
9 10 *Subject Section*

11 12 Single-cell and spatial multiomic inference of gene 13 regulatory networks using SCRIPPro

14
15 Zhanhe Chang^{2,3,6}, Yunfan Xu^{1,3,6}, Xin Dong^{1,3,6}, Yawei Gao^{2,3}, Chenfei
16 Wang^{1,3,4,5*}

17
18 ¹Key Laboratory of Spine and Spinal Cord Injury Repair and Regeneration of Ministry of Education,
19 Department of Orthopedics, Tongji Hospital, School of Life Sciences and Technology, Tongji University,
20 200092, China,

21
22 ²Institute for Regenerative Medicine, Shanghai East Hospital, Shanghai Key Laboratory of Signaling
23 and Disease Research, School of Life Sciences and Technology, Tongji University, Shanghai, China,

24
25 ³Frontier Science Center for Stem Cell Research, Tongji University, Shanghai, China,

26
27 ⁴National Key Laboratory of Autonomous Intelligent Unmanned Systems, Tongji University, Shanghai
28 200120, China,

29
30 ⁵Frontier Science Center for Intelligent Autonomous Systems, Tongji University, Shanghai 200120,
31 China

32
33 ⁶These authors contributed equally to this work

34
35 *To whom correspondence should be addressed.

36 Associate Editor: Dr. Macha Nikolski

37 Received on XXXXX; revised on XXXXX; accepted on 11-Jul-2024

38 Abstract

39 **Motivation:** The burgeoning generation of single-cell or spatial multiomic data allows for the characterization of
40 gene regulation networks (GRNs) at an unprecedented resolution. However, the accurate reconstruction of GRNs
41 from sparse and noisy single-cell or spatial multiomic data remains challenging.

42 **Results:** Here, we present SCRIPPro, a comprehensive computational framework that robustly infers GRNs for
43 both single-cell and spatial multi-omics data. SCRIPPro first improves sample coverage through a density
44 clustering approach based on multiomic and spatial similarities. Additionally, SCRIPPro scans transcriptional
45 regulator (TR) importance by performing chromatin reconstruction and *in silico* deletion analyses using a
46 comprehensive reference covering 1,292 human and 994 mouse TRs. Finally, SCRIPPro combines TR-target
47 importance scores derived from multiomic data with TR-target expression levels to ensure precise GRN
48 reconstruction. We benchmarked SCRIPPro on various datasets, including single-cell multiomic data from human
49 B-cell lymphoma, mouse hair follicle development, Stereo-seq of mouse embryos, and Spatial-ATAC-RNA from
50 mouse brain. SCRIPPro outperforms existing motif-based methods and accurately reconstructs cell type-specific,
51 stage-specific, and region-specific GRNs. Overall, SCRIPPro emerges as a streamlined and fast method capable
52 of reconstructing TR activities and GRNs for both single-cell and spatial multi-omic data.

53 **Availability:** SCRIPPro is available at <https://github.com/wanglabtongji/SCRIPPro>.

54 **Contact:** 08chenfeiwang@tongji.edu.cn

55 **Supplementary information:** Supplementary data are available at *Bioinformatics* online.

1 Introduction

2 Transcription regulators (TR), including transcription factors (TF) and
3 chromatin regulators (CR), play a crucial role in gene regulation by
4 influencing transcription rates through mechanisms like recruiting
5 transcriptional initiating complexes and modulating chromatin
6 accessibility(Buenrostro, et al., 2015). TRs form complex gene
7 regulation networks (GRNs) with their target genes, also known as
8 regulons, which are highly dynamic in different cellular contexts and
9 serve as the foundation for various biological processes. Traditional
10 methods for inferring GRNs, such as GENIE3(Huynh-Thu, et al., 2010),
11 LISA(Qin, et al., 2020), GRNBoost2(Moerman, et al., 2019),
12 TIGRESS(Haury, et al., 2012), ppcor(Kim, 2015), and
13 NIMEFI(Ruyssinck, et al., 2014), were primarily designed for bulk
14 samples from mixed cell type tissues. However, these approaches are
15 limited in their ability to accurately capture the regulatory programs
16 operating in different cell types and states. With the advent of single-cell
17 technologies, there has been a surge in the development of methods for
18 inferring GRNs from single-cell transcriptome or epigenome data,
19 including SCENIC(Aibar, et al., 2017), PIDC(Chan, et al., 2017),
20 SCODE(Matsumoto, et al., 2017), SINCERITIES(Papili Gao, et al.,
21 2018), chromVAR(Schep, et al., 2017), and SCRIP(Dong, et al., 2022).

22 However, methods that solely focus on single-cell transcriptome for
23 predictions not only neglect the genuine chromatin accessibility state but
24 also fail to achieve single-cell resolution in GRN predictions, only
25 attaining a cluster level. Additionally, many of these methods heavily rely
26 on motif references to identify potential targets, which results in the loss
27 of the cell type specificity and cannot robustly predict TR activity without
28 motifs, particularly for chromatin regulators. To address these limitations,
29 we previously developed SCRIP(Dong, et al., 2022), a method that
30 reconstructs single-cell TR activity and GRNs from scATAC-seq data by
31 integrating extensive collections of TR ChIP-seq and motif references.
32 Nonetheless, SCRIP's effectiveness is influenced by the universality and
33 the quality of scATAC-seq data. Recent advancements in single-cell
34 multi-omics data have led to the development of new tools for predicting
35 TR activity. For example, tools such as FigR(Kartha, et al., 2022),
36 GRaNIE(Kamal, et al., 2023), DIRECT-NET(Zhang, et al., 2022) and
37 GLUE(Cao and Gao, 2022) utilize paired or integrated multiome data as
38 inputs and employ linear/non-linear regression methods to construct gene
39 regulatory networks. Other tools, like CellOracle(Kamimoto, et al.,
40 2023), offer pre-built GRNs or the ability to create custom-defined GRNs
41 using scATAC-seq data. Meanwhile, tools like MICA(Alanis-Lobato, et
42 al., 2024) use bulk ATAC-seq to identify potential TR binding sites,
43 applying this landscape to refine single-cell transcriptomic data (Badia,
44 et al., 2023). However, it is worth noting that the aforementioned methods
45 still heavily depend on the motif information, and further exploration is
46 required to enhance their accuracy and coverage.

47 The flourishing development of spatial omics enables precise analysis
48 of cellular structures within complex tissues and the spatial interactions
49 between cells. Techniques such as Stereo-seq(Chen, et al., 2022) and
50 STARmap plus(Shi, et al., 2023), have achieved single-cell-level
51 resolution. These methods have significantly enhanced our understanding
52 of gene regulation in specific microenvironments, providing valuable
53 insights into crosstalk between microenvironment interactions and
54 intracellular GRNs. However, most existing tools for predicting TR

55 activity do not consider cellular spatial location. They overlook the
56 impact in expression similarity of different cells or spots within the same
57 microenvironment, leading to inaccurate TR predictions based on spatial
58 transcriptomics data. Furthermore, the emergence of spatial multi-omics
59 technologies, such as spatial ATAC-RNA-seq(Zhang, et al., 2023),
60 provides paired chromatin accessibility states with gene expression. This
 presents an opportunity for accurate TR activity and GRN prediction
 using spatial multiomic information.

 In this study, we have developed SCRIPPro, a computational framework
 designed to predict TR activity and reconstruct TR-centered GRNs for
 both single-cell and spatial multiomic data. SCRIPPro addresses the
 challenge of sparse single-cell or spatial multiomic signals by employing
 a density clustering approach that considers either expression or spatial
 similarities. Additionally, SCRIPPro leverages a comprehensive TR
 reference compiled from TR ChIP-seq peaks obtained from Cistrome
 DB(Zheng, et al., 2019), along with motifs for 1,252 human TRs and 994
 mouse TRs. Finally, SCRIPPro combines TR-target importance from
 epigenomic data with TR-target expression from transcriptomic data to
 construct the GRNs. We demonstrate the robustness and versatility of
 SCRIPPro by applying it to various datasets, including human B-cell
 lymphoma, mouse hair follicle development single-cell multi-ome data,
 mouse embryo Stereo-seq datasets at consecutive developmental time
 points, and P22 mouse brain Spatial-ATAC-RNA data. The results
 showcase the superior performance and utility of SCRIPPro in diverse
 biological contexts.

2 Methods

Overview of the SCRIPPro Method

To decrease computational requirements and to alleviate the impact of
dropout events inherent in single-cell sequencing data, as well as to
improve the stability of the inferred GRNs, we adopted a divide-and-
conquer approach. SCRIPPro begins by constructing SuperCells(Littman,
et al., 2023), which are aggregates of gene expression profiles from
clusters of individual cells exhibiting similar transcriptional activity.
Following this initial step, SCRIPPro identifies a set of marker genes for
each SuperCell. These marker genes serve as representative features of
their respective SuperCells and are subsequently utilized as inputs for
the LISA(Qin, et al., 2020) framework. Within LISA, SCRIPPro
implements ISD analyses to assess the effects of transcriptional regulator
(TR) perturbations on the expression of these marker genes. The
outcome of this process is a GRN constructed at the SuperCell level,
which provides a higher-order representation of the regulatory landscape
across the pooled cellular subpopulations.

In SCRIPPro, for the transcriptomic-only data, we will use similar
strategies from LISA to reconstruct the chromatin landscapes from bulk
DNase/H3K27ac reference. For epigenomic-only data, we will use
SCRIP to infer potential regulators. For multi-omics data, we will
perform alignment and use matched or paired epigenome datasets as
chromatin landscape. Besides, SCRIPPro considers spatial adjacency
when generating SuperCells. We believe that with these features,
SCRIPPro could provide a robust and flexible solution for GRN inference
supporting different single-cell and spatial modalities.

Reference Dataset

ChIP-seq collection We employed the identical ChIP-seq datasets for
SCRIP, which were sourced from the Cistrome Data Browser(Zheng, et
al., 2019). Following a meticulous reorganization and refinement of

Article short title

annotations pertaining to factors, cell types, and tissues, we applied stringent filtering criteria: a median quality score for raw sequences exceeding 25, a uniquely mapped reads proportion surpassing 50%, a PCR Bottleneck Coefficient (PBC) greater than 0.8, an enrichment of peaks by a minimum of 10-fold in quantities exceeding 100, a Fraction of Reads in Peaks (FRIP) above 0.01, and an overlap of the top 5000 peaks with joint DNase I hypersensitive sites (DHS) exceeding 70%. We preserved only those peaks demonstrating a minimum 5-fold enrichment within each dataset. Datasets comprising fewer than 1000 peaks were subsequently excluded. After this filtration process, we amassed a total of 2314 human and 1920 mouse TR ChIP-seq datasets, encompassing 671 and 440 TRs, respectively.

Motif collection In an effort to enhance the representation of transcription factors, we also synthesized pseudo-peaks data by conducting motif scanning. We procured the transcription factor position weight matrix (PWM) motifs for both human and mouse, amalgamated and transformed the data formats, and subsequently utilized the HOMER(Heinz, et al., 2010) software to scan the genome for motif-associated genomic intervals. We juxtaposed the scanning outcomes with the Encyclopedia of DNA Elements (ENCODE)(2012) candidate cis-regulatory elements (ccREs) and the Cistrome Union DHS compilation, excluded any intersections with known blacklisted regions, and augmented the length of each motif locus to 340 base pairs to facilitate comparison with the ChIP-seq datasets. By imposing a filter based on the P-value, we retained the top 25,000 binding sites, ultimately yielding 916 human and 816 mouse pseudo-peaks derived from motif scanning.

Synthetic reference dataset Subsequently, we integrated these meticulously curated SCRIPro datasets into the LISA framework, generating reference HDF5 (Hierarchical Data Format version 5) files for both human and mouse datasets, which comprised the count of ChIP-seq peaks and the metadata associated with the datasets. In summation, the constructed human TR index encompasses 1252 TRs, while the mouse TR index includes 997 TRs.

Regulatory potential model

To assess the influence of transcriptional regulators on their target genes, we conducted a detailed analysis of ChIP-seq datasets contained in HDF5 files using the Regulatory Potential (RP) model(Wang, et al., 2020). The computational formula for RP, adhering to the SCRI methodology, is as follows:

$$S_g = \sum_{i=1}^n 2^{-\frac{d_i}{d_0}}(1)$$

In this equation, n denotes the count of TR binding sites proximal to the TSS of gene g, while d_i signifies the distance from the i-th peak's center to the TSS. For TRs exhibiting more than 20% of peaks in the promoter region, they are categorized as promoter-centric TRs, and the half-decay distance d_0 is set at 1 kb. In contrast, TRs with fewer promoter-localized peaks are classified as enhancer-centric, with their half-decay distance established at 10 kb. To enhance computational efficiency, the analysis was restricted to genes located within a half-maximal regulatory range (half-decay distance) of 15 d_0 , as peak scores beyond this threshold diminish to less than 0.0005.

An enhanced version of the RP model was employed to explore the potential target genes regulated by transcription factors. This model extends beyond incorporating exon information by also accounting for

the regulatory impact of adjacent genes. Specifically, if a peak is detected within the exon region of a gene, the corresponding score is assigned the value of 1, which is then normalized relative to the gene's total exon length. Conversely, should a peak reside in the promoter or exon region of a neighboring gene, its score is designated as 0.

Preprocessing transcriptomic input data

Single-cell RNA-seq data We first preprocess counts matrix for each cell following Scanpy workflow(Wolf, et al., 2018). Next, principal component analysis (PCA) is performed using `scanpy.tl.pca()` function to reduce data dimensionality. We compute neighboring cells for each cell using `scanpy.pp.neighbors()` function, setting the number of neighbors to `N_neighbors` (default 10) and the number of principal components to `N_pcs` (default 40). Finally, we apply `scanpy.tl.umap()` function for UMAP dimensionality reduction and perform Leiden clustering using `scanpy.tl.leiden()` function with a resolution parameter set to resolution (default 0.8) to stratify cellular populations.

Spatial transcriptomics RNA-seq data After normalizing the count matrix, SCRIPro utilizes the approach from STAGATE(Dong and Zhang, 2022) to build a spatial neighbor network (SNN), integrating the similarity between adjacent spots of a given location, and subsequently transforms this spatial data into an undirected neighbor network based on a predetermined radius r . Subsequently, utilizing a cell type-aware module, the SNN is pruned based on pre-clustered gene expression. After constructing SNN, SCRIPro employs a graph attention auto-encoder to integrate gene expression and spatial location. In cell type-aware module, SCRIPro employs a self-attention mechanism for both types of SNNs. The learned spatial similarities from the standard SNN and the cell type-aware SNN are denoted as $\text{att}^{\text{spatial}}$ and $\text{att}^{\text{aware}}$, respectively. The final spatial attribution used is a linear combination of these two (where α , the default hyperparameter set at 0.5, represents the weight of the cell type-aware SNN):

$$\text{att} = (1 - \alpha)\text{att}^{\text{spatial}} + \alpha\text{att}^{\text{aware}}(2)$$

The output of the decoder is considered as the reconstructed normalized expressions. We then perform dimensionality reduction on the integrated data using `scanpy.pp.neighbors()` and `scanpy.tl.umap()`, following the same methodology as above.

