Briefings in Bioinformatics, 2024, **25(5)**, bbae369
https://doi.org/10.1093/bib/bbae369
Advance access publication date 25 July 2024
Problem Solving Protocol

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

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

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

*Corresponding authors. School of Mathematics, Shandong University, Jinan, Shandong 250100, China. E-mail: bingqiang@sdu.edu.cn; E-mail: Qin.Ma@osumc.edu

Abstract

Deciphering the intricate relationships between transcription factors (TFs), enhancers, and genes through the inference of enhancer-driven gene regulatory networks (eGRNs) is crucial in understanding gene regulatory programs in a complex biological system. This study introduces STREAM, a novel method that leverages a Steiner forest problem model, a hybrid biclustering pipeline, and submodular optimization to infer eGRNs from jointly profiled single-cell transcriptome and chromatin accessibility data. Compared to existing methods, STREAM demonstrates enhanced performance in terms of TF recovery, TF–enhancer linkage prediction, and enhancer–gene relation discovery. Application of STREAM to an Alzheimer's disease dataset and a diffuse small lymphocytic lymphoma dataset reveals its ability to identify TF-enhancer–gene relations associated with pseudotime, as well as key TF-enhancer–gene relations and TF cooperation underlying tumor cells.

Keywords: scRNA-seq; scATAC-seq; data integration; biological network; Steiner forest problem model; submodular optimization

Introduction

Recent single-cell sequencing technologies, such as scRNA-seq and scATAC-seq, have advanced our understanding of gene regulatory networks (GRNs) at single-cell resolution [1]. Approaches like SCENIC use random forest algorithms for GRN construction and cell state identification [2], and DIRECT-NET employs gradient boosting to map *cis*-regulatory element-target gene relationships [3]. Integrating scRNA-seq and scATAC-seq data overcomes these limitations, enhancing GRN inference by reducing noise and improving TF-gene prediction accuracy through regulatory relationship cross-validation [1, 3–19]. This approach broadens motif discovery beyond restricted promoter regions, preserving regulatory sequence diversity. Furthermore, incorporating chromatin accessibility data facilitates the construction of enhancer-driven GRNs (eGRNs) [5], where transcription factor (TF)-target gene connections involve enhancer regions critical for regulation, offering a more comprehensive view of gene regulation mechanisms.

Elucidating eGRNs reveals cell-type-specific and conserved TF regulatory patterns, highlighting the diversity in TF–enhancer and enhancer–gene interactions, which deepens our understanding of gene regulation dynamics [20, 21]. Janssens *et al*., advanced this field by mapping eGRNs for 40 cell types in the fly brain, employing a bioinformatics framework that includes cell clustering, motif discovery, network prediction, and deep learning [5]. A key tool in this progress is SCENIC+ [16], integrating pySCENIC and pycisTopic to predict enhancers, identify regulating TFs, and link enhancers with target genes, utilizing a motif collection of over 30,000 motifs for improved accuracy. SCENIC+ has shown its efficacy across various species and data types, including human peripheral blood mononuclear cells and *Drosophila* retinal development. Additionally, GLUE uses deep learning for eGRN inference from multi-modal single-cell data [4], and Pando models gene expression through TF-peak interactions, demonstrating the diverse applications of these tools in genomic research [14].

Inferencing eGRNs encounters three primary challenges. Firstly, methods like SCENIC+ [16], GLUE [4], and DIRECT-NET [3], Pando [14], and scMEGA [17], which predict enhancer–gene relationships based on accessibility and expression correlations, often lead to high false-positive rates due to not accounting for the interdependence of multiple enhancer and gene interactions. Secondly, biases from initial cell clustering can impact the accuracy of subsequent TF-enhancer–gene relationship predictions. Lastly, the regulatory complexity within cells complicates the extraction of pivotal TF-enhancer–gene relationships, hindered further by the vast array of potential regulatory combinations influencing cell states.

Motivated by the three challenges, we introduce STREAM (Single-cell enhancer regulaTory netwoRk inference from gene Expression And chroMatin accessibility), a computational framework for inferring eGRNs from paired scRNA-seq and scATAC-seq data, utilizing the Steiner forest problem model and submodular optimization (Fig. 1A, Supplementary Note S1, and

Received: April 2, 2024. **Revised:** June 19, 2024. **Accepted:** July 15, 2024

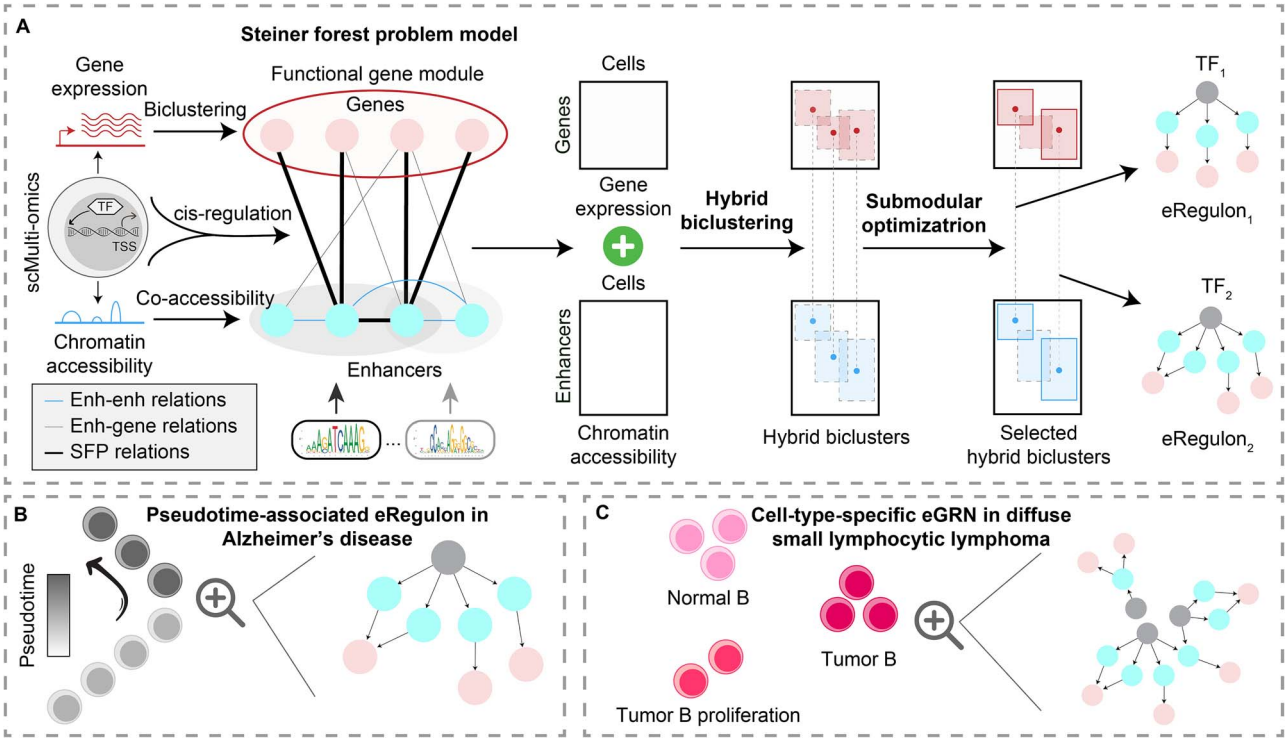

Figure 1. STREAM method overview and its applications in Alzheimer's disease and diffuse small lymphocytic lymphoma. (A) An outline of the STREAM framework for eRegulon identification. An eRegulon comprises a TF, its set of binding enhancers, and target genes. (B) STREAM's application in predicting pseudotime-associated eRegulons in Alzheimer's disease. (C) The use of STREAM for inferring cell-type-specific eGRNs in diffuse small lymphocytic lymphoma. Abbreviations: enh, enhancers; SFP, Steiner forest problem; TSS, transcriptional start site.

Supplementary Note S2). To disentangle the interdependence of multiple enhancer and gene interactions, STREAM constructs a heterogeneous graph that depicts enhancer–enhancer and enhancer–gene relations. This approach leverages the Steiner forest problem model to identify robust relations within context-specific gene modules [22]. To avoid biases from pre-defined cell clusters, STREAM detects hybrid biclusters comprising genes, enhancers, and cells. Each hybrid bicluster includes genes co-regulated by a shared TF through binding to co-accessible enhancers within these cells, thereby eliminating the need for prior cell clustering. To extract pivotal TF–enhancer–gene relationships, STREAM utilizes submodular optimization to prioritize the most representative hybrid biclusters, thereby highlighting key TF–enhancer–gene interactions [23]. Consequently, these interactions within a hybrid bicluster define an enhancer-driven Regulon (eRegulon), forming the foundation of an eGRN for a specific cell subpopulation. The time complexity of these operations is detailed in Supplementary Note S3.

Assessing STREAM on six paired scRNA-seq and scATAC-seq datasets and comparing it with six methods (SCENIC+, SCENIC, GLUE, DIRECT-NET, Pando, and scMEGA), we found STREAM superior in eGRN inference, particularly in TF recovery, TF–enhancer relation identification, and enhancer–gene relation prediction. Its application on Alzheimer's disease (Fig. 1B) and diffuse small lymphocytic lymphoma datasets (Fig. 1C) revealed dynamic eRegulons and underscored STREAM's efficacy in uncovering disease-specific gene regulation. These findings highlight STREAM's robustness in eGRN analysis across complex biological contexts, available for further exploration as an R package on GitHub (https://github.com/OSU-BMBL/STREAM). Meanwhile, there is still much room for improvement in STREAM, including integrating 3-D genome structures, protein data, regulatory perturbations, causality inference, and valid benchmarking based on bulk data.

Materials and Methods

The STREAM framework

Step 1: functional gene module prediction

A functional gene module represents a set of genes exhibiting structured expression patterns, often related or co-regulated within specific cell subpopulations [24]. To identify these modules, STREAM uses gene expression ($X_{n\times o}$) and chromatin accessibility ($Y_{m\times o}$) matrices, post-quality control, indicating n genes' expression and m peaks' accessibility across o cells, respectively (Fig. 2A and Supplementary Fig. S1A). By transforming $X_{n\times o}$ into a discretized matrix $X'_{n\times o}$ using a left-truncated mixture Gaussian model [25], it captures diverse gene expression regulated by transcriptional inputs. sRNA-seq analysis on $X'_{n\times o}$ identifies biclusters B_k ($k = 1, \ldots, l$), where each gene set represents a functional gene module, and each cell set denotes the cells in which the functional gene module is active, preparing for Step 2.

