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, *], James J. Cai[1,3,4, *]

[1]Department of Veterinary Integrative Biosciences; [2]Department of Statistics; [3]Department of Electrical and Computer Engineering; [4]Interdisciplinary Program of Genetics; Texas A&M University, College Station, TX 77843, USA.

[*]Corresponding authors: J.Z.H. (jianhua@stat.tamu.edu) and J.J.C. (jcai@tamu.edu)

Keywords: sing-cell RNAseq, gene regulatory network, machine learning, principal component regression, tensor decomposition, manifold alignment

Abstract

Constructing and comparing gene regulatory networks (GRNs) from single-cell RNA sequencing (scRNAseq) data present formidable computational challenges. Many existing solutions lack effectiveness or efficiency, due to technical and analytical issues of scRNAseq such as random dropout, uncharacterized cell states, and heterogeneous samples. Here, we present a robust, unsupervised machine learning workflow, called scTenifoldNet, to improve on existing solutions. The scTenifoldNet workflow combines principal component regression, low-rank tensor approximation, and manifold alignment. It constructs and compares transcriptome-wide single-cell GRNs (scGRNs) from different samples to identify gene expression signatures shifting with cellular activity changes such as pathophysiological processes and responses to environmental perturbations. We used simulated data to benchmark the performance of scTenifoldNet. Application of scTenifoldNet on three real data sets shows it to be a powerful tool to reveal biological insights, which are otherwise difficult to obtain. In particular, scTenifoldNet identified highly specific shifts in transcriptional regulation associated with aging and acute morphine responses in mouse cortical neurons, as well as those associated with double-stranded RNA-induced immune responses in human dermal fibroblasts.

Highlights

- scTenifoldNet constructs and compares single-cell gene regulatory networks (scGRNs)
- scTenifoldNet is built upon principal component regression (PCR), low-rank tensor approximation, and manifold alignment

- We applied scTenifoldNet to single-cell RNA sequencing (scRNAseq) data from mouse cortical neurons and human dermal fibroblasts
- Specific gene expression signatures, undetected using other methods, were identified in all case studies

Short abstract

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) from single-cell RNAseq (scRNAseq) data. By comparing scGRNs, scTenifoldNet reveals regulatory changes in gene expression between samples. With real data, scTenifoldNet identifies specific gene expression programs associated with aging and morphine responses in mouse cortical neurons, and with induced immune responses in human dermal fibroblasts. These results provide critical biological insights into the underlying regulatory networks governing cellular transcriptional activities.

Main

Gene regulatory network (GRN) is a graph representation of intricate interactions between transcription factors (TFs), associated proteins, and their target genes, reflecting the current physiological condition of given cells. Studying GRNs facilitates the understanding of cell state, cell function, and the regulatory mechanisms underlying the complexity of cell behaviors. Therefore, to this end, several methods for constructing GRNs from gene expression data have been developed [1-4]. Comparing GRNs constructed from different data sets is important since comparative analysis can determine which parts of transcriptome change. The results will help to understand what is the most significant regulatory shift between samples, how genetic and environmental signals are integrated to regulate physiological responses of a cell population and how cells behavior upon different perturbations. All of these are key questions in the study of the functional involvement of a given GRN. Despite the critical importance of comparative GRN analysis, relatively few methods for comparing GRNs have been developed [5].

In the past few years, single-cell RNAseq (scRNAseq) technology has revolutionized biomedical sciences. This technology brings an unprecedented level of resolution with rich information for studying transcriptional regulation, cellular history, and cell interactions. It transforms previous whole tissue-based assays to single-cell transcriptomic readouts and greatly improves our understanding of cell development, homeostasis, and disease. Current scRNAseq systems (e.g., 10x Genomics) can profile transcriptomes for thousands of cells per experiment. This 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 to reveal fundamental changes in regulatory networks and unravel the differences in network dynamics that govern the behaviors of cells. Because our ability to generate scRNAseq data has outpaced our ability to extract information from it, there is a clear need to develop efficient computational algorithms and novel statistical methods to analyze and exploit information embedded within GRNs [9].

Constructing single-cell GRNs (scGRNs) from scRNAseq data and effectively comparing them present great analytical challenges [9, 10]. A meaningful comparison between scGRNs requires a robust construction of scGRN. Comparing biased scGRNs constructed using any unstable solution would result

in misleading outcomes and inappropriate conclusions. The pronounced cell-to-cell variability, high dimensionality, and sparsity of scRNAseq data complicate the construction of unbiased scGRNs. Often, the expression of a gene is governed by stochastic processes and also influenced by transcriptional activities of many other genes, making it 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 scRNAseq data, as the information is deeply embedded in differences in highly complex scGRNs between two samples.

In this paper, we introduce a workflow for constructing and comparing scGRNs using scRNAseq data collected from two samples. The workflow, which we call scTenifoldNet, integrates several unsupervised machine learning techniques. scTenifoldNet can be used as a very powerful framework for comparative analyses of scGRNs to identify specific changes in gene expression signatures and regulatory network rewiring events. The inputs of scTenifoldNet are two scRNAseq expression matrices from different samples. For example, the two samples subject to comparison may be from healthy and diseased donors. The two input expression matrices are processed simultaneously through a multistep procedure. The outputs are a list of genes regulated differently in the two tested samples. The constructed scGRN can also be used to identify functionally significant modules, i.e., subsets of closely regulated genes.

scTenifoldNet is an innovative method. We are not aware of any prior work that uses similar design or achieves the same levels of accuracy and efficiency. It overcomes technical challenges in implementing effective and efficient GRN methods for scRNAseq data. Here, we first benchmarked and demonstrated the utility of scTenifoldNet across synthetic data sets and then applied scTenifoldNet to real data sets [11-13]. Our real data analyses showed the power of scTenifoldNet in identifying significant genes and modules whose regulatory patterns are greatly shifted between the two samples. None of these findings were reported in the respective original studies, in which the data sets 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 unsupervised 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 component representations of scGRNs constructed from different scRNAseq expression matrices. **Fig. 1** shows the main components of scTenifoldNet architecture. The whole workflow contains five key steps: subsampling cells, constructing multilayer scGRNs, denoising, manifold alignment, and module detection. In order to produce a desirable result, 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 **Methods**.

