Enhancer-driven gene regulatory networks inference from single-cell RNA-seq and ATAC-seq data

Yang Li[1], Anjun Ma[1,4], Yizhong Wang[2], Cankun Wang[1], Shuo Chen[3], Hongjun Fu[3], Bingqiang Liu[2,*], Qin Ma[1,4,*]

[1]Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH 43210, USA,
[2]School of Mathematics, Shandong University, Jinan 250100, China,
[3]Department of Neuroscience, College of Medicine, The Ohio State University, Columbus, OH 43210, USA
[4]Pelotonia Institute for Immuno-Oncology, The James Comprehensive Cancer Center, The Ohio State University, Columbus, OH 43210, USA.

ABSTRACT

We present an algorithm, STREAM, for enhancer-driven gene regulatory network (eGRN) inference from transcriptome and chromatin accessibility profiled in the same cells. The algorithm improves the prediction accuracy of relations among transcription factors (TFs), enhancers, and genes by two new ideas: (i) a Steiner forest problem model to identify the set of highly confident enhancer-gene relations underlying a context-specific functional gene module; and (ii) a hybrid biclustering pipeline integrated with submodular optimization for inferring eGRNs by identifying the optimal set of hybrid biclusters, each of which represents co-regulated genes by the same TF and co-accessible enhancers bound by the same TF over a cell subpopulation. These two ideas are embedded in an iterative framework by finding patterns in a pair of transcriptome and chromatin accessibility matrices. Benchmarking analysis shows that the performance, assessed by f-scores, precision, or recall, was significantly improved by our program compared to four state-of-the-art tools on nine datasets. Besides, by implementing STREAM on an Alzheimer's disease dataset, we identified TF-enhancer-gene relations associated with pseudotime and investigated the changing of enhancer-gene relations alongside cell lineages. Additionally, by implementing STREAM on a diffuse small lymphocytic lymphoma dataset, we excavated key TF-enhancer-gene relations and TF cooperation underlying tumor cells.

INTRODUCTION

Recent advances in single-cell sequencing technologies of RNA (scRNA-seq) and Assay for Transposase-Accessible Chromatin using sequencing (scATAC-seq) have provided plentiful data in exploring gene regulatory networks (GRNs) at single-cell resolution (1). Existing transcriptome-based methods construct GRNs by combining gene co-expression analysis with transcription factor (TF) motif discovery. However, the motif discovery requires searching for TF binding sites in a large sequence space around each co-expressed gene module, covering a small portion of a gene's putative regulatory space. Also, motif discovery is often based on the hypothesis that TF

binding sequences of the same gene are conserved across different cell types, which loses the heterogeneity of regulatory sequences and may result in high false positives of TF predictions in heterogeneous single-cell data. Fortunately, the integrative analysis of transcriptome and chromatin accessibility data can address the above issue. With the direct measurement of transcriptome and chromatin accessibility profiled in the same cells, the motif findings are no longer restricted to limited promoter regions, and the regulatory sequence heterogeneity can be retained. Hence, inferring GRNs via integrative analysis decreases the influence of noise due to sparse sequencing and raises the performance of TF-target predictions by cross-validating the regulatory relations through different modalities (1-8).

The increasing availability to profile transcriptome and chromatin accessibility has been breaking through the limitation of constructing GRNs and enables the inference of enhancer-driven GRNs (eGRNs) (3). An eGRN is a GRN in which the edges from TFs to target genes pass through its binding enhancers (3). Specifically, accessible regions detected by scATAC-seq often contain enhancers for TF binding and gene regulation, allowing us to explore changes in TF-enhancer and enhancer-gene relations in eGRNs (9,10). Intuitively, the elucidation of eGRNs enables the identification of cell-type-specific or constant enhancers, variation of TFs binding enhancers, and changes in enhancer accessibility. Janssens *et al.*, put forward the concept of eGRN and built a massive resource of eGRNs for 40 cell types in the fly brain through a bioinformatic framework equipped with motif discovery, network prediction, and deep learning models (3). First, this framework applied i-cisTarget or RcisTarget to identify enriched TF motifs against differentially accessible regions (DARs) and then filter the TF-DAR pairs by retaining the TFs expressed in the same cell type as DARs (11). Second, enhancer-gene relations were predicted using Pearson correlation coefficients and GENIE3 for matched cell types between transcriptome and chromatin accessibility (12). This framework is the only available tool specifically for eGRN prediction. Its accuracy of inferred TF-enhancer-gene relations will be influenced by cell clustering, cell cluster alignment, and cell type annotation in transcriptome and chromatin accessibility.

Currently, no approach has been developed for directly inferring eGRNs free from cell clustering of scRNA-seq and scATAC-seq data. eGRN inference raised three challenges for method development. First, existing methods for enhancer-gene relation prediction are based on the profile correlation between enhancer accessibility and gene expression. These approaches suffer from a high false positive rate in complex situations, where a group of enhancer-enhancer and enhancer-gene relations coordinately participate in gene regulation. Second, current approaches identify co-expressed or co-regulated genes and linked enhancers in a cell type based on prior cell clustering in a sequential manner. Biases that occur in the previous clustering step will impact the accuracy in the subsequent prediction of TF-enhancer-gene relations. Finally, numerous combinations of TF-enhancer-gene relations govern the state and behavior of cells. This regulatory heterogeneity complicates efforts to globally extract the combination of key TF-enhancer-gene relations across all cells.

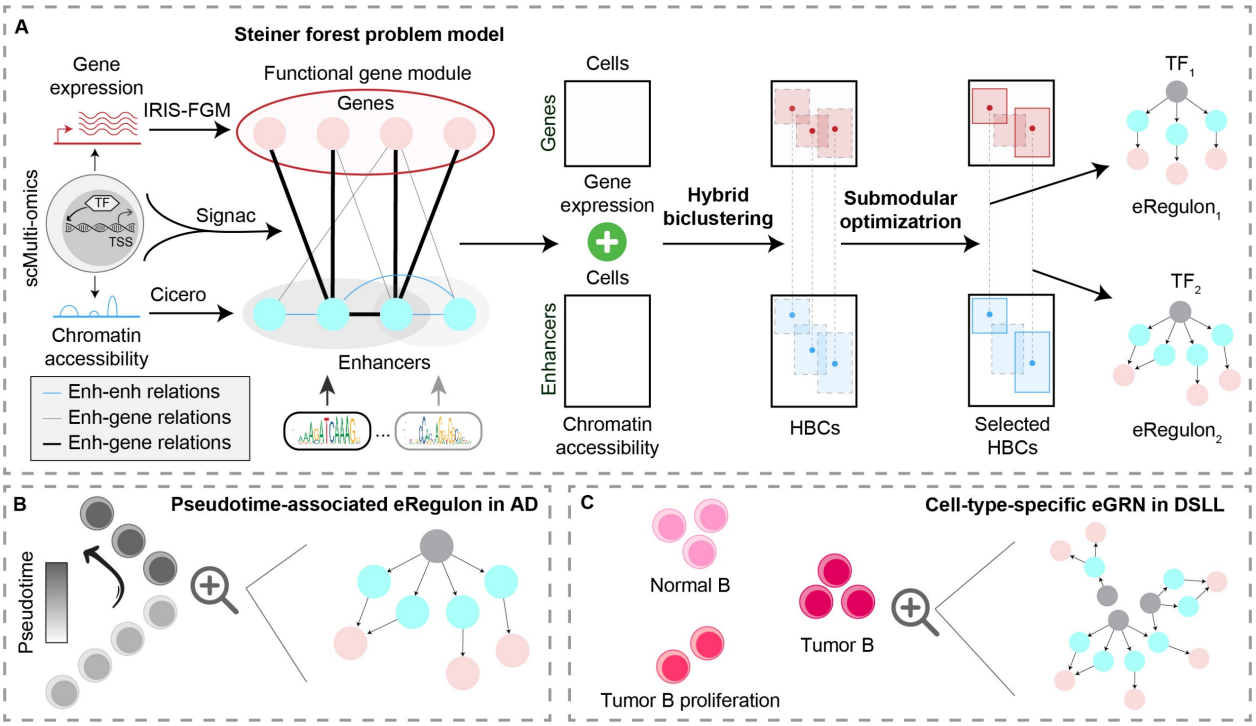


Figure 1. Introduction of the STREAM method and applications in DSLL and AD. **(A)** Overview of the STREAM framework to identify eRegulons. An eRegulon is a TF with its set of binding enhancers and target genes. **(B)** Prediction of pseudotime-associated eRegulons in AD using STREAM. **(C)** Inference of cell-type-specific eGRNs in DSLL via STREAM.

To tackle the above three challenges, we propose **STREAM** (**S**ingle-cell enhancer regula**T**ory netwo**R**k inference from gene **E**xpression **A**nd Chro**M**atin accessibility), a computational framework to infer eGRNs from jointly profiled scRNA-seq and scATAC-seq data (**Figure 1** and **Supplementary Figure S1**). STREAM combines the Steiner forest problem (SFP) model and submodular optimization, respectively, to discover the enhancer-gene relations and TF-enhancer-gene relations in a global optimization manner (**Figure 1A**). First, STREAM builds a heterogeneous graph representing the coordinated relationships among enhancer-enhancer and enhancer-gene relations. Based on the heterogeneous graph, STREAM utilizes the SFP model to identify the set of highly confident enhancer-gene relations which underlie a context-specific functional gene module (FGM) (**Figure 1A** and **Supplementary Figure S1A-S1B**) (13). Second, STREAM formulates the eGRN inference by detecting a set of hybrid biclusters (HBCs), each representing co-regulated genes by the same TF and co-accessible enhancers bound by the same TF over a subset of cells (**Figure 1A** and **Supplementary Figure S1C**). TF-enhancer-gene relations and the cell subpopulation of an HBC could be unveiled without dependency on prior cell clusters. Third, STREAM performs submodular optimization on the HBCs to select the optimal subset of HBCs, composed of the key TF-enhancer-gene relations globally (**Figure 1A** and **Supplementary Figure S1D**) (14). In the outcome, TF, enhancers, and genes within the

same HBC, as well as their relations compose an enhancer Regulon (eRegulon) (**Figure 1A** and **Supplementary Figure S1D**). eRegulons that are active in a cell type compose an eGRN.

To evaluate the performance of STREAM, we applied it to nine jointly profiled RNA and ATAC datasets generated using various technologies and compared the results with SCENIC (11), Signac (6), and GLUE (2). The evaluation criteria include f-scores of pathway enrichment, precisions of chromatin hub interactions (15), and overlap ratios with enhancer-target (ET) pairs in benchmarking analyses. The comparison results revealed that the eGRNs predicted by STREAM showcased higher functional coherency, structural coordination, and *cis*-regulatory interactions than other methods. We then applied STREAM to an Alzheimer's disease (AD) dataset over a time course and unveiled pseudotime-associated eRegulons and changing enhancer-gene relations alongside cell lineages (**Figure 1B**). Additionally, by running STREAM on a diffuse small lymphocytic lymphoma (DSLL) dataset, we demonstrated its capacity to identify the eRegulons underlying diseased cell type and showcase the cooperation among TFs in gene regulation (**Figure 1C**).

MATERIALS AND METHODS

Overview of the STREAM pipeline

The STREAM framework is composed of three parts (**Figure 2**): (i) an SFP model for mining the set of highly confident enhancer-gene relations in support of an FGM (**Figure 2A**), (ii) a hybrid biclustering algorithm to discover genes co-regulated by the same TF and linked co-accessible enhancers (**Figure 2B**), and (iii) a submodular optimization approach to obtain the set of key eRegulons (**Figure 2C**). Based on these functions, STREAM enables the construction of eGRN via connecting the eRegulons active in a cell type.

Data preparation

A total of 11 published scRNA-seq and scATAC-seq datasets were collected in this study (**Supplementary Table S1**). In the benchmarking analyses, we used nine scRNA-seq and scATAC-seq datasets in human and mouse that span eleven tissues/cell lines (lung, A549, skin, blood, embryo, bone marrow, K562, BJ, JK1, Jurkat, and embryonic stem cells) and seven technologies, i.e., scCAT-seq, sci-CAR, SNARE-seq, SHARE-seq, ASTAR-seq, scNMT-seq, and 10X Genomics Multiome (16-23). These data are available either at the NCBI Gene Expression Omnibus (GEO) website, the official website of 10X Genomics, or literature in the public domain. In case studies, we first selected a 10X Genomics Multiome dataset, which has 14k sorted nuclei from DSLL of the lymph node (**Supplementary Table S1**), and another 10X Genomics Multiome dataset of AD mouse model brain coronal sections from one hemisphere over 2.5 months, 5.7 months, and 13+ months (**Supplementary Table S1**).

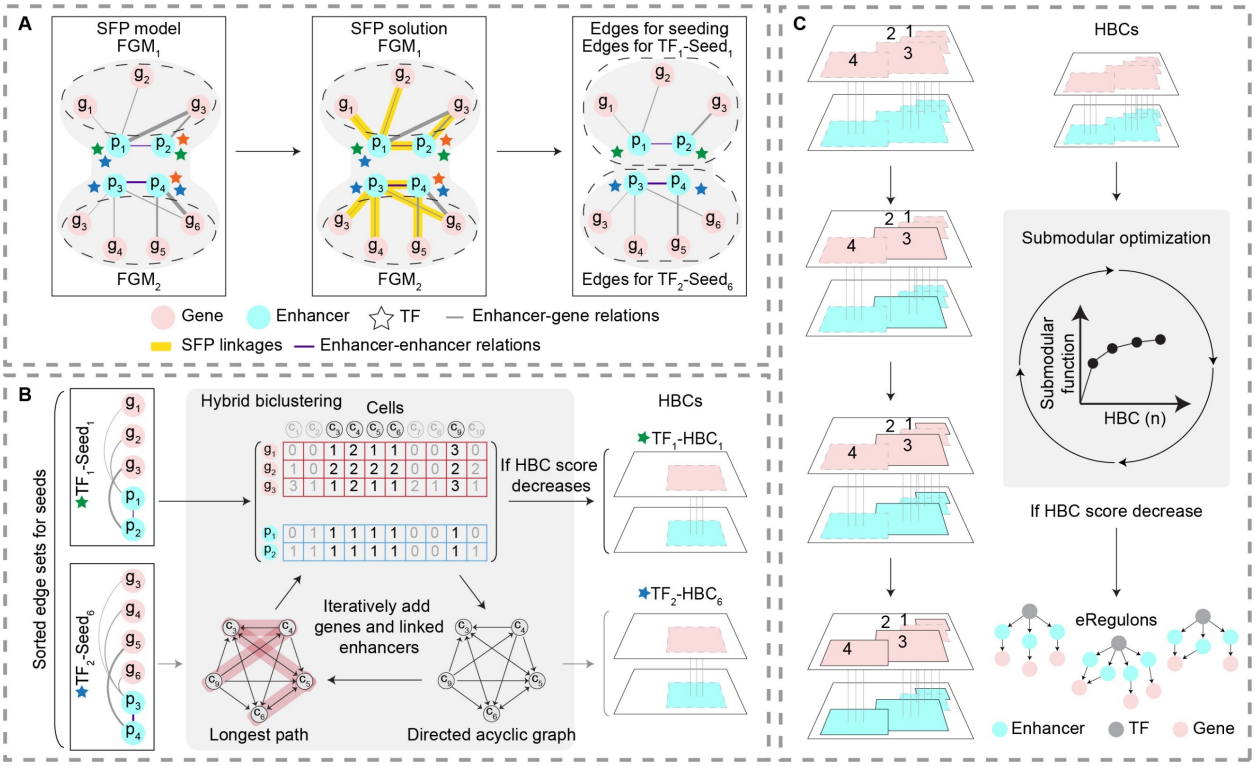


Figure 2. Details of the STREAM framework. **(A)** The SFP model for extracting highly confident enhancer-gene relations. **(B)** The pipeline of hybrid biclustering to identify HBCs. **(C)** Submodular optimization based on HBCs to construct eRegulons.

Data preprocessing

For all the scRNA-seq and scATAC-seq data used in this study, we used the functions *Read10X_h5*, *read_10x*, and *read.table*, respectively, to load count matrices into the program. For scRNA-seq, we utilized the *CreateSeuratObject* function of Seurat (v.4.0.5) to convert the count matrix into a Seurat object. The percentage of mitochondria RNAs were calculated via the *PercentageFeatureSet* function. For scATAC-seq, we retained the enhancers belonging to standard chromosomes using the *standardChromosomes* function. Then, we generated the genomic annotations using the *GetGRangesFromEnsDb* function by setting parameter *ensdb* as *EnsDb.Hsapiens.v86* (human) and *EnsDb.Mmusculus.v79* (mouse), respectively. For all datasets, we utilized the *intersect* function to retain the common cells included in both scRNA-seq and scATAC-seq matrices. For the 10X Genomics Multiome datasets, we incorporated the fragments of scATAC-seq into the Seurat objects via the *CreateChromatinAssay* function. Finally, we use the *subset* function in Seurat for quality control (QC) based on the percentage of mitochondria RNA as well as the count of scRNA-seq and scATAC-seq (**Supplementary Table S2**).

