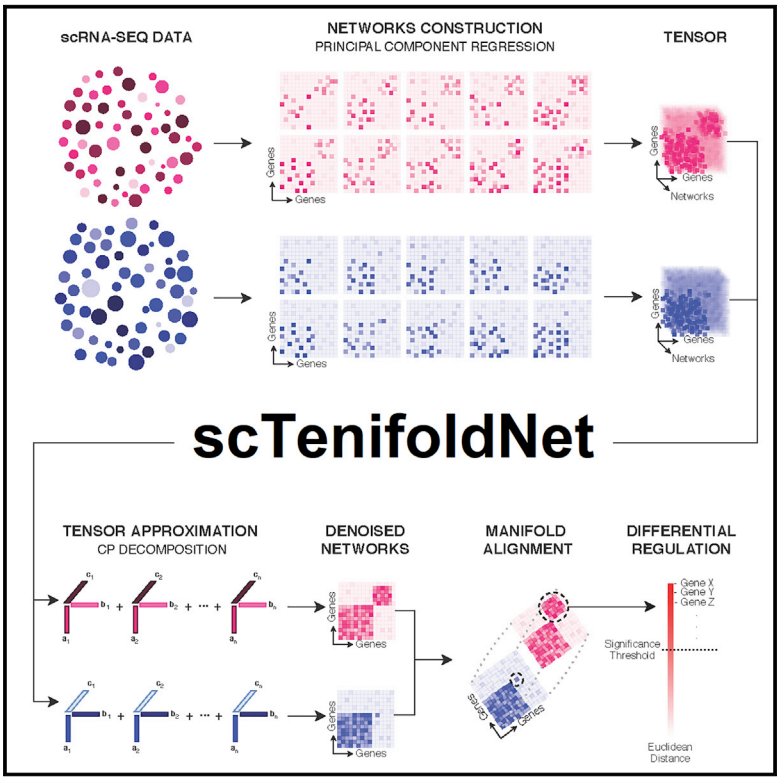
Descriptor

Patterns

scTenifoldNet: A Machine Learning Workflow for Constructing and Comparing Transcriptome-wide Gene Regulatory Networks from Single-Cell Data

Graphical Abstract



Authors

Daniel Osorio, Yan Zhong, Guanxun Li, Jianhua Z. Huang, James J. Cai

Correspondence

jianhua@stat.tamu.edu (J.Z.H.), jcai@tamu.edu (J.J.C.)

In Brief

scTenifoldNet is a machine learning workflow built upon principal-component regression, low-rank tensor approximation, and manifold alignment. It uses single-cell RNA sequencing data to construct single-cell gene regulatory networks (scGRNs) and compares scGRNs of different samples to identify differentially regulated genes. Real-data applications demonstrate that scTenifoldNet accurately detects specific signatures of gene expression relevant to the cellular systems tested.

Highlights

- scTenifoldNet is a machine learning tool for comparative single-cell network analysis
- It is built upon PC regression, tensor decomposition, and manifold alignment
- It constructs and compares gene regulatory networks from scRNA-seq data
- It accurately identifies differentially regulated genes between single-cell samples

Osorio et al., 2020, Patterns *1*, 100139
December 11, 2020
https://doi.org/10.1016/j.patter.2020.100139

CellPress

Patterns

CellPress
OPEN ACCESS

Descriptor

scTenifoldNet: A Machine Learning Workflow for Constructing and Comparing Transcriptome-wide Gene Regulatory Networks from Single-Cell Data

Daniel Osorio,[1] Yan Zhong,[2] Guanxun Li,[2] Jianhua Z. Huang,[2,*] and James J. Cai[1,3,4,5,*]
[1]Department of Veterinary Integrative Biosciences, Texas A&M University, College Station, TX 77843, USA
[2]Department of Statistics, Texas A&M University, College Station, TX 77843, USA
[3]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843, USA
[4]Interdisciplinary Program of Genetics, Texas A&M University, College Station, TX 77843, USA
[5]Lead Contact
*Correspondence: jianhua@stat.tamu.edu (J.Z.H.), jcai@tamu.edu (J.J.C.)
https://doi.org/10.1016/j.patter.2020.100139

THE BIGGER PICTURE Understanding the functions of genes requires the investigation of the structure of their regulatory networks of interactions. Single-cell RNA sequencing (scRNA-seq) brings new challenges and opportunities to the study of such networks. Here, we present a machine learning tool for constructing and comparing single-cell gene regulatory networks. Our algorithm, scTenifoldNet, can be used to identify differentially regulated genes between two scRNA-seq samples. It complements and enhances the commonly used differential expression analysis by revealing differences between samples in the regulatory relationships among genes, rather than the expression level. We anticipate that, by deciphering the complexity of data that surpasses human interpretative ability, scTenifoldNet can help achieve breakthroughs in understanding regulatory mechanisms underlying cell behaviors.

Proof-of-Concept: Data science output has been formulated, implemented, and tested for one domain/problem

SUMMARY

We present scTenifoldNet—a machine learning workflow built upon principal-component regression, low-rank tensor approximation, and manifold alignment—for constructing and comparing single-cell gene regulatory networks (scGRNs) using data from single-cell RNA sequencing. scTenifoldNet reveals regulatory changes in gene expression between samples by comparing the constructed scGRNs. With real data, scTenifoldNet identifies specific gene expression programs associated with different biological processes, providing critical insights into the underlying mechanism of regulatory networks governing cellular transcriptional activities.

INTRODUCTION

A gene regulatory network (GRN) is a graph depicting the intricate interactions between transcription factors (TFs), associated proteins, and their target genes, reflecting the physiological condition of the cells in question. The analysis of GRNs promotes the interpretation of cell states, cell functions, and regulatory mechanisms that underlie the dynamics of cell behaviors. Multiple methods have been developed to build GRNs from data of gene expression.[1–4] It is important to compare GRNs constructed using datasets from different samples because the comparison may reveal regulatory mechanisms leading to transcriptomic changes. In particular, the comparison results may help us understand what is the most significant shift in regulatory mechanisms between samples, as well as how genetic and environmental signals are integrated to regulate a cell population's physiological responses and how cell behavior is affected by various perturbations. All of these are key questions in the study of the functional participation of given GRNs. Despite the critical importance of comparative GRN analysis, relatively few methods have been established to compare GRNs.[5]

Single-cell RNA-sequencing (scRNA-seq) technology has been revolutionizing the biomedical sciences in recent years. New research provides an unparalleled degree of precision in

analyzing transcriptional regulation, cell history, and cell interactions with rich knowledge. It transforms previous entirely tissue-based assays into transcriptomic single-cell measurements and greatly enhances our understanding of cell development, homeostasis, and disease. Current scRNA-seq systems (e.g., 10× Genomics) can profile transcriptomes for thousands of cells per experiment. The sheer number of measured cells can be leveraged to construct GRNs. Advanced computational methods can facilitate such an effort to reach unprecedented resolution and accuracy, revealing the network state of given cells.[6–8] Furthermore, comparative analyses among GRNs of different samples will be extremely powerful in revealing fundamental changes in regulatory networks and unraveling the transcriptional programs that govern the behaviors of cells. Since our ability to generate scRNA-seq data has outpaced our ability to extract information from it, there is a clear need to develop effective computational algorithms and novel statistical methods for analyzing and exploiting information embedded within GRNs.[9]

Constructing single-cell GRNs (scGRNs) using data from scRNA-seq and then effectively comparing constructed scGRNs presents significant analytical challenges.[9,10] A meaningful comparison of scGRNs first requires a robust construction of a GRN from scRNA-seq data. Comparing scGRNs built via an unstable solution would cause misleading results and inappropriate conclusions. The vast number of different cellular states in a sample and the technical and biological noise, as well as the sparsity of scRNA-seq data, complicate the process of scGRN construction. Often, the expression of a gene is governed by stochastic processes and also influenced by transcriptional activities of many other genes. Thus, it is difficult to tease out subtle signals and infer true connections between genes. Furthermore, a direct comparison between two scGRNs is difficult; e.g., comparing each edge of the graph between scGRNs would be ill powered when scGRNs involve thousands of genes. Taken together, the key challenge in conducting comparative scGRN analysis is to extract meaningful information from noisy and sparse scRNA-seq data, since the information is deeply embedded in the differences between highly complex scGRNs of two samples.

In this paper, we introduce a workflow for constructing and comparing scGRNs using data from scRNA-seq of different samples. The workflow, which we call scTenifoldNet, is built upon several machine learning algorithms, including principal-component (PC) regression, low-rank tensor approximation, and manifold alignment. Through several examples, we show that scTenifoldNet is a sensitive tool to detect specific changes in gene expression signatures and the regulatory network rewiring events. The input of scTenifoldNet is a pair of expression matrices from scRNA-seq of two different samples. For instance, one sample may come from a healthy donor and the other from a diseased donor. In scTenifoldNet, the two input expression matrices are simultaneously processed through a multistep procedure. The final output is a list of ranked genes, sorted according to the differential regulation level of each gene. The ranked gene list can be used to perform functional enrichment analysis to detect the enriched molecular functions and involved biological processes. The constructed scGRN can also be used to identify functionally significant modules, i.e., subsets of tightly regulated genes.

scTenifoldNet includes an innovative method for comparing two scGRNs. We are not aware of any prior work using a similar design to achieve the same analytical goal. scTenifoldNet overcomes several methodological challenges, resulting in an effective and efficient scGRN comparison method. Here, we first benchmark and demonstrate the utility of scTenifoldNet across synthetic datasets and then apply scTenifoldNet to real datasets. Our real data analyses showed scTenifoldNet's power in identifying significant genes and network modules whose regulatory patterns shift greatly between samples. Some of these findings have not been reported in the respective original studies in which the datasets were generated.

RESULTS

The scTenifoldNet Architecture

To enable comparative scGRN analysis in a robust and scalable manner, we base our method on a series of machine learning methods. A key challenge of our comparative analysis is to extract meaningful differences in regulatory relationships between two samples from noisy and sparse data. Specifically, we seek to contrast scGRNs constructed from different scRNA-seq expression matrices. Figure 1 shows the main components of scTenifoldNet architecture. The whole workflow contains five key steps: subsampling cells, constructing multilayer scGRNs, denoising, manifold alignment, and differential regulation (DR) test. To produce biologically meaningful results, we made dedicated design decisions for the task in each of these steps. Next, we briefly describe the numerical methods implemented in scTenifoldNet. More technical details are presented in the Experimental Procedures.

Numerical Methods

The numerical methods used to construct and compare scGRNs involve the following five steps:

Step 1. Pre-processing Data and Subsampling Cells. The input data are two scRNA-seq expression data matrices, $\boldsymbol{X}$ and $\boldsymbol{Y}$, containing expression values for n genes in m_1 and m_2 cells from two different samples. Next, m cells in $\boldsymbol{X}$ and $\boldsymbol{Y}$ are randomly sampled to form $\boldsymbol{X}'$ and $\boldsymbol{Y}'$. This subsampling process is repeated t times to create two collections of subsampled cells, $\{\boldsymbol{X}'_i\}$ and $\{\boldsymbol{Y}'_i\}$, where $i = 1, 2, \ldots, t$.

Step 2. Constructing Initial Networks. For each $\boldsymbol{X}'_i \in \{\boldsymbol{X}'_i\}$, $i = 1, 2, \ldots, t$, PC regression is used to construct a GRN. The constructed GRN from $\boldsymbol{X}'_i$ is stored as a weighted graph represented with an $n \times n$ weighted adjacency matrix $\boldsymbol{X}'_i$. Similarly, for each $\boldsymbol{Y}'_i \in \{\boldsymbol{Y}'_i\}$, $i = 1, 2, \ldots, t$, we construct a GRN and represent it with an $n \times n$ weighted adjacency matrix $\boldsymbol{W}^y_i$. Diagonal values of each adjacency matrix are set to zeros, and other values are normalized by dividing by their maximal absolute value. Each normalized adjacency matrix is then filtered by retaining only the top 5% of edges ranked using the absolute edge weight, resulting in a sparse adjacency matrix.

Step 3. Denoising. Tensor decomposition[11] is used to denoise the adjacency matrices obtained in Step 2. The collection of t scGRNs for each sample, $\{\boldsymbol{W}^x_i\}$ or $\{\boldsymbol{W}^y_i\}$, is processed separately as a third-order tensor, denoted as $\mathfrak{T}^x$ or $\mathfrak{T}^y$, each containing $n \times n \times t$ elements. The CANDECOMP/PARAFAC (CP) decomposition is applied to decompose $\mathfrak{T}^x$ and $\mathfrak{T}^y$ into components. Next, $\mathfrak{T}^x$ and $\mathfrak{T}^y$ are reconstructed using the top r

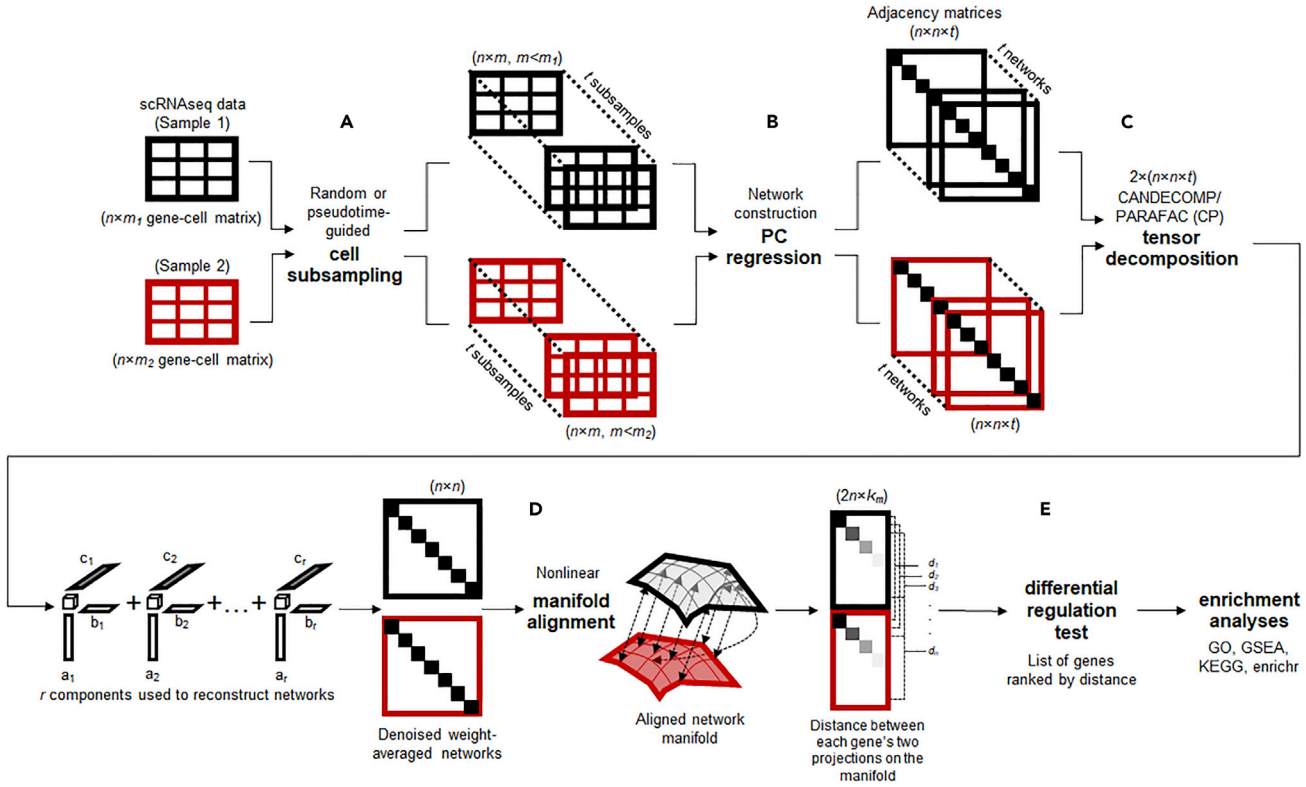


Figure 1. Overview of the scTenifoldNet Workflow

scTenifoldNet is a machine learning framework that uses a comparative network approach with scRNA-seq data to identify regulatory changes between samples. scTenifoldNet is composed of five major steps.

