Single-cell gene regulation network inference by large-scale data integration

Xin Dong [1,2,†], Ke Tang [1,2,†], Yunfan Xu [1,2], Hailin Wei [1,2], Tong Han [1,2] and Chenfei Wang [1,2,*]

[1]Key Laboratory of Spine and Spinal Cord Injury Repair and Regeneration of Ministry of Education, Department of Orthopedics, Tongji Hospital, School of Life Sciences and Technology, Tongji University, Shanghai 200092, China and [2]Frontier Science Center for Stem Cells, School of Life Sciences and Technology, Tongji University, Shanghai 200092, China

Received July 06, 2022; Revised August 11, 2022; Editorial Decision September 05, 2022; Accepted September 14, 2022

ABSTRACT

Single-cell ATAC-seq (scATAC-seq) has proven to be a state-of-art approach to investigating gene regulation at the single-cell level. However, existing methods cannot precisely uncover cell-type-specific binding of transcription regulators (TRs) and construct gene regulation networks (GRNs) in single-cell. ChIP-seq has been widely used to profile TR binding sites in the past decades. Here, we developed SCRIP, an integrative method to infer single-cell TR activity and targets based on the integration of scATAC-seq and a large-scale TR ChIP-seq reference. Our method showed improved performance in evaluating TR binding activity compared to the existing motif-based methods and reached a higher consistency with matched TR expressions. Besides, our method enables identifying TR target genes as well as building GRNs at the single-cell resolution based on a regulatory potential model. We demonstrate SCRIP's utility in accurate cell-type clustering, lineage tracing, and inferring cell-type-specific GRNs in multiple biological systems. SCRIP is freely available at https://github.com/wanglabtongji/SCRIP.

INTRODUCTION

Gene regulation is the basis of many biological processes, including development, differentiation, and disease occurrence and progression. Recently, many single-cell technologies have been developed to investigate gene regulation mechanisms from diverse genomic aspects, such as transcriptomes (1), epigenomes (2) or 3D structures (3). Among them, the single-cell sequencing assay for transposase-accessible chromatin (scATAC-seq) has enabled the profiling of the genome-wide chromatin accessibility landscapes in single cells (4). The most powerful application of the scATAC-seq data is to understand how specific transcription regulators (TR), including transcription factors (TF) and chromatin regulators (CR), bind to the genome and regulate their target genes. Constructing the gene regulatory networks (GRNs) is crucial for understanding the roles of different TRs in regulating development trajectories and disease traits.

Although scATAC-seq has been widely used to tackle gene regulation and their association with phenotypes, several questions remain unsolved. First, the chromatin accessibility captured by scATAC-seq only reflects the overall regulatory potential and cannot identify the binding of exact TRs. Existing methods like chromVAR (5), scFAN (6) and SCENIC (7) integrate sequence features like motifs to evaluate TF activity in each cell. However, motif-based methods cannot discriminate factors of the same TF family that have similar motifs, and also failed to evaluate factors with indirect DNA binding such as CRs. Second, the scATAC-seq data is very sparse and noisy as only two strands of DNA can be captured within a cell. Methods like Signac (8), EpiScanpy (9), MAESTRO (10) and SCALE (11) enhanced the signals by using different latent features, however, the algorithm-defined features were mostly analyzed at the cell type level and cannot be directly linked to the single-cell TR activity. Last but most important, none of these methods can identify the TR targets in each cell, and constructing the GRNs at the single-cell level is still not feasible with scATAC-seq data alone. Therefore, new methods with the potential to address TR binding enrichment and identify its associated target genes at the single-cell level are highly needed for scATAC-seq data.

Chromatin Immunoprecipitation Sequencing (ChIP-seq) (12,13) is a direct way to uncover TRs binding in the genome and determine their target genes at the bulk cell level. Compared to motifs, ChIP-seq is more accurate in defining cell-type-specific TR binding sites and investigating the genomic distribution for many non-DNA-binding CRs. In the past decades, numerous TR ChIP-seq data have been generated for different cell lines, tissues, and species (14–16). Several projects, such as Cistrome DB (16), ENCODE (15) and

*To whom correspondence should be addressed. Tel: +86 21 65981195; Fax: +86 21 65981195; Email: 08chenfeiwang@tongji.edu.cn
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Epigenome Roadmap (17) have curated a large collection of high-quality TR ChIP-seq data. Integrating the large-scale ChIP-seq datasets with the motif information will definitely improve the prediction of TR enrichment in the scATAC-seq data. However, several issues need to be addressed before integration. First, the large-scale ChIP-seq reference should be uniformly processed with standard quality control metrics to remove the potential low-quality data. Besides, while the antibody-affinity and the signal-to-noise ratio might be diverse for different TRs, the enrichment based on TR ChIP-seq peaks should be carefully adjusted and normalized. Finally, an efficient interval searching algorithm is needed for identifying enriched TRs from a large-scale genome-wide TR reference (18).