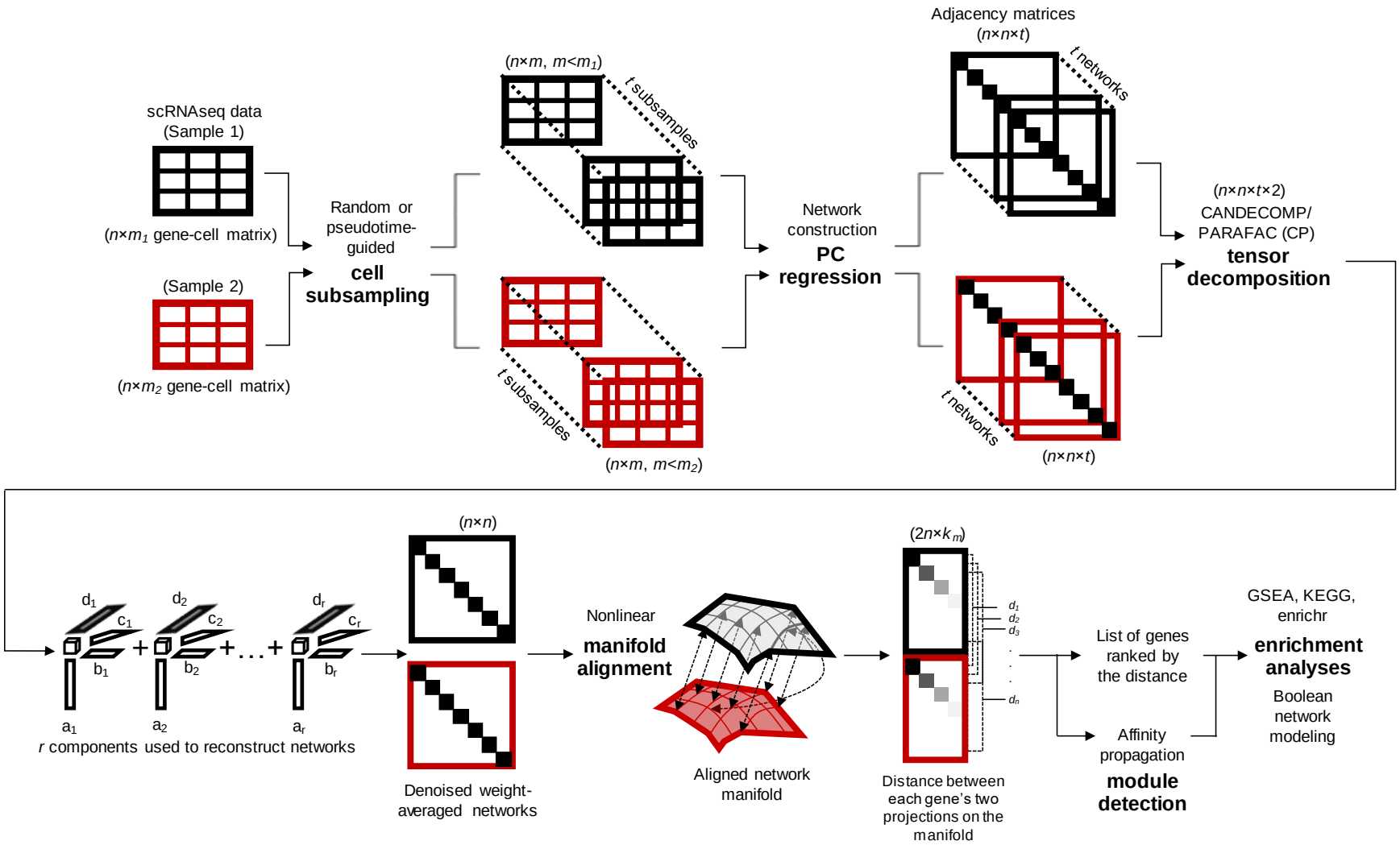

Fig. 1. Overview of the scTenifoldNet workflow. scTenifoldNet is a machine learning framework that uses the scRNAseq network approach to identify regulatory changes between samples. scTenifoldNet is composed of five major steps, starting with subsampling cells in scRNAseq expression matrices. When two samples are analyzed, each of the two samples is subsampled (either randomly or following a pseudotime trajectory of cells, see **Discussion**), independently. 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. Principal component (PC) regression is used for scGRN construction; each scGRN is represented as a weighted adjacency matrix. Two samples produce two multilayer GRNs, form a fourth-order tensor (or two three-order tensors), which is subsequently decomposed into multiple components. Top components of tensor decomposition are then used to reconstruct two denoised multilayer scGRNs. Then, two denoised multilayer scGRNs are collapsed by taking average weight across layers, respectively. 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. The distance between the two data points is the relative difference of the gene in its regulatory relationships in the two scGRNs. Finally, genes are ranked by distance, and an affinity propagation algorithm is used to detect modules with top-ranked genes as seeds in the scGRNs.

Numerical methods

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

Step 1. Pre-processing data and subsampling cells: The input data are two scRNAseq expression data matrices, $\boldsymbol{X}$ and $\boldsymbol{Y}$, containing expression values *x* and y for *n* genes in m_1 and m_2 cells from two different samples, respectively. Next, *m* cells in $\boldsymbol{X}$ and $\boldsymbol{Y}$ are randomly sampled to form $\boldsymbol{X}'$ and $\boldsymbol{Y}'$, where $m < \min(m_1, m_2)$. This subsampling process repeats *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 $\boldsymbol{X}'_i \in \{\boldsymbol{X}'_i\}$ and each $\boldsymbol{Y}'_i \in \{\boldsymbol{Y}'_i\}$, $i = 1, 2, \ldots, t$, principal component regression (PCR) is used to construct a GRN. Constructed GRNs from $\boldsymbol{X}'_i$ and $\boldsymbol{Y}'_i$ are stored as signed, weighted and directional graphs represented with *n*×*n* adjacency matrices, $\boldsymbol{W}_i^x$ and $\boldsymbol{W}_i^y$, respectively. Diagonal values of each matrix are set to zeros, and other values are normalized by dividing by their maximal absolute value. Each adjacency matrix is then filtered, retaining only the top 5% of edges ranked using the absolute weight. This results in a sparse adjacency matrix.

Step 3. Denoising: Tensor decomposition [14] is used to denoise the adjacency matrices obtained in Step 2. The collection of t scGRNs for each sample, $\{\boldsymbol{W}_i^x\}$ and $\{\boldsymbol{W}_i^y\}$, is processed independently as a third-order tensors, $\mathfrak{T}^x$ and $\mathfrak{T}^y$, each containing *n*×*n*×*t* elements. Alternatively, $\{\boldsymbol{W}_i^x\}$ and $\{\boldsymbol{W}_i^y\}$ can be processed jointly as a fourth-order tensor, $\mathfrak{T}^{xy}$ containing *n*×*n*×*t*×*2* 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 top k_t 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 $\boldsymbol{W}_d^y$, then normalized as in step 2.

Step 4. Aligning genes onto a manifold: $\boldsymbol{W}_d^x$ and $\boldsymbol{W}_d^y$ are regarded as two pointwise similarity matrices for a nonlinear manifold alignment. The alignment is done by solving an eigenvalue problem with a Laplacian graph of the joint matrices: $\boldsymbol{W} = [\boldsymbol{W}_d^x, \lambda \boldsymbol{I}/2; \lambda \boldsymbol{I}^T/2, \boldsymbol{W}_d^y]$, where $\boldsymbol{I}$ is the identity matrix and λ is a tuning parameter, which reflects the binary correspondence between genes in $\boldsymbol{X}$ and $\boldsymbol{Y}$. The algorithm projects each gene's *n* features in $\boldsymbol{W}_d^x$ and $\boldsymbol{W}_d^y$ as two data vectors on the shared manifold with a lower dimension $k_m << n$.

Step 5. Ranking genes and detecting modules: For each gene j, d_j is the Euclidean distance between the gene's two projections on the shared manifold: one is from $\boldsymbol{W}_d^x$ and the other is from $\boldsymbol{W}_d^y$. Genes are sorted according to the distance. The greater the distance, the greater the regulatory shift. An affinity propagation algorithm is employed on the low dimensional representations of $\boldsymbol{W}_d^x$ and $\boldsymbol{W}_d^y$ to detect modules.

In the next few sections, we explain the rationale behind each step, the selection of machine learning methods, and implementation details.

Subsampling of cells

Our rationale for subsampling cells is similar to that of ensemble learning, a strategy in which the decisions from multiple models are combined to improve the overall performance. Instead of trying to construct a single scGRN, scTenifoldNet focuses on subsampling cells from a given scRNAseq expression matrix, constructing a number of 'low-accuracy' scGRNs from the subsampled data sets, and then

combining these scGRNs to obtain a high-accuracy scGRN. As mentioned, current scRNAseq technology can produce thousands of cells' transcriptome profiles from each sample. Processing scRNAseq data of high dimensionality and large scale is inherently difficult, especially considering the degree of heterogeneity in the scRNAseq data from many such cells. For example, the presence of outlying cells, i.e., cells showing expression profiles deviate from those of most other cells, may influence scGRN construction. Subsampling, therefore, offers promise as a technique for handling the noise in the input data sets. When the number of cells is small, resampling of the input data matrix with replacement can be applied.

Construction of scGRNs using principal component regression (PCR)

Although many GRN construction methods have been developed [1, 2, 4], it is unclear which one is suitable for constructing a large number of scGRNs from the subsampled data [9]. When dealing with multiple sets of input data, both accuracy and computational efficiency of the algorithms have to be considered. After performing a systematic evaluation of existing methods, we chose to use the PCNet method [5], which is based on PCR [15]. The PCR extracts the first few (e.g., $k = 3$) principal components and then use 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 main use of PCR 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 data sets is regarded as a rank-three tensor. To reduce the noise in the multilayer scGRN, we decompose the tensor and regenerate the multilayer scGRN using truncated components. The idea is similar to that of the truncation of a full-rank singular value decomposition (SVD) matrix by lower ranks. After cutting a larger portion of the noise spread over the lowest singular value components, the regenerated data matrix based on the truncated SVD would, therefore, approximate the original data with reduced noise. Indeed, tensor decomposition has been used in video data analyses for denoising and information extracting purposes [16]. It has also been used to impute missing data [17]. We use the CANDECOMP/PARAFAC (CP) algorithm [18] to factorize the two multilayer scGRNs independently and regenerate all adjacency matrices using top 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 for two purposes: denoising and enhancing, i.e., making main signals stronger and making less important signals weaker.