(A) Cell subsampling. scTenifoldNet starts with subsampling cells in the scRNA-seq expression matrices. When two samples are analyzed, each of the two samples is subsampled either randomly or following a pseudotime trajectory of cells. The subsampling is repeated multiple times to create a series of subsampled cell populations, which are subject to network construction and form a multilayer scGRN.

(B) Network construction. PC regression is used for scGRN construction; each scGRN is represented as a weighted adjacency matrix.

(C) Tensor denoising. Two samples produce two multilayer GRNs and form two three-order tensors, which are subsequently decomposed into multiple components. The top components of tensor decomposition are then used to reconstruct two denoised multilayer scGRNs. Then, two denoised multilayer scGRNs are collapsed by taking the average weight across layers.

(D) Manifold alignment. The two single-layer average scGRNs are then aligned with respect to common genes using a nonlinear manifold alignment algorithm. Each gene is projected to a low-rank manifold space as two data points, one from each sample.

(E) Differential regulation test. The distance between the two data points is the relative difference of the gene in its regulatory relationships in the two scGRNs. Ranked genes are subject to tests for their significance in differential regulation between scGRNs.

components to obtain denoised tensors: $\mathfrak{T}_d^x$ and $\mathfrak{T}_d^y$. Denoised $\{\boldsymbol{W}_i^x\}$ and $\{\boldsymbol{W}_i^y\}$ in $\mathfrak{T}_d^x$ and $\mathfrak{T}_d^y$ are collapsed by taking the average of edge weights for each edge to form two denoised, averaged matrices, $\boldsymbol{W}_d^x$ and $\boldsymbol{W}_d^y$, which are subsequently normalized as in step 2 and then symmetrized.

Step 4. Aligning Genes onto a Manifold. The weighted adjacency matrices $\boldsymbol{W}_d^x$ and $\boldsymbol{W}_d^y$ are regarded as two similarity matrices for a nonlinear manifold alignment procedure. The alignment is done by solving an eigenvalue problem with a Laplacian matrix derived from the joint matrices, $\boldsymbol{W} = [\boldsymbol{W}_d^x, \lambda \boldsymbol{I}/2; \lambda \boldsymbol{I}^T/2, \boldsymbol{W}_d^y]$, where λ is a tuning parameter and $\boldsymbol{I}$ is the identity matrix that reflects the binary correspondence between genes in the samples, $\boldsymbol{X}$ and $\boldsymbol{Y}$. As the result of manifold alignment, all genes in the samples, $\boldsymbol{X}$ and $\boldsymbol{Y}$, are projected on a shared, low-dimensional manifold with a dimension $k_m << n$. The projections of each gene j from the samples, $\boldsymbol{X}$ and $\boldsymbol{Y}$, are two k_m-dimensional vectors, F_j^x and F_j^y.

Step 5. Ranking Genes. For each gene j, let d_j be the Euclidean distance between the gene's two projections F_j^x and F_j^y on the shared manifold: one is from the sample $\boldsymbol{X}$, and the other is from the sample $\boldsymbol{Y}$. Genes are sorted according to this distance. The greater the distance, the greater the regulatory shift.

In the following sections, we explain the rationale behind each step of scTenifoldNet, as well as the selection of machine learning algorithms and implementation details.

Subsampling of Cells

The rationale for randomly subsampling cells is close to that of ensemble learning. Ensemble learning is a technique where multimodel decisions are merged to improve overall performance. Similarly, instead of attempting to build a single scGRN, scTenifoldNet randomly samples subsets of cells from the given scRNA-seq expression matrix and builds a series of "low-precision" scGRNs with the subsampled datasets. These low-precision scGRNs are then combined to obtain a "high-precision"

scGRN. As mentioned above, current scRNA-seq technology can produce the transcriptome profiles of thousands of cells from each sample. It is fundamentally challenging to process high-dimensionality and large-scale scRNA-seq data, especially since there can be a substantial variation among cells. This happens even in a group of highly homogeneous cells of the same type.[12] The presence of so-called outlying cells, i.e., cells showing profiles of expression that deviate from those of most other cells, may influence the construction of high-precision scGRNs. Therefore, subsampling offers promise as a technique for handling the noise in the input datasets. When the number of cells is small, the input data matrix may be resampled with replacement.[13]

Constructing scGRNs Using PC Regression

Although many GRN construction methods have been developed,[1,2,4] it is unclear which one is suitable for constructing multiple large scGRNs from the subsampled data.[9] When dealing with multiple sets of input data, both the accuracy and the computational efficiency of these algorithms have to be considered. After conducting a thorough review of existing methods, we opted to adopt PCNet,[5] a method of network construction using PC regression.[14] The PC regression method extracts the first few (e.g., $k = 3$) PCs and then uses these components as the predictors in a linear regression model fitted using ordinary least squares. The values of the transformed coefficients of genes are treated as the strength and regulatory effect between genes to generate the network. The utilization of PC regression in scTenifoldNet lies in its ability to surpass the multicollinearity problem that arises when two or more explanatory variables are linearly correlated.

Denoising via Low-Rank Tensor Approximation

Removing the noise from constructed scGRNs is an important step of scTenifoldNet. Here the term "noise" is used in a broad sense to refer to any outlier or interference that is not the quantity of interest, i.e., the true regulatory relationship between genes. For each sample, the multilayer scGRN constructed from multiple subsampled datasets is regarded as a rank 3 tensor. To reduce the noise in the multilayer scGRN, we decompose the tensor and reconstruct the multilayer scGRN using leading components. The idea is similar to that of denoising using truncated singular value decomposition (SVD). After cutting a larger portion of the noise spread over the lowest singular value components, the reconstructed data matrix based on the truncated SVD would, therefore, represent the original data with reduced noise. Indeed, tensor decomposition has been used in video data analyses for denoising and information-extracting purposes.[15] It has also been used to impute missing data.[16] Using the CP algorithm,[17] we factorize the two multilayer scGRNs separately and regenerate all adjacency matrices using leading components. The number of components used for reconstruction can be specified and is set to 3 by default. In the real data applications, we find the tensor GRN regeneration serves two purposes, denoising and enhancing, i.e., making main signals stronger and making less important signals weaker.

Manifold Alignment of Two scGRNs

For a gene, its position in one of the two scGRNs (i.e., denoised adjacency matrices from the two samples) is determined by its regulatory relationships with all other genes. Here we regard each gene as a data point in a high-dimensional space where components of the data point are the features, i.e., weights between the gene and all other genes in the scGRN adjacency matrix. To compare the same gene's positions in the two scGRNs, we first align the two scGRNs. To do so, we take a popular and effective approach for processing high-dimensional data, intuitively modeling the intrinsic geometry of the data as being sampled from a low-dimensional manifold, commonly referred to as the manifold assumption.[18] This assumption essentially means that local regions in the data can be mapped to low-dimensional coordinates, while the nonlinearity and high dimensionality in the data come from the curvature of the manifold. Manifold alignment produces projections between sets of data, given that the original datasets lie on a common manifold.[19–22] Manifold alignment matches the local and nonlinear structures among the data points from multiple sources and projects them to the same low-dimensional space while maintaining their local manifold structure of each source. The ability to flexibly learn and accurately represent the structure in the data with manifold alignment has been demonstrated in applications in automatic machine translation, face recognition, and so on.[23,24] Here, we use manifold alignment to match genes in the two denoised scGRNs, one from each sample, to identify cross-network linkages. Consequently, the information of genes stored in two scGRNs is aligned, meaning points close together in the low-dimensional space are more similar than points that are farther apart.

Ranking Genes and Reporting Differentially Regulated Genes

To identify genes whose regulatory status differs between the two samples, we calculate the distance between projected data points in the manifold alignment subspace. For each gene, if the gene appears in scGRNs of both samples, there are two data points for the same gene, one from each sample. We compute the Euclidean distance between the two data points of the gene and use the distance to measure the dissimilarity in the gene's regulatory status in two scGRNs.[25] We do this for all genes shared between two samples and then rank the genes by the distance. The larger the distance, the more different the gene in two samples. In this way, we obtain a list of ranked genes. These ranked genes are subject to functional annotation, such as by using the pre-ranked gene set enrichment analysis (GSEA)[26] to assess the enriched functions associated with top genes. To avoid choosing the number of selected genes arbitrarily, we compute p values for genes using χ^2 tests, adjust the p values with a multiple testing correction, and select significant genes using a 5% false discovery rate (FDR) cutoff.

Benchmarking the Performance of scTenifoldNet Using Simulated Data

Precision and Recall of the Network Construction Method Adopted in scTenifoldNet

PC regression is the method we adopted for scTenifoldNet to construct scGRN. It is important to ensure that scTenifoldNet/PC regression is an effective and efficient network construction method for our purpose.

To this end, we conducted a systematic comparison between network construction algorithms using a published evaluation

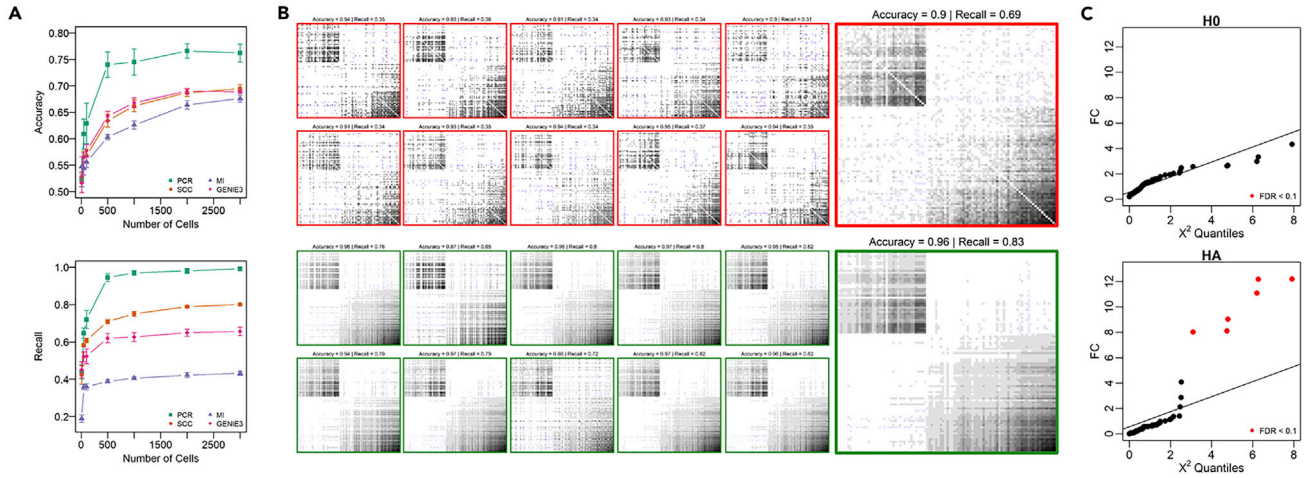


Figure 2. Benchmarking the Performance of scTenifoldNet Using Simulated Data

(A) The accuracy and recall of scGRN construction using different methods, PC regression, SCC, MI, and GENIE3, as functions of the number of cells used in the analysis. Error bar is the SD of the computed values after 10 bootstrapped evaluations. PCR, PC regression; SCC, Spearman's correlation coefficient; MI, mutual information; GENIE3, a random-forest-based network construction method.

(B) Visualization of the effect of tensor denoising on accuracy and recall of multilayer scGRNs. Each subpanel is a heatmap of a 100 × 100 adjacency matrix constructed using PC regression over the counts of 500 randomly subsampled cells. Gray scale indicates the relative strength of regulatory relationships between genes. Top part includes networks before tensor denoising (adjacency matrices in heatmap with red box); bottom part includes corresponding networks after tensor denoising (adjacency matrices in heatmap with green box). In each part, adjacency matrices of networks of 10 subsamples (10 small heatmaps) and their average adjacency matrix (one large heatmap) are shown.

(C) Evaluation of the sensitivity of scTenifoldNet in identifying punctual changes in the regulatory profiles. Top: evaluation of the original data matrix against itself. Bottom: evaluation of the original matrix against the perturbed matrix. Significant genes identified using the differential regulation test (FDR <0.1, B-H correction) are indicated in red. All significant genes are perturbed in simulation and thus are expected to be identified.

tool package called BEELINE.[10] We benchmarked scTenifoldNet/PC regression and compared it with 11 other algorithms (see Experimental Procedures). We chose to reuse a reference dataset called the gonadal sex determination (GSD) in the BEELINE package to perform the benchmarking. In the BEELINE package, GSD is the largest curated reference dataset, and it contains 19 genes and 2,000 cells. We compared different algorithms jointly using the area under the precision-recall curve (AUPRC), area under the receiver operating characteristic curve (AUROC), and computation time, and found that scTenifoldNet/PC regression and partial information decomposition and context (PIDC)[27] outperformed other algorithms (see Figure S1 for details).

We also simulated scRNA-seq data using a parametric method with a predefined scGRN model (see Experimental Procedures for details).[28] With the simulated data, which contain 100 genes and up to 3,000 cells, we compared constructed scGRNs against the ground truth (i.e., the simulated scGRN) to estimate the accuracy of reconstruction. We tested the accuracy of scTenifoldNet/PC regression against methods based on Spearman's correlation coefficient (SCC) and mutual information (MI)[1] and on GENIE3.[2] The SCC and MI methods are computationally efficient, whereas GENIE3 is not, but GENIE3 is the top-performing method for network inference in the DREAM challenges.[3] For each method, their performance in recovering gene regulatory relationships was compared with the ground-truth interactions between genes, which were generated according to pre-setting parameters. We found that scTenifoldNet/PC regression produced more specific (better accuracy) and more sensitive (better recall) scGRNs than other methods (Figure 2A). This is true across a wide range of settings of cell numbers in input scRNA-seq expression matrices. scTenifoldNet/PC regression is also much faster than GENIE3 (running time information is available in Table S1).