The major obstacle to investigating single-cell GRNs is the lack of single-cell ChIP-seq data. While several recently developed techniques such as scCUT&RUN (19), scCUT&Tag (20) and scCUT&Tag-pro (21) could successfully generate ChIP-seq profiles at the single-cell level, however, most of the data were generated for high abundant histone modifications (HMs) rather than TRs. Although several attempts have been performed on specific TRs (22), they are highly dependent on the quality of the TR antibodies and usually have extremely fuzzy signals at the single-cell level. Regulatory potential (RP) models have been widely used to identify TR targets for bulk ChIP-seq samples (23–25). The integration of TR ChIP-seq and scATAC-seq data has the potential for evaluating the single-cell TR binding site using the RP model, which could be based on the imputed TR ChIP-seq peaks at the single-cell level.

Here, we present a computational method SCRIP (**S**ingle-**C**ell gene **R**egulation network **I**nference using Ch**IP**-seq and motif), which integrates a large-scale TR and motif reference for evaluating TR activity as well as constructing single-cell GRNs based on scATAC-seq data. SCRIP includes a high-quality TR reference covering 1,252 human TRs and 997 mouse TRs. Based on this large-scale reference, SCRIP showed superior performance in evaluating single-cell TR activity, performing TR-based clustering, and lineage tracing analyses. In addition, SCRIP could accurately reconstruct the single-cell GRNs based on imputed ChIP-seq peaks at the single-cell level. We demonstrated the usability of SCRIP on multiple biological systems including peripheral blood mononuclear cell (PBMC), hematopoietic stem cell (HSC) differentiation, human organ development, and basal cell carcinomas (BCC).

MATERIALS AND METHODS

Data collection and generation of SCRIP index

We downloaded the uniformly processed ChIP-seq datasets from the Cistrome Data Browser (16,26,27) through the 'batch download' function. In total, we obtained bed files of 11 348 human and 9060 mouse TR ChIP-seq datasets, and 11 079 human and 10 944 mouse HM ChIP-seq datasets. Since the auto-parsed metadata included mistakes, we systematically curated the annotation of factors, cell types, and tissues. To ensure the quality of the datasets, we used the following criteria to filter the TR datasets: the raw sequence median quality score was >25, the percent of uniquely mapped reads was >50%, PBC (PCR bottleneck coefficient) was >0.8, the number of fold 10 peaks was >100, FRiP (Fraction of Reads in Peaks) was >0.01 and the number of top 5000 peaks overlapping with union DHS is >70%. To acquire the high confidence peaks, we only kept the 5-fold enrichment peaks in each peak set. Then, we removed the datasets with <1000 peaks. After filtering, we obtained 2314 human and 1920 mouse TR ChIP-seq datasets, covering 671 and 440 TRs respectively (Supplementary Figure S1). We also filtered histone modification datasets with the same criteria as above. According to previous reports, active histone modifications such as H3K4me1/2/3 or H3K27ac tends to be presented in the open chromatin region, while repressive histone modifications such as H3K27me3 or H3K9me3 are enriched at the heterochromatin region and have very few overlaps with the scATAC-seq peaks (28–30). Thus, we only retained the histone modifications with active functions, including H3K4me1/2/3, H3K9ac and H3K27ac. We obtained 1678 human and 1013 mouse HM ChIP-seq datasets covering five HMs (Supplementary Figure S2).

To improve the coverage of transcription factors, we downloaded the motif information, including 7704 human and 7000 mouse TF PWMs (position weight matrix), from the *cis*-BP database (31). We combined PWMs from the same TF and converted the format to the HOMER (32) format. Then we scanned the motifs on the hg38 or mm10 genome with the HOMER and obtained the genome intervals where motifs appeared. We overlapped the scanning intervals with the ENCODE ccRE (candidate *cis*-regulatory elements) (33) list and Cistrome union DHS (DNase-I hypersensitive sites) list and removed intervals with the intersection of the blacklist. To make the motif sites comparable to the ChIP-seq datasets, we extended the length of each scanned motif site to 340 bp, which is the average length of ChIP-seq peaks. For those motifs that have much more binding sites than others, we only kept the top 25k binding sites, which were the average of filtered peaks in ChIP-seq, by filtering out the low confidence motif sites using *P*-values. In total, we obtained 916 human and 816 mouse motif-scanned pseudo peaks (Supplementary Figure S1e, f).

Next, we combined the ChIP-seq peak sets and motif-scanned pseudo peak sets as reference datasets to build the search index. To calculate the similarity between reference datasets and the scATAC-seq datasets, we introduced GIGGLE (18), a fast genomics search engine, into SCRIP. We sorted the bed files, compressed them into gz format, and built the index with GIGGLE. In addition, we also included the peaks number, metadata of the datasets, and original bed files in the index. The reference processing codes are provided in the Code availability part. Overall, the human TR index covers 1252 TRs and the mouse covered 997 TRs (Supplementary Figure S1e, f).