Manifold alignment of scGRNs from two samples

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—i.e., commonly referred to as the manifold assumption [19]. 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 data sets lie on a common manifold [20-23]. 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 [24, 25]. 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 detecting modules in scGRN

To identify genes whose regulatory status differs between the two samples, we calculated 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 genes, one from each sample. We computed the Euclidean distance between the two data points of the gene and used the distance to measure the dissimilarity in the gene's regulatory status in two scGRNs. We do this for all genes shared between two samples and then rank genes by the distance. The larger the distance, the more different the gene in two samples. In this way, we were able to produce a list of ranked genes. These ranked genes are subject to functional annotation such as using the pre-ranked gene set enrichment analysis (GSEA) [26] to assess the enriched functions associated with the top genes. To avoid choosing the number of selected genes arbitrarily, we computed *p*-values for genes using Chi-square tests, adjusted *p*-values with a multiple testing correction, and selected significant genes using the 10% FDR cutoff. Furthermore, we used the coordinates in the computed aligned manifolds for the identified genes with significant regulatory shifts to generate a similarity matrix $\boldsymbol{S}$ using a linear kernel, and then used the affinity propagation algorithm [27, 28] to detect modules in $\boldsymbol{S}$. Affinity propagation assigns labels to previously unlabeled data points, providing a way for expanding significant individual genes into modules.

Benchmarking the performance of scTenifoldNet using simulated data

Accuracy and recall of PCR for network construction

To show the effectiveness of PCR, we simulated scRNAseq data using a parametric method with a pre-defined scGRN model (see **Methods** for details). With these simulated data whose generating parameters are known, we compared the scGRNs constructed using PCR against the scGRN predefined (i.e., ground truth) to estimate the accuracy of PCR. Similarly, we tested the accuracy of other methods, including the Spearman correlation coefficient (SCC), mutual information (MI) [1], and GENIE3 (a random forest-based method [2, 7]). 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 performances in recovering gene regulatory relationships were compared. We found that the PCR method tended to produce more specific (better accuracy) scGRNs more sensitively (better recall) than the other three methods across a wide range of settings of cell numbers in the input scRNAseq expression matrix (**Fig. 2A**). PCR is also much faster than GENIE3. For instance, on

a typical workstation, our implementation of PCR can construct a GRN for all-by-all 15,000 genes in less than 50 minutes, whereas GENIE3 requires more than 24 hours (data is available in **Supplementary Table S1**). Thus, using the ground-truth interactions between genes generated according to pre-setting parameters, we confirmed that PCR outperforms to all other methods tested.

Effect of tensor denoising

To show the effect of tensor denoising, we simulated scRNAseq data (see **Methods**) and processed the data using the first two steps of scTenifoldNet, i.e., cell subsampling followed by the construction of scGRNs using PCR. We subsampled 500 cells each time and generated 10 scGRNs. The 10 scGRNs are 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. **Fig. 2B** contains the heatmaps of adjacency matrices of the 10 scGRNs before and after denoising (small panels). We also show two collapsed scGRNs (**Fig. 2B**, large panels), 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%. Thus, tensor denoising proves to be a solution for removing impacts of random dropout and other noise issues afflicting scRNAseq data on the scGRN construction.

Detecting power of the scTenifoldNet workflow illustrated with a toy data set

We used simulated data to show the capability of scTenifoldNet in detecting differentially regulated (DR) genes. We fist used the negative binomial distribution to generate a sparse synthetic scRNAseq data set (an expression matrix including 67% zeros in its values). This toy data set 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) against itself (sample 1 vs. sample 1) and also against the manually perturbed version (sample 1 vs. 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 DR genes (**Fig. 2C**, $P < 0.05$). These results are expected and support the sensitivity of scTenifoldNet in identifying subtly shifted gene expression programs.

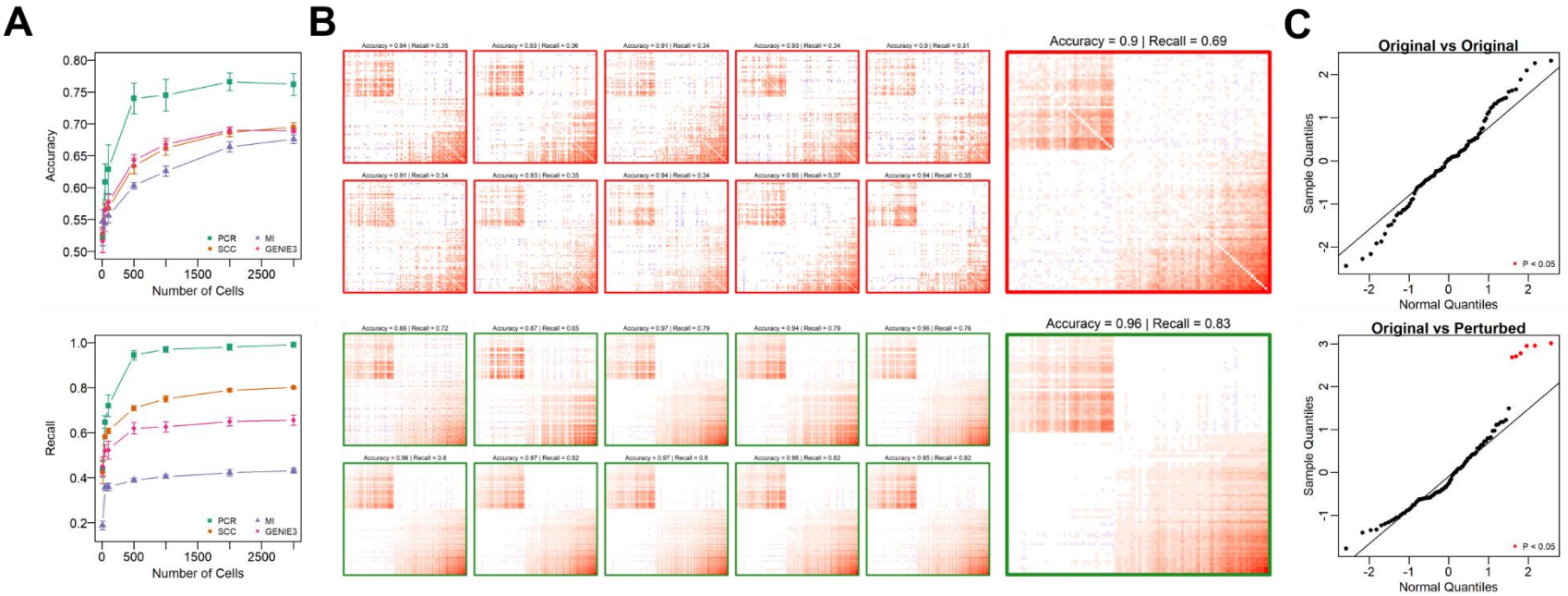

Fig. 2. Benchmarking the performance of scTenifoldNet using simulated data. (**A**) The accuracy and recall of scGRN construction using different methods: PCR, SCC, MI, and GENIE3, as functions of the number of cells used in the analysis. Error bar is the standard deviation of the computed values after 10 bootstrapped evaluations. Accuracy is defined as = (TP+TN)/(TP+TN+FP+FN), and recall = TP/(TP+FN), where T, P, F, and N stands for true, positive, false and negative, respectively. PCR – PC regression; SCC – Spearman 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 PCR over the counts of 500 randomly subsampled cells. Colors indicate the relative strength of regulatory relationships between genes. (**C**) Evaluation of the sensibility of scTenifoldNet to identify punctual changes in the regulatory profiles, on the left, evaluation of the original matrix against itself, on the right, evaluation of the original matrix against the perturbed one. Identified genes under differential regulation (FDR < 0.05, the B–H correction) are labeled in red.

Real data analyses

Analysis of transcriptional responses to morphine in mouse cortical neurons

To illustrate the use of scTenifoldNet, we first applied it to a scRNAseq data set from [11]. This is a study on the mouse neural cells' transcriptional responses to morphine. The authors performed scRNAseq experiments with the nucleus accumbens (NAc) of mice after four hours of the morphine treatment, using mice treated with saline as mock controls. The authors obtained expression data for neurons from four morphine- and four mock-treated mice. Differential expression (DE) analysis showed that 256 genes are differentially expressed between morphine- and mock-treated NAc (see **Table S2** of [11]). However, no specific function was found to be enriched with these DE genes. Also, the difference in the expression level of these genes was only observed in an uncharacterized subpopulation of neurons. Overall, it seems that this scRNAseq data does not support the hypothesis that molecular and behavioral responses to opioids are primarily mediated by neurons.

We were intrigued by such a negative result and set out to work on the data. We first reproduce 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 tSNE plot; expression levels of several known morphine responsive genes, e.g., *Adcy5*, *Ppp1r1b,* and *Ppp3ca*, show no difference (**Fig. 3A**). Thus, we confirmed the result from the original authors, that is, a direct comparison of gene expression between neurons using the DE method could not identify relevant genes known to be involved in the morphine response.

