] and its accelerated version PX-EM [26]. Finally, considering the computational requirements, we use the PX-EM algorithm to estimate the two parameters $V_g$ and $V_e$ until the convergence condition is reached.

## Data preprocessing

Our first application utilized the dataset processed by SCODE [13], while the remaining datasets were processed as described below. Initially, we filtered out genes with expression levels of zero in the majority of cells (∼95%). Typically, genes displaying significant variability in expression levels within a cell population are of particular interest, as such variability is often driven by underlying biological factors. To capture this, we calculated the variance of expression counts for each gene across all cells, subsequently selecting those with the highest variability for further analysis. Given the computational demands associated with large gene sets, we ultimately narrowed the focus to 100–500 candidate genes. Next, we normalized the count matrix by the library size of each cell, ensuring that all cells contained $M$ reads post-normalization. Common choices for $M$ include the median library size or a predetermined constant (e.g. $10^5$) [27]. Denoting the normalized matrix as $C$, to reduce the influence of extreme values, we then applied a log10 transformation to the normalized count matrix, resulting in a log-transformed gene expression matrix $Y$, where $Y_{ij} = \log_{10}(C_{ij} + 1)$, for $i = 1, 2, \ldots, p$ and $j = 1, 2, \ldots, n$.

## Selection of genes

The BEELINE framework [28] provided reference ground truth networks for the datasets we used. In the Application 2, to facilitate more reliable performance verification, we used this reference network to calculate each gene’s degree (both in-degree and out-degree). We then selected the top 100 genes with the highest degree as candidate genes for further analysis. Relying solely on variance as a filtering criterion may result in a situation where only a small number of predicted edges correspond to those in the ground truth, leading to a significant imbalance between positive and negative samples.

## Calculation of pseudotime

Among methods we compared, some, like SCODE, require pseudotime information for analysis. To calculate pseudotime for each cell, we utilized the R package monocle3 (version 1.0.0) [29]. Pseudotime can be derived by inputting the expression matrix along with cell and gene information into the package. One challenge in this process is that monocle3 requires the specification of a starting cell, but our datasets lack explicit labels to identify such a cell. To address this, we select several marker genes based on literature research and experience within the dataset and determined the starting cell based on the expression levels of these markers. For instance, in tumor cells, the expression levels of certain proto-oncogenes are elevated. The pseudotime dimensionality reduction trajectory for the cells in Application 3 is illustrated in Supplementary Fig. S1.

## Enrichment analysis

Following gene clustering, we conducted enrichment analysis for each cluster using the R package clusterProfiler (version 3.18.1) [30]. For human datasets, we employed the org.Hs.eg.db database (version 3.12.0) [31] as the input parameter, while for mouse datasets, we used the org.Mm.eg.db database (version 3.12.0) [32].

## Methods and parameters for comparison

In this study, datasets were analyzed with several state-of-the-art methods. The usage details of each method are as follows:

- GENIE3 was obtained from Bioconductor and run using the default parameters of the R implementation version. The GENIE3() function was executed to obtain the weights between all gene pairs.
- SCODE was obtained by running the command ‘git clone https://github.com/hmatsu1226/SCODE’ and was executed with the command ‘Rscript SCODE.R <Input_file1><Input_file2><Output_dir><G><D><C><I>’. The expression matrix and pseudotime matrix were processed into text files and used as Input_file1 and Input_file2. G represented the number of genes, Z was set to the example parameter 4, C was the number of cells, and I was set to the example parameter 100.
- scLink was run using the R package “scLink” (v0.0.1). Specifically, we used the sclink_cor() function, where expr was the preprocessed expression matrix, ncore was set to 8, and nthre and dthre were set to the default parameters.

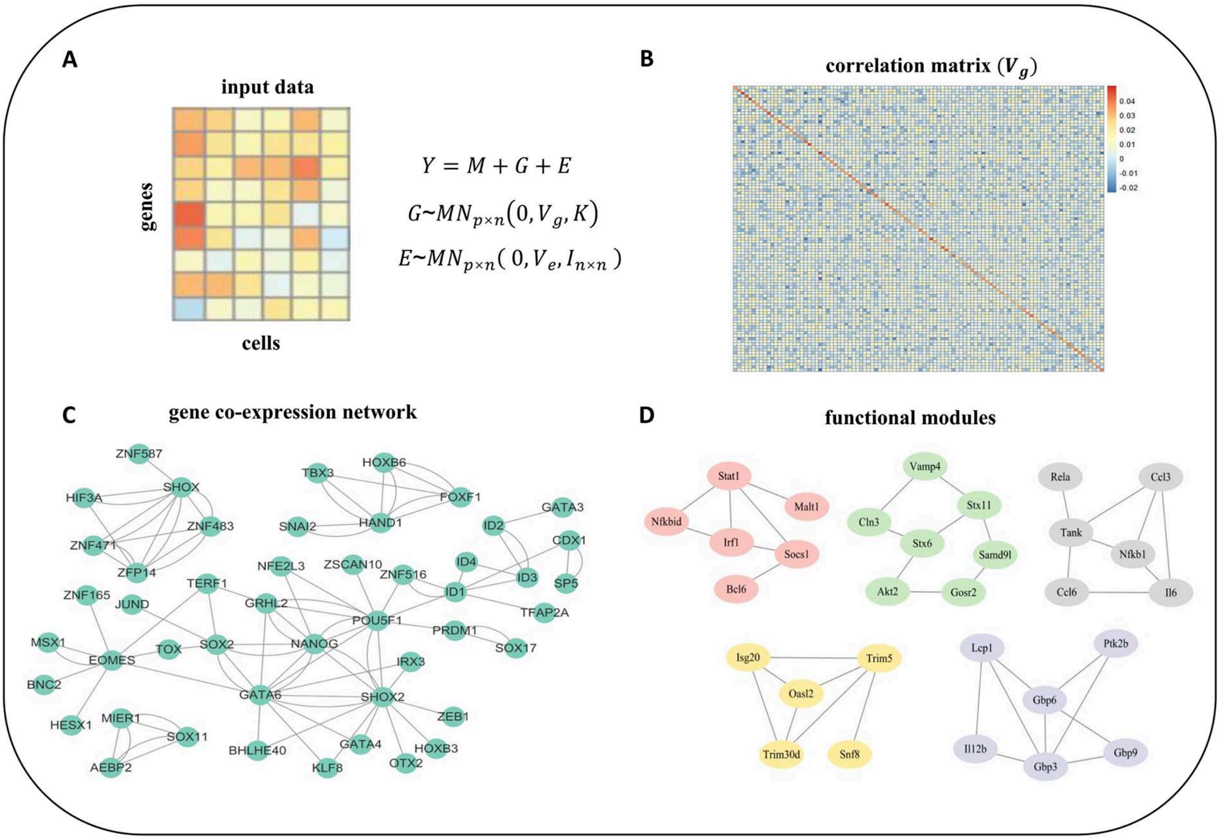

Figure 1. Method schematic for GCNVDA. (A) GCNVDA first models the scRNA-seq data with the formula $Y = M + G + E$. Each row of M is the mean of the corresponding row of $Y$. $G$ and $E$ are two random matrix variables that obey matrix normal distribution, respectively. (B) The covariance matrix that characterizes the correlation between genes is obtained by estimating the parameter $V_g$, and the greater its value, the stronger the interaction relationship between genes. (C) The GCN is obtained by converting $V_g$ into a correlation coefficient matrix and setting a threshold to filter some edges with small correlation coefficients. (D) We use the correlation coefficient as a distance measure to cluster and get several gene sets. Then we did enrichment analysis of these gene sets and selected the functional modules that were both significant ($P$-value $\leq 0.05$) and consistent with the characteristics of the data.

