Genomics 114 (2022) 110480

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

INSISTC: Incorporating network structure information for single-cell type classification

Hansi Zheng [a], Saidi Wang [a], Xiaoman Li [b,*], Haiyan Hu [c,*]

[a] Department of Computer Science, University of Central Florida, Orlando, FL 32816, USA
[b] Burnett School of Biomedical Science, College of Medicine, University of Central Florida, Orlando, FL 32816, USA
[c] Department of Computer Science, Genomics and Bioinformatics Cluster, University of Central Florida, Orlando, FL 32816, USA

ARTICLE INFO

Keywords:
Single-cell type classification
Gene regulatory network
Gene regulatory

ABSTRACT

Uncovering gene regulatory mechanisms in individual cells can provide insight into cell heterogeneity and function. Recent accumulated Single-Cell RNA-Seq data have made it possible to analyze gene regulation at single-cell resolution. Understanding cell-type-specific gene regulation can assist in more accurate cell type and state identification. Computational approaches utilizing such relationships are under development. Methods pioneering in integrating gene regulatory mechanism discovery with cell-type classification encounter challenges such as determine gene regulatory relationships and incorporate gene regulatory network structure. To fill this gap, we developed INSISTC, a computational method to incorporate gene regulatory network structure information for single-cell type classification. INSISTC is capable of identifying cell-type-specific gene regulatory mechanisms while performing single-cell type classification. INSISTC demonstrated its accuracy in cell type classification and its potential for providing insight into molecular mechanisms specific to individual cells. In comparison with the alternative methods, INSISTC demonstrated its complementary performance for gene regulation interpretation.

1. Introduction

Understanding gene regulatory mechanism in a cell-specific manner is a fundamental task in molecular biology. Genes are regulated at different stages, such as transcriptional and post-transcriptional gene regulation. During gene transcriptional regulation, transcription factors (TFs) and their cofactors interact with the DNA regulatory elements to regulate the gene expression levels of their target genes. Many algorithms have been developed to identify gene regulatory mechanisms through TF-target finding [15,45]. Many public resources have been available to store TF-target information [17,24,57].

Rapidly advanced Single-cell RNA sequencing (scRNA-Seq) enables genome-wide gene expression measurements in individual cells. scRNA-Seq data has numerous applications and has been utilized to study complicated biological processes at the single-cell resolution. For example, studying the transcriptional similarities and differences using scRNA-Seq data revealed cell-to-cell gene expression heterogeneity across species and tissues [7,10,20,42]. The recently accumulated scRNA-Seq-based transcriptomics data also create opportunities to understand transcriptional gene regulation at the single-cell level [9]. Computational methods have been developed to identify gene regulatory networks (GRNs), and some of them are in the context of single-cell transcriptomics [8,12,36,48,51]. Besides, unsupervised methods such as clustering have become common to discover cell types and cell states from scRNA-Seq experiments in heterogeneous tissues [20,33,55]. Many clustering algorithms have been developed for cell-type classification using scRNA-Seq data [22,25,27,32,52,58]. For example, Seurat V3 uses a graph clustering approach [6]. This method projects single cells into a graph structure. Graph partitioning algorithms are then used to identify clusters. GiniClust aims to use the Gini index to identify rare cell types from scRNA-Seq data [27]. Although these clustering algorithms have shown their capability in detecting cell types from scRNA-Seq data, they are often challenged by the lack of consistency with each other [31]. Single-Cell Consensus Clustering (SC3) attempted to conquer this challenge through consensus identification. To do that, SC3 combines multiple clustering solutions to derive a consensus matrix indicating whether two cells are in one cluster [32]. Hierarchical clustering is further applied to this matrix to obtain the final clusters.

Classification-based machine learning algorithms have also been proposed for single-cell type classification based on scRNA-Seq data

* Corresponding authors.
E-mail addresses: xiaoman@mail.ucf.edu (X. Li), haihu@cs.ucf.edu (H. Hu).

https://doi.org/10.1016/j.ygeno.2022.110480
Received 24 July 2022; Received in revised form 30 August 2022; Accepted 4 September 2022
Available online 6 September 2022

[1,3,34,50]. These classification algorithms apply various machine learning methods such as Support Vector Machine (SVM), Random Forests, and deep learning. For example, scPred first applies a singular value decomposition-based dimension reduction approach to obtain low-dimensional principle component representations for gene expression levels [3]. Specific feature selection criteria are then applied to select informative principle components for further SVM model training and prediction. Automated Cell Type Identification using Neural Networks (ACTINN) is a recent example of deep learning-based methods for single-cell type classification [34]. ACTINN trained its deep neural network model using the Tabula Muris Atlas (a mouse cell type atlas) and a human immune cell dataset. The prediction capability was demonstrated using immune-related cell types such as mouse leukocytes and human T cell subtypes. These classification-based methods usually require training samples and specific feature selection protocols. Most of these methods utilize only the scRNA-Seq measurements of individual genes' expression levels without considering the underlying cellular mechanisms.

Another computational method called Single-Cell rEgulatory Network Inference and Clustering (SCENIC) takes a different approach for cell-type classification. SCENIC aims to infer single-cell-resolution gene regulatory information from scRNA-Seq data and then use this information for cell-type classification [48]. SCENIC uses GEne Network Inference with Ensemble of trees (GENIE3) to infer transcription factor (TF) target co-expression relationships and uses the motif finding algorithm named RcisTarget to determine direct TF target genes. A TF and its identified targets together are defined as a regulon. SCENIC then uses the AUCell algorithm to score regulon activities based on the gene expression measurements in individual cells. Using gene regulatory information to classify cell types is beneficial in two aspects. One is that integrating regulatory information is likely to help the cell type and state discovery. This is because the sensitivity of scRNA-Seq technology can result in transcriptional noise [39], and low-expression genes are difficult to detect, causing dropouts in the data [38], which may be alleviated by integrating expression data with regulatory information. The other is that cell types inferred from the underlying regulatory states can provide insight into the cell-type-specificity of gene regulatory mechanisms. However, gene regulatory relationships forming GRNs are complex such that one TF can have many targets, and multiple TFs can collaboratively regulate the same target genes. The network structure properties in a GRN have not been taken into account for single-cell data analysis.