Step 2: steiner forest problem model

The STREAM methodology employs a Steiner forest problem model to deduce eGRNs by identifying enhancer–gene and enhancer–enhancer relationships conducive to gene co-expression in functional gene modules (Fig. 2A and Supplementary Fig. S1B) [24]. For each module B_k, discrete expression (X'_k) and chromatin accessibility (Y_k) submatrices are constructed, restricted to the cells of B_k. The rows of X'_k correspond to the genes of B_k. Using Signac and Cicero [26, 27], a heterogeneous graph $G^{(k)} = (V^{(k)}, E^{(k)})$ is constructed. Nodes $V^{(k)}$ represent genes/enhancers from X'_k and Y_k, and edges $E^{(k)}$ indicate enhancer–gene *cis*-regulatory and enhancer–enhancer co-accessibility

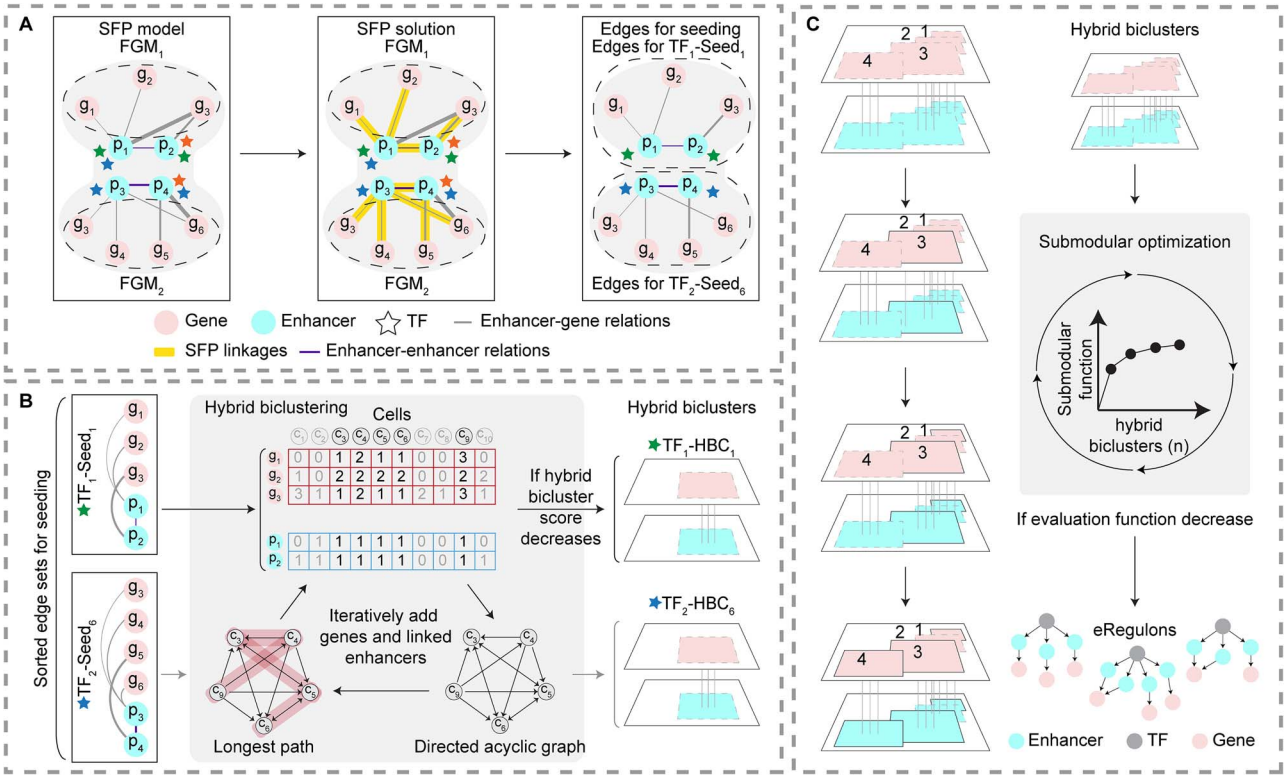

Figure 2. Detailed overview of the STREAM framework. (A) The Steiner forest problem model is utilized to extract highly confident enhancer–gene relations. (B) The hybrid biclustering pipeline is used for the identification of hybrid biclusters. (C) The construction of eRegulons using a submodular optimization approach based on hybrid biclusters. Abbreviations: FGM, functional gene module; SFP, Steiner forest problem.

connections derived from these matrices. Edge weights are quantified using Pearson correlation coefficients (enhancer–gene) and covariances (enhancer–enhancer) as defined by Signac and Cicero, respectively. Additionally, partial correlation is used as an alternative metric to provide a more nuanced understanding of relationships by accounting for the influence of other variables [28]. Edge costs are calculated by subtracting the min-max normalized edge weight from one. Nodes in $V^{(k)}$ representing genes are segmented into subsets $V_s^{(k)}$ ($s = 1, \ldots, t$), termed terminal nets, where nodes within the same subset belong to the same connected component in $G^{(k)}$.

Within the weighted undirected graph $G^{(k)}$ and t terminal nets $V_s^{(k)}$, the Steiner forest model seeks a minimum-cost forest $F^{(k)}$, connecting nodes within each $V_s^{(k)}$ via forest edges. Given the Steiner forest problem's NP-hard nature, a heuristic strategy is adopted. $F^{(k)}$ comprises multiple trees $T_s^{(k)}$, linking nodes in each terminal net. Identifying $F^{(k)}$ involves finding each $T_s^{(k)}$, starting with the gene pair from $V_s^{(k)}$ that are co-regulated across the maximum number of cells. The shortest path between them initiates $T_s^{(k)}$. Subsequently, the gene outside $T_s^{(k)}$ closest to it is iteratively incorporated, extending $T_s^{(k)}$ until it includes $V_s^{(k)}$. All $T_s^{(k)}$, for $s = 1, \ldots, t$, merge to form $F^{(k)}$

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

Within $F^{(k)}$, enhancer–gene relations are selected and grouped by TF binding sites downloaded from JASPAR [29], associating each subset with enhancers bound by the same TF. Identifying TF–enhancer relations using experimentally verified TF binding sites curated in JASPAR, instead of performing a motif scan, reduces computational resources and time while decreasing false positives. Only subsets linked to ≥ 2 genes are retained. Each enhancer–gene subset is converted into a hybrid bicluster, comprising gene, enhancer, and cell subsets, denoted as $I_i^{(k)}$, $J_i^{(k)}$, and $K_i^{(k)}$

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

In each hybrid bicluster $H_i^{(k)}$, cells in $K_i^{(k)}$ are ranked by descending average expression of genes in $I_i^{(k)}$, creating trend $M_i^{(k)}$. For each enhancer in $J_i^{(k)}$, we compute the average ratio $r_i^{(k)}$ of cells where the enhancer is accessible compared to all cells in $K_i^{(k)}$. The hybrid bicluster score is defined as the minimum of $\left|I_i^{(k)}\right|$ and $\left|K_i^{(k)}\right|$, used to rank biclusters. All hybrid biclusters are saved in S, serving as seeds for hybrid biclustering in Step 3,

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

Step 3: hybrid biclustering

Starting with seeds from Step 2, STREAM applies hybrid biclustering to identify co-regulated genes and co-accessible enhancers within specific cell subpopulations. Expanding from a seed in set S, we grow the hybrid bicluster vertically (genes and enhancers) and horizontally (cells), maintaining a monotonic expression trend (Fig. 2B and Supplementary Fig. S1C). This process, akin to simultaneously biclustering two matrices, is computationally intensive due to its NP-hard nature [30], necessitating a heuristic approach. Expansion ceases when no further growth is possible, resulting in the final hybrid bicluster and delete this seed from S. The process stops if S is empty;

otherwise, we evaluate the first seed in S for eligibility; unqualified seeds are removed. Qualified seeds are expanded via the longest path identification in directed acyclic graphs, representing gene expression monotonicity. A seed qualifies if it shares less than an α (0.5 by default) proportion of submatrix defined by genes and cells with prior hybrid biclusters (Supplementary Fig. S2).

Selecting the seed $H_i^{(k)}$ with the maximum hybrid bicluster score from S, a directed acyclic graph $\hat{G}_i^{(k)} = (\hat{V}_i^{(k)}, \hat{E}_i^{(k)})$ is constructed, with nodes for cells in $K_i^{(k)}$ and directed edges from higher- to lower-ranking cells according to the trend $M_i^{(k)}$. To prevent infinite paths, looping nodes condense into one, and the hybrid bicluster score is defined as

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

We define the candidate gene set as all genes excluding those in $I_i^{(k)}$ and including those linked to at least one enhancer bound by the same TF, referred to as supporting enhancers. For each candidate gene g, we refine $\hat{G}_i^{(k)}$ to build another directed acyclic graph $\tilde{D}_i^{(k)} = (\tilde{V}_i^{(k)}, \tilde{E}_i^{(k)})$ by excluding non-expressing cells, edges violating $M_i^{(k)}$, and genes without accessible supporting enhancers in $\geq r_i^{(k)} \cdot |\tilde{V}_i^{(k)}|$ cells. Using dynamic programming, we identify the gene yielding the longest path $\tilde{P}_i^{(k)}$ from $\tilde{D}_i^{(k)}$ [31]. We select the candidate gene g yielding the longest path and form a new hybrid bicluster

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

The updated hybrid biclusters $\tilde{H}_i^{(k)}$ are constructed by: (i) adding gene g to $I_i^{(k)}$ to form $\tilde{I}_i^{(k)}$; (ii) if g's supporting enhancer is not in $J_i^{(k)}$, add the enhancer, accessible in the maximum number of cells of $\tilde{V}_i^{(k)}$, e to $J_i^{(k)}$ to form $\tilde{J}_i^{(k)}$, otherwise, $J_i^{(k)}$ remains unchanged; (iii) constructing $\tilde{K}_i^{(k)}$ with cells from the longest path $\tilde{P}_i^{(k)}$. If $\tilde{H}_i^{(k)'}$s score is $\geq H_i^{(k)'}$s, assign $H_i^{(k)}$ as $\tilde{H}_i^{(k)}$. If not, end the expansion for this seed, output $H_i^{(k)}$, and remove the seed from S. If S is not empty, proceed with the next seed.

Finally, hybrid biclusters are ranked by descending score, and within each, enhancer–gene relations ($C_i^{(k)}$) link genes to enhancers within 250 kb of their TSS [21]. Each hybrid bicluster is denoted as

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

We send all the hybrid biclusters for optimization in Step 4.

Step 4: hybrid bicluster optimization

Hybrid bicluster optimization seeks a subset that enhances diversity and minimizes redundancy, outputting eRegulons. This involves three elements (Fig. 2C and Supplementary Fig. S1D): (i) an evaluation function determining eRegulon number, (ii) a submodular objective function measuring hybrid biclusters' informativeness, and (iii) a submodular optimization algorithm selecting highly ranked hybrid biclusters [23].

eRegulon number identification. In our model, the evaluation function is formulated as follows:

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

R^{in} represents the number of enhancer–gene pairs linked in selected hybrid biclusters, while R^{out} counts the number of pairs within 250 kb without established connections in selected hybrid biclusters.

Submodular function. Using the facility location function, we quantify the data fraction in the entire set $\mathcal{U}$ captured by subset $\mathcal{W}$ of hybrid biclusters [23], where f maps $\mathcal{U}$'s power set to real numbers

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

where $r_{H',H}$ determines the pairwise similarity between hybrid biclusters $H' = (I', J', K', C')$ and $H = (I, J, K, C)$, and is given by

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

Intuitively, the facility location function achieves a high value when every hybrid bicluster in $\mathcal{U}$ has at least one representative in $\mathcal{W}$ that is similar. We define the conditional gain of f as:

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

Submodular optimization. Submodular optimization starts with an empty set $\mathcal{W}_0 = \varnothing$ and iteratively selects a hybrid bicluster H_i maximizing conditional gain, updating $\mathcal{W}_i$ to $\mathcal{W}_{i-1} \cup \{H_i\}$. The process ends when there is no unselected hybrid bicluster. Finally, select $\mathcal{W}_i$ that yields the maximum ΔR as the set of eRegulons.

