scDirect: key transcription factor identification for directing cell state transitions based on single-cell multi-omics data

Chen Li[1], Sijie Chen[1], Yixin Chen[1], Haiyang Bian[1], Minsheng Hao[1], Lei Wei[1,*] and Xuegong Zhang[1,2]

Abstract

Cell state transitions are complicated processes that occur in various life activities. Understanding and artificially manipulating them have been longstanding challenges. Substantial experiments reveal that the transitions could be directed by several key transcription factors (TFs), and some computational methods have been developed to alleviate the burden of biological experiments on identifying key TFs. However, most existing methods employ data with resolution on the cell population, instead of individual cells, which will influence the identification quality due to cell heterogeneity. Besides, they require collecting abundant samples for candidate cell states and are not available for unknown cell states. As for the alternative single-cell analysis methods, they generally concentrate on differentially expressed genes between cell states but can hardly identify key TFs responsible for directing cells state transition. Here we present scDirect, a computational framework to identify key TFs based on single-cell multi-omics data. scDirect models the TF identification task as a linear inverse problem, and solve it with gene regulatory networks enhanced by a graph attention network. Through a benchmarking on a single-cell human embryonic stem cell atlas, we systematically demonstrate the robustness and superiority of scDirect against other alternative single-cell analysis methods on TF identification. With application on various single-cell datasets, scDirect exhibits high capability in identifying key TFs in cell differentiation, somatic cell conversion, and cell reprogramming. Furthermore, scDirect can quantitatively identify experimentally validated reprogramming TF combinations, which indicates the potential of scDirect to guide and assist the experimental design in cellular engineering. We envision that scDirect could utilize rapidly increasing single-cell datasets to identify key TFs for directing cell state transitions, and become an effective tool to facilitate regenerative medicine and therapeutic discovery.

[1]MOE Key Laboratory of Bioinformatics and Bioinformatics Division of BNRIST, Department of Automation, Tsinghua University, Beijing 100084, China
[2]Center for Synthetic and Systems Biology, School of Life Sciences and School of Medicine, Tsinghua University, Beijing 100084, China
[*]To whom correspondence should be addressed. Email: weilei92@tsinghua.edu.cn

1 Introduction

Transitions between different cellular states are essential in various biological processes, such as cell differentiation, development, and disease. The widespread practice of cell engineering has successfully implemented transitions of different cellular states, including reprogramming of somatic cells to pluripotent stem cells, directed differentiation of pluripotent stem cells to somatic cells, and direct conversions between somatic cells [1-4]. These transitions are commonly implemented with forced expression of several specific transcription factors (TFs). It has been proved that a small number of key TFs is sufficient to direct cell state transitions [5-7].

There are roughly 2,000 different TFs in humans [8, 9], composing a large space of plausible sets of TFs. It is time-consuming and labor-intensive to identify key TFs that can direct a certain cell state transition from the space experimentally. Researchers have developed a series of computational methods to help identify key TFs in cell state transitions [10-12]. Most of these methods are designed to predict TFs based on bulk omics data generated from a mixed population of cells or tissues. D'Alessio et al. [13] proposed an entropy-based method to calculate a score for each TF from gene expression data. CellNet [14] incorporates the gene regulatory network (GRN) structure inferred from gene expression and calculates a network influence score for each TF. The score represents how well the TF-perturbed GRN matches the target-cell-state GRN. As GRNs inferred from gene expression data may bring some false positive TF-target pairs [10], methods like Mogrify [15] introduce the external DNA-protein interactions from STRING [16] to construct an experimentally validated GRN. Moreover, ANANSE [17] builds GRNs from both gene expression and chromatin immunoprecipitation sequencing (ChIP-seq) data, and TFs are ranked according to their degree of regulating genes vital to cell reprogramming. Besides methods primarily based on gene expression, there are some other methods that only employed chromatin accessibility profiles also achieving good performance [11, 18, 19].

Although the methods based on bulk data provide abundant resources to identify the regulatory TFs and reveal the mechanism of cell state transition, these methods requires abundant samples of cells with candidate states [14], and thus is unusable for rare cell types or undefined cell states. Additionally, the heterogeneity of the bulk data may reduce the quality of constructed GRNs and the performance of TF identification [20]. Yet, the development of single-cell sequencing technologies [21-23] has provided an unprecedent opportunity to facilitate study of characterization of cell identity, inference of GRNs, and transitions between different cell states at the single-cell level. As for the GRN construction, the methods based on both single-cell RNA sequencing (scRNA-seq) data and single-cell ATAC sequencing (scATAC-seq) data exhibit better performance than previous methods based on single-omics data [24-26]. The constructed GRN can be utilized for TF perturbation, the inverse task of TF identification, which perturbs cells by the influence of given specific TFs' variation propagating on the GRN [25]. Regarding TF identification in cell state transitions, some methods like differential expression analysis (DEA) and Dynamo can be employed as alternative solutions [27, 28], while these methods only output important genes during cell state transitions instead of key TFs responsible for directing cell state transitions.

To better leverage the significantly increasing single-cell datasets for TF identification, we present scDirect, a computational framework to identify key TFs in cell state transitions. scDirect models the TF identification task as a linear inverse problem, and leverage the graph attention network (GAT)- enhanced GRN inferred from single-cell multi-omics data to solve the alterations of total TFs. We benchmark scDirect on a TF-overexpression human embryonic stem cell (hESC) atlas [29], and the results show a superior performance of scDirect. Then we apply scDirect on a mouse single-cell multi-omics dataset [23], and known key TFs of different lineages of mouse hair follicle development are identified by scDirect. Furthermore, we curate 5 single-cell datasets of different

experimentally validated reprogramming cases, and scDirect identifies most key TFs of these cases and outperforms existing methods. Moreover, scDirect could quantitatively identify feasible reprogramming TF combinations among thousands of possible combinations. Together, these results suggest scDirect is a reliable and effective tool to identify key TFs directing cell state transitions at the single-cell level.

2 Methods

2.1 Overview of the scDirect framework

scDirect is a computational framework to identify key TFs directing cell state transitions (Fig. 1). We simplify the transition as a linear process. Taking scRNA-seq data and scATAC-seq data as input, scDirect first constructs a primary GRN with scRNA-seq data and scATAC-seq data, and then enhances the GRN with GAT to obtain more putative TF-target pairs with high confidence. After, scDirect models the TF identification task as a linear inverse problem and solves the expected alteration of each TF with Tikhonov regularization. Then the directing score of each TF directing the source cell type to the target cell type is calculated to provide a ranked TF list.

2.2 Primary GRN construction with CellOracle

As we model cell state transition as a linear process, we expect a GRN maintaining linear relations between TFs and target genes. CellOracle [25] is a reliable tool to construct linear GRNs with a regression method, and it is proved to be more accurate than tree-based ensemble methods, such as SCENIC [30] and GENIE3 [31]. CellOracle takes both scRNA-seq data and scATAC-seq data as input. The scRNA-seq data should contain cells of both the source and target state. The scATAC-seq data are not required to be simultaneously sequenced in the same cells of scRNA-seq data but are the best to be derived from similar tissues to guarantee the reliability.

Following the CellOracle workflow, we first use scATAC-seq data to construct a base GRN matrix $\widetilde{\boldsymbol{A}} = [\tilde{a}_{i,j}] \in \mathbb{R}^{\tilde{m}\times\tilde{n}}$, where $\tilde{m}$ and $\tilde{n}$ denote the number of targets and TFs, respectively. The target here means a gene that is regulated by a specific TF, and may also be a TF itself. $\tilde{a}_{i,j} \in [0,1]$ represents whether TF j regulates gene i.