We developed a method called Incorporate Network Structure Information for Single cell Type Classification (INSISTC) to utilized biological network information to facilitate cell type classification and interpretation. INSISTC utilizes the Systematic Identification Of Motifs In ChIP-Seq data (SIOMICS) approach to generate a GRN with its TF-target relationships identified through de novo DNA regulatory motif discovery [13,14]. SIOMICS is capable of considering both TFs and their cofactors for motif prediction and has demonstrated good performance. Besides, to take the structural properties of the GRN, INSISTC adopts a random-walk-based graph algorithm to represent the GRN structural information. INSISTC incorporates genes and GRN structural information by creating a Latent Dirichlet Allocation (LDA)-based topic model. The model generates cell-type-specific topics used for cell-type classification and regulatory mechanism discovery. We compared our method to SCENIC and alternative topic model construction of INSISTC. We showed that INSISTC can accurately perform cell-type classification for single cells. We also demonstrated that INSISTC could uncover cell-specific gene regulatory mechanisms.

2. Material and methods

2.1. Overview of INSISTC

INSISTC is a topic-model-based computational framework developed to detect single cell types from scRNA-Seq measurements while providing insight into cell-type-specific gene regulatory mechanisms. For a given scRNA-Seq dataset, INSISTC consists of four steps (Fig. 1). First, INSISTC provides data pre-processing and filtering. Second, based on GRN generated by SIOMICS, INSISTC executes a random walk-based graph algorithm to generate word representation for all scRNA-Seq samples. Third, INSISTC applies the LDA topic model to generate a topic representation for each scRNA-Seq sample. Fourth, INSISTC performs single-cell clustering and visualization, where the SC3 and UMAP algorithms can be applied. In the following, we describe each of these four steps in more detail.

2.2. Data collection and pre-processing

To evaluate the performance of INSISTC, we obtain three scRNA-Seq datasets: the mouse cerebral cortex data (GSE60361) [4], the mouse skeletal muscle data (GSE143437: GSM4259473, GSM4259476, GSM4259478, GSM4259481) [11], and the mouse embryo data (Array express E-MTAB-3321) [19]. The corresponding cell numbers are 3005, 14,242 and 124. The mouse cortex data is annotated to seven cell types, including Interneuron, pyramidal SS, pyramidal CA1, oligodendrocytes, endothelial, astrocytes ependymal, and microglia cells. The muscle data are annotated with the following 12 cell types: Mature skeletal muscle, B/T/NK cells, MuSCs and progenitors, Monocytes/Macrophages/Platelets, Endothelial, Fibro-adipogenic progenitors (FAPs), Anti-inflammatory macrophages, Resident Macrophages/APCs, Pro-inflammatory macrophages, Neural/Glial/Schwann cells, Tenocytes, and Smooth muscle cells. The mouse embryo data are annotated with 5 cell stages including 2-cell stage, 4-cell stage, 8-cell stage, 16-cell stage, and 32-cell stage cells.

To filter the most likely unreliable genes that may provide only noise, we apply two layers of filters on the expression matrix. The first filter is based on the total number of sequencing reads per gene. If a gene does not fit the following requirement, we remove the column of this gene from the expression matrix. The thresholds of this filter are based on each scRNA-Seq dataset, calculated by Eq. (1). We use the expression threshold 3 in our experiments, but other values can be explored based on the mean or median of non-zero expression level. If a gene has a total number of reads less than this threshold 3, it will be removed from the dataset. To further remove genes only expressed in one or very few cells, we apply the second filter such that genes detected in at least 1 % of the total cells are kept.

$$\text{threshold}_{\text{filter1}} = expression\_threshold \times 1\% \text{of cells} \quad (1)$$

2.3. Computational identification of TF-target relationships

INSISTC leverages the relationship between the TFs and their target genes, constructs the gene relation graph, and traverses along the edges to infer the gene-gene connection. INSISTC uses SIOMICS v3 to obtain gene regulatory relationships [13,14]. SIOMICS is a computational tool for de novo discovery of motifs and TF binding sites in a set of DNA sequences such as those from all peak regions of a ChIP-seq experiment. SIOMICS simultaneously considers motifs of a TF and those of its co-factors to discover motifs, which enables it to discover combinations of any number of co-occurring motifs and significantly reduce false-positive predictions compared with tools considering individual motifs separately. We call the significant motif combinations output from SIOMICS motif modules, which describes the binding pattern of a group of TFs and cofactors that co-regulate their target genes under the corresponding experimental conditions. We construct the GRN with the motif modules predicted by SIOMICS. The TF-target gene pairs for each motif were obtained by comparing the predicted motifs in motif modules with the known motifs in the JASPAR2020 vertebrate database [17] using the tool STAMP [35] with an *E*-value cutoff of 1E-5. In this way, we obtain 430 unique TFs and 20,006 unique targets, corresponding to a