A limitation of our simulation-based evaluation is that simulated scGRNs are much simpler than GRNs in reality, which may contain hundreds or thousands of genes. It is a challenge to simulate such a realistic GRN and, for GRN inference algorithms, to figure out the key regulators and their targets. In this study, we chose to apply our method directly to real datasets and evaluate the biological relevance of results, rather than explore the impact of the size and diversity of synthetic GRNs on the results.

Effect of Denoising with Tensor Decomposition

To show the effect of tensor denoising, we simulated scRNA-seq data (see Experimental Procedures) and processed the data using the first two steps of scTenifoldNet, i.e., cell subsampling followed by the construction of scGRNs using PC regression. We subsampled 500 cells each time and generated 10 scGRNs. The 10 scGRNs were treated as a multilayer network or a tensor to be denoised. For each scGRN, we kept the top 20% of the links. The presence and absence of links in each scGRN were compared with those in the simulated, ground-truth scGRN to estimate the accuracy of recovery and the rate of recall. Figure 2B contains the heatmaps of adjacency matrices of the 10 scGRNs before and after denoising (small heatmaps). We also show two collapsed scGRNs (Figure 2B, large heatmaps), which were generated by averaging link weights across the 10 scGRNs

before and after denoising. These results illustrate the ability of scTenifoldNet to denoise multilayer scGRNs. For instance, tensor denoising improves the recall rate of regulatory relationships between genes by 25%. This simulation study suggests that tensor denoising could be useful for removing the impacts of random dropout and other noise issues affecting the scGRN construction using scRNA-seq data.

Detecting Power Illustrated with a Simulated Dataset

We used simulated data to show the capability of scTenifoldNet in detecting differentially regulated genes. We first used the negative binomial distribution to generate a sparse synthetic scRNA-seq dataset (an expression matrix including 67% zeros in its values). This toy dataset includes 2,000 cells and 100 genes. We called it sample 1. We then duplicated the expression matrix of sample 1 to make sample 2. We modified the expression matrix of sample 2 by swapping expression values of three randomly selected genes with those of another three randomly selected genes. Thus, the differences between samples 1 and 2 are restricted in these six genes. Using scTenifoldNet with the default parameter setting, we compared the originally generated expression matrix (sample 1) with itself (sample 1 versus sample 1) and also with the manually perturbed version (sample 1 versus sample 2). As expected, when comparing the original matrix against itself, none of the genes was identified to be significant. However, when samples 1 and 2 were compared, the six genes whose expression values were swapped were identified as significant differentially regulated genes (Figure 2C, FDR <0.1). These results are expected and support the sensitivity of scTenifoldNet in identifying subtly shifted gene expression programs.

Real Data Analyses

Practical Considerations of Real Data Analysis Using scTenifoldNet

First of all, we address several practical questions regarding the application of scTenifoldNet to real scRNA-seq data. (1) What are the input expression matrices to be compared? The input to scTenifoldNet is two matrices of gene expression values (e.g., unique molecular identifier [UMI] counts) as measured in two samples to be compared. In each matrix, columns represent cells, and rows represent genes. We assume that each input matrix contains a sizable number of cells. For example, a typical input matrix may contain UMI counts for 5,000 genes and 2,000 cells. Whether a gene is expressed among cells can be determined by examining if this gene has a nonzero UMI count in more than 5% of cells. Scaling normalization (e.g., the library size normalization) of the input UMI count matrix does not seem to affect the construction of scGRNs (Figure S2). In contrast, imputing the UMI count matrix using an imputation algorithm (e.g., MAGIC)[29] may have an impact on the performance of scTenifoldNet (Figure S3). (2) How does scTenifoldNet handle cell heterogeneity? Heterogeneity in expression among cells is inevitable. scTenifoldNet is designed to tolerate a certain level of such heterogeneity as long as the cells are of the same type. scTenifoldNet is not a data preparation tool. It also does not perform any clustering analysis for cells; it does not assign cells to cell types. We assume all cells in both input matrices are of the same type. Otherwise, the results would be difficult to interpret. To solve this problem, a specific tool (to prioritize cell types most responsive to biological perturbations) has been developed elsewhere.[30] (3) What if the number of cells is too small? We expect that each input matrix contains a sizable number of cells (e.g., n > 2,000). If this is the case, the jackknife method (subsampling without replacement) is adapted by default: m = 500 cells are subsampled each time. Alternatively, an m-out-of-n bootstrap method (subsampling with replacement) can be used.[13] When the number of cells is small (e.g., n = 500), a full bootstrap method can be used, i.e., resampling 500 cells each time out of 500 given cells with replacement.[13,31] scTenifoldNet is robust against unbalanced cell numbers in the two samples for comparison (Figure S4). (4) What is the relationship between scTenifoldNet analysis and differential expression (DE) analysis? scTenifoldNet analysis should be used as a complementary analysis method in addition to DE analysis, rather than replacing DE analysis. DE analysis (using, e.g., MAST,[32] edgeR,[33] or SCDE)[34] is still a widely used method for understanding the difference between two scRNA-seq samples.[35] scTenifoldNet is designed based on a principle different from that underlying DE analysis. Thus, the results of scTenifoldNet analysis and DE analysis are not supposed to be compared side by side. It is not uncommon that scTenifoldNet and DE analyses report the same genes to be significant. This is because the change in the regulatory pattern of a gene in scGRNs may be associated with the change in the gene's expression level. To evaluate the influence of gene expression level on scGRN construction, we calculated the correlation between the average gene expression level and the average weighted degree of nodes in scGRNs, which are constructed using scTenifoldNet/PC regression and other algorithms in the BEELINE package.[10] If the weighted degree of nodes in an scGRN constructed using a method is correlated with the expression level of genes, then it indicates that the method is likely to be biased toward highly expressed genes during the process of scGRN construction. We found that all evaluated algorithms produced results showing a certain level of such a correlation (Figure S5). However, compared with all other algorithms, scTenifoldNet/PC regression produced the smallest correlation value and thus is most robust against the bias toward highly expressed genes.

Analysis of Transcriptional Responses of Neurons to Acute Morphine Treatment

To illustrate the use of scTenifoldNet, we first applied scTenifoldNet to an scRNA-seq dataset from Avey and colleagues[36] This is a study on transcriptional responses of mouse neural cells to morphine (Figure 3A). In the study, Avey and colleagues performed scRNA-seq experiments with the nucleus accumbens of mice after 4 h of morphine treatment and used mice treated with saline as mock controls. Single-cell expression data were obtained for 11,171 and 12,105 cells from four morphine- and four mock-treated mice, respectively.[36] The measured cells were clustered to identify neurons (7,972 and 8,912 from morphine- and mock-treated samples, respectively); the identified neurons were then subgrouped into 11 clusters, including major clusters of D1 and D2 medium spiny neurons (MSNs), comprising ~95% of the neurons in the nucleus accumbens. Using DE analysis implemented in SCDE,[34] Avey et al. identified several hundred genes that are differentially expressed between

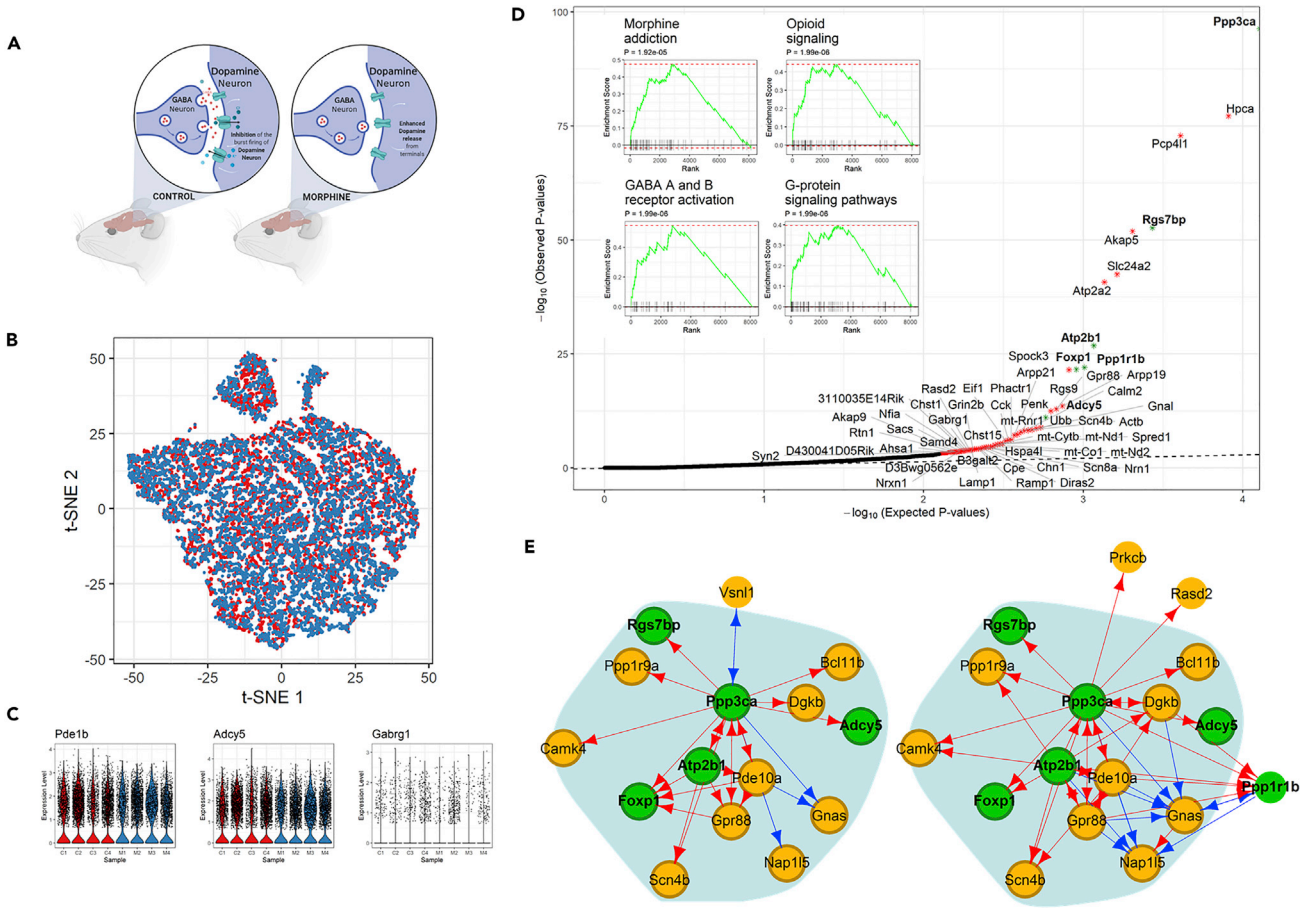


Figure 3. Analysis of Transcriptional Responses to Morphine in Mouse Cortical Neurons

(A) Illustration of experimental design and data collection of the morphine response study.[36]

(B) t-SNE plot of 7,972 and 8,912 neurons from morphine-treated (blue) and mock-treated (red) mice, respectively.

(C) Violin plots showing the log-normalized expression levels of representative differentially regulated and/or differentially expressed genes in four (M) morphine- and four (C) mock-treated mice.

(D) Quantile-quantile (Q-Q) plot for observed and expected p values of the 8,138 genes tested. Genes (n = 65) with FDR <0.1 are shown in red; genes (n = 56) with FDR <0.05 are labeled with an asterisk. Inset shows results of the GSEA for genes ranked by their distances in manifold aligned scGRNs from morphine- and mock-treated mice.

(E) The module enriched with differentially regulated genes and the corresponding subnetworks in two scGRNs. For illustrative purposes, the module is centered on the differentially regulated gene *Ppp3ca*. Significantly differentially regulated genes (FDR <0.05) in the module are highlighted in green. Edges are color-coded: red indicates a positive association, and blue indicates negative. Weak edges are filtered out by thresholding for clear visualization, and the background shadow indicates the shared portion of the module in the two scGRNs.

morphine- and mock-treated samples (Table S2 of Avey et al.).[36] Although this result is intriguing, we argue that it seems that when so many genes are identified as "significant players," it is difficult to interpret the result and to pinpoint the specific regulatory mechanism underlying the true response. Indeed, instead of functional enrichment analysis with identified differentially expressed genes, the subsequent analyses in the study of Avey and coworkers[36] were refocused on a tiny portion of D1 MSNs, called activated MSNs. It was only when activated MSNs were compared with all other D1 MSNs that 256 differentially expressed genes were identified (SCDE, $p < 0.001$, Table S2 of Avey et al.).[36] These genes were then found to be associated with several terms related to opioid addiction, including morphine dependence and opioid-related disorders (Table S3 of Avey et al.).[36] In the morphine-treated sample, less than 4.5% of D1 MSNs were activated MSNs; in the mock-treated sample, less than 2% (see Figure S2B of Avey et al.).[36] In view of this, we point out here that while relevant signals can be detected using traditional DE analysis, the analytical method involves extensive human intervention; i.e., an iterative clustering procedure is needed to identify a final population of cells (in this case, activated MSNs). The cell population size is small, making the analysis result potentially variable.

We were motivated by these considerations and set out to reanalyze the data. We first reproduced the results of the DE analysis. We found that the mock- and morphine-treated neurons indeed exhibited a striking similarity. For example, mock- and morphine-treated neurons are indistinguishable in a t-distributed

stochastic neighbor embedding (t-SNE) plot (Figure 3B); expression levels of several known morphine-responsive genes, e.g., *Adcy5*, *Ppp1r1b*, and *Ppp3ca*, show no difference (Figure 3C). Thus, a direct comparison of gene expression between neurons using the DE method may have limited power to identify relevant genes involved in the morphine response.

Next, using scTenifoldNet, we identified 56 genes showing significant differences in their transcriptional regulation between mock- and morphine-treated neurons (Table S2). Compared with other genes, these genes have a significantly greater distance between their positions in two scGRNs aligned into the manifold (FDR <0.05, χ^2 test with Benjamini-Hochberg (B-H) multiple test adjustment, see Experimental Procedures for details). GSEA showed that these differentially regulated genes are enriched for *opioid signaling*, *signaling by G protein-coupled receptors*, *reduction of cytosolic calcium levels*, and *morphine addiction* (Figure 3D, inset, see also Table S3). It is known that morphine binds to the opioid receptors on the neuronal membrane. The signal is then transmitted through the G-protein-signaling system, inhibiting the adenylyl cyclase in the cytoplasm and decreasing the levels of cAMP and calcium-channel conduction.[37–39] Furthermore, 21 of 56 (38%) identified differentially regulated genes were found to be targets of *RARB* (adjusted $p < 0.01$, Enrichr enrichment test based on the chromatin immunoprecipitation sequencing [ChIP-seq] data).[40] *RARB* plays a role in synaptic transmission in dopaminergic neurons and the adenylate cyclase-activating dopamine receptor signaling pathway.[41,42] Thus, these enriched functions are relevant to the morphine stimulus, which is known to induce the disinhibition of dopaminergic neurons by GABA transmission, enhance dopamine release, and cause addiction.[43,44] Using the constructed scGRN, we were able to trace differentially regulated genes back to their topological positions in the network and examine their interacting genes. Figure 3E shows such a network module, including multiple differentially regulated genes.