After, we process scRNA-seq data following the standard pipeline of Scanpy [27]. We select highly variable genes (HVGs) between source and target state cells with a fixed dispersion threshold. As the feature selection with HVGs

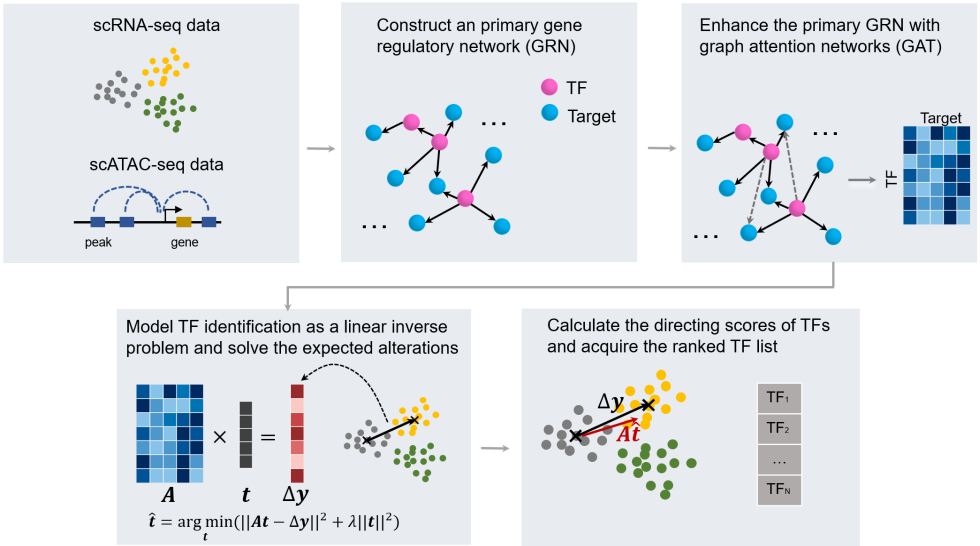

Fig. 1. A graphical illustration of scDirect.

would miss some important TFs, we further select TFs to supplement the interested gene set. We calculate three sets of TFs: (1) Top 400 TFs highly variable in source and target cells; (2) Top 100 TFs differentially expressed between source and target cells; (3) Top 100 TFs highly expressed in target cells. We take the intersection of these sets as supplementary TFs and append them into the interested gene list. The final interested gene list usually includes 2,000 to 3,000 genes.

We then use CellOracle to obtain a refined GRN matrix $\widehat{\boldsymbol{A}} = [\hat{a}_{i,j}] \in \mathrm{R}^{m\times n}$ based on $\widetilde{\boldsymbol{A}}$ and the scRNA-seq data of the interested genes, where m and n denote number of genes and TFs in the interested gene set, respectively, and $\hat{a}_{i,j}$ represents the regulatory coefficient of TF j to target i. For each transition from a source cell state to a target cell state we calculate a GRN matrix $\widehat{\boldsymbol{A}}$, and the calculation of $\widetilde{\boldsymbol{A}}$ and $\widehat{\boldsymbol{A}}$ follows the default parameters of CellOracle.

2.3 GRN enhancement with graph attention networks (GAT)

Though incorporating the chromatin accessibility data with transcriptomics data is helpful to refine the TF-target links, the dropout events of scATAC-seq data and the incomplete TF binding motif knowledge may lead to the missing of important TF-target links. Following the idea that graph neural networks can predict the interactions of genes [32, 33], we employee GAT to recover missing TF-target links.

As shown in Supplementary Fig. 1, the GAT prediction model consists of two parts, the GAT encoder and the multilayer perceptron (MLP) predictor. The input of the encoder is the node feature matrix $\boldsymbol{X} \in \mathbb{R}^{c,m}$ and the adjacent matrix $\boldsymbol{M} \in \mathbb{R}^{m\times m}$, where c is the total number of source and target cells. $\boldsymbol{X}$ is the processed log-normalized scRNA-seq expression matrix, and $\boldsymbol{M}$ is transformed from $\widehat{\boldsymbol{A}}$ by connecting each TF-target pair whose value is not zero in $\widehat{\boldsymbol{A}}$. $\boldsymbol{X}$ and $\boldsymbol{M}$ comprise the input graph. In the input graph, each node is a gene represented by a vector $\boldsymbol{x}_{\cdot,\boldsymbol{j}} \in \mathbb{R}^c$, a column of $\boldsymbol{X}$. The encoder consists of three graph attentional layers, each layer learns a shared weight matrix $\boldsymbol{W} \in \mathbb{R}^{F'\times F}$ and takes a set of node features as input, $\boldsymbol{G} = \{\boldsymbol{g}_1, \boldsymbol{g}_2, \dots, \boldsymbol{g}_m\}, \boldsymbol{g}_i \in \mathbb{R}^F$, and produces a new set of node features, $\boldsymbol{G}' = \{\boldsymbol{g}'_1, \boldsymbol{g}'_2, \dots, \boldsymbol{g}'_m\}, \boldsymbol{g}_i \in \mathbb{R}^{F'}$, where F and F' are feature numbers. In the first layer, F is equal to c. The attention coefficient of gene j to gene i is calculated by:

$$e_{i,j} = a(\boldsymbol{W}\boldsymbol{g}_i, \boldsymbol{W}\boldsymbol{g}_j) \tag{1}$$

where the attention mechanism a is a single layer feedforward network. Only neighbors of gene i are used to calculate the attention coefficients, and then the coefficient is normalized by the Softmax function:

$$\alpha_{i,j} = \mathrm{softmax}_j(e_{i,j}) = \frac{\exp\left(\mathrm{LeakyRelu}(e_{i,j})\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(\mathrm{LeakyRelu}(e_{i,k})\right)} \tag{2}$$

where $\mathcal{N}_i$ is the first-order neighbors of gene i, including i, and LeakyRelu is a non-linear activation function. Multi-head attention mechanism is then applied to stabilize the learning process, and the output of gene i is:

$$\boldsymbol{g}'_i = \|_{k=1}^{K}\ \sigma\left(\sum_{j\in\mathcal{N}_i}\alpha_{i,j}^k \boldsymbol{W}^k \boldsymbol{g}_j\right) \tag{3}$$

where $\|$ is concatenation operation, and σ is the ELU activation function. For the last graph attention layer, the concatenation is replaced by averaging:

$$\boldsymbol{g}'_i = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in\mathcal{N}_i}\alpha_{i,j}^k \boldsymbol{W}^k \boldsymbol{g}_j\right). \tag{4}$$

After the GAT encoder, the gene i and gene j are encoded to $\widehat{\boldsymbol{g}}_i$ and $\widehat{\boldsymbol{g}}_j$, respectively. The predicted score of gene i regulating gene j is calculated by:

$$s_{i,j} = f(\widehat{\boldsymbol{g}}_i \,\|\, \widehat{\boldsymbol{g}}_j) \tag{5}$$

where f is a three-layer MLP. Here we denote $\boldsymbol{L} = [l_{i,j}] \in \mathbb{R}^{m\times n}$ as a binary label matrix indicating the regulatory relations between TFs and targets, and L is transformed from $\widehat{\boldsymbol{A}}$ by converting all the non-zero values in $\widehat{\boldsymbol{A}}$ to 1.

Here we use the binary cross entropy loss to optimize the model:

$$\mathcal{L} = -\sum_{i=1}^{\hat{n}}\sum_{j=1}^{\hat{m}}\left(l_{i,j}\log\left(s_{i,j}\right) + \left(1 - l_{i,j}\right)\log\left(1 - s_{i,j}\right)\right). \tag{6}$$

The details of model training and predicted link filtering are shown in Supplementary Note 1. After adding the recovered links, we acquire an enhanced GRN matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$.