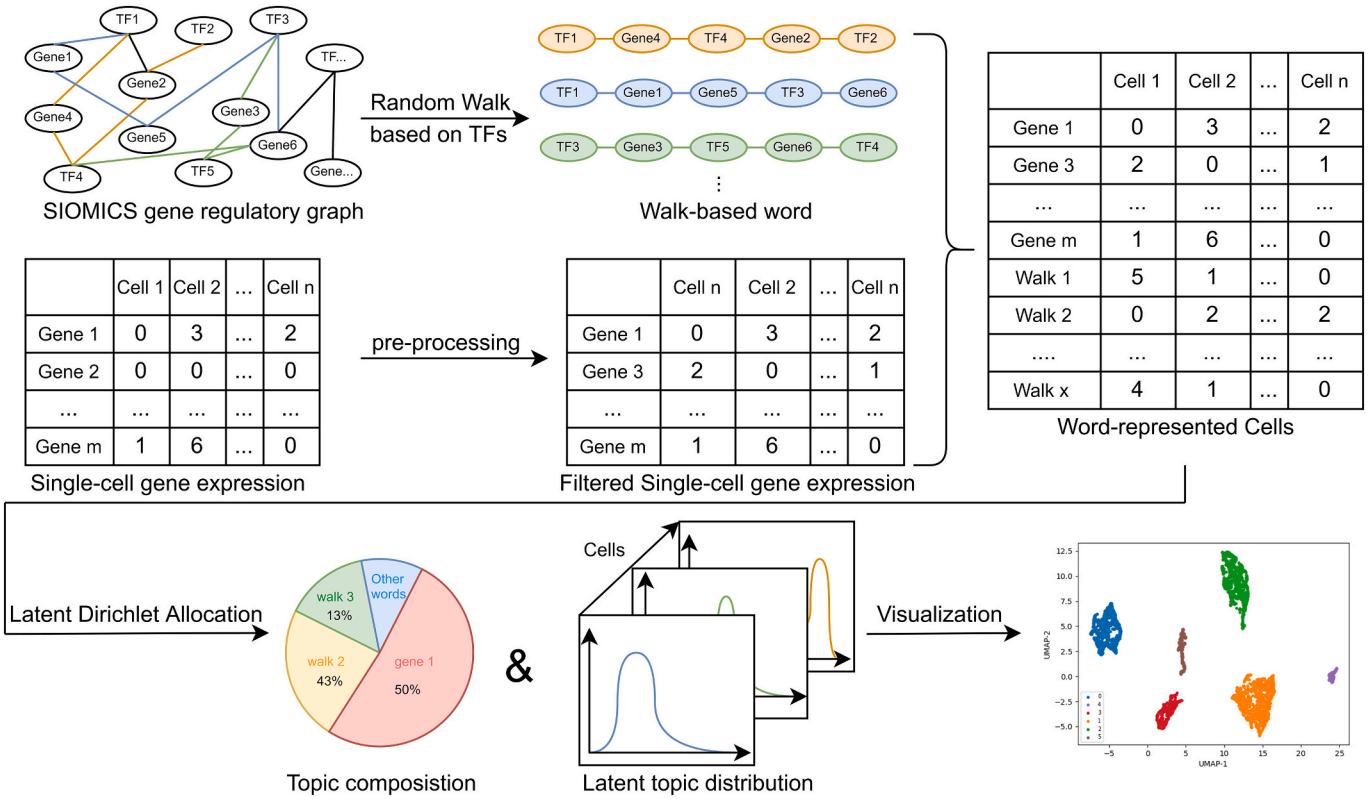

Fig. 1. The pipeline of INSISTC includes four steps: data pre-processing and filtering; a graph algorithm to generate word representation for all scRNA-Seq samples; a LDA topic model to generate a topic representation for each scRNA-Seq sample; and single-cell clustering and visualization.

GRN where TFs and genes are identified as nodes, and TFs and their targets are connected by edges.

2.4. LDA topic model of scRNA-Seq data

INSISTC uses the LDA topic model to model a scRNA-Seq dataset. LDA is a generative probabilistic model commonly used for topic modeling [5]. LDA is motivated by the need to model a collection of discrete data. When applied to text corpora, LDA represents a document as a collection of words, and the whole word collection is defined as word vocabulary. A document can then be modeled as a finite mixture over an underlying set of topics, and a topic can be modeled as a finite mixture over an underlying set of words.

To apply the LDA to model the collections of individual cells in a scRNA-Seq dataset, we need to define the corresponding words and documents. Intuitively, we can consider the single-cell sample as a document with each gene as a word. However, defining words based on genes alone does not consider gene regulatory relationships. To account for the gene regulatory relationship, we can define each TF-target pair as a word. Nevertheless, this definition ignores the interactions between different TFs and their target genes, i.e., the structural properties of a GRN.

To incorporate the structural properties of GRNs properly into the word definition of the LDA model, INSISTC uses an anchor-based random walk with a forest fire mechanism [23,40]. Briefly, each TF serves as an anchor for the beginning of a random walk, and each anchor is subject to a maximum of five walks. For every step of the random walk, the edge that connects the nodes of the current step with the nodes of the following step is removed to avoid a redundant walk path. The forest fire method provides for a more thorough traversal than a standard random walk, as well as a more accurate representation of the graph structure and the retention of only the unique random walk result. Each obtained random walk path is then defined as a word, named as a walk-based word. The collection of all the genes and walk-based words is designated as the INSISTC vocabulary.

To further describe a single cell sample as a document with the above word definitions, we need to specify the occurrence of a specific word. INSISTC measures the occurrence of a gene based on its expression level and defines the occurrence of a walk-based word using the AUCell scoring schema [2]. Briefly, for all the genes in a walk-based word, AUCell uses the "Area Under the Curve" (AUC) to calculate whether a critical subset of the input gene set is enriched within the expressed genes for a given single cell sample. The AUCell scores are further scaled by a constant coefficient of 100 to represent the word occurrences.

2.5. Comparison with SCENIC and alternative approaches

To evaluate INSISTC performance, we compare INSISTC with a popular method, SCENIC. We also compare INSISTC results based on alternative word and vocabulary definitions. We introduce alternative definitions including "gene-only", "walk-only", and "TF-target-based" vocabulary. The "gene-only" and "walk-only" are straightforward, meaning the vocabulary only contains genes and walk-based words. To define TF-target-based vocabulary, we first filter the TF-target gene pairs based on both genes' expression levels; TF-target gene pairs are considered words if and only if both genes' expression levels in the scRNA-Seq expression matrix are non-zero. For TF-target-based vocabulary, the word occurrence is the average expression between TF and target genes. The occurrence of a word for other definitions is the same as described in the above section.

To evaluate the cell type classification accuracy between any two given methods, we compare the results from different approaches with the cell type annotation from the reference publications using the adjusted rand index (ARI) [46]. The Rand Index (RI) can measure the similarity of two clustering results by considering the different ways of their assignments of objects to clusters. The ARI is the corrected-for-

chance version of the RI. The ARI score is close to 0 if the clustering results are in a random agreement and close to 1 when the clustering results are nearly identical. Therefore, the ARI scores based on INSISTC resulted clusters and annotated cell type clusters is able to show the how consistent INSISTC results are with the cell type annotations. The ARI is calculated based on the following equation,

$$\mathrm{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{b_j}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{n}{2}} \quad (2)$$

where n_{ij}denotes the number of objects in common between two clusters, $a_i$and $b_j$denote the sum of elements for each model.

3. Results

3.1. INSISTC reliably classifies different cell types in comparison with alternative methods

To evaluate INSISTC in terms of cell type classification accuracy, we run INSISTC on three datasets with previously annotated cell types, including mouse cortex, mouse skeletal muscle, and mouse embryo datasets (See "Materials and methods" section). The vocabulary of the topic model involved in INSISTC was defined as the union of both genes and walks (gene-walk-based vocabulary). For example, 13,063 gene-based words and 1982 walk-based words constitute the 15,045-word vocabulary for the mouse cortex data. Similarly, for the mouse skeletal muscle data, there are 13,701 genes-based words and 2035 walk-based words leading to 15,736 word vocabulary. The INSISTC model output major topics covered by the input single-cell samples for a specified vocabulary and topic number. Each topic contains a mixture of words that are either genes or walks. INSISTC represents each single cell sample as a mixture of the topics. Take the mouse cortex data as one example. We found 1223 walk-based words and 1560 gene-based words with the mixture proportion cutoff $p > 0.0005$. We observed that 2886 out of 3005 cells have at least one topic with $p > 25\%$, 1238 cells have at least one topic with $p > 50\%$, and 196 cells have at least one topic with $p > 75\%$. Of the 45 topics, 40 have at least one cell with $p > 25\%$, 34 have at least one cell with $p > 50\%$, and 21 have at least one cell with $p > 75\%$. A clustering algorithm was then applied to the topics-represented single-cell samples to obtain cell type classification.