# Results

## GCNVDA method overview

A brief method workflow diagram is provided in Fig. 1. Specifically, we first model the normalized expression data and express it as the sum of three matrices: $M$, $G$, and $E$. Among these, $G$ is the random effect term, $E$ is the residual errors term, and both follow the matrix normal distribution. By estimating the parameters of the row covariance matrices of these two random matrix variables, we obtain the desired correlation matrix $V_g$. Further, this matrix is transformed into a correlation coefficient matrix and used as the final predicted network. In this process, we account for both the random effect ($K$) between cells and random noise, thereby fully utilizing the data. Next, we cluster all genes according to the Euclidean distance of the correlation coefficients. In each cluster, the R package clusterProfiler provides a ranking of GO terms based on significance and identifies a set of functionally related genes for each GO term. We then use these most prominent gene sets to construct gene modules. To test the ability of GCNVDA in constructing and analyzing gene coexpression networks across datasets, we applied GCNVDA in three real data applications. The following subsections are organized as follows: (i) GCNVDA infers a gene coexpression network from human embryonic stem (ES) cells; (ii) GCNVDA captures the response of mouse dendritic cells (mDCs) stimulated by lipopolysaccharide; and (iii) GCNVDA recognizes differences in gene regulation between tumor cells and normal cells.

## GCNVDA accurately infers gene coexpression network from definitive endoderm cells

In this section, we aim to evaluate the accuracy and functional relevance of gene coexpression networks inferred by GCNVDA in definitive endoderm (DE) cells [13, 33], comprising 758 cells and 100 genes. In this application, we compared GCNVDA's performance with other methods using receiver operating characteristic (ROC) and precision-recall (PRC) curves (Fig. 2A and B). The results indicate that GCNVDA outperforms other methods in terms of ROC, with an area under the ROC curve (AUROC) of 0.582 for GCNVDA, compared to 0.483 for GENIE3, 0.518 for SCODE, and 0.543 for scLink. In network inference problems, it is common to encounter a significant imbalance between the number of negative and positive samples, with negative samples typically far outnumbering positive ones. As a consequence, the PR curve may not yield a high overall area under the curve (AUC), even if the classifier performs reasonably well in distinguishing between the two classes. This occurs because the abundance of negative samples dilutes the influence of the fewer positive instances, thus suppressing the AUC in comparison to other metrics that might account for class imbalance differently. The PR curve demonstrates that despite the inherent class imbalance in network inference problems, GCNVDA still achieves a relatively higher precision in identifying true gene interactions, suggesting that it is more robust in handling sparse regulatory connections. Additionally, we extracted the top 1000 edges predicted by each method and used their true positives as a precision metric. As shown in Fig. 2C, GCNVDA achieves the best score (67), followed by scLink (59), GENIE3 (50), and SCODE (41). After confirming the network's accuracy, we proceeded to assess its functional relevance by clustering genes and analyzing enriched GO terms. To further evaluate the utility of the predicted correlation coefficients, we clustered the 100 genes into five clusters using K-means [34] clustering based on the correlation coefficient matrix obtained by GCNVDA. The heatmap (Fig. 2D) shows that genes in the first and fifth clusters exhibit relatively higher within-cluster correlation coefficients than genes in the other three clusters. Notably, key genes such as NANOG, SOX2, and POU5F1, which are crucial for embryonic stem cell development [35], are prominently featured in the first and fifth clusters. We then performed gene enrichment analysis on the gene sets from these five clusters to better understand their functional roles. Subsequently, we refined the gene coexpression networks constructed for each cluster by applying a selective pruning approach. Initially, we identified the most biologically relevant Gene Ontology (GO) terms to ensure the focus remained on significant functional categories. Within these GO categories, we retained only the strongest correlations, thereby enhancing the functional relevance and interpretability of the network structure. Specifically, we first prioritized the top ten edges based on their correlation strength, discarding any edges that did not meet this threshold. Additionally, we removed disconnected edges, thereby further simplifying the network. This pruning process effectively yielded smaller, functionally coherent gene modules, each representing distinct biological pathways or processes (Fig. 2E). These curated modules allow for more focused analyses and interpretations of gene interactions within each cluster, potentially uncovering unique functional insights. Results show that clusters 1, 2, and 5 are significantly enriched in GO terms related to embryonic development, while clusters 3 and 4 are enriched in digestive system development (Figure S2), suggesting that GCNVDA can effectively distinguish functional modules based on gene expression patterns, as evidenced by the distinct enrichment of GO terms between clusters. To emphasize the differences in gene functional modules predicted by GCNVDA, we presented a heatmap of the top eight significantly enriched functions for each cluster. Results in Fig. 3C show that there are only a few overlapped GO terms across clusters. It further indicates that clustering based on our predicted correlation coefficients can effectively distinguish different functional modules.

To explore dynamic gene regulation, we examined how gene relationships change across different time points. We constructed gene coexpression networks for cells at 0, 12, 24, 72, and 96 h, and tracked the regulatory activity of each gene over time. To quantify these regulatory dynamics, we introduced a new metric to evaluate the activity of individual genes throughout the time course. For a given gene $i$, its gene activity score, denoted as $GAS_i$, is defined as follows:

$$\mathrm{GAS_i} = \frac{\sum_{j=1}^{p} \mathbb{1}(|W_{ij}| > 0.25)}{p}, \tag{6}$$

where $p$ represents the number of genes, and $W_{ij}$ denotes the strength of the correlation between genes $i$ and $j$. This metric reflects the proportion of genes for which the absolute value of their correlation coefficient with gene $i$ exceeds 0.25 at a given time point. The threshold of 0.25 is set based on the overall average correlation to filter out genes with low correlation. The regulatory activity of a gene ranges from 0 to 1, with values closer to 1 indicating higher activity and those closer to 0 indicating weaker interactions. Our analysis revealed that the regulatory activities of genes such as SP6, ZFX, ID1, AEBP2, and NANOG fluctuated significantly over time, suggesting their pivotal roles in the differentiation of embryonic stem cells into endoderm cells (Fig. 3A). For example, SP6 is known for its maternally derived expression and is essential for embryonic and extra-embryonic tissue development [36]. We also observed dramatic changes in the correlation coefficients of selected genes, such as SOX2, NANOG, and ZFX, over time, with the correlation between ZFX and NANOG shifting from 0.26 to –0.22, and between ZFX and SOX2 shifting from 0.20 to –0.33. These changes suggest a transition from mutual activation to mutual inhibition over time. This observation was further corroborated by comparing their expression levels over pseudotime (Fig. 3B). Initially, these genes were highly expressed, but as time progressed, the expression levels of SOX2 and NANOG declined, consistent with our findings. A literature review further supports the functional interactions between these genes; for instance, higher ZFX expression has been linked to the loss of NANOG expression during endoderm differentiation, with ZFX overexpression reducing spontaneous differentiation while permitting directed differentiation, thereby maintaining hESC pluripotency [37]. This illustrates GCNVDA's capability to uncover functional relationships between genes and facilitate subsequent functional analyses. Some other functional interactions, such as TERF1 and POU5F1, have also been identified, as shown in Supplementary Fig. S3. It can be concluded that GCNVDA effectively captures dynamic gene regulatory networks, revealing distinct functional modules and time-dependent interactions essential to DE cell differentiation.