SuperCell Construction

After employing leiden clustering(Traag, et al., 2019) to identify cell subsets at a resolution parameter of 0.8, each cluster is treated as an independent RNA-seq dataset. Using a binary search method, we iteratively increase the secondary resolution to obtain more refined leiden cluster classifications until the average number of cells in each small leiden cluster reaches a user-specified N. SuperCells within each large Leiden cluster that contain fewer than 30 cells are merged with the nearest SuperCell to ensure a minimum of 30 cells per SuperCell.

Marker genes are then computed for each SuperCell:

For smaller scRNA-seq datasets, we identify the top 500 marker genes per SuperCell using the `scanpy.get.rank_genes_groups_df()` method. For larger datasets (cell number > 150,000), we recommend you to use large-scale marker gene selection strategy. We extract top 1500 genes in each leiden cluster, and then identify genes expressed above the 60% percentile in each SuperCell as marker genes for subsequent analyses.

Article short title

1
2
3
4
5
6
7 SuperCells with fewer than 35 selected marker genes are excluded from
8 further calculations.

7
8 Chromatin landscape Construction
9

10 SCRIPPro has two different strategies for constructing chromatin
11 landscapes:
12

13 (1) For scRNA-seq data, SCRIPPro use chromatin landscapes constructed
14 in “Reference Dataset” (see Methods). Then SCRIPPro performs LISA’s
15 ISD calculations for each SuperCell’s marker genes within 8 chunks (by
16 default), obtaining results for each SuperCell.

17 (2) In the context of multi-omics data, there are generally two scenarios:
18 whether the barcodes from scRNA-seq and scATAC-seq are matchable.
19 If the barcodes match, such as with 10x Chromium, clustering is first
20 performed using scRNA-seq data to delineate SuperCells, and then the
21 scATAC-seq data is used to construct the landscape for each SuperCell
22 during the ISD step. If the barcodes do not match, as with scMultiome-
23 seq, SCRIPPro initially employs GLUE(Cao and Gao, 2022) to integrate
24 the two datasets, matching RNA and ATAC omics information within
25 the same cell. Subsequently, the reconstruction of the ATAC landscape
26 is carried out following the steps described above. To prepare the
27 scATAC-seq data, we sort it using the bedtools(Quinlan and Hall, 2010)
28 sort command and merge intervals within each sorted TSV file, ensuring
29 that they do not exceed 1000 bases, using the bedtools merge command.
30 Each merged TSV file is then converted into a bigwig format file using
31 the *bedGraphToBigWig* command(Kent, et al., 2002). These bigwig
32 files, corresponding to each SuperCell, are used as the landscape during
33 the ISD step.

34 **In silico deletion in SCRIPPro**

35 SCRIPPro performs *in silico* deletion with the same strategy as LISA.
36 LISA’s chromatin landscape model uses L1-regularized logistic
37 regression to select an optimum sample set for H3K27ac ChIP-seq or
38 DNase-seq samples. LISA first calculated the chrom-RP for each
39 RefSeq gene. The chrom-RP for gene k in sample j is defined as

$$j = \sum_{i \in [t_k - L, t_k + L]} w_i s_{ji}$$

40 L is set to 100 kb, The weight w_i represents the regulatory impact of a
41 locus at position i on the gene k ’s transcription start site at genomic
42 position t_k . s_{ji} is the signal of chromatin profile j at position i .

43
44 LISA also calculates peak-RP for each set of ChIP-seq data. The
45 definition of peak-RP is same to what has mentioned above. Then the
46 ISD method recalculates the chrom-RP after erasing the signal in all 1-
47 kb windows containing at least one peak from a putative regulatory
48 cistrome, and then comparing the model RPs with and without deletion
49 to produce a Δ RP value for each gene. The combined statistics method
50 for TR ranking compares the peak-RPs or Δ RPs of the query gene set
51 with that of the background gene set. It uses the one-sided Wilcoxon
52 rank-sum test and combines peak-RP, DNase-seq, and H3K27ac chrom-
53 RP for ChIP-seq-based methods. The Cauchy combination test is used
54 to compute a summary p value for each TR.

55
56 **Calculate SCRIPPro TR score**
57

58 SCRIPPro applies a negative logarithmic transformation to the summary
59 p-values obtained from LISA’s ISD results, serving as the score for

60 individual TRs within each SuperCell.

$$N = -\log(p_{summary})$$

61 Additionally, for each TR and its corresponding targets within a
62 SuperCell, we calculate a z-score relative to the mean:

$$M_{TR} = \frac{E_{i, m} - \text{mean}(E_m)}{\text{std}(E_m)} \quad (5)$$

$$M_{TR} = \begin{cases} 4 & \text{if } x > 4 \\ M_{TR} & \text{if } -4 \leq x \leq 4 \\ -4 & \text{if } x < -4 \end{cases} \quad (6)$$

$$M_{Target} = \frac{\sum_1^k \left(\frac{E_{i, n} - \text{mean}(E_n)}{\text{std}(E_n)} \right)}{k} \quad (7)$$

$$M_{Target} = \begin{cases} 4 & \text{if } x > 4 \\ M_{Target} & \text{if } -4 \leq x \leq 4 \\ -4 & \text{if } x < -4 \end{cases} \quad (8)$$

$$M_{exp} = (M_{TR} + M_{Target}) \quad (9)$$

$$S = N \times \frac{M_{exp} - \min(M_{exp})}{\max(M_{exp}) - \min(M_{exp})} \quad (10)$$

where E represents the expression value of TRs (represented by m) or target genes (represented by n), while i denotes each SuperCell. More specifically, k represents the target genes regulated by TR m. For each set of ChIP-seq data, we select genes with an RP score > 5 as the target genes for this TF. If the number of genes with an RP score > 5 is less than 300, then we choose the top 300 genes in RP score ranking as the target genes. For each TR within a SuperCell, we first calculate the z-score of the TR value to obtain M_{TR} (Equation 5). Then, We average the Z-scores of each target gene to get M_{Target} (Equation 7). Both M_{TR} and M_{Target} clip with (-4,4) as the threshold(Equation 6), and the influence of extreme values on them has been subtracted. Then M_{TR} and M_{Target} is added to obtain the M_{exp} (Equation 9), which is the coefficient for LISA Pvalue correction. The TR activity score (S) is equal to N(negative logarithmic transformation of the summary p-values of each TR) multiplied by the max-min normalized M_{exp} (Equation 10). These steps ensure that both the expression of TR and target genes were included when calculating the final TR activity score.

50 **Dataset analysis methods**51 **(1) CRISPRa human T Cells Dataset**

52 **Preprocessing** We downloaded this dataset (PRJNA787633) from
53 <http://www.perturbbase.cn/download>. The downloaded data has already
54 undergone preprocessing and mixscape analysis. We selected cells with
55 the ‘gene’ label in the SCRIPPro reference TR dataset for downstream
56 AUPRC and AUROC analysis. We applied SCRIPPro (SCRIPPro
57 reference), SCENIC, SCING and LISA (LISA reference) to this dataset
58 with default parameters, and then performed downstream performance
59 testing using the TR enrichment scores inferred by these methods.
60 Specifically, SCRIPPro uses the tf_score matrix, SCENIC uses the
AUCell matrix, and SCING uses the aucs_mtx matrix for TR in the
corresponding cluster of gene_membership, where the LISA P value is
P_value_matrix calculated by SCRIPPro.

Article short title

Performance comparison of different SuperCell settings We tested the performance under different supercell settings by adjusting the clustering parameters of SCRIPro. Specifically, we modified the Cell_num parameter in Ori_Data to values of (10, 20, 30, 50, 65, 100, 200). For Metacell, we set the target_metacell_size to 50, which resulted in 581 metacells with an average of 28.7 cells per metacell. Similarly, we computed supercells using the Cell_num parameter, setting cell_num to 30 to achieve the same granularity as the metacells, resulting in an average of 29.78 cells per supercell. We then calculated the AUROC and AUPRC for each setting, and recorded the number of supercells and the runtime. These results were then visualized using Matplotlib to create line charts.

Performance comparison of different methods We use the ‘gene’ labels in the dataset as the gold standard to compute the AUROC and AUPRC. Specifically, if a cell’s gene label corresponds to a particular TR, it is labeled as 1; otherwise, it is labeled as 0. Using this data and the TR enrichment scores obtained from different methods, we calculate AUPRC and AUROC using the precision_recall_curve and auc functions from the sklearn package.

(2) B-Cell Lymphoma Dataset

Preprocessing In the scRNA-seq dataset, cells with fewer than 200 genes and genes present in fewer than 3 cells were removed. We retained only cells that had both RNA-seq and ATAC-seq counts. We applied SCRIPro with default parameters to the filtered scMultiome-seq peak count matrix to evaluate the activity of transcriptional regulators. For each TR within a SuperCell, we first calculate the z-score of the TR value to obtain M_{TR} (Equation 5). Then, We average the Z-scores of each target gene to get M_{Target} (Equation 7). Both M_{TR} and M_{Target} clip with (-4,4) as the thre-sold(Equation 6),, and the influence of extreme values on them has been subtracted. Then M_{TR} and M_{Target} is added to obtain the M_{exp} (Equation 9), which is the coefficient for LISA Pvalue correction. The TR activity score (S) is equal to N(negative logarithmic transformation of the summary p-values of each TR) multiplied by the the max-min normalized M_{exp} (Equation 10). These steps ensure that both the expression of TR and target genes were included when calculating the final TR activity score. The activity scores of different TRs were visualized as heatmaps with seaborn’s clustermap and projected onto UMAP plots with Scipy. After integrating the original scATAC-seq data with SCRIPro, the resulting landscape was displayed using the IGV genome browser(Robinson, et al., 2011).

Clustering Performance Comparison For the assessment of SCRIPro, we chose to compare it with SCENIC+(Bravo González-Blas, et al., 2023). The SCENIC algorithm(Aibar, et al., 2017), which is the precursor to SCENIC+, is a widely recognized method for GRN inference. SCENIC+ builds upon SCENIC by providing single-cell resolution transcription factor activity scores, allowing for a comprehensive performance comparison with SCRIPro. Since SCRIPro relies on ChIP-seq datasets, it is not suitable for direct comparison with traditional methods that are based on synthetic GRN datasets. Instead, we compared SCRIPro with SCENIC+ in terms of gene expression correlation. Specifically, SCRIPro selects the TR activity score in SuperCells and calculates the correlation with gene expression within those SuperCells. On the other hand, SCENIC+’s

results are correlated with the original cell expression. Both methods were used with their default parameters.

To visualize and cluster the number of target genes of TFs in Tumor B Cells and B Cells, we utilized the `sns.clustermap()` function in the Seaborn library, using the default parameters. Additionally, we employed Metascape(Zhou, et al., 2019) to calculate the GO terms of the overlapping genes identified by the two methods. Finally, we visualized the GO terms using an R script.

(3) Hair Follicle Development Dataset

Preprocessing We annotated SHARE-seq data cell types using labels from the original study. SCRIPro was applied with default parameters to the SHARE-seq data to assess the activity of TRs in each cell.

Pseudotime Analysis We utilized MIRA(Lynch, et al., 2022) for trajectory analysis of the SHARE-seq data. We designated ORS as the starting cell type and IRS, Cortex, and Medulla as the terminal cell types to study the differentiation trajectory in hair follicle development. Cells were ordered by pseudotime using the ‘mira_pseudotime’ column from MIRA results. The activity score of each TR in each terminal cell type (Medulla_prob/IRS_prob/Cortex_prob > 0.8) was visualized in ternary plots.

ATAC-RNA Analysis SCRI(Dong, et al., 2022) was used to perform TR activity analysis on the scATAC-seq data from the SHARE-seq data, applying the SCRI enrichment function with default parameters to the peak count matrix. Subsequently, we transformed the numerical scores from two distinct matrices—single-cell RNA sequencing data (RNA-infer) and single-cell ATAC sequencing data (ATAC-infer)—into their corresponding ranks. We then proceeded to compute the discrepancies between the two matrices by subtracting the RNA-infer ranks from the ATAC-infer ranks, thereby deriving a set of differential rankings that culminate in final score. These scores were clustered and compared according to the pseudotime obtained from MIRA(Lynch, et al., 2022).

(4) Mouse Embryonic Development Dataset

Preprocessing The Stereo-seq dataset detailing mouse embryonic development was procured from the specified database (<https://db.cngb.org/search/project/CNP0001543/>). SCRIPro build SNN and perform graph attention auto-encoder using STAGATE strategy. We utilized SCRIPro to process embryonic data at E16.5 to identify 28 spatial clusters employing a large dataset gene selection strategy. Standard preprocessing steps, including quality control, normalization, and data filtering, were executed prior to downstream analyses. Motif pattern was downloaded from JASPAR(<https://jaspar.elixir.no/>)(Fornes, et al., 2020).

Identification of TR Activity Co-expression Modules The R package Giotto `binSpect()` function was employed to discern TRs exhibiting spatial coherence in their activity scores as determined by SCRIPro. Giotto(Chen, et al., 2023) identified 40 distinct spatial TR modules in E16.5 embryo datatset, which were subsequently analyzed using heatmap clustering. Centrality metrics were calculated using networkx Python package `degree_centrality()` function.

Article short title

TR Activity Analysis and Comparison To assess the cluster purity of the identified spatial domains, we adopted ROGUE score (<https://github.com/PaulingLiu/ROGUE>). Transcription factor activity scores were computed using both SCENIC and SCRIPPro, adhering to their default parameters. The re-clustering of these scores, integrating spatial context, facilitated the computation of the ROGUE score via the R package ROGUE. Normalized Mutual Information (NMI) scores were derived by comparing the clustering outcomes of TR activity scores obtained from both methods against those from STAGATE clustering.

(5) P22 Mouse Brain Spatial Multi-omics Dataset Analysis

Preprocessing The spatial ATAC-RNA-seq multi-omics dataset of the P22 mouse brain was retrieved from the cited source (GSE205055). The `scripro.cal_ISD_parallel()` function from SCRIPPro was then employed to analyze the multi-omics data, resulting in the determination of TR activity scores for each SuperCell. We employed the `binSpect()` function in Giotto to identify spatially variable TRs and subsequently conducted heatmap clustering to discern cell type-specific TRs. The pseudotime spatial trajectory analysis of neural stem cell emergence was explored by spaceFlow(Ren, et al., 2022).