2.4 Modelling TF identification as a linear inverse problem

Suppose the transition is from source cell state $\mathcal{S}$ to target cell state $\mathcal{T}$. The processed gene expression data of $\mathcal{S}$ is $\boldsymbol{X_0} \in \mathbb{R}^{c_0,m}$, and processed gene expression data of $\mathcal{T}$ is $\boldsymbol{X_1} \in \mathbb{R}^{c_1,m}$, where c_0 and c_1 are the cell number of $\mathcal{S}$ and $\mathcal{T}$, respectively. We then average the expression across all the cells of $\mathcal{S}$ and $\mathcal{T}$, respectively, to acquire mean expression vectors $\boldsymbol{y_0} \in \mathbb{R}^{m\times 1}$ and $\boldsymbol{y_1} \in \mathbb{R}^{m\times 1}$. The state transition vector can be calculated by:

$$\Delta\boldsymbol{y} = \boldsymbol{y_1} - \boldsymbol{y_0} \tag{7}$$

where $\Delta\boldsymbol{y}$ indicates the direction that governs the cell state transition. We assume the alteration of TFs will direct cells to deviate from the original state by influencing the regulated target genes. With a linear GRN matrix A, we can model the transition process by the regulating relations of TFs to targets:

$$\boldsymbol{At} = \Delta\boldsymbol{y} \tag{8}$$

where $\boldsymbol{t} \in \mathbb{R}^{n\times 1}$ represents the variations of TFs. The solving of $\boldsymbol{t}$ with known $\boldsymbol{A}$ and $\Delta\boldsymbol{y}$ is commonly referred to as a linear inverse problem [34]. This equation describes the state transition that TFs only transmit one-step signals to the regulated targets, while in the GRN the TFs can propagate regulatory signals with multiple iterations. Following the setting of CellOracle, we set the propagation iterations to 3, and the GRN matrix is modified as:

$$\boldsymbol{A}' = \lambda_1\boldsymbol{A} + \lambda_2\boldsymbol{BA} + \lambda_3\boldsymbol{B}^2\boldsymbol{A} \tag{9}$$

where $\boldsymbol{B} = [b_{i,j}] \in \mathbb{R}^{m\times m}$ is the GRN matrix transformed from A, for gene j that regulates gene i the $b_{i,j}$ is set as the value in $\boldsymbol{A}$, and otherwise is set to 0. λ_1, λ_2 and λ_3 are fixed to 0.6, 0.2 and 0.2, respectively. In our case, the row number m of $\boldsymbol{A}'$ is larger than the column number n, which leads to an over-determined problem and may not have a solution. Consequently, we aim to find a $\hat{\boldsymbol{t}}$ that ensures $\boldsymbol{A}'\hat{\boldsymbol{t}}$ closely resembles $\Delta\boldsymbol{y}$. Additionally, we add the Tikhonov regularization to restrict the range of $\hat{t}$, and it can be solved by:

$$\hat{\boldsymbol{t}} = \underset{\boldsymbol{t}}{\operatorname{argmin}}(||\boldsymbol{A}'\boldsymbol{t} - \Delta\boldsymbol{y}||^2 + \lambda||\boldsymbol{t}||^2) \tag{10}$$

where $\lambda||\boldsymbol{t}||^2$ is a Tikhonov regularization, also called L2 regularization, and λ controls the degree of the regularization. We apply the Ridge regression function in the scikit-learn library [35] to solve the problem, and the alpha parameter is set to 1. The solved $\hat{\boldsymbol{t}}$ represents the expected alterations of TFs.

After acquiring the expected alteration vector $\hat{\boldsymbol{t}}$, for each TF we calculate a directing score to measure the importance in the transition procedure. For each TF, keep the corresponding value and set other values to 0 in $\hat{\boldsymbol{t}}$ to get $\hat{\boldsymbol{t}}'$, and the Pearson correlation coefficient (PCC) between $\boldsymbol{A}'\hat{\boldsymbol{t}}'$ and $\Delta\boldsymbol{y}$ is calculated as the directing score. The calculated vector $\boldsymbol{A}'\hat{\boldsymbol{t}}'$ is denoted as perturbed transition vector (PTV). The final TF set is selected based on both expected alterations and directing scores with a same quantile threshold, and ranked by directing scores. In order to acquire a candidate TF list with a specific number, the quantile threshold is calculated with iterative searching. The directing score of a TF combination could be calculated in a same way.

2.5 Benchmark construction on a TF-overexpression hESC atlas

To our knowledge, there is not a quantitatively benchmarking of TF identification on single-cell datasets. We adopt a single-cell atlas built by Joung et al. to construct the benchmark of TF identification [29]. They infected hESCs with a lentivirus library to perform single TF overexpression in each cell. These hESCs were then profiled by scRNA-seq

to get a TF overexpression single-cell atlas. After down sampling, the atlas retains 671,453 cells covering 3,266 TFs. With TF overexpression directing hESCs to differentiate into different cell states, this atlas can be an experimentally validated ground truth to evaluate the key TF identification in cell state transitions.

The original atlas is clustered to 9 major clusters. Clusters 6-8 are confirmed to be differentiated cells and were further subclustered to 25 minor clusters [29]. The major cluster 5 is the most adjacent cluster to the differentiated cells in the original low-dimensional space, and we assume it as the source cell state to be converted to those differentiated cells. Then we randomly sample 2,000 cells in major cluster 5 and combine them to the 25 minor clusters to form the benchmarking dataset. The 2,000 sampled cells are regarded as the source cell state that possibly transits to the 25 target cell states (Supplementary Fig. 2). Each target cell state comprises distinct groups of overexpressed TFs, and these TFs are considered as ground-truth key TFs directing differentiation from the source cell state to the target cell states. For each TF in each target cell state, we divide the number of cells that indicate the TF ORF in the state by number of cells that indicate this TF ORF in the whole atlas, to calculate the percentage. Then we filter out the TFs of percentage less than 5% and counting cells less than 5 in each target cell state, and 15 target cell states are retained with key TFs counting from 1 to 8 (Fig. 2a).

Then for each target cell state we define a TF identification score (TIS) between the ground-truth key TF list $\mathcal{K}$ and identified candidate ranked TF list $\mathcal{C}$:

$$\mathrm{TIS} = \frac{1}{N_{\mathcal{K}}} \sum_{i=1}^{N_{\mathcal{K}}} \frac{N_{\mathcal{C}} - \mathrm{rank}_{\mathcal{C}}(k_i)}{N_{\mathcal{C}}} \tag{11}$$

where $N_{\mathcal{K}}$ is length of $\mathcal{K}$, $N_{\mathcal{C}}$ is the length of $\mathcal{C}$, and k_i is the ith TF in $\mathcal{K}$. If k_i is in $\mathcal{C}$, $\mathrm{rank}_{\mathcal{C}}(k_i)$ means the rank of k_i in $\mathcal{C}$, and the rank starts from zero not including the former elements in $\mathcal{C}$. For example, if top three elements in $\mathcal{C}$ is k_1, c_1, c_2, and k_2 (c_1 and c_2 are not in $\mathcal{K}$), and $\mathrm{rank}_{\mathcal{C}}(k_1)$ and $\mathrm{rank}_{\mathcal{C}}(k_2)$ are 0 and 2, respectively. For the k_i not in $\mathcal{C}$, the $\mathrm{rank}_{\mathcal{C}}(k_i)$ will be $N_{\mathcal{C}}$. Consequently, TIS can be a comprehensive metric ranging from 0 to 1 and considering both the ranks and numbers of identified TFs.