We performed SC3 clustering on topic-represented single cells to understand INSISTC results in terms of cell-type classification. SC3 is a supervised clustering tool that utilizes a consensus strategy to combine multiple clustering solutions for single-cell samples. Specification of the number of clusters is not required. To further evaluate the cell classification accuracy, we defined true positives as pairs of cells with the same annotated cell type and fall into the same SC3 cluster. True negatives are cell pairs with different cell type annotations and fall into different SC3 clusters. Similarly, false negatives are cell pairs with the same cell type annotations but fall into different SC3 clusters. False positives have different cell type annotations but fall into the same cluster.

We run INSISTC under various settings of topic numbers ranging from 15 to 60 and found incorporating walk-based words in INSISTC topic discovery, in general, provides sufficient cell type classification accuracy. For the mouse cortex, skeletal muscle, and embryo samples, the best ARI achieved based on INSISTC results is 0.83, 0.67, and 0.77, respectively. In contrast, the best ARI achieved running SC3 directly on the original scRNA-Seq samples is 0.49, 0.31 and 0.58, respectively. The performance of INSISTC, in terms of additional metrics, is in general superior to clustering-based cel type classification based on all three datasets. For example, the average sensitivity, specificity and F1 scores for the mouse cortex data are 0.73, 0.91 and 0.65. In contrast, the SC3 clustering based on the original scRNA-Seq data has the corresponding sensitivity, specificity and F1 scores as 0.28, 0.99 and 0.43 (Table 1 & supplementary Table S1). Therefore, applying the walk-incorporated topic model has a measurable impact on the ability to cluster cells belonging to the same cell types.

We also compared INSISTC results with SCENIC in terms of cell-type classification. SCENIC has recently demonstrated its capability to successfully uncover gene regulatory information and classify cell states from scRNA-Seq data. Both INSISTC and SCENIC can identify gene regulatory mechanism from single-cell transcriptomic data. INSISTC differs from SCENIC in two major aspects. One is that INSISTC focuses on network-structure incorporation using graph algorithms. The other is that in terms of regulatory motif finding, INSISTC uses SIOMICS to identify regulatory motifs that take into account binding cofactors, while SCENIC utilizes computationally defined TF-targeting relationships called regulons. We ran SCENIC (version 0.11.2) using the provided TFs and cis-regulatory database from the pySCENIC tutorial [2,48]. We obtained 422 regulons corresponding to 422 TFs and 11,897 genes. We then performed SC3 clustering based on regulon activities inferred by SCENIC. SC3 predicted 14 clusters, based on which SCENIC corresponds to an ARI value of 0.67 comparing with the seven cell type annotations. SCENIC clustering has its overall sensitivity, specificity and F1 scores as 0.44, 0.98 and 0.59, respectively. Therefore, INSISTC has better sensitivity and F1 scores while having a slight disadvantage in specificity compared to SCENIC.

To understand how different GRN inference methods impact the results, we compared INSISTC on the GRNs inferred by PIDC, GENIE3 and GRNBoost2 [8,26,37]. PIDC utilizes partial information decomposition to infer gene regulatory relationships efficiently. GENIE3 infers the relationship between a gene pair based on the feature importance of one gene for predicting the other gene's expression. GRNBoost2 has a similar rationale as GENIE3 but improves the efficiency by adopting stochastic Gradience Boosting Machine regression. These three GRNs have been ranked top and consistent performers according to a recent single-cell transcriptomic data-based GRN benchmark [41]. We run each of these three methods for each single cell dataset to generate a GRN. We then applied INSISTC and SC3 clustering on the GRNs under the same topic settings previously described.

We found that INSISTC's performance on the three GRNs is consistent with that on the SIOMICS-generated GRN. As stated previously, the best ARI achieved running SC3 directly on the original scRNA-Seq samples is 0.49, 0.31 and 0.58, respectively. The best ARIs achieved based on INSISTC results on the three GRNs indicate more accurate cell type classification (Supplementary Table S2). Briefly, for PIDC-based GRN, for the mouse cortex, skeletal muscle, and embryo samples, the best ARI achieved based on INSISTC results is 0.91, 0.55, and 0.79, respectively. As to GENIE3-based GRN, corresponding to the mouse cortex, skeletal muscle, and embryo samples, the best ARI achieved based on INSISTC results is 0.91, 0.56, and 0.65, respectively. Also, for GRNBoost-based GRN, for the mouse cortex, skeletal muscle, and embryo samples, the best ARI achieved based on INSISTC results is 0.97, 0.58, and 0.55, respectively. In terms of additional metrics, INSISTC on the three alternative GRNs also shows advantages compared to clustering-based cell type classification (Supplementary Table S2). For example, for the mouse cortex data, the SC3 clustering based on the original scRNA-Seq data has the corresponding sensitivity, specificity

Table 1
The performance of INSISTC on mouse cortex data.

Topic num	Sensitivity	Specificity	F1-score	Cluster num	ARI
15	0.90452	0.64029	0.55778	4	0.37418
20	0.83717	0.93290	0.80231	6	0.74606
25	0.88362	0.93989	0.83861	6	0.79217
30	0.77198	0.95326	0.79342	7	0.73952
35	0.72438	0.94701	0.75393	7	0.69080
40	0.58524	0.96586	0.68357	9	0.79773
45	0.57523	0.96536	0.67508	10	0.81575
50	0.53188	0.97448	0.65387	11	0.83474
SCENIC	0.27795	0.99441	0.42803	7	0.49000

and F1 scores as 0.28, 0.99 and 0.43. In contrast, for PIDC-based GRN, the average sensitivity, specificity and F1 scores for the mouse cortex data are 0.74, 0.92 and 0.73. As to GENIE3-based GRN, the average sensitivity, specificity and F1 scores are 0.95, 0.67 and 0.57. Also, for GRNBoost-based GRN, the average sensitivity, specificity and F1 scores are 0.73, 0.92 and 0.73. These results suggest walk-incorporated topic model's effect on cell type classification is robust to typical GRN inference methods.

3.2. Network structure incorporation enhances the accuracy of cell type classifications

INSISTC defines the vocabulary of its topic model as the collection of genes and walks. To investigate how alternative vocabulary definition affects cell topic discovery and cell-type classification, we specified three alternative definitions to compare INSISTC results: "gene-only", "walk-only", and "TF-target-based". Briefly, "gene-only" means that only genes are considered as words for the topic model in INSISTC, and "walk-only" means that only walk-based words are considered. TF-target-based means the TF and one of its target genes form a word to define the vocabulary.