TR activity score calculation

Normalization for removing biases. The SCRIP takes the scATAC-seq peak by count matrix or bin count matrix as input. For the sake of getting the comparable TR activity of each reference dataset in each cell, SCRIP first calculates the number of peak overlaps between each cell and the ChIP-seq peaks set or motif-scanned intervals set by GIGGLE. SCRIP records the number of overlap peaks to

build the matrix M, where the column is cell i and the row is dataset j, and the content is the number of overlapped peaks. The definition of peaks overlap is the same as the bedtools (34) and GIGGLE definition, for which a single bp overlap between TR ChIP-seq peaks/motif sites and the scATAC-seq peaks will be counted as one overlap. To remove the bias from the peak number of the datasets and the total length of single-cell peaks, SCRIP normalizes the matrix by:

$$N = D \times Q$$

$$M'_{i,j} = \frac{M_{i,j}}{N_{i,j}}$$

where D is a $j \times 1$ matrix that records the number of peaks of each TR ChIP-seq dataset or motif sites, and Q is a $1 \times i$ matrix that records the number of base-pair in each cell per 100 million. Then N is a normalization matrix that is the matrix product of D and Q, and the M' matrix is the normalized peak overlap matrix, which scores the relative enrichment of TRs in different cells. M' is further normalized by the average score of each TR dataset to scale the enrichment for different TRs:

$$M''_j = M'_j - \text{mean}\left(M'_j\right)$$

Deduplication for redundant TR datasets. As M'' still contains duplicate ChIP-seq datasets or motifs for the same TR, we remove the duplicate datasets and only keep the column with the largest score for the same TR k:

$$Y_{i,k} = \text{argmax}\left\{ M''_{i,j} \mid j \in k \right\}$$

where the $Y_{i,k}$ is the TR enrichment score matrix. In this step, the best-matched dataset of each cell is determined independently according to the largest score within each cell. Here, we set this maximum strategy as default in SCRIP due to its superior performance (Supplementary Figure S3), but also provide an average strategy option, which uses the average of all same TR datasets to represent the TR enrichment score.

Scaling the TR enrichment scores. Next, to stabilize the TR enrichment score and compress the outliers, we introduce the logistic sigmoid function to each TR. The z-score is a step of the sigmoid function, which will shift the data range around 0 for effective logistic sigmoid transformation:

$$z_{i,k} = \frac{Y_{i,k} - \text{mean}\left(Y_k\right)}{\text{std}\left(Y_k\right)}$$

$$Y = \frac{1}{1 + e^{-z}}$$

Finally, SCRIP applies z-score normalization to each cell to ensure that the TR activity scores after sigmoid normalization has a similar dynamic range and achieve better clustering performance:

$$Y_{i,k} = \frac{Y_{i,k} - \text{mean}\left(Y_i\right)}{\text{std}\left(Y_i\right)}$$

To better understand the TR enrichment score calculation and normalization, the distribution changes along with score calculation and normalization were shown in Supplementary Figure S4.

TR targets modeling. To acquire scaled and dependable targets of a TR, SCRIP first imputes the potential binding sites of each TR in each cell. For a specific TR, the best match ChIP-seq dataset of each cell was determined in the TR activity score calculation step. SCRIP then imputes the potential TR binding sites by overlapping the ChIP-seq peak sets with scATAC-seq peaks or intervals of each cell. Since some TR ChIP-seq datasets do not have a sufficient number of peaks and ChIP-seq peak sets are performed on bulk tissue, which may include the peaks from other cell-type, SCRIP further provides a function that uses all the best match TR peak sets of this data found to include other potential binding sites.

With the TR potential binding sites, we can measure the effect on other genes of this TR, or determine the target of this TR, with the RP model. The RP score of a gene is its likelihood of being regulated by a TR and has been used in several previous studies (10,23–25). In general, the RP models could be classified as signal RP model and peak RP model, which are based on scaled signals or binarized peaks, separately. Compared to the signal RP model, the peak RP model used in Cistrome-GO (25) and MAESTRO (10) is more compatible with the binarized signal of scATAC-seq and thus is introduced here. In SCRIP, the formula of the RP score calculation is shown below:

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

where the d_0 is the half-decay distance or regulatory range, and it can be changed by users. The n denotes the number of binding sites near the TSS of gene g. To save the computation time, SCRIP only takes genes within $15d_0$ into account as the score will be less than 0.0005 if a peak is over $15d_0$. The d_i is the distance between the ith peak's center and TSS. This is defined as the simple RP model in SCRIP. The enhanced RP model takes the exon information and nearby genes into account. If a peak is located at the exon's region of a gene, the score of this peak is set to 1 and further normalized by the total exon length of the gene. If a peak locates in the promoter or exon regions of any nearby genes, then the score of this peak is set to 0. Also, for specific TRs, we added a 'auto' mode to automatically determine the d_0 by the percentage of the TR peaks on promoter regions (1 kb around TSS). TRs with >20% peaks on promoter regions are defined as promoter-type TRs and use 1k as the half-decay distance. Other TRs are defined as enhancer-type TRs and use 10k as the half-decay distance.