Model specification

Step 1. FGM prediction. The input to STREAM is an expression matrix ($X_{n\times o}$) and a chromatin accessibility matrix ($Y_{m\times o}$) after QC, measuring the expression level (numeric) of n genes and chromatin accessibility (binary) of m peaks, respectively, in o cells (**Figure 2A**). To unveil the gene regulatory information underlying gene expression, we transform the expression matrix $X_{n\times o}$ into a discretized matrix $X'_{n\times o}$ using a left truncated mixture Gaussian (LTMG) model and we call it LTMG matrix (24). LTMG models scRNA-seq data to capture the diversity of gene expression states regulated by different transcriptional regulatory inputs across individual cells (24). Then, we apply a FGM prediction tool, IRIS-FGM, on $X'_{n\times o}$ to obtain biclusters $B_k, k = 1, \dots, l$. Each bicluster, e.g., B_k, represents an FGM.

Step 2. SFP model. We design an SFP model to reverse-engineer the underlying eGRNs by extracting the highly confident enhancer-gene and enhancer-enhancer relations that lead to FGMs (**Figure 2A**). For each bicluster B_k, we construct an LTMG submatrix (X'_k) whose rows correspond to the genes of B_k and a chromatin accessibility submatrix (Y_k) whose rows correspond to all the enhancers. For the columns of both X'_k and Y_k, we only keep the cells in B_k. Based on X'_k and Y_k, we construct a heterogeneous graph $G^{(k)} = (V^{(k)}, E^{(k)})$ using Signac (6) and Cicero (15) to compute the enhancer-gene edges and enhancer-enhancer edges, respectively. We calculate the "cost" of enhancer-gene edges weighted by Pearson correlation coefficients and enhancer-enhancer edges weighted by covariances by subtracting the min-max normalized edge weight by one. Then, we divide the set of nodes denoting genes in $V^{(k)}$ into several subsets $V_s^{(k)}, s = 1, \dots, t$, so that gene nodes belonging to the same subset are connected by a path in $G^{(k)}$. We call these subsets terminal sets. We will not distinguish between genes/enhancers and nodes representing them in graphs in the remaining text.

Given the weighted undirected graph $G^{(k)}$and t terminal sets $V_s^{(k)}, s = 1, \dots, t$, the objective of the SFP model is to identify a minimum cost forest $F^{(k)}$ in $G^{(k)}$, i.e., Steiner forest (SF), such that the nodes inside each $V_s^{(k)}$ are connected via the edges of the forest. Since SFP is a strongly NP-hard problem (13), we utilized heuristic to solve it. Obviously, $F^{(k)}$ is composed of one or more trees $T_s^{(k)}, s = 1, \dots, t$, through each of which nodes within each terminal set are connected. We will identify $F^{(k)}$ by discovering each $T_s^{(k)}$, respectively. For each terminal set $V_s^{(k)}$, we start from the pair of genes that are co-regulated across the maximum number of cells and then identify the shortest path in $G^{(k)}$ between them. We initialize $T_s^{(k)}$ as this shortest path and keep: (i) selecting the gene from $V_s^{(k)} \backslash T_s^{(k)}$, which yields the shortest path to $V_s^{(k)} \backslash T_s^{(k)}$, and (ii) adding nodes of the shortest path to $T_s^{(k)}$, until $V_s^{(k)} \subseteq T_s^{(k)}$. Finally, we merge all $T_s^{(k)}, s = 1, \dots, t$ to obtain the SF, i.e.,

$$F^{(k)} = \bigcup_{s=1}^{t} T_s^{(k)}. \tag{1}$$

We extract the Enhancer-gene relations from edges of $F^{(k)}$ and classify them into enhancer-gene relation subsets via JASPAR TF binding sites (25), such that the enhancers of all enhancer-gene

relations in each subset are bound by the same TF. We only retain the enhancer-gene relation subsets, which link to at least two genes.

Then, we convert each enhancer-gene subset into an initial HBC, i.e., a pair of biclusters whose columns correspond to the same set of cells, by denoting an HBC as

$$H_i^{(k)} = \left(I_i^{(k)}, J_i^{(k)}, K_i^{(k)}\right), i = 1, \dots, p; k = 1, \dots, l, \tag{2}$$

in which $I_i^{(k)}$ and $J_i^{(k)}$ represent the genes and enhancers linked by the enhancer-gene relations in the enhancer-gene subset, respectively, $K_i^{(k)}$ correspond to the cells of B_k, and p refers to the number of HBCs identified from $G^{(k)}$. In each HBC $H_i^{(k)}$, we sort the cells in $K_i^{(k)}$ in decreasing order of the average gene expression values across $I_i^{(k)}$, yielding a monotonical trend $M_i^{(k)}$. Additionally, for each enhancer in $J_i^{(k)}$, we calculate the average ratio $r_i^{(k)}$ of cells in which the enhancer is accessible against all cells in $K_i^{(k)}$. We define $\min(\left|I_i^{(k)}\right|, \left|K_i^{(k)}\right|)$ as the HBC score, based on which we pool and sort all HBCs in decreasing order of HBC scores. The sorted HBCs will act as the set of seeds for hybrid biclustering, and we denote these HBCs as

$$S = \left\{H_i^{(k)} = \left(I_i^{(k)}, J_i^{(k)}, K_i^{(k)}\right) : i = 1, \dots, p; k = 1, \dots, l\right\}. \tag{3}$$

Step 3. Hybrid biclustering. Based on a seed in S, we expand it along with both the vertical (genes and enhancers) and horizontal directions (cells) without violating the monotonical trend of gene expression throughout cells of the HBC (**Figure 2B**). Hybrid biclustering equals simultaneous biclustering on two matrices, each of which is NP-hard (26). Therefore, we employed a heuristic to perform hybrid biclustering. When the HBC cannot be further expanded, output it as the final HBC. Details follow. If S is empty, stop hybrid biclustering; otherwise, check if the first HBC in S is a qualified seed. If it is not, remove it from S, and repeat this step; otherwise, expand it via iteratively finding the longest path from directed acyclic graphs (DAGs) representing the monotonicity of gene expression among cells. Herein, we define a seed as qualified if the seed is overlapped in less than α genes ($\alpha = 0.50$ by default) with any previously identified HBCs.

For the current bicluster $H_i^{(k)}$, we construct a DAG $D_i^{(k)} = (V_i^{(k)}, E_i^{(k)})$, in which nodes represent cells in $K_i^{(k)}$, and each directed edge connects from a higher-ranking cell to a lower-

ranking cell according to the monotonical trend $M_i^{(k)}$. A pair of nodes belonging to a loop or circle is shrunk as a single node. Hybrid biclustering attempts to increase the HBC score, i.e.,

$$score\left(H_i^{(k)}\right) = score\left(\left(I_i^{(k)}, J_i^{(k)}, K_i^{(k)}\right)\right) = \min\left(\left|I_i^{(k)}\right|, \left|J_i^{(k)}\right|, \left|K_i^{(k)}\right|\right), \quad (4)$$

while maintaining the monotonical trend $M_i^{(k)}$ among cells of $K_i^{(k)}$. As known, trend-preserving gene may compose a co-regulated gene module under a subset of conditions or cells (27,28). In each iteration, we build a candidate gene set by incorporating the genes such that: (i) they are not included in $I_i^{(k)}$, (ii) they are linked to at least one enhancer bound by the same TF as $I_i^{(k)}$, and we call these enhancers as supporting enhancers of the gene. For each gene g from the candidate gene set, we reconstruct a DAG $\widetilde{D}_i^{(k)} = (\tilde{V}_i^{(k)}, \tilde{E}_i^{(k)})$ by: (i) removing the cells from $V_i^{(k)}$ in which g is not expressed and meanwhile deleting their incident edges from $E_i^{(k)}$; (ii) excluding the edges in $E_i^{(k)}$ which violate the monotonical trend of expression values of g across cells in $\tilde{V}_i^{(k)}$; (iii) eliminating the genes which have no supporting enhancers that are accessible in $\geq \left|r_i^{(k)} \cdot \tilde{V}_i^{(k)}\right|$ cells. We identify the longest path from each $\widetilde{D}_i^{(k)}$ using dynamic programming and determine the gene (optimal gene) which yields the longest path $\tilde{P}_i^{(k)}$ among all genes in the candidate gene set (29). We still use $\widetilde{D}_i^{(k)}$ to denote the DAG constructed based on the optimal gene. Then, we build a new HBC

$$\widetilde{H}_i^{(k)} = \left(\tilde{I}_i^{(k)}, \tilde{J}_i^{(k)}, \widetilde{K}_i^{(k)}\right) \quad (5)$$

by: (i) setting $\tilde{I}_i^{(k)} = I_i^{(k)} \cup \{g\}$, (ii) if no supporting enhancer of g is included in $J_i^{(k)}$, letting $\tilde{J}_i^{(k)} = J_i^{(k)} \cup \{e\}$, where e is the supporting enhancer of g accessible in the maximum number of cells of $\tilde{V}_i^{(k)}$; otherwise, keeping $\tilde{J}_i^{(k)} = J_i^{(k)}$; (iii) setting $\widetilde{K}_i^{(k)}$ as the nodes in $\tilde{P}_i^{(k)}$. If $score\left(\widetilde{H}_i^{(k)}\right) \geq score\left(H_i^{(k)}\right)$, set $H_i^{(k)}$ as $\widetilde{H}_i^{(k)}$, and $D_i^{(k)}$ as $\widetilde{D}_i^{(k)}$; otherwise, terminate the iteration, output $H_i^{(k)}$, and delete the current seed from S. If S is empty, stop hybrid biclustering; otherwise, select the next qualified seed for hybrid biclustering.

Finally, we sort the HBCs in decreasing order of HBC score. In addition, within each HBC, we build Enhancer-gene relations ($C_i^{(k)}$) by linking each gene to the enhancers which are within 250 kb from the transcription start site (TSS) (10). Correspondingly, we denote each HBC as

$$H_i^{(k)} = \left(I_i^{(k)}, J_i^{(k)}, K_i^{(k)}, C_i^{(k)}\right). \tag{6}$$

We send all the HBCs for submodular optimization.

Step 4. HBC optimization. The goal of HBC optimization is to identify a subset of HBCs that yield complementary views of both expression and enhancer throughout cells in a dataset, which will be output as the set of eRegulons. The solution to HBC optimization is composed of three parts (**Figure 2C**): (i) an evaluation function to determine the number of eRegulons in a dataset; (ii) an objective function that defines the informativeness of a subset of HBCs; and (iii) a submodular optimization algorithm that identifies a subset of HBCs that scores highly according to the objective function (14).

The *evaluation function* is defined as

$$\Delta \mathrm{R} = R^{in} - R^{out}, \tag{7}$$

in which R^{in} represents the number of enhancer-gene pairs linked by enhancer-gene relations in HBCs, while R^{out} denotes the number of enhancer-gene pairs that are within 250 kb by default to each other, but no enhancer-gene relation in HBCs connects them.

We propose to borrow the objective function named *facility location*, which measures what fraction of the information available in the complete set of HBCs is contained within the HBC subset. Using U and W to denote the set and subset of HBCs, respectively, the facility location function $f: 2^U \rightarrow \mathbb{R}$ is defined as follows:

$$f(W) = \sum_{h' \in U} \max_{h \in W} r_{h',h}, \tag{8}$$

where $r_{h',h}$ quantifies the pairwise similarity between HBCs $h' = (I', J', K', C')$ and $h = (I, J, K, C)$, i.e.,

$$r_{h',h} = \min\left(\frac{|I \cap I'| \cdot |K \cap K'|}{\min(|I| \cdot |K|, |I'| \cdot |K'|)}, \frac{|J \cap J'| \cdot |K \cap K'|}{\min(|J| \cdot |K|, |J'| \cdot |K'|)}\right). \tag{9}$$

Intuitively, the facility location function takes a high value when every HBC in U has at least one similar representative in W. We define the *conditional gain* of f as

$$f(h|W) = f(h \cup W) - f(W). \tag{10}$$

Submodular optimization starts with the empty set $W_0 = \emptyset$ and at each iteration i adds the HBC h_i that maximizes the conditional gain $f(h_i|W_{i-1})$ with ties broken arbitrarily (i.e., finding $h_i \in arg \max_{h \in U \setminus W_{i-1}} f(h|W_{i-1})$) and then updates $W_i \leftarrow W_{i-1} \cup \{h_i\}$. The algorithm stops when the

number of HBCs (to be determined be the evaluation function) to select is met. We output these HBCs as eRegulons.

Comparison with four existing eGRN inference tools

We compared STREAM with four existing methods for GRN inference under a wide array of parameter settings. These methods include (i) SCENIC v.1.1.2, which is an R package to reconstruct GRNs from scRNA-seq (11); (ii) Signac v.1.7.0, a comprehensive R toolkit for ATAC (Signac-ATAC) and (iii) RNA and ATAC analyses (Signac-RNA-ATAC) (6); (iv) GLUE v.0.2.2, a modular framework for integrating unpaired scRNA-seq and scATAC-seq data and inferring regulatory interactions simultaneously (2). As to the retained methods, we followed tutorials and recommendations of these methods and implemented each of them by trying 2-4 values for each of their 3-4 parameters, leading to ~ 16 combinations of parameters for grid search (**Supplementary Table S3**).

Workflow introduction of SCENIC. SCENIC is a framework developed to infer GRNs based solely on scRNA-seq, by integrating three R packages: (i) GENIE3, to excavate potential TF-gene relations based merely on gene coexpression (30); (ii) RcisTarget, to determine the direct TF-gene relations via motif enrichment analysis (11); and (iii) AUCell, to quantify the activity of regulons in single cells (11).

Framework description of Signac based on ATAC (Signac-ATAC). Signac-ATAC could identify eRegulons by predicting TF-enhancer and enhancer-gene relations via motif enrichment analysis and discovering enhancers co-accessible with gene promoters, respectively (6). We infer eRegulons using Signac on scATAC-seq by predicting the set of DARs overrepresented in a cell cluster and bound by the same TF. We perform normalization and dimension reduction using the LSI method in Signac and then find cell clusters with the *FindClusters* function. Meanwhile, we utilize Cicero to construct enhancer-enhancer relations (15). For each cell cluster, we retain the enhancer-enhancer relations, of which both enhancers are DARs specific in this cell cluster. Then, we link the enhancers to their bound TFs and to the genes whose promoters are overlapped with the enhancers, leading to an eRegulon. Finally, we sort the eRegulons in decreasing order of the number of genes and output them.

Method details of Signac based on RNA and ATAC (Signac-RNA-ATAC). Signac-RNA-ATAC is able to construct eRegulons via predicting TF-enhancer and enhancer-gene relations using motif enrichment analysis and correlation calculation, respectively (6). We reconstruct eRegulons using Signac on scRNA-seq and scATAC-seq via combining the prediction of enhancer-gene relations and TF binding site annotation. We perform normalization scRNA-seq and scATAC-seq, respectively, using Signac. Then, using the *LinkPeaks* function of Signac, we predict the enhancer-gene relations. The TF-enhancer relations are constructed by annotating enhancers

using their bound TFs. Finally, we pool the TF-enhancer and enhancer-gene relations to construct eRegulons.

Model overview of GLUE. GLUE is a modular framework based on deep learning for integrating scRNA-seq and scATAC-seq data for inferring *cis*-regulatory relations (2). After QC, we convert the final Seurat objects into the standard file format of 10X genomics (Feature-Barcode Matrices), for compliance with GLUE (2). GLUE models cell states as low-dimensional cell embeddings through different variational autoencoders for each modality (31), respectively. Then, GLUE utilizes prior biological knowledge regarding locations of enhancers and genes to construct a guidance graph. Based on the guidance graph, adversarial multimodal alignment is performed iteratively to refine the enhancer-gene relations. Finally, SCENIC is used to construct TF-gene relations (11). By filtering out the enhancers linked to target genes of a TF, we can obtain the TF-enhancer-gene relations.

Evaluation metrics for eGRN inference. We compared the performance of different eGRN/GRN identification methods from three viewpoints: pathway enrichment (32), overlaps with chromatin hubs (15), and intersections with ET pairs (33,34). Pathway enrichment measures how functionally synergistic genes of an eRegulon are; the overlaps with chromatin hubs assess how coordinated in histone marks enhancers of an eRegulon undergo; the intersection with ET pairs evaluates how coherence expression and accessibility of an eRegulon are. For pathway enrichment, we utilized precision, recall, and f-score to quantify the percentage of eRegulons significantly enriched (p-value < 0.05 adjusted via Bonferroni correction) in KEGG pathways against all eRegulons, ratio of significantly enriched pathways among all pathways, and the prediction accuracy, respectively. The definition of f-score is

$$f\text{–score} = 2 \cdot precision \cdot recall/(precision + recall)\,. \tag{11}$$

enhancers in a chromatin hub should satisfy four criteria (15): (i) greater physical proximity than expected based on linear distance on genome; (ii) interaction with the common set of protein complexes; (iii) concordant epigenetic modification; (iv) contribution to gene regulation. The second and fourth criteria are obvious for eRegulons. Due to the limitation of chromatin interaction data available in the public domain, we focused on measuring the concordance in gaining H3K27 acetylation (H3K27ac), which is associated with the higher activation of transcription and therefore defined as an active enhancer mark. Therefore, to assess the abundance of chromatin hubs in eRegulons, we computed the percentage of enhancers of each eGRN that was overlapped with peaks of all Cistrome Data Browser (CistromeDB) datasets of chromatin immunoprecipitation assays with sequencing, i.e., ChIP sequencing (ChIP-seq) for H3K27ac (35).