Benchmarks

Datasets

This study benchmarks against six datasets from human cell lines, combining scRNA-seq and scATAC-seq data, accessible through NCBI GEO, 10x Genomics, or literature (Supplementary Table S1). Since the evaluation of performance depends on TF ChIP-seq and chromatin interaction data from the same cell lines in ENCODE, we downloaded all jointly profiled datasets and retained those with both supporting TF ChIP-seq and chromatin interaction data from the same cell lines (Supplementary Table S2-S3). Therefore, technologies used include 10x Genomics Multiome, scCAT-seq [32], SHARE-seq [33], and SNARE-seq2 [34]. Case studies feature a 10x Genomics Multiome dataset from an Alzheimer's mouse model at various ages and another dataset from a diffuse small lymphocytic lymphoma model with ~14,000 sorted nuclei.

Preprocessing scRNA-seq and scATAC-seq datasets

This study processed scRNA-seq and scATAC-seq data using Read10X_h5, read_10x, and read.table functions for loading. scRNA-seq matrices were transformed into Seurat objects with CreateSeuratObject (Seurat v.4.0.5), and mitochondrial RNA percentages calculated via PercentageFeatureSet. For scATAC-seq, we retained enhancers on standard chromosomes (standardChromosomes) and annotated genomes using GetGRangesFromEnsDb (EnsDb.Hsapiens.v86 for hg38, EnsDb.Hsapiens.v75 for hg19). Common cells across matrices were identified using intersect. scATAC-seq fragments in 10x Genomics Multiome datasets were integrated into Seurat with CreateChromatinAssay. Quality control was performed using subset based on mitochondrial RNA content and counts.

Benchmark methods

We compared STREAM against six (e)GRN inference methods (Supplementary Note S4): (i) SCENIC (pySCENIC v.0.12.1) for fast TF, GRN, and cell type deduction from scRNA-seq [2]; (ii) SCENIC+ v.1.0.1 for constructing eGRNs from scRNA-seq and scATAC-seq datasets [16]; (iii) GLUE v.0.3.2, a deep learning framework for regulatory interaction inference from scRNA-seq and scATAC-seq [4]; (iv) DIRECT-NET v.1.0.0, for GRN construction from scRNA-seq and scATAC-seq [3]; (v) Pando v.1.0.5, using multi-modal single-cell data for GRN inference [14]; (vi) scMEGA v.1.0.1, inferring GRNs with Seurat, Signac, and ArchR [17].

Evaluation metrics of eGRN inference

We evaluate the effectiveness of various eGRN/GRN identification techniques, and examined their performance following SCENIC+ using three perspectives: TF recovery, TF–enhancer relation prediction, and enhancer–gene relation discovery [16].

TF recovery. To assess TF identification accuracy, we obtained TF ChIP-seq bed files for six cell lines from ENCODE (Supplementary Table S2). We analyzed method accuracy by identifying overlaps between TFs in the (e)GRNs and ENCODE data, considering these overlaps as true positives. *f* scores were calculated to compare TF recovery effectiveness across methods, aiming to comprehensively evaluate each method's ability to accurately identify and recover TFs for the specified cell lines.

TF–enhancer relation prediction. To evaluate the accuracy of predicted TF–enhancer associations by benchmarked methods, we sourced TF ChIP-seq bed files for six cell lines from ENCODE (Supplementary Table S3). For comprehensiveness, we selected the bed file with the most signal peaks per TF. We then compared the predicted TF-binding enhancers from our methods to the ENCODE ChIP-seq peaks, employing precision as our metric. This approach allowed us to assess the accuracy and relevance of our TF–enhancer predictions, emphasizing quality and significance over sheer quantity.

Enhancer–gene relation discovery. To assess the accuracy of inferred enhancer–gene connections by various methods, we utilized chromatin interaction, e.g., Hi-C data, for six cell lines from ENCODE, selecting the largest chromatin contact matrix in .hic format for each. Using strawr v.0.0.91, we converted .hic files to contact matrices with 2500 kb bins, indicating chromatin contact frequency. Gene locations were determined using gene annotations from EnsDb.Hsapiens.v86 (hg38) and EnsDb.Hsapiens.v75 (hg19). We compared our predicted enhancer–gene associations with ENCODE Hi-C contact data, using *f* scores to quantify the prediction accuracy of enhancer–gene connections.

Case studies

In the case study of Alzheimer's disease, cell-type-specific trajectories from scRNA-seq and scATAC-seq data were inferred using Monocle3 v.1.0.0, designating 2.5-month stage cells as trajectory roots [35]. In the case study of diffuse small lymphocytic lymphoma, we created cell-type-specific eGRNs for diffuse small lymphocytic lymphoma via a two-step approach. Initially, cell-type-specific eRegulons were identified using a hypergeometric test ($p < 0.05$), adjusted for multiple tests with the Bonferroni correction. An eRegulon was considered cell-type-specific if its active cell set was significantly enriched in that cell type, merging eRegulons under the same TF for each cell type. Subsequently, we assembled the eGRN by integrating these eRegulons and their TF–enhancer–gene links. Differentially expressed genes or differentially accessible regions were identified using Seurat v.4.0.5's FindMarkers function, with significance set at adjusted *p*-values < 0.05.

Results

Benchmark evaluations of STREAM

We benchmarked STREAM using six datasets from 10x Genomics Multiome, scCAT-seq [32], SHARE-seq [33], and SNARE-seq [34], covering cell lines like bone marrow and K562 (Fig. 3A). Our goal was to validate predictions on TFs, enhancers, and their interactions. Unlike SCENIC+'s simulated approach, we used real datasets to better understand method efficacy and sequencing nuances [16].

We compared STREAM to six (e)GRN construction tools: SCENIC+ [16], SCENIC [2], GLUE [4], DIRECT-NET [3], Pando [14], and scMEGA [17] (see Materials and Methods for details). STREAM identified 143–159 TFs across datasets, comparable to SCENIC and GLUE's 229–356, but higher than SCENIC+'s 36, DIRECT-NET's 89, Pando's 147–407, and scMEGA's 15–84 (Fig. 3B and Supplementary Table S1). SCENIC+ struggled with detecting differentially accessible regions, affecting eRegulon inference. STREAM covered 70.4–78.3% of JASPAR-confirmed TFs (203) [29]. GLUE and SCENIC reported the highest TF counts due to their methodology, often exceeding STREAM. On average, STREAM identified 124 genes per (e)regulon, contrasting with SCENIC+'s 9, DIRECT-NET's 28, Pando's 2, SCENIC's 486, GLUE's 157, and scMEGA's 403 (Fig. 3C). Additionally, STREAM found an average of 69 enhancers per eRegulon, against SCENIC+'s 10, GLUE's 184, and Pando's 3 (Fig. 3D).

We evaluated the biological relevance of identified TFs against 690 ENCODE TF ChIP-seq datasets for the same cell lines (Figs. 3E-3J and Supplementary Table S2). STREAM showed the highest recovery rate of TFs regulating (e)regulons, as indicated by its *f* score. DIRECT-NET excelled in 10x Genomics Multiome datasets, with SCENIC and GLUE performing similarly across various platforms. SCENIC+ detected a limited number of eRegulons (36) in the 10x Genomics Multiome dataset (Fig. 3E), none of which matched ENCODE ChIP-seq TFs. While DIRECT-NET was strong on the 10x platform, its performance dropped on others (Figs. 3E-3J). Pando stood out in scCAT-seq analysis, surpassing other methods (Fig. 3F).

We assessed the accuracy of predicted TF target enhancers against ENCODE TF ChIP-seq data specific to the same cell line, excluding SCENIC, DIRECT-NET, and scMEGA for their lack of TF–enhancer inference. STREAM led in precision (Figs. 3K-3P), with GLUE close behind. SCENIC+ failed to align with ENCODE's TF binding peaks due to non-detection of TFs (Fig. 3K). Consistent with TF recovery, Pando excelled in the scCAT-seq dataset analysis (Fig. 3L).

In our final analysis of enhancer–gene relationship accuracy against Hi-C data, we focused on STREAM, GLUE, and SCENIC+, excluding SCENIC, DIRECT-NET, and Pando (Supplementary Table S3). STREAM led with *f* scores of 0.4 to 0.5 across datasets (Figs. 3Q-3V), while GLUE followed with 0.3 to 0.4. SCENIC+ had an *f* score of 0.12 in the 10x Genomics Multiome dataset, indicative of its limited enhancer–gene detection (Fig. 3Q). Consistently, Pando excelled in the scCAT-seq dataset, aligning with previous findings on TF recovery and TF–enhancer predictions (Fig. 3R).

In conclusion, STREAM excels in eRegulon inference across cell lines and sequencing techniques, outshining rivals in TF recovery, TF–enhancer relationship prediction, and enhancer–gene connection identification. Its success stems from globally optimizing enhancer–gene interactions and predicting co-regulated gene and