Wasserstein distance calculation To investigate the impact of signaling pathways on TR expression in spatial contexts, we used the Wasserstein distance(Flamary, et al., 2021) to measure differences in TR and gene expression across brain regions, utilizing spot positions as coordinates and expression levels as values. For a ligand expressed in m spots and a receptor expressed in n spots, we formed a matrix $D \in R^{m \times n}$ to record the Euclidean distances between spots, based on spatial coordinates. By identifying an optimal transport $\gamma' \in R^{m \times n}$ that minimizes the total transport cost of ligand and receptor expression distributions: $L \in R^{m \times 1}$ and $R \in R^{n \times 1}$. This total transport cost is determined by summation of the products of the transport value and the Euclidean distance between each spot. Based on the optimal transport plan, the Wasserstein distance can be computed as follows:

$$W(L, R) = \min_{\gamma' \in \tau(L, R)} \langle \gamma', D \rangle = \langle \gamma', D \rangle$$

We filter target gene expression presented in the L-R interactions collection (collected from NicheNet (Browaeys, et al., 2020)), considering these as signaling pathways in cell-cell interactions. Next, we screened the L-R interaction pair based on the expression pair of TR and its target genes in adjacent cells, and considered it to be a potential signaling pathway for cell-cell interaction.

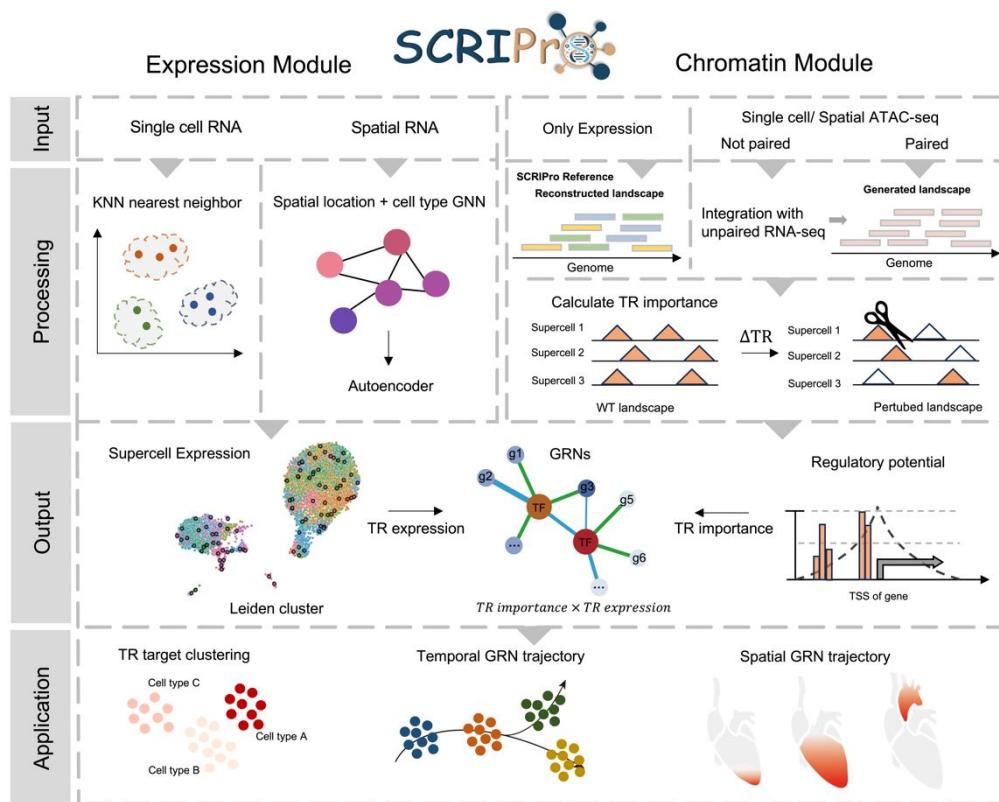


Fig. 1 Overview of SCRIPPro.

SCRIPPro takes single cell RNA-seq or spatial RNA-seq as input. SCRIPPro first employs density clustering using a high coverage SuperCell strategy. While for spatial data, SCRIPPro combines gene expression and cell spatial similarity information to a latent low-dimension embeddings via a graph attention auto-encoder. Then SCRIPPro conducts in silico deletion analyses, utilizing matched scATAC-seq or reconstructed chromatin landscapes from public chromatin accessibility data, to assess the regulatory significance of TRs by RP model in each SuperCell. At last, SCRIPPro combines TR expression and TR to generate TR-centered GRNs at the SuperCell resolution. The output of SCRIPPro can be applied for TR target clustering, temporal GRN trajectory and spatial GRN trajectory.

3 Results

3.1 SCRIPro combines comprehensive chromatin and TR binding references to predict GRNs for both single-cell and spatial multi-omics data

SCRIPro comprises expression and chromatin modules corresponding to input transcriptomic and epigenomic data (Fig. 1). (1) The expression module accepts single-cell or spatial transcriptome data. To overcome gene coverage limitations and minimize drop-out effects, SCRIPro utilizes a density clustering approach based on K-nearest neighbor (KNN) and a graph attention auto-encoder to generate SuperCells(Littman, et al., 2023) with consistent expression patterns or spatial coordinates(Dong and Zhang, 2022) (Fig. 1, Supplementary Fig. S1A-B). The SuperCell expression is then used to evaluate TR expression levels and co-expression patterns within GRNs. For expression-only data, SCRIPro curates an extensive collection of public chromatin references encompassing 1,471 DNase-seq data and 2,575 H3K27ac data for human and mouse samples (Zheng, et al., 2019) (Supplementary Fig. S1C-E). Subsequently, SCRIPro employs a logistic regression-based approach to scan the chromatin reference and reconstruct in silico chromatin landscapes that best match resemble marker genes identified in SuperCells (Supplementary Fig. S1A). (2) The chromatin module of SCRIPro can utilize experimentally paired or GLUE-integrated single-cell or spatial ATAC-seq inputs to build chromatin landscapes. SCRIPro then uses the in silico or paired chromatin landscapes to assess the importance of TRs.

SCRIPro compiles a comprehensive TR reference dataset comprising 2,314 human TR ChIP-seq data and 1,920 mouse TR ChIP-seq data (Fig. S2A-B, Table S1). This reference dataset is enriched with high-quality motifs from cis-BP and HOMER, covering 1,252 human TRs and 994 mouse TRs (Supplementary Fig. S2C-F). SCRIPro conducts in silico deletion analyses of TR binding sites to evaluate the TR impact on the expression of marker genes within SuperCells (Fig. 1 and Supplementary Fig. S1A). The potential TR targets are determined based on the best-matched TR ChIP-seq data for a given SuperCell using a regulatory potential (RP) model (Supplementary Fig. S1A). Subsequently, SCRIPro integrates TR-target expression from transcriptome data and TR-target importance from epigenome data to compute the final TR activity within a SuperCell, identify regulated genes, and construct TR-centered GRNs. Moreover, SCRIPro can directly apply the SCRI method for scATAC-seq-only data to predict TRs (Supplementary Fig. S1B).

SCRIPro's performance is further evaluated through downstream benchmarking, including assessing TR activity accuracy, clustering analysis, identifying cell type or stage-specific GRNs, and region-specific GRNs across various biological systems.

3.2 Performance evaluation and parameter selection

To quantitatively evaluate SCRIPro's performance, we used a published single-cell CRISPR activation dataset that screens for T-cell stimulation regulators in primary human T-cells as the benchmark for GRN inference methods (Fig. 2A). We specifically analyzed cells containing sgRNAs targeting 15 TRs from the original 70 hits to benchmark performance. Our hypothesis was that cells with introduced TR sgRNAs would demonstrate increased TR activity, detectable

through GRN inference algorithms. Initially, we benchmarked various parameters and selected 30 cells per SuperCell, balancing performance and computational efficiency (Fig. 2B-C). While different methods of generating SuperCells showed no significant differences in performance, the SuperCell strategy proved to be more computationally efficient than MetaCell(Persad, et al., 2023) (Fig. 2D). Furthermore, we evaluated the reconstructed TR activities using different methods. Both AUROC and AUPRC metrics indicated that SCRIPro outperformed SCING, SCENIC, and the LISA methods (which SCRIPro was modified from) (Fig. 2E-G). The consistent results were observed at both single-cell and supercell levels, as well as when using different TR references. These findings collectively demonstrate the robust and accurate performance of SCRIPro compared to existing methods.

3.3 SCRIPro identifies tumor-specific GRNs in the human B-cell lymphoma 10X multi-ome dataset

We benchmarked the performance of SCRIPro on a human B-cell lymphoma (small lymphocytic lymphoma, SLL) dataset using 10X single-cell multi-ome containing 14,566 cells. We annotated 9 major cell types based on the expressed marker genes of each lineage, and tumor B-cells form a slightly different cluster compared to normal B-cells (Fig. 3A). For multiome data, SCRIPro first integrates RNA and ATAC data to obtain integrated clusters (Supplementary Fig. S3A). SCRIPro could accurately predicts well-known master regulators including SPI1, IRF1, FLI1, and STAT2 for mono/macrophages, GATA3, RUNX3, SMARCA4, JUND, and MYB for T-cells, PAX5, BCL2 and IRF4 for B and tumor B-cells (Fig. 3B, Supplementary Fig. S3B). The identified TR ChIP-seq reference corresponds well to the chromatin landscape from scATAC-seq at the SuperCell level, suggesting that the ChIP-seq reference is informative in predicting TR binding events (Fig. 3C). We compared the performance of SCRIPro with SCENIC+, a widely used algorithm that infers GRNs based on a combination of motif enrichment and GRNBoost2(Moerman, et al., 2019). SCRIPro successfully identifies IRF8 in plasmacytoid dendritic cells (pDCs), which has been reported to be essential for the development of pDC and type 1 conventional dendritic cells(Sichien, et al., 2016). However, SCENIC+ fails to assign IRF8 scores, possibly due to the limited number of pDC cells. SCRIPro specifically enriches SPIB, a driver regulator that mediates apoptosis through the PI3K-AKT pathway in diffuse B-cell lymphoma(Takagi, et al., 2016), in tumor B-cells but not normal cells, while SCENIC+ fails to predict the SPIB activity in all the B-cells (Fig. 3D). Finally, we assume the inferred TR activity should be highly correlated with their gene expression, with a positive correlation for activators and a negative correlation for repressors. SCRIPro shows a significantly high concordance between TR activity and TR expression compared to SCENIC+, with over 80% (81/105) of the factors showing better consistency (Fig. 3E, Supplementary Fig. S3C). Importantly, SCRIPro predicts activity scores for nearly 800 TRs, while SCENIC+ is only able to evaluate over 100 TR activities (Supplementary Fig. S3D). In summary, these analyses suggest that SCRIPro can accurately infer GRNs globally and outperforms existing methods in terms of consistency between TR activity and TR expression.

B-cell lymphoma develops as a result of abnormal interactions between B cells and the microenvironment during development(Garaud, et al., 2019). Given the accurate identification of TR activities specific to

Article short title

tumor B-cells, including SPIB, using SCRIPro, we conducted a systematic analysis to identify tumor-specific GRNs that potentially drive malignancy. Unsupervised clustering of TR activity identifies three independent clusters in tumor B-cells, compared to two clusters in normal B-cells (Fig. 3F). Both Group 1 and 2 were shared between tumor and normal B-cells, with Group 1 enriched in B-cell activation and differentiation, and Group 2 enriched in cytoplasmic translation that is important for B-cell development (Fig. 3G, Supplementary Fig. S5A-D). Notably, the majority of TRs in Group 3 of tumor B-cells

belonged to the ZNF family, which has been reported to silence retrotransposons and regulate epithelial proliferation(Cassandri, et al., 2017; Imbeault, et al., 2017) (Fig. 3F-H). These analyses suggest that tumor B-cells may employ alternative proliferation strategies, such as activating epithelial proliferating genes. Collectively, our analyses demonstrate the superior accuracy and sensitivity of SCRIPro in identifying cell-type-specific TRs, even enabling the discrimination between similar cell types.

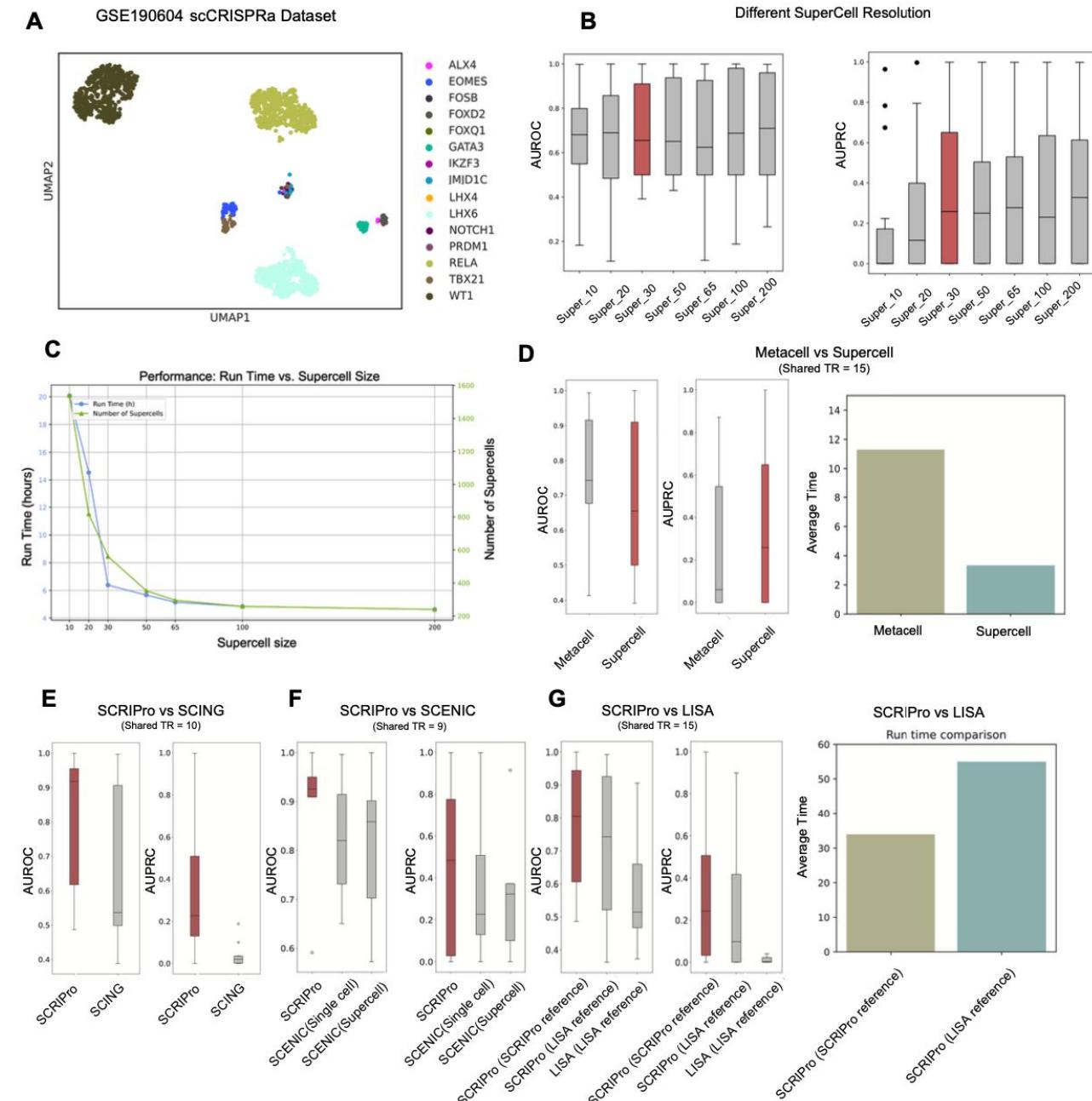


Fig. 2 Benchmark of SCRIPro using single cell CRISPR activation dataset.