## GCNVDA captures the response of bone marrow-derived dendritic cells stimulated by lipopolysaccharide

To assess GCNVDA's capability to capture temporal dynamics in gene coexpression relationships, we conducted analyses with GCNVDA and several benchmark methods on a dataset comprising 7371 genes from 1700 bone marrow-derived dendritic cells. These cells were subjected to stimulation with lipopolysaccharide (LPS) at four distinct time points—1, 2, 4, and 6 hours post-stimulation [38, 39]. A list of gene interactions used as ground truth for validating the predicted networks is provided in Supplementary Table S1. These reference interactions were selected based on established biological evidence, serving as a benchmark to assess the accuracy and relevance of the network predictions generated by each method. Results in Fig. 4A show that GCNVDA achieved the highest AUROC score (0.561), followed by GENIE3 (0.550), SCODE (0.541), and scLink (0.501). In a manner consistent with the previous application, we quantified the true positive counts of the top 50 000 edges to assess their accuracy in predicting biologically meaningful gene interactions. GCNVDA identified 82 true positive edges, compared to 68 by scLink, and 65 each by GENIE3 and SCODE (Fig. 4B). Furthermore, to investigate the biological functions of the genes, we applied clustering techniques [34] to the predicted correlation matrix, resulting in the formation of five distinct gene clusters. Figure 4C presents a heatmap illustrating the gene correlation patterns within the identified clusters. Among them, the correlation between the first and fourth clusters is relatively high compared to other clusters, corresponding to the functions of regulating T cell differentiation and regulating the lifecycle of viruses, respectively. Regulation of T cell differentiation is an important mechanism for the host immune system to resist viruses, directly affecting the

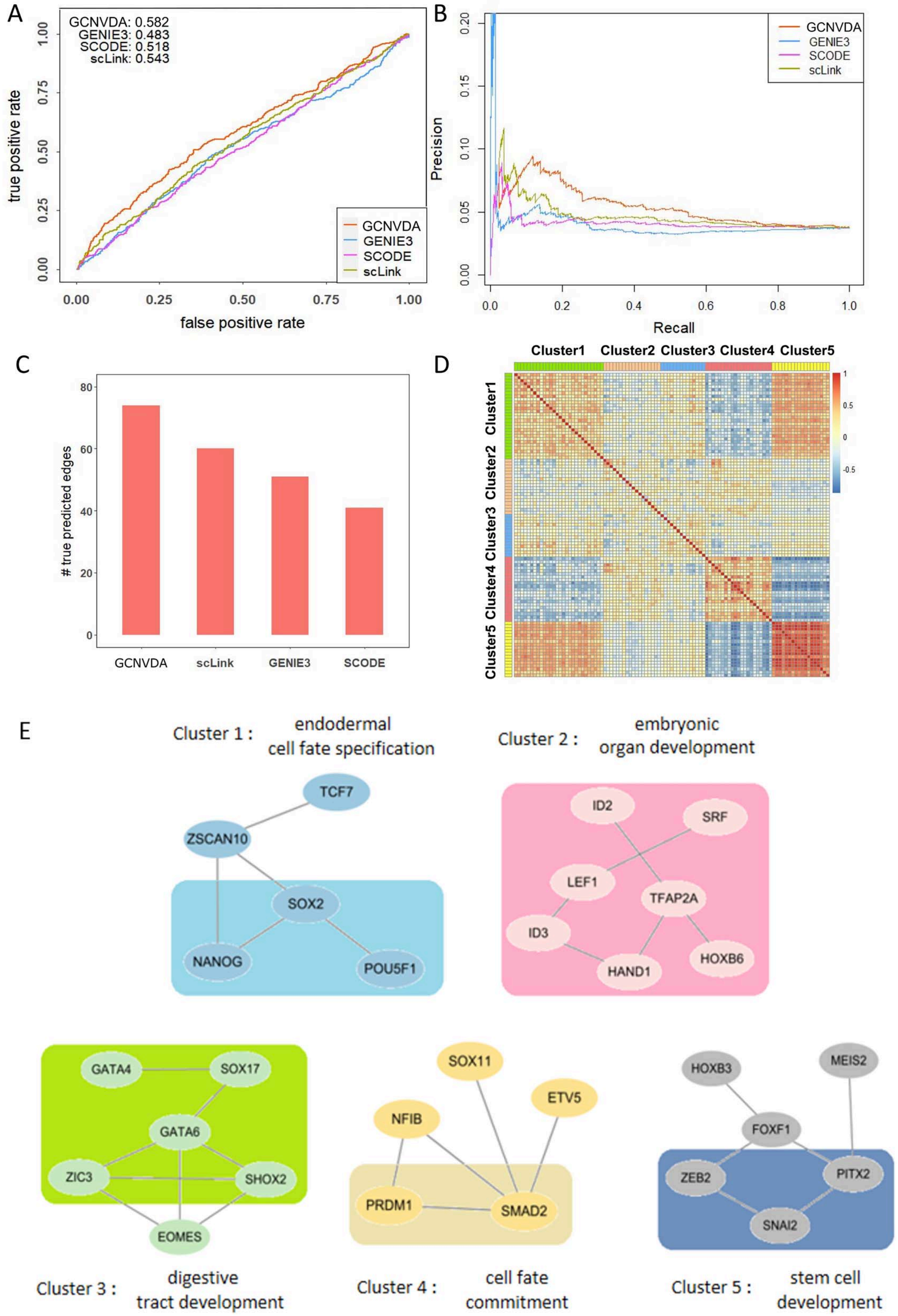

Figure 2. The performance of GCNVDA in inferring gene coexpression networks and the effect of clustering with its predicted correlation. (A) ROC of the four methods and their respective AUC values. (B) PRC of the four methods. (C) The height of the histogram represents the number of coexpression relationships in the first 1000 interactions predicted by this method. (D) The strength of the correlation coefficient is depicted in the form of a heatmap. These genes' clusters are marked on the top and left, respectively. (E) The sub-networks are constituted by the five clusters, respectively. The part circled by a rectangle in the figure is a functional module. Its functions are marked around the figure.