In this case, scTenifoldNet was used as an unsupervised tool, which needs no human interference to operate. This feature is critical when referring to this specific set of data because where the signal is limited to rare types of cells, there is a chance that a less sensitive approach would miss the signal, especially when human interference is not provided. It is ideal to have an unsupervised tool that is sensitive to signals and, at the same time, robust to variation between cells. We note that scTenifoldNet is a tool different from conventional DE analysis tools: —scTenifoldNet reported fewer differentially regulated genes, in terms of the number of genes, compared with differentially expressed genes identified in the original study.[36] Among the 56 differentially regulated genes that scTenifoldNet detected, 11 (*Actb*, *Adcy5*, *Akap9*, *D430041D05Rik*, *Eif1*, *Pcp4l1*, *Penk*, *Phactr1*, *Rasd2*, *Scn4b*, and *Ubb*) are among the 256 differentially expressed genes reported in Table S2 of Avey and colleagues[36] The number of overlapping genes is not significantly higher than expected by random according to a hypergeometric test ($p = 0.29$) with a total of 1,432 genes (from Table S2 of Avey et al.)[36] included in the test. Figure 3C shows expression levels of three representative genes, *Pde1b*, *Adcy5*, and *Gabrg1*, in neurons from mock- and morphine-treated mice. All three genes are known to be involved in the morphine response,[45–47] but only when DE and DR tests were applied jointly were all three genes identified: *Pde1b* is a differentially expressed but not a differentially regulated gene, *Adcy5* a differentially regulated and a differentially expressed gene, and *Gabrg1* a differentially regulated but not a differentially expressed gene.

Analysis of Transcriptional Responses of a Carcinoma Cell Line to Cetuximab

To further illustrate the power of scTenifoldNet in identifying genes associated with specific perturbations, we applied scTenifoldNet to another published set of scRNA-seq data.[48] In this study,[48] Kagohara et al. use scRNA-seq to study mechanisms underlying the development of resistance to cetuximab in head and neck squamous cell carcinoma (HNSCC) (Figure 4A). Cetuximab is a human-murine chimeric monoclonal antibody used to treat metastatic colorectal cancer, metastatic non-small cell lung cancer, and head and neck cancer. In conjunction with radiotherapy, cetuximab improves the objective response rate in first-line treatment of recurrent or metastatic squamous cell carcinoma of the head and neck.[49] Cetuximab binds to the extracellular domain of the epidermal growth factor receptor (EGFR) on both normal and tumor cells.[50] EGFR is overexpressed in many cancers. Competitive binding of cetuximab to EGFR blocks the phosphorylation and activation of receptor-associated kinases and their downstream targets, e.g., MAPK, PI3K/Akt, and Jak/Stat pathways,[51] thereby reducing their effects on cell growth and metastatic spread. It is known that blocking EGFR activation also affects cellular processes such as apoptosis, cell growth, and vascular endothelial growth factor production.[52] Cetuximab is also known to cause degradation of the antibody-receptor complex and the downregulation of *EGFR1* expression.[53]

Kagohara et al. sequenced the transcriptome profile of cells before and after cetuximab treatment for 120 h in three different HNSCC cell lines: SCC1, SCC6, and SCC25.[48] They found that SCC6 is the most sensitive to the cetuximab treatment, reporting 8,389 genes as differentially expressed (including 4,166 upregulated and 4,223 downregulated ones with $p < 0.05$; Table S4 of Kagohara et al.).[48] Such a large number of differentially expressed genes makes it difficult to identify genes directly associated with the molecular mechanism through which cetuximab acts.

We extracted scRNA-seq data for 4,507 and 5,217 SCC6 cells treated with and without cetuximab, respectively (Figure 4B). Expression levels of three genes, *DuSP4*, *TIGA3*, and *LIF*, in cells of two treatment groups are shown in Figure 4C. All three genes are in the EGFR pathway. We used scTenifoldNet to reanalyze the data and identified 125 differentially regulated genes (FDR <0.05, Figure 4D and Table S4). These genes are enriched with those (39 of 125) that are under the regulation of TFs: *SMAD2* and *SMAD3*. GSEA showed that these differentially regulated genes are associated with the *EGFR1* pathway, *regulation of apoptosis*, *cell-cycle checkpoints*, *G1 cell-cycle arrest*, and *regulation of apoptosis* (Figure 4D inset, Table S5). Once again, scTenifoldNet identified a much smaller set of significant genes compared with those reported in the original paper:[48] 125 differentially regulated genes versus 8,389 differentially expressed genes. Nevertheless, functional analyses show that scTenifoldNet identified a more specific gene set relevant to cetuximab's mechanism of action. Further scrutinization of enriched molecular functions of these differentially regulated genes will help to

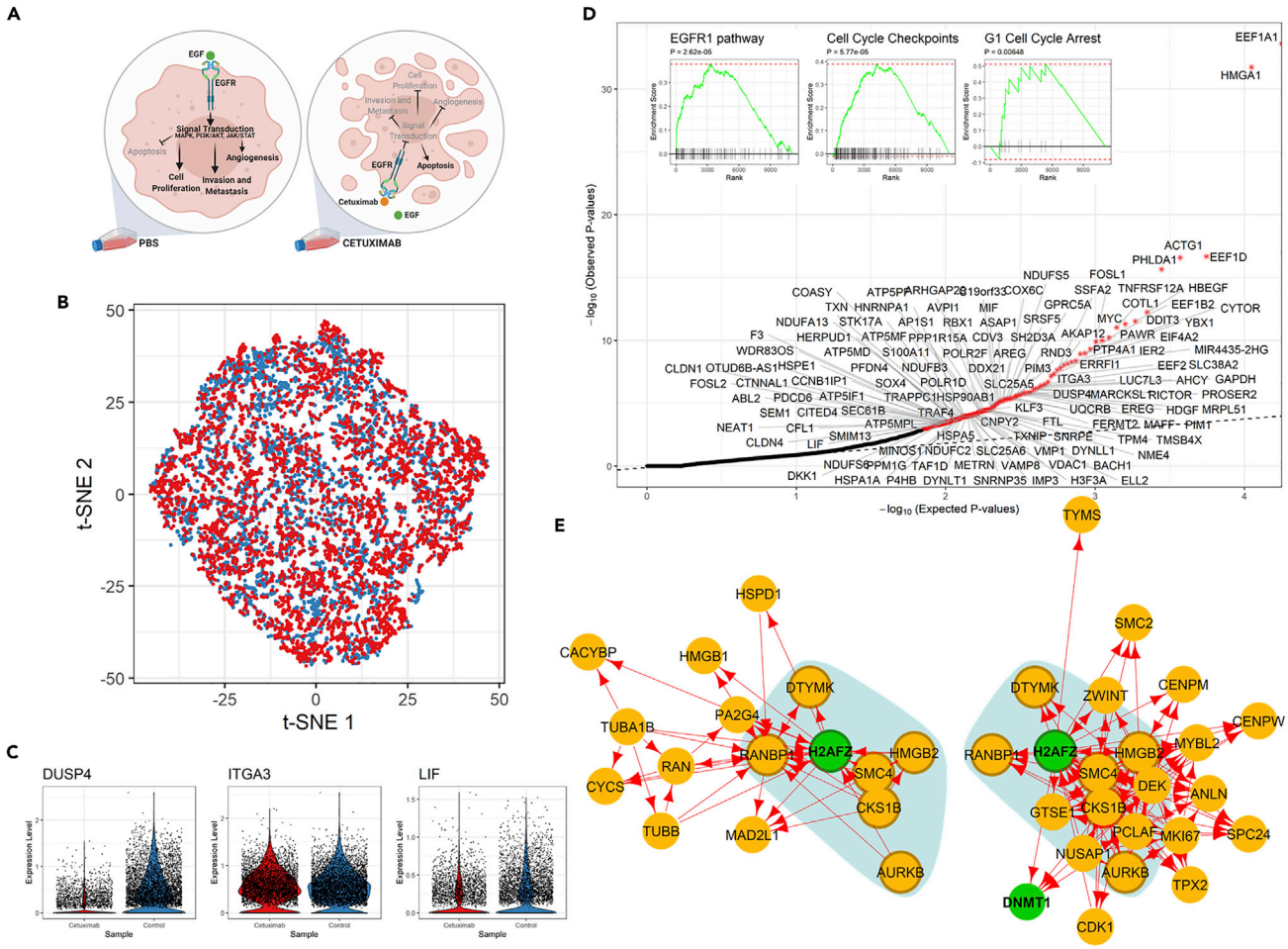


Figure 4. Analysis of Transcriptional Responses of a Carcinoma Cell Line to Cetuximab

(A) Illustration of experimental design, including sample groups and the known mechanism of drug action, in the study of cetuximab resistance of HNSCC cell lines.[48]

(B) t-SNE plot of 5,217 and 4,507 HNSCC-SCC6 cells treated with cetuximab (red) and PBS (blue), respectively.

(C) Violin plots showing the log-normalized expression levels of selected differentially regulated genes in SCC6 cells with and without cetuximab treatment.

(D) Q-Q plot for observed and expected p values of the 7,503 genes tested. Genes (n = 25) with FDR <0.05 are labeled with an asterisk. Inset shows the results of the GSEA for genes ranked by their distances in manifold aligned scGRNs from young and old mice.

(E) A representative module with differentially regulated genes and corresponding subnetworks in two scGRNs. The module is enriched with differentially regulated genes and the corresponding subnetworks in two scGRNs. For illustrative purposes, the module is centered on the differentially regulated gene *H2AFZ*. The colors, edges, and marks are presented as in Figure 3E.

identify more regulatory targets induced by cetuximab in HNSCC cells.

Analysis of Transcriptional Responses of Alveolar Type 1 Cells to Nkx2-1 *Gene Knockout*

In the third example, we applied scTenifoldNet to another published set of scRNA-seq data from type 1 alveolar (AT1) cells.[54] AT1 cells are responsible for gas exchange, the physiological function of the lung.[55] Little et al. found that NK homeobox 2-1 (*Nkx2-1*) is expressed in AT1 cells and thought *Nkx2-1* might be essential to the development and maintenance of AT1 cells.[54] To determine the function of NKX2-1 during the development of AT1 cells, they performed scRNA-seq experiments to obtain the transcriptome profile of cells from the lungs of *Nkx2-1*$^{CKO/CKO}$; *Aqp5*$^{Cre/+}$ mutant mice (i.e., knockout [KO] mice) and littermate controls (i.e., wild-type [WT] mice). They used early infant mice (postnatal day 10, P10), because P10 represents an intermediate time point when individual AT1 cells in the mutant lung are expected to collectively feature the full range of transcriptomic phenotypes. They reported 3,622 differentially expressed genes (2,105 upregulated and 1,517 downregulated, Supplementary Dataset S1 of Little et al.)[54] between the KO and the WT mice. Their analyses suggest that, without *Nkx2-1*, developing AT1 cells lose their molecular markers, morphology, and cellular quiescence, leading to aberrant expression of gastrointestinal (GI) genes, alveolar simplification, and lethality (Figure 5A).

To evaluate the power of scTenifoldNet in identifying regulatory changes caused by gene KO, we reanalyzed the transcriptional profiles of 2,397 mutant AT1 cells from the

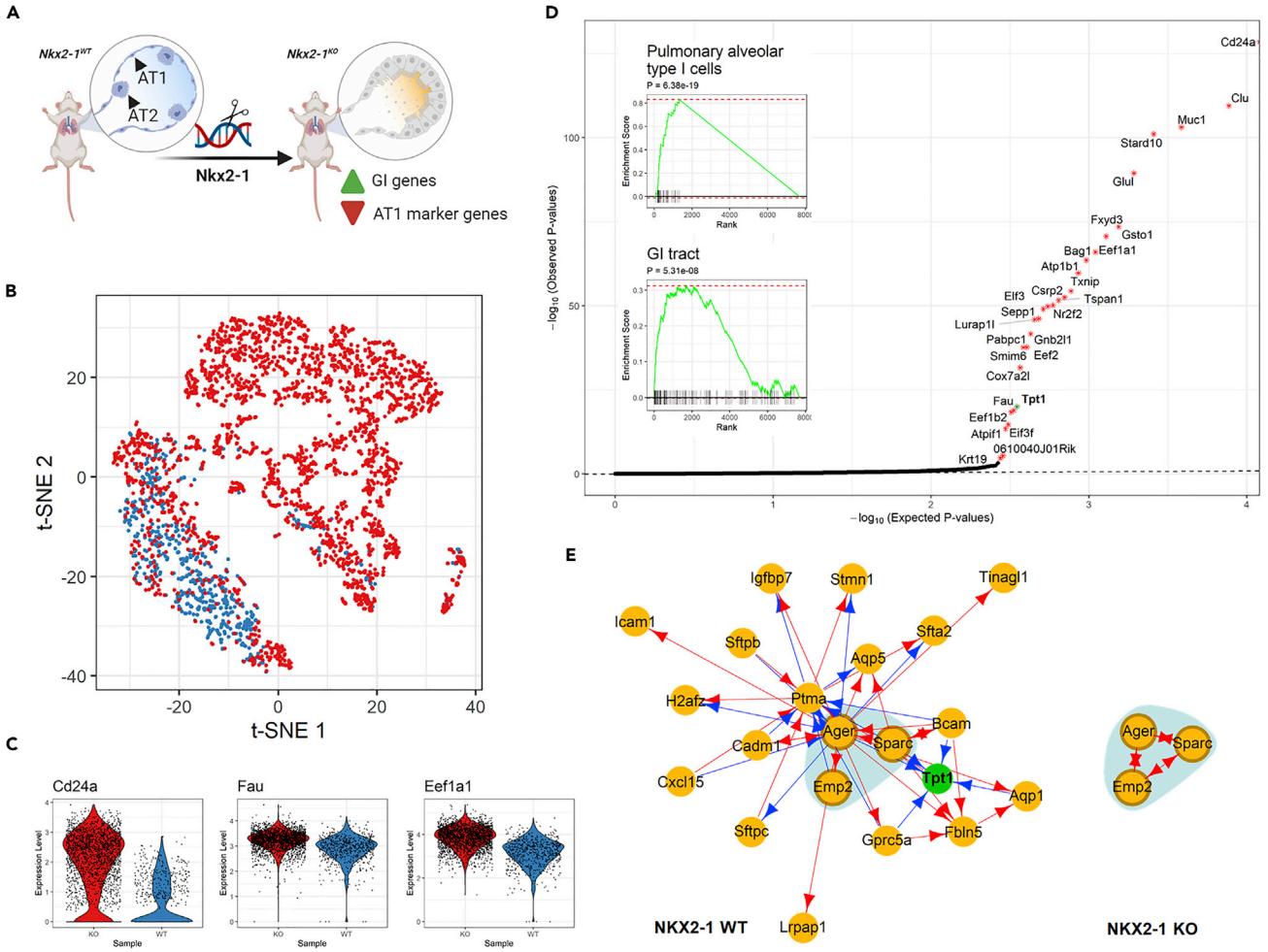


Figure 5. Analysis of Transcriptional Responses of Alveolar Type 1 Cells to Nkx2-1 Gene Knockout

(A) Illustration of experimental design and data collection of the KO experiment.[54]

(B) t-SNE plot of 2,397 and 638 AT1 cells from Nkx2-1 KO mice (red) and WT mice (blue).