- UMAP of GSE190604 single cell CRISPRa dataset.
- AUROC and AUPRC of different SuperCell resolution.
- Running time and SuperCell size of different SuperCell resolution.
- G. Comparison of Metacell and SuperCell (shared TR = 15), SCRIPro and SCING (shared TR = 10), SCRIPro and SCENIC (shared TR = 9), SCRIPro and LISA (shared TR = 15).

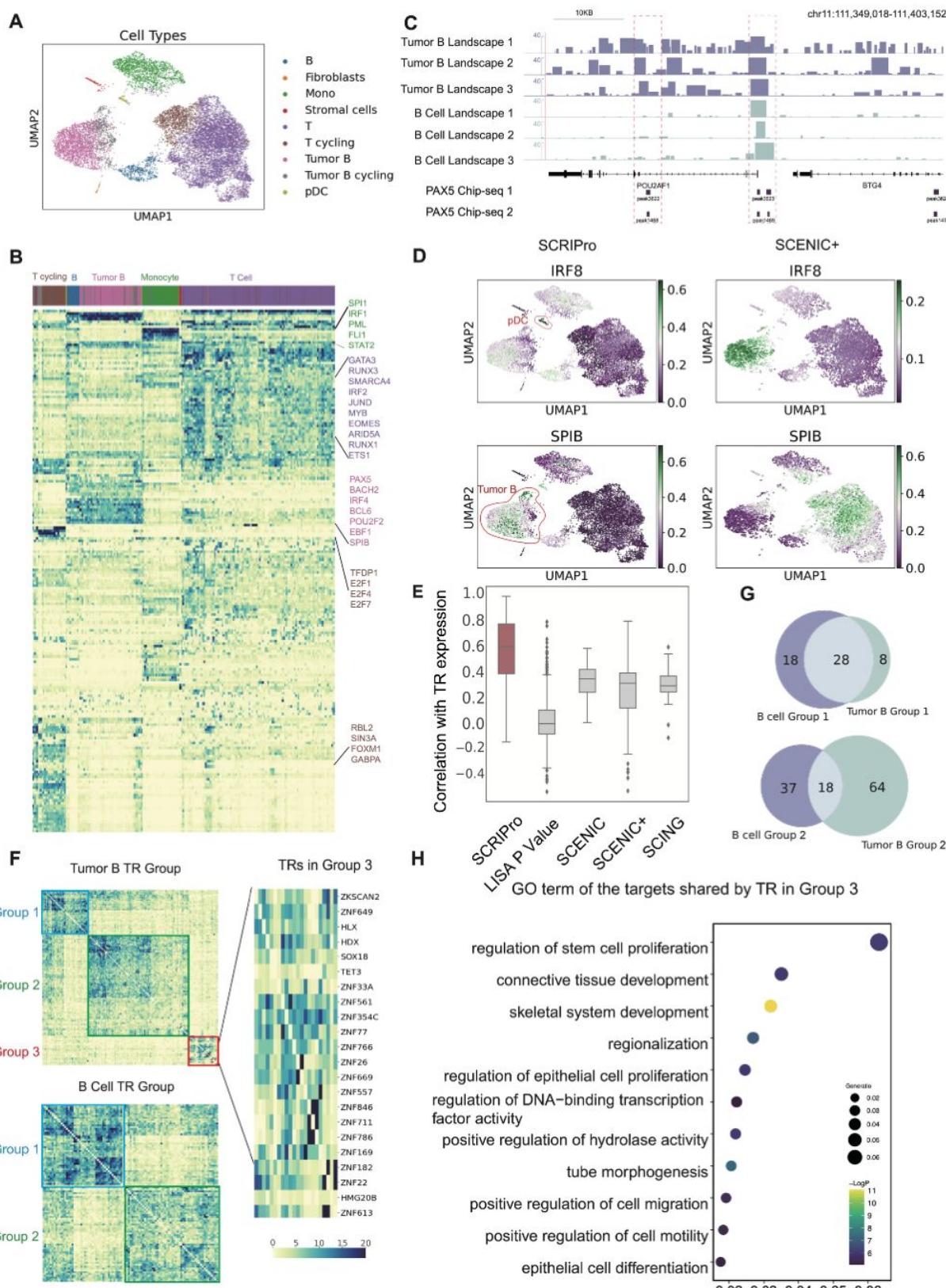
Article short title

3.4 SCRIPro reveals epigenetic priming effects of mouse hair
follicle differentiation

TRs play a crucial role in driving cell type differentiation, compared to static datasets that only contain differentiated cells, reconstructing GRNs from developmental datasets presents challenges due to subtle differences along the development trajectories. We next benchmarked the performance of SCRIPro on a hair follicle differentiation dataset generated using the SHARE-seq protocol(Ma, et al., 2020). The cells were annotated into 7 major cell types including outer root sheath (ORS), transit-amplifying cells-1 (TAC-1), and TAC-2, medulla, hair shaft-cuticle cortex (cortex), inner root sheath (IRS), and mix cells (Supplementary Fig. S6A). Starting from ORS cells, pseudo time analyses suggest three distinct differentiation paths that led to the formation of medulla, cortex and IRS cells (Supplementary Fig. S6B). SCRIPro robustly identifies TRs enriched in the initial ORS cell type and the three different trajectories (Supplementary Fig. S6C, Supplementary Fig. S7A and Supplementary Fig. S8A). For instance, medulla cells exhibited unique TR activity including Prdm1, Pbx3, and Rnf2, known for their significant regulatory roles in medullary-related functions(Roberts, et al., 2017)(Rhee, et al., 2004). Cortex cells were characterized by specific activities of Lef1(Zhang, et al., 2013) and Rora(Steinmayr, et al., 1998). Furthermore, IRS cells exhibited high activity of Gata3, an important TR in the skin stem cell lineage during the initiation of epidermal stratification and hair follicle IRS patterning(Kaufman, et al., 2003) (Supplementary Fig. S6D). These analyses demonstrate the capability of SCRIPro to identify lineage-specific GRNs along the developmental trajectory even with subtle differences.

Traditional GRN prediction methods designed for scRNA-seq datasets, such as SCENIC, PIDC, and SCODE, primarily rely on the co-expression information to infer TR regulation. However, the prediction of TR activity based on gene expression and chromatin accessibility can be decoupled due to the influence of epigenetic priming. Since SCRIPro could both reconstruct the TR activity from expression (transcriptome-only module) and chromatin accessibility (epigenome-only or paired module), we compared the difference between RNA versus ATAC inferred TR activity to systematically evaluate the potential priming effect. We focused on specific developmental lineage, for example, the ORS to medulla path. Encouragingly, our analysis revealed a TR group in which chromatin-inferred activity preceded expression-inferred activity (Supplementary Fig. S6E-F, Group 1), indicating a strong epigenetic priming effect that leads to a delay in target gene expression. Conversely, Group 2 exhibited an opposite trend, with RNA-inferred activity preceding ATAC-inferred activity (Supplementary Fig. S6G). Notably, many factors in this group possess repressive functions including Jarid2 and Mtf2, subunits of the PRC2 complex reported in mouse embryonic stem cells(Zhang, et al., 2011), thus validating the priming effect of repressive H3K27me3 modifications. Group 3, comprising the majority of TRs, displayed largely synchronous patterns regardless of whether they were predicted using RNA or ATAC data (Supplementary Fig. S6E, H). Similar patterns were observed for the ORS to cortex and ORS to IRS directions (Supplementary Fig. S8B). In summary, these analyses confirm the epigenetic priming effect using the hair follicle development SHARE-seq data and also underscore the importance of chromatin landscapes in accurately predicting the TRs required for the differentiation of future trajectories.

Article short title

**Fig. 3** SCRIPro identified tumor-specific GRNs in the human B-cell lymphoma 10X multi-ome dataset.

- A. UMAP of 9 cell types identified in human B-cell lymphoma dataset. Mono: Monocytes. pDC: Plasmacytoid dendritic cells.
- B. Heatmap showing the clustering of TRs by cell type. Top: Cell types annotated in Figure A. Right: Highlighted TRs.
- C. The PAX5 ChIP-seq signal landscape identified by SCRIPro, which is aggregated based on SuperCells, show across tumor B and B cell types for POU5F1 and BTG4 genes on chr11:111349018-111403152.

Article short title

- 1 D. UMAP showing the predicted distributions of IRF8 and SPIB by SCRIPPro and SCENIC+. IRF8 is highlighted in the pDC cell type in SCRIPPro, while SPIB is prominent
2 in Tumor B cells.
- 3 E. Box plot depicting the Pearson correlation between TRs and gene expression by SCRIPPro, LISA P Value, SCENIC, SCENIC+ and SCING.
- 4 F. Heatmap clustering of TRs in tumor B and B cell types on the SCRIPPro TR activity scale. Top: TR heatmap in the tumor B cell type, showing 3 clusters. Bottom: TR
5 heatmap in the B cell type, showing 2 clusters. Right: An enlarged view of the group 3 (outlined in red) of the heatmap for tumor B, with all TRs labeled on the right.
- 6 G. Top: Venn diagram showing the overlap of TRs between B cell group 1 and tumor B cell group 1. Bottom: Venn diagram showing the overlap of TRs between B cell group
7 2 and tumor B cell group 2.
- 8 H. GO terms of the targets shared by TRs in Group 3.

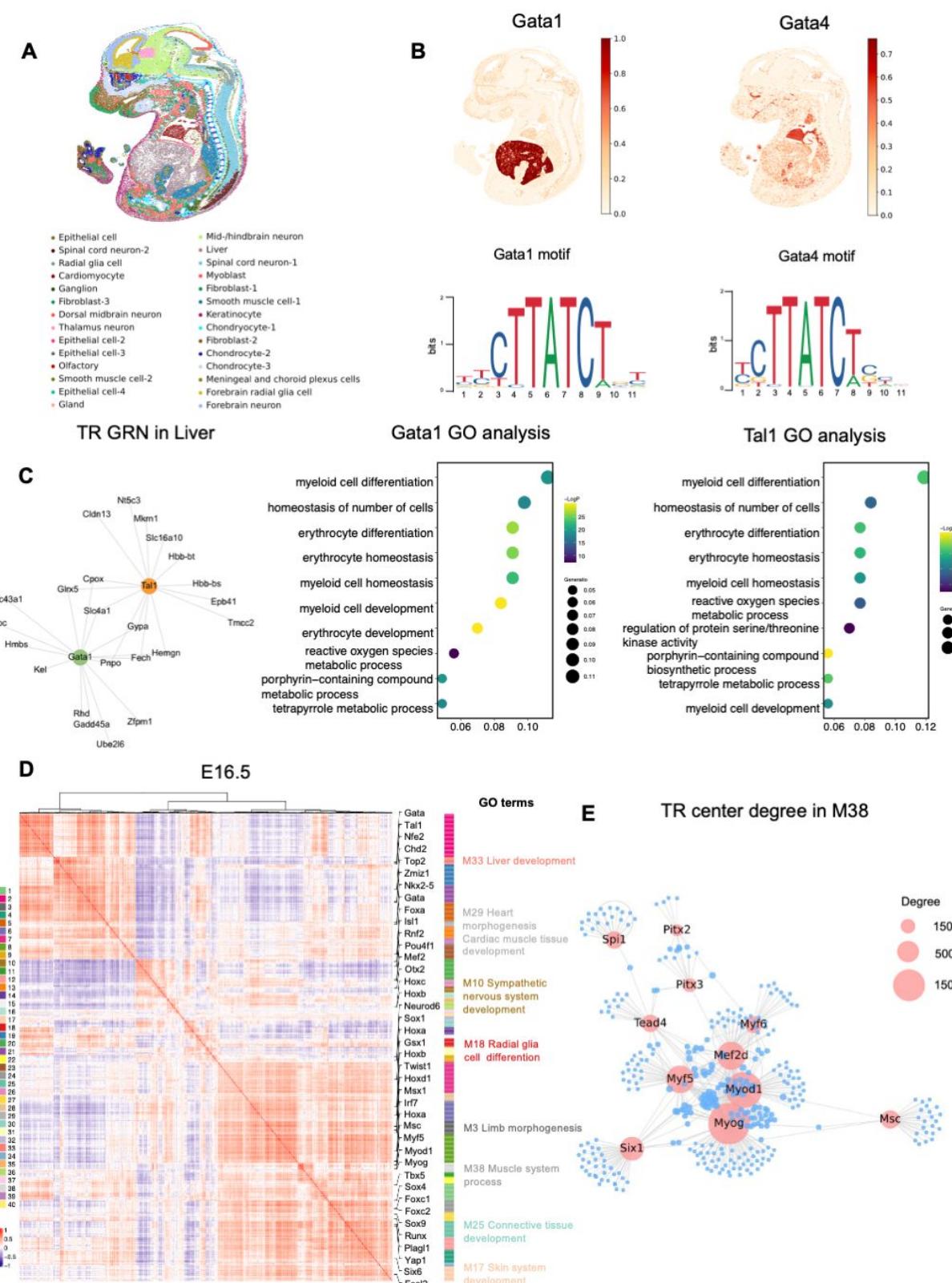
9 **3.5 Integration of spatial information enhances TR activity and**
10 **GRN prediction on E16.5 mouse embryo Stereo-seq data**