2.6 Datasets and analysis methods

In this paper we used a series of single-cell datasets, mainly including the scRNA-seq and scATAC-seq datasets. Every scRNA-seq dataset is accompanied with a scATAC-seq dataset, and ensure both are from the same samples or the homologous tissues. Details and accessions of the datasets are listed in Supplementary Table 1. The computational tools used for data processing and analysis are summarized in Supplementary Table 2.

3 Results

3.1 Graph attention networks enhance the gene regulatory signals

We adopted the GRN benchmark in CellOracle to validate the improvement on the original GRN by GAT. Five tissues available both in the Tabula Muris scRNA-seq dataset [36] and mouse sci-ATAC-seq atlas data [37] were selected as the ground-truth datasets: heart, kidney, liver, lung and spleen. The ground-truth GRNs were curated using 1,298 chromatin immunoprecipitation followed by sequencing (ChIP–seq) datasets for 80 regulatory factors across these tissues [25]. The TF number of each ground-truth GRN ranges from 7 to 44, and the TF-target connection number ranges from 340 to 33,247. The base GRN was provided by mouse sci-ATAC-seq atlas [37], and 13 samples across 5 tissues in the scRNA-seq dataset were utilized to conduct the GRN benchmarking.

For each sample, we used CellOracle to calculate a primary GRN, and then GAT was applied to enhance the GRN. As shown in Fig. 2a, the GAT prediction model robustly learns the regulatory relations between TFs and targets for each fold, and gives the prediction result with a high area under the receiver operating characteristic (AUROC) score

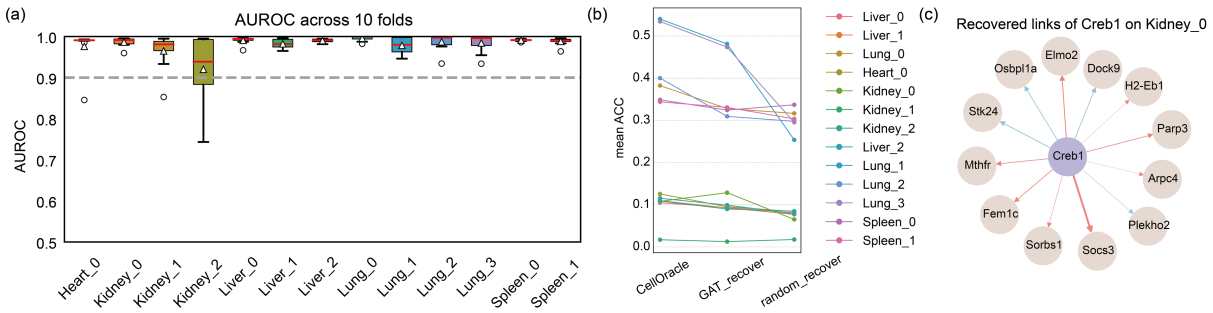

Fig. 2. Graph attention network (GAT) module enhances the regulatory signal of original GRN. (a) Boxplots of area under the receiver operating characteristic (AUROC) on the test set across 13 samples. The box plots indicate the medians (centerlines), means (triangles), first and third quartiles (bounds of boxes). (b) Mean TF-target prediction accuracy comparison across CellOracle, GAT recovering, and random sampling. (d) GAT recovered network of Creb1. Ground-truth links are annotated in red.

on the test set. For example, in sample Liver_0 and Kidney_0, the prediction model achieves AUROC over 0.980 across all folds. The fold-average AUROC score ranges from 0.920 (Kidney_2) to 0.996 (Lung_0) across samples, which indicates GAT model is highly effective in capturing the regulatory signals of different tissues.

To further investigate the GAT recovered TF-target links, we calculated the accuracy (ACC) of predicted links compared to the ground-truth GRN for each TF in different samples. Then we randomly sampled unconnected links as the control group, and the number of samples links is equal to the number of recovered links. The link prediction accuracy comparison among CellOracle, GAT recovering and random sampling are illustrated in Fig. 2b. The performance of GAT recovered links predominantly depends on the quality of original GRN, and among 11 of 13 samples, the predicted links exhibited higher mean ACC scores than the random links.

We illustrated the recovered links of a specific TF Creb1, a kind of cAMP responsive element modulator (Fig. 2d). 8 of 12 links (red lines) are in the ground-truth kidney GRN, and Arpc4 and Mthrf are demonstrated to be targets of Creb1 in Jaspar database [38]. All the results demonstrated that the GAT module captures the regulatory relations and enhances the regulatory signals of the primary GRN.

3.2 Benchmarking TF identification on the TF-overexpression hESC atlas

We benchmarked scDirect on the TF-overexpression hESC atlas (Methods). Each target state contains a set of ground-truth key TFs, and the number of the ground-truth TFs ranges from 1 to 8 (Fig. 3a). We compared scDirect with some baseline methods. Differential expression (DE) was performed by the Wilcoxon rank-sum test. Existing methods for TF identification such as Mogrify and ANANSE are not practicable in the benchmarking as they require to collect datasets for specific cell states. Besides, the design for bulk data also hinders the application of these methods. To overcome the limitations, we replace the GRN inference part of ANANSE by CellOracle, and ANANSE is then used to calculate the importance score of each TF in the differential GRN of the target and source state. We also included scDirect without GAT enhancement (scDirect_WOE) to evaluate the effect of the GAT module.

We acquired a list of ranked top 10 TFs for each method to perform the benchmarking. As shown in Fig. 3c, scDirect significantly outperformed DE and ANANSE and achieved higher TISs (defined in Section 2.5). Moreover, the result suggests a slight improvement of scDirect to scDirect_WOE. scDirect outperforms DE in 16 of 18 target states except target state 4 and target state 14 (Fig. 3d). In most target states, such as target state 8-1 and 16, the ground-truth key TFs ranked high in scDirect identified TF list, while they may not appear in the DE identified list. As for target state 4 and 14, scDirect identified MITF in a low rank, and KLF17 is one of a few TFs that lacks the binding motif in the