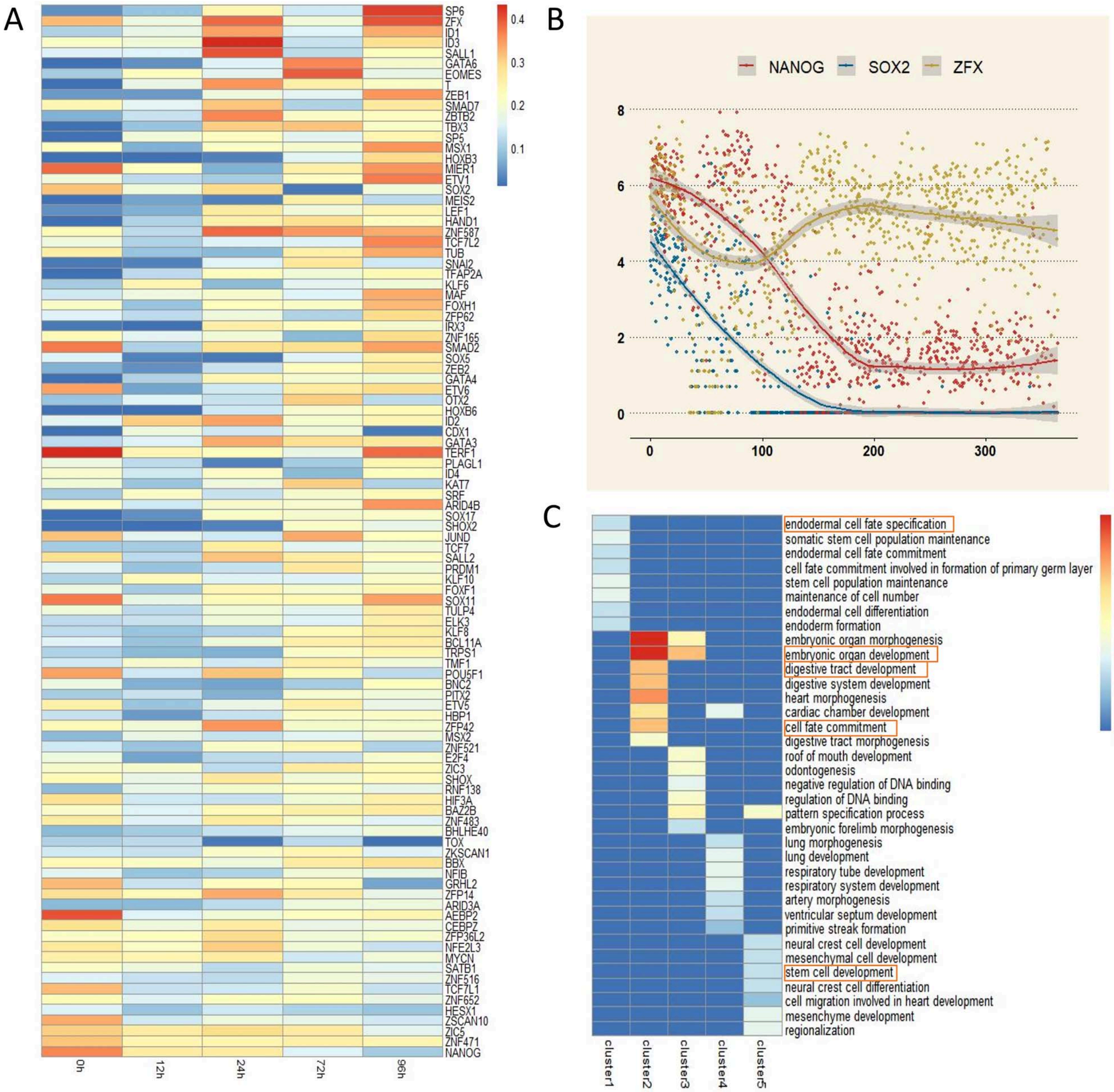

Figure 3. (A) Heatmap depicting the value of gene regulatory activity of 100 genes in five different periods. The genes on the right are arranged in a gradual gradient. (B) The dots in the figure represent cells. The fitting line indicates the expression trend of the gene with pseudotime. (C) The value of the heat map represents the number of genes enriched in this function in the corresponding cluster.

regulation of the virus lifecycle. And viruses interfere with T cell differentiation in various ways, prolonging their own life cycle [40, 41]. Similarly, we constructed gene function modules (Fig. 4D) for each cluster using the selective pruning approach of the previous application. The significant biological functions identified across the five classes encompass key processes such as the regulation of alpha-beta T cell differentiation, vesicle fusion, cellular responses to interleukin-1, regulation of the viral life cycle, and defense responses to protozoan infections. These functions are consistent with the anticipated immune and inflammatory responses triggered by lipopolysaccharide stimulation, as documented in previous studies [42–46]. As shown in Fig. 4E, there was no overlap in their enriched GO terms of the five distinct groups, which further demonstrates that clustering based on our predicted correlations can accurately identify modules with relatively independent functions. Next, we predicted the gene coexpression networks for datasets at each time point and analyzed the temporal changes in gene correlations and functional activities. Using the equation (1) described above, we calculated the regulatory activity of each gene across these four time points (Fig. 4F). The results reveal a clear gradient in gene regulatory activity across the time series. Notably, the regulatory activity of Gbp2, Gbp6, Cd40, and Ccl22 increased significantly with prolonged stimulation times. Gbp2 and Gbp6 are involved in various biological processes, including the cellular response to lipopolysaccharide, a phenomenon accurately captured by our method. Furthermore, lipopolysaccharide stimulation is known to promote cytokine secretion, triggering specific immune responses, while Cd40 enhances antigen-binding activity, and Ccl22 participates in multiple processes, including the cellular