Regarding the intersection with ET pairs, we pooled all datasets of ET pairs from EnhancerAtlas 2.0 (34), and then computed the ratio of enhancer-gene pairs in eRegulons that were intersected with ET pairs in the database among all enhancer-gene relations in eRegulons.

Prediction of enhancer-gene relations associated with the trajectory in AD

We inferred trajectory for each cell type, respectively, based on the transcriptomics part of the scRNA-seq and scATAC-seq data by running monocle3 v.1.0.0 with cells labeled by 2.5 months as the root of trajectory (36).

Cell-type-specific eGRN construction in DSLL

We constructed eGRN specific in a cell type of DSLL via two steps: (i) identified cell-type-specific eRegulons in a cell type by hypergeometric test (p-value < 0.05) followed by Bonferroni correction; (ii) reconstructed cell-type-specific eGRN by pooling all eRegulons specific in this cell type and connecting all TF-enhancer-gene relations. DEGs or DARs overrepresented in each cell type were discovered using the *FindMarkers* function of Seurat v.4.0.5 under default parameters. We selected the DEGs/DARs with adjusted p-values < 0.05 as significant ones.

RESULTS

STREAM accurately predicts eRegulons from scRNA-seq and scATAC-seq data

We collected nine scRNA-seq and scATAC-seq datasets for benchmarking the performance of eGRN inference (**Figures 3A**). These datasets were composed of 35-14,104 cells (**Figure 3B**), in which 18,666-63,568 genes were expressed (**Figure 3C**) and 4,706-288,662 enhancers were accessible (**Figure 3D**). On these datasets, we carried out benchmarking analysis between STREAM and four GRN inference approaches on scRNA-seq (SCENIC), scATAC-seq (Signac-ATAC), and scRNA-seq together with scATAC-seq (Signac-RNA-ATAC and GLUE). To measure the functional homogeneity of co-regulated genes in eRegulons, we benchmarked if STREAM could infer eRegulons enriched in KEGG pathways (**Figures 3E-3M** and **Methods**). We found that STREAM achieved the best mean f-scores among all methods on eight datasets except the sci-CAR dataset, which is inferior to SCENIC. Among the eight datasets, STREAM achieved significantly higher f-scores (p-value < 0.05) than all other approaches on seven of them except the SNARE-seq dataset. The lower performance of STREAM on both sci-CAR and SNARE-seq could be due to the low sequencing quality and throughput (19,20). Generally, SCENIC was the runner-up in this comparison and performed well on the nine datasets except for ASTAR-seq and scNMT-seq. SCENIC performed stably across the nine datasets, as it only depends on one modality. Besides, SCENIC achieved the lowest f-score on the first ASTAR-seq dataset because of the small number of cells. Note that GLUE, Signac-ATAC, and Signac-RNA-ATAC failed under several parameter settings on the nine datasets because of small numbers of cells or dimensions and failure to conform to distribution (**Supplementary Table S4**). We compared the f-scores under the default parameter setting, median f-scores, precisions, and recalls (**Figures 3E-3M**

and **Materials and Methods**). STREAM exceeded other methods evaluated via these four metrics, and it possessed particularly higher precision on nine datasets (**Figures 3E-3M**).

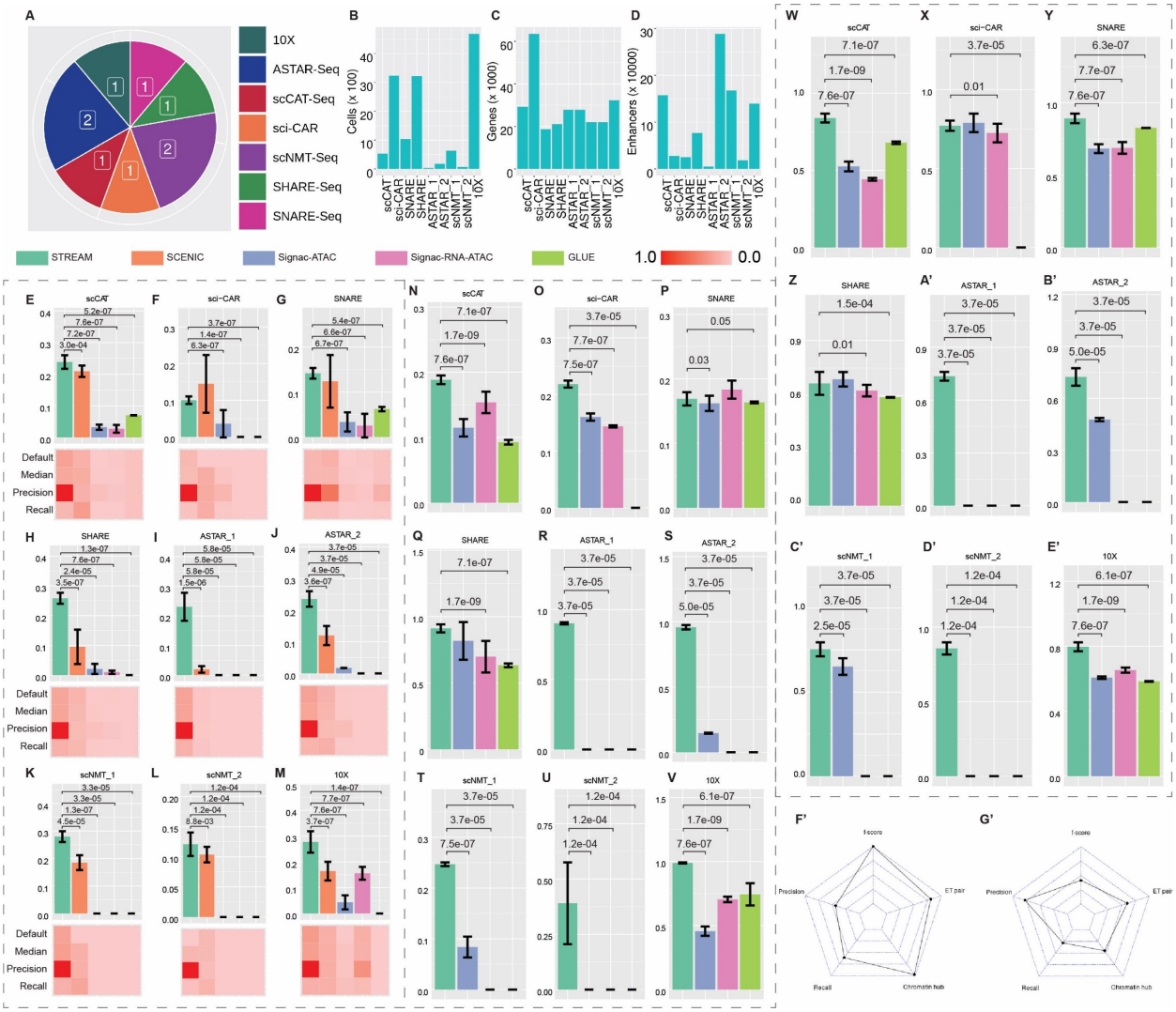


Figure 3. Benchmarking of eRegulon prediction accuracy. **(A)** Datasets and corresponding techniques used in benchmarking. **(B-D)** The numbers of cells, genes, and enhancers in the benchmarking datasets. **(E-M)** The mean f-scores (bar plots), default and median f-scores, precisions, and recalls of pathway enrichment analysis for all the five methods. The Wilcoxon rank sum test p-values of the mean f-scores between STREAM and other four methods were showcased. **(N-V)** Proportions of eRegulon enhancers that are overlapped with H3K27ac ChIP-seq peaks. **(W-E')** Fractions of enhancer-gene relations that are intersected with enhancer-target pairs curated in EnhancerAtlas 2.0. **(F'-G')** Radar charts of the above evaluation metrics of STREAM on the AD and DSLL datasets used in two case studies, respectively.

As co-accessible enhancers form a chromatin hub and undergo coordinated changes in H3K27ac, we calculated the abundance of enhancers of an eRegulon in H3K27ac ChIP-seq

peaks. Herein the peak set in a ChIP-seq dataset acts as a proxy of chromatin hubs (**Figures 3N-3V** and **Methods**). STREAM achieved the highest abundance among the four approaches on eight of the nine datasets, with the SNARE-seq dataset as the only exception. The possible reason could be the low throughput of SNARE-seq. We also observed significantly higher precisions of STREAM compared to other methods, assessed by Wilcoxon rank sum test p-values.

To further demonstrate the advantages of STREAM, we mapped the enhancer-gene relations to curated ET pairs and compared the ratio of enhancer-gene relations overlapped with ET pairs (**Figures 3W-3E'** and **Materials and Methods**). From the results, STREAM exceeded the other three methods on seven of the nine datasets, with the sci-CAR and SHARE-seq datasets as exceptions. It is clear more than 50% of the enhancer-gene relations in STREAM eRegulons are overlapped with ET pairs in the EnhancerAtlas 2.0 database (34). We verified the significantly higher precision of STREAM compared to existing tools using Wilcoxon rank sum test p-values.

STREAM uncovers pseudotime-associated eRegulons and changing enhancer-gene relations alongside trajectories in AD

To demonstrate the capacity of temporal scRNA-seq and scATAC-seq analysis, we present another case study that applies STREAM to a published scRNA-seq and scATAC-seq dataset ($n = 21{,}374$) composed of 32,286 genes and 66,861 enhancers of AD mouse model brain over a time course (2.5 months: 7,517; 5.7 months: 2,357; and 13+ months: 11,500) generated using 10X Genomics Multiome. We identified 27 cell clusters using Seurat and then manually annotated each cluster by visualizing the expression and activity level of marker genes (**Supplementary Table S5**), obtaining seven cell types, including oligodendrocytes (Oligo), oligodendrocyte progenitor (OPC), inhibitory neuron (IN), excitatory neuron (EX), astrocyte (AG), microglia (MG), and endothelial & pericyte (**Figures 4A-4B**). We ran STREAM on this dataset and identified 81 eRegulons of TFs important in AD, e.g., androgen receptor (AR) (37), JUN family proteins (38), ESR2 (39,40), FOSL2 (41), PLAG1 (42), RUNX1 (43), RORA (44), and STAT2 (45), etc. We assessed the similarity between eRegulons identified at different stages via the fraction of overlapped enhancer-gene relations, and then visualized their relationships using a graph (**Figure 4C**). On the one hand, high similarity existed between pairs of eRegulons regulated by the same TF, e.g., NR2C2, NFE2, JUND, ELK4, AR, FOSL2 at three stages, RUNX1 at 2.5 and 5.7 months, and PLAG1 at 5.7 and 13+ months, etc. On the other hand, some eRegulons regulated by different TFs also had a large proportion of intersected enhancer-gene relations, e.g., FOS (2.5 months) v.s. FOSL2 (5.7 months), FOSL2 (5.7 months) v.s. FOS (13+ months), NR2C2 (5.7 months) v.s. NFE2 (13+ months), JUN (2.5 months) v.s. FOSL2 (5.7 months), and FOSL2 (5.7 months) v.s. JUN (13+ months), etc. Based on these eRegulons, we identified cell-type-specific ones in EX ($n = 18$), IN ($n = 11$), and Oligo ($n = 17$), respectively.

To investigate the changing in regulatory strengths of eRegulons alongside pseudotime in EX cells, we extracted the EX cells and inferred trajectory and calculated pseudotime (**Figure 4D**). We calculated the mean regulatory strength of enhancer-gene relations within each eRegulon

and transformed them into z-scores (**Figure 4E**). We observed monotonical trends in the regulatory strengths of some eRegulons alongside the pseudotime, including both increasing and decreasing trends. eRegulons regulated by RUNX1, FOS, NFE2, JUND, and FOSL2 showcased decreasing trend in gene expression. Hence, they have high regulatory strengths in cells aligned at the beginning of the pseudotime. Besides, high regulatory strengths were observed for eRegulons regulated by NR2C2, ESR2, RUNX1, and FOSL2 in cells aligned in the middle of the pseudotime. Additionally, the regulatory strengths of eRegulons regulated by NKX3-1, ELK4, FOSL2, AR, and ESR2 were increasing alongside the pseudotime. As known, RUNX1 exerts a pro-neurogenic function in neural precursor cells (46); the decreasing concentration of AR contributes to cognitive deficits in the elderly and AD patients by mediating signaling pathways (47); JUN and FOS could influence the initiation and execution of apoptosis and thus are associated with neuronal risk and/or cell death in AD (48,49); variation in ESR2 has the potential to be linked with increased AD susceptibility (50). JUND belongs to the JUN protein family, possessing the potential to impact the cell apoptosis (48,49). ELK4 was proved to be associated with glia or neuron functions in AD (51); FOSL2, a member of the FOS protein family, participates in gene regulation through collaboratively forming TF complex AP-1 with JUN protein family (52). We focused on chr11-70646709-70647600, which was bound by FOS and FOLS2, and found that it had decreasing regulatory strength on *Mink1* (**Figure 4F**). Mink1 is a cognition-associated gene that is critical in AD (53). Moreover, chr12-110696009-110696853, regulated by ELK4 and ESR2, exerted increasing regulatory strength on the AD-associated gene *Dync1h1* (**Figure 4G**) (54).

We analyzed the changes in gene expression of eRegulons alongside pseudotime in IN cells using the same method as EX cells (**Figure 4H-4I**). We also discovered monotonical trends in regulatory strengths of eRegulons in IN cells, which were regulated by ESR2, TLX1, JUN, JUND, NFE2, NR2C2, ELK4, RORA, and KAISO (encoded by Zbtb33). RORA emerges as a gene with a probable central role in the AD pathology/etiology (44), whose expression is distinctly upregulated in the AD hippocampus (44). KAISO regulates the development of the central nervous system (55). ESR2, ESR2, and JUN bind chr2-34372425-34373312 to regulate *Pbx3*, an important gene in the central nervous system (56), with a decreasing regulatory strength alongside pseudotime (**Figure 4J**). On the contrary, chr2-121806663-121807507 was found to exert an increasing regulatory strength on *Frmd5*, which was co-regulated by ELK4 and JUN (**Figure 4K**).

In addition to neuron cells, we conducted the same analysis on Oligo cells (**Figure 4L**). We presented example eRegulons of which regulatory strength was monotonical alongside pseudotime (**Figure 4M**). These eRegulons were regulated by a similar list of TFs with EX cell types (**Figure 4E**) and IN (**Figure 4I**) analyses. We also focused on two enhancers to explore their regulatory strengths on target genes. chr17-65884560-65885475 was bound by FOSL2 and FOS and exerted a decreasing regulatory strength on *Rab31*, which often exists as a target of RUNX1 in AD (**Figure 4N**) (57). Moreover, the regulatory strength of chr10-79681469-79682367

on *Bsg* was increasing alongside pseudotime under the regulation by AR and JUN (**Figure 4O**). As known, *Bsg* is involved in learning and memory (58), whose dysregulation may lead to abnormalities of sensory and memory functions (59).