Next, using scTenifoldNet, we identified 59 genes (***Hpca***, ***Ppp3ca***, *Pcp4l1,* ***Akap5***, *Rgs7bp, Atp2a2, Slc24a2, Foxp1,* ***Atp2b1***, *Spock3,* ***Ppp1r1b***, ***Adcy5***, ***Rgs9***, *Arpp19,* ***Gpr88***, *Actb, Ubb,* ***Calm2***, *Phactr1, Rnr1,* ***Penk***, ***Arpp21***, ***Gnal***, *Cck,* ***Eif1***, *Nd1, Cytb, Spred1,* ***Scn4b***, *Nd2,* ***Cadm2***, *Co1,* ***Rasd2***, *Nrn1, Hspa4l,* ***Chn1***, *Diras2, Cpe,* ***Ramp1***, ***Grin2b***, ***Chst15***, *B3galt2, Lamp1,* ***Pde1b***, *Gabrg1, Chst1, 3110035e14rik, Nfia,* ***Scn8a***, *Sacs, Samd4, Nrxn1,* ***D3bwg0562e***, *Rtn1, Akap9, Ahsa1, D430041d05rik, Syn2,* and ***Klf9***) showing significant differences in their regulatory patterns, indicated by greater distances between genes' positions in the aligned manifold of two scGRNs (FDR < 0.05, Chi-square test with B-H multiple test adjustment, see **Methods** for details). These DR genes are regulated differentially between two samples (**Fig. 3B**). These genes are enriched for *opioid signaling, signaling by G protein-coupled receptors*, *reduction of cytosolic Calcium levels,* and *morphine addiction*. 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 the calcium-channel conduction [29-31]. Furthermore, 24 (highlighted in bold) of the 59 identified DR genes were found to be targets of RARB (40%, adjusted P-value < 0.01, enrichment test by Enrichr [32] based on results of ChIP-seq studies [33]). RARB encodes for a plastic TF playing a role in synaptic transmission in dopaminergic neurons and the adenylate cyclase-activating dopamine receptor signaling pathway [34, 35]. Thus, all these enriched functions are highly relevant to morphine. Indeed, morphine stimulus is known to induce the disinhibition of dopaminergic neurons by GABA transmission, enhancing the dopamine release and causing addiction [36, 37]. Using scTenifoldNet, we also identified modules that are enriched with DR genes (**Fig. 3C**).

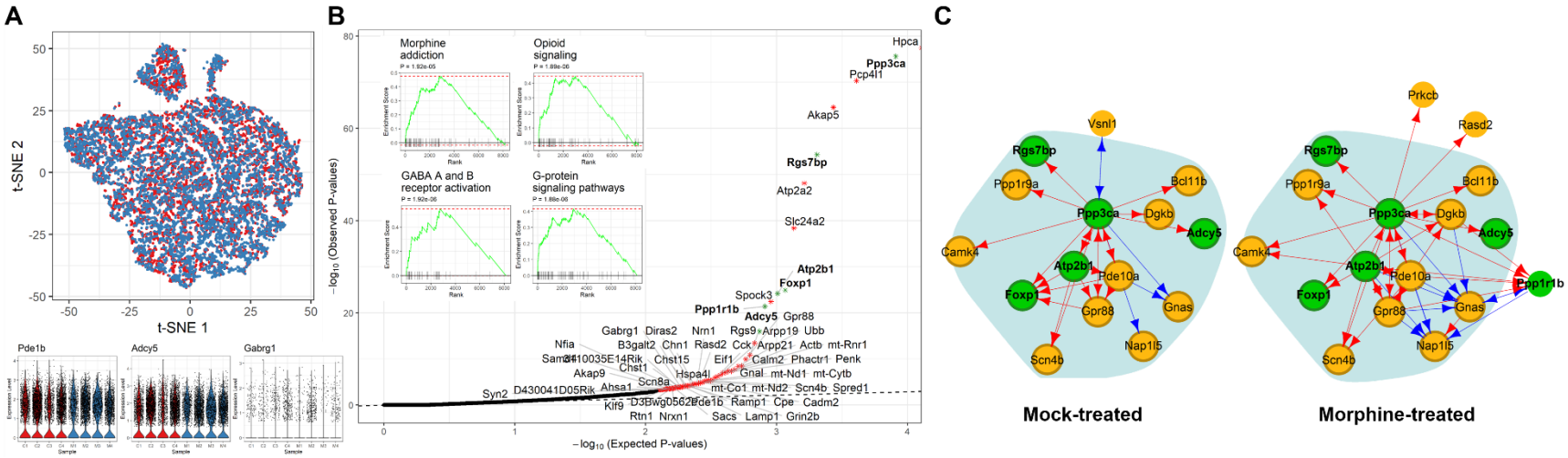

Fig. 3. Analysis of transcriptional responses to morphine in mouse cortical neurons. (**A**) A t-SNE plot of 7,972 and 8,912 neurons from morphine-treated (blue) and mock-treated (red) mice, respectively. Violin plots show the log-normalized expression levels of selected DR genes in four (M) morphine- and four (C) mock-treated mice. (**B**) Quantile-quantile (Q-Q) plot for observed and expected p-values of the 8,138 genes tested. Genes (n=67) with FDR < 0.1 are shown in red; genes (n=59) with FDR < 0.05 are labeled with asterisk. Inset shows results of the GSEA analysis for genes ranked by their distances in manifold aligned scGRNs from morphine- and mock-treated mice. (**C**) The module enriched with DR genes and the corresponding subnetworks in two scGRNs. The module is centered on the DR gene, *Ppp3ca*. All DR genes in the module are in green and bold. Edges are color-coded: red indicates a positive association, and blue indicates negative. Weak edges are filtered out from two scGRNs for visualization, and the background shadow indicates the shared genes.

Analysis of transcriptional response to aging in mouse cortical neurons

To further illustrate the power of scTenifoldNet in identifying gene modules associated with specific conditions, we applied scTenifoldNet to another published data set [12]. In that study [12], the authors focused on the aging-related gene expression changes that occur in the mouse brain. The authors obtained single-cell transcriptomes of 50,212 neural cells (24,401 young and 25,811 old) derived from the brains of eight young (2–3 months) and eight old (21–23 months) mice. These neural cells were classified into four major subtypes: oligodendrocyte, astrocyte, neuron, and microglia. In neurons, 206 DE genes that passed the 10% fold-change (FC) threshold were identified (**Supplementary Table** 8 of [12]). However, among these DE genes, only one gene (*Rpl8*) is known to be associated with aging. As the authors pointed out: the calculation of FC is dependent on several factors, including the number of cells within each population, the level of transcription, and the algorithm for analysis. For instance, most genes show no differences in their expression levels across cells between samples (**Fig. 4A**). Therefore, it is difficult to improve the power of DE analysis methods in analyzing scRNAseq data.

We were again inspired and re-analyzed the data using scTenifoldNet. We identified 25 DR genes: ***Gria2**, **Gad1**, **Pbx1**, Inpp5j, Ptpro, **Malat1**, **Meis2**, **Atp1b1**, **Ndrg4**, **Camk2b**, Nrxn3, **Celf2**, **Calm2**, Shisa8, Cox7c, **Prkca**, Rps8, **Dclk1**, **Cox4i1**, Rps19, **Ckb**, Ubb, Itm2b, Gpsm1*, and Rps*27a*, showing significant differences in their transcriptional regulation patterns between scGRNs constructed using brain samples from old and young mice (FDR < 0.05, **Fig. 4B**). Among these DR genes, 14 (highlighted in bold) are targets of the same master TF—*RARB*, which play a role in aging responses through regulating the scavenging of reactive oxygen species [38-40]. The enrichment is significant (14 out of 25 = 56%, adjusted *p*-value < 0.01, enrichment test by Enrichr [32]). Among them, *Malat1* is a long non-coding RNA gene known to be associated with the regulation of oxidative stress [41-43]. Taken together, these DR genes are functionally enriched for *circadian entrainment, nitric oxide signaling pathway, unblocking of NMDA receptor, glutamate binding and activation, Parkinson's disease,* and *dopaminergic synapse*. The GSEA analysis shows that shifted regulatory mechanisms are also associated with the *oxidative phosphorylation pathway* and the *electron transport chain*. Thus, scTenifoldNet enabled the identification of gene expression signatures in neurons associated with mouse brain aging. Many of these signatures are well known. For example, aging causes dysregulation of the oxidative phosphorylation, lower cellular energy production, and higher oxygen free radicals [44, 45], as seen in Parkinson's and Alzheimer's diseases [44-47]. It is also known that aging is associated with lower nitric oxide synthase activity and impaired NMDA receptor signaling [48].