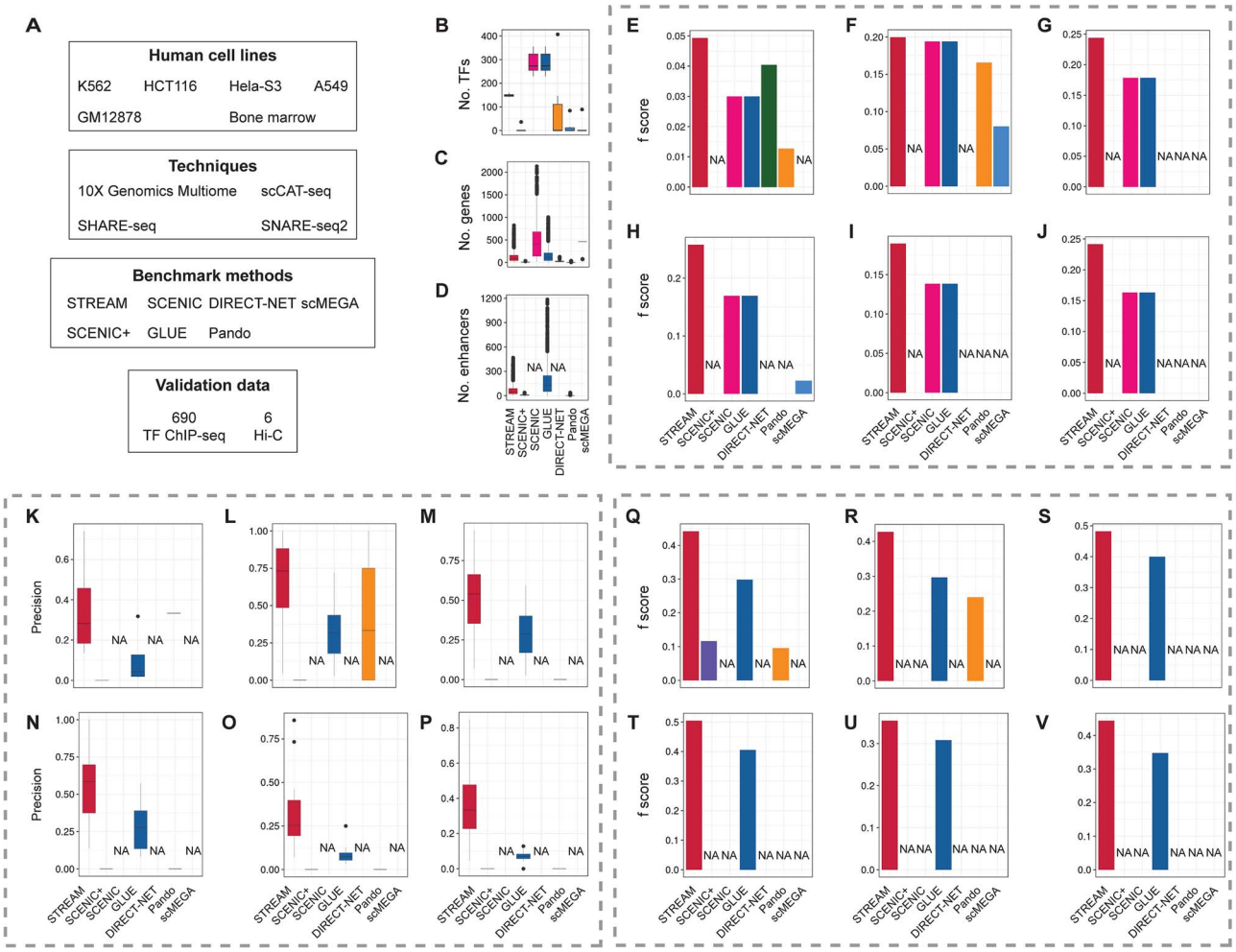

Figure 3. Evaluation of STREAM in comparison to other single-cell RNA-seq and ATAC-seq GRN inference methods using mainstream jointly profiled data. (A) Chart of benchmarking strategy. (B-D) Number of TFs (B) identified per method and distributions of the number of target genes (C), and enhancers per regulon and method (D). (E-J) *f* score distributions from the comparison of TF recovery per method for datasets of human bone marrow (10x Multiome, E); mixture of K562, HCT116, and Hela-S3 (scCAT-seq, F); GM12878 (SHARE-seq, G-H); A549 (SNARE-seq2, I); and GM12878 (SNARE-seq2, J). (K-P) Overlap between Hi-C links and predicted enhancer–gene relations per method for datasets of human bone marrow (10x Multiome, K); mixture of K562, HCT116, and Hela-S3 (scCAT-seq, L); GM12878 (SHARE-seq, M-N); A549 (SNARE-seq2, O); and GM12878 (SNARE-seq2, P). (Q-V) *f* score distributions from the comparison of regulon target genes, per method for datasets of human bone marrow (10x Multiome, K); mixture of K562, HCT116, and Hela-S3 (scCAT-seq,L); GM12878 (SHARE-seq, M-N); A549 (SNARE-seq2, O); and GM12878 (SNARE-seq2, P).

enhancer pairs (Supplementary Table S4). STREAM's comprehensive approach, leveraging the Steiner forest problem model and submodular optimization, offers a unified view of TF–enhancer and enhancer–gene relationships, enhancing the accuracy and depth of regulatory network mapping.

STREAM reveals pseudotime-linked eRegulons and dynamic enhancer–gene relationships in Alzheimer's disease trajectories

Using STREAM, we analyzed a scRNA-seq and scATAC-seq dataset from an Alzheimer's disease mouse model (n = **21,374** cells, **32,286** genes, **66,861** enhancers) across three stages (2.5, 5.7, and 13+ months) generated by 10x Genomics Multiome. We identified 27 cell clusters using Seurat v.4.0.5 and manually annotated seven cell types: oligodendrocytes, oligodendrocyte progenitors, inhibitory neurons, excitatory neurons, astrocytes, microglia, and endothelial & pericytes (Supplementary Table S5 and Figs. 4A-4B). STREAM revealed 81 eRegulons linked to Alzheimer's TFs, including androgen receptor [36], JUN [37], ESR2 [38, 39], FOSL2 [40], PLAG1 [41], RUNX1 [42], RORA [43], and STAT2 [44]. We assessed eRegulon overlap across stages (Fig. 4C), noting similarities within and across different TF-regulated eRegulons. Specifically, in excitatory neurons, inhibitory neurons, and oligodendrocytes, we found 18, 11, and 17 cell-type-specific eRegulons, respectively, highlighting STREAM's capacity to reveal temporal dynamics in Alzheimer's disease progression.

To understand the temporal dynamics of eRegulon regulatory strengths in excitatory neurons, we isolated excitatory neuron cells, mapped their developmental trajectory (Fig. 4D), and computed pseudotime. We quantified each eRegulon's enhancer–gene regulatory strength as the cell proportion showing accessible enhancers and gene expression within pseudotime segments, normalizing these values into z-scores to identify temporal patterns (Fig. 4E). Some eRegulons exhibited monotonically changing regulatory strengths across pseudotime, with trends of both increase and decrease. Specifically, eRegulons under RUNX1, FOS, NFE2, JUND, and FOSL2 showed diminishing expression, indicating strong early pseudotime activity. In contrast, regulatory strengths for NR2C2, ESR2, RUNX1, and FOSL2 peaked

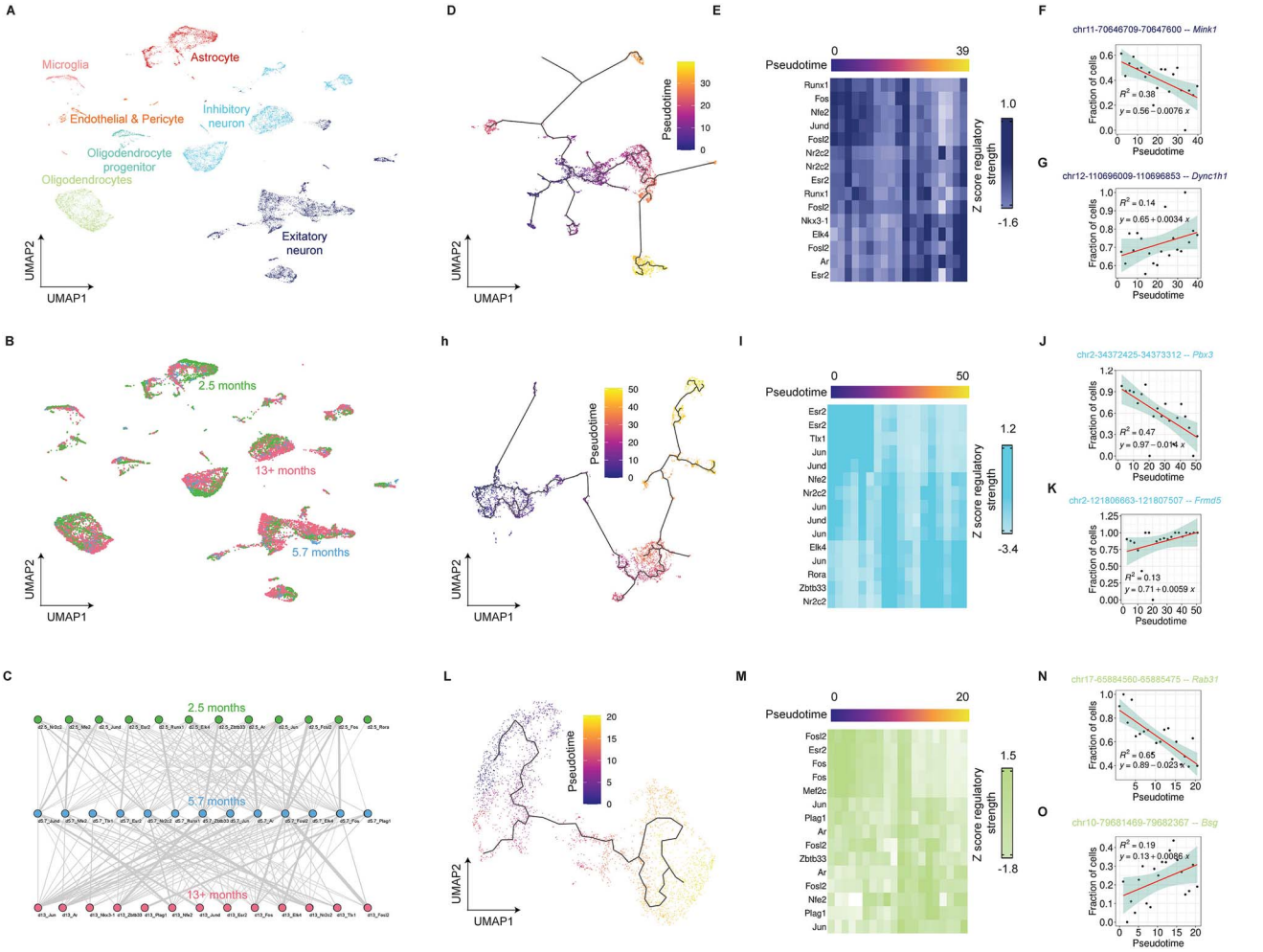

Figure 4. Analysis of pseudotime-associated eRegulons and changing trends of enhancer–gene relations. (A-B) UMAP plots color-coded by cell types (A) and stages (B). (C) Graph visualizing eRegulon similarity identified at three stages, with nodes representing eRegulons and weighted edges indicating pairwise similarity defined as the Jaccard index of enhancer–gene relations of two eRegulons. (D) UMAP plot with pseudotime color-coding in excitatory neuron cells. (E) Heatmap showing mean regulatory strengths of enhancer–gene relations in eRegulons specific to excitatory neuron cells over pseudotime. (F-G) Exemplary enhancer–gene relations demonstrating a monotonic trend in regulatory strengths over pseudotime in excitatory neuron cells. Similar plots for Inhibitory neuron cells (H-K) and oligodendrocytes (L-O).

mid-trajectory, while NKX3–1, ELK4, FOSL2, androgen receptor, and ESR2 eRegulons increased over pseudotime. These patterns align with the neurological roles of these TFs, such as RUNX1's pro-neurogenic function [45], androgen receptor's link to cognitive deficits [46], JUN and FOS's involvement in apoptosis [47, 48], ESR2's Alzheimer's disease susceptibility [38], JUND's apoptotic impact, ELK4's and FOSL2's (AP-1 component) neuronal functions [47–49]. Notably, the enhancer bound by FOS and FOLS2 on chr11–70646709–70647600 showed decreasing regulatory strength on Mink1 (Fig. 4F), important for cognition and Alzheimer's disease. Similarly, an enhancer regulated by ELK4 and ESR2 on chr12–110696009–110696853 exhibited growing influence on Alzheimer's-associated gene Dync1h1 (Fig. 4G) [38, 50], highlighting STREAM's ability to capture significant regulatory changes over time in disease progression.

Our analysis extended to inhibitory neurons, revealing monotonic trends in eRegulon regulatory strengths similar to those in excitatory neurons (Figs. 4H-4I). eRegulons controlled by ESR2, TLX1, JUN, JUND, NFE2, NR2C2, ELK4, RORA, and KAISO (Zbtb33) exhibited distinct patterns. Notably, RORA [43], associated with Alzheimer's disease pathology, showed increased expression in disease contexts. KAISO plays a role in central nervous system development. ESR2 and JUN [51], binding to chr2–34372425–34373312, regulated Pbx3 [52], essential for the central nervous system, with diminishing regulatory strength over pseudotime (Fig. 4J). Conversely, ELK4 and JUN enhanced Frmd5's regulatory impact from chr2–121806663–121807507, indicating an increasing influence over time (Fig. 4K).