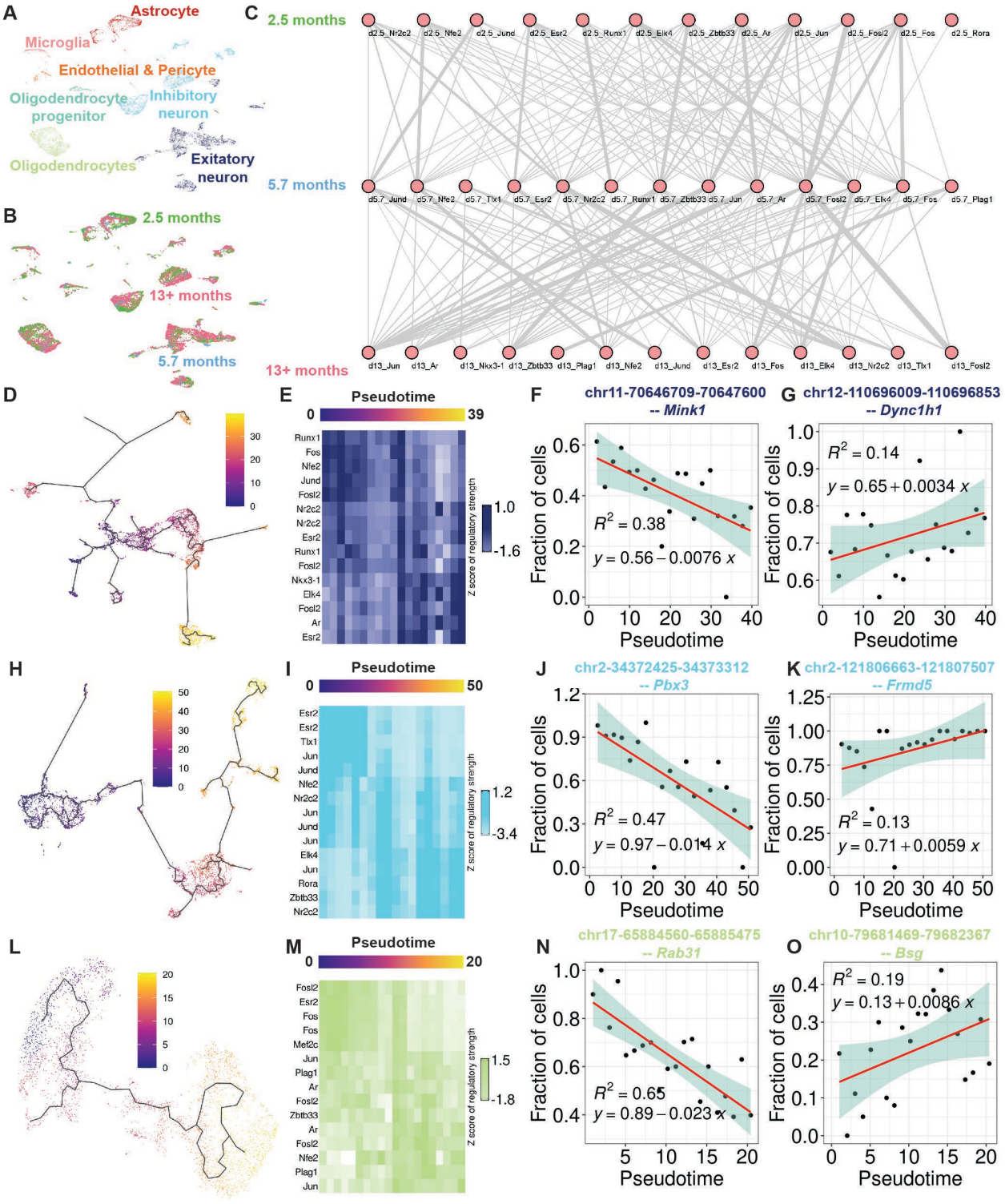


Figure 4. Analyses of pseudotime-associated eRegulons and the changing trend of enhancer-gene relations. **(A-B)** UMAP plots with colors denoting cell types **(A)** and stages **(B)**. **(C)** the graph to visualize the similarity between eRegulons identified in three stages, in which nodes represent eRegulons, while weighted edges denote pairwise similarity. **(D)** UMAP plot with colors indicating pseudotime in excitatory neuron cells. **(E)** Heatmap of mean regulatory strengths of enhancer-gene relations in eRegulons specific to excitatory neuron cells alongside pseudotime. **(F-G)** Representative enhancer-gene relations, which showcase a monotonical trend in regulatory strengths alongside pseudotime in excitatory neuron cells. **(H)** UMAP plot with colors indicating pseudotime in inhibitory neuron cells. **(I)** Heatmap of mean regulatory strengths of enhancer-gene relations in eRegulons specific to inhibitory neuron cells alongside pseudotime. **(J-K)** Representative enhancer-gene relations, which showcase a monotonical trend in regulatory strengths alongside pseudotime in inhibitory neuron cells. **(L)** UMAP plot with colors indicating pseudotime in oligodendrocytes cells. **(M)** Heatmap of mean regulatory strengths of enhancer-gene relations in eRegulons specific to oligodendrocytes alongside pseudotime. **(N-O)** Representative enhancer-gene relations, which showcase a monotonical trend in regulatory strengths alongside pseudotime in oligodendrocytes.

In general, through inferring the regulatory strength of eRegulons or enhancer-gene relations alongside trajectory in EX (**Figures 4D-4G**), IN (**Figures 4H-4K**), Oligo (**Figures 4L-4O**), MG (**Supplementary Figure S2**), AG (**Supplementary Figure S3**), and OPC (**Supplementary Figure S4**), STREAM showcased its potential to infer eRegulons or enhancers playing critical role in cell developmental process.

STREAM unveils underlying eRegulons in diseased B cells in DSLL

To illustrate the application power of STREAM, we utilized a DSLL dataset sequenced using 10X Genomics Multiome, containing 14,104 cells, 36,601 genes, and 70,469 enhancers. We performed unsupervised clustering followed by manual annotation of cell types using curated gene markers (60). We subset this dataset to only retain the cells annotated by 11 cell types (60) (**Figure 5A**), and then ran STREAM on the scRNA-seq and scATAC-seq dataset of DSLL, leading to 50 eRegulons composed of 19-290 genes and 9-70 enhancers active in 34-703 cells. Within these eRegulons, 99.6% (4,632) genes were linked to 1-10 enhancers, whereas 47.0% (4,658) enhancers were only linked to the closest gene (**Figure 5B**).

Among the identified eRegulons, we extracted 37 cell-type-specific eRegulons in Effector-like CD4+ T cell type 1 (EL CD4+ T1), Effector-like CD4+ T cell type 2 (EL CD4+ T1), Exhaustive CD4+ T cells (Exh CD4+ T), Effector-like CD8+ T cells (EL CD8+ T), Central memory-like CD8+ T cells (CML CD8+ T), Exhaustive CD8+ T cells (Exh CD8+ T), Dendritic cells (DCs), Normal B cells (Normal B), Tumor B proliferation cells (Tumor B pro), and Tumor B cells (Tumor B), which were specific to eight cell types. We observed obvious co-binding among TFs of these cell-type-specific eRegulons (**Figure 5C**). In addition, we found some well-known TFs regulated these cell-type-specific eRegulons in the DSLL dataset, e.g., TCF7 (61) and LHX2 in Normal B (62), SPI1 (63,64), TCF12 (65), STAT3 (66,67), NFIC (68), NFIB (69), and MYF5 (70) in both Tumor B pro

and Tumor B. On the one hand, we observed obvious changing consistency between expression and chromatin accessibility in most eRegulons regulated by these TFs (**Figure 5D**). On the other hand, we also found that no variation was observed in chromatin accessibility despite the changes in expression for some eRegulons, e.g., LHX2, SPI1, TCF7, and TCF12. Hence, chromatin accessibility influences transcription within distinct eRegulons in different cell types (71).

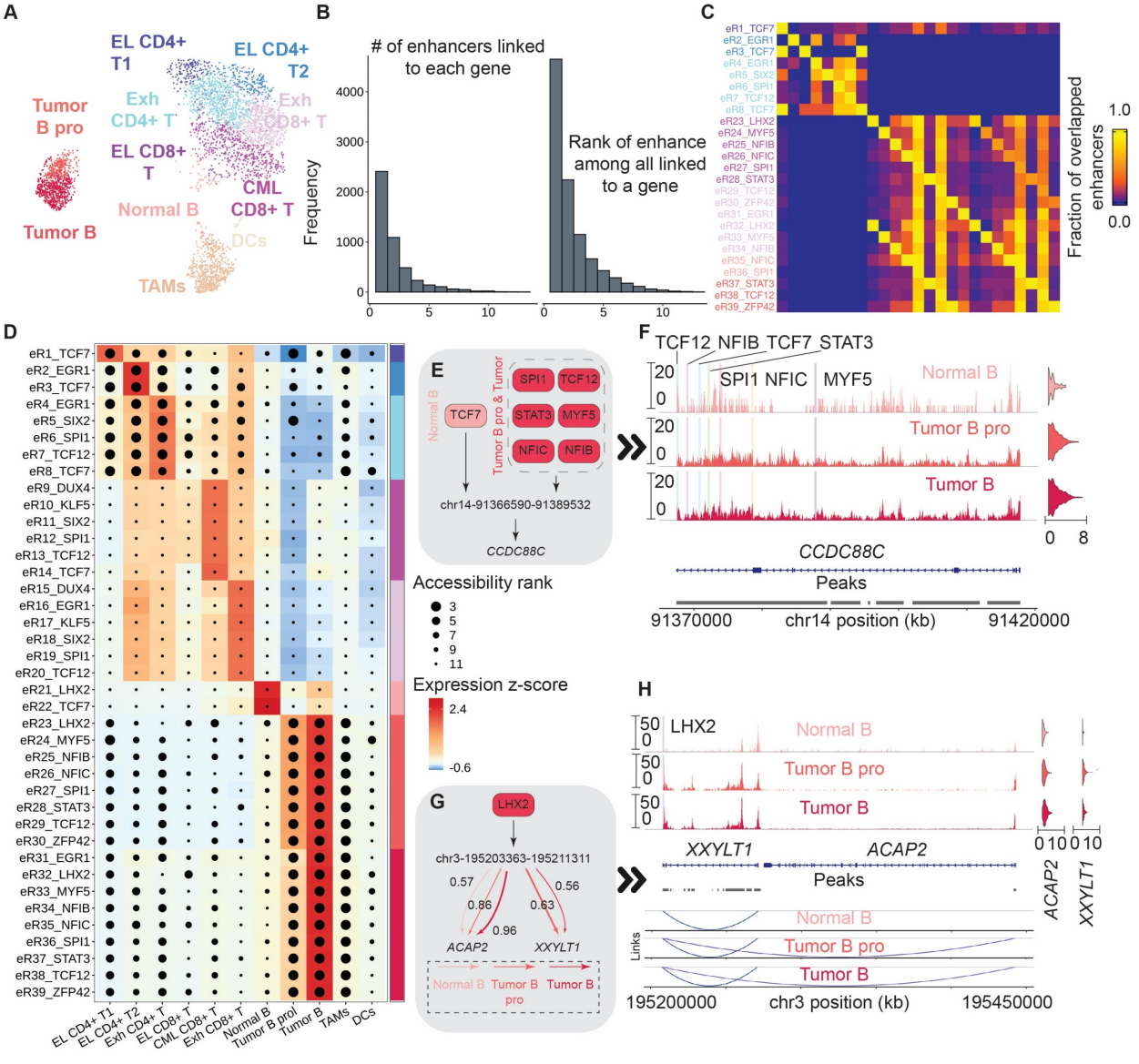


Figure 5. STREAM eRegulons reveal key enhancer and enhancer-gene relations underlying tumor development in B cells. **(A)** UMAP plot of cell types in DSLL. **(B)** Distribution of the number of enhancers linked to each gene and distribution of rank based on the absolute distance between enhancer and gene for each enhancer. **(C)** Fraction of overlapped enhancers between a pair of eRegulons divided by the number of enhancers in the eRegulon in each row. **(D)** Heatmap-dotplot showing gene expression of the eRegulon on a color scale and the ranks of chromatin accessibility of the eRegulon on a size scale. **(E)** The

schematic overview showcases the variation of TFs that regulate *CCDC88C* via binding chr14-91366590-91389532 among three B cell types, i.e., Normal B, Tumor B proliferation, and Tumor B cells. **(F)** Chromatin accessibility profiles labeled by the binding sites of the TFs in **(E)** across three B cell types within the region chr14-91370000-91420000. **(G)** The schematic overview showcases the variation of enhancer-gene relations between chr3-195203363-195211311 and a pair of genes (*ACAP2* and *XXYLT1*) across three B cell types. **(H)** Chromatin accessibility profiles labeled by the binding sites of the LHX2 across three B cell types within the region chr3-195203363-195211311. Enhancer-gene links are shown using arcs, and the color of arcs denotes the fraction of cells, in which the enhancer is accessible while the gene is expressed.

To investigate the underlying TF-enhancer-gene relations in DSLL, we focused on the cell-type-specific eRegulons in three B cell types, i.e., Normal B, Tumor B proliferation, and Tumor B. We identified two, eight, and nine cell-type-specific eRegulons in Normal B, Tumor B proliferation, and Tumor B, respectively. Two cell-type-specific eRegulons regulated by TCF7 and LHX2, respectively, were discovered in Normal B cells. In the LHX2 and TCF7 eRegulons, 14 and 73 genes were regulated by TFs binding 11 and 82 enhancers, respectively. LHX2, which codes for a regulatory protein involved in the differentiation of lymphoid multipotent progenitor cells with self-renewal properties, was found to be expressed early during B cell differentiation in a subepithelial lymphoid compartment (61). TCF7 was proven to play an important role in B cell and T cell development (62). Regarding the eight cell-type-specific eRegulons in Tumor B proliferation cells, TFs (LHX2, MYF5, NFIB, NFIC, SPI1, STAT3, TCF12, and ZFP42) regulated 14-478 genes via binding 11-1,097 enhancers. As a member of myogenic regulatory factors, MYF5 plays an important role in myoblast specificity and cell proliferation (70). In leukemia/lymphoma, abnormal myoblasts accumulate in the bone marrow and crowd out healthy blood cells (72). NFIB was proven to facilitate lymphoma cell proliferation, migration, and invasion (69). NFIC is a lymphoid-specific TF in the lymphoma/leukemia (68). PU. 1, encoded by SPI1, was demonstrated to activate the transcription of genes required for B cell differentiation and function (63,64). STAT3 is active in activated B-cell-like diffuse large B-cell lymphomas (66,67). HEB, encoded by TCF12, plays an important role in B-/T- cell commitment (65). Although REX1, encoded by ZFP42, is a known marker of pluripotency and is usually found in undifferentiated embryonic stem cells, its mutation was found to be associated with implicating pathways involved in lymph node oncogenesis (73). As to the nine cell-type-specific eRegulons, a similar scenario with Tumor B proliferation cells were obtained (i.e., 14-478 genes and 11-1,097 enhancers) in Tumor B cells. The major difference between eRegulons specific to Tumor B proliferation and Tumor B cells was that an EGR1 (Early growth response gene) eRegulon was discovered in Tumor B cells. EGR1 is a TF known to be a downstream effector of B-cell receptor signaling and Janus kinase 1 signaling in B-cell lymphoma (74). In sum, STREAM could discover trustworthy TF-enhancer-gene relations that can explain two Tumor B cell types.

To inspect the impact of TF binding variation on critical gene expression across various B cell types, we concentrated on one marker gene of DSLL, *CCDC88C*, and one enhancer (chr14-

91366590-91389532), through which *CCDC88C* was regulated by various TFs. As known, Daple (encoded by *CCDC88C*) modulates and activates Wnt signaling (75), which decreases the cell apoptosis (76). Additionally, *CCDC88C* was proved to be up-regulated in diseased B cells in lymphoma/leukemia compared to healthy ones (77). We observed variation of TFs binding chr14-91366590-91389532 among three B cell types (**Figure 5E**). Specifically, this enhancer was bound by TCF7 in Normal B cells while co-bound by six different TFs (SPI1, TCF12, STAT3, MYF5, NFIC, and NFIB) in Tumor B proliferation and Tumor B cells. Co-operativity was observed among these TFs on this enhancer (**Figure 5F**). Finally, violin plots indicated up-regulation of *CCDC88C* between Tumor B proliferation or Tumor B cells versus (v.s.) Normal B cells (**Figure 5F**). In conclusion, chr14-91366590-91389532 can participate in the regulation of *CCDC88C* by different combinations of TFs, such as TCF7, SPI1, TCF12, STAT3, MYF5, NFIC, and NFIB, leading to the changes of *CCDC88C* expression and cell differentiation.