(C) Violin plots showing the log-normalized expression levels of selected differentially regulated genes in KO (red) and WT (blue) mice.

(D) Q-Q plot for observed and expected p values of tested genes. Genes (n = 29) with FDR <0.05 are labeled with an asterisk. Inset shows the results of GSEA for genes ranked by their distances in manifold aligned scGRNs.

(E) A representative module that contains the differentially regulated gene *Tpt1* in the WT mice. Most parts of the module disappear in the KO mice. The colors, edges, and marks are presented as in Figure 3E.

Nkx2-1$^{CKO/CKO}$; *Aqp5*$^{Cre/+}$ mice and 638 AT1 cells from the WT mice (Figure 5B). Expression levels of *Cd24a*, *Fau*, and *Eef1a1* in AT1 cells of KO and WT mice are shown in Figure 5C. *Cd24a* is a marker gene for AT1 cells; *Fau* and *Eef1a1* are GI genes, known to be highly expressed in the GI tissues. Using scTenifoldNet, we identified 29 genes exhibiting significant difference in their regulation between the two samples: KO versus WT (FDR <0.05, Figure 5D). These 29 genes are ***Cd24a***, *Clu*, ***Muc1***, *Stard10*, *Glul*, ***Fxyd3***, ***Gsto1***, *Eef1a1*, *Bag1*, ***Atp1b1***, *Txnip*, *Csrp2*, ***Tspan1***, *Nr2f2*, ***Elf3***, *Sepp1*, *Pabpc1*, *Lurap1l*, *Gnb2l1*, *Eef2*, *Smim6*, *Cox7a2l*, ***Tpt1***, *Fau*, *Eef1b2*, *Eif3f*, *Atpif1*, ***0610040J01Rik***, and ***Krt19***. Targets of *Sox2* are highlighted in bold.[56] As reported,[54] this gene list is enriched with genes highly expressed in the intestine. Using GSEA, we showed the significant enrichment of GI marker genes (Figure 5D, insets),[57] which confirmed the effect of *Nkx2-1* KO on the cellular identity of AT1 cells.

Analysis of Transcriptional Responses of Human Dermal Fibroblasts to a Double-Stranded RNA Stimulus

Next, we show the use of scTenifoldNet on an scRNA-seq dataset from human dermal fibroblasts.[58] In the original paper, Hagai et al. focused on single-cell transcriptional responses induced by the stimulus of polyinosinic-polycytidylic acid (polyI:C), a synthetic double-stranded RNA (dsRNA) (Figure 6A).[58] They obtained and compared the transcriptomes of 2,553 unstimulated and 2,130 stimulated cells and identified 875 differentially expressed genes (Table S3 of Hagai et al.).[58] These differentially expressed genes include *IFNB*, *TNF*, *IL1A*, and *CCL5*, encoding antiviral and inflammatory gene products, and are enriched for *inflammatory response*, *positive regulation of immune system*

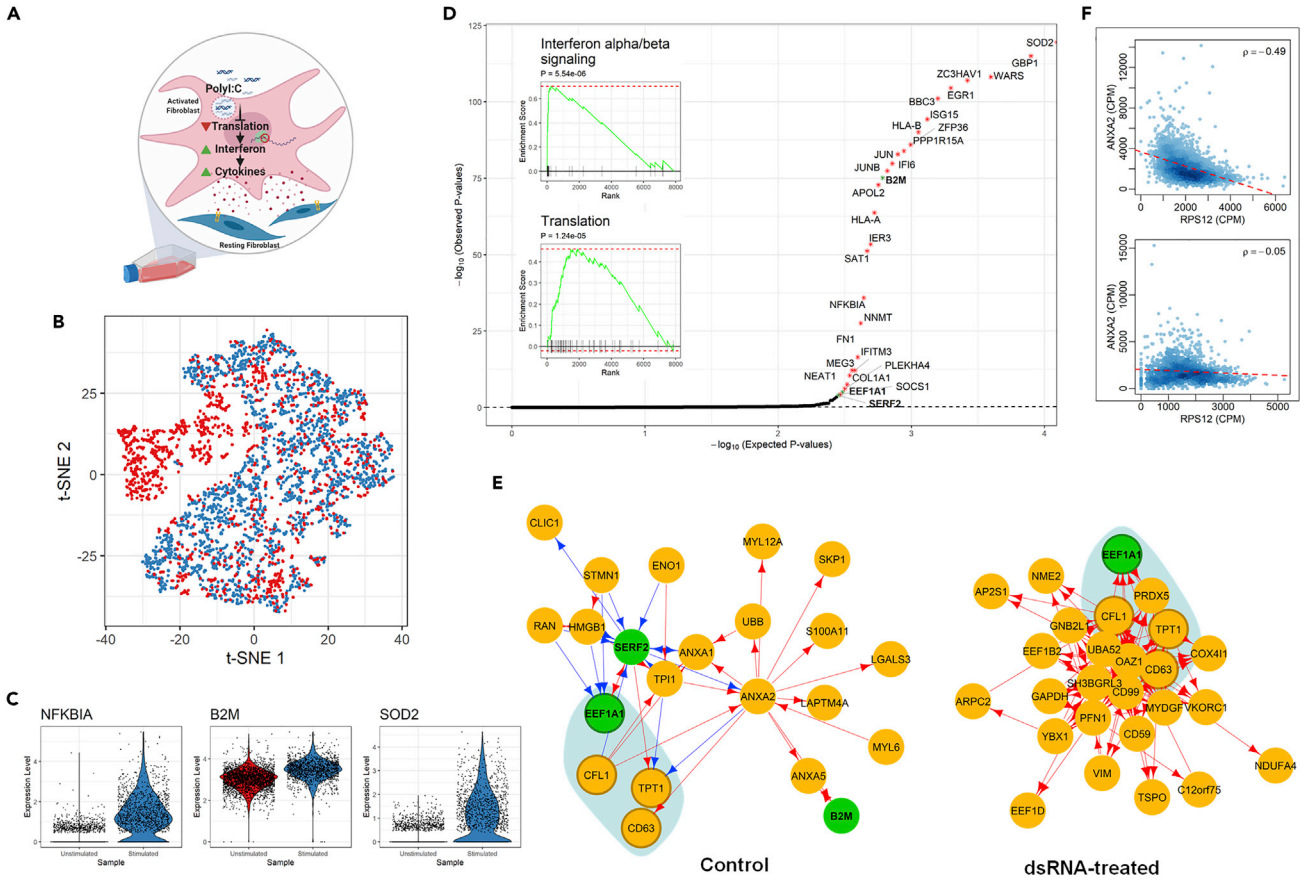


Figure 6. Analysis of Transcriptional Responses of Human Dermal Fibroblasts to a Double-Stranded RNA Stimulus

(A) Illustration of experimental design and tested mechanism of transcriptional responses.[58]

(B) t-SNE plot of human dermal fibroblasts before (blue) and after (red) dsRNA stimulus.

(C) Violin plots showing the log-normalized expression levels of selected differentially regulated genes before (blue) and after (red) stimulus.

(D) Q-Q plot for observed and expected p values of tested genes. Genes ($n = 29$) with FDR <0.05 are labeled with an asterisk. Inset shows the results of GSEA for genes ranked by their distances in manifold aligned scGRNs.

(E) Comparison of a representative module that contains three differentially regulated genes in the control sample. The colors, edges, and marks are presented as in Figure 3E.

(F) Scatterplots showing the correlation between *TPT1* and *ANXA2* before (top) and after (bottom) dsRNA stimulus.

process, and *response to cytokine*, among many other biological processes and pathways. We found that the original scRNA-seq data have a batch effect between two samples, but the global batch effect can be removed using Harmony,[59] as shown in the t-SNE plot of cells of two samples (Figure 6B). Nevertheless, the differences in the expression levels between samples can still be detected in selected genes with Harmony-processed data (Figure 6B).

Applying scTenifoldNet to the processed data, we identified 29 differentially regulated genes: ***SOD2***, ***GBP1***, *WARS*, *ZC3HAV1*, ***EGR1***, ***BBC3***, *ISG15*, ***HLA-B***, *ZFP36*, ***PPP1R15A***, ***JUN***, *IFI6*, ***JUNB***, ***B2M***, *APOL2*, ***HLA-A***, ***IER3***, *SAT1*, ***NFKBIA***, ***NNMT***, ***FN1***, *IFITM3*, *MEG3*, *NEAT1*, *COL1A1*, *PLEKHA4*, *EEF1A1*, *SOCS1*, and *SERF2* (Figure 6D, Table S6). Among them, 14 (highlighted in bold) are targets of *RELA* (48%, adjusted $p < 0.01$, enrichment test by Enrichr).[60] These differentially regulated genes are functionally enriched for *interferon signaling*, *immune system*, *interleukin-1 regulation of extracellular matrix*, and others (Table S7).

Once again, scTenifoldNet reports fewer genes than DE analysis does in the original paper.[58] By comparing the differentially regulated genes with the differentially expressed genes, we found that enriched functions of differentially expressed genes reflect the differences between unstimulated cells and cells that have completed an initial response to a dsRNA stimulus and reached a final phase of the response, whereas the enriched functions of differentially regulated genes reflect ongoing activities associated with regulatory changes and immune responses to the stimulus. In this sense, differentially regulated genes are valuable for informing about mechanisms through which the dsRNA acts to induce immunological responses.[61–63] For example, it is known that dsRNA inhibits the translation of mRNA to proteins[62] and leads to the synthesis of interferon, which induces the synthesis of ribosomal units that are able to distinguish between cell mRNA and viral RNA.[63] Interferon also promotes cytokine production that activates the immune response and induces inflammation.[61] To further illustrate the changes in the regulatory patterns between samples, we

plotted the GRN module around *EEF1A1*. It can be seen that, before and after the dsRNA treatment, the interacting partnership of the genes is changed substantially (Figure 6E). Two scatterplots show the change in correlation between *TPT1* and *ANXA2*, as an example (Figure 6F). The negative correlation between the two genes' expression among cells disappears after the dsRNA treatment and, thus, the two genes are not linked in the scGRN constructed using the after-treatment data.

Analysis of Transcriptional Responses of Mouse Neurons in Alzheimer Disease

Last, we applied scTenifoldNet to scRNA-seq data of isolated single nuclei from the brains of the WT and 5xFAD mice.[64] The 5xFAD strain recapitulates the major features of Alzheimer disease amyloid pathology. The genotype of these mice contains several familial Alzheimer disease (FAD) mutations in *APP* and *PSEN1*, causing the overexpression of mutant human amyloid-β (Aβ) precursor protein and human presenilin 1. The 5xFAD model rapidly develops amyloid pathology, with high levels of intraneuronal Aβ accumulation beginning around 1.5 months of age and extracellular Aβ deposition beginning around 2 months.[65]

In the original paper,[64] Zhou et al. compared scRNA-seq data between 6-month-old WT mice with 6-month-old 5xFAD mice. They found that neurons show limited responses to Aβ peptides: compared with microglia and oligodendrocytes, neurons show minimal transcriptional changes (149 differentially expressed genes) between WT and 5xFAD mice. To test whether scTenifoldNet can detect genes whose expression is differentially regulated between WT and 5xFAD mice, we decided to apply our method to these scRNA-seq data exclusively in neurons. We downloaded expression data matrices from the GEO database using accession no. GSE140511 and extracted expression data of neurons from two samples: WT2 (GSM4173505) and WT_5XFAD2 (GSM4173511) (Figure 7A).

After reanalyzing the data using scTenifoldNet, we identified 18 differentially regulated genes: *Zdhhc17*, *Chl1*, *Abhd17b*, *Rchy1*, *Stmn2*, *Tjp1*, *Nrbp2*, *Ly6h*, *Smarcd1*, *Rhbdd2*, *Ndfip1*, *Mark2*, *Icam5*, *Fam92a*, *Rgl1*, *Gmcl1*, *Daam1*, and *Fxr1* (FDR <0.05, Figure 7D). For functional enrichment analysis, we relaxed the significant-gene cutoff to include 57 additional genes with FDR $\geq$0.05 but nominal $p < 0.05$. These additional genes include *Apoe* and *Bin1*. *Bin1* encodes bridging integrator 1 (also known as amphiphysin 2), which is the second most important risk locus (after *Apoe*) for late-onset Alzheimer disease.[66,67] *Apoe* and *Bin1* rank 25th and 61th, respectively, in the list of 75 significant genes (18 genes with FDR <0.05 followed by 57 genes with nominal $p < 0.05$), and both play a role in *negative regulation of amyloid precursor protein catabolic process* and *tau protein binding*. Enrichr analysis reported the following top gene ontology terms: *regulation of neuron projection development*, *positive regulation of cell projection organization*, *phosphatidylserine metabolic process*, *protein acylation*, *potassium channel activity*, and *methylation-dependent protein binding*. GSEA showed that regulatory changes are associated with *integrin signaling pathway*, *serotonin HTR1 group and FOS pathway*, and *glutamate neurotransmitter release cycle* (Figure 7D, insets).

DISCUSSION

We present scTenifoldNet, a robust, unsupervised machine learning workflow that streamlines comparative GRN analyses with data from scRNA-seq. The key feature of scTenifoldNet is to apply comparative network analysis with scRNA-seq data. It detects differences in the cell population's state between two samples in a sensitive and scalable manner. It provides the function of DR analysis, which can be used to reveal subtle regulatory shifts of genes.

Today, DE analysis is still the primary method for the purpose of comparative analysis between scRNA-seq samples (see Avey et al., Hagai et al., Ximerakis et al.).[36,58,68] As scRNA-seq datasets are becoming widely available, there will be more and more interest in comparing samples. The scTenifoldNet-based DR analysis is expected to be adapted in more scenarios wherever DE analysis is applicable. scTenifoldNet learns and contrasts high-dimensional features of genes in scGRNs by examining global interactions between the genes. scTenifoldNet is more suitable for comparing highly similar samples, such as two populations of cells of the same type. scTenifoldNet is built as a robust, sensitive tool that can capture signals that are even confined to rare cell types.