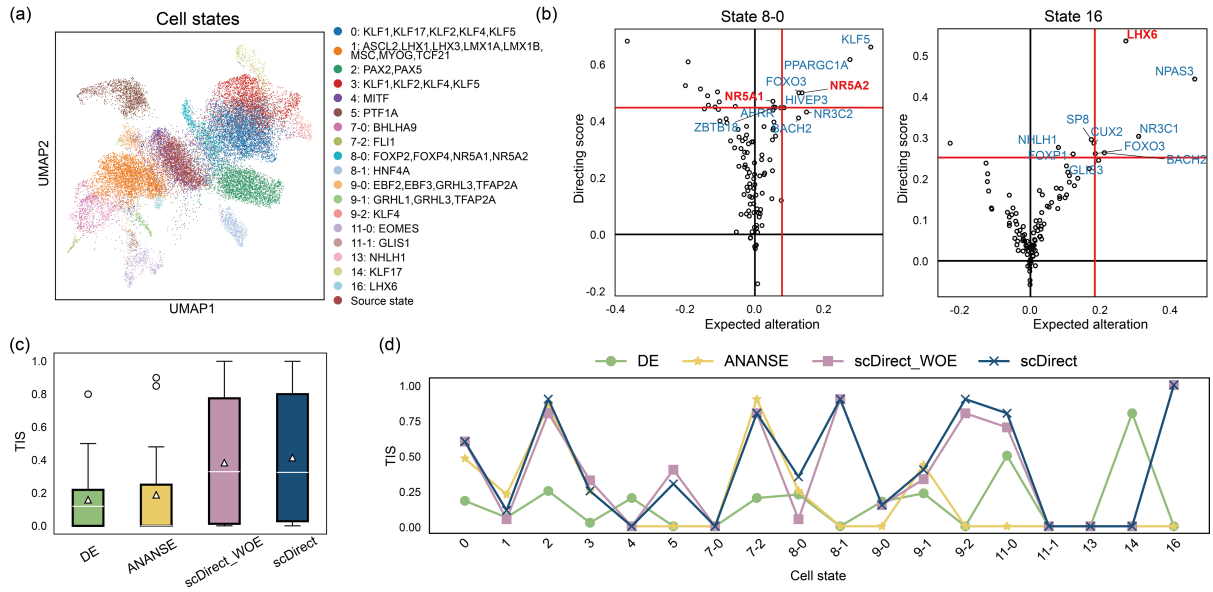

Fig. 3. TF identification benchmarking on a single-cell human embryonic stem cell (hESC) atlas. (a) UMAP visualization of the single-cell hESC atlas. (b) scDirect TF identification plot on target state 2 and 16. Red lines are the quantile thresholds to filter 10 TFs. Key TFs are annotated in red. (c) TF identification score (TIS) comparison across 18 target states. The box plots indicate the medians (centerlines), means (triangles), first and third quartiles (bounds of boxes). (d) Line plots of TIS comparison across 18 target states.

database so we missed it at the beginning of GRN inference. ANANSE achieved best performance in target state 1, 7-2, and 9-1, while in most of remaining states it showed poor performance. It is reasonable of the instability of ANANSE because it ignores part information of TFs whose difference between target state and source state is not statistically significant. Compared to scDirect_WOE, scDirect obtained better performance in 6 target states (1, 2, 8-0, 9-1, 9-2, and 12-0), comparable performance in 10 target states, and poorer performance in 2 target states (3 and 5). Taking state 8-0 as an example (Fig. 3b), scDirect identified NA5R2 and NA5R1 with the 4th and the 5th rank, respectively, while scDirect_WOE only identified NA5R2 with the 9th rank. Besides, scDirect raised the ranks of key TFs like LMX1A, PAX2, GRHL1, KLF4, and EOMES (Supplementary Fig. 3).

Focusing on the scDriect identified TF list for each target state, each TF is given two biologically meaningful scores of the expected alteration and the directing score (Fig. 3b and Supplementary Fig. 3, Section 2.4). The expected alteration represents the expression variation that the TF is supposed to change to direct the source state to the target state, and the directing score indicates how similar the perturbed cell state to target cell state after changing the TF with expected alteration, which is used to rank the TFs. The identified key TFs generally exhibited significantly higher directing scores than remaining TFs (Fig. 3b and Supplementary Fig. 3), which indicates their importance in directing cell state transitions. For instance, scDirect identified key TF LHX6 in target 16 with the highest expected alteration and directing score.

3.3 scDirect identifies the TFs directing mouse hair follicle development

We then tried to demonstrate the reliability of scDirect in identifying key TFs in cell development. We collected a SHARE-seq dataset profiling mouse skin with paired scRNA-seq data and scATAC-seq data [23]. This dataset comprises cells of mouse hair follicle development system. In this system, Transit-Amplifying Cells (TACs)

differentiate into different lineages: inner root sheath (IRS), hair shaft, or medulla. These lineages have been utilized to evaluate various computational methods [39, 40]. Due to a large imbalance in cell types of original dataset, we subsampled equal number of each cell type in the hair follicle development system and obtained a dataset of 2,688 cells (Fig. 4a). 2,748 genes were retained after gene filtering. We applied scDirect on each lineage and selected top 10 TFs as the candidate TFs (Fig. 4b).

In the IRS lineage, Gata3 is the top 1 identified TF with expected alteration 0.422 and directing score 0.691. It has been reported that Gata3 plays a direct role in differentially regulating cell lineages of hair follicle [41], and loss of Gata3 negatively impacts the formation of IRS [42]. Our identification result is consistent with existing findings, and indicates Gata3 is a key TF directing the IRS lineage (Fig. 4c). Among other identified TFs in the IRS lineage, Tfap2a is considered to be up-regulated in IRS during early hair follicle morphogenesis [43], and Maf is also involved in hair morphogenesis of IRS layers [44].

As for the hair shaft lineage, scDirect identified Lef1 as the top 1 TF. Lef1 is demonstrated as a key regulator of Wnt/β-catenin signaling in hair shaft differentiation [45]. Besides, Runx1 is co-localized with Lef1 and is a specific marker of hair shaft [46], and Hoxc13 is reviewed to affect hair shaft differentiation with various alternative mechanisms [47].

Though medulla lineage has been less explored compared to other two lineages, we can still find some existing literature to support our identified TFs. For instance, the deletion of Prdm1 can result in aberrant medulla cell organization [48], which indicates the important role of Prdm1 in medulla formation. Additionally, Foxq1 is discovered to control medulla differentiation through a common regulatory pathway as Hoxc13 [49]. Even Foxq1 is not a differentially expressed gene (DEG), it can still be identified by scDirect, which suggests scDirect can identify

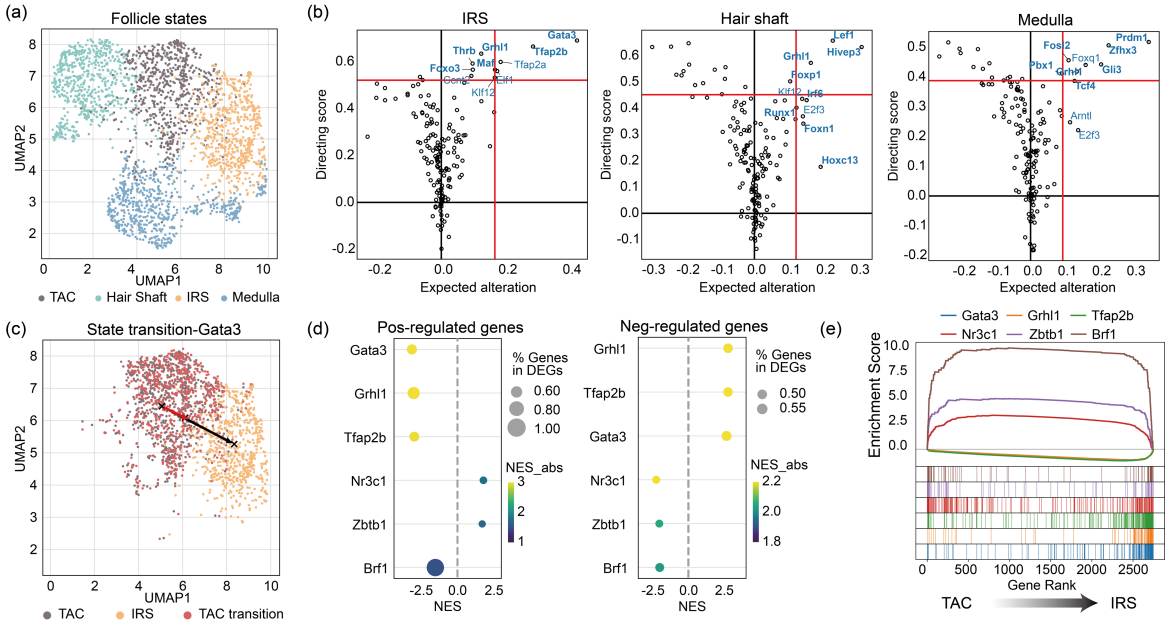

Fig. 4. scDirect identifies key TFs in mouse hair follicle development. (a) UMAP visualization of down-sampled mouse hair follicle scRNA-seq data. (b) scDirect TF identification plot on mouse hair developing lineages of IRS, hair shaft, and medulla. Red lines are the quantile thresholds to filter 10 TFs. DEGs are annotated in bold. (c) UMAP visualization of TAC transiting towards IRS with Gata3 changing the expected variation. (d) Dot plots of gene set enrichment analysis (GSEA) results, showing enrichment of the positively and negatively regulated target sets of scDirect identified TFs within DEGs between TAC and IRS. TFs that rank top 3 and bottom 3 are selected. (e) Enrichment plot of TF positively regulated gene sets within DEGs between TAC and IRS.

key TFs despite the difference between target state and source state is relatively small.