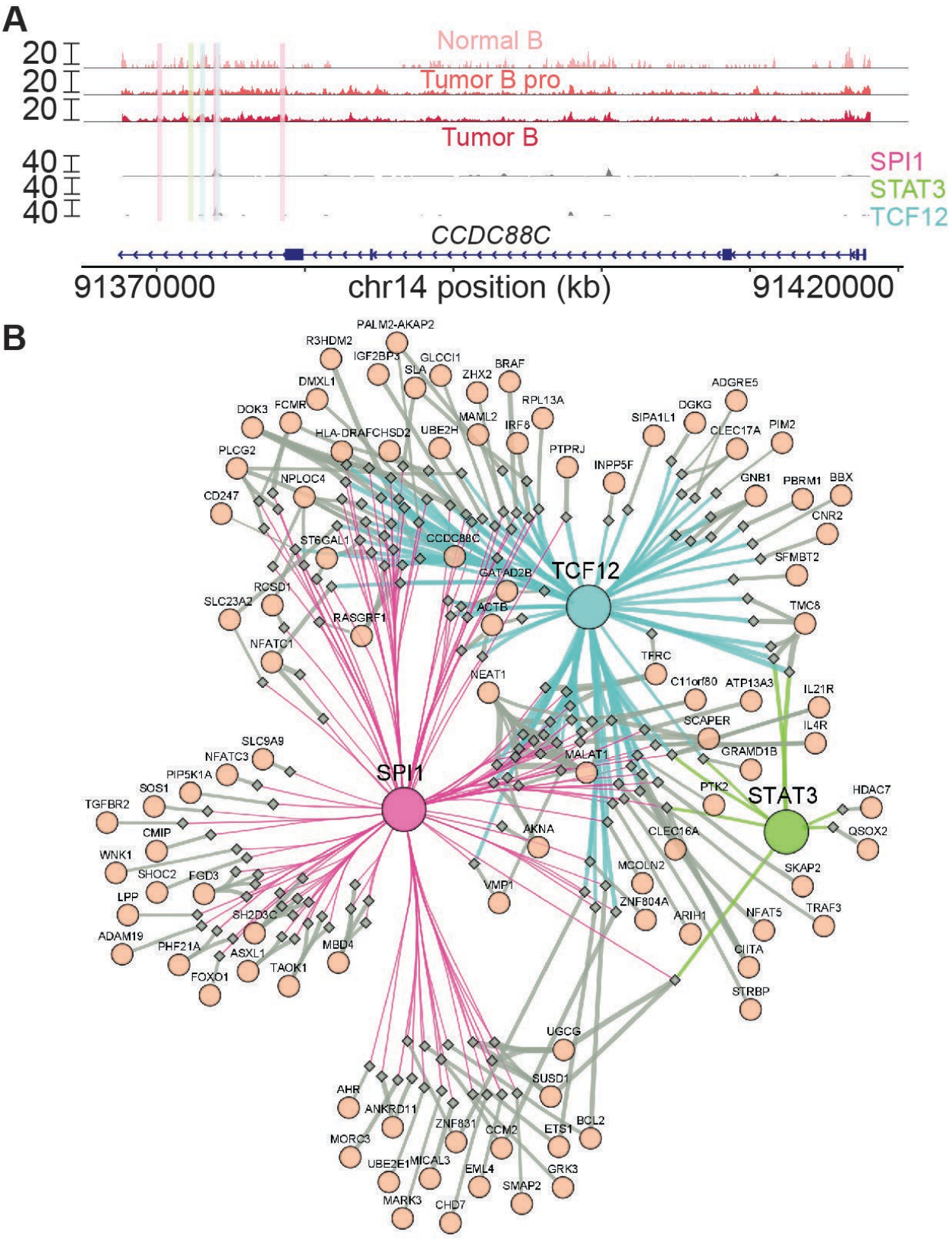
Besides TF binding, we also investigated how enhancer-gene relations influenced marker gene expression. We selected two marker genes, *ACAP2* and *XXYLT1*, which were regulated by LHX2 via binding chr3-195203363-195211311 (**Figure 5G**). *ACAP2* is an over-represented marker gene in lymphoma (78), whereas *XXYLT1* negatively regulates Notch signaling activation (79). Notch signaling could switch an oncogene to a tumor suppressor (80). We calculated the fraction of cells where the enhancer is accessible, and the gene is expressed, denoting the regulatory strength. We found that LHX2 regulated *XXYLT1* via binding chr3-195203363-195211311 only in Tumor B proliferation and Tumor B cells with regulatory strengths: 0.63 and 0.56, respectively, rather than in Normal B cells. As to *ACAP2*, LHX2 regulated *ACAP2* through the enhancer chr3-195203363-195211311 with different regulatory strengths (i.e., Normal B: 0.57; Tumor B proliferation: 0.86; Tumor B: 0.96) in three cell types. We generated a coverage plot within the region incorporating chr3-195203363-195211311 and the two markers and observed increased events of Tn5 insertions in Tumor B proliferation and Tumor B cells v.s. Normal B cells. Hence, it is probably the increasing regulatory strengths between chr3-195203363-195211311 and the two markers that lead to the over-representation of marker expression, thus, the tumor development of B cells (**Figure 5H**). The variation in regulatory strengths might be due to the changing chromatin accessibility. In summary, chr3-195203363-195211311 could be engaged in regulating *ACAP2* and *XXYLT1* by LHX2 with different regulatory intensities in different B cell types, which results in the alternation of marker gene expression, cell functions, and differentiation.

STREAM constructs comprehensive eGRNs showcasing the co-operativity among TFs in gene regulation across cell types

To globally illustrate various TF-enhancer-gene relations in B cells of DSLL, we selected SPI1, STAT3, and TCF12, of which ChIP-seq signal files in GM12878 could be obtained from ENCODE (81), to reconstruct an eGRN in Normal B, Tumor B proliferation, and Tumor B cells, respectively. SPI1 and TCF12 play an important role in B cell development and commitment, respectively (63-65), while STAT3 is actively involved in B cell lymphoma (66,67). Additionally, SPI1 and STAT3

collaboratively regulate several pathways important to the tumorigenesis (82). Similarly, as a member of the basic helix-loop-helix family, TCF12 promotes metastasis in, e.g., lymph nodes (83). Moreover, the three TFs participate in the regulation of *CCDC88C*, an activator of the Wnt signaling pathway. We highlighted the binding sites of them within the region chr14-91370000-91420000 in three B cell types as well as ChIP-seq datasets of SPI1, STAT3, and TCF12 in the GM12878 cell line (**Figure 6A**). We observed overlaps between the TF binding sites on the enhancer and ChIP-seq signals, especially increased chromatin accessibility and ChIP-seq signals at regions commonly targeted by SPI1 and TCF12 (**Figure 6A**). In conclusion, eGRNs inferred by STREAM can be utilized to exploit co-operativity among multiple TFs in gene regulation.

We pooled all TF-enhancer-gene relations based on a top-ranked variable based on the three TFs and their variable enhancers and genes and calculated the weight of TF-enhancer and enhancer-gene relations via the fraction of cells in which the gene (target genes or gene encoding TFs) was expressed and the enhancer was accessible. We obtained three eGRNs composed of 276 nodes and 457 edges with different edge weights in three B cell types, respectively (**Figure 6B**, **Supplementary Figures S5-S6**). For the eGRN in Tumor B proliferation cells, we discovered 253 enhancers co-bound by SPI1 and STAT3 regulating 148 genes, 297 enhancers co-bound by SPI1 and TCF12 regulating 94 genes, and 208 enhancers co-bound by STAT3 and TCF12 regulating 69 genes, respectively (**Supplementary Table S6**). We performed pathway enrichment for the three groups of regulated genes against KEGG (32), resulting in 45, 43, and 19 pathways (p-value < 0.05 adjusted via Bonferroni correction), respectively (**Supplementary Table S7**). Co-regulated genes included by these pathways showcased higher regulatory strengths in Tumor B Proliferation cells compared with Normal B and Tumor B cells. Genes co-regulated by SPI1 and STAT3 were enriched in pathways including Leukocyte transendothelial migration with regulatory strengths 0.55 (Normal B), 0.99 (Tumor B Proliferation), and 0.73 (Tumor B), Toxoplasmosis with regulatory strengths 0.60 (Normal B), 0.95 (Tumor B Proliferation), and 0.76 (Tumor B), Human papillomavirus infection with regulatory strengths 0.26 (Normal B), 0.31 (Tumor B Proliferation), and 0.27 (Tumor B), and Thyroid hormone signaling pathway with regulatory strengths 0.55 (Normal B), 1.0 (Tumor B Proliferation), 0.71 (Tumor B), etc. Leukocyte transendothelial migration is a process induced by intracellular signaling in lymphoma (84); Toxoplasmosis is a complication of lymphoma (85); Persistent human papillomavirus infection may promote carcinogenesis by hyperactivation of the immune system and is associated with an increased risk of lymphoma (86); Thyroid hormone stimulate cancer cell proliferation via dysregulation of molecular and signaling pathways (87). Genes co-regulated by SPI1 and TCF12 were enriched in pathways such as Acute myeloid leukemia with regulatory strengths 0.16 (Normal B), 0.33 (Tumor B Proliferation), and 0.28 (Tumor B), Human cytomegalovirus infection with regulatory strengths 0.41 (Normal B), 0.83 (Tumor B Proliferation), and 0.64 (Tumor B), Phospholipase D signaling pathway with regulatory strengths 0.33 (Normal B), 0.63 (Tumor B Proliferation), and 0.41 (Tumor B), and Choline metabolism in cancer with regulatory strengths


A
Normal B
Tumor B pro
Tumor B
SPI1
STAT3
TCF12
CCDC88C
91370000
chr14 position (kb)
91420000
B
TCF12
SPI1
STAT3
PALM2-AKAP2
R3HDM2
IGF2BP3
GLCCI1
SLA
ZHX2
BRAF
DMXL1
RPL13A
ADGRE5
DGKG
DOK3
FCMR
MAML2
IRF8
SIPA1L1
CLEC17A
PIM2
HLA-DRAFCHSD2
UBE2H
PTPRJ
PLCG2
INPP5F
NPLOC4
GNB1
PBRM1
BBX
CD247
CNR2
ST6GAL1
CCDC88C
SFMBT2
GATAD2B
ROSD1
SLC23A2
ACTB
TMC8
RASGRF1
NFATC1
TFRC
NEAT1
C11orf80
ATP13A3
IL21R
SCAPER
IL4R
SLC9A9
NFATC3
MALAT1
GRAMD1B
PIP5K1A
SOS1
PTK2
HDAC7
TGFBR2
AKNA
CLEC16A
QSOX2
CMIP
WNK1
SKAP2
SHOC2
FGD3
MCOLN2
LPP
VMP1
ZNF804A
TRAF3
SH2D3C
ARIH1
NFAT5
ADAM19
PHF21A
ASXL1
MBD4
CIITA
FOXO1
TAOK1
STRBP
UGCG
SUSD1
AHR
ANKRD11
ZNF831
CCM2
ETS1
BCL2
MORC3
UBE2E1
MICAL3
EML4
GRK3
SMAP2
MARK3
CHD7

1

Figure 6. Cooperation among SPI1, STAT3, and TCF12 in B cells. **(A)** Chromatin accessibility profiles across three B cell types and ChIP-seq signals of SPI1, STAT3, and TCF12 within the region chr14-91370000-91420000. **(B)** Visualization of eGRN specific to Tumor B proliferation cells, which are formed by SPI1, STAT3, and TCF12 and restricted to highly variable genes/enhancers. The thickness of the edge between a pair of nodes represents the fraction of cells where both nodes are accessible (enhancers) or expressed (genes encoding TFs or targets).

0.12 (Normal B), 0.33 (Tumor B proliferation), and 0.21 (Tumor B), etc. Leukemia and lymphoma are both forms of blood cancer, affecting blood and lymph nodes, respectively (88); Albeit the role of cytomegalovirus infection in the development of lymphoma is unclear, studies implied a potential association between them (89-91); Phospholipase D signaling contributes to the progression of cancer, e.g., lymphoma (92); lymphoma features the dysregulation of Choline metabolism (93). Pathways enriched by genes co-regulated by STAT3 and TCF12 were incorporated by that enriched by genes co-regulated by SPI1 and STAT3 or SPI1 and TCF12, respectively, e.g., Th1 and Th2 cell differentiation with regulatory strengths 0.36 (Normal B), 0.70 (Tumor B proliferation), and 0.54 (Tumor B) and Inflammatory bowel disease with regulatory strengths 0.47 (Normal B), 0.94 (Tumor B proliferation), and 0.73 (Tumor B). The balance between Th1 and Th2 activity is critical in lymphoid cell development and differentiation (94); the risk of lymphoma is associated with inflammatory bowel disease (94).

DISCUSSION

We developed STREAM, a computational method to infer eRegulons and eGRNs in scRNA-seq and scATAC-seq data. We performed a comprehensive benchmarking analysis between STREAM and four existing GRN inference models dedicated to gene expression modality (SCENIC), chromatin accessibility modality (Signac-ATAC), and both modalities (Signac-RNA-ATAC and GLUE). By comparing the performance under a wide array of parameter combinations, strong evidence has been showcased to validate the effectiveness of STREAM, assessed by pathway enrichment, chromatin hub overlap, and ET pair intersection. Besides, we investigated the capacity of STREAM in temporal scRNA-seq and scATAC-seq analysis. The results indicated that STREAM could identify the eRegulons of which genes showcased monotonical trends alongside the pseudotime. Then, STREAM computed the key enhancer-gene relations showcasing a monotonical trend quantified by the fraction of cells where the enhancer was accessible, and the gene was expressed. These predictions now await experimental validation. Finally, by applying STREAM on a DSLL scRNA-seq and scATAC-seq dataset, we identified eRegulons specific to both healthy and diseased cell types. The case study revealed the capacity of STREAM to excavate the variation in binding TFs and enhancer accessibility implying tumor development. Moreover, STREAM reconstructed eGRNs based on eRegulons in various cell types, by which cooperation among TFs in gene regulation was explored. These predictions now await experimental validation. Overall, the implementation and application of the STREAM

framework provide a solid foundation for the elucidation of gene regulatory mechanisms across different cell types or alongside cell lineages.

STREAM has the following intrinsic limitations. First, like many other GRN prediction methods, STREAM depends on hyperparameters (**Supplementary Table S3**). For example, STREAM is sensitive to the consistency level in hybrid biclustering. Second, STREAM did not perform well on scRNA-seq and scATAC-seq data of which the chromatin accessibility is of low quality, e.g., Paired-seq (95), which sequences chromatin accessibility using single-ended technology. Third, STREAM relies on annotations, including TF binding sites in JASPAR (25), so STREAM is not suitable for constructing eGRNs containing TFs not included in JASPAR.

Overall, STREAM is an effective tool for eGRN inference. eGRNs identified by STREAM can provide a characterization of complex regulatory mechanisms from multiple modalities in diverse application scenarios. The power of STREAM could be further enhanced to identify eRegulons or eGRNs as well as cell types, via borrowing methodologies from our in-house cell-type-specific GRN prediction approach, DeepMAPS (60).

DATA AND CODE AVAILABILITY

All data used for benchmarking and case studies are collected from the public domain and can be retrieved using links or accession numbers provided in **Supplementary Table S1**. STREAM is available at the GitHub repository (https://github.com/YangLi-Bio/stream_core).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank Qiuqin Wu, Xiaoying Wang, Zhenyu Wu, Yujia Xiang, and Cindy Tong for their assistance in data collection, pipeline development, and tool testing.

Authors' contributions: Q.M. and B.L. conceived the basic idea and designed the overall analyses. H.F. provided a guide for biological annotation in the AD case study. Y.L. designed the specific experiments and developed the framework. Y.L., A.M., Y.W., and C.W. conducted the computational analysis and data interpretation. S.C. annotated the biological data for the AD case study. Y.L., A.M., Q.M., and B.L. wrote the manuscript.

FUNDING

This work was supported by awards R01GM131399 and U54AG075931 from the National Institute of General Medical Sciences of the National Institutes of Health. The work was also supported by the award NSF1945971 from the National Science Foundation. This work was supported by the Pelotonia Institute of Immuno-Oncology (PIIO). The content is solely the responsibility of the authors and does not necessarily represent the official views of the PIIO.

Conflict of interest statement. No author has competing interests.

REFERENCES

1. Jin, T., Rehani, P., Ying, M., Huang, J., Liu, S., Roussos, P. and Wang, D. (2021) scGRNom: a computational pipeline of integrative multi-omics analyses for predicting cell-type disease genes and regulatory networks. *Genome Med*, **13**, 95.
2. Cao, Z.-J. and Gao, G. (2022) Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*, 1-9.
3. Janssens, J., Aibar, S., Taskiran, I.I., Ismail, J.N., Gomez, A.E., Aughey, G., Spanier, K.I., Rop, F.V.D., González-Blas, C.B., Dionne, M. *et al.* (2022) Decoding gene regulation in the fly brain. *Nature*, **601**, 630-636.
4. Duren, Z., Lu, W.S., Arthur, J.G., Shah, P., Xin, J., Meschi, F., Li, M.L., Nemec, C.M., Yin, Y. and Wong, W.H. (2021) Sc-compReg enables the comparison of gene regulatory networks between conditions using single-cell data. *Nat Commun*, **12**, 4763.
5. Duren, Z., Chen, X., Zamanighomi, M., Zeng, W., Satpathy, A.T., Chang, H.Y., Wang, Y. and Wong, W.H. (2018) Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. *Proceedings of the National Academy of Sciences*, **115**, 7723-7728.
6. Stuart, T., Srivastava, A., Madad, S., Lareau, C.A. and Satija, R. (2021) Single-cell chromatin state analysis with Signac. *Nat Methods*, **18**, 1333-1341.
7. Duren, Z., Chang, F., Naqing, F., Xin, J., Liu, Q. and Wong, W.H. (2022) Regulatory analysis of single cell multiome gene expression and chromatin accessibility data with scREG. *Genome Biol*, **23**, 114.
8. Yuan, H. and Kelley, D.R. (2022) scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural networks. *Nat Methods*, **19**, 1088-1096.
9. Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y. and Greenleaf, W.J. (2015) Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, **523**, 486-490.
10. Granja, J.M., Corces, M.R., Pierce, S.E., Bagdatli, S.T., Choudhry, H., Chang, H.Y. and Greenleaf, W.J. (2021) ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature Genetics*, **53**, 403-411.
11. Aibar, S., González-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J. *et al.* (2017) SCENIC: single-cell regulatory network inference and clustering. *Nat Methods*, **14**, 1083-1086.
12. Wang, Y.K., Hurley, D.G., Schnell, S., Print, C.G. and Crampin, E.J. (2013) Integration of steady-state and temporal gene expression data for the inference of gene regulatory networks. *Plos One*, **8**, e72103.