11 To showcase the superior performance of SCRIPPro in reconstructing
12 spatial GRNs by integrating spatial locations and neighborhood
13 information, we applied it to a Stereo-seq dataset derived from E16.5
14 mouse embryos(Chen, et al., 2022) (Fig. 4A). Our spatial clustering
15 strategy successfully identified 28 cell types with unique spatial
16 locations (Fig.4A, Supplementary Fig. S9A), and exhibiting denser cell
17 positioning and higher within-cell type homogeneity compared to other
18 spatial-based methods, including BayesSpace(Zhao, et al., 2021),
19 Giotto(Dries, et al., 2021), and the raw annotations from Stereo-seq
20 (using Squidpy) (Chen, et al., 2022)(Supplementary Fig. S9B). For
21 instance, in the heart region, the BayesSpace clustering mixed different
22 heart cell types, while Giotto, the Stereo-seq raw annotations, and
23 SCRIPPro all revealed distinct spatial patterns corresponding to the
24 various heart cell populations (Supplementary Fig. S9C). However, in
25 the brain and spinal cord regions, Giotto failed to differentiate thalamus
26 neurons, and the raw annotations could not separate the spinal cord from
27 the mid/hindbrain. In contrast, the spatial domains identified by
28 SCRIPPro delineated these anatomical sub-regions (Supplementary Fig.
29 S9D, indicated by black arrows). Furthermore, the ROGUE score
30 analyses indicated that the spatial domains defined by SCRIPPro were
31 more homogeneous compared to the other methods (Fig. 9E).We next
32 quantified the TR activity in different cell types and re-clustered the
33 cells by TR activity. SCRIPPro demonstrated a significantly higher
34 consistency with the cell types generated by spatial clustering compared
35 to SCENIC (Supplementary Fig. S10A-B). For instance, we examined
36 the TRs Foxo1 and Foxo3, known to be enriched in the brain and facial
37 regions including the striatum, anterior thalamic nucleus, olfaction, and
38 dental epithelium(Hoekman, et al., 2006). SCRIPPro robustly predicted
39 the TR activity in these regions, with Foxo3 exhibiting a dispersed
40 distribution compared to Foxo1(Hoekman, et al., 2006). In contrast,
41 SCENIC failed to enrich either factor in the brain and facial regions
42 (Supplementary Fig. S10B). Gata6 plays a critical role in cardiac
43 function with pronounced enrichment in the outflow tract and aortic
44 arch(Xin, et al., 2006), SCRIPPro distinctly identified the Gata6 activity
45 across different spatial subsets of the heart (Supplementary Fig. S10B).
46 Additionally, SCRIPPro exhibited a broader TR coverage and
47 successfully predicted the spatial TR activity for Neurod2 and Otx2 in
48 distinct brain regions, which were not covered by SCENIC
49

50 (Supplementary Fig. S10C). Finally, we benchmarked the performance
51 of SCRIPPro on homologous factors with similar motifs. SCRIPPro
52 accurately localized Gata4 in the heart region and Gata1 in the
53 developing liver region, despite the highly resemblant motifs of these
54 two factors (Fig. 4B). Similarly, SCRIPPro demonstrated the ability to
55 distinguish between Tcf12 and Tcf7, demonstrating the ability in
56 identifying cell type-specific binding patterns of homologous factors
57 through incorporating ChIP-seq data (Supplementary Fig. S10D).
58 Collectively, these results suggest SCRIPPro could accurately predict cell
59 type and region-specific GRNs than existing tools on the spatial
60 transcriptomic-only dataset, with the ability to distinguish TRs with
similar motif patterns.

61 TRs and their associated cofactors collaborate to modulate downstream
62 genes, thereby establishing GRNs instrumental in determining cell
63 phenotypes. To evaluate the effectiveness of SCRIPPro in identifying TR
64 regulons, we focused on factors enriched in the embryonic liver,
65 specifically Gata1(Papadopoulos, et al., 2013) and Tal1(Elefanty, et al.,
66 1999). We conducted a screening of their target genes and pruned the
67 network based on TR-target co-expression, resulting in the generation of
68 co-regulation regulons (Fig. 4C). Functional analyses revealed that both
69 TRs are linked to myeloid cell and erythrocyte differentiation and
70 homeostasis, indicating a potential co-binding of these two TRs in
71 regulating hematopoietic function in the embryonic liver (Fig. 4C).
72 Similarly, by iteratively clustering the spatially variable TRs and regulon
73 co-expression patterns, we successfully identified 40 co-regulated gene
74 modules in the mouse E16.5 embryo (Fig. 4D and Supplementary Fig.
75 S11A-B). These modules show highly specific spatial patterns, including
76 Pax5 for forebrain (M1), Neurog2 for mid/hindbrain (M10), Myog for
77 muscle cells (M38), Msx1 in maxillary and limb mesenchymal
78 cells(Jumlongras, et al., 2001) (M3), and Hoxc8 in the mouse embryonic
79 spine(Blackburn, et al., 2009) (M36) (Fig. 4D and Supplementary Fig.
80 S11B). Within each module, SCRIPPro also constructed TR-centered
81 GRNs to identify crucial factors. For example, Myod1, Myog, Mef2d,
82 and Myf5 were identified as key nodes in the M38 muscle module,
83 aligning well with their known roles as muscle master regulators (Fig. 4E).
84 In summary, these analyses demonstrate the ability of SCRIPPro to
85 accurately identify TR target genes and construct cell type-specific GRNs,
86 thereby enabling the identification of potential novel master regulators for
87 each lineage.

Article short title

**Fig. 4** SCRIPPro detected cell type specific GRNs in E16.5 mouse embryo Stereo-seq data.

- A. SCRIPPro identified 28 cell types on E16.5 mouse embryo stereo-seq dataset.
- B. SCRIPPro can distinguish different TRs spatial distribution with similar motif in same family.
- C. SCRIPPro is capable of predicting Gata1 and Tal1 target genes and build GRNs, and utilizes these target genes for GO analysis.
- D. Heatmap showing the modules with significant spatial autocorrelation that are clustered into different modules based on spatial co-expression of E16.5 embryo.
- E. Calculate TR center degree in module 38 (muscle) and find out important TRs.

Article short title

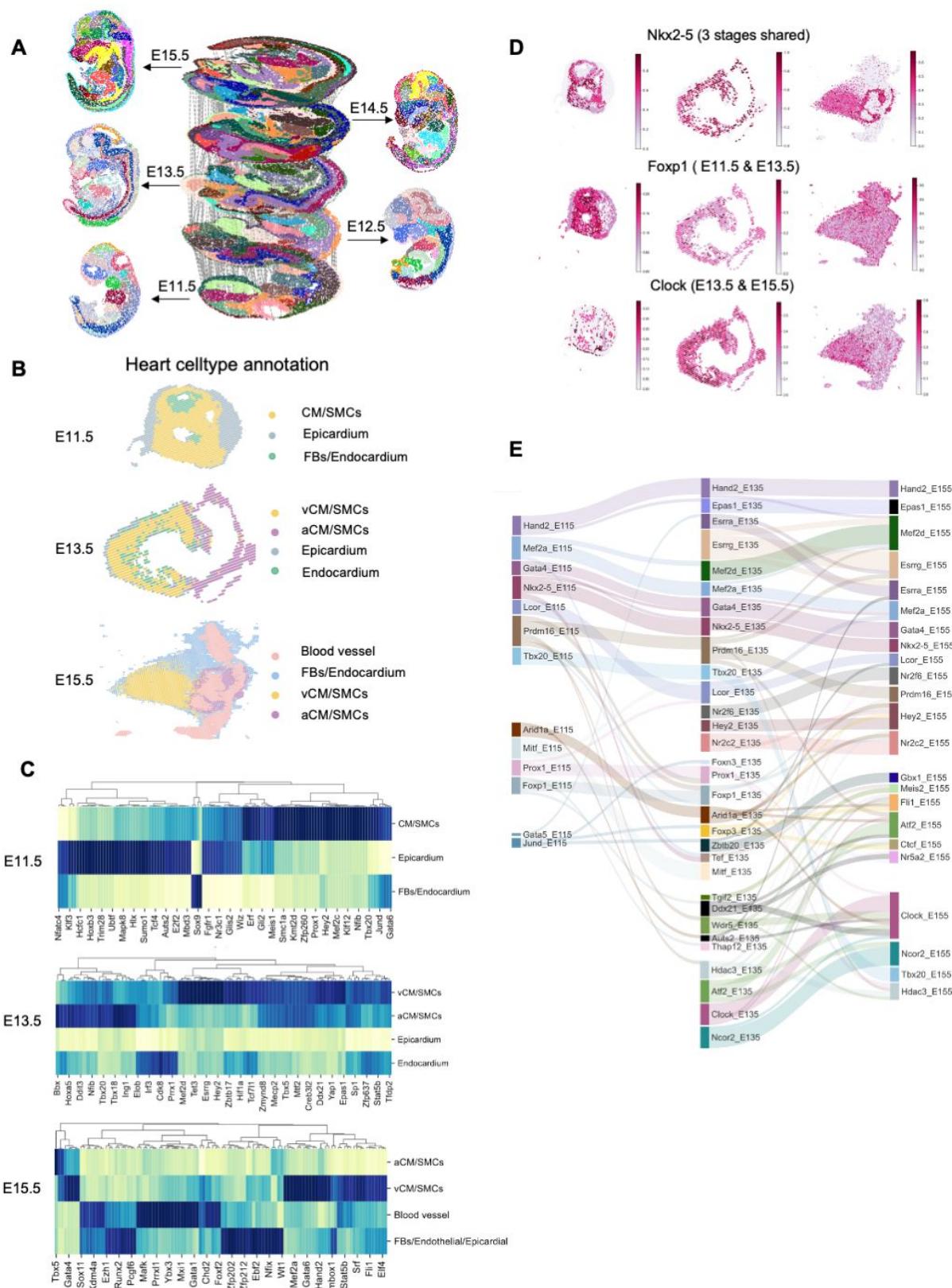
3.6 SCRIPro detects stage-specific GRNs in cardiomyocytes across mouse embryonic heart sections

Our previous analyses demonstrate that SCRIPro could identify lineage- and stage-specific GRNs in the hair follicle development dataset, we next evaluated whether it is suitable for analyzing time-series spatial datasets. We applied SCRIPro on mouse embryo stereo-seq data(Chen, et al., 2022) spanning from E11.5 to E15.5 across five continuous stages (Fig. 5), and aligned these spatial slides using SLAT(Chen-Rui, et al., 2023) (Fig. 5A). Our focus was on heart development, given its early formation in mammalian embryos and the diverse changes observed in cardiomyocytes from E11.5 to E15.5 (Fig. 5B). We categorized cells in the heart region into different celltypes from E11.5 to E15.5 (Supplementary Fig. S12A-B). Encouragingly, SCRIPro identifies highly specific TR activity among different cell types starting from E11.5, including Gata6, Jund, and Mef2c for cardiomyocytes, Tcf4 and Hoxb3 for epicardium, and Sox9 for fibroblasts (Fig. 5C). To further elucidate the dynamics of TRs and their regulons during cardiomyocyte development, we constructed a cross-stage GRN based on enriched TRs in cardiomyocytes (Supplementary Fig. S13B). Our analyses identified stage-shared regulators including Nkx2-5, Hand2, Gata4, and Mef2d. Most of these factors are supported by existing literatures(Jumlongras, et al., 2001) (Blackburn, et al., 2009)(Fig. 5D and Supplementary Fig. S13A). We applied the same analysis to single-cell mouse embryo hearts across the same developmental stages. Focusing on cardiomyocytes (CM) as an example, we found that the inferred transcriptional regulators (TRs) at different stages were generally consistent between the scRNA-seq and spatial transcriptomics (ST) data (Supplementary Fig. S12C-D). For instance, the TR Nkx2-5 was identified as important across all three stage. Additionally, stage-specific TRs, such as Foxp1 in E11.5-E13.5 CM and Clock in E13.5-E15.5 CM, as well as the cell-type-specific TR Tbx5 in

atrial cardiomyocytes (Supplementary Fig. S12D, S13A), exhibited similar dynamics in both scRNA-seq and ST data. However, we also observed differences in the inferred TRs between the two data modalities. The scRNA-seq-specific TRs at E11.5, including Hdac2, Dicer1, and Kdm2b, were primarily associated with DNA repair, epigenetic regulation, and metabolic functions, and were often nucleus-enriched. In contrast, the ST-specific TRs across the three stages, such as Stat3, Smad4, Tgif1, and Foxp3, were involved in intercellular signaling and immune-related functions (Supplementary Fig. S12D).

In addition to TR activity, SCRIPro robustly identified the TR regulons, for which can be used to evaluate the stable or dynamic regulation of different TRs (Fig. 5E). Most TRs showed conserved regulation of its regulon among different stages. However, Prdm16, a regulator for brown adipocyte differentiation(Harms, et al., 2014), shows remarkable dynamics in its regulon(Cibi, et al., 2020; Wu, et al., 2022) (Fig. 5E). We performed functional analyses of different subsets of the Prdm16 regulon and found that the constant Prdm16 regulon was enriched in fatty acid beta-oxidation, consistent with its well-known regulatory functions (Supplementary Fig. S13C). Interestingly, the genes lost at E11.5 were highly enriched in glycogen metabolism, while the unique genes at E13.5 were specifically enriched in NADP metabolism (Supplementary Fig. S13C). These analyses suggest that Prdm16 may play a critical role in the metabolic reprogramming from glycolysis to oxidative phosphorylation in cardiomyocytes, which has been reported to be connected with the proliferation ability of cardiomyocytes(Li, et al., 2023; Puente, et al., 2014). In summary, SCRIPro effectively tracks and analyzes the dynamics of TRs as well as their regulons in mouse embryonic heart development, enabling future identification of novel lineage regulators from spatial transcriptomics-only datasets.

Article short title

**Fig. 5** SCRIPPro identified stage-specific GRNs in consecutive mouse embryonic heart sections.

- Alignment of embryonic sections from five developmental stages, E11.5, E12.5, E13.5, E14.5, and E15.5, based on clusters predicted by SCRIPPro.
- Annotation of embryonic hearts at three developmental stages: E11.5, E13.5, and E15.5.
- Highly specific TR activity among different cell types starting from E11.5 to E15.5.
- TR spatial distribution in 3 stages. Nkx2-5: 3 stages shared. Foxp1: E11.5 and E13.5 shared. Clock: E13.5 and E15.5 shared.
- The stage specific GRNs correlation across three developmental stages.

Article short title

1
2
3
4
5 **3.7 Spatial multi-omic prediction of GRNs reveals crosstalk**
6 **between intra-cellular gene regulation and extra-cellular**
7 **interactions in the P22 mouse brain**

8
9
10 Finally, we applied SCRIPPro to a spatial ATAC-RNA-seq dataset from
11 the mouse brain on postnatal day 22, which provided paired expression
12 and chromatin accessibility data along with spatial location. SCRIPPro
13 successfully identified 10 distinct spatial domains that corresponded
14 well with specific anatomical structures (Fig. 6A). Notably, domains 0,
15 1, and 3 represented Cortex (CT) regions associated with advanced
16 neural and emotional processing functions(Pessoa and Adolphs, 2010).
17 Domain 6 corresponded to the Corpus Callosum (CC) region, crucial
18 for interhemispheric communication, while domain 9 aligned with the
19 Lateral Ventricle (LV) region, known for neural stem cell origins.
20 Pseudo-time analysis revealed an inward-to-outward neural stem cell
21 emergence pattern consistent with the differentiation trajectories of
22 neuron cells (Supplementary Fig. S14A). SCRIPPro accurately predicted
23 spatial variable TRs for each region. For instance, it identified Sox6
24 and Sox10 as significant TRs in the CC region, and Sox2, Sox4, and
25 Sox11 in the LV regions, highlighting the pivotal role of the SOX
26 family in neurogenesis(Stevanovic, et al., 2021) (Fig. 6B). Additionally,
27 Bcl11b and Foxp1(Tamura, et al., 2004) exhibit strong TR activity
28 within the striatum, where Bcl11b plays a critical role in the
29 differentiation of medium spiny neurons important to motor
30 control(Arlotta, et al., 2008) (Fig. 6B). Moreover, Mef2c(Barbosa, et
31 al., 2008), Neurod2, and Neurod6(Bormuth, et al., 2013; Lin, et al.,
32 2004) have been previously demonstrated to regulate gene expression
33 in the CT region (Fig. 6C and Supplementary Fig. S14B). These results
34 validate the accuracy of SCRIPPro in predicting region-specific TRs
35 using sparse spatial multiomic datasets.