It is worth noting that most of DR genes showed no differences in their expression levels across neuronal cells between young and old mice. That is to say, scTenifoldNet detects the shift of regulatory patterns in genes, but not the difference in genes' expression level. Many DR genes are closely regulated and appear in the same modules (e.g., **Fig. 4C**). These modules contain enriched information regarding specific regulatory relationships between relevant genes.

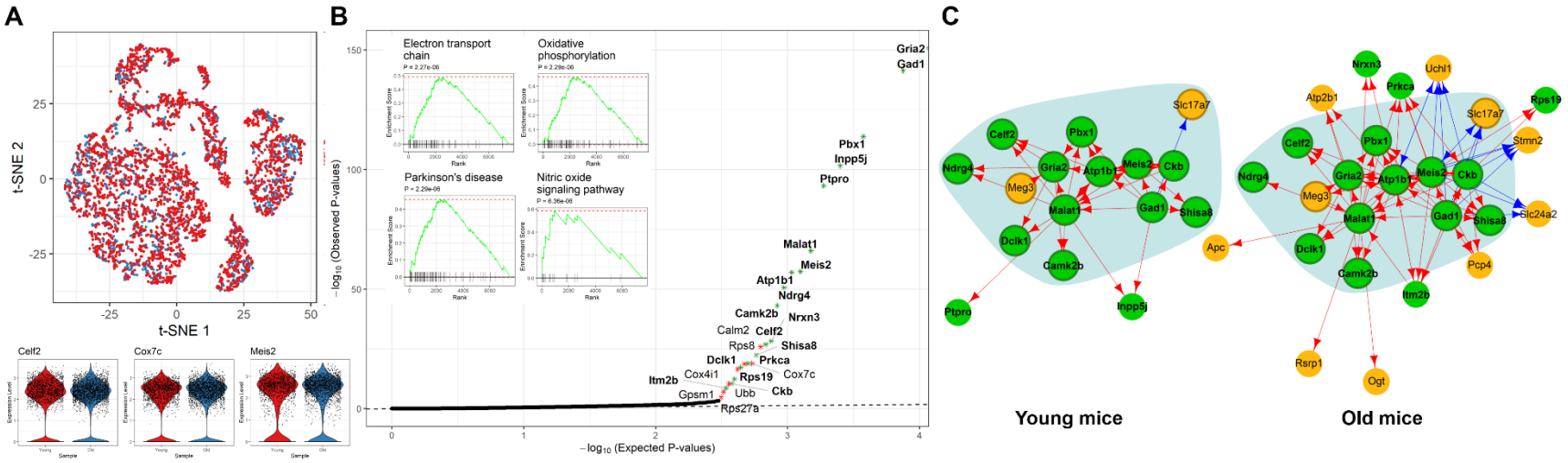

Fig. 4. Analysis of transcriptional responses to aging in mouse cortical neurons. (**A**) t-SNE plot of 2,047 and 3,082 neurons from young (red) and old (blue) mice, respectively. Violin plots show the log-normalized expression levels of selected DR genes in young (red) and old (blue) mice. (**B**) Q-Q plot for observed and expected p-values of the 7,503 genes tested. Genes (n = 25) with FDR < 0.05 are labeled with asterisk. Inset shows the results of the GSEA analysis for genes ranked by their distances in manifold aligned scGRNs from young and old mice. (**C**) A representative module enriched with DR genes and the corresponding subnetworks in two scGRNs. The module is centered on the DR gene, *Meis2*. All DR genes in the module are in bold. Edges are color-coded: red indicates a positive association, and blue indicates negative. Weak edges are filtered out from two scGRNs for visualization, and the background shadow indicates the shared genes.

Analysis of transcriptional response to double-stranded RNA in human dermal fibroblasts

Finally, we show the use of scTenifoldNet to scRNAseq data from human dermal fibroblasts [13]. The authors focused on single-cell transcriptional responses induced by the stimulus of the synthetic double-stranded RNA (dsRNA) (**Fig. 5A**). They obtained and compared transcriptomes of 2,553 unstimulated and 2,130 stimulated cells and identified 875 DE genes (Table S3 in [13]). These DE genes include *IFNB, TNF, IL1A*, and *CCL5*, encoding antiviral and inflammatory gene products, and are enriched for *inflammatory response, positive regulation of immune system process,* and *response to cytokine*, among many others biological processes and pathways.

After re-analyzing the data using scTenifoldNet, we identified 28 DR genes: ***RPL3****, RPL13A,* ***RPLP1****,* ***RPS6****,* ***RPS14****,* ***EEF1A1****,* ***RPL6****,* ***RPL15****, RPS4X,* ***RPS12****,* ***RPL10A****,* ***RPS15A****,* ***RPS18****,* ***RPL10****,* ***RPL13****,* ***RPL18A****,* ***RPL27A****,* ***GNB2L1****,* ***RPS3A****, RPS19,* ***RPS9****, SELM,* ***TPT1****,* ***RPL12****,* ***RPL11****,* ***RPL28****, NUPR1,* and *FTL*. These genes are functionally enriched for *response to an infectious disease*, *reduction of the peptide chain elongation*, *eukaryotic translation elongation*, and *viral mRNA translation*. GSEA analysis shows that regulatory changes are associated with the *interferon alpha/beta signaling*, and the *viral RNA transcription and replication* pathways (**Fig. 5B**). Furthermore, 23 (highlighted in bold) of the 32 identified DR genes were found to be targets of TF gene *MYC* [49] (71%, adjusted p-value < 0.01, enrichment test by Enrichr [32]).

Through comparing DR genes with the DE genes identified in the original paper [13], we found that enriched functions of DE genes reflect merely a final status of cells after cells responding to the dsRNA stimulus, whereas the enriched functions of DR genes reflect the ongoing regulatory processes and immune responses to the stimulus, which are valuable for informing the mechanisms through which the dsRNA acts to induce immunological responses [50-52]. For example, it is known that the dsRNA inhibits the translation of mRNA to proteins [51] and leads the synthesis of interferon, which induces the synthesis of ribosomal units that are able to distinguish between cell mRNA and viral RNA [52]. Interferon also promotes cytokine production that activates the immune responses and induces inflammation [50]. To further illustrate the changes in the regulatory mechanisms between the unstimulated and stimulated samples identified by scTenifoldNet, we present a module containing significant DR genes centered at *RPL13*, as well as scatter plots for a pair of DR genes (*ANXA2-RPS12*), showing the changes of co-expression patterns cause by the stimulus of dsRNA (**Fig. 5C**).