13. Gassner, E. (2010) The Steiner Forest Problem revisited. *Journal of Discrete Algorithms*, **8**, 154-163.
14. Wei, K., Libbrecht, M.W., Bilmes, J.A. and Noble, W.S. (2016) Choosing panels of genomics assays using submodular optimization. *Genome Biol*, **17**, 229.
15. Pliner, H.A., Packer, J.S., McFaline-Figueroa, J.L., Cusanovich, D.A., Daza, R.M., Aghamirzaie, D., Srivatsan, S., Qiu, X., Jackson, D., Minkina, A. *et al.* (2018) Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Molecular cell*, **71**, 858-871.e858.
16. Liu, L., Liu, C., Quintero, A., Wu, L., Yuan, Y., Wang, M., Cheng, M., Leng, L., Xu, L., Dong, G. *et al.* (2019) Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity. *Nat Commun*, **10**, 470.
17. Clark, S.J., Argelaguet, R., Kapourani, C.-A., Stubbs, T.M., Lee, H.J., Alda-Catalinas, C., Krueger, F., Sanguinetti, G., Kelsey, G., Marioni, J.C. *et al.* (2018) scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat Commun*, **9**, 781.
18. Argelaguet, R., Clark, S.J., Mohammed, H., Stapel, L.C., Krueger, C., Kapourani, C.-A., Imaz-Rosshandler, I., Lohoff, T., Xiang, Y., Hanna, C.W. *et al.* (2019) Multi-omics profiling of mouse gastrulation at single cell resolution. *Nature*, **576**, 487-491.
19. Cao, J., Cusanovich, D.A., Ramani, V., Aghamirzaie, D., Pliner, H.A., Hill, A.J., Daza, R.M., McFaline-Figueroa, J.L., Packer, J.S., Christiansen, L. *et al.* (2018) Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, **361**, 1380-1385.
20. Chen, S., Lake, B.B. and Zhang, K. (2019) High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature Biotechnology*, **37**, 1452-1457.
21. Xing, Q.R., Farran, C.A.E., Zeng, Y.Y., Yi, Y., Warrier, T., Gautam, P., Collins, J.J., Xu, J., Dröge, P., Koh, C.-G. *et al.* (2020) Parallel bimodal single-cell sequencing of transcriptome and chromatin accessibility. *Genome Res*, **30**, 1027-1039.
22. Ma, S., Zhang, B., LaFave, L.M., Earl, A.S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V.K., Tay, T. *et al.* (2020) Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell*, **183**, 1103-1116.e1120.
23. Genomics, X. (2020), *10X Genomics*.
24. Wan, C., Chang, W., Zhang, Y., Shah, F., Lu, X., Zang, Y., Zhang, A., Cao, S., Fishel, M.L., Ma, Q. *et al.* (2019) LTMG: a novel statistical modeling of transcriptional expression states in single-cell RNA-Seq data. *Nucleic Acids Res*, **47**, e111.
25. Castro-Mondragon, J.A., Riudavets-Puig, R., Rauluseviciute, I., Lemma, R.B., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Pérez, N.M. *et al.* (2022) JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic acids research*, **50**, D165-D173.

26. Li, G., Ma, Q., Tang, H., Paterson, A.H. and Xu, Y. (2009) QUBIC: a qualitative biclustering algorithm for analyses of gene expression data. *Nucleic acids research*, **37**, e101.
27. Ben-Dor, A., Chor, B., Karp, R. and Yakhini, Z. (2003) Discovering local structure in gene expression data: the order-preserving submatrix problem. *Journal of computational biology : a journal of computational molecular cell biology*, **10**, 373-384.
28. Liu, X., Li, D., Liu, J., Su, Z. and Li, G. (2020) RecBic: a fast and accurate algorithm recognizing trend-preserving biclusters. *Bioinformatics*, **36**, 5054-5060.
29. Eddy, S.R. (2004) What is dynamic programming? *Nature Biotechnology*, **22**, 909-910.
30. Huynh-Thu, V.A., Irrthum, A., Wehenkel, L. and Geurts, P. (2010) Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *Plos One*, **5**, e12776.
31. Kingma, D.P. and Welling, M.J.a.p.a. (2013) Auto-encoding variational bayes.
32. Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, **27**, 29-34.
33. Gao, T., Zheng, Z., Pan, Y., Zhu, C., Wei, F., Yuan, J., Sun, R., Fang, S., Wang, N., Zhou, Y. *et al.* (2021) scEnhancer: a single-cell enhancer resource with annotation across hundreds of tissue/cell types in three species. *Nucleic Acids Research*, **50**, D371-D379.
34. Gao, T. and Qian, J. (2020) EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Research*, **48**, D58-D64.
35. Zheng, R., Wan, C., Mei, S., Qin, Q., Wu, Q., Sun, H., Chen, C.-H., Brown, M., Zhang, X., Meyer, C.A. *et al.* (2018) Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Research*, **47**, D729-D735.
36. Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J. *et al.* (2019) The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, **566**, 496-502.
37. Ferrari, R., Dawoodi, S., Raju, M., Thumma, A., Hynan, L.S., Maasumi, S.H., Reisch, J.S., O'Bryant, S., Jenkins, M., Barber, R. *et al.* (2013) Androgen receptor gene and sex-specific Alzheimer's disease. *Neurobiology of aging*, **34**, 2077.e2019-2020.
38. MacGibbon, G.A., Lawlor, P.A., Walton, M., Sirimanne, E., Faull, R.L., Synek, B., Mee, E., Connor, B. and Dragunow, M. (1997) Expression of Fos, Jun, and Krox family proteins in Alzheimer's disease. *Experimental neurology*, **147**, 316-332.
39. Pirskanen, M., Hiltunen, M., Mannermaa, A., Helisalmi, S., Lehtovirta, M., Hänninen, T. and Soininen, H. (2005) Estrogen receptor beta gene variants are associated with increased risk of Alzheimer's disease in women. *European Journal of Human Genetics*, **13**, 1000-1006.
40. Zhao, L., Woody, S.K. and Chhibber, A. (2015) Estrogen receptor β in Alzheimer's disease: From mechanisms to therapeutics. *Ageing research reviews*, **24**, 178-190.

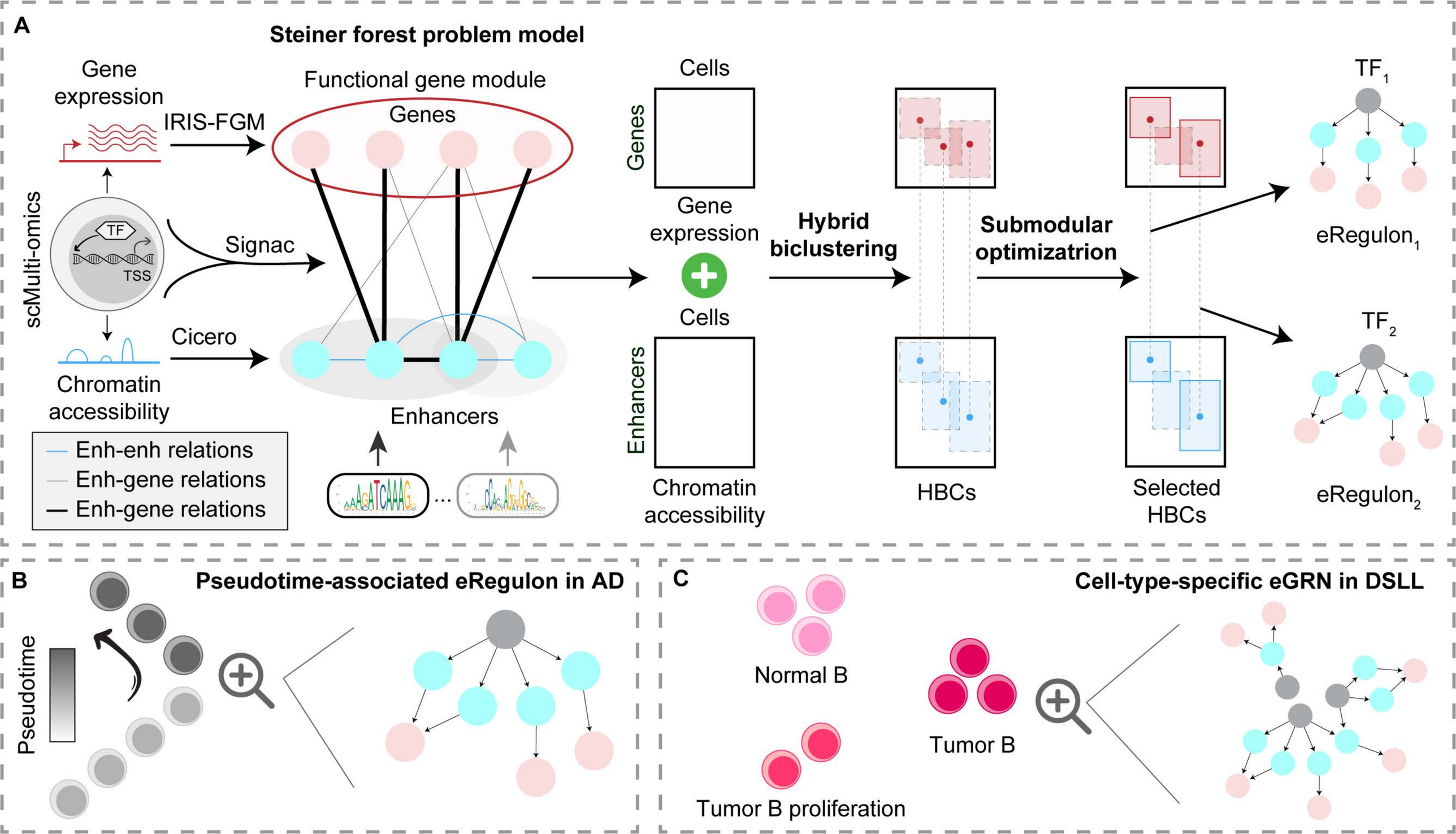
41. Morabito, S., Miyoshi, E., Michael, N., Shahin, S., Martini, A.C., Head, E., Silva, J., Leavy, K., Perez-Rosendahl, M. and Swarup, V. (2021) Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's disease. *Nat Genet*, **53**, 1143-1155.
42. Liu, C., Chyr, J., Zhao, W., Xu, Y., Ji, Z., Tan, H., Soto, C. and Zhou, X. (2018) Genome-Wide Association and Mechanistic Studies Indicate That Immune Response Contributes to Alzheimer's Disease Development. *Frontiers in genetics*, **9**, 410.
43. Patel, A., Rees, S.D., Kelly, M.A., Bain, S.C., Barnett, A.H., Thalitaya, D. and Prasher, V.P. (2011) Association of variants within APOE, SORL1, RUNX1, BACE1 and ALDH18A1 with dementia in Alzheimer's disease in subjects with Down syndrome. *Neuroscience letters*, **487**, 144-148.
44. Acquaah-Mensah, G.K., Agu, N., Khan, T. and Gardner, A. (2015) A regulatory role for the insulin- and BDNF-linked RORA in the hippocampus: implications for Alzheimer's disease. *Journal of Alzheimer's disease : JAD*, **44**, 827-838.
45. Sierksma, A., Lu, A., Mancuso, R., Fattorelli, N., Thrupp, N., Salta, E., Zoco, J., Blum, D., Buée, L., De Strooper, B. *et al.* (2020) Novel Alzheimer risk genes determine the microglia response to amyloid-β but not to TAU pathology. **12**, e10606.
46. Fukui, H., Rünker, A., Fabel, K., Buchholz, F. and Kempermann, G. (2018) Transcription factor Runx1 is pro-neurogenic in adult hippocampal precursor cells. *Plos One*, **13**, e0190789.
47. Raber, J. (2004) Androgens, ApoE, and Alzheimer's Disease. **2004**, re2-re2.
48. Anderson, A.J., Su, J.H. and Cotman, C.W. (1996) DNA damage and apoptosis in Alzheimer's disease: colocalization with c-Jun immunoreactivity, relationship to brain area, and effect of postmortem delay. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, **16**, 1710-1719.
49. Marcus, D.L., Strafaci, J.A., Miller, D.C., Masia, S., Thomas, C.G., Rosman, J., Hussain, S. and Freedman, M.L. (1998) Quantitative neuronal c-Fos and c-Jun expression in Alzheimer’s disease11To whom correspondence should be addressed. *Neurobiology of aging*, **19**, 393-400.
50. Pirskanen, M., Hiltunen, M., Mannermaa, A., Helisalmi, S., Lehtovirta, M., Hänninen, T. and Soininen, H. (2005) Estrogen receptor beta gene variants are associated with increased risk of Alzheimer's disease in women. *European journal of human genetics : EJHG*, **13**, 1000-1006.
51. Barrera, J., Song, L., Gamache, J.E., Garrett, M.E., Safi, A., Yun, Y., Premasinghe, I., Sprague, D., Chipman, D., Li, J. *et al.* (2021) Sex dependent glial-specific changes in the chromatin accessibility landscape in late-onset Alzheimer’s disease brains. *Molecular Neurodegeneration*, **16**, 58.
52. Hüttenrauch, M., Salinas, G. and Wirths, O.J.F.i.m.n. (2016) Effects of long-term environmental enrichment on anxiety, memory, hippocampal plasticity and overall brain gene expression in C57BL6 mice. **9**, 62.

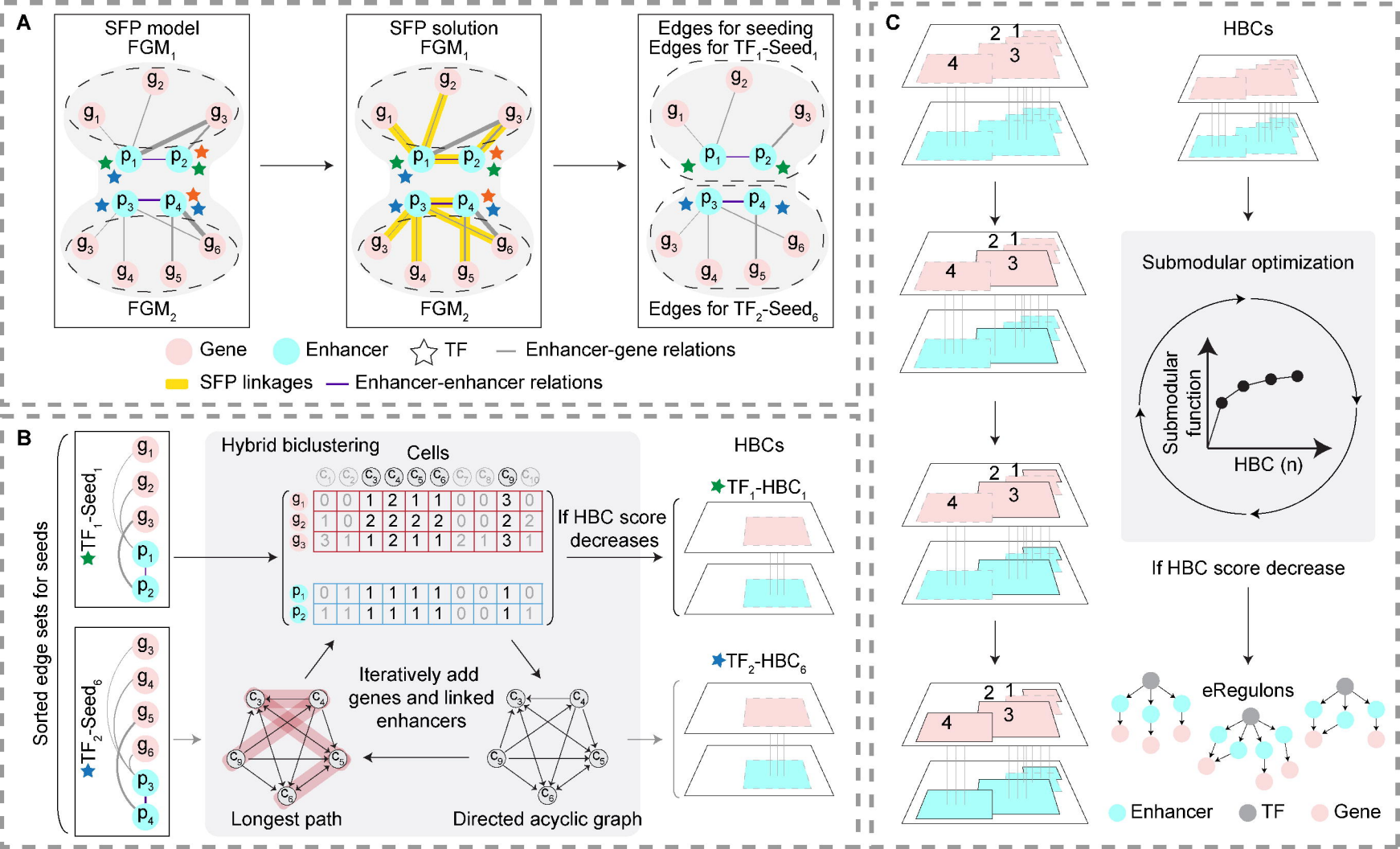
53. Vigneswaran, J., Muthukumar, S.A., Shafras, M. and Pant, G. (2021) An insight into Alzheimer's disease and its on-setting novel genes. *The Egyptian Journal of Neurology, Psychiatry and Neurosurgery*, **57**, 160.
54. Mentis, A.A., Vlachakis, D., Papakonstantinou, E., Zaganas, I., Patrinos, G.P., Chrousos, G.P. and Dardiotis, E. (2022) A novel variant in DYNC1H1 could contribute to human amyotrophic lateral sclerosis-frontotemporal dementia spectrum. *Cold Spring Harbor molecular case studies*, **8**.
55. Illarionova, N.B., Borisova, M.A., Bazhenova, E.Y., Zabelina, D.S., Fursenko, D.V. and Kulikov, A.V. (2021) Zbtb33 Gene Knockout Changes Transcription of the Fgf9, Fgfr3, c-Myc and FoxG1 Genes in the Developing Mouse Brain. *Molecular Biology*, **55**, 363-371.
56. Rhee, J.W., Arata, A., Selleri, L., Jacobs, Y., Arata, S., Onimaru, H. and Cleary, M.L. (2004) Pbx3 deficiency results in central hypoventilation. *Am J Pathol*, **165**, 1343-1350.
57. Kunkle, B.W., Vardarajan, B.N., Naj, A.C., Whitehead, P.L., Rolati, S., Slifer, S., Carney, R.M., Cuccaro, M.L., Vance, J.M., Gilbert, J.R. *et al.* (2017) Early-Onset Alzheimer Disease and Candidate Risk Genes Involved in Endolysosomal Transport. *JAMA Neurology*, **74**, 1113-1122.
58. Naruhashi, K., Kadomatsu, K., Igakura, T., Fan, Q.-W., Kuno, N., Muramatsu, H., Miyauchi, T., Hasegawa, T., Itoh, A., Muramatsu, T. *et al.* (1997) Abnormalities of Sensory and Memory Functions in Mice LackingBsgGene. *Biochemical and Biophysical Research Communications*, **236**, 733-737.
59. Najyb, O., Brissette, L. and Rassart, E. (2015) Apolipoprotein D Internalization Is a Basigin-dependent Mechanism. *The Journal of biological chemistry*, **290**, 16077-16087.
60. Ma, A., Wang, X., Wang, C., Li, J., Xiao, T., Wang, J., Li, Y., Liu, Y., Chang, Y., Wang, D. *et al.* (2021) DeepMAPS: Single-cell biological network inference using heterogeneous graph transformer. *Biorxiv*, 2021.2010.2031.466658.
61. Rosén, A., Bergh, A.-C., Gogok, P., Evaldsson, C., Myhrinder, A.L., Hellqvist, E., Rasul, A., Björkholm, M., Jansson, M., Mansouri, L. *et al.* (2012) Lymphoblastoid cell line with B1 cell characteristics established from a chronic lymphocytic leukemia clone by in vitro EBV infection. *OncoImmunology*, **1**, 18-27.
62. Wu, J.Q., Seay, M., Schulz, V.P., Hariharan, M., Tuck, D., Lian, J., Du, J., Shi, M., Ye, Z., Gerstein, M. *et al.* (2012) Tcf7 is an important regulator of the switch of self-renewal and differentiation in a multipotential hematopoietic cell line. *PLoS genetics*, **8**, e1002565.
63. Solomon, L.A., Li, S.K.H., Piskorz, J., Xu, L.S. and DeKoter, R.P. (2015) Genome-wide comparison of PU.1 and Spi-B binding sites in a mouse B lymphoma cell line. *BMC Genomics*, **16**, 76.
64. Torlakovic, E., Tierens, A., Dang, H.D. and Delabie, J. (2001) The Transcription Factor PU.1, Necessary for B-Cell Development Is Expressed in Lymphocyte Predominance, But Not Classical Hodgkin's Disease. *The American Journal of Pathology*, **159**, 1807-1814.