Taking identified TFs of different lineages to conduct a comparison, we found an interesting TF Grhl1 identified in all the lineages. Grhl1 is known to be dynamically expressed in the interfollicular epidermis (IFE) differentiation [50], and is reported to associated with all the three lineages [51], which is in agreement with our findings. We hypothesize Grhl1 may be an important regulator in hair follicle development. With considerable and reliable support of existing literature, we verified that scDriect can identify the key TFs in mouse hair follicle development.

We further characterized the identified TFs in detail taking the IRS lineage as a case. We investigated the relation between regulated targets of identified TFs and state DEGs. The 2,748 genes were ordered from TAC DEGs to IRS DEGs, and we divided regulated targets of each TF into positively regulated part and negatively regulated part according to the regulatory values. We conducted gene enrichment analysis (GSEA) to evaluate enrichment of the regulated target sets within TAC DEGs or IRS DEGs [52]. From the total ranked candidate TFs identified by scDirect, top 3 TFs (Gata3, Grhl1, and Tfap2b) and bottom 3 TFs (Nr3c1, Zbtb1, and Brf1) were selected. Target sets of these TFs were utilized to calculate normalized enrichment scores (NESs), which quantitatively measure the enrichment degree of gene sets. The NESs indicated the positively regulated target sets and negatively regulated target sets of top TFs are enriched in target IRS state and source TAC state, respectively (Fig. 4d), while regulated target sets of bottom TFs did not exhibit such enrichment profiles. Furthermore, we illustrated the enrichment result of positively regulated target set for each TF (Fig. 4e). For the top 3 TFs, most positively regulated targets were enriched in top IRS DEGs, while there was not a significant enrichment bias for bottom TFs regulated targets between the TAC and IRC state (Fig. 4e).

3.4 scDirect identifies the TFs directing cell reprogramming across different scenarios

We collected single-cell datasets for 5 different cell reprogramming scenarios of human and mouse: fibroblasts to keratinocytes [53], fibroblasts to macrophages [54], fibroblasts to cardiomyocytes [55], and B cells to macrophages in mouse [56], and fibroblasts to induced pluripotent stem cells (iPSCs) [2, 57, 58] in human (Supplementary Fig. 4). Each reprogramming case has been extensively investigated with several reported key TFs (Fig. 5a). We denoted these cases from s1 to s5 for brevity (Fig. 5a). We compared scDirect with D'Alessio method, Mogrify, and ANANSE,

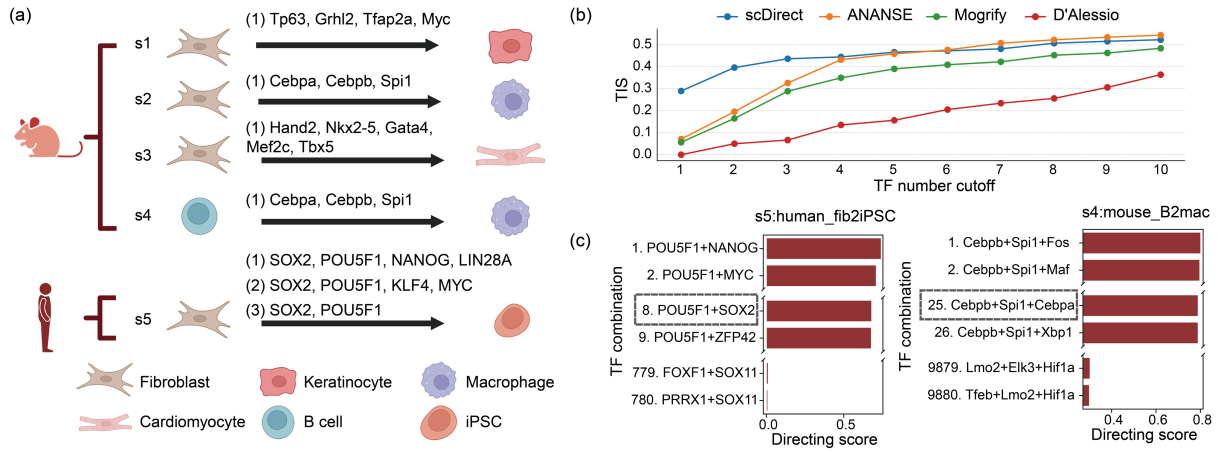

Fig. 5. scDirect identifies the TFs directing cell reprogramming across different scenarios. (a) 5 reprogramming scenarios (s1-s5) with experimentally validated reprogramming TF combinations. (b) The line plots show the comparison of different methods based on the TIS calculated between identified TFs and ground-truth reprogramming TFs. (d) Ranked TF combinations with scDirect predicted directing scores. the ground-truth combinations of case s4 (Cebpa, Cebpb, and Spi1) and s5 (POU5F1 and SOX2) are marked with dashed boxes.

which are methods applied on bulk-omics data to identify key TFs. Mogrify prediction results are downloaded at https://mogrify.net/, and results of D'Alessio et al. and ANANSE are obtained from the original paper [13, 17]. The methods that did not report the TF list of our reprogramming scenarios were excluded in the comparison.

To comprehensively analyze the performance of methods, we calculated the average TIS across different scenarios for each method, and illustrated results with rank cutoff of identified TF list from 1 to 10. With TF rank cutoff from 1 to 3, scDirect outperformed other methods with significantly higher HISs. With TF rank cutoff from 4 to 10, scDirect and ANANSE achieve higher HIS than Mogrify and D'Alessio method, and the performance of scDirect and ANANSE is comparable. Besides, scDirect identifies key TFs with the highest rank in 4 of 5 reprogramming cases (Supplementary Fig. 5). These results demonstrate that scDirect outperforms baselines and identifies key reprogramming TFs with high rank.

In addition to the better performance than existing methods, to the best of our knowledge, scDirect is the only method capable of quantitatively identifying possible TF combinations in cell state transition. For each combination of TFs, scDirect can calculate a directing score (Methods). We took reprogramming case s4 and s5 as examples. We applied scDirect to acquire a ranked list of 40 top TFs for each case, and then calculated and ranked all the possible combinations of 3 TFs and 2 TFs for case s4 and s5, respectively. As shown in Fig. 5c, the ground-truth combination of case s4 (Cebpa, Cebpb, and Spi1) ranks the 25th in total 9,880 combinations, and the ground-truth combination of case s5 (POU5F1 and SOX2) ranks the 8th in total 780 combinations.

4 Discussion

Understanding the cell state transitions could greatly contribute biomedical research and drug discovery. Many researchers have been dedicated to discovering effective TF combinations directing cell state transitions [7, 54, 59]. The efforts on engineering cells towards expected states significantly facilitate regenerative medicine. In this paper, we presented scDirect, a computational framework to identify key TFs in state transitions. As far as we know, scDirect is the first method specially designed for identifying key TFs for directing cell state transitions at the single-cell level, and can be directly applied on single-cell datasets with either annotated cell types or unnamed clusters. scDirect quantitatively identifies key TFs between cell state transitions and assigns a directing score to arbitrary TF combination. We demonstrate that scDirect exhibit superior performance in cell differentiation, somatic cell conversion, and cell reprogramming.