We compared the performance of the simple RP model and enhanced RP model, which use different half-decay distances for different types of TRs (1k for promoter-type TRs, and 10k for enhancer-type TRs). We applied different RP models to identify the targets for 140 TRs from the PBMC scATAC-seq dataset. To evaluate the performance, we calculated the expression correlation between TR and its top

500 targets using matched PBMC scRNA-seq and used the number of true targets (abs correlation ≥ 0.3 and P-value ≤ 0.01) as the evaluation metric. We first compared the performance of using different regulatory ranges. 1k group represents the half-decay for all TRs were set as 1k, similar for the 10k group. For the auto group, the half-decay is determined by the percentage of the TR peaks on promoter regions. Clearly, the auto half-decay strategy shows an overall higher number of true targets compared to 1k and 10k strategies (Supplementary Figure S5a). We also compared the performance of the simple versus enhanced RP model. Interestingly, the simple RP model seem to have slightly better performance than the enhancer RP model considering all TRs (Supplementary Figure S5b). However, for factors like IRF8 and CEBPA, the enhanced RP model shows better performance, while TBX20 and EBF1 have a different trend. Besides, the two models share a great number of target genes for the same factor (Supplementary Figure S5c, d). These results suggest that the simple and enhanced models do not show a significant difference in identifying target genes. We set the simple RP model and auto half-decay distance as the default parameter for SCRIP, but also provided the enhanced model for users to choose from. With the RP score, we can rank the target genes for each TR, and obtain the top-ranked target genes to build the gene regulatory network.

DATA PROCESSING ON DIFFERENT SCATAC-SEQ DATASETS

PBMC multiome dataset

Preprocessing and TRs activity. The PBMC multiome data was downloaded on the 10X genomics website (https://www.10xgenomics.com/resources/datasets/pbmc-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-10-k-1-standard-2-0-0). The peaks were called using Cell Ranger ATAC 2.0 by fitting the peak signals using a Zero-Inflated Negative Binomial model, combining scATAC-seq reads as a bulk sample. In the scRNA-seq data, cells with <200 genes and genes with less than three cells were removed. We only retained the cells with both RNA-seq counts and ATAC-seq counts. We clustered the scRNA-seq data with the Louvain clustering algorithm and annotated the cell type by cell markers (Supplementary Figure S6a, c, d). Then, we transferred the cell type labels to scATAC-seq data by the matched cell barcodes (Supplementary Figure S6b, e). We applied the SCRIP to the filtered scATAC-seq peak count matrix with the default parameters to evaluate the activity of TRs. The activity scores of different TRs were used to draw heatmap with clustermap of seaborn and project to UMAP with scanpy (Figure 2A, Supplementary Figure S7).

Clustering performance comparison. We compared the clustering performance between TR-based tools (SCRIP, chromVAR), peak-based tools (SCALE, Signac, CisTopic), and bin-based tools (ArchR and SnapATAC) in the PBMC dataset. The scATAC-seq data was preprocessed into TR-cell, peak-cell, or bin-cell matrix to meet the requirements of each tool. All tools were applied to scATAC-seq data with their default parameters. In addition, we use the chromVAR motifs as the motif reference in the process of chromVAR. The Louvain algorithm in Seurat or scanpy was used to perform clustering with the tools. Normalized mutual information score (NMI) and Adjusted rand index (ARI) were calculated with the python package sklearn (Figure 2B).

TR activity and expression correlation. SCRIP TR activity score and chromVAR z-score were used to calculate the Spearman correlation coefficients (SCC) with gene expression. Only the 468 TRs that appear in SCRIP, chromVAR, and gene expression were used to calculate the correlation (Figure 2C, Supplementary Figure S8). We also separated the positive and negative correlations and compared them individually. We defined the TRs with SCC > 0.3 with their expression, and the P-value <0.01 as high-confidence positive regulators, and with SCC <-0.3 and the P-value <0.01 as high-confidence negative regulators. We count the number of high-confidence regulators for both positive and negative regulators between SCRIP and chromVAR, respectively.