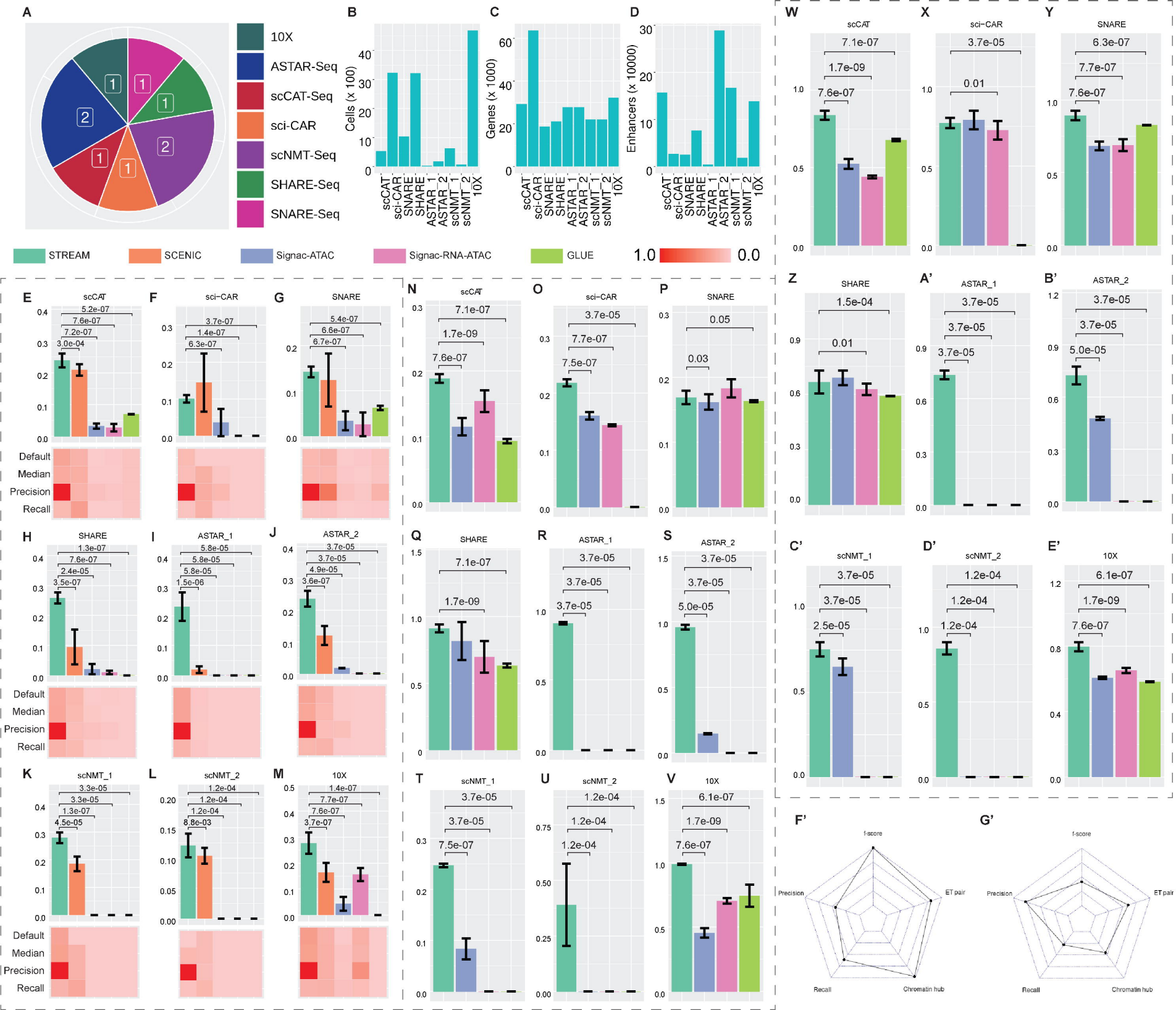
65. Galbiati, M., Lettieri, A., Micalizzi, C., Songia, S., Morerio, C., Biondi, A., Dufour, C. and Cazzaniga, G. (2013) Natural history of acute lymphoblastic leukemia in neurofibromatosis type 1 monozygotic twins. *Leukemia*, **27**, 1778-1781.
66. Scuto, A., Kujawski, M., Kowolik, C., Krymskaya, L., Wang, L., Weiss, L.M., DiGiusto, D., Yu, H., Forman, S. and Jove, R. (2011) STAT3 Inhibition Is a Therapeutic Strategy for ABC-like Diffuse Large B-Cell Lymphoma. *Cancer Research*, **71**, 3182-3188.
67. Huang, X., Meng, B., Iqbal, J., Ding, B.B., Perry, A.M., Cao, W., Smith, L.M., Bi, C., Jiang, C., Greiner, T.C. *et al.* (2013) Activation of the STAT3 signaling pathway is associated with poor survival in diffuse large B-cell lymphoma treated with R-CHOP. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, **31**, 4520-4528.
68. Schmidl, C., Vladimer, G.I., Rendeiro, A.F., Schnabl, S., Krausgruber, T., Taubert, C., Krall, N., Pemovska, T., Araghi, M., Snijder, B. *et al.* (2019) Combined chemosensitivity and chromatin profiling prioritizes drug combinations in CLL. *Nature chemical biology*, **15**, 232-240.
69. Tian, M., Li, Y., Zheng, W., Liu, Q.Q., Zhang, X.L., Liu, J.L., Zhang, S., Sheng, Y.X., Fan, C.B. and Zhang, W.L. (2021) LncRNA PCAT1 enhances cell proliferation, migration and invasion by miR-508-3p/NFIB axis in diffuse large B-cell lymphoma. *European review for medical and pharmacological sciences*, **25**, 2567-2576.
70. Harada, A., Okada, S., Odawara, J., Kumamaru, H., Saiwai, H., Aoki, M., Nakamura, M., Nishiyama, Y. and Ohkawa, Y. (2010) Production of a rat monoclonal antibody specific for Myf5. *Hybridoma (2005)*, **29**, 59-62.
71. Lynch, A.W., Theodoris, C.V., Long, H.W., Brown, M., Liu, X.S. and Meyer, C.A. (2022) MIRA: joint regulatory modeling of multimodal expression and chromatin accessibility in single cells. *Nat Methods*, **19**, 1097-1108.
72. Bernard, S.C., Abdelsamad, E.H., Johnson, P.A., Chapman, D.L. and Parvathaneni, M.J.J.o.C.C.t. (2017) Pediatric leukemia: Diagnosis to treatment–A review. **2**, 1.
73. Penther, D., Viailly, P.J., Latour, S., Etancelin, P., Bohers, E., Vellemans, H., Camus, V., Menard, A.L., Coutant, S., Lanic, H. *et al.* (2019) A recurrent clonally distinct Burkitt lymphoma case highlights genetic key events contributing to oncogenesis. *Genes, chromosomes & cancer*, **58**, 595-601.
74. Kimpara, S., Lu, L., Hoang, N.M., Zhu, F., Bates, P.D., Daenthanasanmak, A., Zhang, S., Yang, D.T., Kelm, A., Liu, Y. *et al.* (2021) EGR1 Addiction in Diffuse Large B-cell Lymphoma. *Molecular cancer research : MCR*, **19**, 1258-1269.
75. Kurihara, Y., Mizuno, H., Honda, A., Shimura, A., Fujioka, Y., Maki, H. and Kurokawa, M. (2022) CCDC88C-FLT3 gene fusion in CD34-positive haematopoietic stem and multilineage cells in myeloid/lymphoid neoplasm with eosinophilia. *Journal of cellular and molecular medicine*, **26**, 950-952.

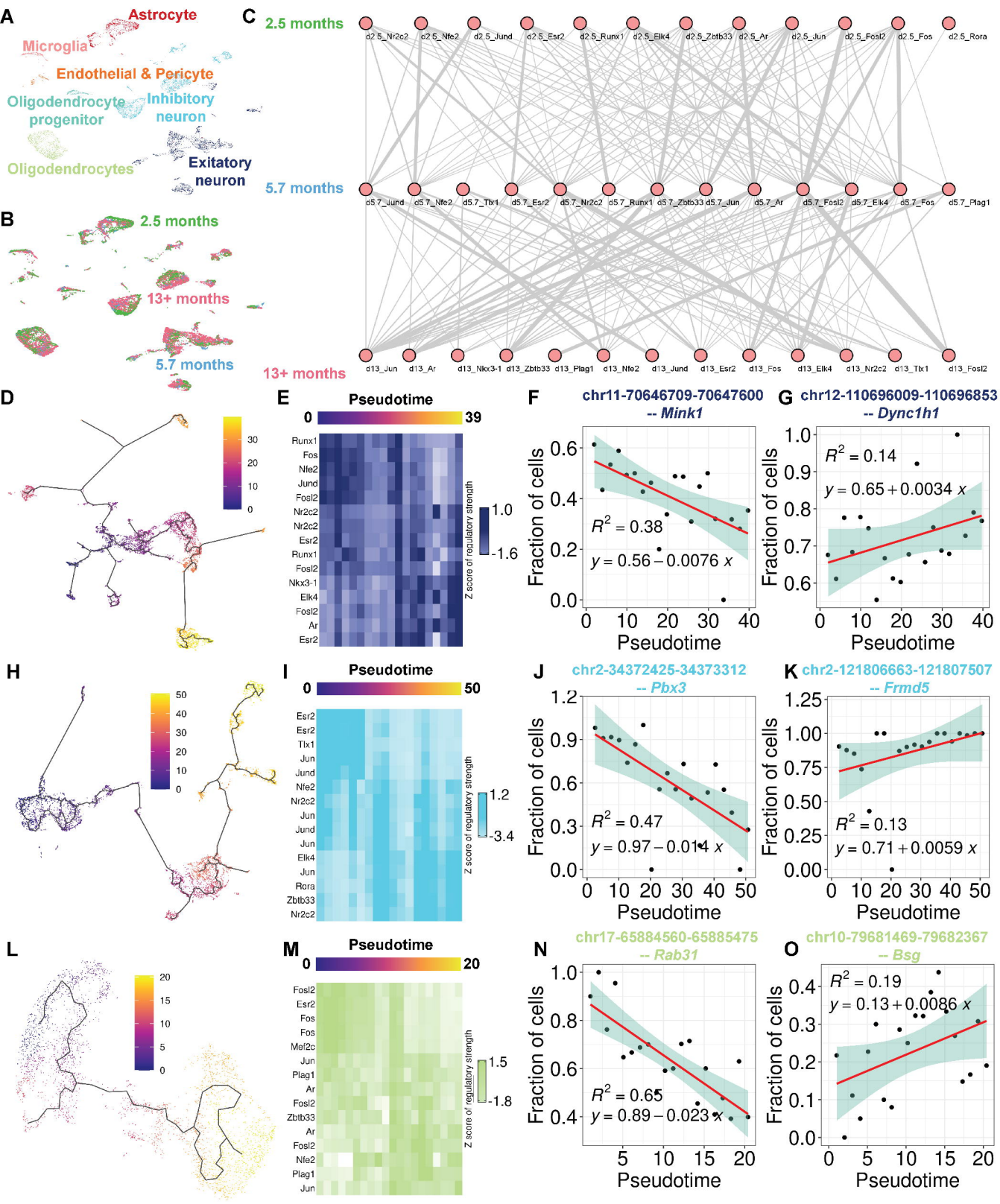
76. Román-Gómez, J., Cordeu, L., Agirre, X., Jiménez-Velasco, A., San José-Eneriz, E., Garate, L., Calasanz, M.a.J., Heiniger, A., Torres, A. and Prosper, F.J.B. (2007) Epigenetic regulation of Wnt-signaling pathway in acute lymphoblastic leukemia. **109**, 3462-3469.
77. Patel, M.S., Kendall, E.K., Ondrejka, S., Mian, A., Sawalha, Y., Hu, B., Hsi, E.D., Hill, B.T. and Gupta, N. (2020) Gene Expression and Epigenetic Analysis in Relapsed/Refractory Diffuse Large B Cell Lymphoma Provides Insights into Evolution of Treatment Resistance to R-CHOP. *Blood*, **136**, 26.
78. Rouillard, A.D., Gundersen, G.W., Fernandez, N.F., Wang, Z., Monteiro, C.D., McDermott, M.G. and Ma'ayan, A. (2016) The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database*, **2016**.
79. Yu, H., Takeuchi, M., LeBarron, J., Kantharia, J., London, E., Bakker, H., Haltiwanger, R.S., Li, H. and Takeuchi, H. (2015) Notch-modifying xylosyltransferase structures support an SNi-like retaining mechanism. *Nature chemical biology*, **11**, 847-854.
80. Lobry, C., Oh, P., Mansour, M.R., Look, A.T. and Aifantis, I. (2014) Notch signaling: switching an oncogene to a tumor suppressor. *Blood*, **123**, 2451-2459.
81. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57-74.
82. Huang, C. and Xie, K. (2012) Crosstalk of Sp1 and Stat3 signaling in pancreatic cancer pathogenesis. *Cytokine & growth factor reviews*, **23**, 25-35.
83. Yang, J., Zhang, L., Jiang, Z., Ge, C., Zhao, F., Jiang, J., Tian, H., Chen, T., Xie, H., Cui, Y. *et al.* (2019) TCF12 promotes the tumorigenesis and metastasis of hepatocellular carcinoma via upregulation of CXCR4 expression. *Theranostics*, **9**, 5810-5827.
84. Tsuzuki, S., Toyama-Sorimachi, N., Kitamura, F., Tsuboi, H., Ando, J., Sakurai, T., Morii, N., Narumiya, S. and Miyasaka, M. (1998) Intracellular Signal-transducing Elements Involved in Transendothelial Migration of Lymphoma Cells. **89**, 571-577.
85. Dudley, N.J. and Balfour, A.H. (1988) Non-Hodgkin's lymphoma presenting as 'chronic active toxoplasmosis'. *Postgraduate medical journal*, **64**, 883-885.
86. Intaraphet, S., Farkas, D.K., Johannesdottir Schmidt, S.A., Cronin-Fenton, D. and Søgaard, M. (2017) Human papillomavirus infection and lymphoma incidence using cervical conization as a surrogate marker: a Danish nationwide cohort study. **35**, 172-176.
87. Liu, Y.C., Yeh, C.T. and Lin, K.H. (2019) Molecular Functions of Thyroid Hormone Signaling in Regulation of Cancer Progression and Anti-Apoptosis. *International journal of molecular sciences*, **20**.
88. Bispo, J.A.B., Pinheiro, P.S. and Kobetz, E.K.J.C.S.H.p.i.m. (2020) Epidemiology and etiology of leukemia and lymphoma. **10**, a034819.
89. Mehravaran, H., Makvandi, M., Samarbaf Zade, A., Neisi, N., Kiani, H., Radmehr, H., Shahani, T., Hoseini, S.Z., Ranjbari, N. and Nahid Samiei, R. (2017) Association of Human Cytomegalovirus with Hodgkin's Disease and Non-Hodgkin's lymphomas. *Asian Pacific journal of cancer prevention : APJCP*, **18**, 593-597.