INSISTC was run under eight topic number settings ranging from 15 to 60 for the three scRNA-Seq datasets. SC3 clustering was performed on the INSISTC topic-represented single cells. ARI was used to evaluate the cell clustering consistency with the cell-type annotation in the reference paper. The ARIs corresponding to alternative and gene-walk-based vocabulary were then compared. We found the gene-walk-based vocabulary for INSISTC, in general, resulted in more accurate cell type classification than alternative vocabulary-based INSISTC versions did (Fig. 2).

The averaged ARIs for the cortex dataset over eight topic settings are 0.72, 0.54, 0.59, and 0.54 for gene-walk, gene-only, walk-only, and TF-target-based versions. For the mouse skeletal muscle dataset, the averaged ARI over the eight topic settings are 0.55, 0.48, 0.39, and 0.09 for the same four versions. The same scenario is for the embryo dataset. The corresponding averaged ARIs are 0.57, 0.31, 0.30, and 0.29. This result shows that the network structure incorporation in the clustering procedure generally enhances the accuracy of cell type classifications.

3.3. INSISTC reveals marker topics contributing to cell type classification

To investigate the capability of INSISTC in interpreting the single cell type classification, we studied cell-type-specific topics (CSTs) that significantly contribute to the cell type classification. We identified CSTs based on their potential to distinguish a cell cluster from others. Using the SC3 package, we obtained CSTs as marker topics with p-values smaller than 0.01. A p-value was calculated based on Wilcoxon signed-rank test. For comparison, we also performed SC3 clustering on SCENIC results based on regulon activity scores.

For the mouse cortex dataset, we obtained 14 CSTs out of 45 topics (Fig. 3). Among these topics, topic 32 can distinguish oligodendrocytes from other cell types. We found that 603 out of the 820 oligodendrocytes (74%) have topic 32 as their most enriched topic. The average enrichment proportion of the topic 32 in all the oligodendrocytes is 0.56. We performed GO analysis with Gorilla [16] using 100 top-contributor words in topic 32 and found the enrichment of "response to interleukin-1" (GO:0070555, corrected p-value: 7.39E-2) and "regulation of gliogenesis" (GO:0014013, corrected p-value: 0.15). Gliogeneisis is directly relevant to oligodendroctye generation, while Interleukin-1 has been found to regulate the proliferation and differentiation of oligodendrocytes [49]. In contrast, besides the zinc finger and BTB domain containing 33 gene (*Zbtb33*), SCENIC-based SC3 clustering results do not show other regulons that overlap with genes or walks in the topic 32 (Supplementary Fig. S1). *Zbtb33* is involved in oligodendroglial maturation [56].

We also inspected multiple CSTs identified from the same cluster but correspond to multiple cell types. For example, topics 8, 34 and 43 were all selected as CSTs that can distinguish astrocytes and the endothelial cell type from others. Although the majority of astrocytes and endothelial cells are clustered together, topic 8 can tell astrocytes apart from others, while topic 43 is significantly enriched in endothelial cells. In both topics, Malat1 is the most enriched gene. However, a close investigation of topic 43 shows a number of top walks connected by the Kruppel-like factor 6 (*Klf6*) gene. This TF was reported to regulate target genes in endothelial injury recovery [18]. It is interesting to see that, although topics 22, 27, 39 and 44 were all CSTs corresponding to pyramidal CA1 cell types, they actually fell into three different SC3 clusters indicating potential subtypes.

Fig. 2. The ARIs corresponding to SCENIC and four different vocabulary-based INSISTC results on mouse cortex, skeletal muscle and embryo datasets, respectively. For a given topic number, the bars from top to bottom are in the order of SCENIC, TF-target, Gene-walk, Gene-only, and Walk-only. Here, "Gene-walk" is when both genes and walk-based words are considered as words for the topic model in INSISTC. "walk-only" is when only walk-based words are considered. "gene-only" is when only genes are considered. TF-target-based means the TF and one of its target genes form a word to define the vocabulary.

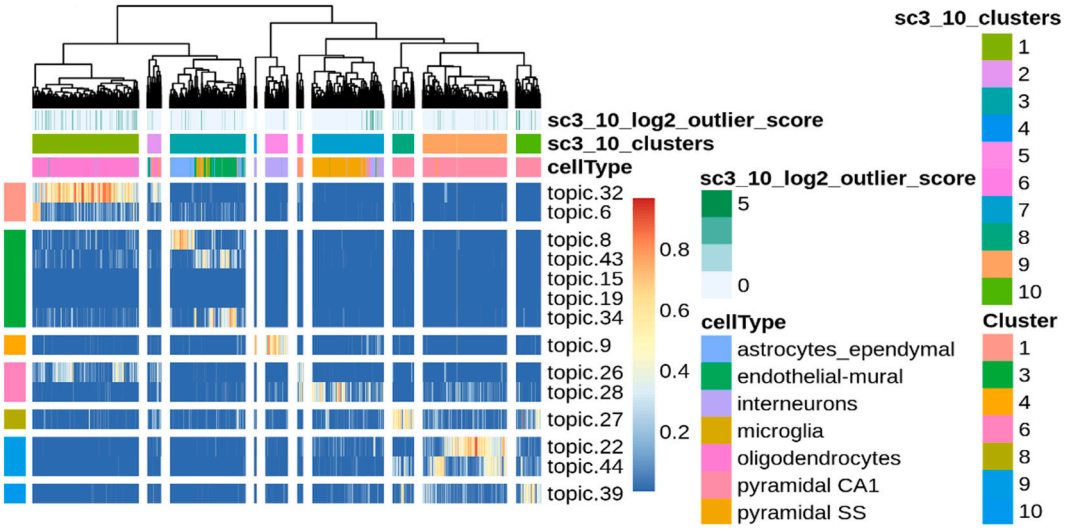

Fig. 3. The CSTs identified in mouse cortex, illustrated using SC3 package. The SC3 cell outlier scores indicate how well a cell fit into its cluster, calculated based on the minimum covariance determinant. Cells that fit well into their clusters receive an outlier score of 0, whereas high values indicate that the cell should be considered an outlier. The scale indicates the topics distribution for the cells.