Datasets selection evaluation. We evaluated SCRIP's ability for identifying the best-matched dataset using POLR2A, a TR that has the most abundant ChIP-seq references (Supplementary Figure S1c, d, >150 high-quality datasets). We count the number of identified TR datasets for each cell type and the number of cells for specific TR datasets (Supplementary Figure S9). Considering that not all of the single-cells, especially for the rare population such as mast cells, could find TR datasets with matched cell types. We did further analyses to compare the performance of using cell-type-matched strategy, average strategy (average the score for the same TR), and maximum strategy (highest score within the same TR, current model in SCRIP). Only 340 TRs were used for the cell-type-matched strategy due to the relatively low cell type coverage of many TRs. Finally, we checked the TR enrichment distributions of several cell-type-specific TRs using different strategies (Supplementary Figure S3).

H3K27ac targets determination

To capture the loci of rare population cell types, we convert the 10X scATAC-seq data and scCUT&Tag-pro data to the bin-cell matrix with 500 bp. We applied the SCRIP impute function to the scATAC-seq data with HM reference to impute the loci of H3K27ac modifications. The genome track was built by merging the same cell type and normalizing by min-max normalization (Figure 3B, Supplementary Figure S10). With the imputed H3K27ac bin-count matrix, we applied the SCRIP target function to calculate the RP score of each gene and determine the key affected genes with the H3K27ac modification. We applied the same algorithm to calculate the RP score of the scCUT&Tag-pro dataset and the original scATAC-seq dataset. We merged the T cells and monocytes with the maximum RP score, which reflects the driven peaks in the cell type. Also, a bulk T cell and a monocyte dataset's RP score were obtained from Cistrome DB. Then, we calculated the Spearman correlation with the scCUT&Tag-pro dataset and showed the overlap of target genes between the top 1000 RP score targets of each dataset (Figure 3C).