In oligodendrocytes, we observed eRegulons with monotonic changes in regulatory strength over pseudotime (Fig. 4L), controlled by TFs similar to those in excitatory and inhibitory neuron analyses (Fig. 4M). We explored two enhancers: chr17–65884560–65885475, regulated by FOSL2 and FOS, showed decreasing regulatory influence on Rab31 (Fig. 4N) [53], a RUNX1 target in Alzheimer's. Meanwhile, chr10–79681469–79682367, under androgen receptor and JUN (Fig. 4O), revealed increasing impact on Bsg, linked to learning, memory [54], and potential sensory and memory function abnormalities [55].

STREAM elucidates the role of eRegulons or enhancer–gene relationships in developmental trajectories across excitatory neurons, inhibitory neurons, oligodendrocytes, microglia, astrocytes, and oligodendrocyte progenitors (Figs. 4D-4O **and** Supplementary Figs. S3, S4, and S5), showcasing its ability to infer critical regulatory dynamics.

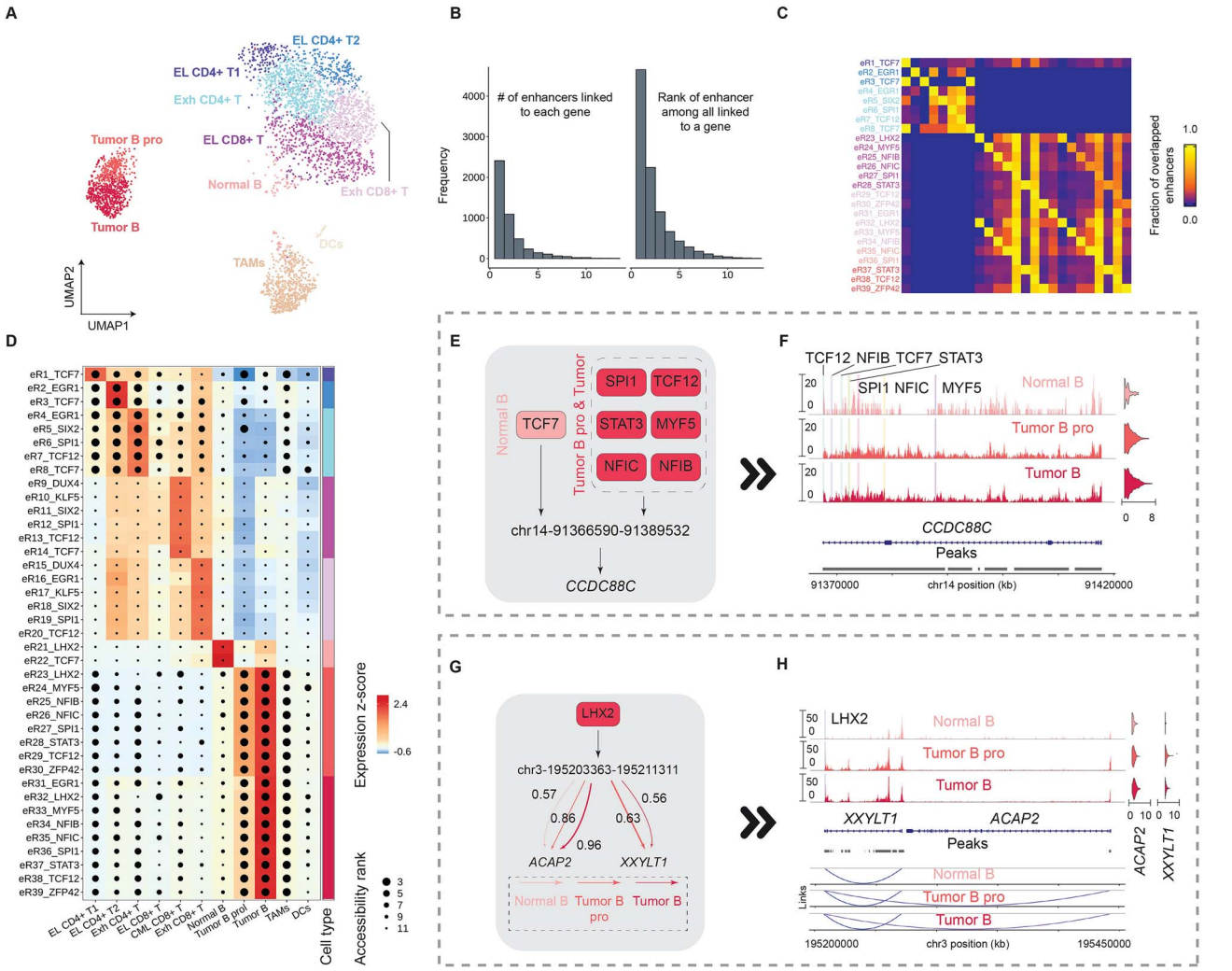

Figure 5. Key enhancer and enhancer–gene relations in B cell tumor development revealed by STREAM eRegulons. (A) UMAP plot of cell types in diffuse small lymphocytic lymphoma. (B) Distribution of the number of enhancers linked to each gene and rank distribution based on absolute distance between enhancer and gene. (C) Overlap fraction of enhancers between eRegulon pairs, normalized by enhancer count in each row's eRegulon. (D) Heatmap-dotplot displaying eRegulon gene expression (color scale) and ranks of chromatin accessibility (size scale). (E) Schematic showcasing TF variation regulating CCDC88C via chr14–91366590-91389532 binding among Normal B, Tumor B proliferation, and Tumor B cells. (F) Chromatin accessibility profiles across three B cell types within chr14–91370000–91420000, labeled by TF binding sites from (E). (G) Schematic highlighting enhancer–gene relation variation between chr3–195203363–195211311 and gene pair (ACAP2 and XXYLT1) across three B cell types. (H) Chromatin accessibility profiles within chr3–195203363–195211311 across three B cell types, labeled by LHX2 binding sites. Arcs show enhancer–gene links with color denoting the fraction of cells showing concurrent enhancer accessibility and gene expression. Abbreviations: Tumor B pro, Tumor B proliferation cell; TAM, Tumor-associated macrophage; DC, Dendritic cell; EL CD4+ T, Effector-like CD4+ T cell; EL CD4+ T1, Effector-like CD4+ T cell type 1; EL CD4+ T2, Effector-like CD4+ T cell type 2; Exh CD4+ T, Exhaustive CD4+ T cell; EL CD8+ T, Effector-like CD8+ T cell; Exh CD4+ T, Exhaustive CD4+ T cell.

STREAM reveals eRegulons in diseased B cells of diffuse small lymphocytic lymphoma

Demonstrating STREAM's utility in cancer research, we analyzed a diffuse small lymphocytic lymphoma dataset from 10x Genomics Multiome with 14,104 cells, 36,601 genes, and 70,469 enhancers. Post-unsupervised clustering and manual cell type annotation using Seurat v.4.0.5, we refined the dataset to include 11 cell types (Fig. 5A) [11]. STREAM identified 50 eRegulons across this dataset, each containing 19–290 genes and 9–70 active enhancers affecting 34–703 cells. Notably, 99.6% of genes were linked to 1–10 enhancers, and 47.0% of enhancers were exclusively associated with their nearest gene (Fig. 5B).

We identified 37 eRegulons specific to certain cell types in a diffuse small lymphocytic lymphoma dataset, including Effector-like and Exhaustive CD4+ T cells, Effector-like and Central memory-like CD8+ T cells, Dendritic cells, Normal and Tumor B cells. These eRegulons, exclusive to eight cell types, showed TF co-binding patterns (Fig. 5C). Notable TFs regulating these eRegulons included TCF7 and LHX2 in Normal B cells [56, 57], along with SPI1 [58, 59], TCF12 [60], STAT3 [61, 62], NFIC [63], NFIB [64], and MYF5 [65] in Tumor B cells. We observed a concordance between expression and chromatin accessibility in eRegulons regulated by these TFs, highlighting chromatin's role in transcription across different cell types (Fig. 5D). However, for eRegulons like those regulated by LHX2 (eR21), SPI1 (eR12), TCF7 (eR22), and TCF12 (eR13), chromatin accessibility changes did not always accompany expression variations, indicating the complex interplay of chromatin accessibility in transcription regulation within specific eRegulons [66].

In exploring TF–enhancer–gene relationships in diffuse small lymphocytic lymphoma, we analyzed cell-type-specific eRegulons within Normal B cells, Tumor B proliferation cells, and Tumor