There are also several avenues for improving scDirect. First, cell state transition is a continuous biological process, and TF variations could be modelled as a changing tendency during the whole transition. Second, not only key TFs but also some other regulation components, such chromatin regulators, splicing factors, and microRNAs [60], could be incorporated into our framework. Third, incorporating with non-linear regulatory relations may also contribute to the performance of scDirect.

Acknowledgements

The work is supported in part by National Natural Science Foundation of China (grants 62373210, 62250005, 61721003), and National Key R&D Program of China (grant 2021YFF1200900).

Reference

1. Wichterle H, Lieberam I, Porter JA, Jessell TM: **Directed differentiation of embryonic stem cells into motor neurons.** *Cell* 2002, **110:**385-397.
2. Takahashi K, Yamanaka S: **Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors.** *Cell* 2006, **126:**663-676.
3. Marson A, Foreman R, Chevalier B, Bilodeau S, Kahn M, Young RA, Jaenisch R: **Wnt signaling promotes reprogramming of somatic cells to pluripotency.** *Cell Stem Cell* 2008, **3:**132-135.
4. Takahashi K, Yamanaka S: **A decade of transcription factor-mediated reprogramming to pluripotency.** *Nature Reviews Molecular Cell Biology* 2016, **17:**183-193.
5. Graf T, Enver T: **Forcing cells to change lineages.** *Nature* 2009, **462:**587-594.
6. Buganim Y, Faddah DA, Jaenisch R: **Mechanisms and models of somatic cell reprogramming.** *Nature Reviews Genetics* 2013, **14:**427-439.
7. Morris SA, Daley GQ: **A blueprint for engineering cell fate: current technologies to reprogram cell identity.** *Cell Research* 2013, **23:**33-48.
8. Fulton DL, Sundararajan S, Badis G, Hughes TR, Wasserman WW, Roach JC, Sladek R: **TFCat: the curated catalog of mouse and human transcription factors.** *Genome Biology* 2009, **10**.
9. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM: **A census of human transcription factors: function, expression and evolution.** *Nature Reviews Genetics* 2009, **10:**252-263.
10. Tran A, Yang PY, Yang JYH, Ormerod J: **Computational approaches for direct cell reprogramming: from the bulk omics era to the single cell era.** *Briefings in Functional Genomics* 2022, **21:**270-279.
11. Hammelman J, Patel T, Closser M, Wichterle H, Gifford D: **Ranking reprogramming factors for cell differentiation.** *Nature Methods* 2022, **19:**812-+.
12. Wang HF, Yang YC, Liu JD, Qian L: **Direct cell reprogramming: approaches, mechanisms and progress.** *Nature Reviews Molecular Cell Biology* 2021, **22:**410-424.
13. D'Alessio AC, Fan ZP, Wert KJ, Baranov P, Cohen MA, Saini JS, Cohick E, Charniga C, Dadon D, Hannett NM, et al: **A Systematic Approach to Identify Candidate Transcription Factors that Control Cell Identity.** *Stem Cell Reports* 2015, **5:**763-775.
14. Cahan P, Li H, Morris SA, da Rocha EL, Daley GQ, Collins JJ: **CellNet: Network Biology Applied to Stem Cell Engineering.** *Cell* 2014, **158:**903-915.
15. Rackham OJL, Firas J, Fang H, Oates ME, Holmes ML, Knaupp AS, Suzuki H, Nefzger CM, Daub CO, Shin JW, et al: **A predictive computational framework for direct reprogramming between human cell types.** *Nature Genetics* 2016, **48:**331-335.
16. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin JY, Minguez P, Bork P, von Mering C, Jensen LJ: **STRING v9.1: protein-protein interaction networks, with increased coverage and integration.** *Nucleic Acids Research* 2013, **41:**D808-D815.
17. Xu Q, Georgiou G, Frolich S, van der Sande M, Veenstra GJC, Zhou HQ, van Heeringen SJ: **ANANSE: an enhancer network-based computational approach for predicting key transcription factors in cell fate determination.** *Nucleic Acids Research* 2021, **49:**7966-7985.
18. Bailey TL: **DREME: motif discovery in transcription factor ChIP-seq data.** *Bioinformatics* 2011, **27:**1653-1659.

19. Berest I, Arnold C, Reyes-Palomares A, Palla G, Rasmussen KD, Giles H, Bruch PM, Huber W, Dietrich S, Helin K, Zaugg JB: **Quantification of Differential Transcription Factor Activity and Multiomics-Based Classification into Activators and Repressors: diffTF.** *Cell Reports* 2019, **29:**3147-+.
20. Morris SA, Cahan P, Li H, Zhao AM, Roman AKS, Shivdasani RA, Collins JJ, Daley GQ: **Dissecting Engineered Cell Types and Enhancing Cell Fate Conversion via CellNet.** *Cell* 2014, **158:**889-902.
21. Shalek AK, Satija R, Shuga J, Trombetta JJ, Gennert D, Lu DN, Chen PL, Gertner RS, Gaublomme JT, Yosef N, et al: **Single-cell RNA-seq reveals dynamic paracrine control of cellular variation.** *Nature* 2014, **510:**363-+.
22. Pott S, Lieb JD: **Single-cell ATAC-seq: strength in numbers.** *Genome Biology* 2015, **16**.
23. Ma S, Zhang B, LaFave LM, Earl AS, Chiang Z, Hu Y, Ding JR, Brack A, Kartha VK, Tay T, et al: **Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin.** *Cell* 2020, **183:**1103-+.
24. Bravo Gonzalez-Blas C, De Winter S, Hulselmans G, Hecker N, Matetovici I, Christiaens V, Poovathingal S, Wouters J, Aibar S, Aerts S: **SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks.** *Nature methods* 2023, **20:**1355-1367.
25. Kamimoto K, Stringa B, Hoffmann CM, Jindal K, Solnica-Krezel L, Morris SA: **Dissecting cell identity via network inference and in silico gene perturbation.** *Nature* 2023, **614:**742-+.
26. Wang LF, Trasanidis N, Wu T, Dong GL, Hu M, Bauer DE, Pinello L: **Dictys: dynamic gene regulatory network dissects developmental continuum with single-cell multiomics.** *Nature Methods* 2023, **20:**1368-+.
27. Wolf FA, Angerer P, Theis FJ: **SCANPY: large-scale single-cell gene expression data analysis.** *Genome Biology* 2018, **19**.
28. Qiu X, Zhang Y, Martin-Rufino JD, Weng C, Hosseinzadeh S, Yang D, Pogson AN, Hein MY, Min KH, Wang L, et al: **Mapping transcriptomic vector fields of single cells.** *Cell* 2022, **185:**690-+.
29. Joung J, Ma S, Tay T, Geiger-Schuller KR, Kirchgatterer PC, Verdine VK, Guo BL, Arias-Garcia MA, Allen WE, Singh A, et al: **A transcription factor atlas of directed differentiation.** *Cell* 2023, **186:**209-+.
30. Aibar S, Gonzalez-Blas CB, Moerman T, Van Anh H-T, Imrichova H, Hulselmans G, Rambow F, Marine J-C, Geurts P, Aerts J, et al: **SCENIC: single-cell regulatory network inference and clustering.** *Nature Methods* 2017, **14:**1083-+.
31. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P: **Inferring Regulatory Networks from Expression Data Using Tree-Based Methods.** *Plos One* 2010, **5**.
32. Chen G, Liu Z-P: **Graph attention network for link prediction of gene regulations from single-cell RNA-sequencing data.** *Bioinformatics* 2022, **38:**4522-4529.
33. Li H, Sun Y, Hong H, Huang X, Tao H, Huang Q, Wang L, Xu K, Gan J, Chen H, Bo X: **Inferring transcription factor regulatory networks from single-cell ATAC-seq data based on graph neural networks.** *Nature Machine Intelligence* 2022, **4:**389-+.
34. Bai Y, Chen W, Chen J, Guo W: **Deep learning methods for solving linear inverse problems: Research directions and paradigms.** *Signal Processing* 2020, **177**.
35. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al: **Scikit-learn: Machine Learning in Python.** *Journal of Machine Learning Research* 2011, **12:**2825-2830.
36. Schaum N, Karkanias J, Neff NF, May AP, Quake SR, Wyss-Coray T, Darmanis S, Batson J, Botvinnik O,

Chen MB, et al: **Single-cell transcriptomics of 20 mouse organs creates a *Tabula Muris*.** *Nature* 2018, **562:**367-+.