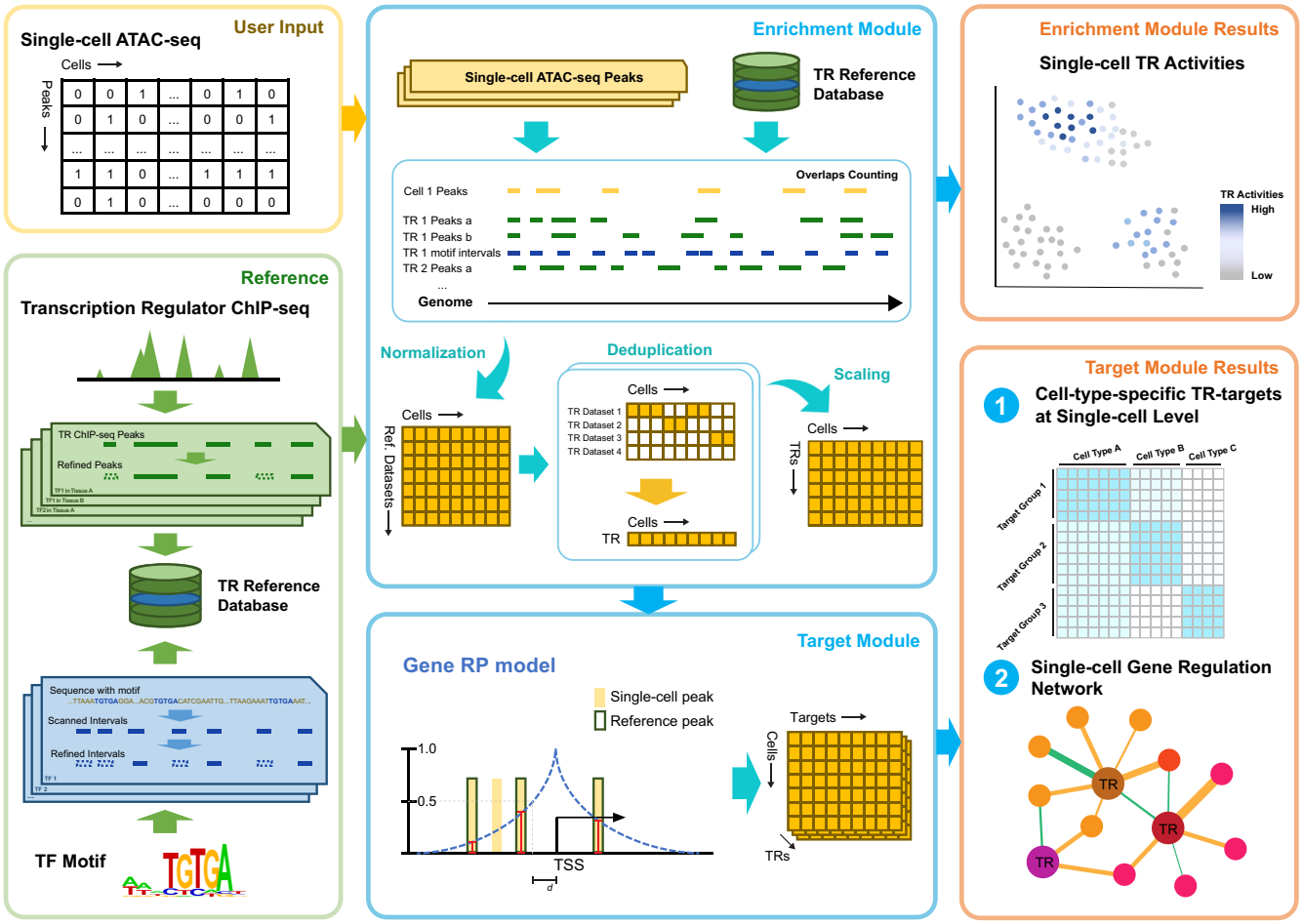

Figure 1. Workflow of SCRIP. Schematic of SCRIP workflow. SCRIP takes the feature count matrix of scATAC-seq as input. The TR ChIP-seq and motif reference datasets were built based on Cistrome ChIP-seq data and Cis-BP motifs, with careful curation. For the enrichment module, the overlaps of scATAC-seq and reference datasets are firstly counted using GIGGLE and further normalized. Then the scores for the same TR were merged and only kept the score of the best-matched dataset for every TR in every cell. Finally, the TR scores were scaled and output. For the Target Module, the best-matched ChIP-seq peaks are combined with scATAC-seq peaks to determine the target gene using the RP model. SCRIP outputs the TR activity, differential targets between cell types, and GRNs at the single-cell level.

HSC dataset

Preprocessing. The HSC scATAC-seq peak count matrix was obtained from the GEO (GSE96769). LiftOver (35) was used to convert genome build from hg19 to hg38. The HSC peaks were generated by MACS2 (36) using bulk hematopoietic data from the same study. We annotated the cell type with the labels from the original study. To evaluate the activity of TRs in each cell, we applied the SCRIP enrich function to the peak count matrix with the default parameters. We performed the unsupervised clustering with the TR activity score and calculated the NMI and ARI with the cell type annotations (Supplementary Figure S11e, Supplementary Table S1). The TRs' activity score matrix was used to do the following analysis.

Trajectory analysis. We applied the R packages destiny (37) to perform the trajectory analysis of HSC. To meet the requirement of data distribution of the destiny, we did an extra normalization step that centralized the activity score at the TR level after the deduplication. The top 600 most variable TRs were used to reconstruct the differential path. The k was set to 4 for the k-nearest neighbor (KNN) algorithm in destiny. The tip was set to 1 to calculate the diffusion pseudo time (DPT). The first two diffusion components were used to draw the diffusion map. We evaluate performances of trajectory between SCRIP, chromVAR, and peak count with the relative distance between starting cell types (HSC) and terminal cell types (monocytes, MEP, CLP and pDC). The start position was indicated using the HSC's average coordinates. We carry out the 0–1 normalization for the four terminal cell types using the start position and every terminal cell type cell. The cell type position was then determined by averaging the coordinates of its cells. Between the start position and cell type position, the Euclidean distance was determined. We projected the HOXA9, GATA1, CEBPB, TCF4 and other TRs activity scores to the diffusion map to show the TRs activity on each branch (Figure 4B, D, Supplementary Figure S11f–j). To visualize the dynamic changes in the TRs' activity in the differential lineages, we showed the TR's activity score with the cell's DPT (Supplementary Figure S11k–z).