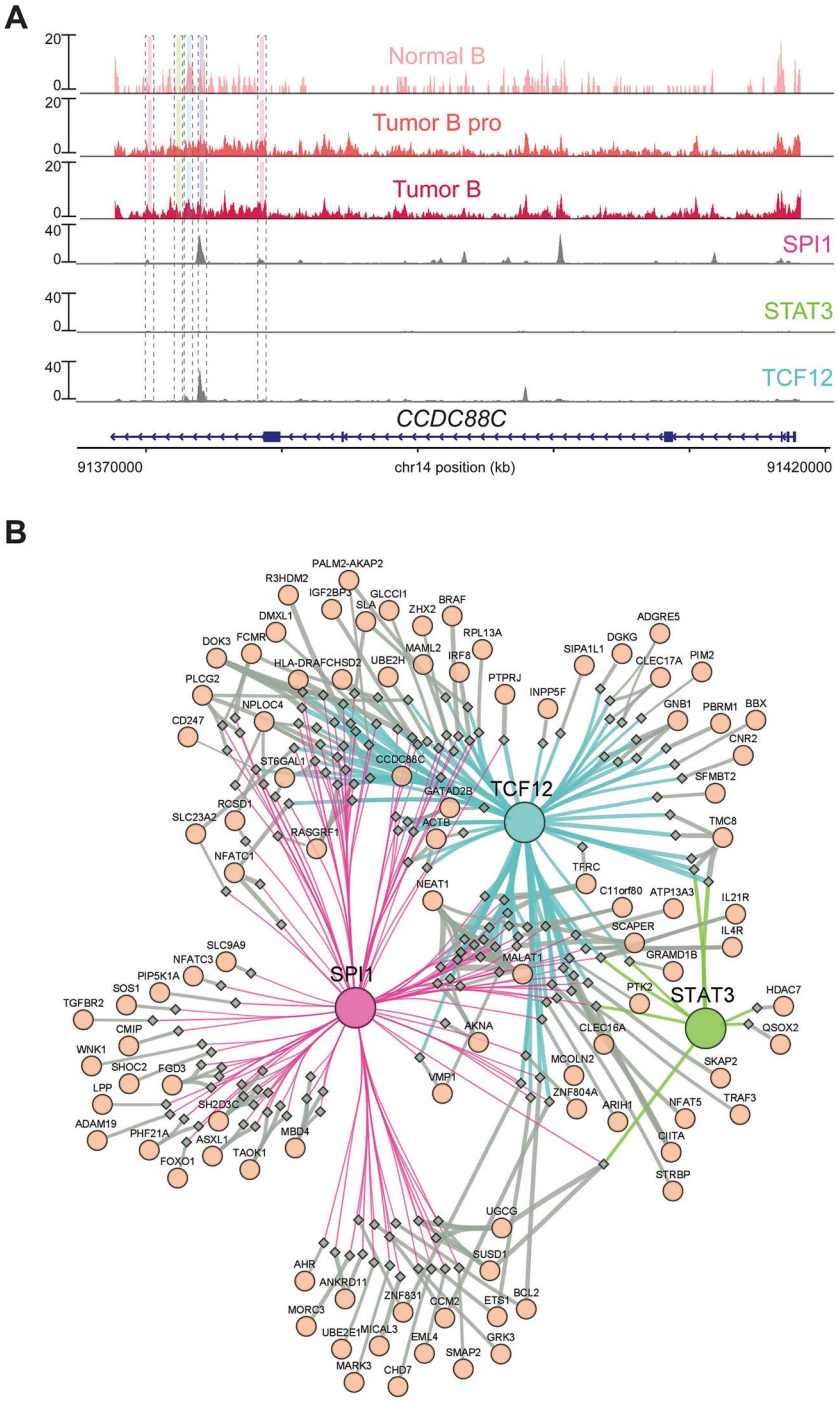

Figure 6. Interaction of SPI1, STAT3, and TCF12 in B cells. (A) Chromatin accessibility profiles across three distinct B cell types, complemented by ChIP-seq signals for SPI1, STAT3, and TCF12 in the specified chromosomal region (chr14–91370000–91420000). (B) Visualization of an eGRN exclusive to proliferating Tumor B cells, constructed by the interplay of SPI1, STAT3, and TCF12 and confined to highly variable genes or enhancers. The thickness of the line connecting two nodes signifies the proportion of cells in which both nodes are either accessible (in the case of enhancers) or expressed (in the case of genes encoding TFs or their targets). Abbreviations: Tumor B pro, Tumor B proliferation cell.

B cells (Figs. 5E-5H). We discovered cell-type-specific eRegulons: two in Normal B, eight in Tumor B proliferation, and nine in Tumor B cells. In Normal B cells, eRegulons regulated by TCF7 and LHX2 targeted 14 and 73 genes through 11 and 82 enhancers [56], highlighting LHX2's role in B cell differentiation and TCF7's in B cell lineage commitment [56]. The Tumor B proliferation cells–cells presented eight eRegulons governed by TFs such as LHX2, MYF5, NFIB, NFIC, SPI1, STAT3, TCF12, and ZFP42, affecting 14–478 genes via 11–1,097 enhancers. This diversity underlines B cell differentiation's molecular complexity, with MYF5 linked to myoblast proliferation and NFIB/NFIC associated with lymphoma cell characteristics [63–65]. SPI1 and TCF12 are essential for B cell functionality, while ZFP42 (REX1) is connected to lymph node oncogenesis [58–60, 67, 68]. Tumor B cells shared similarities with Tumor B proliferation cells, with the addition of an EGR1 eRegulon, a gene implicated in B-cell lymphoma signaling pathways [69]. STREAM's

analysis reveals intricate TF–enhancer–gene dynamics, offering insights into the molecular underpinnings of Tumor B cell biology.

To assess the impact of TF binding variation on gene expression in B cells, we studied CCDC88C, regulated by an enhancer (chr14–91366590–91389532) across diffuse small lymphocytic lymphoma marker genes (Fig. 5E). CCDC88C encodes Daple, which activates Wnt signaling [70], reducing apoptosis [71], and is upregulated in lymphoma/leukemia B cells versus healthy counterparts [72]. Variation in TF binding to this enhancer was noted among B cell types: TCF7 exclusively in Normal B, with SPI1, TCF12, STAT3, MYF5, NFIC, and NFIB co-binding in Tumor B proliferation and Tumor B cells (Fig. 5F), indicating cooperative TF interactions. Violin plots showed CCDC88C upregulation in Tumor B cells compared to Normal B, highlighting the role of diverse TF combinations in CCDC88C regulation and cell differentiation (Fig. 5F).

Exploring the impact of enhancer–gene relationships on lymphoma marker gene expression, we analyzed ACAP2 and XXYLT1, regulated by LHX2 at chr3–195203363–195211311 (Fig. 5G). ACAP2, overrepresented in lymphoma [73], and XXYLT1, a Notch signaling regulator [74, 75], illustrate the complexity of gene regulation in cancer. Regulatory strength, defined by the fraction of cells showing both enhancer accessibility and gene expression, highlighted LHX2's distinct influence on these genes across B cell types. XXYLT1 showed regulatory strengths of 0.63 and 0.56 in Tumor B proliferation and Tumor B cells, respectively, absent in Normal B cells (Fig. 5H). ACAP2 displayed varying strengths across cell types, peaking in Tumor B cells (0.96). Increased Tn5 insertion events in tumor cells suggest enhanced marker expression driven by regulatory strength changes, possibly due to altered chromatin accessibility. Thus, chr3–195203363–195211311, under LHX2 regulation, varies in regulatory intensity across B cell types, affecting marker gene expression and cell differentiation.

STREAM constructs comprehensive eGRNs highlighting TF cooperativity in gene regulation across cell types

To elucidate TF-enhancer–gene interactions in B cells of diffuse small lymphocytic lymphoma, we focused on SPI1, STAT3, and TCF12, leveraging ChIP-seq data from ENCODE's GM12878 cell line [76]. This facilitated eGRN reconstruction in Normal B, Tumor B proliferation, and Tumor B cells. SPI1 and TCF12 are pivotal for B cell development and commitment [58–60], while STAT3 is crucial in B cell lymphoma [61, 62]. SPI1 and STAT3 collaboratively influence tumorigenesis pathways, and TCF12 [77], belonging to the basic helix–loop–helix family, contributes to metastasis, including in lymph nodes [78]. These TFs regulate CCDC88C, a Wnt signaling activator. Analyzing their binding within chr14–91370000–91420000 across B cell types (Fig. 6A), we noted significant binding site and ChIP-seq signal overlaps, indicating heightened chromatin accessibility and TF activity, especially for SPI1 and TCF12 (Fig. 6A). STREAM's eGRNs thus reveal the cooperative regulation by multiple TFs, offering insights into their roles in gene regulation and lymphoma pathology.

To map TF-enhancer–gene interactions in diffuse small lymphocytic lymphoma B cells, we focused on SPI1, STAT3, and TCF12, analyzing their enhancer and gene interactions. We quantified TF–enhancer and enhancer–gene relationship strengths based on gene activity and enhancer accessibility across B cell types. This approach generated three eGRNs, each comprising 276 nodes and 457 edges, with edge weights differing by cell type (Fig. 6B and Supplementary Figs. S6-S7). Notably, in Tumor B proliferation cells, we found 253 enhancers jointly targeted by SPI1 and STAT3 regulating 148 genes, 297 by SPI1 and TCF12 for 94 genes, and 208 by STAT3 and TCF12 for 69 genes (Supplementary Table S6). Pathway enrichment analyses identified significant pathways related to lymphoma pathogenesis (Supplementary Table S7) [79], including leukocyte migration, toxoplasmosis, and thyroid hormone signaling [80–84]. Genes co-regulated by SPI1 and TCF12 were linked to acute myeloid leukemia and cancer metabolism, indicating their role in lymphoma [84–89]. Similarly, genes targeted by STAT3 and TCF12 were involved in immune differentiation and inflammatory diseases, highlighting their contribution to lymphoma risk [90]. These findings illustrate STREAM's capacity to elucidate the cooperative gene regulation by TFs in lymphoma, offering insights into the underlying biological pathways and mechanisms.

Conclusion

STREAM is a robust framework for inferring eGRNs from scRNA-seq and scATAC-seq data. By leveraging the Steiner forest problem model, hybrid biclustering, and submodular optimization, STREAM elucidates regulatory mechanisms by accounting for inter-dependencies among multiple enhancer–enhancer and enhancer–gene interactions. This approach eliminates the impact of pre-defined cell clusters on eGRN inference and globally optimizes TF-enhancer–gene relationships through a global optimization perspective. Benchmarking analyses reveal that STREAM outperforms six established (e)GRN inference methods across various datasets in TF recovery, TF–enhancer linkage prediction, and enhancer–gene relationship identification. Applied to Alzheimer's disease and diffuse small lymphocytic lymphoma datasets, STREAM effectively identified eRegulons and their dynamics over pseudotime, highlighting its utility in revealing disease-specific gene regulation patterns and the interplay among TFs in gene regulation. Despite its strengths, STREAM has limitations that need addressing, such as developing customized approaches or parameter settings for different data qualities or sequencing techniques, integrating additional modalities (e.g., 3-D genome structures, protein expression, protein–protein interactions), distinguishing positive/negative *trans*- and *cis*-regulatory mechanisms, inferring causality among TFs, enhancers, and genes, and developing valid benchmarking schemes using bulk data. Moreover, STREAM must identify critical states and transitions in complex biological systems and disease progression through graph entropy analyses [91–94]. Nonetheless, STREAM's robust performance underscores its potential as a complementary tool in gene regulation analysis, with future enhancements anticipated through integration with other GRN prediction approaches, promising deeper insights into cellular dynamics and regulatory networks.

Key Points

- The paper introduces STREAM, an innovative algorithm designed for the inference of enhancer-driven gene regulatory networks.
- STREAM leverages global optimization strategies, incorporating a Steiner forest problem model and a hybrid biclustering pipeline integrated with a framework of submodular optimization.
- Evaluated against six established methods across benchmark datasets from six cell lines, STREAM demonstrates superior performance.

- STREAM adeptly identifies relationships among transcription factors, enhancers, and genes relevant to Alzheimer's disease and diffuse small lymphocytic lymphoma.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Funding

This work was supported by award R01GM131399 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. The authors thank Qiuqin Wu, Xiaoying Wang, Zhenyu Wu, Yujia Xiang, Shuangquan Zhang, Shuo Chen, and Cindy Tong for their assistance in data collection, pipeline development, tool testing, and cell type annotation.

Data availability

Detailed tutorials and documentation on the STREAM workflow are available at https://github.com/OSU-BMBL/STREAM. Scripts to reproduce the analyses presented in this manuscript are available at https://github.com/OSU-BMBL/stream_analyses. All datasets analyzed in this study were published previously. The corresponding descriptions and pre-processing steps are described in the Materials and Methods.