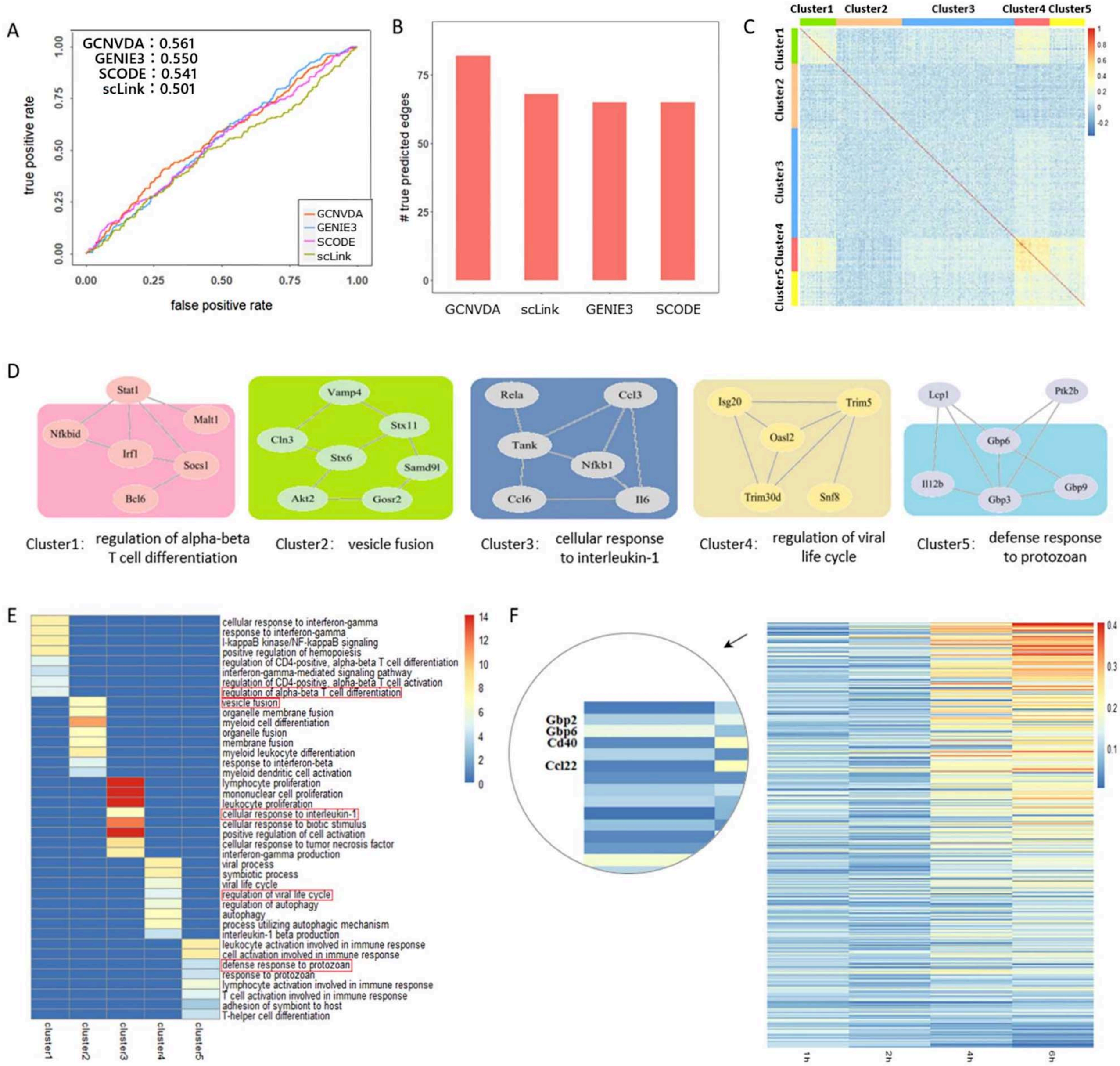

Figure 4. Results of mDC dataset. (A) ROC of the four methods and their respective AUC values. (B) PRC of the four methods. (C) The strength of the correlation coefficient is depicted in the form of a heatmap. These genes' clusters are marked on the top and left, respectively. (D) The sub-networks are constituted by the five clusters, respectively. The part circled by a rectangle in the figure is a functional module. Its functions are marked around the figure. (E) The value of the heat map represents the number of genes enriched in this function in the corresponding cluster. (F) Heatmap depicting the value of gene regulatory activity of 414 genes at four different times after stimulation by lipopolysaccharide. The genes on the right are arranged in a gradual gradient. Some of the top-ranked genes are shown in the left magnifying glass.

response to cytokine stimulation. These findings demonstrate that regulatory activity calculated by our method effectively captures the biological essence that changes over time. Hence, we concluded that GCNVDA effectively captures time-dependent shifts in gene coexpression networks and identifies distinct functional modules responsive to lipopolysaccharide stimulation.

## GCNVDA discovers novel transcription factors in tumor cells

To assess GCNVDA's ability in predicting transcription factors uniquely activated in tumor cells, we analyzed the GSE182434 dataset [47], which comprises 49 632 genes from a patient (ID: DLBCL002B) diagnosed with diffuse large B-cell lymphoma (DLBCL), including 1568 tumor B cells and 82 normal B cells. To do this, we first used GCNVDA to calculate two correlation matrices $X$ and $Y$, for all pairs of genes in tumor and normal cells, respectively. Then, we applied paired t-test for $(X_j,Y_j)$, $j \in 1, \cdots, p$, to determine whether each gene maintains the same regulatory role in both coexpression networks. If genes in the tumor network show significant differences (adjust-$P$ value $<0.05$) compared to their counterparts in the normal network, they may play potential markers of tumor development. Consequently, our analysis identified three candidate marker genes, FOSB (adjust-$P$ $= 1.08 \times 10^{-7}$), JUNB (adjust-$P = 6.13 \times 10^{-7}$), and JUN (adjust-$P$ $= 1.59 \times 10^{-3}$), which have been previously implicated in tumor progression [48–52]. The $t$-test results of all genes are shown in