To achieve technical requirements, we overcome several analytical barriers in developing scTenifoldNet. First, constructing scGRN from scRNA-seq data, which consists of cells in many different states, is challenging at present. It is also difficult to control for technical noise in the data. To address these issues, we let scTenifoldNet begin with random cell subsampling. It is worth noting that not only can random cell subsampling help in dealing with the problem of cell heterogeneity, but additional information of cells can be incorporated into subsampling schema. More specifically, in addition to the random subsampling using jackknife and bootstrap methods, we can adapt a semirandom subsampling schema, if cells in an input matrix are sorted according to pseudotime.[69] These cells can be subsampled using a pseudotime-guided method, with which sorted cells are sampled along the pseudotime trajectory. In such a way, the subsamples contain pseudotime information, and the multilayer scGRN constructed from these subsamples will contain the pseudotime-series information. In machine learning, many multilayer network analysis algorithms have been proposed.[70–72] With our pseudotime-series scGRN data, these algorithms will be relevant and applicable. Second, regulatory relationships between genes from scRNA-seq data are difficult to establish, even though the data may theoretically capture a complete picture of the regulatory gene landscape. We consider PC regression to stand out as a crucial method of building scGRNs. PC regression significantly outperforms the other GRN construction algorithms in all aspects of methodology metrics, including specificity, sensitivity, computational efficiency, and the required minimum number of cells. Importantly, PC regression explicitly projects thousands of gene expression measurements into a low-dimensional space to capture much of the observed variation. PC regression, therefore, establishes the relationship for each pair of genes after controlling for the most important background interactions. Third, in scTenifoldNet, the tensor denoising procedure effectively smooths edge weights across all networks in multilayer scGRNs. Fourth, scTenifoldNet performs

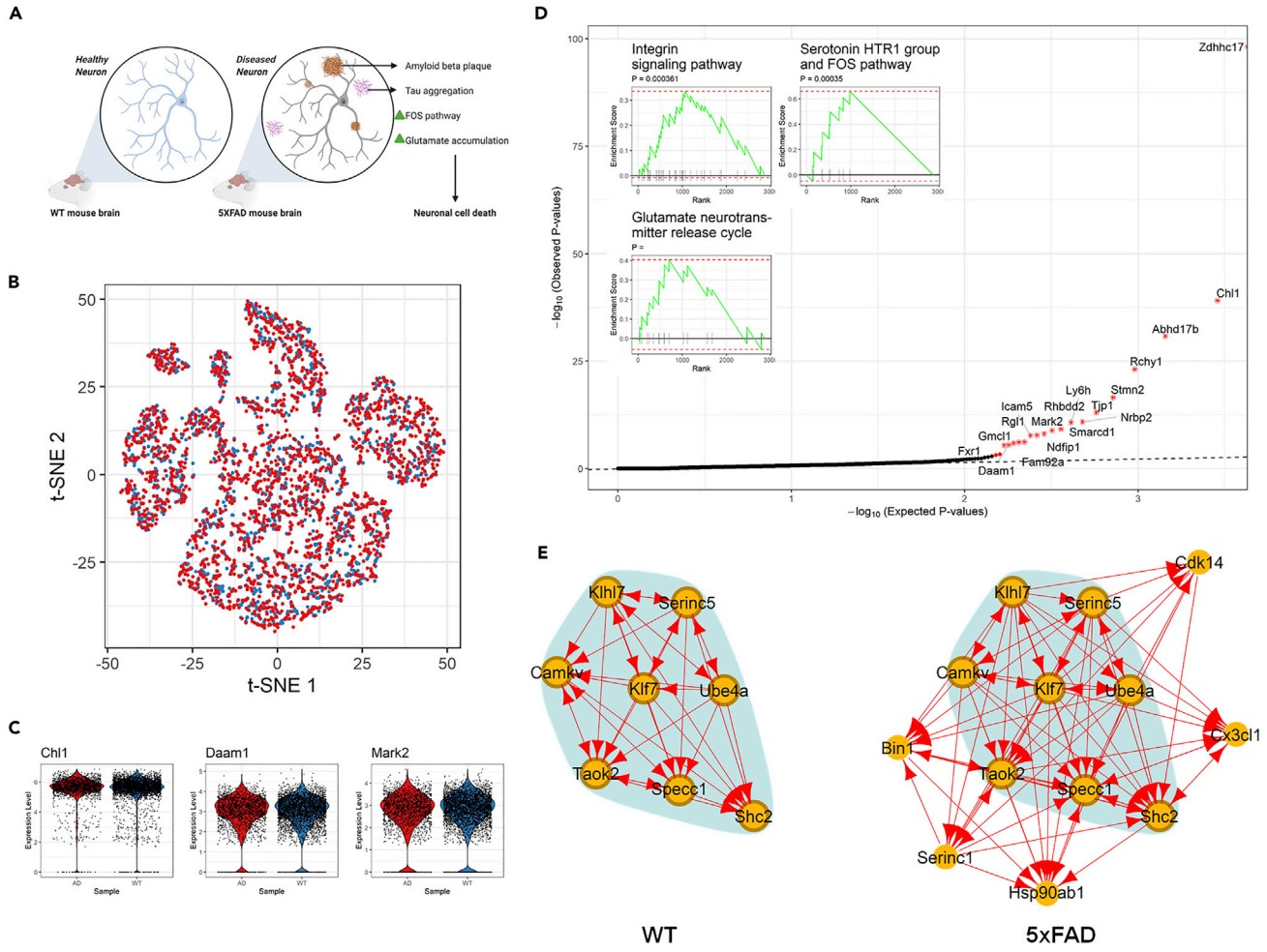


Figure 7. Analysis of Transcriptional Responses of Neurons to Aβ Peptides in 5xFAD Mice, a Model of Alzheimer Disease

(A) Illustration of experimental design and data collection of the 5xFAD mouse study.[64] (B) t-SNE plot of neurons of the 5xFAD (red) and WT (blue) mice.

(C) Violin plots showing the log-normalized expression levels of selected differentially regulated genes in neurons of the 5xFAD (red) and WT (blue) mice.

(D) Q-Q plot for observed and expected p values of tested genes. Genes (n = 18) with FDR <0.05 are labeled with an asterisk. Inset shows the results of the GSEA for genes ranked by their distances in manifold aligned scGRNs.

(E) Comparison of a representative module that contains top-ranked differentially regulated genes between the two scGRNs. The colors, edges, and marks are presented as in Figure 3E.

nonlinear manifold alignment to align two networks. As such, two networks can be contrasted directly, and differentially regulated genes could be detected using distance in new coordinates of data in a low-dimensional space.

We validate the power of scTenifoldNet using real datasets coming from various studies and demonstrate that scTenifoldNet is sensitive to signals. Five real scRNA-seq datasets are involved (Table 1). These five datasets have one thing in common: they all have two sets of scRNA-seq data: one from a treated group and the other from a control/untreated group. More importantly, in all five cases, we have sufficient prior knowledge about the biological system from which the data were collected. Therefore, we have hypotheses about what transcriptional changes we are expected to see before doing the analysis. For example, in the morphine response analysis, the causal factor of transcriptional responses, i.e., the morphine stimulus, is known, and thus, we know what should be recovered through the analysis. Similarly, we had some clues in the examples of cetuximab and fibroblasts about what transcriptional changes we might be able to retrieve. By compiling all the findings from scTenifoldNet applications, we tested scTenifoldNet and showed that scTenifoldNet provides findings that are precise, specific, and relevant to the biological systems and questions in the test. This is of significance to building a specific and sensitive tool like scTenifoldNet for the purpose of molecular mechanism studies using scRNA-seq. This is because causal factors and their target genes remain unknown in many biological systems studied. If this is the case, it is crucial to apply a sensitive approach like scTenifoldNet, which may be in addition to the DE analysis, to unveil more gene candidates. Only then will we be able to scrutinize identified genes further to learn the mechanisms behind their actions

Table 1. Summary of Real-Data Applications of scTenifoldNet Analysis

Study	Reference	Species	Cell Type	Perturbation Type	Number of Genes Included in Analysis	Number of Cells in Two Groups	Number of Differentially Regulated Genes	Enriched Functions of Differentially Regulated Genes
1	Avey et al.[36]	mouse	neurons	morphine	8,138	mock-treated 8,912; morphine-treated, 7,972	56	opioid signaling; signaling by G-protein-coupled receptors; reduction of cytosolic calcium levels; morphine addiction
2	Kagohara et al.[48]	human	carcinoma cell line	cetuximab	11,140	untreated, 5,217; treated, 4,507	125	EGFR1 pathway; regulation of apoptosis; cell-cycle checkpoints; G1 cell-cycle arrest, regulation of apoptosis
3	Little et al.[54]	mouse	lung alveolar cells	Nkx2-1 gene KO	7,842	WT, 638; KO, 2,397	29	gastrointestinal marker genes; *Sox2* target genes
4	Hagai et al.[58]	human	dermal fibroblasts	dsRNA immune stimulus	7,904	unstimulated, 2,553; stimulated, 2,130	29	interferon signaling; immune system; interleukin-1 regulation of extracellular matrix
5	Zhou et al.[64]	mouse	neurons	Alzheimer disease	2,869	WT, 4,561; 5xFAD, 2,423	18	*Apoe* and *Bin1*; regulation of neuron projection development; positive regulation of cell projection organization; phosphatidylserine metabolic process; protein acylation; potassium channel activity; methylation-dependent protein binding; integrin signaling pathway; serotonin HTR1 group and FOS pathway; glutamate neurotransmitter release cycle

in the whole system. We face such a challenge in many studies from unknown factors that cause a disorder. It is therefore critical that we adopt tools such as scTenifoldNet, instead of relying solely on conventional DE analysis, to tackle this big data analysis problem.

In summary, scRNA-seq enables the study of cellular and molecular components and the dynamics of complex biological systems at single-cell resolution. To unravel the regulatory mechanisms underlying cell behaviors, novel computational methods are essential for understanding the complexity in scRNA-seq data (e.g., scGRNs) that surpasses human interpretative ability. We anticipate that, when applied to real scRNA-seq data, our machine learning workflow implemented in scTenifoldNet can help achieve breakthroughs by deciphering the full cellular and molecular complexity of the data by constructing and comparing scGRNs.

EXPERIMENTAL PROCEDURES

Resource Availability

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, James J. Cai (jcai@tamu.edu).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

scTenifoldNet has been implemented in R. The source code is available at https://github.com/cailab-tamu/scTenifoldNet, which also includes the code of the benchmarking method, auxiliary functions, and example datasets (including the simulated data used to generate Figure 2). The scTenifoldNet R package is available at the CRAN repository: https://cran.r-project.org/web/packages/scTenifoldNet/.

The scTenifoldNet Workflow

The scTenifoldNet workflow takes two scRNA-seq expression matrices as inputs. The two matrices are supposed to be obtained from two samples of the same type of cell, such as those of different treatments or from diseased and healthy subjects. The purpose of the analysis is to identify genes whose transcriptional regulation is shifted between the two samples. The whole workflow consists of five steps: cell subsampling, network construction, network denoising, manifold alignment, and module detection.

Cell Subsampling

Instead of using all cells of each sample to construct a single GRN, we randomly subsample cells multiple times to obtain a set of subsampled cell populations. This subsampling strategy is to ensure the robustness of results against cell heterogeneity in samples. Subsampling of each sample is performed as follows: assuming the sample has M cells, m cells ($m < M$) are randomly selected to form a subsampled cell population. The process is repeated with cell replacement t times to produce a set of t subsampled cell populations.

Network Construction

For a given expression matrix, a PC-regression network construction method[5] is adopted to construct scGRN. PC regression is a popular multiple regression method, where the original explanatory variables are first subjected to a PC analysis (PCA) and then the response variable is regressed on the few leading PCs. By regressing on M PCs ($M << n$, where n is the total number of genes in the expression matrix), PC regression mitigates the overfitting and reduces the computation time. To build an scGRN, each time we focus on one gene (referred to as the target gene) and apply the PC-regression method, treating the expression level of the target gene as the response variable and the expression levels of other genes as the explanatory variables. The regression coefficients from PC regression are then used to measure the strength of the association of the target gene and other genes and to construct the scGRN. We repeat this process n times, each time with one gene as the target gene. At the end, the interaction strengths between all possible gene pairs are obtained and an adjacency matrix is formed. The details of applying the PC-regression method to a scRNA-seq expression data matrix are described as follows.

More specifically, suppose $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is the gene expression matrix with n genes and p cells. The ith row of $\boldsymbol{X}$, denoted by $X_i \in \mathbb{R}^p$, represents the gene expression level of the ith gene in p cells. We construct a data matrix, $\boldsymbol{X}_{-i} \in \mathbb{R}^{(n-1)\times p}$, by deleting X_i from $\boldsymbol{X}$. To estimate the effects of the other $n-1$ genes to the ith gene, we build a PC-regression model for X_i. First, we apply PCA to $\boldsymbol{X}_{-i}^T$ and take the first M leading PCs to construct $\boldsymbol{Z}^i = (Z_1^i, \cdots, Z_M^i) \in \mathbb{R}^{p\times M}$, where $Z_m^i \in \mathbb{R}^p$ is the mth PC of $\boldsymbol{X}_{-i}^T$, $m = 1, 2, \ldots, M$. Mathematically, $\boldsymbol{Z}^i = \boldsymbol{X}_{-i}^T \boldsymbol{V}^i$, where $\boldsymbol{V}^i \in \mathbb{R}^{(n-1)\times M}$ is the PC loading matrix for the first M leading PCs, satisfying $(\boldsymbol{V}^i)^T \boldsymbol{V}^i = \boldsymbol{I}_M$. Second, the PC-regression method regresses $\boldsymbol{X}_i$ on $\boldsymbol{Z}^i$ and solves the following optimization problem:

$$\hat{\beta}^i = \arg\ min_{\beta^i \in R^M} \boldsymbol{X}_i - \boldsymbol{Z}^i \beta^i{}_2^2.$$

Then, $\hat{\alpha}^i = \boldsymbol{V}^i \hat{\beta}^i \in \mathbb{R}^{n-1}$ quantifies the effects of the other $n-1$ genes to the ith gene. After performing PC regression on each gene, we collect $\{\hat{\alpha}^i\}_{i=1}^n$ together and construct an $n \times n$ weighted adjacency matrix $\boldsymbol{W}$ of the gene-gene interaction network. The ith row of $\boldsymbol{W}$ is $\hat{\alpha}^i$, and the diagonal entries of $\boldsymbol{W}$ are all 0. Then we retain interactions with the top α% (= 5% by default) absolute value in the matrix to obtain the scGRN adjacency matrix.

Tensor Decomposition

For each of the t subsamples of cells obtained in the cell subsampling step, we construct a network using PC regression, as described above. Each network is represented as an $n \times n$ adjacency matrix; the adjacency matrices of the t networks can be stacked to form a third-order tensor, $\mathfrak{T} \in \mathbb{R}^{(n\times n\times t)}$. To remove the noise in the adjacency matrices and extract important latent factors, the CP tensor decomposition is applied. Similar to the truncated SVD of a matrix, the CP decomposition approximates the tensor by a summation of multiple rank 1 tensors.[11] More specifically, for our problem:

$$\mathfrak{T} \approx \mathfrak{T}_d = \sum_{r=1}^{d} \lambda_r a_r \circ b_r \circ c_r,$$

where $\circ$ denotes the outer product, $a_r \in \mathbb{R}^n, b_r \in \mathbb{R}^n$ and $c_r \in \mathbb{R}^t$ are unit-norm vectors, and λ_r is a scalar. In the CP decomposition, $\mathfrak{T}_d$ is the denoised tensor of $\mathfrak{T}$, which assumes that the valid information of $\mathfrak{T}$ can be described by d rank 1 tensors, and the remaining part $\mathfrak{T} - \mathfrak{T}_d$ is mostly noise.