For the mouse skeletal muscle data, INSISTC identified 2035 walk-based words and 13,071 gene-based words. Under the setting of gene-walk-based vocabulary and 55 topics, INSISTC resulted in 14 clusters corresponding to 15 cell types. We identified CSTs specific for monocytes/macrophage, endothelial, FAPs, anti-inflammatory macrophages and resident macrophages/APCs. Almost all the CSTs are supported by the GO annotation of the top-contributor words. For example, topic 14 is the CST for the anti-inflammatory macrophages. Significant GO annotation terms enriched in topic 14 include "antigen processing and presentation of peptide antigen" (GO:0048002, corrected *p*-value: 2.09E-7), "immune response" (GO:0006955, corrected p-value: 1.18E-3), "defense response" (GO:0006952, corrected p-value: 0.238), and others. Similarly, topic 1 is the CST for the FAPs. The most significant GO terms include "regulation of angiogenesis" (GO:0045765, corrected p-value: 0.31), "animal organ development" (GO: 0048513, corrected p-value: 0.183), and "positive regulation of vasculature development" (GO: 1904018, corrected p-value: 0.339).

3.4. INSISTC reveals cell-type-specific regulatory mechanisms

We also explored the walks in the top 100 words of CSTs to identify cell-specific regulatory mechanisms. The walk-based words ranked top according to their mixture proportions in a CST are named its top-contributor walks. We found that top-contributor walks provide insights into the regulatory mechanisms of specific cell types. Take the topic 32 identified from the cortex data for example, the GLI Family Zinc Finger 3 (*Gli3*) induced a number of top-contributor walks, including *Rai1*, *Fev*, *Tspan2* genes. It has been shown *Gli3* is important for developing mature oligodendrocytes [47]. Similarly, in topic 9, which was found to be a major marker topic for interneuron cells, we found *Tcf4* involved in a few top-contributor walks that form a small network connecting *Mtfr1*, *Dnajc4*, *Irf2*, *Foxo1*, *Ndufa4*, *Prox1*, *Chd1l*, *Pmaip1*, *Thy1* and others (Fig. 4a). These relevant walks include TFs such as *Zscan4*, *Hmx2*, *Irf2*, *Foxo1*, and *Znf16*. Studies have demonstrated that *Tcf4* plays an important role in the interneuron function and has been shown in interneuron dysfunction associated disorders [29]. For the mouse skeletal muscle data, we found the CST 39 for endothelial cells. The TF Kruppel-like factor 4 (*Klf4*) engaged walks were observed in the top words. *Klf4* connects *Tead3*, *E2f2*, *Fgf7*, *Irf1*, *Hic1*, *Egr2*, *Kif26* and other genes. Several of them, such as *E2f2* and *Irf1*, have well-studied roles in endothelial cell growth and angiogenesis [28,53]. Meanwhile, *Klf4* plays an important role in endothelial transcriptome regulation and greatly impacts endothelial functions [43]. In addition, *Egr1* centered regulatory network was also revealed by the top-contributor walks involving multiple TFs such as *Tgif2lx*, *Hoxa13*, *Hnf4a* and *Znf460* (Fig. 4b). Most of these TFs participate in endothelial proliferation and angiogenesis [44,53]. *Egr1* itself is essential to endothelial gene expression [30]. Similarly, the CST 14 is a marker topic for anti-inflammatory macrophages. The *Irf7* and *Spi1* are connected through a subnetwork that emerged from the top-container walks. *Irf7* and *Spi1* both play key roles in macrophage phenotype formulation and function [21,54].

4. Conclusion and discussion

The availability of a large amount of scRNA-Seq data enables the study of gene regulatory mechanisms at single-cell resolution. Meanwhile, the discovery of underlying gene regulatory mechanisms can benefit more accurate cell type and state discovery from scRNA-Seq data. Methods have emerged recently to integrate gene regulatory mechanism discovery with cell-type classification. However, such method development is still at its beginning stage, and there is still space for improvement in terms of GRN construction and strategies for utilizing such GRN information. The INSISTC method was developed to overcome current challenges. INSISTC takes advantage of a de novo motif analysis that considers both TFs and their cofactors. Most importantly, INSISTC considers the graph structure of the GRN and uses a graph algorithm to incorporate this network structure. INSISTC further applies a topic model to identify particular cell-enriched topics involving cell-relevant genes and regulatory mechanisms. Such topics can be further examined for cell-specific gene regulatory mechanisms and also can be grouped, e.g., by SC3, for cell-type classification. INSISTC demonstrated sufficient cell type classification accuracy and cell-type-specific gene regulatory mechanism discovery. Compared with the recent method SCENIC, INSISTC demonstrated its complementary performance for gene regulation interpretation.

At the final stage, INSISTC runs a clustering algorithm on the identified topics to identify cell types and states. We illustrated INSISTC here using SC3 algorithm because the SC3 algorithm offers cluster number estimation, while most clustering algorithms do not have such a function. However, any clustering algorithms can be plugged into the pipeline to derive final single-cell clusters. Besides, although we used three mouse datasets and mouse GRN to illustrate the usage of INSISTC here, users can apply INSISTC to scRNA-Seq data of other species and other