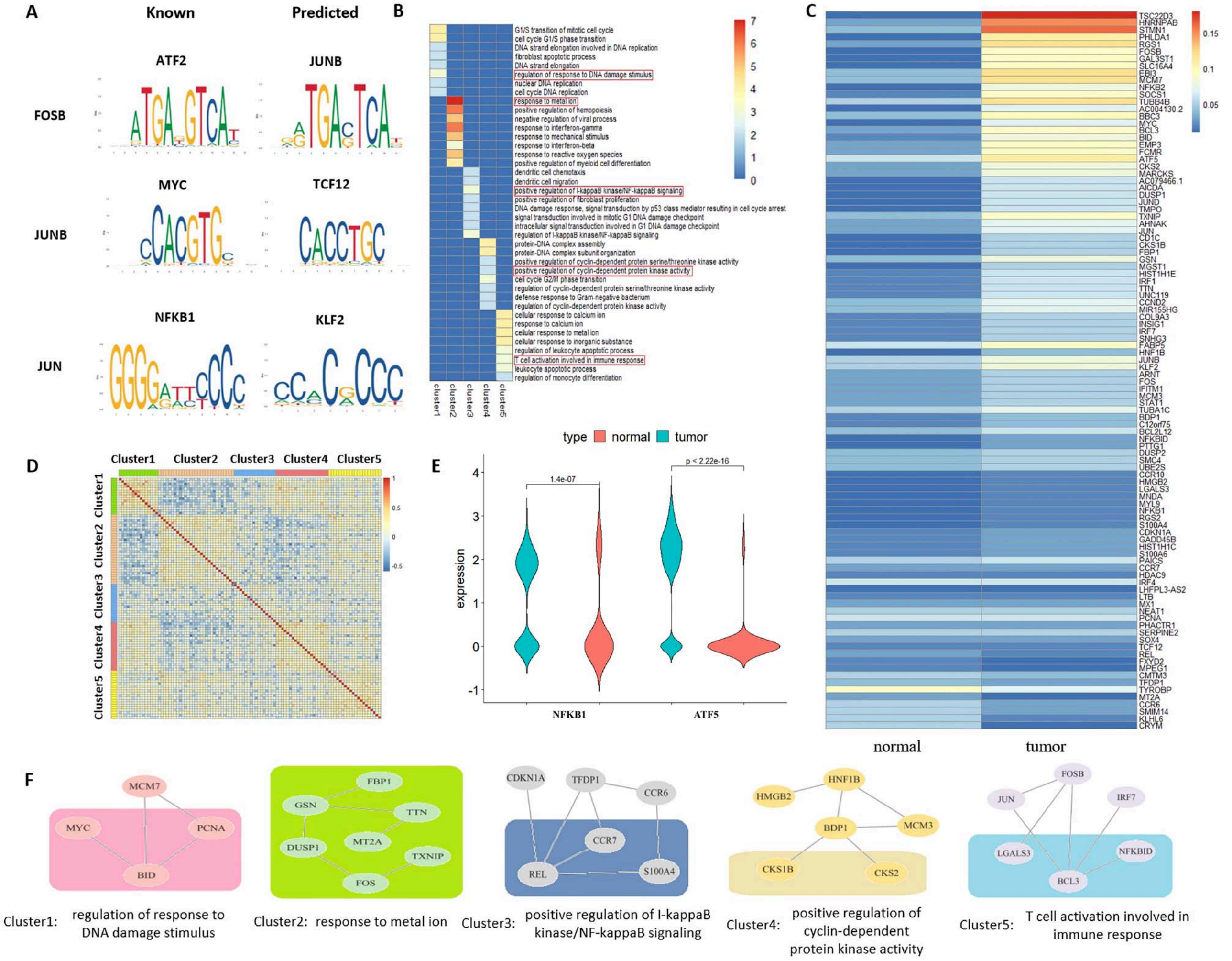

Figure 5. (A) The binding sites of the known transcription factors and potential transcription factors predicted by us of the three genes. (B) The value of the heat map represents the number of genes enriched in this function in the corresponding cluster. (C) Heatmap depicting the value of gene regulatory activity of 100 genes in normal cells and tumor cells. The genes on the right are arranged in a gradual gradient. (D) The strength of the correlation coefficient is depicted in the form of a heatmap. These genes' clusters are marked on the top and left, respectively. (E) Violin diagram of the expression levels of two genes in two different states. The *P*-value of the Wilcox test in two states is marked in the figure. (F) The sub-networks are constituted by the five clusters, respectively. The part circled by a rectangle in the figure is a functional module. Its functions are marked around the figure.

Supplementary Fig. S4. From Supplementary Fig. S4B–D, it can be seen that compared to the negative correlation in normal cells, FOSB, JUNB, and JUN all exhibit co expression with their transcription factors in tumor cells.

To identify candidate transcription factors potentially regulating these genes, we examined our predicted gene coexpression network in tumor cells and selected JUNB, TCF12, and KLF2 as the most strongly co-expressed genes with FOSB, JUNB, and JUN, respectively.

To verify whether these genes could function as potential transcription factors, we compared their motif sequences with the motifs of known corresponding transcription factors. we consulted the Transcription Factor Regulatory Network database (http://www.regulatorynetworks.org). This database is built using DNaseI footprints and TF-binding motifs [53]. From the database, we identified ATF2, MYC, and NFKB1 as known transcription factors regulating FOSB, JUNB, and JUN, respectively. Notably, JASPAR analysis (Fig. 5A) revealed strong motif similarity between each candidate transcription factor and its corresponding known regulator: JUNB with ATF2, TCF12 with MYC, and KLF2 with NFKB1. These findings demonstrate that GCNVDA effectively identifies novel transcription factors associated with tumorigenic activity and highlight its utility in identifying key regulatory elements in tumor cell development. Building on these findings, we next investigated the functional organization of tumor-specific regulatory networks. Clustering the top 100 genes based on their correlation matrices revealed five distinct functional modules, as visualized in a heatmap (Fig. 5B). Notably, clusters 1 and 4 exhibited strong correlations and shared functions such as regulating DNA damage response and modulating cyclin-dependent kinase activity (Fig. 5D). Specifically, functional enrichment analysis confirmed that these clusters contain key regulators of tumor progression, including genes involved in DNA repair and kinase activity, both of which are critical in DLBCL pathogenesis [54, 55]. These findings validate GCNVDA's ability to distinguish functionally relevant gene modules in tumor cells. We next evaluated GCNVDA's ability to detect differential gene activity between tumor and normal cells. Our analysis identified TSC22D3 as the gene with the most significant change in regulatory activity. TSC22D3 encodes the anti-inflammatory protein GC-induced leucine zipper, which plays a crucial role in immunosuppression. Additionally, FOSB, a member of the AP-1 family [51, 56], exhibited altered activity, supporting its role in inflammation and tumorigenesis [48] (Fig. 5C). These observations confirm that GCNVDA effectively captures differential regulatory activity between normal and tumor states. To further explore transcriptional shifts, we identified differentially expressed genes between normal and tumor cells, focusing on NFKB1 and ATF5. Both genes displayed lower expression in normal cells but were significantly upregulated in tumor cells (Fig. 5E). The correlation between NFKB1 and ATF5 shifted from −0.097 in normal cells to 0.325 in tumor cells. Validation using the transcription factor regulatory network database confirmed that NFKB1 regulates ATF5, reinforcing the relevance of these transcriptional changes. These results highlight the capacity of GCNVDA to detect tumor-specific regulatory alterations.r0. Finally, we examined the biological functions of genes from the five clusters, revealing key roles in immune response mechanisms. Genes in cluster 3 were enriched for I-kappaB kinase and NF-kappaB signaling pathways, which are essential for immune regulation and cancer development [57]. Cluster 5 contained genes involved in T-cell activation, illustrating the critical role of B cells in immune processes (Fig. 5F). To identify enriched pathways, we conducted pathway enrichment analyses using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database and GO enrichment analysis. The detailed results of the GO enrichment and KEGG pathway analyses for all five clusters are presented in Supplementary Figs S5–S7. These findings reinforce GCNVDA's ability to uncover immune-related regulatory mechanisms in tumor cells. Overall, our findings highlight GCNVDA's effectiveness in uncovering tumor-specific regulatory mechanisms and immune-related functional modules. Importantly, by screening for highly correlated genes, GCNVDA has the potential to discover new transcription factors related to tumor cell development.

## Discussion

The advancement of single-cell sequencing technology has revolutionized the inference of gene coexpression networks, enabling unprecedented resolution. In this study, we introduce GCNVDA, a novel method specifically designed to leverage scRNA-seq data for more accurate and reliable network predictions. By addressing cellular heterogeneity through modeling random effects, GCNVDA reduces errors, enhances the identification of functional modules, and avoids biases associated with predefined assumptions.

Our findings highlight three key strengths of GCNVDA: its precision in inferring gene coexpression networks, its ability to classify functionally distinct gene sets, and its use of a statistical model that directly analyzes gene expression data. These features allow GCNVDA to uncover biologically relevant functional modules and identify potential transcription factors, making it a valuable tool for exploring gene regulatory mechanisms and functional relationships.

Despite its strengths, GCNVDA is best suited for analyses focusing on highly variable genes or targeted subsets of the transcriptome due to computational demands. Future work aims to integrate pseudotime and spatial transcriptomics data to further refine network predictions and capture dynamic regulatory changes, potentially broadening its applicability and improving performance in large-scale datasets.

**Key Points**

- We present gene coexpression networks via variance decomposition analysis (GCNVDA), a novel computational framework for inferring gene coexpression networks from single-cell RNA sequencing data. Unlike existing methods, GCNVDA is specifically designed to address cellular heterogeneity without relying on restrictive modeling assumptions commonly used in regression-based approaches.
- GCNVDA integrates a random effects model with a precomputed cell-to-cell similarity matrix, K, which encodes prior information about cellular relationships. This integration effectively mitigates the confounding influence of cell state variability, leading to more accurate and biologically meaningful GCNs.
- We demonstrate the utility and robustness of GCNVDA across multiple scRNA-seq datasets, including human embryonic stem cells and tumor-infiltrating B cells. Empirical evaluations show that GCNVDA outperforms state-of-the-art methods in detecting functional gene modules and recovering known gene regulatory interactions.
- Our results highlight GCNVDA's potential as a general-purpose tool for single-cell network inference, offering improved resolution and interpretability in studies of cellular differentiation, disease progression, and regulatory dynamics.