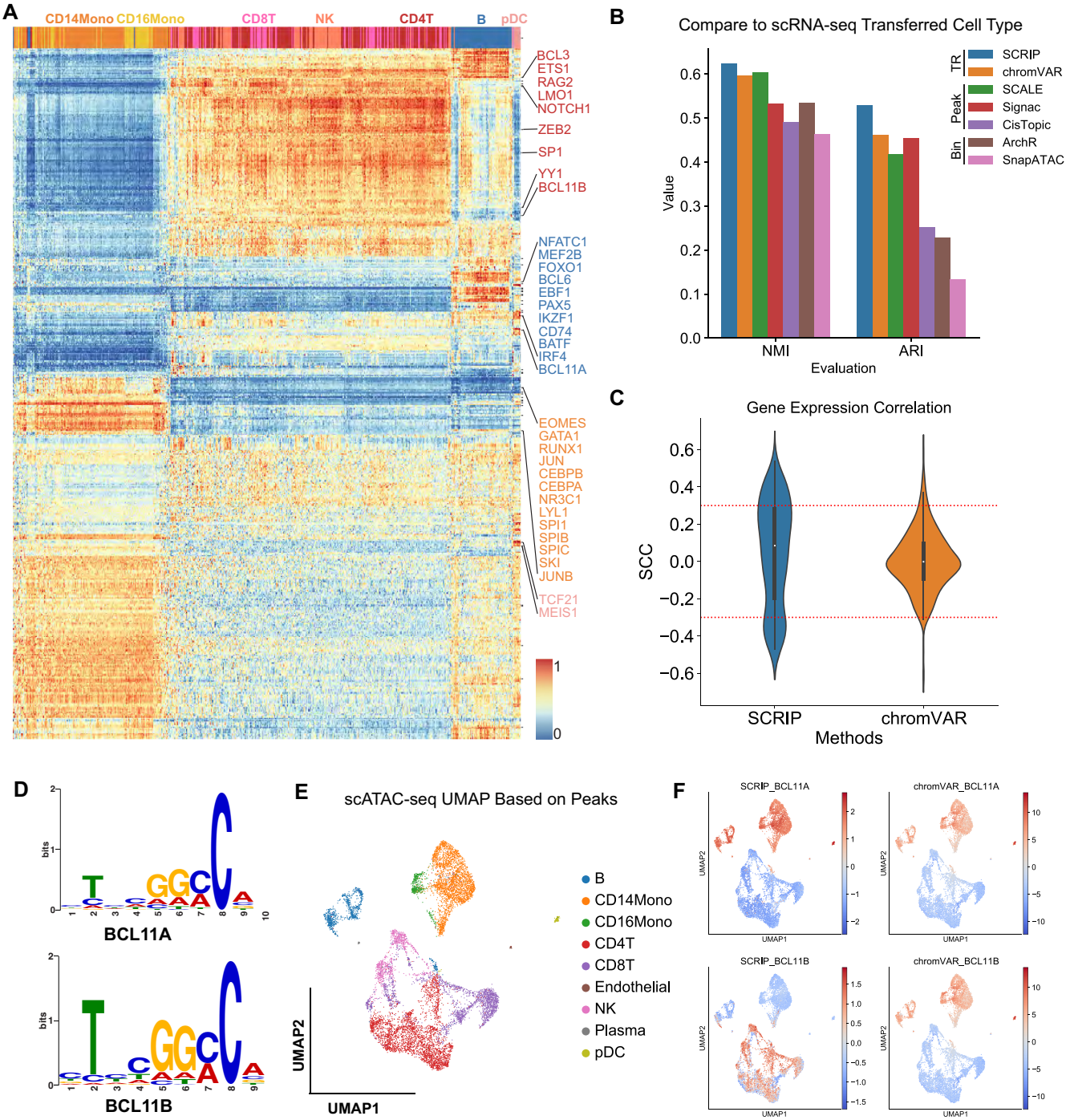

Figure 2. SCRIP achieves better performance on the PBMC dataset. (**A**) Heatmap of TR activity with different cell types. The X-axis denotes the unsupervised clustering results with the TR activity score. (**B**) The clustering performance among SCRIP, chromVAR, SCALE, Signac, CisTopic, ArchR and SnapATAC. Y-axis: NMI scores or ARI scores. (**C**) Distribution of Spearman's correlation coefficients of TR activity and TR expression from SCRIP and chromVAR. (**D**) The SeqLogo of BCL11A and BCL11B motifs. (**E**) Annotation of scATAC-seq with scRNA-seq transferred labels. UMAP was generated by the peak count matrix. (**F**) SCRIP and chromVAR TR activity results of BCL11A and BCL11B in PBMC dataset.