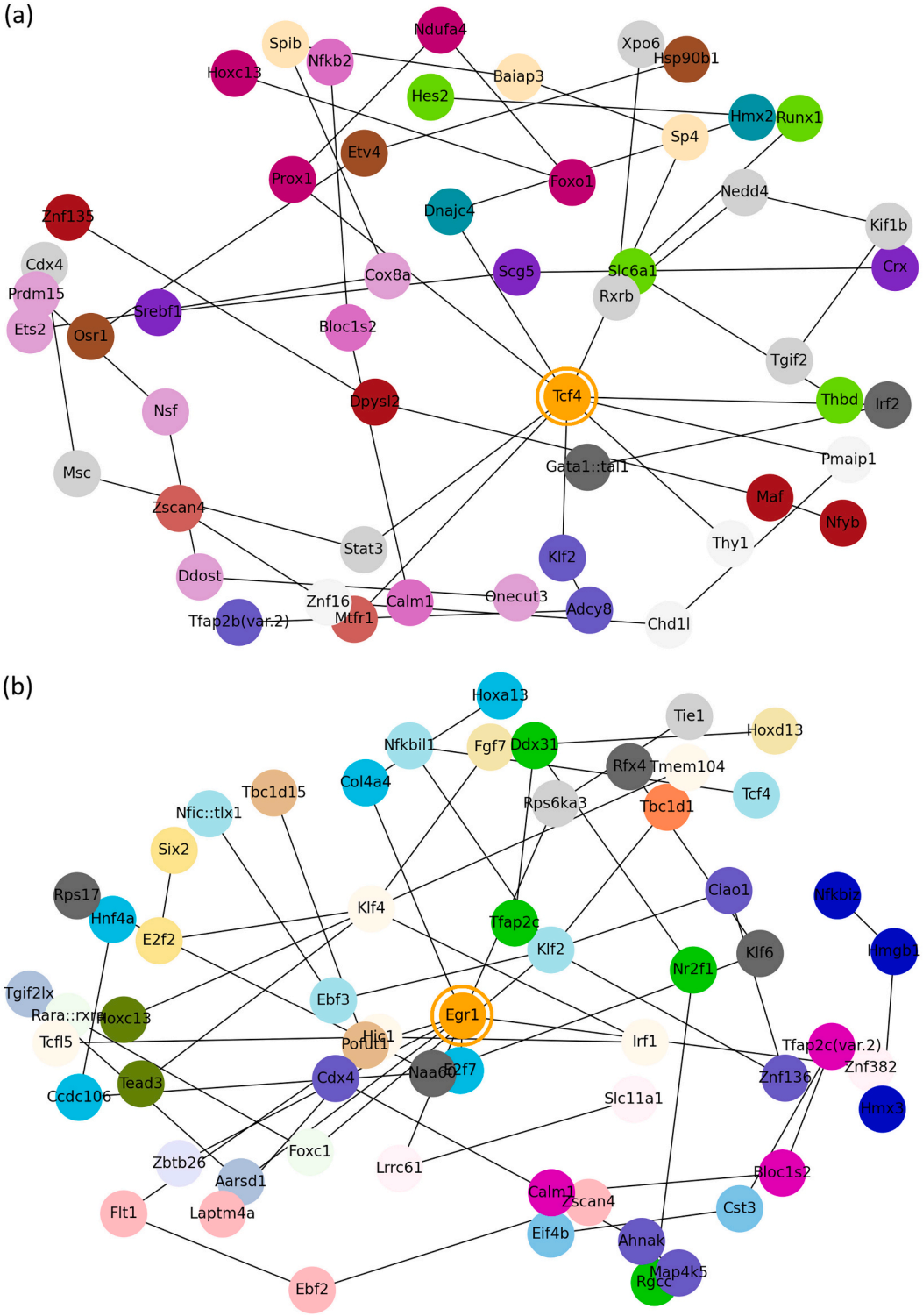

Fig. 4. Illustration of top-contributor walks in CSTs. (a) The Tcf4 network from topic 9 in cortex. (b) the Egr1 network from topic 39 in muscle. The network was built and plotted with the Python packages networkx (https://networkx.org/) and matplotlib (https://matplotlib.org/), respectively.

available biological networks of interest. It is also worth noting that, the three datasets we discussed here contain different number of cell numbers, ranging from hundreds to thousands. The number of cells can have various impacts on the cell type classification results. Low cell number datasets such as the mouse embryo datasets might not have the same discrimination power for cell classification as the other two datasets we used. In addition, for the topic model that is the essential part of INSISTC, the users need to specify a topic number. There are multiple ways to determine topic numbers. For example, the topic coherence and perplexity metrics are often applied in the context of language modeling. However, it is common to run a set of topic numbers to observe biological interpretability.

INSISTC runs SIOMICS to generate the TF-target relationship because SIOMICS considers both TFs and their cofactors in de novo motif discovery. However, with more TF-target information such as ChIP-Seq data available, the performance of INSISTC in terms of gene regulatory

mechanism discovery can be further improved. Finally, although the usage of INSISTC was illustrated on GRNs, INSISTC is flexible to incorporate other types of biological networks such as pathways, protein interaction networks and gene co-expression networks. It is also possible for INSISTC to identify cell-type-specific mechanisms from a properly integrated network.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2022.110480.

Author statement

H.H. and X.L. conceived the idea and designed the study. H.Z. and S. W. implemented the idea and generated results. H. Z., S. W., X.L. and H. H. analyzed the results and wrote the manuscript. All authors reviewed the manuscript.

Funding

This work was supported by the United States National Science Foundation [1661414, 2015838, 2120907].

Availability

Source code and manual of INSISTC are available at https://hulab.ucf.edu/research/projects/INSISTC/

Declaration of Competing Interest

The authors declare that they have no competing interests.

Data availability

Data will be made available on request.