## Author contributions

Jialu Hu (Conceptualization, Resources, Supervision, Funding acquisition and Project administration), Jialu Hu and Bin Lian (Software, Formal analysis, Methodology, Writing – original draft, and Writing – review & editing) and Bin Lian (Data curation, Validation, Investigation, Visualization). Ahmad Aziz, Xuequn Shang, Tao Wang, and Yongtian Wang contributed to many discussions to improve the manuscript.

## Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

## Conflict of interest

None declared.

## Funding

National Natural Science Foundation of China (Grant No. 62072374). This work has also been supported by the China Scholarship Council (CSC).

## Data availability

**Application 1:** The first dataset is a scRNA-seq time course (at 0, 12, 24, 36, 72, and 96 hours) derived from DE cells differentiated from human ES cells, comprising 758 cells. The original dataset is available from GEO (GSE75748) [33] and can be downloaded at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE75748. In this application, we utilized the processed version provided by SCODE [13]. SCODE preprocessed this dataset and selected 100 highly variable genes for network inference. The processed dataset can be accessed at https://github.com/hmatsu1226/SCODE/tree/master/data3.

**Application 2:** In the second application, datasets were obtained from BEELINE and downloaded from https://zenodo.org/record/3701939. The dataset contains 7371 genes and 383 cells, with cells stimulated using lipopolysaccharides for 1, 2, 4, and 6 hours. BEELINE provided the ground truth for the dataset. After preprocessing, we selected 414 genes (using a degree threshold of 7) based on the degree of nodes in the graph constructed by the reference network, in descending order, for subsequent inference. This threshold was chosen to ensure that the number of selected genes fell within the range of 100–500 while maximizing the number of included genes (see more in Selection of genes).

**Application 3:** In the third application, we analyzed a dataset comprising 24 379 tumor cells from diffuse large B-cell lymphoma (DLBCL) and 4037 normal cells, obtained from the GEO database (GSE182434) [47] and available for download at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE182434. To minimize variability introduced by different patients, we focused on data from the same patient within the dataset. Specifically, we selected 1568 tumor cells and 82 normal cells, all derived from patient DCBCL002 and identified as B cells. The raw data were normalized using counts per million (CPM), followed by a $log(1 + CPM)$ transformation. Given the large number of genes in the dataset (49 632 genes), we selected the top 100 most highly variable genes to ensure computational efficiency.

The R source code of GCNVDA for reproducing the results of this paper can be accessible at https://github.com/jhu99/GCNVDA.

## References

1. Mackay TFC, Stone EA, Ayroles JF. The genetics of quantitative traits: challenges and prospects. *Nat Rev Genet* 2009;**10**:565–77. https://doi.org/10.1038/nrg2612
2. Choobdar S, Ahsen ME, Crawford J. *et al.* Assessment of network module identification across complex diseases. *Nat Methods* 2019;**16**:843–52. https://doi.org/10.1038/s41592-019-0509-5
3. Bailey P, Chang DK, Nones K. *et al.* Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature* 2016;**531**:47–52. https://doi.org/10.1038/nature16965
4. Jeremy A, Miller SH, Geschwind DH. Divergence of human and mouse brain transcriptome highlights alzheimer disease pathways. *Proc Natl Acad Sci* 2010;**107**:12698–703.
5. Voineagu I, Wang X, Johnston P et al. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* 2011;**474**:380–4. https://doi.org/10.1038/nature10110
6. Li WV, Li JJ. Modeling and analysis of RNA-seq data: a review from a statistical perspective. *Quant Biol* 2018;**6**:195–209.
7. Kolodziejczyk AA, Kim JK, Svensson V. *et al.* The technology and biology of single-cell RNA sequencing. *Mol Cell* 2015;**58**:610–20. https://doi.org/10.1016/j.molcel.2015.04.005
8. Zhang H, Wang Y, Lian B. *et al.* Scbean: a python library for single-cell multi-omics data analysis. *Bioinformatics* 2024;**40**:btae053.
9. Wang Y, Lian B, Zhang H. *et al.* A multi-view latent variable model reveals cellular heterogeneity in complex tissues for paired multimodal single-cell data. *Bioinformatics* 2023;**39**:btad005.
10. Jialu H, Zhong Y, Shang X. A versatile and scalable single-cell data integration algorithm based on domain-adversarial and variational approximation. *Brief Bioinform* 2022;**23**:bbab400. https://doi.org/10.1093/bib/bbab400
11. Jialu H, Chen M, Zhou X. Effective and scalable single-cell data alignment with non-linear canonical correlation analysis. *Nucleic Acids Res* 2022;**50**:e21–1.
12. Chang S, Zichun X, Shan X. *et al.* Cell-type-specific co-expression inference from single cell RNA-sequencing data. *Nat Commun* 2023;**14**:4846. https://doi.org/10.1038/s41467-023-40503-7
13. Matsumoto H, Kiryu H, Furusawa C. *et al.* Scode: an efficient regulatory network inference algorithm from single-cell RNA-seq during differentiation. *Bioinformatics* 2017;**33**:2314–21. https://doi.org/10.1093/bioinformatics/btx194
14. Aubin-Frankowski P-C, Vert J-P. Gene regulation inference from single-cell RNA-seq data with linear differential equations and velocity inference. *Bioinformatics* 2020;**36**:4774–80. https://doi.org/10.1093/bioinformatics/btaa576
15. Gao NP, Minhaz Ud-Dean SM, Gandrillon O et al. Sincerities: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics* 2018;**34**:258–66.
16. Huynh-Thu VA, Irrthum A, Wehenkel L. *et al.* Inferring regulatory networks from expression data using tree-based methods. *PloS One* 2010;**5**:e12776. https://doi.org/10.1371/journal.pone.0012776
17. Specht AT, Li J. Leap: constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering. *Bioinformatics* 2017;**33**:764–6. https://doi.org/10.1093/bioinformatics/btw729
18. Li WV, Li Y. Sclink: inferring sparse gene co-expression networks from single-cell expression data. *Genom Proteom Bioinform* 2021;**19**:475–92.
19. Chan TE, Stumpf MPH, Babtie AC. Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Syst* 2017;**5**:251–267.e3. https://doi.org/10.1016/j.cels.2017.08.014
20. Langfelder P, Horvath S. Wgcna: an r package for weighted correlation network analysis. *BMC Bioinform* 2008;**9**:1–13.
21. Iacono G, Massoni-Badosa R, Heyn H. Single-cell transcriptomics unveils gene regulatory network plasticity. *Genome Biol* 2019;**20**:1–20.
22. Coifman RR, Lafon S, Lee AB. *et al.* Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proc Natl Acad Sci* 2005;**102**:7426–31.
23. McLachlan GJ, Krishnan T. *The EM Algorithm and Extensions*. 2nd ed. Hoboken, NJ: John Wiley & Sons; 2007. https://doi.org/10.1002/9780470191613.
24. Meyer K. Estimating variances and covariances for multivariate animal models by restricted maximum likelihood. *Genet Select Evol* 1991;**23**:67–83. https://doi.org/10.1186/1297-9686-23-1-67
25. DempsterAP L. Rubindb. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Ser B (Methodol)* 1977;**39**:1r38.
26. Liu C, Rubin DB, Ying Nian W. Parameter expansion to accelerate EM: the PX-EM algorithm. *Biometrika* 1998;**85**:755–70. https://doi.org/10.1093/biomet/85.4.755
27. Azizi E, Prabhakaran S, Carr A. *et al.* Bayesian inference for single-cell clustering and imputing. *Genom Comput Biol* 2017;**3**:e46–6.
28. Pratapa A, Jalihal AP, Law JN. *et al.* Benchmarking algorithms for gene regulatory network inference from single-cell transcrip-

tomic data. *Nat Methods* 2020;**17**:147–54. https://doi.org/10.1038/s41592-019-0690-6