36
37 To further assess SCRIPPro's accuracy in predicting TR regulons from
38 spatial multiomic data, we constructed GRNs for Sox2 and Mef2c,
39 which were enriched in LV and CT regions, respectively (Fig. 6D). The
40 targets and functions of these two factors exhibited distinct
41 characteristics. The Sox2 regulon was notably enriched in the WNT
42 signaling pathway and neural tube/epithelial tube
43 development(Mercurio, et al., 2022), while the MEF2C regulon
44 predominantly converged on signal transduction, synapse organization,
45 and hindbrain development(Harrington, et al., 2020). These analyses
46 emphasize the specific functions of these TRs and demonstrate the
47 accuracy of SCRIPPro in identifying TR-specific regulons. In addition
48 to intrinsic gene regulation, cellular crosstalk could also regulate cell
49 type-specific TR expression. We then investigated whether spatial
50 GRN analyses could be used to identify extracellular regulations on
51 TRs such as cell-cell interactions. Specifically, we focused on the TRs
52 with documented ligand-receptor pairs (see Methods). For example, we
53 observed a high enrichment of NOTCH1 activity in the LV region (D9),
54 which is known to harbor neural stem cells. Interestingly, its upstream
55 regulator Dlk1/2(Sánchez-Solana, et al., 2011) and Jagged1(Fissel and
56 Farah, 2021) were found to have closer associations with the NOTCH1
57 regulon in the LV region compared to the CC region (D6) and CT
58 region (D0) (Fig. 6E). Similar analyses using FGFR1 expression did
59 not yield significant differences (Supplementary Fig. S14C). In
60 summary, SCRIPPro effectively utilizes spatial multiomic data to
construct detailed GRNs in the mouse brain, revealing distinct TR
specificity across different spatial regions and facilitating the
exploration of extracellular regulations influencing TR expression in
different regions.

Article short title

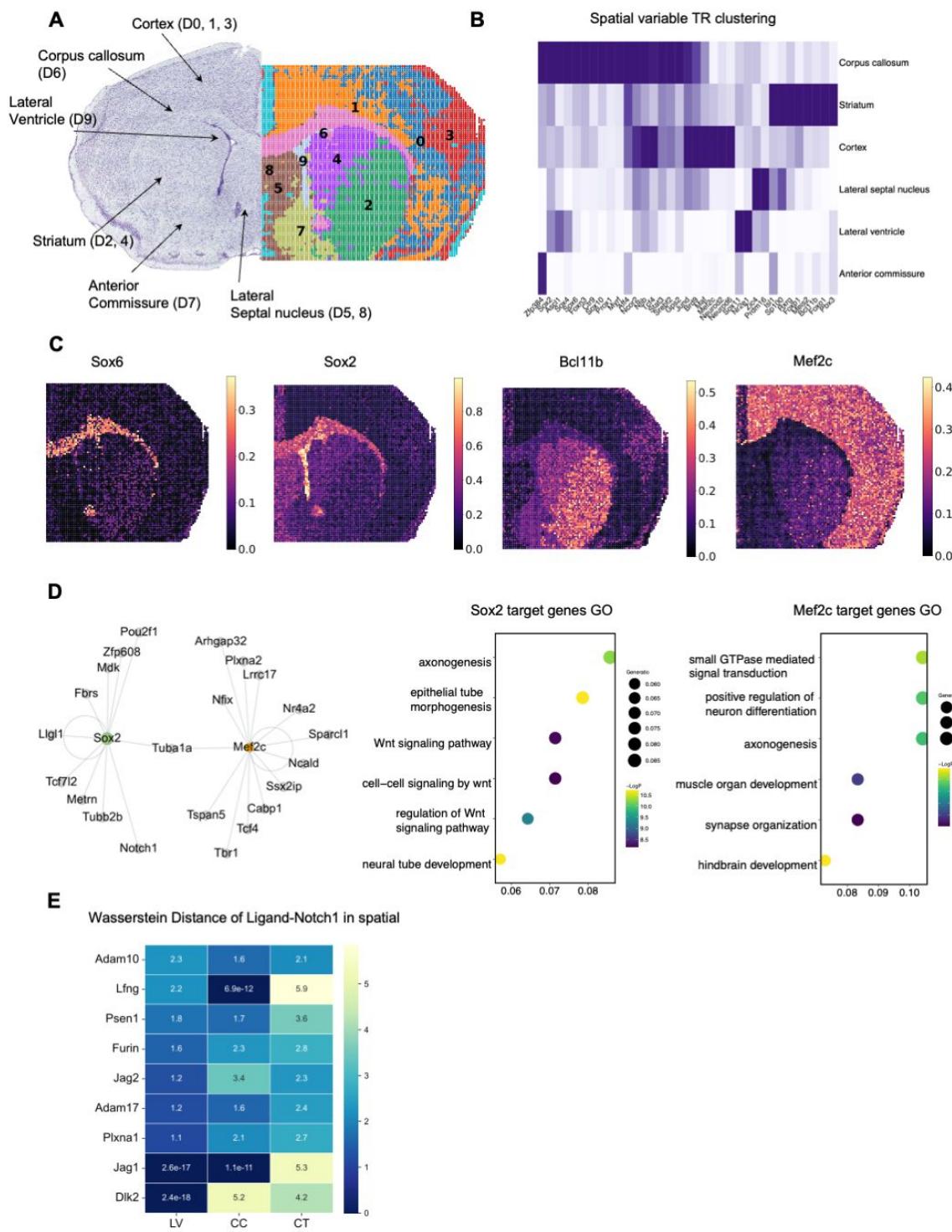


Fig. 6 Spatial multi-omic prediction of GRNs reveals crosstalk between intra-cellular gene regulation and extra-cellular interactions in the P22 mouse brain.

- SCRIPPro identified 10 clusters on mouse brain spatial ATAC-RNA-seq (RNA).
- Heatmap of spatial variable TR clustering by 6 cell type regions.
- SCRIPPro activity score of selected marker TR in spatial.
- SCRIPPro predicts Sox2 and Mef2c target genes then builds GRNs, and utilizes these target genes for GO analysis.
- Wasserstein distance of Ligand and Notch1 receptor in different regions.

Article short title

4 Discussion

Constructing high-resolution GRNs from extensive single-cell and spatial transcriptomics data is crucial for understanding gene regulation mechanisms in cell fate determination and disease development. In this study, we developed SCRIPro, a rapid and user-friendly method that accurately predicts GRNs. SCRIPro addresses the challenge of sparse signals at the SuperCell(Littman, et al., 2023) resolution by accounting for both expression and spatial similarity. Notably, SCRIPro includes a chromatin reconstruction step designed for scRNA-seq and spatial transcriptomics datasets without paired ATAC-seq data, significantly enhancing its usability. Through in-silico deletion TR analyses using a comprehensive TR ChIP-seq reference, SCRIPro outperforms existing motif-based methods in various systems, including human B-cell lymphoma, mouse hair follicle development, mouse developing embryos, and brain at P22. We anticipate that SCRIPro will be widely utilized by researchers to identify crucial TRs underlying novel differentiation and disease mechanisms.

SCRIPro integrates rich information from external ChIP-seq datasets, addressing a significant limitation in most existing GRN inference methods for single cells, such as SCENIC(Aibar, et al., 2017), SCING(Littman, et al., 2023), chromVAR(Schep, et al., 2017), SCENIC+(Bravo González-Blas, et al., 2023), Dictys(Wang, et al., 2023), and CellOracle(Kamimoto, et al., 2023), which rely heavily on motif information. Using only motif information can lead to a loss of cell-type-specific secondary TF binding in about 45% of cases, as estimated from ENCODE data. By including external ChIP-seq datasets, SCRIPro significantly enhances GRN inference performance, similar to recent methods using pre-trained neural networks from ENCODE data(2012). Moreover, SCRIPro is optimized for single-cell datasets. Unlike the LISA method, which was designed for bulk RNA-seq data and shows poor performance when applied directly to single-cell multi-ome data, SCRIPro uses the SuperCell strategy to overcome data sparsity, prunes targets using TR-TG expression correlation, and improves reference quality with stringent QC filters. These optimizations result in significantly improved performance compared to the existing LISA method. SCRIPro also offers flexibility with variable input data. For transcriptomic-only data, it uses strategies similar to LISA to reconstruct chromatin landscapes from bulk DNase/H3K27ac references. For epigenomic-only data, it employs SCRIP(Dong, et al., 2022) to infer potential regulators. For multi-omics data, SCRIPro aligns and uses matched or paired epigenome datasets as the chromatin landscape. Additionally, SCRIPro considers spatial adjacency when generating SuperCells, providing a robust and flexible solution for GRN inference(Yuan and Duren, 2024) across different single-cell modalities.

Despite its current superiority over existing methods, SCRIPro has some limitations. Its performance relies heavily on the quality of public ChIP-seq datasets, which can affect its robustness. To address this, SCRIPro integrates motif scanning results into the TR reference to mitigate the reduction in TR reference coverage after filtering out low-quality datasets. Furthermore, the rapid development of single-cell/spatial epigenomics technologies such as sciATAC-

seq3(Domcke, et al., 2020), sciMAP-ATAC(Thornton, et al., 2021), and some multiome epigenomics technologies such as Paired-Tag(Zhu, et al., 2021), scCUT&Tag-pro(Zhang, et al., 2022) and DOGMA-seq(Mimitou, et al., 2021), accelerates the accumulation of high-quality epigenome data in the field. SCRIPro aims to incorporate these data into the chromatin and TR reference, expanding coverage to more cell types and improving performance, particularly on rare cell types. While SCRIPro adeptly predicts TRs at a SuperCell resolution, its precision at the single-cell level is still being established due to challenges related to background noise and data quality. To overcome these limitations, we are exploring the use of machine learning models, such as diffusion and autoencoders, to improve the quality of single-cell and spatial-omics data, thereby enhancing both the resolution and accuracy of TR prediction. Finally, the cellular spatial locations from spatial-omics data are essential for understanding in situ gene expression regulation, cell interactions, and signal transduction within spatial microenvironments. Currently, SCRIPro calculates upstream signaling pathways by constructing spatial GRNs and filtering ligand-receptor (L-R) pair expressions. However, this approach is limited by the coverage of spatial-omics data and the number of L-R pairs in the database. Future development integrating GRNs, protein-protein interactions (PPI), and L-R co-occurrence is expected to significantly increase the connections between intracellular GRNs and extracellular cell-cell interactions (CCIs).

In summary, SCRIPro is a promising tool that enables researchers to leverage extensive single-cell or spatial transcriptomics data to identify driver TR regulations, with or without paired epigenome information. It facilitates the interpretation of GRNs across diverse cell types, cellular trajectories, and spatial domains.

Acknowledgement

The authors thank the Bioinformatics Supercomputer Center of Tongji University for offering computing resource

Funding

This work was supported by the National Key R&D Program of China [2022YFA1106000, 2020YFA0113200], the National Natural Science Foundation of China [32222026, 32170660, 92168205, 62088101]. Shanghai Municipal Science and Technology Major Project [2021SHZDZX0100]. Shanghai Pilot Program for Basic Research. The Fundamental Research Funds for the Central Universities [20002150110, 22120230292].

Conflict of Interest: none declared.

Author information

Zhanhe Chang, Yunfan Xu and Xin Dong contributed equally to this work.

Authors and Affiliations

Key Laboratory of Spine and Spinal Cord Injury Repair and Regeneration of Ministry of Education, Department of Orthopedics, Tongji Hospital, School of Life Science and Technology, Tongji University, China.

Yunfan Xu, Xin Dong & Chenfei Wang

Article short title

Institute for Regenerative Medicine, Shanghai East Hospital, Shanghai Key Laboratory of Signaling and Disease Research, School of Life Sciences and Technology, Tongji University, Shanghai, China.
Zhanhe Chang, Yunfan Xu, Xin Dong & Yawei Gao

Frontier Science Center for Stem Cell Research, Tongji University, Shanghai, China.
Zhanhe Chang, Yawei Gao, Chenfei Wang

Contributions

C.W. conceived and supervised the whole project. Y.X. and Z.C. designed and implemented the SCRIPro algorithm. X.D. collected and preprocessed the ChIP-seq and motif datasets. Z.C. evaluated the performance on mouse embryo stereo-seq datasets, mouse embryo single cell datasets and mouse brain spatial ATAC-RNA-seq datasets. Y.X. performed the analysis of scCRISPRa dataset, B-cell lymphoma 10X Genomics multiomic datasets and the hair follicle single-cell multiomic datasets. Z.C., Y.X., X.D., Y.G. and C.W. wrote the manuscript with the help of other authors. All authors read and approved the final manuscript.