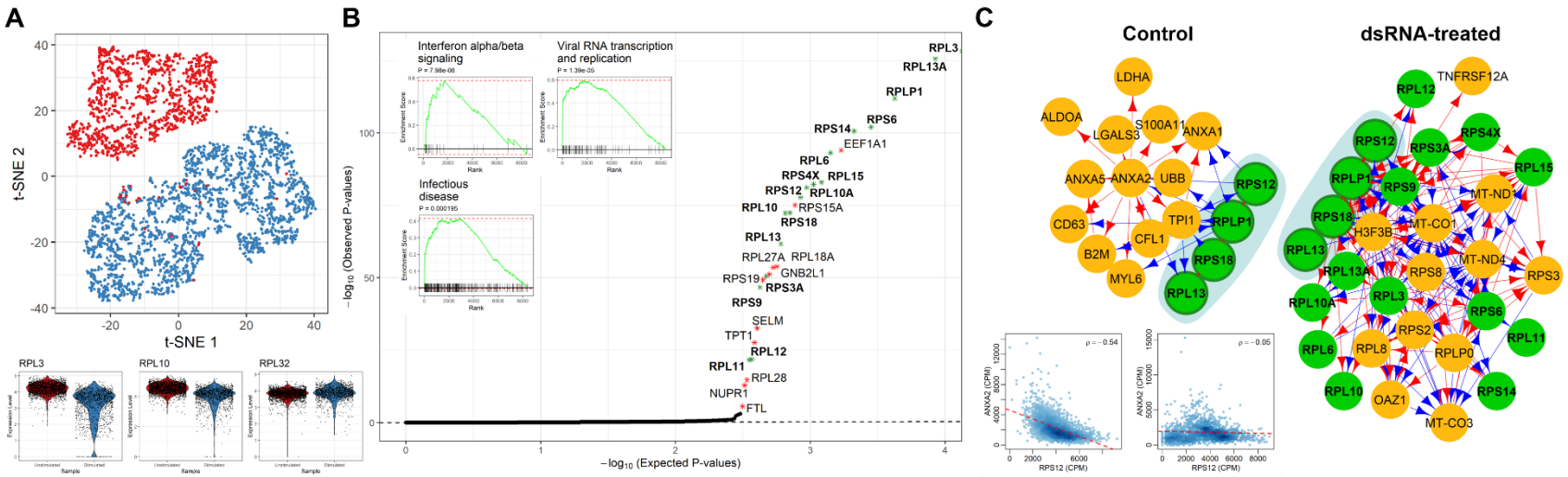

Fig. 5. Analysis of transcriptional responses to dsRNA stimulus in human dermal fibroblasts. (**A**) t-SNE plot of 2,478 and 1,358 fibroblast cells before (red) and after (blue) the dsRNA stimulus, respectively. Violin plots show the log-normalized expression levels of selected DR genes before (red) and after (blue) stimulus. (**B**) Q-Q plot for observed and expected p-values of 8,454 genes tested. Genes (n = 28) with FDR < 0.05 are labeled with asterisk. Inset shows the results of the GSEA analysis for genes ranked by their distances in manifold aligned scGRNs. (**C**) A representative module enriched with DR genes and the corresponding subnetworks in two scGRNs. The module is centered on the DR gene, *RPL13*. All DR genes in the module are in bold. Edges are color-coded: red indicates positive association and blue indicates negative. Weak edges are filtered out from two scGRNs for visualization, and the background shadow indicates the shared genes.

Discussion

Here we presented scTenifoldNet, a machine learning workflow, which streamlines comparative scGRN analyses. As a methodological advance, scTenifoldNet presents a new and efficient way of using machine learning in the analysis of scRNAseq data. The key feature of scTenifoldNet is its ability to filter out noise in heterogenous input scRNAseq data and identify gene expression signatures in a sensitive and scalable fashion. It can be used to reveal subtle regulatory shifts of genes that cause differences in cell population state between samples, making differential regulation (DR) analysis more readily to be adapted. This is significant because the differential expression (DE) analysis is still the primary method for the purpose of comparative analysis between scRNAseq samples (e.g., [11-13]). As scRNAseq data sets are becoming widely available, there will be more and more interest in comparing between samples. The DR analysis using scTenifoldNet is methodologically superior to the DE analysis, given that scTenifoldNet learns features of genes from scGRNs through examining global gene-gene interactions.

To achieve the technical requirements, scTenifoldNet is designed to overcome several analysis hurdles. First, it is currently challenging to construct scGRN from scRNAseq data consisting of cells in many different states. It is also difficult to control for technical noise in the data. To deal with this issue, we let scTenifoldNet start from random subsampling of cells, followed by scGRN construction and network denoising. Second, it is difficult to establish regulatory relationships between genes from scRNAseq data that can capture a relatively complete picture of the gene regulatory landscape. We found that PCR stands out as a crucial scGRN construction method. PCR outperforms the other GRN construction algorithms substantially in all aspects of methodology metrics, including specificity, sensitivity, computational efficiency, and the minimal number of cells required. Importantly, PCR explicitly projects thousands of gene expression measurements into a low dimensional space to capture much of the observed variation. Thus, PCR establishes the relationship for each pair of genes controlling for the most important background interactions. Third, the tensor denoising procedure in scTenifoldNet effectively smooths edge weights across all networks in multilayer scGRNs. Finally, scTenifoldNet performs nonlinear manifold alignment to align two networks. As such, two networks can be contrasted directly and DR genes could be detected using distance in new coordinates of data in a low-dimensional space.

In addition to the function of DR gene identification, we also provide scTenifoldNet the function of module detection. This function is optional but can be useful in many circumstances. The module detection method we adapted is the affinity propagation algorithm. The parameters of the affinity propagation algorithm have an advantage that they can be adjusted to control the size of detected modules, which would benefit the further investigation of the causal relationships between genes. By using the affinity propagation clustering algorithm over the low dimensional representation of the networks, we can identify modules of small size, with size and number defined automatically by optimization procedures. For example, compared to a gigantic module of hundreds of genes, modules of moderate sizes such as those containing 10 to 20 genes, are more ready to be analyzed using network learning algorithms, e.g., boolean network model-based algorithms [53]. Thus, the small number of genes in each module makes many algorithms, which are difficult to scale up, become applicable without concern about the cost of computation.

We show that scTenifoldNet is sensitive to signals that are otherwise undetectable using other analysis methods. This feature makes scTenifoldNet suitable for comparing highly similar samples, such as two populations of cells of the same type. We validate the power of scTenifoldNet with real data sets

generated in mouse neurons and human fibroblasts, from three different studies. Two of the example data sets were from morphine and aging studies, respectively. The third one was from human dermal fibroblasts. With each of these data sets, we found a great discrepancy between our results obtained using scTenifoldNet and the results obtained by authors who produced the original data set. This discrepancy leads to different conclusions, albeit the same set of scRNAseq data was analyzed in each case. This underscores the necessity of the adoption of systems and network-based approaches in scRNAseq data analyses; otherwise, fundamental signals embedded in large data sets are much likely to be overlooked. In the morphine response example, the causal factor of transcriptional responses was pre-defined, i.e., the morphine stimulus, and thus it was known that what we were supposed to recover.

Similarly, in the aging and fibroblast examples, we had some clues on what transcriptional changes we might be able to recover. However, in many biological systems studied, causal genes are unknown. If this is the case, it is crucial to apply the most sensitive approach to unveil those associated genes. Only then will we be able to scrutinize these identified genes further to learn the mechanisms behind their actions in the whole system. In many studies, we are facing such a challenge from unknown factors that cause the disorder. It is, therefore, critical that we use tools such as scTenifoldNet to tackle this big-data analysis problem.

Our method can be extended to adapt a non-random subsampling schema. For example, if cells in an input scRNAseq matrix contain structure shown as a long trajectory in the pseudotime analyses [54], then these cells can be subsampled using a pseudotime-guided method, with which cells are ordered according to their pseudotime and sampled by using sliding windows. In this 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 [55-57]. With our pseudotime-series scGRN data, these algorithms will be relevant and applicable.

In summary, scRNAseq enables the study of cellular, molecular components, and 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 scRNAseq data (e.g., scGRNs) surpass the human capacity for interpretation. We anticipate that our machine learning workflow implemented in scTenifoldNet will help to achieve breakthroughs by deciphering the full cellular and molecular complexity in scRNAseq data through constructing and comparing scGRNs.

Methods

The scTenifoldNet workflow takes two scRNAseq expression matrices as inputs. The two matrices are supposed to be from two samples, 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 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 for t times to produce a set of t subsampled cell populations.

Network construction

For a given expression matrix, a PCR-based network construction method [5] is adopted to construct scGRN. PCR is a popular multiple regression method, where the original explanatory variables are first subject to a PC analysis (PCA) and then the response variable is regressed on the few leading PCs. By regressing on n PCs ($n \ll N$, where N is the total number of genes in the expression matrix), PCR mitigates the over-fitting 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 PCR, 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 the PCR 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 $N-1$ 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 PCR to an scRNAseq expression data matrix is described as follows.

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 the 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 PCR model for X_i. First, we apply PCA to $\boldsymbol{X}_{-i}^T$ and take the first M leading PCs to construct $\boldsymbol{Z}^i = \left(Z_1^i, \cdots, Z_M^i\right) \in \mathbb{R}^{p\times M}$, where $Z_m^i \in \mathbb{R}^p$ is the mth PC of X_{-i}^T. Mathematically, $\boldsymbol{Z}^i = X_{-i}^T \boldsymbol{W}^i$, where $\boldsymbol{W}^i \in \mathbb{R}^{(n-\mathbb{1})\times M}$ is the PC loading matrix for the first M leading PCs, satisfying $\left(\boldsymbol{W}^i\right)^T \boldsymbol{W}^i = \boldsymbol{I}_M$. Secondly, the PCR regresses X_i on $\boldsymbol{Z}^i$ and solves the following optimization problem:

$$\hat{\beta}^i = \arg min_{\beta^i \in R^M} \left\| X_i - \boldsymbol{Z}^i \beta^i \right\|_2^2.$$

Then, $\hat{\alpha}^i = \boldsymbol{W}^i \hat{\beta}^i \in \mathbb{R}^{n-1}$ quantifies the effects of the other $n-1$ genes to the ith gene. After performing PCR on each gene, we collect $\{\hat{\alpha}^i\}_{i=1}^n$ together and construct an $n \times n$ gene-gene interaction matrix, whose diagonal entries are all 0. Then we retain interactions with top α% (=3% 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 the PCR as described above. Each network is represented as a $n \times n$ 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 CANDECOMP/PARAFAC (CP) tensor decomposition is applied. Similar to the truncated singular value decomposition (SVD) of a matrix, the CP decomposition approximates the tensor by a summation of multiple rank-one tensors [14]. 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-one tensors, and the rest part $\mathfrak{T} - \mathfrak{T}_d$ is mostly noise.