29. Cao J, Spielmann M, Qiu X. *et al*. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* 2019;**566**: 496–502. https://doi.org/10.1038/s41586-019-0969-x
30. Guangchuang Y, Wang L-G, Han Y. *et al*. Clusterprofiler: an r package for comparing biological themes among gene clusters. *Omics: J Integr Biol* 2012;**16**:284–7.
31. Carlson M. org.Hs.eg.db: Genome wide annotation for Human. R package version 3.12.0. Bioconductor. 2020. https://doi.org/10.18129/B9.bioc.org.Hs.eg.db
32. Carlson M. org.Mm.eg.db: Genome wide annotation for Mouse. R package version 3.12.0. Bioconductor. 2020. https://doi.org/10.18129/B9.bioc.org.Mm.eg.db
33. 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.
34. Lloyd S. Least squares quantization in PCM. *IEEE Trans Inform Theory* 1982;**28**:129–37. https://doi.org/10.1109/TIT.1982.1056489
35. Lunde K, Belting H-G, Driever W. Zebrafish pou5f1/pou2, homolog of mammalian oct4, functions in the endoderm specification cascade. *Curr Biol* 2004;**14**:48–55. https://doi.org/10.1016/j.cub.2003.11.022
36. Parker-Katiraee L, Carson AR, Yamada T. *et al*. Identification of the imprinted klf14 transcription factor undergoing human-specific accelerated evolution. *PLoS Genet* 2007;**3**:e65. https://doi.org/10.1371/journal.pgen.0030065
37. Harel S, Tu EY, Weisberg S. *et al*. ZFX controls the self-renewal of human embryonic stem cells. *PLoS One* 2012;**7**:e42302. https://doi.org/10.1371/journal.pone.0042302
38. Shalek AK, Satija R, Shuga J. *et al*. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* 2014;**510**: 363–9. https://doi.org/10.1038/nature13437
39. Qiu X, Rahimzamani A, Wang L. *et al*. Inferring causal gene regulatory networks from coupled single-cell expression dynamics using scribe. *Cell Syst* 2020;**10**:265–274.e11. https://doi.org/10.1016/j.cels.2020.02.003
40. Zhou X, Shuyang Y, Zhao D-M. *et al*. Differentiation and persistence of memory cd8+ t cells depend on t cell factor 1. *Immunity* 2010;**33**:229–40. https://doi.org/10.1016/j.immuni.2010.08.002
41. Grakoui A, Bromley SK, Sumen C. *et al*. The immunological synapse: a molecular machine controlling t cell activation. *Science* 1999;**285**:221–7. https://doi.org/10.1126/science.285.5425.221
42. Kawai T, Akira S. The role of pattern-recognition receptors in innate immunity: update on toll-like receptors. *Nat Immunol* 2010;**11**:373–84. https://doi.org/10.1038/ni.1863
43. Lustig A, Liu HB, Jeffrey Metter E. *et al*. Telomere shortening, inflammatory cytokines, and anti-cytomegalovirus antibody follow distinct age-associated trajectories in humans. *Front Immunol* 2017;**8**:1027. https://doi.org/10.3389/fimmu.2017.01027
44. Dinarello CA. Immunological and inflammatory functions of the interleukin-1 family. *Annu Rev Immunol* 2009;**27**:519–50. https://doi.org/10.1146/annurev.immunol.021908.132612
45. Sadler AJ, Williams BRG. Interferon-inducible antiviral effectors. *Nat Rev Immunol* 2008;**8**:559–68. https://doi.org/10.1038/nri2314
46. Nathan C, Shiloh MU. Reactive oxygen and nitrogen intermediates in the relationship between mammalian hosts and microbial pathogens. *Proc Natl Acad Sci* 2000;**97**:8841–8.
47. Steen CB, Luca BA, Esfahani MS. *et al*. The landscape of tumor cell states and ecosystems in diffuse large b cell lymphoma. *Cancer Cell* 2021;**39**:1422–1437.e10. https://doi.org/10.1016/j.ccell.2021.08.011
48. Eferl R, Wagner EF. Ap-1: a double-edged sword in tumorigenesis. *Nat Rev Cancer* 2003;**3**:859–68. https://doi.org/10.1038/nrc1209
49. Ozanne BW, Spence HJ, McGarry LC. *et al*. Transcription factors control invasion: Ap-1 the first among equals. *Oncogene* 2007;**26**: 1–10. https://doi.org/10.1038/sj.onc.1209759
50. Shaulian E, Karin M. Ap-1 as a regulator of cell life and death. *Nat Cell Biol* 2002;**4**:E131–6. https://doi.org/10.1038/ncb0502-e131
51. Zarubin T, Han J. Activation and signaling of the p38 map kinase pathway. *Cell Res* 2005;**15**:11–8. https://doi.org/10.1038/sj.cr.7290257
52. Ren B, Cam H, Takahashi Y. *et al*. E2f integrates cell cycle progression with DNA repair, replication, and g2/m checkpoints. *Genes Dev* 2002;**16**:245–56. https://doi.org/10.1101/gad.949802
53. Neph S, Stergachis AB, Reynolds A. *et al*. Circuitry and dynamics of human transcription factor regulatory networks. *Cell* 2012;**150**:1274–86. https://doi.org/10.1016/j.cell.2012.04.040
54. He W, Zhijian X, Song D. *et al*. Antitumor effects of rafoxanide in diffuse large b cell lymphoma via the pten/pi3k/akt and jnk/c-Jun pathways. *Life Sci* 2020;**243**:117249. https://doi.org/10.1016/j.lfs.2019.117249
55. Morin RD, Mendez-Lago M, Mungall AJ. *et al*. Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature*. 2011;**476**:298–303. https://doi.org/10.1038/nature10351
56. Ye N, Ding Y, Wild C. *et al*. Small molecule inhibitors targeting activator protein 1 (ap-1) miniperspective. *J Med Chem* 2014;**57**: 6930–48. https://doi.org/10.1021/jm5004733
57. Prescott JA, Mitchell JP, Cook SJ. Inhibitory feedback control of nf-κb signalling in health and disease. *Biochem J* 2021;**478**:2619–64. https://doi.org/10.1042/BCJ20210139

*Briefings in Bioinformatics*, 2025, **26(4)**, bbaf309
https://doi.org/10.1093/bib/bbaf309
Problem Solving Protocol