References

- An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;489(7414):57-74.
- Aibar, S., et al. SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* 2017;14(11):1083-1086.
- Alanis-Lobato, G., et al. MICA: a multi-omics method to predict gene regulatory networks in early human embryos. *Life Sci Alliance* 2024;7(1).
- Arlotta, P., et al. Ctip2 controls the differentiation of medium spiny neurons and the establishment of the cellular architecture of the striatum. *J Neurosci* 2008;28(3):622-632.
- Badia, I.M.P., et al. Gene regulatory network inference in the era of single-cell multi-omics. *Nat Rev Genet* 2023.
- Barbosa, A.C., et al. MEF2C, a transcription factor that facilitates learning and memory by negative regulation of synapse numbers and function. *Proc Natl Acad Sci U S A* 2008;105(27):9391-9396.
- Blackburn, J., et al. Generation of conditional Hoxc8 loss-of-function and Hoxc8->Hoxc9 replacement alleles in mice. *Genesis* 2009;47(10):680-687.
- Bormuth, I., et al. Neuronal basic helix-loop-helix proteins Neurod2/6 regulate cortical commissure formation before midline interactions. *J Neurosci* 2013;33(2):641-651.
- Bravo González-Blas, C., et al. SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nat Methods* 2023;20(9):1355-1367.
- Browaeys, R., Saelens, W. and Saeys, Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nature Methods* 2020;17(2):159-162.
- Buenrostro, J.D., et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 2015;523(7561):486-490.
- Cao, Z.J. and Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat Biotechnol* 2022;40(10):1458-1466.
- Cassandri, M., et al. Zinc-finger proteins in health and disease. *Cell Death Discov* 2017;3:17071.
- Chan, T.E., Stumpf, M.P.H. and Babtie, A.C. Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. *Cell Syst* 2017;5(3):251-267.e253.
- Chen, A., et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* 2022;185(10):1777-1792.e1721.
- Chen, J.G., et al. Giotto Suite: a multi-scale and technology-agnostic spatial multi-omics analysis ecosystem. *bioRxiv* 2023.
- Chen-Rui, X., et al. Spatial-linked alignment tool (SLAT) for aligning heterogenous slices properly. *bioRxiv* 2023;2023.2004.2007.535976.
- Cibi, D.M., et al. Prdm16 Deficiency Leads to Age-Dependent Cardiac Hypertrophy, Adverse Remodeling, Mitochondrial Dysfunction, and Heart Failure. *Cell Rep* 2020;33(3):108288.
- Domcke, S., et al. A human cell atlas of fetal chromatin accessibility. *Science* 2020;370(6518).
- Dong, K. and Zhang, S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nature Communications* 2022;13(1):1739.
- Dong, K. and Zhang, S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nat Commun* 2022;13(1):1739.
- Dong, X., et al. Single-cell gene regulation network inference by large-scale data integration. *Nucleic Acids Res* 2022;50(21):e126.
- Dries, R., et al. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biology* 2021;22(1):78.
- Elefanti, A.G., et al. SCL expression in the mouse embryo detected with targeted lacZ reporter gene demonstrates its localization to hematopoietic, vascular, and neural tissues. *Blood* 1999;94(11):3754-3763.
- Fissel, J.A. and Farah, M.H. The influence of BACE1 on macrophage recruitment and activity in the injured peripheral nerve. *J Neuroinflammation* 2021;18(1):71.
- Flamary, R., et al. POT: Python optimal transport. *J. Mach. Learn. Res.* 2021;22(1):Article 78.
- Fornes, O., et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2020;48(D1):D87-d92.
- Garaud, S., et al. Tumor infiltrating B-cells signal functional humoral immune responses in breast cancer. *JCI Insight* 2019;5(18).
- Harms, M.J., et al. Prdm16 is required for the maintenance of brown adipocyte identity and function in adult mice. *Cell Metab* 2014;19(4):593-604.
- Harrington, A.J., et al. MEF2C Hypofunction in Neuronal and Neuroimmune Populations Produces MEF2C Haploinsufficiency Syndrome-like Behaviors in Mice. *Biol Psychiatry* 2020;88(6):488-499.
- Haury, A.C., et al. TIGRESS: Trustful Inference of Gene REgulation using Stability Selection. *BMC Syst Biol* 2012;6:145.
- Heinz, S., et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 2010;38(4):576-589.
- Hoekman, M.F., et al. Spatial and temporal expression of FoxO transcription factors in the developing and adult murine brain. *Gene Expr Patterns* 2006;6(2):134-140.
- Huynh-Thu, V.A., et al. Inferring regulatory networks from expression data using tree-based methods. *PLoS One* 2010;5(9).
- Imbeault, M., Helleboid, P.Y. and Trono, D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* 2017;543(7646):550-554.
- Jumlongras, D., et al. A nonsense mutation in MSX1 causes Witkop syndrome. *Am J Hum Genet* 2001;69(1):67-74.
- Kamal, A., et al. GRaNIE and GRaNPA: inference and evaluation of enhancer-mediated gene regulatory networks. *Mol Syst Biol* 2023;19(6):e11627.
- Kamimoto, K., et al. Dissecting cell identity via network inference and in silico gene perturbation. *Nature* 2023;614(7949):742-751.
- Kartha, V.K., et al. Functional inference of gene regulation using single-cell multi-omics. *Cell Genom* 2022;2(9).

Article short title

- Kaufman, C.K., et al. GATA-3: an unexpected regulator of cell lineage determination in skin. *Genes Dev* 2003;17(17):2108-2122.
- Kent, W.J., et al. The human genome browser at UCSC. *Genome Res* 2002;12(6):996-1006.
- Kim, S. ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients. *Commun Stat Appl Methods* 2015;22(6):665-674.
- Li, X., et al. Inhibition of fatty acid oxidation enables heart regeneration in adult mice. *Nature* 2023;622(7983):619-626.
- Lin, C.H., et al. Regulation of neuroD2 expression in mouse brain. *Dev Biol* 2004;265(1):234-245.
- Littman, R., et al. SCING: Inference of robust, interpretable gene regulatory networks from single cell and spatial transcriptomics. *iScience* 2023;26(7):107124.
- Lynch, A.W., et al. MIRA: joint regulatory modeling of multimodal expression and chromatin accessibility in single cells. *Nat Methods* 2022;19(9):1097-1108.
- Ma, S., et al. Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* 2020;183(4):1103-1116.e1120.
- Matsumoto, H., et al. SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics* 2017;33(15):2314-2321.
- Mercurio, S., et al. Deconstructing Sox2 Function in Brain Development and Disease. *Cells* 2022;11(10).
- Mimitou, E.P., et al. Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nat Biotechnol* 2021;39(10):1246-1258.
- Moerman, T., et al. GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics* 2019;35(12):2159-2161.
- Papadopoulos, G.L., et al. GATA-1 genome-wide occupancy associates with distinct epigenetic profiles in mouse fetal liver erythropoiesis. *Nucleic Acids Res* 2013;41(9):4938-4948.
- Papili Gao, N., et al. SINCERITIES: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics* 2018;34(2):258-266.
- Persad, S., et al. SEACells infers transcriptional and epigenomic cellular states from single-cell genomics data. *Nature Biotechnology* 2023.
- Pessoa, L. and Adolphs, R. Emotion processing and the amygdala: from a 'low road' to 'many roads' of evaluating biological significance. *Nat Rev Neurosci* 2010;11(11):773-783.
- Puente, B.N., et al. The oxygen-rich postnatal environment induces cardiomyocyte cell-cycle arrest through DNA damage response. *Cell* 2014;157(3):565-579.
- Qin, Q., et al. Lisa: inferring transcriptional regulators through integrative modeling of public chromatin accessibility and ChIP-seq data. *Genome Biol* 2020;21(1):32.
- Quinlan, A.R. and Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;26(6):841-842.
- Ren, H., et al. Identifying multicellular spatiotemporal organization of cells with SpaceFlow. *Nat Commun* 2022;13(1):4076.
- Rhee, J.W., et al. Pbx3 deficiency results in central hypoventilation. *Am J Pathol* 2004;165(4):1343-1350.
- Roberts, N.A., et al. Prdm1 Regulates Thymic Epithelial Function To Prevent Autoimmunity. *J Immunol* 2017;199(4):1250-1260.
- Robinson, J.T., et al. Integrative genomics viewer. *Nat Biotechnol* 2011;29(1):24-26.
- Ruyssinck, J., et al. NIMEFI: gene regulatory network inference using multiple ensemble feature importance algorithms. *PLoS One* 2014;9(3):e92709.
- Sánchez-Solana, B., et al. The EGF-like proteins DLK1 and DLK2 function as inhibitory non-canonical ligands of NOTCH1 receptor that modulate each other's activities. *Biochim Biophys Acta* 2011;1813(6):1153-1164.
- Schep, A.N., et al. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat Methods* 2017;14(10):975-978.
- Shi, H., et al. Spatial atlas of the mouse central nervous system at molecular resolution. *Nature* 2023.
- Sichien, D., et al. IRF8 Transcription Factor Controls Survival and Function of Terminally Differentiated Conventional and Plasmacytoid Dendritic Cells, Respectively. *Immunity* 2016;45(3):626-640.
- Steinmayer, M., et al. staggerer phenotype in retinoid-related orphan receptor alpha-deficient mice. *Proc Natl Acad Sci U S A* 1998;95(7):3960-3965.
- Stevanovic, M., et al. SOX Transcription Factors as Important Regulators of Neuronal and Glial Differentiation During Nervous System Development and Adult Neurogenesis. *Front Mol Neurosci* 2021;14:654031.
- Takagi, Y., et al. SPIB is a novel prognostic factor in diffuse large B-cell lymphoma that mediates apoptosis via the PI3K-AKT pathway. *Cancer Sci* 2016;107(9):1270-1280.
- Tamura, S., et al. Foxp1 gene expression in projection neurons of the mouse striatum. *Neuroscience* 2004;124(2):261-267.
- Thornton, C.A., et al. Spatially mapped single-cell chromatin accessibility. *Nat Commun* 2021;12(1):1274.
- Traag, V.A., Waltman, L. and van Eck, N.J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* 2019;9(1):5233.
- Wang, C., et al. Integrative analyses of single-cell transcriptome and regulome using MAESTRO. *Genome Biol* 2020;21(1):198.
- Wang, L., et al. Dictys: dynamic gene regulatory network dissects developmental continuum with single-cell multiomics. *Nat Methods* 2023;20(9):1368-1378.
- Wolf, F.A., Angerer, P. and Theis, F.J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* 2018;19(1):15.
- Wu, T., et al. PRDM16 Is a Compact Myocardium-Enriched Transcription Factor Required to Maintain Compact Myocardial Cardiomyocyte Identity in Left Ventricle. *Circulation* 2022;145(8):586-602.
- Xin, M., et al. A threshold of GATA4 and GATA6 expression is required for cardiovascular development. *Proc Natl Acad Sci U S A* 2006;103(30):11189-11194.
- Yuan, Q. and Duren, Z. Inferring gene regulatory networks from single-cell multiome data using atlas-scale external data. *Nature Biotechnology* 2024.
- Zhang, B., et al. Characterizing cellular heterogeneity in chromatin state with scCUT&Tag-pro. *Nat Biotechnol* 2022;40(8):1220-1230.
- Zhang, D., et al. Spatial epigenome-transcriptome co-profiling of mammalian tissues. *Nature* 2023;616(7955):113-122.
- Zhang, L., Zhang, J. and Nie, Q. DIRECT-NET: An efficient method to discover cis-regulatory elements and construct regulatory networks from single-cell multiomics data. *Sci Adv* 2022;8(22):eabl7393.
- Zhang, Y., et al. Lefl contributes to the differentiation of bulge stem cells by nuclear translocation and cross-talk with the Notch signaling pathway. *Int J Med Sci* 2013;10(6):738-746.
- Zhang, Z., et al. PRC2 complexes with JARID2, MTF2, and esPRC2p48 in ES cells to modulate ES cell pluripotency and somatic cell reprogramming. *Stem Cells* 2011;29(2):229-240.
- Zhao, E., et al. Spatial transcriptomics at subspot resolution with BayesSpace. *Nature Biotechnology* 2021;39(11):1375-1384.
- Zheng, R., et al. Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Res* 2019;47(D1):D729-D735.

1
2
3
*Article short title*4
5
Zhou, Y., et al. Metascape provides a biologist-oriented resource for the analysis
of systems-level datasets. *Nat Commun* 2019;10(1):1523.6
7
Zhu, C., et al. Joint profiling of histone modifications and transcriptome in single
cells from mouse brain. *Nat Methods* 2021;18(3):283-292.8
Supplementary Tables9
Supplementary Table 1 Datasets information used in this study.10
Supplementary Table 2 Summary of human TR reference datasets used in this study.11
Supplementary Table 3 Summary of mouse TR reference datasets used in this study.

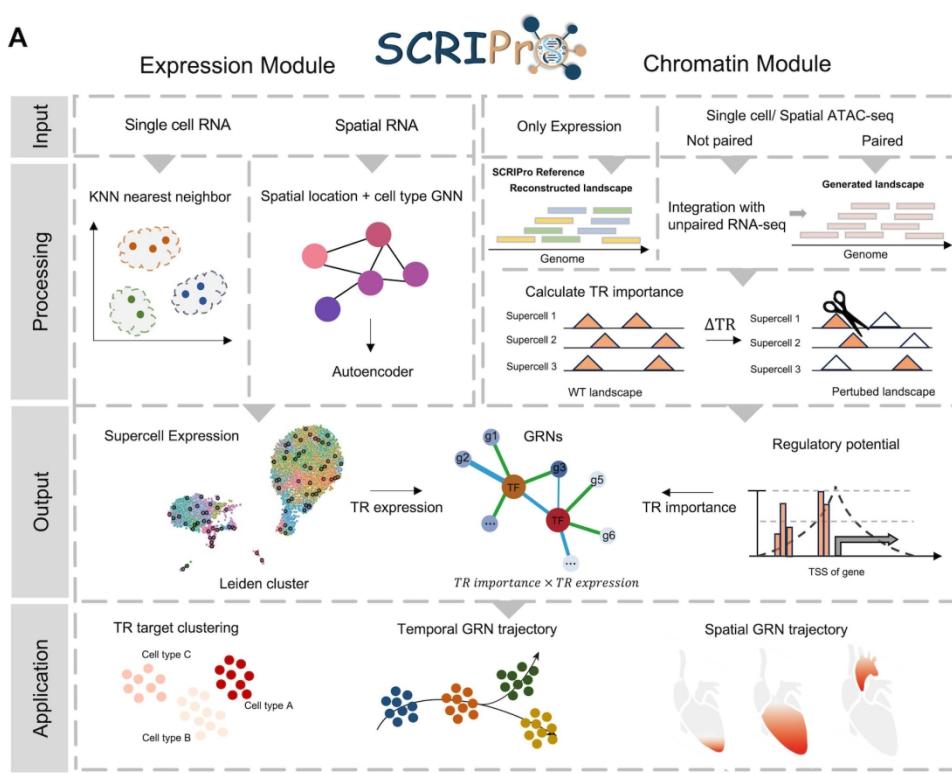


Figure 1

162x126mm (300 x 300 DPI)

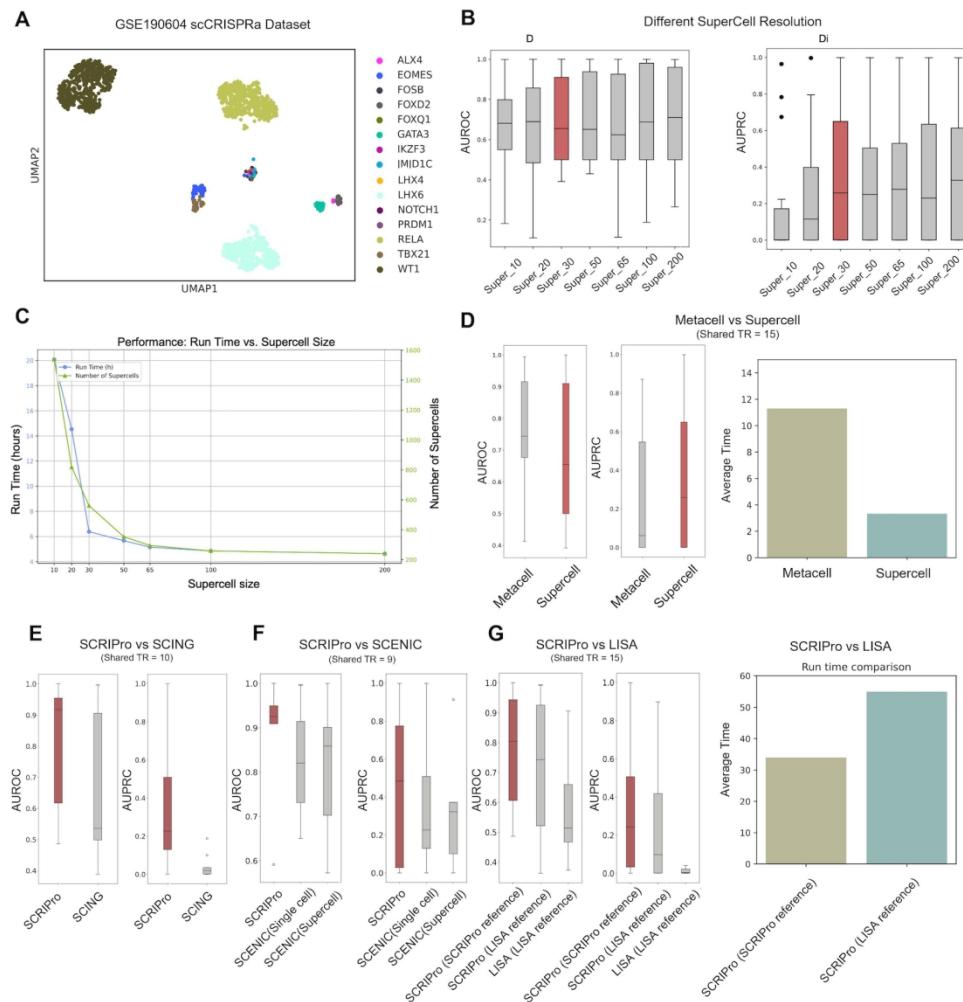


Figure 2

152x162mm (300 x 300 DPI)