37. Cusanovich DA, Hill AJ, Aghamirzaie D, Daza RM, Pliner HA, Berletch JB, Filippova GN, Huang X, Christiansen L, DeWitt WS, et al: **A Single-Cell Atlas of *In Vivo* Mammalian Chromatin Accessibility.** *Cell* 2018, **174:**1309-+.
38. Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, Lemma RB, Turchi L, Blanc-Mathieu R, Lucas J, Boddie P, Khan A, Perez NM, et al: **JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles.** *Nucleic Acids Research* 2022, **50:**D165-D173.
39. Tran A, Yang P, Yang JYH, Ormerod JT: **scREMOTE: Using multimodal single cell data to predict regulatory gene relationships and to build a computational cell reprogramming model.** *Nar Genomics and Bioinformatics* 2022, **4**.
40. Lynch AW, Theodoris CV, Long H, Brown M, Liu XS, Meyer CA: **MIRA: joint regulatory modeling of multimodal expression and chromatin accessibility in single cells.** *Nature Methods* 2022, **19:**1097-+.
41. Kurek D, Garinis GA, van Doorninck JH, van der Wees J, Grosveld FG: **Transcriptome and phenotypic analysis reveals Gata3-dependent signalling pathways in murine hair follicles.** *Development* 2007, **134:**261-272.
42. Kaufman CK, Zhou P, Pasolli HA, Rendl M, Bolotin D, Lim KC, Dai X, Alegre ML, Fuchs E: **GATA-3: an unexpected regulator of cell lineage determination in skin.** *Genes & Development* 2003, **17:**2108-2122.
43. Panteleyev AA, Mitchell PJ, Paus R, Christiano AM: **Expression patterns of the transcription factor AP-2α during hair follicle morphogenesis and cycling.** *Journal of Investigative Dermatology* 2003, **121:**13-19.
44. Miyai M, Tanaka YG, Kamitani A, Hamada M, Takahashi S, Kataoka K: **c-Maf and MafB transcription factors are differentially expressed in Huxley's and Henle's layers of the inner root sheath of the hair follicle and regulate cuticle formation.** *Journal of Dermatological Science* 2010, **57:**178-182.
45. Narhi K, Jarvinen E, Birchmeier W, Taketo MM, Mikkola ML, Thesleff I: **Sustained epithelial β-catenin activity induces precocious hair development but disrupts hair follicle down-growth and hair shaft formation.** *Development* 2008, **135:**1019-1028.
46. Raveh E, Cohen S, Levanon D, Negreanu V, Groner Y, Gat U: **Dynamic expression of Runx1 in skin affects hair structure.** *Mechanisms of Development* 2006, **123:**842-850.
47. Awgulewitsch A: ***Hox* in hair growth and development.** *Naturwissenschaften* 2003, **90:**193-211.
48. Telerman SB, Rognoni E, Sequeira I, Pisco AO, Lichtenberger BM, Culley OJ, Viswanathan P, Driskell RR, Watt FM: **Dermal Blimp1 Acts Downstream of Epidermal TGFβ and Wnt/β-Catenin to Regulate Hair Follicle Formation and Growth.** *Journal of Investigative Dermatology* 2017, **137:**2270-2281.
49. Potter CS, Peterson RL, Barth JL, Pruett ND, Jacobs DF, Kern MJ, Argraves WS, Sundberg JP, Awgulewitsch A: **Evidence that the satin hair mutant gene *Foxq1* is among multiple and functionally diverse regulatory targets for *Hoxc13* during hair follicle differentiation.** *Journal of Biological Chemistry* 2006, **281:**29245-29255.
50. Joost S, Zeisel A, Jacob T, Sun XY, La Manno G, Lönnerberg P, Linnarsson S, Kasper M: **Single-Cell Transcriptomics Reveals that Differentiation and Spatial Signatures Shape Epidermal and Hair Follicle Heterogeneity.** *Cell Systems* 2016, **3:**221-+.
51. Joost S, Annusver K, Jacob T, Sun XY, Dalessandri T, Sivan U, Sequeira I, Sandberg R, Kasper M: **The**

Molecular Anatomy of Mouse Skin during Hair Growth and Rest. *Cell Stem Cell* 2020, **26:**441-+.

52. Fang Z, Liu X, Peltz G: **GSEApy: a comprehensive package for performing gene set enrichment analysis in Python.** *Bioinformatics* 2023, **39**.
53. Kurita M, Araoka T, Hishida T, O'Keefe DD, Takahashi Y, Sakamoto A, Sakurai M, Suzuki K, Wu J, Yamamoto M, et al: **In vivo reprogramming of wound-resident cells generates skin epithelial tissue.** *Nature* 2018, **561:**243-+.
54. Feng R, Desbordes SC, Xie H, Tillo ES, Pixley F, Stanley ER, Graf T: **PUA and C/EBPα/β convert fibroblasts into macrophage-like cells.** *Proceedings of the National Academy of Sciences of the United States of America* 2008, **105:**6057-6062.
55. Addis RC, Ifkovits JL, Pinto F, Kellam LD, Esteso P, Rentschler S, Christoforou N, Epstein JA, Gearhart JD: **Optimization of direct fibroblast reprogramming to cardiomyocytes using calcium activity as a functional measure of success.** *Journal of Molecular and Cellular Cardiology* 2013, **60:**97-106.
56. Xie HF, Ye M, Feng R, Graf T: **Stepwise reprogramming of B cells into macrophages.** *Cell* 2004, **117:**663-676.
57. Yu J, Vodyanik MA, Smuga-Otto K, Antosiewicz-Bourget J, Frane JL, Tian S, Nie J, Jonsdottir GA, Ruotti V, Stewart R, et al: **Induced pluripotent stem cell lines derived from human somatic cells.** *Science* 2007, **318:**1917-1920.
58. Huangfu D, Maehr R, Guo W, Eijkelenboom A, Snitow M, Chen AE, Melton DA: **Induction of pluripotent stem cells by defined factors is greatly improved by small-molecule compounds.** *Nature Biotechnology* 2008, **26:**795-797.
59. Fu J-D, Stone NR, Liu L, Spencer CI, Qian L, Hayashi Y, Delgado-Olguin P, Ding S, Bruneau BG, Srivastava D: **Direct Reprogramming of Human Fibroblasts toward a Cardiomyocyte-like State.** *Stem Cell Reports* 2013, **1:**235-247.
60. Badia-i-Mompel P, Wessels L, Mueller-Dott S, Trimbour R, Flores RRO, Argelaguet R, Saez-Rodriguez J: **Gene regulatory network inference in the era of single-cell multi-omics.** *Nature Reviews Genetics* 2023.