References

1. Jin T, Rehani P, Ying M. *et al.* scGRNom: a computational pipeline of integrative multi-omics analyses for predicting cell-type disease genes and regulatory networks. *Genome Med* 2021;**13**:95. https://doi.org/10.1186/s13073-021-00908-9.
2. Aibar S, González-Blas CB, Moerman T. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* 2017;**14**:1083–6. https://doi.org/10.1038/nmeth.4463.
3. Zhang L, Zhang J, Nie Q. DIRECT-NET: An efficient method to discover cis-regulatory elements and construct regulatory networks from single-cell multiomics data. *Sci Adv* 2022;**8**:eabl7393. https://doi.org/10.1126/sciadv.abl7393.
4. Cao Z-J, Gao G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat Biotechnol* 2022;**40**:1458–66. https://doi.org/10.1038/s41587-022-01284-4.
5. Janssens J, Aibar S, Taskiran II. *et al.* Decoding gene regulation in the fly brain. *Nature* 2022;**601**:630–6. https://doi.org/10.1038/s41586-021-04262-z.
6. Duren Z, Lu WS, Arthur JG. *et al.* Sc-compReg enables the comparison of gene regulatory networks between conditions using single-cell data. *Nat Commun* 2021;**12**:4763. https://doi.org/10.1038/s41467-021-25089-2.
7. Duren Z, Chen X, Zamanighomi M. *et al.* Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. *Proc Natl Acad Sci* 2018;**115**:7723–8. https://doi.org/10.1073/pnas.1805681115.
8. Duren Z, Chang F, Naqing F. *et al.* Regulatory analysis of single cell multiome gene expression and chromatin accessibility data with scREG. *Genome Biol* 2022;**23**:114. https://doi.org/10.1186/s13059-022-02682-2.
9. Badia-i-Mompel P, Wessels L, Müller-Dott S. *et al.* Gene regulatory network inference in the era of single-cell multi-omics. *Nat Rev Genet* 2023;**24**:739–54. https://doi.org/10.1038/s41576-023-00618-5.
10. Kamimoto K, Stringa B, Hoffmann CM. *et al.* Dissecting cell identity via network inference and in silico gene perturbation. *Nature* 2023;**614**:742–51. https://doi.org/10.1038/s41586-022-05688-9.
11. Ma A, Wang X, Li J. *et al.* Single-cell biological network inference using a heterogeneous graph transformer. *Nat Commun* 2023;**14**:964. https://doi.org/10.1038/s41467-023-36559-0.
12. Kartha VK, Duarte FM, Hu Y. *et al.* Functional inference of gene regulation using single-cell multi-omics. *Cell Genom* 2022;**2**:100166. https://doi.org/10.1016/j.xgen.2022.100166.
13. Kamal A, Arnold C, Claringbould A. *et al.* GRaNIE and GRaNPA: Inference and evaluation of enhancer-mediated gene regulatory networks applied to study macrophages. *Mol Syst Biol* 2023;**19**:e11627. https://doi.org/10.15252/msb.202311627.
14. Fleck JS, Jansen SMJ, Wollny D. *et al.* Inferring and perturbing cell fate regulomes in human brain organoids. *Nature* 2023;**621**:365–72. https://doi.org/10.1038/s41586-022-05279-8.
15. Duren Z, Chen X, Jiang R. *et al.* Modeling gene regulation from paired expression and chromatin accessibility data. *Proc Natl Acad Sci* 2017;**114**:E4914–23. https://doi.org/10.1073/pnas.1704553114.
16. González-Blas CB, De Winter S, Hulselmans G. *et al.* SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nat Methods* 2023;**20**:1355–67. https://doi.org/10.1038/s41592-023-01938-4.
17. Li Z, Nagai JS, Kuppe C. *et al.* scMEGA: single-cell multi-omic enhancer-based gene regulatory network inference. *Bioinform Adv* 2023;**3**:vbad003. https://doi.org/10.1093/bioadv/vbad003.
18. Duren Z, Chen X, Xin J. *et al.* Time course regulatory analysis based on paired expression and chromatin accessibility data. *Genome Res* 2020;**30**:622–34. https://doi.org/10.1101/gr.257063.119.
19. Zhang Q, Teng P, Wang S. *et al.* Computational prediction and characterization of cell-type-specific and shared binding sites. *Bioinformatics* 2022;**39**:btac798. https://doi.org/10.1093/bioinformatics/btac798.
20. Buenrostro JD, Wu B, Litzenburger UM. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 2015;**523**:486–90. https://doi.org/10.1038/nature14590.
21. Granja JM, Corces MR, Pierce SE. *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat Genet* 2021;**53**:403–11. https://doi.org/10.1038/s41588-021-00790-6.
22. Gassner E. The Steiner Forest Problem revisited. *J Discrete Algorithms* 2010;**8**:154–63. https://doi.org/10.1016/j.jda.2009.05.002.
23. Wei K, Libbrecht MW, Bilmes JA. *et al.* Choosing panels of genomics assays using submodular optimization. *Genome Biol* 2016;**17**:229. https://doi.org/10.1186/s13059-016-1089-7.
24. Chang Y, Allen C, Wan C. *et al.* IRIS-FGM: an integrative single-cell RNA-Seq interpretation system for functional gene module analysis. *Bioinformatics* 2021;**37**:3045–7. https://doi.org/10.1093/bioinformatics/btab108.
25. Wan C, Chang W, Zhang Y. *et al.* LTMG: a novel statistical modeling of transcriptional expression states in single-cell

RNA-Seq data. *Nucleic Acids Res* 2019;**47**:e111. https://doi.org/10.1093/nar/gkz655.
26. Stuart T, Srivastava A, Madad S. *et al.* Single-cell chromatin state analysis with Signac. *Nat Methods* 2021;**18**:1333–41. https://doi.org/10.1038/s41592-021-01282-5.
27. Pliner HA, Packer JS, McFaline-Figueroa JL. *et al.* Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol Cell* 2018;**71**:858–871.e8. https://doi.org/10.1016/j.molcel.2018.06.044.
28. Abbaszadeh O, Azarpeyvand A, Khanteymoori A. *et al.* Data-Driven and Knowledge-Based Algorithms for Gene Network Reconstruction on High-Dimensional Data. *IEEE/ACM Trans Comput Biol Bioinform* 2022;**19**:1545–57. https://doi.org/10.1109/TCBB.2020.3034861.
29. Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I. *et al.* JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2022;**50**:D165–73. https://doi.org/10.1093/nar/gkab1113.
30. Li G, Ma Q, Tang H. *et al.* QUBIC: a qualitative biclustering algorithm for analyses of gene expression data. *Nucleic Acids Res* 2009;**37**:e101. https://doi.org/10.1093/nar/gkp491.
31. Eddy SR. What is dynamic programming? *Nat Biotechnol* 2004;**22**:909–10. https://doi.org/10.1038/nbt0704-909.
32. Liu L, Liu C. *et al.* Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity. *Nat Commun* 2019;**10**:470. https://doi.org/10.1038/s41467-018-08205-7.
33. Ma S, Zhang B, LaFave LM. *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* 2020;**183**:1103–1116.e20. https://doi.org/10.1016/j.cell.2020.09.056.
34. Plongthongkum N, Diep D, Chen S. *et al.* Scalable dual-omics profiling with single-nucleus chromatin accessibility and mRNA expression sequencing 2 (SNARE-seq2). *Nat Protoc* 2021;**16**:4992–5029. https://doi.org/10.1038/s41596-021-00507-3.
35. Cao J, Spielmann M, Qiu X. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* 2019;**566**:496–502. https://doi.org/10.1038/s41586-019-0969-x.
36. Ferrari R, Dawoodi S, Raju M. *et al.* Androgen receptor gene and sex-specific Alzheimer's disease. *Neurobiol Aging* 2013;**34**:2077.e19–20. https://doi.org/10.1016/j.neurobiolaging.2013.02.017.
37. MacGibbon GA, Lawlor PA, Walton M. *et al.* Expression of Fos, Jun, and Krox family proteins in Alzheimer's disease. *Exp Neurol* 1997;**147**:316–32. https://doi.org/10.1006/exnr.1997.6600.
38. Pirskanen M, Hiltunen M, Mannermaa A. *et al.* Estrogen receptor beta gene variants are associated with increased risk of Alzheimer's disease in women. *Eur J Hum Genet* 2005;**13**:1000–6. https://doi.org/10.1038/sj.ejhg.5201447.
39. Zhao L, Woody SK, Chhibber A. Estrogen receptor β in Alzheimer's disease: From mechanisms to therapeutics. *Ageing Res Rev* 2015;**24**:178–90. https://doi.org/10.1016/j.arr.2015.08.001.
40. Morabito S, Miyoshi E, Michael N. *et al.* Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's disease. *Nat Genet* 2021;**53**:1143–55. https://doi.org/10.1038/s41588-021-00894-z.
41. Liu C, Chyr J, Zhao W. *et al.* Genome-Wide Association and Mechanistic Studies Indicate That Immune Response Contributes to Alzheimer's Disease Development. *Front Genet* 2018;**9**:410. https://doi.org/10.3389/fgene.2018.00410.
42. Patel A, Rees SD, Kelly MA. *et al.* Association of variants within APOE, SORL1, RUNX1, BACE1 and ALDH18A1 with dementia in Alzheimer's disease in subjects with Down syndrome. *Neurosci Lett* 2011;**487**:144–8. https://doi.org/10.1016/j.neulet.2010.10.010.
43. Acquaah-Mensah GK, Agu N, Khan T. *et al.* A regulatory role for the insulin- and BDNF-linked RORA in the hippocampus: implications for Alzheimer's disease. *J Alzheimers Dis* 2015;**44**:827–38. https://doi.org/10.3233/JAD-141731.
44. Sierksma A, Lu A, Mancuso R. *et al.* Novel Alzheimer risk genes determine the microglia response to amyloid-β but not to TAU pathology. *EMBO Mol Med* 2020;**12**:e10606. https://doi.org/10.15252/emmm.201910606.
45. Fukui H, Rünker A, Fabel K. *et al.* Transcription factor Runx1 is pro-neurogenic in adult hippocampal precursor cells. *PloS One* 2018;**13**:e0190789. https://doi.org/10.1371/journal.pone.0190789.
46. Raber J. Androgens, ApoE, and Alzheimer's Disease. *Sci Aging Knowledge Environ* 2004;**2004**:re2–2. https://doi.org/10.1126/sageke.2004.11.re2.
47. Anderson AJ, Su JH, Cotman CW. DNA damage and apoptosis in Alzheimer's disease: colocalization with c-Jun immunoreactivity, relationship to brain area, and effect of postmortem delay. *J Neurosci* 1996;**16**:1710–9. https://doi.org/10.1523/JNEUROSCI.16-05-01710.1996.
48. Marcus DL, Strafaci JA, Miller DC. *et al.* Quantitative neuronal c-Fos and c-Jun expression in Alzheimer's disease11To whom correspondence should be addressed. *Neurobiol Aging* 1998;**19**:393–400. https://doi.org/10.1016/S0197-4580(98)00077-3.
49. Hüttenrauch M, Salinas G, Wirths O. Effects of Long-Term Environmental Enrichment on Anxiety, Memory, Hippocampal Plasticity and Overall Brain Gene Expression in C57BL6 Mice. *Front Mol Neurosci* 2016;**9**:62. https://doi.org/10.3389/fnmol.2016.00062.
50. Mentis AA, Vlachakis D, Papakonstantinou E. *et al.* A novel variant in DYNC1H1 could contribute to human amyotrophic lateral sclerosis-frontotemporal dementia spectrum. *Cold Spring Harb Mol Case Stud* 2022;**8**:mcs.a006096. https://doi.org/10.1101/mcs.a006096.
51. Illarionova NB, Borisova MA, Bazhenova EY. *et al.* Zbtb33 Gene Knockout Changes Transcription of the Fgf9, Fgfr3, c-Myc and FoxG1 Genes in the Developing Mouse Brain. *Mol Biol* 2021;**55**:363–71. https://doi.org/10.1134/S0026893321020230.
52. Rhee JW, Arata A, Selleri L. *et al.* Pbx3 deficiency results in central hypoventilation. *Am J Pathol* 2004;**165**:1343–50. https://doi.org/10.1016/S0002-9440(10)63392-5.
53. Kunkle BW, Vardarajan BN, Naj AC. *et al.* Early-Onset Alzheimer Disease and Candidate Risk Genes Involved in Endolysosomal Transport. *JAMA Neurol* 2017;**74**:1113–22. https://doi.org/10.1001/jamaneurol.2017.1518.
54. Naruhashi K, Kadomatsu K, Igakura T. *et al.* Abnormalities of Sensory and Memory Functions in Mice LackingB-sgGene. *Biochem Biophys Res Commun* 1997;**236**:733–7. https://doi.org/10.1006/bbrc.1997.6993.
55. Najyb O, Brissette L, Rassart E. Apolipoprotein D Internalization Is a Basigin-dependent Mechanism. *J Biol Chem* 2015;**290**:16077–87. https://doi.org/10.1074/jbc.M115.644302.
56. Rosén A, Bergh A-C, Gogok P. *et al.* Lymphoblastoid cell line with B1 cell characteristics established from a chronic lymphocytic leukemia clone by in vitro EBV infection. *Onco Targets Ther* 2012;**1**:18–27. https://doi.org/10.4161/onci.1.1.18400.
57. Wu JQ, Seay M, Schulz VP. *et al.* Tcf7 is an important regulator of the switch of self-renewal and differentiation in a multipotential hematopoietic cell line. *PLoS Genet* 2012;**8**:e1002565. https://doi.org/10.1371/journal.pgen.1002565.