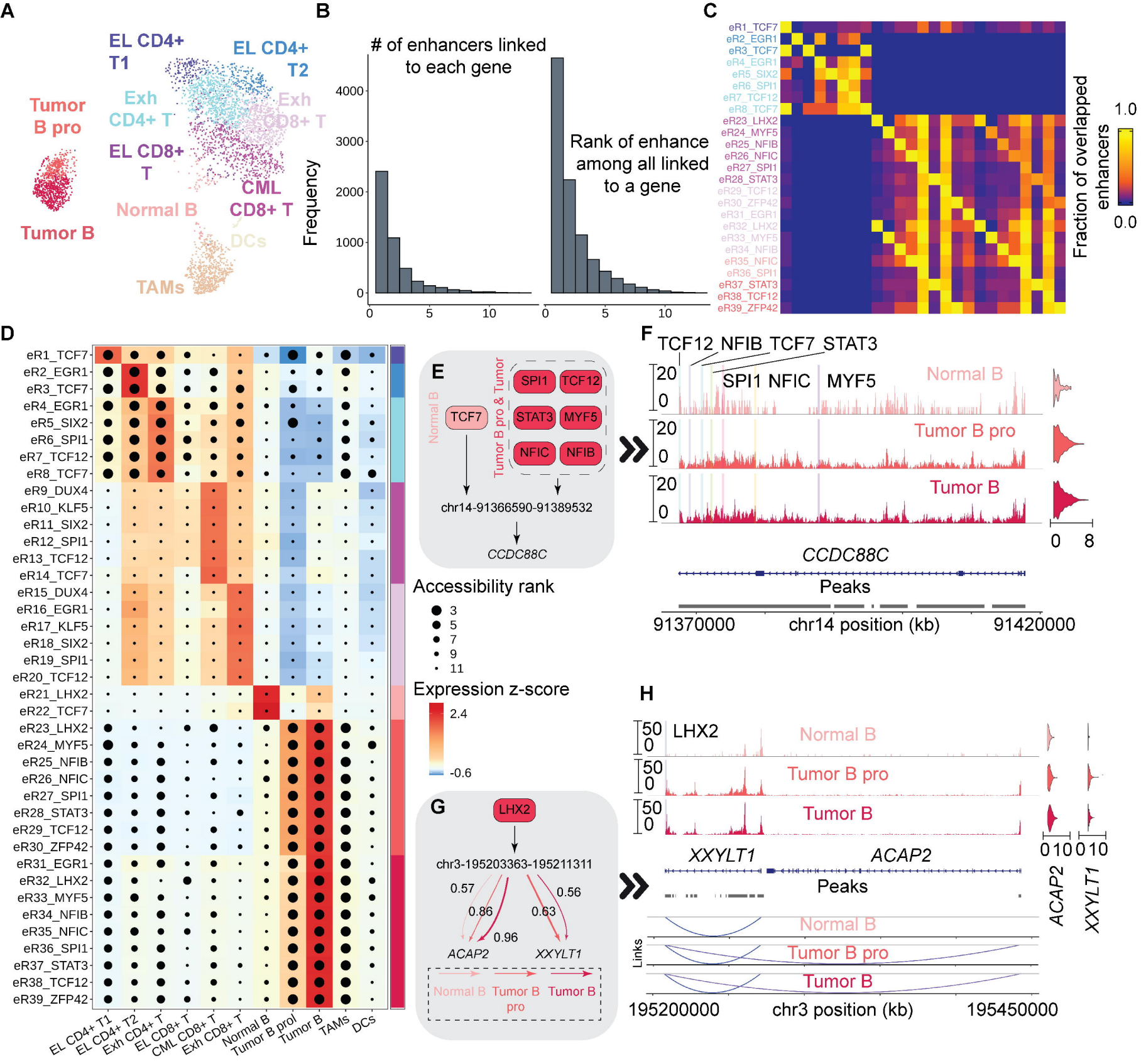
90. Grützmeier, S., Porwit, A., Schmitt, C., Sandström, E., Åkerlund, B. and Ernberg, I. (2016) Fulminant anaplastic large cell lymphoma (ALCL) concomitant with primary cytomegalovirus (CMV) infection, and human herpes virus 8 (HHV-8) infection together with Epstein-Barr-virus (EBV) reactivation in a patient with asymptomatic HIV-infection. *Infectious Agents and Cancer*, **11**, 46.
91. Sato, K., Igarashi, S., Tsukada, N., Inamura, J., Yamamoto, M., Shindo, M., Moriichi, K., Mizukami, Y., Fujiya, M. and Torimoto, Y. (2022) Cytomegalovirus infection in patients with malignant lymphomas who have not received hematopoietic stem cell transplantation. *BMC Cancer*, **22**, 944.
92. Kang, D.W., Choi, K.-Y. and Min, D.S. (2014) Functional Regulation of Phospholipase D Expression in Cancer and Inflammation*. *Journal of Biological Chemistry*, **289**, 22575-22582.
93. Xiong, J., Wang, L., Fei, X.C., Jiang, X.F., Zheng, Z., Zhao, Y., Wang, C.F., Li, B., Chen, S.J., Janin, A. *et al.* (2017) MYC is a positive regulator of choline metabolism and impedes mitophagy-dependent necroptosis in diffuse large B-cell lymphoma. *Blood Cancer Journal*, **7**, e582-e582.
94. Lan, Q., Wang, S.S., Menashe, I., Armstrong, B., Zhang, Y., Hartge, P., Purdue, M.P., Holford, T.R., Morton, L.M., Kricker, A. *et al.* (2011) Genetic variation in Th1/Th2 pathway genes and risk of non-Hodgkin lymphoma: a pooled analysis of three population-based case-control studies. *British journal of haematology*, **153**, 341-350.
95. Zhu, C., Yu, M., Huang, H., Juric, I., Abnousi, A., Hu, R., Lucero, J., Behrens, M.M., Hu, M. and Ren, B. (2019) An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nature Structural & Molecular Biology*, **26**, 1063-1070.

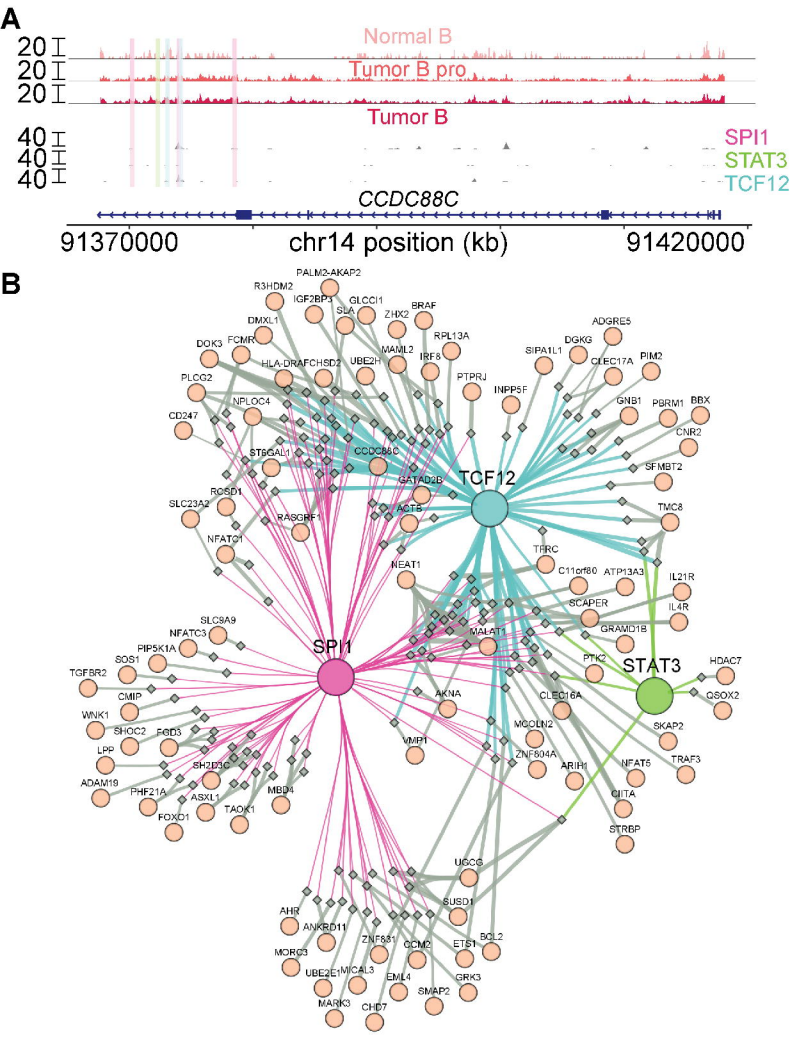

A
Steiner forest problem model
Gene expression
IRIS-FGM
Functional gene module
Genes
scMulti-omics
TF
TSS
Signac
Cicero
Chromatin accessibility
Enhancers
Enh-enh relations
Enh-gene relations
Enh-gene relations
Cells
Genes
Gene expression
Cells
Enhancers
Chromatin accessibility
Hybrid biclustering
HBCs
Submodular optimizatrion
Selected HBCs
TF1
eRegulon1
TF2
eRegulon2
B
Pseudotime-associated eRegulon in AD
Pseudotime
C
Cell-type-specific eGRN in DSLL
Normal B
Tumor B
Tumor B proliferation


A
SFP model
FGM1
FGM2
SFP solution
FGM1
FGM2
Edges for seeding
Edges for TF1-Seed1
Edges for TF2-Seed6
Gene
Enhancer
TF
Enhancer-gene relations
SFP linkages
Enhancer-enhancer relations
B
Sorted edge sets for seeds
TF1-Seed1
TF2-Seed6
Hybrid biclustering
Cells
If HBC score decreases
Iteratively add genes and linked enhancers
Longest path
Directed acyclic graph
HBCs
TF1-HBC1
TF2-HBC6
C
HBCs
Submodular optimization
Submodular function
HBC (n)
If HBC score decrease
eRegulons
Enhancer
TF
Gene


A
10X
ASTAR-Seq
scCAT-Seq
sci-CAR
scNMT-Seq
SHARE-Seq
SNARE-Seq
B
Cells (x 100)
C
Genes (x 1000)
D
Enhancers (x 10000)
scCAT
sci-CAR
SNARE
SHARE
ASTAR_1
ASTAR_2
scNMT_1
scNMT_2
10X
STREAM
SCENIC
Signac-ATAC
Signac-RNA-ATAC
GLUE
1.0
0.0
Default
Median
Precision
Recall
f-score
ET pair
Chromatin hub


A
Astrocyte
Microglia
Endothelial & Pericyte
Oligodendrocyte progenitor
Inhibitory neuron
Oligodendrocytes
Exitatory neuron
B
2.5 months
13+ months
5.7 months
C
2.5 months
5.7 months
13+ months
d2.5_Nr2c2 d2.5_Nfe2 d2.5_Jund d2.5_Esr2 d2.5_Runx1 d2.5_Elk4 d2.5_Zbtb33 d2.5_Ar d2.5_Jun d2.5_Fosl2 d2.5_Fos d2.5_Rora
d5.7_Jund d5.7_Nfe2 d5.7_Tlx1 d5.7_Esr2 d5.7_Nr2c2 d5.7_Runx1 d5.7_Zbtb33 d5.7_Jun d5.7_Ar d5.7_Fosl2 d5.7_Elk4 d5.7_Fos d5.7_Plag1
d13_Jun d13_Ar d13_Nkx3-1 d13_Zbtb33 d13_Plag1 d13_Nfe2 d13_Jund d13_Esr2 d13_Fos d13_Elk4 d13_Nr2c2 d13_Tlx1 d13_Fosl2
D
E
Pseudotime
0
39
Runx1
Fos
Nfe2
Jund
Fosl2
Nr2c2
Nr2c2
Esr2
Runx1
Fosl2
Nkx3-1
Elk4
Fosl2
Ar
Esr2
Z score of regulatory strength
1.0
-1.6
F
chr11-70646709-70647600
-- Mink1
$R^2 = 0.38$
$y = 0.56 - 0.0076\ x$
Fraction of cells
Pseudotime
G
chr12-110696009-110696853
-- Dync1h1
$R^2 = 0.14$
$y = 0.65 + 0.0034\ x$
Fraction of cells
Pseudotime
H
I
Pseudotime
0
50
Esr2
Esr2
Tlx1
Jun
Jund
Nfe2
Nr2c2
Jun
Jund
Jun
Elk4
Jun
Rora
Zbtb33
Nr2c2
Z score of regulatory strength
1.2
-3.4
J
chr2-34372425-34373312
-- Pbx3
$R^2 = 0.47$
$y = 0.97 - 0.014\ x$
Fraction of cells
Pseudotime
K
chr2-121806663-121807507
-- Frmd5
$R^2 = 0.13$
$y = 0.71 + 0.0059\ x$
Fraction of cells
Pseudotime
L
M
Pseudotime
0
20
Fosl2
Esr2
Fos
Fos
Mef2c
Jun
Plag1
Ar
Fosl2
Zbtb33
Ar
Fosl2
Nfe2
Plag1
Jun
Z score of regulatory strength
1.5
-1.8
N
chr17-65884560-65885475
-- Rab31
$R^2 = 0.65$
$y = 0.89 - 0.023\ x$
Fraction of cells
Pseudotime
O
chr10-79681469-79682367
-- Bsg
$R^2 = 0.19$
$y = 0.13 + 0.0086\ x$
Fraction of cells
Pseudotime


A
EL CD4+ T1
EL CD4+ T2
Exh CD4+ T
Exh CD8+ T
Tumor B pro
EL CD8+ T
CML CD8+ T
Normal B
Tumor B
DCs
TAMs
B
# of enhancers linked to each gene
Rank of enhance among all linked to a gene
Frequency
C
Fraction of overlapped enhancers
D
Accessibility rank
Expression z-score
E
TCF7
SPI1
TCF12
STAT3
MYF5
NFIC
NFIB
Normal B
Tumor B pro & Tumor
chr14-91366590-91389532
CCDC88C
F
TCF12 NFIB TCF7 STAT3
SPI1 NFIC MYF5
Normal B
Tumor B pro
Tumor B
CCDC88C
Peaks
91370000
chr14 position (kb)
91420000
G
LHX2
chr3-195203363-195211311
0.57
0.86
0.96
0.56
0.63
ACAP2
XXYLT1
Normal B
Tumor B pro
Tumor B
H
LHX2
Normal B
Tumor B pro
Tumor B
XXYLT1
ACAP2
Peaks
Links
195200000
chr3 position (kb)
195450000


A
Normal B
Tumor B pro
Tumor B
20
20
20
40
40
40
SPI1
STAT3
TCF12
CCDC88C
91370000
chr14 position (kb)
91420000
B
TCF12
SPI1
STAT3
PALM2-AKAP2
R3HDM2
IGF2BP3
GLCCI1
SLA
ZHX2
BRAF
DMXL1
ADGRE5
DOK3
FCMR
MAML2
RPL13A
IRF8
DGKG
SIPA1L1
CLEC17A
PIM2
PLCG2
HLA-DRA
FCHSD2
UBE2H
PTPRJ
INPP5F
NPLOC4
GNB1
PBRM1
BBX
CD247
CNR2
ST6GAL1
CCDC88C
GATAD2B
SFMBT2
RGS1
SLC23A2
RASGRF1
ACTB
TMC8
NFATC1
TRRC
NEAT1
C11orf80
ATP13A3
IL21R
SCAPER
IL4R
SLC9A9
NFATC3
MALAT1
GRAMD1B
PIP5K1A
SOS1
PTK2
HDAC7
TGFBR2
CMIP
AKNA
CLEC16A
QSOX2
WNK1
SHOC2
FGD3
MGOLN2
SKAP2
LPP
WMP1
ZNF804A
ARIH1
NFAT5
TRAF3
SH2D3C
ADAM19
PHF21A
ASXL1
MBD4
CIITA
FOXD1
TADA1
STRBP
UGCG
SUSD1
AHR
ANKRD11
ZNF581
CCM2
ETS1
BCL2
MORC3
UBE2E1
MICAL3
EML4
GRK3
SMAP2
MARK3
CHD7