We use the function `cp` in the 'rTensor' R package 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 the maximum absolute value to obtain the final scGRNs for the given sample. For later use, denote the denoised adjacency matrices for the two samples as $\boldsymbol{W}_d^x$ and $\boldsymbol{W}_d^y$.

Manifold alignment

After obtaining $\boldsymbol{W}_d^x$ and $\boldsymbol{W}_d^y$, we compare them to identify the regulatory changes and associated genes and modules. Instead of directly comparing the two $n \times n$ adjacency matrices, we apply manifold alignment to build comparable low-dimensional features and compare these features of genes between two samples, while maintaining the structural information of the two scGRNs [19]. Manifold alignment is used here to match the local and no-linear structures among the data points of $\boldsymbol{W}_d^x$ and $\boldsymbol{W}_d^y$ and project them to the same low-dimensional space. Specifically, we use $\boldsymbol{W}_d^x$ and $\boldsymbol{W}_d^y$ to denote the pair-wise similarity matrices obtained by applying the PCR network construction and the denosing throught tensor decomposition on the two initial expression matrices, $\boldsymbol{X}$ and $\boldsymbol{Y}$. These similarity matrices serve as the input for manifold alignment 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 $d \ll n$. 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 the similarity matrix contains negative edge weights, properties of the corresponding Laplacian are not entirely well understood [58]. We propose two methods to deal with this problem. The first method is to directly take the absolute value of $\boldsymbol{W}_d^x$ and $\boldsymbol{W}_d^y$ as the similarity matrices, in which we regard that the highly-negative correlated genes also support a strong functional relationship. In the second method, 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 $(1,2]$. In this case, 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 data sets.

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 i^{th} gene in two samples should be relatively close in the projected space; and (2) if i^{th} gene and j^{th} 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:

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

where $W_{i,j}^{x}$ and $W_{i,j}^{y}$ 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. By 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: $\boldsymbol{F} = \begin{bmatrix} \boldsymbol{F}^x \\ \boldsymbol{F}^y \end{bmatrix}$, $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$, $\boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}^x & \frac{\lambda}{2}\boldsymbol{I} \\ \frac{\lambda}{2}\boldsymbol{I} & \boldsymbol{W}^y \end{bmatrix}$, and $\boldsymbol{D}$ is a diagonal matrix with $D_{ii} = \sum_j W_{ij}$. $\boldsymbol{L}$ is called a graph Laplacian matrix. 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 $\boldsymbol{L}$ [59].

Determination of p-value of DR genes

With $F = \begin{bmatrix} F^x \\ F^y \end{bmatrix} = [f_1, f_2, \cdots, f_d]$ 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 the Chi-square distribution to determine significant genes. Specifically, d_j^2 is derived from the summation of squares of the differences of projected representations of two samples, whose distribution could be approximately Chi-square. To adjust the scale of the distribution, we compute the scaled fold-change $df \cdot d_j^2/\overline{d^2}$ for each gene j, where $\overline{d^2}$ denotes the average of d_j^2 among all the tested genes. The scaled fold-change approximately follows the Chi-square 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 Chi-square distribution, we assign the *P*-values for genes and adjust them for multiple testing using the Benjamini-Hochberg (B-H) FDR correction [60]. 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.

Affinity propagation clustering

Module detection were conducted among genes with P < 0.05. This set of genes are focused as they are more functionally informative with detectable shifting in their regulatory patterns. These genes' coordinates in the projected low dimensional manifold (d = 30) were used to compute the pairwise similarities. The similarity matrix S is constructed as the scalar products of data vectors (linear kernel), and scaled into a [-1, 1] range by using $S = (x^T y)/(\|x\|\|y\|)$. Then, affinity propagation clustering [27] is applied over S as implemented in the 'apcluster' R package [28] using the default parameters. Affinity propagation clustering provide and optimized c number of clusters with variable sizes, that are determined automatically as result of an optimization function. Once the clusters are defined, their gene composition is split in function of the source sample. Similarity in gene composition between modules is measured using Jaccard coefficient.

Functional enrichment analyses

Functional enrichment analysis of gene sets was performed using Enrichr [32], 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 (ChEA) [33] based on

comprehensive results from ChIP-seq studies. Finally, predefined gene sets from the REACTOME, BioPlanet and KEGG databases were tested for the enriched functions using gene set enrichment analysis (GSEA) [26].

Simulations of scRNAseq data and benchmarking of network methods

To test the performance of our workflow, we generated synthetic data sets using SERGIO, a single-cell expression simulator guided by gene regulatory networks [61]. SERGIO allows for the simulation of scRNAseq 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:

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

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. In order to obtain the mRNA concentrations as a function of time, the above stochastic differential equation is integrated for all genes as follows:

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

The simulation was focused on testing and comparing the performance of PCR 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 one of the tested methods, we randomly select $n = \{10, 50, 100, 500, 1000, 2000, 3000\}$ number of cells from the simulated data for ten times and build ten scGRN. For each n, relevance measurements (accuracy and recall) were evaluated for each of the ten 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 $Recall = TP/(TP + FN)$, where T, P, F, and N stands for true, positive, false and negative, respectively. For the MI and GENIE3 methods that only provide 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.

Code availability

scTenifoldNet has been implemented in R and Matlab. 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 Fig. 2). The 'scTenifoldNet' R package is available at the CRAN repository, https://cran.r-project.org/web/packages/scTenifoldNet/.

Acknowledgments

This research was funded by Texas A&M University 2019 X-Grants and the NIH grant R21AI126219 for J.J.C.

References

1. Margolin, A.A., et al., *ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.* BMC Bioinformatics, 2006. **7 Suppl 1**: p. S7.
2. Huynh-Thu, V.A., et al., *Inferring regulatory networks from expression data using tree-based methods.* PLoS One, 2010. **5**(9).
3. Marbach, D., et al., *Wisdom of crowds for robust gene network inference.* Nat Methods, 2012. **9**(8): p. 796-804.
4. Friedman, N., et al., *Using Bayesian networks to analyze expression data.* J Comput Biol, 2000. **7**(3-4): p. 601-20.
5. Gill, R., S. Datta, and S. Datta, *A statistical framework for differential network analysis from microarray data.* BMC Bioinformatics, 2010. **11**: p. 95.
6. Todorov, H., et al., *Network Inference from Single-Cell Transcriptomic Data.* Methods Mol Biol, 2019. **1883**: p. 235-249.
7. Aibar, S., et al., *SCENIC: single-cell regulatory network inference and clustering.* Nat Methods, 2017. **14**(11): p. 1083-1086.
8. Fiers, M., et al., *Mapping gene regulatory networks from single-cell omics data.* Brief Funct Genomics, 2018. **17**(4): p. 246-254.
9. Chen, S. and J.C. Mar, *Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data.* BMC Bioinformatics, 2018. **19**(1): p. 232.
10. Pratapa, A., et al., *Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data.* bioRxiv, 2019: p. 642926.
11. Avey, D., et al., *Single-Cell RNA-Seq Uncovers a Robust Transcriptional Response to Morphine by Glia.* Cell Rep, 2018. **24**(13): p. 3619-3629 e4.
12. Ximerakis, M., et al., *Single-cell transcriptomic profiling of the aging mouse brain.* Nat Neurosci, 2019. **22**(10): p. 1696-1708.
13. Hagai, T., et al., *Gene expression variability across cells and species shapes innate immunity.* Nature, 2018. **563**(7730): p. 197-202.
14. Rabanser, S., O. Shchur, and S. Günnemann *Introduction to Tensor Decompositions and their Applications in Machine Learning*. arXiv e-prints, 2017.
15. Kendall, M.G., *A course in multivariate analysis*. Griffin's statistical monographs & courses,. 1957, New York,: Hafner Pub. Co. 185 p.
16. Baburaj, M. and S.N. George, *Reweighted Low-Rank Tensor Decomposition based on t-SVD and its Applications in Video Denoising.* arXiv, 2016. **1611**: p. 05963.
17. Yuan, L., et al., *High-dimension Tensor Completion via Gradient-based Optimization Under Tensor-train Format.* arXiv:1804.01983, 2018.
18. Battaglino, C., G. Ballard, and T.G. Kolda, *A Practical Randomized CP Tensor Decomposition.* arXiv:1701.06600, 2017.
19. Moon, K.R., et al., *Manifold learning-based methods for analyzing single-cell RNA-sequencing data.* Current Opinion in Systems Biology, 2018. **7**: p. 36-46.
20. Roscher, R., F. Schindler, and W. F, *High dimensional correspondences from low dimensional manifolds: an empirical comparison of graph-based dimensionality reduction algorithms*, in