58. Solomon LA, Li SKH, Piskorz J. *et al.* Genome-wide comparison of PU.1 and Spi-B binding sites in a mouse B lymphoma cell line. *BMC Genomics* 2015;**16**:76. https://doi.org/10.1186/s12864-015-1303-0.
59. Torlakovic E, Tierens A, Dang HD. *et al.* The Transcription Factor PU.1, Necessary for B-Cell Development Is Expressed in Lymphocyte Predominance, But Not Classical Hodgkin's Disease. *Am J Pathol* 2001;**159**:1807–14. https://doi.org/10.1016/S0002-9440(10)63027-1.
60. Galbiati M, Lettieri A, Micalizzi C. *et al.* Natural history of acute lymphoblastic leukemia in neurofibromatosis type 1 monozygotic twins. *Leukemia* 2013;**27**:1778–81. https://doi.org/10.1038/leu.2013.55.
61. Scuto A, Kujawski M, Kowolik C. *et al.* STAT3 Inhibition Is a Therapeutic Strategy for ABC-like Diffuse Large B-Cell Lymphoma. *Cancer Res* 2011;**71**:3182–8. https://doi.org/10.1158/0008-5472.CAN-10-2380.
62. Huang X, Meng B, Iqbal J. *et al.* Activation of the STAT3 signaling pathway is associated with poor survival in diffuse large B-cell lymphoma treated with R-CHOP. *J Clin Oncol* 2013;**31**:4520–8. https://doi.org/10.1200/JCO.2012.45.6004.
63. Schmidl C, Vladimer GI, Rendeiro AF. *et al.* Combined chemosensitivity and chromatin profiling prioritizes drug combinations in CLL. *Nat Chem Biol* 2019;**15**:232–40. https://doi.org/10.1038/s41589-018-0205-2.
64. Tian M, Li Y, Zheng W. *et al.* LncRNA PCAT1 enhances cell proliferation, migration and invasion by miR-508-3p/NFIB axis in diffuse large B-cell lymphoma. *Eur Rev Med Pharmacol Sci* 2021;**25**:2567–76. https://doi.org/10.26355/eurrev_202103_25420.
65. Harada A, Okada S, Odawara J. *et al.* Production of a rat monoclonal antibody specific for Myf5. *Hybridoma (Larchmt)* 2010;**29**:59–62. https://doi.org/10.1089/hyb.2009.0066.
66. Lynch AW, Theodoris CV, Long HW. *et al.* MIRA: joint regulatory modeling of multimodal expression and chromatin accessibility in single cells. *Nat Methods* 2022;**19**:1097–108. https://doi.org/10.1038/s41592-022-01595-z.
67. Bernard SC, Abdelsamad EH, Johnson PA. *et al.* Pediatric leukemia: Diagnosis to treatment–A review. *J Cancer Clin Trials* 2017;**2**:1.
68. Penther D, Viailly PJ, Latour S. *et al.* A recurrent clonally distinct Burkitt lymphoma case highlights genetic key events contributing to oncogenesis. *Genes Chromosomes Cancer* 2019;**58**:595–601. https://doi.org/10.1002/gcc.22743.
69. Kimpara S, Lu L, Hoang NM. *et al.* EGR1 Addiction in Diffuse Large B-cell Lymphoma. *Mol Cancer Res* 2021;**19**:1258–69. https://doi.org/10.1158/1541-7786.MCR-21-0267.
70. Kurihara Y, Mizuno H, Honda A. *et al.* CCDC88C-FLT3 gene fusion in CD34-positive haematopoietic stem and multilineage cells in myeloid/lymphoid neoplasm with eosinophilia. *J Cell Mol Med* 2022;**26**:950–2. https://doi.org/10.1111/jcmm.17143.
71. Román-Gómez J, Cordeu L, Agirre X. *et al.* Epigenetic regulation of Wnt-signaling pathway in acute lymphoblastic leukemia. *Blood* 2007;**109**:3462–9. https://doi.org/10.1182/blood-2006-09-047043.
72. Patel MS, Kendall EK, Ondrejka S. *et al.* Gene Expression and Epigenetic Analysis in Relapsed/Refractory Diffuse Large B Cell Lymphoma Provides Insights into Evolution of Treatment Resistance to R-CHOP. *Blood* 2020;**136**:26. https://doi.org/10.1182/blood-2020-138645.
73. Rouillard AD, Gundersen GW, Fernandez NF. *et al.* The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* 2016;**2016**:baw100. https://doi.org/10.1093/database/baw100.
74. Yu H, Takeuchi M, LeBarron J. *et al.* Notch-modifying xylosyltransferase structures support an SNi-like retaining mechanism. *Nat Chem Biol* 2015;**11**:847–54. https://doi.org/10.1038/nchembio.1927.
75. Lobry C, Oh P, Mansour MR. *et al.* Notch signaling: switching an oncogene to a tumor suppressor. *Blood* 2014;**123**:2451–9. https://doi.org/10.1182/blood-2013-08-355818.
76. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**:57–74. https://doi.org/10.1038/nature11247.
77. Huang C, Xie K. Crosstalk of Sp1 and Stat3 signaling in pancreatic cancer pathogenesis. *Cytokine Growth Factor Rev* 2012;**23**:25–35. https://doi.org/10.1016/j.cytogfr.2012.01.003.
78. Yang J, Zhang L, Jiang Z. *et al.* TCF12 promotes the tumorigenesis and metastasis of hepatocellular carcinoma via upregulation of CXCR4 expression. *Theranostics* 2019;**9**:5810–27. https://doi.org/10.7150/thno.34973.
79. Ogata H, Goto S, Sato K. *et al.* KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 1999;**27**:29–34. https://doi.org/10.1093/nar/27.1.29.
80. Tsuzuki S, Toyama-Sorimachi N, Kitamura F. *et al.* Intracellular Signal-transducing Elements Involved in Transendothelial Migration of Lymphoma Cells. *Jpn J Cancer Res* 1998;**89**:571–7. https://doi.org/10.1111/j.1349-7006.1998.tb03299.x.
81. Dudley NJ, Balfour AH. Non-Hodgkin's lymphoma presenting as 'chronic active toxoplasmosis. *Postgrad Med J* 1988;**64**:883–5. https://doi.org/10.1136/pgmj.64.757.883.
82. Intaraphet S, Farkas DK, Johannesdottir Schmidt SA. *et al.* Human papillomavirus infection and lymphoma incidence using cervical conization as a surrogate marker: a Danish nationwide cohort study. *Hematol Oncol* 2017;**35**:172–6. https://doi.org/10.1002/hon.2270.
83. Liu YC, Yeh CT, Lin KH. Molecular Functions of Thyroid Hormone Signaling in Regulation of Cancer Progression and Anti-Apoptosis. *Int J Mol Sci* 2019;**20**:4896. https://doi.org/10.3390/ijms20204986.
84. Bispo JAB, Pinheiro PS, Kobetz E. Epidemiology and Etiology of Leukemia and Lymphoma. *Cold Spring Harb Perspect Med* 2020;**10**:a034819. https://doi.org/10.1101/cshperspect.a034819.
85. Mehravaran H, Makvandi M, Samarbaf Zade A. *et al.* Association of Human Cytomegalovirus with Hodgkin's Disease and Non-Hodgkin's lymphomas. *Asian Pac J Cancer Prev* 2017;**18**:593–7.
86. Grützmeier S, Porwit A, Schmitt C. *et al.* 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. *Infect Agent Cancer* 2016;**11**:46. https://doi.org/10.1186/s13027-016-0094-5.
87. Sato K, Igarashi S, Tsukada N. *et al.* Cytomegalovirus infection in patients with malignant lymphomas who have not received hematopoietic stem cell transplantation. *BMC Cancer* 2022;**22**:944. https://doi.org/10.1186/s12885-022-10008-5.
88. Kang DW, Choi K-Y, Min DS. Functional Regulation of Phospholipase D Expression in Cancer and Inflammation*. *J Biol Chem* 2014;**289**:22575–82. https://doi.org/10.1074/jbc.R114.569822.
89. Xiong J, Wang L, Fei XC. *et al.* MYC is a positive regulator of choline metabolism and impedes mitophagy-dependent necroptosis in diffuse large B-cell lymphoma. *Blood Cancer J* 2017;**7**:e582–2. https://doi.org/10.1038/bcj.2017.61.
90. Lan Q, Wang SS, Menashe I. *et al.* Genetic variation in Th1/Th2 pathway genes and risk of non-Hodgkin lymphoma: a pooled analysis of three population-based case-control studies. *Br J Haematol* 2011;**153**:341–50. https://doi.org/10.1111/j.1365-2141.2010.08424.x.

91. Zhong J, Tang H, Huang Z. *et al.* Uncovering the pre-deterioration state during disease progression based on sample-specific causality network entropy (SCNE). *Research* 2024;**7**:0368. https://doi.org/10.34133/research.0368.
92. Zhong J, Han C, Chen P. *et al.* SGAE: single-cell gene association entropy for revealing critical states of cell transitions during embryonic development. *Brief Bioinform* 2023;**24**:bbad366. https://doi.org/10.1093/bib/bbad366.
93. Zhong J, Han C, Wang Y. *et al.* Identifying the critical state of complex biological systems by the directed-network rank score method. *Bioinformatics* 2022;**38**:5398–405. https://doi.org/10.1093/bioinformatics/btac707.
94. Zhong J, Han C, Zhang X. *et al.* scGET: Predicting Cell Fate Transition During Early Embryonic Development by Single-cell Graph Entropy. *Genomics Proteomics Bioinformatics* 2021;**19**:461–74. https://doi.org/10.1016/j.gpb.2020.11.008.