We use the function `cp` in the R package "rTensor" to do the CP decomposition. For each sample, the reconstructed tensor $\mathfrak{T}_d$ includes t denoised scGRNs. We then calculate the average of associated t denoised networks to obtain the overall stable network. We further normalize entries by dividing them by their maximum absolute value to obtain the final scGRNs for the given sample. For later use, we denote the denoised adjacency matrices for the two samples as $\boldsymbol{W}_d^x$ and $\boldsymbol{W}_d^y$.

Manifold Alignment

$\boldsymbol{W}_d^x$ and $\boldsymbol{W}_d^y$ are then compared to identify changes in regulatory relationships among genes and identify significantly affected genes. Instead of directly comparing these two $n \times n$ adjacency matrices, manifold alignment is applied to match the local and non-linear structures among the data points of $\boldsymbol{W}_d^x$ and $\boldsymbol{W}_d^y$, project them to the same low-dimensional space, and build comparable low-dimensional features. These features between two samples can then be compared while maintaining the structural information of the two scGRNs. Specifically, $\boldsymbol{W}_d^x$ and $\boldsymbol{W}_d^y$ serve as the input for a manifold alignment algorithm to find the low-dimensional projections $\boldsymbol{F}^x \in \mathbb{R}^{n\times d}$ and $\boldsymbol{F}^y \in \mathbb{R}^{n\times d}$ of genes from each sample, where $d << n$. In terms of the underlying matrix representation, we use $F_i^x \in \mathbb{R}^d$ and $F_i^y \in \mathbb{R}^d$ to denote the ith row of $\boldsymbol{F}^x$ and $\boldsymbol{F}^y$ that reflect the features of the ith gene in $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively.

We note that $\boldsymbol{W}_d^x$ and $\boldsymbol{W}_d^y$ may include negative values, which means genes are negatively correlated. When an adjacency matrix contains negative edge weights, the properties of the corresponding Laplacian are not entirely well understood.[73] To mitigate this problem, we add 1 to all entries in $\boldsymbol{W}_d^x$ and $\boldsymbol{W}_d^y$, transforming the range of $\boldsymbol{W}_d^x$ and $\boldsymbol{W}_d^y$ from $[-1,1]$ to $[0,2]$. As a result, all original negative relationships have a transformed value in $[0,1)$ and all original positive relationships have a transformed value in $(1,2]$. The projected features of two genes with a positive correlation will be closer than those with a negative correlation. For convenience, we still use $\boldsymbol{W}_d^x$ and $\boldsymbol{W}_d^y$ to denote the transformed similarity matrices of two datasets.

Now we propose a specific manifold alignment method to find appropriate low-dimensional projections of each gene. Our manifold alignment should trade off the following two requirements: (1) the projections of the same ith gene in two samples should be relatively close in the projected space and (2) if the ith gene and jth gene in sample 1 are functionally related, their projections F_i^x and F_j^x should be close in the projected space, and the same is true for sample 2. We minimize the following loss function:

$$Loss(\boldsymbol{F}^x, \boldsymbol{F}^y) = \lambda \sum_{i=1}^{n} \|F_i^x - F_i^y\|_2^2 + \sum_{i,j=1}^{n} \|F_i^x - F_j^x\|_2^2 W_{ij}^x + \sum_{i,j=1}^{n} \|F_i^y - F_j^y\|_2^2 W_{ij}^y,$$

where W_{ij}^x and W_{ij}^y denote the (i, j) entry of $\boldsymbol{W}_d^x$ and $\boldsymbol{W}_d^y$, respectively. The first term of the loss function requires the similarity between corresponding genes across two samples; the second and third terms are regularizers preserving the local similarity of genes in each of the two networks. λ is an allocation parameter to balance the effects of two requirements.

One way to minimize the loss function is by using an algorithm similar to Laplacian eigenmaps,[74] which requires the adjacency matrix to be symmetry, but in our case both $\boldsymbol{W}_d^x$ and $\boldsymbol{W}_d^y$ are asymmetric. Notice that if we symmetrize $\boldsymbol{W}_d^x$ and $\boldsymbol{W}_d^y$ by $\boldsymbol{W}^x = \frac{1}{2}\left((\boldsymbol{W}_d^x)^T + \boldsymbol{W}_d^x\right)$ and $\boldsymbol{W}^y = \frac{1}{2}((\boldsymbol{W}_d^y)^T + \boldsymbol{W}_d^y)$, and again denote W_{ij}^x and W_{ij}^y as the (i, j) entry of $\boldsymbol{W}^x$ and $\boldsymbol{W}^y$, then the value of the loss function will not be changed. Thus, minimizing the loss function based on the symmetrized adjacency matrices, $\boldsymbol{W}^x$ and $\boldsymbol{W}^y$, is equivalent to using the original adjacency matrices, $\boldsymbol{W}_d^x$ and $\boldsymbol{W}_d^y$. Based on this observation, using linear algebra, we can write the loss function into the matrix form as $Loss(\boldsymbol{F}^x, \boldsymbol{F}^y) = 2trace(\boldsymbol{F}^T \boldsymbol{L} \boldsymbol{F})$, where $\boldsymbol{F} = \begin{bmatrix} \boldsymbol{F}^x \\ \boldsymbol{F}^y \end{bmatrix}$, $\boldsymbol{L} = \frac{1}{2}(\boldsymbol{D} - \boldsymbol{W})$, $\boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}^x & \frac{\lambda}{2}\boldsymbol{I} \\ \frac{\lambda}{2}\boldsymbol{I} & \boldsymbol{W}^y \end{bmatrix}$, and $\boldsymbol{D}$ is a diagonal matrix with $D_{ii} = \sum_i W_{ij}$. $\boldsymbol{L}$ is called a graph Laplacian matrix. The default selection of λ is 0.9 times the mean value of the row sums of $\boldsymbol{W}^x$ and $\boldsymbol{W}^y$. By further adding the constraint $\boldsymbol{F}^T \boldsymbol{F} = \boldsymbol{I}$ to remove the arbitrary scaling factor, minimizing $Loss(\boldsymbol{F}^x, \boldsymbol{F}^y)$ is equivalent to solving an eigenvalue problem. The solution for $\boldsymbol{F} = [f_1, f_2, \cdots, f_d]$ is given by d eigenvectors corresponding to the d smallest nonzero eigenvalues of $\boldsymbol{L}$.[75]

Determination of the p Value of Differentially Regulated Genes

With $\boldsymbol{F} = \begin{bmatrix} \boldsymbol{F}^x \\ \boldsymbol{F}^y \end{bmatrix} = [f_1, f_2, \cdots, f_d]$ obtained in manifold alignment, we calculate the distance d_j between projected data points of two samples for each gene. One may declare significant genes according to the ranking of d_j's. To avoid arbitrariness in deciding the number of selected genes, we propose to use χ^2 distribution to determine the significance of genes.[76] Specifically, d_j^2 is derived from the summation of squares of differences of projected representations of gene j for two samples, whose distribution could be approximately χ^2. To adjust the scale of the distribution, we compute the scaled fold change defined as $df \cdot d_j^2 / \overline{d^2}$ for each gene j, where f denotes the average of d_j^2 among all the tested genes. The scaled fold change approximately follows χ^2 distribution with the degree of freedom df if the gene does not perform differently in the

two samples. By using the upper tail ($P[X > x]$) of the χ^2 distribution, we assign p values for genes and adjust them for multiple testing using B-H FDR correction.[77] To determine *df*, since the number of the significant genes will increase as *df* increases, we use $df = 1$ to make a conservative selection of genes with high precision.

Functional Enrichment Analyses

Functional enrichment analysis of gene sets was performed using Enrichr,[78,79] which is a web-based, integrative enrichment analysis application based on more than 100 curated gene set libraries. The test of enriched TF targets was performed using the ChIP-X enrichment analysis[40] based on comprehensive results from ChIP-seq studies. Finally, predefined gene sets from the REACTOME, BioPlanet, and KEGG databases were tested for enriched functions using the pre-ranked GSEA.[26]

Simulations of scRNA-Seq Data and Benchmarking of Network Methods

A systematic evaluation of state-of-the-art algorithms for inferring scGRNs was performed using BEELINE.[10] We applied scTenifoldNet/PC regression and other scGRN inference algorithms to a dataset called GSD, which is derived from a curated Boolean model.[80] These methods include PIDC,[27] PPCOR,[81] LEAP,[82] GRNBOOST2,[83] GENIE3,[2] SCINGE,[84] SINCERITIES,[85] GRISLI,[86] SCODE,[87] GRNVBEM,[88] and SCNS.[89] Due to compatibility issues, Scribe[90] was not included in the comparison. We processed the dataset through the uniform pipeline provided by BEELINE, including (1) data pre-processing, (2) generation of Docker containers for scTenifoldNet/PC regression and the 11 above-mentioned algorithms, (3) parameter estimation, and (4) post-processing and evaluation. Throughout the analysis, no information on TF-target relationships was given to any tested algorithm. We compared algorithms based on their average performance among three different metrics: AUROC, AUPRC, and time of computing. AUROC shows the performance of a tested algorithm by presenting the trade-off between true positive rate TP/(TP + FN) and false positive rate FP/(FP + TN) across different decision thresholds, while AUPRC shows the area under the precision TP/(TP + FP)-recall TP/(TP + FN) curve computed for different decision thresholds between 1 and 0 using $\sum_i (R_i - R_{i-1})P_i$, where P_i and R_i are the precision and recall at the *i*th threshold, summarizing a weighted mean of precisions achieved at each threshold with the increase in recall from the previous threshold used as the weight. TP stands for true positive, TN true negative, FP false positive, and FN false negative.

We generated our own synthetic datasets using SERGIO, a single-cell expression simulator guided by GRNs.[28] SERGIO allows for the simulation of scRNA-seq data while considering the linear and non-linear influences of regulatory interactions between genes. SERGIO takes a user-provided GRN to define the interactions and generates expression profiles of genes in steady state using systems of stochastic differential equations derived from the chemical Langevin equation. The time course of mRNA concentration of gene *i* is modeled by:

$$\frac{\partial X_i}{\partial t} = P_i(t) - \lambda_i x_i(t) + q_i\left(\sqrt{P_i(t)}\alpha\right),$$

where x_i is the expression of gene *i*; P_i is its production rate, which reflects the influence of its regulators as identified by the given GRN; λ_i is the decay rate; q_i is the noise amplitude in the transcription of gene *i*; and α is an independent Gaussian white noise process. To obtain the mRNA concentrations as a function of time, the above stochastic differential equation is integrated for all genes as follows:

$$(X_i)_t = (X_i)_{t_0} + \int_{t_o}^{t} (P_i(t) - \lambda_i x_i(t))\partial t + \int_{t_0}^{t} q_i\left(\sqrt{P_i(t)}\right)\partial W_\alpha.$$

The simulation was focused on testing and comparing the performance of PC regression and several other methods (SCC, MI, GENIE3) using sparse data without imputation. The relationships between 100 genes were simulated as they belong to two major modules containing 40 and 60 genes, respectively. Each module is under the influence of one TF. We used the steady-state simulations to synthesize data to generate expression profiles of 100 genes, according to the parameter setting for two modules.

For each of the tested methods, we randomly selected $n = \{10, 50, 100, 500, 1{,}000, 2{,}000, 3{,}000\}$ cells from the simulated data 10 times and built 10 scGRNs. For each *n*, relevance measurements (accuracy and recall) were evaluated for each of the 10 networks using the match of the sign of the relationships between genes to compute the following formulas: accuracy = (TP + TN)/(TP + TN + FP + FN) and recall = TP/(TP + FN). For the MI and GENIE3 methods that provide only positive values, the median value was used as the center point, and then the values were scaled to [−1,1] by dividing them over the maximum absolute value.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.patter.2020.100139.

ACKNOWLEDGMENTS

We much appreciate insightful views and constructive comments from four anonymous reviewers who helped us improve this paper. Our research was supported by the Texas A&M University 2019 T3- and X-Grants for J.J.C. and the 2020 Award of Texas A&M Institute of Data Science (TAMIDS) Data Resource Development Program for D.O. and Y.Z.

AUTHOR CONTRIBUTIONS

J.J.C. and D.O. designed the workflow and conceptualized the study. D.O. implemented the software. D.O., Y.Z., and G.L. performed data analysis under the supervision of J.J.C. and J.Z.H. All authors contributed to the writing of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 15, 2020
Revised: September 29, 2020
Accepted: October 12, 2020
Published: November 5, 2020

REFERENCES

1. Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., and Califano, A. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics *7* (*Suppl 1*), S7.
2. Huynh-Thu, V.A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. PLoS One *5*, https://doi.org/10.1371/journal.pone.0012776.
3. Marbach, D., Costello, J.C., Kuffner, R., Vega, N.M., Prill, R.J., Camacho, D.M., Allison, K.R., Consortium, D., Kellis, M., Collins, J.J., et al. (2012). Wisdom of crowds for robust gene network inference. Nat. Methods *9*, 796–804.
4. Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. J. Comput. Biol. *7*, 601–620.
5. Gill, R., Datta, S., and Datta, S. (2010). A statistical framework for differential network analysis from microarray data. BMC Bioinformatics *11*, 95.
6. Todorov, H., Cannoodt, R., Saelens, W., and Saeys, Y. (2019). Network inference from single-cell transcriptomic data. Methods Mol. Biol. *1883*, 235–249.
7. Aibar, S., Gonzalez-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.C., Geurts, P., Aerts, J., et al. (2017). SCENIC: single-cell regulatory network inference and clustering. Nat. Methods *14*, 1083–1086.
8. Fiers, M., Minnoye, L., Aibar, S., Bravo Gonzalez-Blas, C., Kalender Atak, Z., and Aerts, S. (2018). Mapping gene regulatory networks from single-cell omics data. Brief. Funct. Genomics *17*, 246–254.