Proceedings of the 2010 international conference on Computer vision - Volume part II. 2011, Springer-Verlag: Queenstown, New Zealand. p. 334-343.
21. Vu, H.T., C.J. Carey, and S. Mahadevan, *Manifold warping: manifold alignment over time*, in *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*. 2012, AAAI Press: Toronto, Ontario, Canada. p. 1155-1161.
22. Wang, C. and S. Mahadevan, *A General Framework for Manifold Alignment*, in *AAAI Fall Symposium: Manifold Learning and Its Applications*. 2009, Association for the Advancement of Artificial Intelligence. p. FS-09-04.
23. Nguyen, N.D., I.K. Blaby, and D. Wang, *ManiNetCluster: a novel manifold learning approach to reveal the functional links between gene networks.* BMC Genomics, 2019. **20**(Suppl 12): p. 1003.
24. Diaz, F. and D. Metzler, *Pseudo-aligned multilingual corpora*, in *Proceedings of the 20th international joint conference on Artifical intelligence*. 2007, Morgan Kaufmann Publishers Inc.: Hyderabad, India. p. 2727-2732.
25. Wang, C. and S. Mahadevan, *Manifold alignment using Procrustes analysis*, in *Proceedings of the 25th international conference on Machine learning*. 2008, ACM: Helsinki, Finland. p. 1120-1127.
26. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.
27. Frey, B.J. and D. Dueck, *Clustering by passing messages between data points.* Science, 2007. **315**(5814): p. 972-6.
28. Bodenhofer, U., A. Kothmeier, and S. Hochreiter, *APCluster: an R package for affinity propagation clustering.* Bioinformatics, 2011. **27**(17): p. 2463-4.
29. Goodsell, D.S., *The molecular perspective: morphine.* Oncologist, 2004. **9**(6): p. 717-8.
30. Tso, P.H. and Y.H. Wong, *Molecular basis of opioid dependence: role of signal regulation by G-proteins.* Clin Exp Pharmacol Physiol, 2003. **30**(5-6): p. 307-16.
31. Jalabert, M., et al., *Neuronal circuits underlying acute morphine action on dopamine neurons.* Proc Natl Acad Sci U S A, 2011. **108**(39): p. 16446-50.
32. Kuleshov, M.V., et al., *Enrichr: a comprehensive gene set enrichment analysis web server 2016 update.* Nucleic Acids Res, 2016. **44**(W1): p. W90-7.
33. Lachmann, A., et al., *ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments.* Bioinformatics, 2010. **26**(19): p. 2438-44.
34. Krezel, W., et al., *Impaired locomotion and dopamine signaling in retinoid receptor mutant mice.* Science, 1998. **279**(5352): p. 863-7.
35. Tafti, M. and N.B. Ghyselinck, *Functional implication of the vitamin A signaling pathway in the brain.* Arch Neurol, 2007. **64**(12): p. 1706-11.
36. Morikawa, H. and C.A. Paladini, *Dynamic regulation of midbrain dopamine neuron activity: intrinsic, synaptic, and plasticity mechanisms.* Neuroscience, 2011. **198**: p. 95-111.
37. Johnson, S.W. and R.A. North, *Opioids excite dopamine neurons by hyperpolarization of local interneurons.* J Neurosci, 1992. **12**(2): p. 483-8.
38. Schafer, F.Q., et al., *Comparing beta-carotene, vitamin E and nitric oxide as membrane antioxidants.* Biol Chem, 2002. **383**(3-4): p. 671-81.
39. Li, Y., et al., *Retinoic acid receptor beta stimulates hepatic induction of fibroblast growth factor 21 to promote fatty acid oxidation and control whole-body energy homeostasis in mice.* J Biol Chem, 2013. **288**(15): p. 10490-504.
40. Niewiadomska-Cimicka, A., et al., *Genome-wide Analysis of RARbeta Transcriptional Targets in Mouse Striatum Links Retinoic Acid Signaling with Huntington's Disease and Other Neurodegenerative Disorders.* Mol Neurobiol, 2017. **54**(5): p. 3859-3878.

41. Chen, J., et al., *Long noncoding RNA MALAT1 regulates generation of reactive oxygen species and the insulin responses in male mice.* Biochem Pharmacol, 2018. **152**: p. 94-103.
42. Wang, X., et al., *Long Noncoding RNAs in the Regulation of Oxidative Stress.* Oxid Med Cell Longev, 2019. **2019**: p. 1318795.
43. Zhang, X., M.H. Hamblin, and K.J. Yin, *The long noncoding RNA Malat1: Its physiological and pathophysiological functions.* RNA Biol, 2017. **14**(12): p. 1705-1714.
44. Lesnefsky, E.J. and C.L. Hoppel, *Oxidative phosphorylation and aging.* Ageing Res Rev, 2006. **5**(4): p. 402-33.
45. Wallace, D.C., *Mitochondrial genetics: a paradigm for aging and degenerative diseases?* Science, 1992. **256**(5057): p. 628-32.
46. Sharma, L.K., J. Lu, and Y. Bai, *Mitochondrial respiratory complex I: structure, function and implication in human diseases.* Curr Med Chem, 2009. **16**(10): p. 1266-77.
47. Lunnon, K., et al., *Mitochondrial genes are altered in blood early in Alzheimer's disease.* Neurobiol Aging, 2017. **53**: p. 36-47.
48. Yamada, K. and T. Nabeshima, *Changes in NMDA receptor/nitric oxide signaling pathway in the brain with aging.* Microsc Res Tech, 1998. **43**(1): p. 68-74.
49. Raveh, T., et al., *Double-stranded RNA-dependent protein kinase mediates c-Myc suppression induced by type I interferons.* J Biol Chem, 1996. **271**(41): p. 25479-84.
50. Kopitar-Jerala, N., *The Role of Interferons in Inflammation and Inflammasome Activation.* Front Immunol, 2017. **8**: p. 873.
51. Gantier, M.P. and B.R. Williams, *The response of mammalian cells to double-stranded RNA.* Cytokine Growth Factor Rev, 2007. **18**(5-6): p. 363-71.
52. Levy, H.B., L.W. Law, and A.S. Rabson, *Inhibition of tumor growth by polyinosinic-polycytidylic acid.* Proc Natl Acad Sci U S A, 1969. **62**(2): p. 357-61.
53. Saadatpour, A. and R. Albert, *Boolean modeling of biological regulatory networks: a methodology tutorial.* Methods, 2013. **62**(1): p. 3-12.
54. Kester, L. and A. van Oudenaarden, *Single-Cell Transcriptomics Meets Lineage Tracing.* Cell Stem Cell, 2018. **23**(2): p. 166-179.
55. Zheng, X., Y. Huang, and X. Zou, *scPADGRN: A preconditioned ADMM approach for reconstructing dynamic gene regulatory network using single-cell RNA sequencing data.* 2019: p. 799189.
56. Ma, X., P. Sun, and G. Qin, *Identifying Condition-Specific Modules by Clustering Multiple Networks.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2018. **15**(5): p. 1636-1648.
57. Ma, X., D. Dong, and Q. Wang, *Community Detection in Multi-Layer Networks Using Joint Nonnegative Matrix Factorization.* IEEE Transactions on Knowledge and Data Engineering, 2019. **31**(2): p. 273-286.
58. Chen, Y.X., S.Z. Khong, and T.T. Georgiou, *On the definiteness of graph Laplacians with negative weights: Geometrical and passivity-based approaches.* 2016 American Control Conference (Acc), 2016: p. 2488-2493.
59. von Luxburg, U. *A Tutorial on Spectral Clustering*. arXiv e-prints, 2007.
60. Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* J Roy Statist Soc Ser B (Methodological), 1995. **57**: p. 289-300.
61. Dibaeinia, P. and S. Sinha, *A single-cell expression simulator guided by gene regulatory networks.* bioRxiv, 2019: p. 716811.