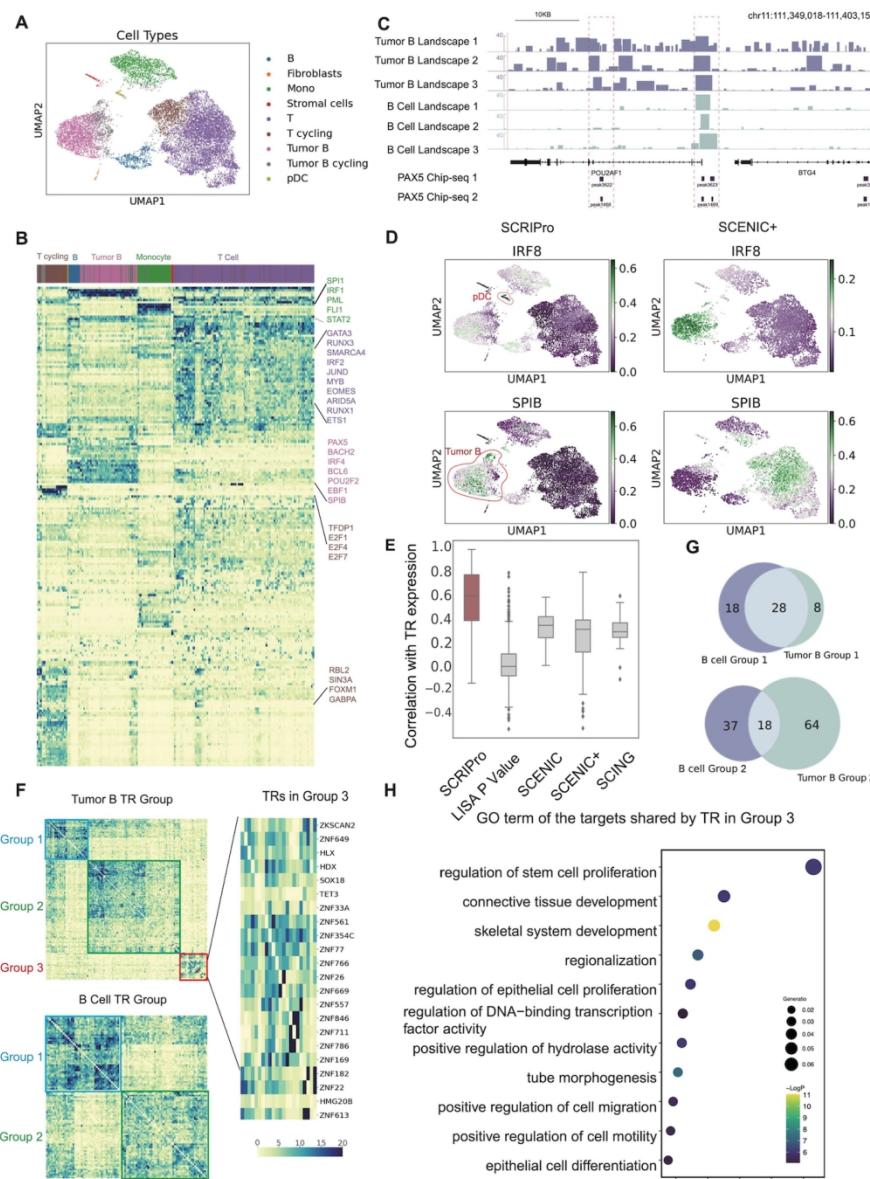


Figure 3

121x162mm (300 x 300 DPI)

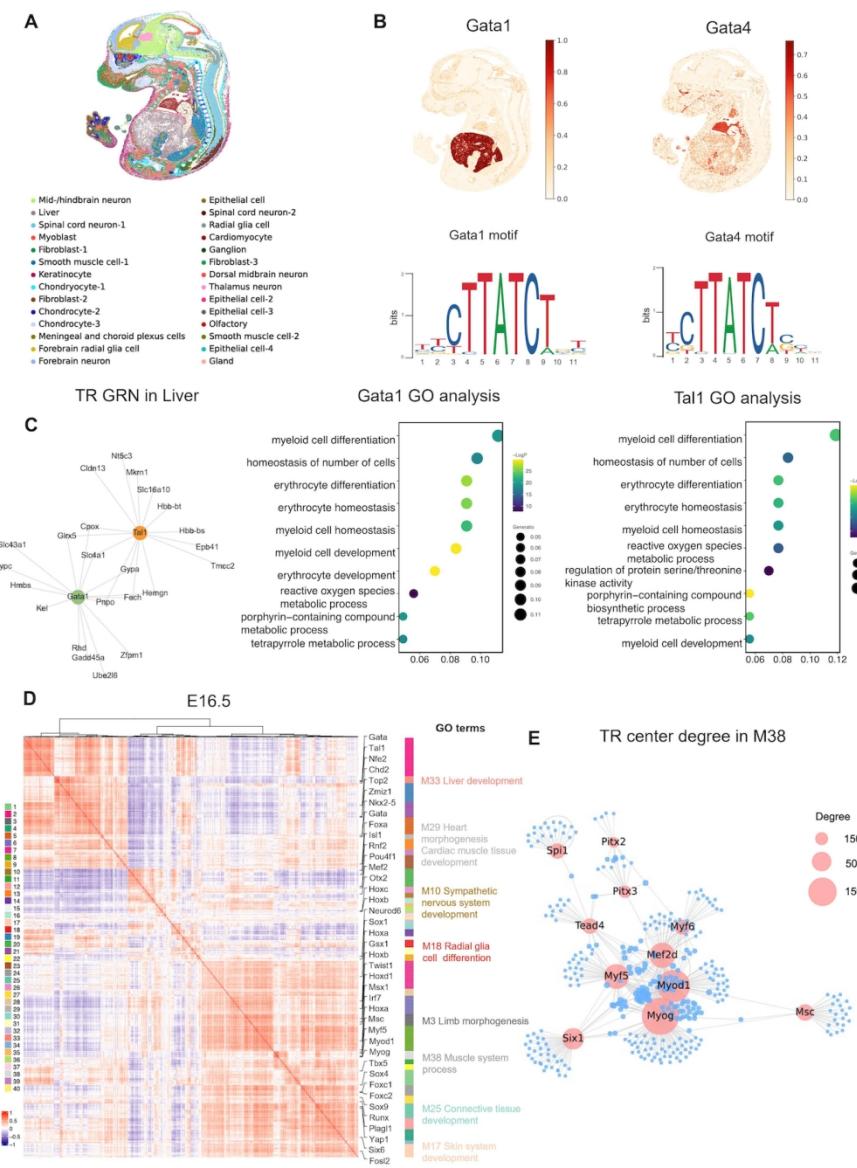


Figure 4

124x162mm (300 x 300 DPI)

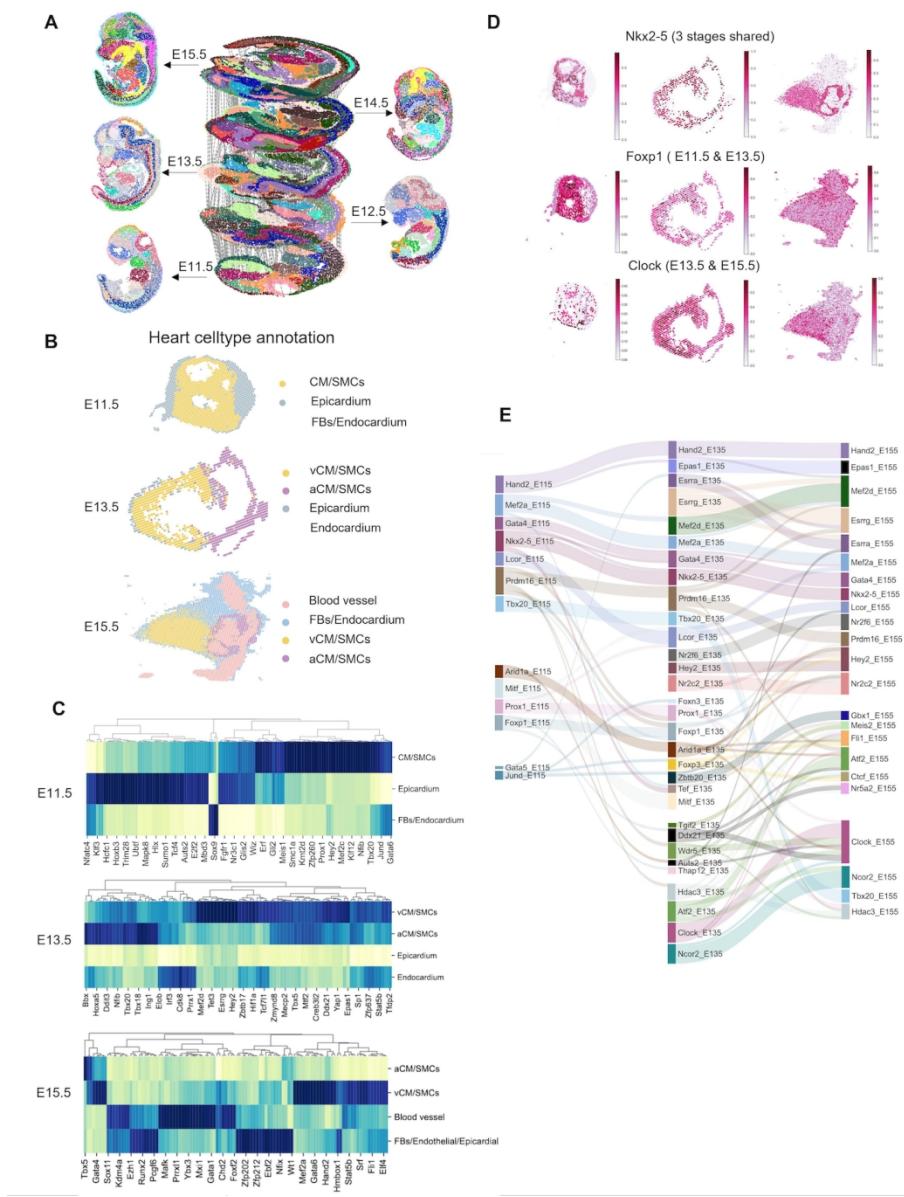


Figure 5

123x162mm (300 x 300 DPI)

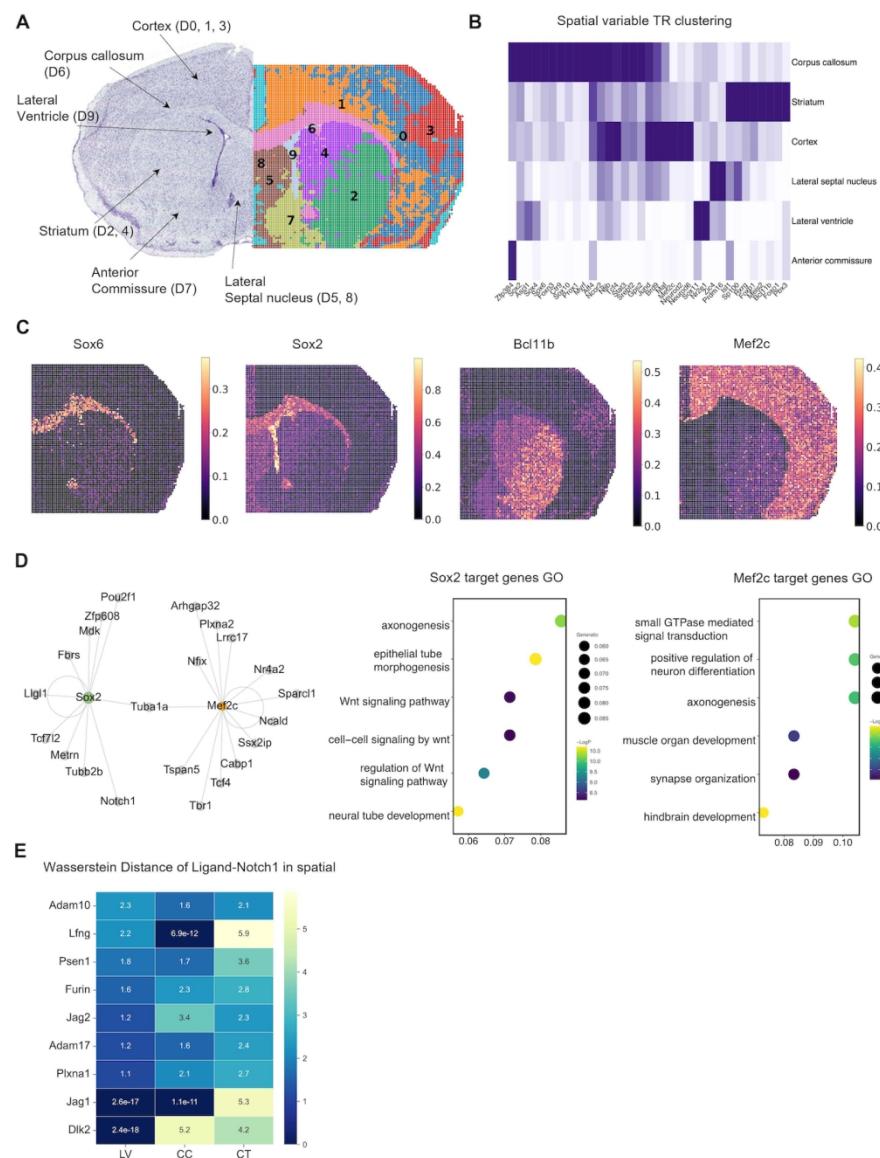


Figure 6

124x162mm (300 x 300 DPI)