9. Chen, S., and Mar, J.C. (2018). Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. BMC Bioinformatics *19*, 232.
10. Pratapa, A., Jalihal, A.P., Law, J.N., Bharadwaj, A., and Murali, T.M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nat. Methods *17*, 147–154.
11. Rabanser, S., Shchur, O., and Günnemann, S. (2017). Introduction to tensor decompositions and their applications in machine learning. arXiv, 1711.10781.
12. Osorio, D., Yu, X., Zhong, Y., Li, G., Yu, P., Serpedin, E., Huang, J.Z., and Cai, J.J. (2019). Single-cell expression variability implies cell function. Cells *9*, 14.
13. Beasley, W.H., and Rodgers, J. (2019). Resampling Methods. In The Sage Handbook of Quantitative Methods in Psychology, *Vol. 9*, R.E. Millsap and A. Maydeu-Olivares, eds. *(Sage), pp. 60–71*.
14. Kendall, M.G. (1957). A Course in Multivariate Analysis (Hafner Pub. Co.).
15. Baburaj, M., and George, S.N. (2016). Reweighted low-rank tensor decomposition based on t-SVD and its applications in video denoising. arXiv, 1611.05963.
16. Yuan, L., Zhao, Q., Gui, L., and Cao, J. (2018). High-dimension tensor completion via gradient-based optimization under tensor-train format. arXiv, 1804.01983.
17. Battaglino, C., Ballard, G., and Kolda, T.G. (2017). A practical randomized CP tensor decomposition. arXiv, 1701.06600.
18. Moon, K.R., Stanley, J.S., Burkhardt, D., van Dijk, D., Wolf, G., and Krishnaswamy, S. (2018). Manifold learning-based methods for analyzing single-cell RNA-sequencing data. Curr. Opin. Syst. Biol. *7*, 36–46.
19. Roscher, R., Schindler, F., and Förstner, W. (2011). In Proceedings of the 2010 International Conference on Computer Vision - Volume Part II 334-343 (Queenstown, New Zealand: Springer-Verlag).
20. Vu, H.T., Carey, C.J., and Mahadevan, S. (2012). In Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence 1155-1161 (Toronto, Ontario, Canada: AAAI Press).
21. Wang, C., and Mahadevan, S. (2009). AAAI Fall Symposium: Manifold Learning and its Applications FS-09-04 (Association for the Advancement of Artificial Intelligence).
22. Nguyen, N.D., Blaby, I.K., and Wang, D. (2019). ManiNetCluster: a novel manifold learning approach to reveal the functional links between gene networks. BMC genomics *20*, 1003.
23. Diaz, F., and Metzler, D. (2007). In Proceedings of the 20th International Joint Conference on Artifical Intelligence 2727-2732 (Hyderabad, India: Morgan Kaufmann Publishers Inc).
24. Wang, C., and Mahadevan, S. (2008). In Proceedings of the 25th International Conference on Machine Learning - ICML '08 1120-1127 (Helsinki, Finland: ACM).
25. Wilson, R.C., and Zhu, P. (2008). A study of graph spectra for comparing graphs and trees. Pattern Recognit. *41*, 2833–2841.
26. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U S A *102*, 15545–15550.
27. Chan, T.E., Stumpf, M.P.H., and Babtie, A.C. (2017). Gene regulatory network inference from single-cell data using multivariate information measures. Cell Syst. *5*, 251–267.e3.
28. Dibaeinia, P., and Sinha, S. (2020). SERGIO: a single-cell expression simulator guided by gene regulatory networks. Cell Syst. *11*, 252–271.
29. van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A.J., Burdziak, C., Moon, K.R., Chaffer, C.L., Pattabiraman, D., et al. (2018). Recovering gene interactions from single-cell data using data diffusion. Cell *174*, 716–729.e27.
30. Skinnider, M.A., Squair, J.W., Kathe, C., Anderson, M.A., Gautier, M., Matson, K.J.E., Milano, M., Hutson, T.H., Barraud, Q., Phillips, A.A., et al. (2020). Cell type prioritization in single-cell data. Nat. Biotechnol. https://doi.org/10.1038/s41587-020-0605-1.
31. Rodgers, J.L. (1999). The bootstrap, the jackknife, and the randomization test: a sampling taxonomy. Multivariate Behav. Res. *34*, 441–456.
32. Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A.K., Slichter, C.K., Miller, H.W., McElrath, M.J., Prlic, M., et al. (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. *16*, 278.
33. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.
34. Kharchenko, P.V., Silberstein, L., and Scadden, D.T. (2014). Bayesian approach to single-cell differential expression analysis. Nat. Methods *11*, 740–742.
35. Soneson, C., and Robinson, M.D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. Nat. Methods *15*, 255–261.
36. Avey, D., Sankararaman, S., Yim, A.K.Y., Barve, R., Milbrandt, J., and Mitra, R.D. (2018). Single-cell RNA-seq uncovers a robust transcriptional response to morphine by glia. Cell Rep. *24*, 3619–3629.e14.
37. Goodsell, D.S. (2004). The molecular perspective: morphine. Oncologist *9*, 717–718.
38. Tso, P.H., and Wong, Y.H. (2003). Molecular basis of opioid dependence: role of signal regulation by G-proteins. Clin. Exp. Pharmacol. Physiol. *30*, 307–316.
39. Jalabert, M., Bourdy, R., Courtin, J., Veinante, P., Manzoni, O.J., Barrot, M., and Georges, F. (2011). Neuronal circuits underlying acute morphine action on dopamine neurons. Proc. Natl. Acad. Sci. U S A *108*, 16446–16450.
40. Lachmann, A., Xu, H., Krishnan, J., Berger, S.I., Mazloom, A.R., and Ma'ayan, A. (2010). ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. Bioinformatics *26*, 2438–2444.
41. Krezel, W., Ghyselinck, N., Samad, T.A., Dupe, V., Kastner, P., Borrelli, E., and Chambon, P. (1998). Impaired locomotion and dopamine signaling in retinoid receptor mutant mice. Science *279*, 863–867.
42. Tafti, M., and Ghyselinck, N.B. (2007). Functional implication of the vitamin A signaling pathway in the brain. Arch. Neurol. *64*, 1706–1711.
43. Morikawa, H., and Paladini, C.A. (2011). Dynamic regulation of midbrain dopamine neuron activity: intrinsic, synaptic, and plasticity mechanisms. Neuroscience *198*, 95–111.
44. Johnson, S.W., and North, R.A. (1992). Opioids excite dopamine neurons by hyperpolarization of local interneurons. J. Neurosci. *12*, 483–488.
45. Laakso, A., Mohn, A.R., Gainetdinov, R.R., and Caron, M.G. (2002). Experimental genetic approaches to addiction. Neuron *36*, 213–228.
46. Kim, K.S., Lee, K.W., Lee, K.W., Im, J.Y., Yoo, J.Y., Kim, S.W., Lee, J.K., Nestler, E.J., and Han, P.L. (2006). Adenylyl cyclase type 5 (AC5) is an essential mediator of morphine action. Proc. Natl. Acad. Sci. U S A *103*, 3908–3913.
47. Korostynski, M., Piechota, M., Kaminska, D., Solecki, W., and Przewlocki, R. (2007). Morphine effects on striatal transcriptome in mice. Genome Biol. *8*, R128.
48. Kagohara, L.T., Zamuner, F., Davis-Marcisak, E.F., Sharma, G., Considine, M., Allen, J., Yegnasubramanian, S., Gaykalova, D.A., and Fertig, E.J. (2020). Integrated single-cell and bulk gene expression and ATAC-seq reveals heterogeneity and early changes in pathways associated with resistance to cetuximab in HNSCC-sensitive cell lines. Br. J. Cancer *123*, 101–113.
49. Blick, S.K., and Scott, L.J. (2007). Cetuximab: a review of its use in squamous cell carcinoma of the head and neck and metastatic colorectal cancer. Drugs *67*, 2585–2607.
50. Harding, J., and Burtness, B. (2005). Cetuximab: an epidermal growth factor receptor chemeric human-murine monoclonal antibody. Drugs Today (Barc) *41*, 107–127.

51. Vincenzi, B., Zoccoli, A., Pantano, F., Venditti, O., and Galluzzo, S. (2010). Cetuximab: from bench to bedside. Curr. Cancer Drug Targets *10*, 80–95.

52. Herbst, R.S., and Shin, D.M. (2002). Monoclonal antibodies to target epidermal growth factor receptor-positive tumors: a new paradigm for cancer therapy. Cancer *94*, 1593–1611.

53. Burtness, B. (2005). The role of cetuximab in the treatment of squamous cell cancer of the head and neck. Expert Opin. Biol. Ther. *5*, 1085–1093.

54. Little, D.R., Gerner-Mauro, K.N., Flodby, P., Crandall, E.D., Borok, Z., Akiyama, H., Kimura, S., Ostrin, E.J., and Chen, J. (2019). Transcriptional control of lung alveolar type 1 cell development and maintenance by NK homeobox 2-1. Proc. Natl. Acad. Sci. U S A *116*, 20545–20555.

55. Desai, T.J., Brownfield, D.G., and Krasnow, M.A. (2014). Alveolar progenitor and stem cells in lung development, renewal and cancer. Nature *507*, 190–194.

56. Tompkins, D.H., Besnard, V., Lange, A.W., Keiser, A.R., Wert, S.E., Bruno, M.D., and Whitsett, J.A. (2011). Sox2 activates cell proliferation and differentiation in the respiratory epithelium. Am. J. Respir. Cell Mol. Biol. *45*, 101–110.

57. Franzen, O., Gan, L.M., and Bjorkegren, J.L.M. (2019). PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database (Oxford) *2019*, baz046.

58. Hagai, T., Chen, X., Miragaia, R.J., Rostom, R., Gomes, T., Kunowska, N., Henriksson, J., Park, J.E., Proserpio, V., Donati, G., et al. (2018). Gene expression variability across cells and species shapes innate immunity. Nature *563*, 197–202.

59. Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.R., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods *16*, 1289–1296.

60. Li, M., Shillinglaw, W., Henzel, W.J., and Beg, A.A. (2001). The Rela(p65) subunit of NF-kappaB is essential for inhibiting double-stranded RNA-induced cytotoxicity. J. Biol. Chem. *276*, 1185–1194.

61. Kopitar-Jerala, N. (2017). The role of interferons in inflammation and inflammasome activation. Front. Immunol. *8*, 873.

62. Gantier, M.P., and Williams, B.R. (2007). The response of mammalian cells to double-stranded RNA. Cytokine Growth Factor Rev. *18*, 363–371.

63. Levy, H.B., Law, L.W., and Rabson, A.S. (1969). Inhibition of tumor growth by polyinosinic-polycytidylic acid. Proc. Natl. Acad. Sci. U S A *62*, 357–361.

64. Zhou, Y., Song, W.M., Andhey, P.S., Swain, A., Levy, T., Miller, K.R., Poliani, P.L., Cominelli, M., Grover, S., Gilfillan, S., et al. (2020). Human and mouse single-nucleus transcriptomics reveal TREM2-dependent and TREM2-independent cellular responses in Alzheimer's disease. Nat. Med. *26*, 131–142.

65. Oakley, H., Cole, S.L., Logan, S., Maus, E., Shao, P., Craft, J., Guillozet-Bongaarts, A., Ohno, M., Disterhoft, J., Van Eldik, L., et al. (2006). Intraneuronal beta-amyloid aggregates, neurodegeneration, and neuron loss in transgenic mice with five familial Alzheimer's disease mutations: potential factors in amyloid plaque formation. J. Neurosci. *26*, 10129–10140.

66. Tan, M.S., Yu, J.T., and Tan, L. (2013). Bridging integrator 1 (BIN1): form, function, and Alzheimer's disease. Trends Mol. Med. *19*, 594–603.

67. Holler, C.J., Davis, P.R., Beckett, T.L., Platt, T.L., Webb, R.L., Head, E., and Murphy, M.P. (2014). Bridging integrator 1 (BIN1) protein expression increases in the Alzheimer's disease brain and correlates with neurofibrillary tangle pathology. J. Alzheimers Dis. *42*, 1221–1227.

68. Ximerakis, M., Lipnick, S.L., Innes, B.T., Simmons, S.K., Adiconis, X., Dionne, D., Mayweather, B.A., Nguyen, L., Niziolek, Z., Ozek, C., et al. (2019). Single-cell transcriptomic profiling of the aging mouse brain. Nat. Neurosci. *22*, 1696–1708.

69. Kester, L., and van Oudenaarden, A. (2018). Single-cell transcriptomics meets lineage tracing. Cell Stem Cell *23*, 166–179.

70. Zheng, X., Huang, Y., and Zou, X. (2020). scPADGRN: a preconditioned ADMM approach for reconstructing dynamic gene regulatory network using single-cell RNA sequencing data. PLoS Comput. Biol. *16*, e1007471.

71. Ma, X., Sun, P., and Qin, G. (2017). Identifying condition-specific modules by clustering multiple networks. IEEE/ACM Trans. Comput. Biol. Bioinform. *15*, 1636–1648.

72. Ma, X.K., Dong, D., and Wang, Q. (2019). Community detection in multi-layer networks using joint nonnegative matrix factorization. IEEE Trans. Knowledge Data Eng. *31*, 273–286.

73. Chen, Y.X., Khong, S.Z., and Georgiou, T.T. (2016). On the definiteness of graph Laplacians with negative weights: geometrical and passivity-based approaches. Proc. Am. Contr. Conf. 2488–2493.

74. Belkin, M., and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. *15*, 1373–1396.

75. von Luxburg, U. (2007). A tutorial on spectral clustering. arXiv, 0711.0189.

76. Brennecke, P., Anders, S., Kim, J.K., Kolodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C., et al. (2013). Accounting for technical noise in single-cell RNA-seq experiments. Nat. Methods *10*, 1093–1095.

77. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Statist. Soc. Ser. B (Methodological) *57*, 289–300.

78. Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. *44*, W90–W97.

79. Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R., and Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics *14*, 128.

80. Rios, O., Frias, S., Rodriguez, A., Kofman, S., Merchant, H., Torres, L., and Mendoza, L. (2015). A Boolean network model of human gonadal sex determination. Theor. Biol. Med. Model. *12*, 26.

81. Kim, S. (2015). Ppcor: an R package for a fast calculation to semi-partial correlation coefficients. Commun. Stat. Appl. Methods *22*, 665–674.

82. Specht, A.T., and Li, J. (2017). LEAP: constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering. Bioinformatics *33*, 764–766.

83. Moerman, T., Aibar Santos, S., Bravo Gonzalez-Blas, C., Simm, J., Moreau, Y., Aerts, J., and Aerts, S. (2019). GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. Bioinformatics *35*, 2159–2161.

84. Deshpande, A., Chu, L.-F., Stewart, R., and Gitter, A. (2019). Network inference with granger causality ensembles on single-cell transcriptomic data. bioRxiv, 534834, https://doi.org/10.1101/534834.

85. Papili Gao, N., Ud-Dean, S.M.M., Gandrillon, O., and Gunawan, R. (2018). SINCERITIES: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. Bioinformatics *34*, 258–266.

86. Aubin-Frankowski, P.-C., and Vert, J.-P. (2020). Gene regulation inference from single-cell RNA-seq data with linear differential equations and velocity inference. Bioinformatics. https://doi.org/10.1093/bioinformatics/btaa576.

87. Matsumoto, H., Kiryu, H., Furusawa, C., Ko, M.S.H., Ko, S.B.H., Gouda, N., Hayashi, T., and Nikaido, I. (2017). SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. Bioinformatics *33*, 2314–2321.

88. Sanchez-Castillo, M., Blanco, D., Tienda-Luna, I.M., Carrion, M.C., and Huang, Y. (2018). A Bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data. Bioinformatics *34*, 964–970.

89. Woodhouse, S., Piterman, N., Wintersteiger, C.M., Gottgens, B., and Fisher, J. (2018). SCNS: a graphical tool for reconstructing executable regulatory networks from single-cell genomic data. BMC Syst. Biol. *12*, 59.

90. Qiu, X., Rahimzamani, A., Wang, L., Ren, B., Mao, Q., Durham, T., McFaline-Figueroa, J.L., Saunders, L., Trapnell, C., and Kannan, S. (2020). Inferring causal gene regulatory networks from coupled single-cell expression dynamics using Scribe. Cell Syst. *10*, 265–274.e11.