References

[1] T. Abdelaal, et al., A comparison of automatic cell identification methods for single-cell RNA sequencing data, Genome Biol. 20 (1) (2019) 194.
[2] S. Aibar, et al., SCENIC: single-cell regulatory network inference and clustering, Nat. Methods 14 (11) (2017) 1083–1086.
[3] J. Alquicira-Hernandez, et al., scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data, Genome Biol. 20 (1) (2019) 264.
[4] F.H. Biase, X. Cao, S. Zhong, Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing, Genome Res. 24 (11) (2014) 1787–1796.
[5] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (null) (2003) 993–1022.
[6] A. Butler, et al., Integrating single-cell transcriptomic data across different conditions, technologies, and species, Nat. Biotechnol. 36 (5) (2018) 411–420.
[7] J. Cao, et al., Comprehensive single-cell transcriptional profiling of a multicellular organism, Science (New York, N.Y.) 357 (6352) (2017) 661–667.
[8] T.E. Chan, M.P.H. Stumpf, A.C. Babtie, Gene regulatory network inference from single-cell data using multivariate information measures, Cell Syst. 5 (3) (2017) 251–267 e253.
[9] G. Chen, B. Ning, T. Shi, Single-cell RNA-Seq technologies and related computational data analysis, Front. Genet. 10 (2019) 317.
[10] G. Chen, et al., Single-cell analyses of X chromosome inactivation dynamics and pluripotency during differentiation, Genome Res. 26 (10) (2016) 1342–1354.
[11] A.J. De Micheli, et al., A reference single-cell transcriptomic atlas of human skeletal muscle tissue reveals bifurcated muscle stem cell populations, Skelet. Muscle 10 (1) (2020) 19.
[12] J. Ding, et al., ChIPModule: systematic discovery of transcription factors and their cofactors from ChIP-seq data, Pac. Symp. Biocomput. (2013) 320–331.
[13] J. Ding, et al., Systematic discovery of cofactor motifs from ChIP-seq data by SIOMICS, Methods 79-80 (2015) 47–51.
[14] J. Ding, H. Hu, X. Li, SIOMICS: a novel approach for systematic identification of motifs in ChIP-seq data, Nucleic Acids Res. 42 (5) (2014), e35.
[15] Z. Duren, et al., Modeling gene regulation from paired expression and chromatin accessibility data, Proc. Natl. Acad. Sci. U. S. A. 114 (25) (2017) E4914–E4923.
[16] E. Eden, et al., GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists, BMC Bioinform. 10 (2009) 48.
[17] O. Fornes, et al., JASPAR 2020: update of the open-access database of transcription factor binding profiles, Nucleic Acids Res. 48 (D1) (2020) D87–D92.
[18] E. Gallardo-Vara, et al., Transcription factor KLF6 upregulates expression of metalloprotease MMP14 and subsequent release of soluble endoglin during vascular injury, Angiogenesis 19 (2) (2016) 155–171.
[19] M. Goolam, et al., Heterogeneity in Oct4 and Sox2 targets biases cell fate in 4-cell mouse embryos, Cell 165 (1) (2016) 61–74.
[20] D. Grun, et al., Single-cell messenger RNA sequencing reveals rare intestinal cell types, Nature 525 (7568) (2015) 251–255.
[21] R. Gunthner, H.J. Anders, Interferon-regulatory factors determine macrophage phenotype polarization, Mediat. Inflamm. 2013 (2013), 731023.
[22] M. Guo, et al., SINCERA: a pipeline for single-cell RNA-Seq profiling analysis, PLoS Comput. Biol. 11 (11) (2015), e1004575.
[23] W.L. Hamilton, R. Ying, J. Leskovec, Inductive representation learning on large graphs, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, Curran Associates Inc., Long Beach, California, USA, 2017, pp. 1025–1035.
[24] H. Han, et al., TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions, Nucleic Acids Res. 46 (D1) (2018) D380–D386.
[25] Y. Hao, et al., Integrated analysis of multimodal single-cell data, Cell 184 (13) (2021) 3573–3587 e3529.
[26] V.A. Huynh-Thu, et al., Inferring regulatory networks from expression data using tree-based methods, PLoS One 5 (9) (2010).
[27] L. Jiang, et al., GiniClust: detecting rare cell types from single-cell gene expression data with Gini index, Genome Biol. 17 (1) (2016) 144.
[28] N.C. Joyce, et al., Effect of overexpressing the transcription factor E2F2 on cell cycle progression in rabbit corneal endothelial cells, Invest. Ophthalmol. Vis. Sci. 45 (5) (2004) 1340–1348.
[29] M. Jung, et al., Analysis of the expression pattern of the schizophrenia-risk and intellectual disability gene TCF4 in the developing and adult brain suggests a role in development and plasticity of cortical and hippocampal neurons, Mol. Autism. 9 (2018) 20.
[30] L.M. Khachigian, et al., Egr-1-induced endothelial gene expression: a common theme in vascular injury, Science (New York, N.Y.) 271 (5254) (1996) 1427–1431.
[31] V.Y. Kiselev, T.S. Andrews, M. Hemberg, Challenges in unsupervised clustering of single-cell RNA-seq data, Nat. Rev. Genet. 20 (5) (2019) 273–282.
[32] V.Y. Kiselev, et al., SC3: consensus clustering of single-cell RNA-seq data, Nat. Methods 14 (5) (2017) 483–486.
[33] A.M. Klein, et al., Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells, Cell 161 (5) (2015) 1187–1201.
[34] F. Ma, M. Pellegrini, ACTINN: automated identification of cell types in single cell RNA sequencing, Bioinformatics (Oxford, England) 36 (2) (2020) 533–538.
[35] S. Mahony, P.V. Benos, STAMP: a web tool for exploring DNA-binding motif similarities, Nucleic Acids Res. 35 (Web Server issue) (2007) W253–W258.
[36] H. Matsumoto, et al., SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation, Bioinformatics (Oxford, England) 33 (15) (2017) 2314–2321.
[37] T. Moerman, et al., GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks, Bioinformatics (Oxford, England) 35 (12) (2019) 2159–2161.
[38] E. Papalexi, R. Satija, Single-cell RNA sequencing to explore immune cell heterogeneity, Nat. Rev. Immunol. 18 (1) (2018) 35–45.
[39] J. Park, et al., Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease, Science (New York, N.Y.) 360 (6390) (2018) 758–763.
[40] K. Pearson, The problem of the random walk, Nature 72 (1865) (1905) 294.
[41] A. Pratapa, et al., Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data, Nat. Methods 17 (2) (2020) 147–154.
[42] A.B. Rosenberg, et al., Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding, Science (New York, N.Y.) 360 (6385) (2018) 176–182.
[43] P. Sangwung, et al., KLF2 and KLF4 control endothelial identity and vascular integrity, JCI Insight. 2 (4) (2017), e91700.
[44] C.A. Shaut, et al., HOXA13 is essential for placental vascular patterning and labyrinth endothelial specification, PLoS Genet. 4 (5) (2008), e1000073.
[45] W. Sikora-Wohlfeld, et al., Assessing computational methods for transcription factor target gene identification based on ChIP-seq data, PLoS Comput. Biol. 9 (11) (2013), e1003342.
[46] D. Steinley, Properties of the Hubert-Arabie adjusted Rand index, Psychol. Methods 9 (3) (2004) 386–396.
[47] M. Tan, et al., Gli3 mutation rescues the generation, but not the differentiation, of oligodendrocytes in Shh mutants, Brain Res. 1067 (1) (2006) 158–163.
[48] B. Van de Sande, et al., A scalable SCENIC workflow for single-cell gene regulatory network analysis, Nat. Protoc. 15 (7) (2020) 2247–2276.
[49] J.M. Vela, et al., Interleukin-1 regulates proliferation and differentiation of oligodendrocyte progenitor cells, Mol. Cell. Neurosci. 20 (3) (2002) 489–502.
[50] T. Wang, J. Bai, S. Nabavi, Single-cell classification using graph convolutional networks, BMC Bioinform. 22 (1) (2021) 364.
[51] Y. Wang, et al., Prognostic cancer gene signatures share common regulatory motifs, Sci. Rep. 7 (1) (2017) 4750.
[52] C. Xu, Z. Su, Identification of cell types from single-cell transcriptomes using a novel clustering method, Bioinformatics (Oxford, England) 31 (12) (2015) 1974–1980.
[53] R. Yan, et al., Endothelial interferon regulatory factor 1 regulates lipopolysaccharide-induced VCAM-1 expression independent of NFkappaB, J. Innate. Immun. 9 (6) (2017) 546–560.

[54] A. Zakrzewska, et al., Macrophage-specific gene functions in Spi1-directed innate immunity, Blood 116 (3) (2010) e1–11.
[55] A. Zeisel, et al., Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq, Science (New York, N.Y.) vol. 347 (6226) (2015) 1138–1142.
[56] C. Zhao, et al., Dual regulatory switch through interactions of Tcf7l2/Tcf4 with stage-specific partners propels oligodendroglial maturation, Nat. Commun. 7 (2016) 10883.
[57] Y. Zheng, X. Li, H. Hu, PreDREM: a database of predicted DNA regulatory motifs from 349 human cell and tissue samples, Database (Oxford) (2015) 2015.
[58] J. Zurauskiene, C. Yau, pcaReduce: hierarchical clustering of single cell transcriptional profiles, BMC Bioinform. 17 (2016) 140.