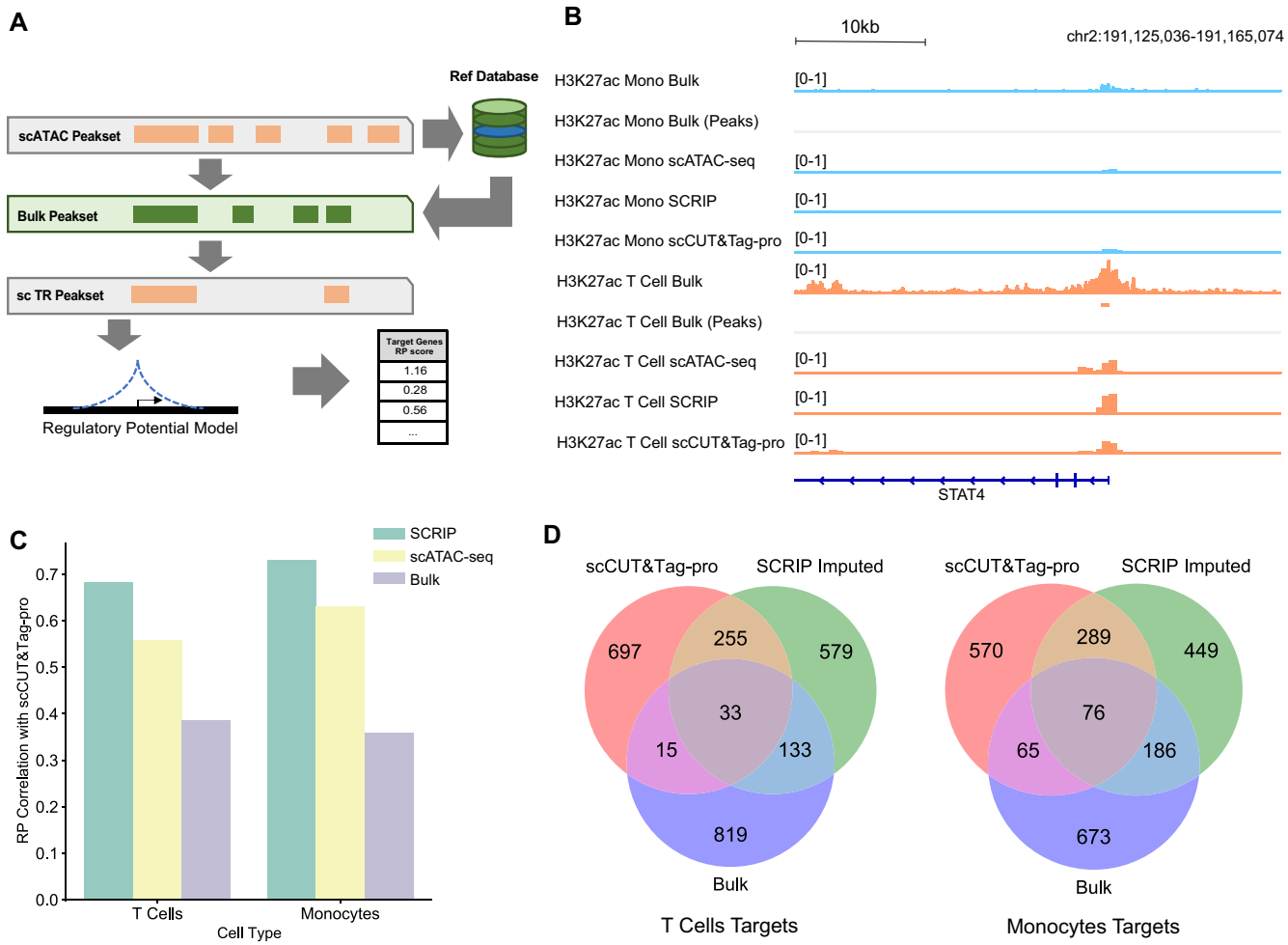

Figure 3. SCRIP enables finding targets from the scATAC-seq data. (**A**) A simple schematic of SCRIP workflow on determining target genes. (**B**) Genome track of monocytes and T cells on H3K27ac signals at STAT4. Light blue: monocytes; orange: T cells. Bulk tracks are read level; scCUT&Tag-pro and SCRIP-inferred tracks are 500 bp bin level. In single-cell tracks, the height of the signal denotes the normalized cell number. (**C**) Spearman correlation coefficients of RP scores between SCRIP imputed, original scATAC-seq and bulk dataset with scCUT&Tag-pro in T cells and monocytes. (**D**) Top 1000 target genes overlap among H3K27ac scCUT&Tag-pro, SCRIP imputed and bulk dataset in T cells and monocytes.

Triangle plot. The activity score of each TR of each cell type was calculated by averaging the TR's activity of all cells in the same cell type. The quantile of the TR in each cell type among all cell types was used to suggest TR's preference for each cell type. The TRs' activity on different lineages were represented by the TR activity score of the terminally differentiated cell type. For example, we used the CLP (common lymphoid progenitor) to represent the lymphoid branch, the monocytes to represent the myeloid branch, and the MEP (megakaryocyte-erythroid progenitor) to represent the erythroid branch. We displayed the positions of TRs on each branch with the ggtern package (Figure 4C).

Human fetal organ datasets

Preprocessing and Clustering. The different human organs scATAC-seq datasets were obtained from the GEO (GSE149683), which provided the filtered peak count matrix and cell labels with the Seurat object format. The peaks of each sample were called using MACS2 by combining scATAC-seq reads as a bulk sample, then the peaks from different samples/organs were merged to generate a union peak set. LiftOver was used to convert genome build from hg19 to hg38. We applied the SCRIP enrich function to the provided scATAC-seq peak count matrix with the default parameters. Then, we used the most variable TRs in each organ and clustered them with R packages ggtree and ComplexHeatmap (Figure 5B–D, Supplementary Figure S12a–c, Supplementary Table S2). We performed the unsupervised clustering with the TR activity score and calculated the NMI with the cell type annotations (Figure 5A, Supplementary Table S1).

Target analysis

We applied the SCRIP impute and target functions with default parameters to determine the GATA3 target genes in the lung, the MYOD1 target genes in the intestine, and the GATA4 and EPAS1 target genes in the liver. To build the credible GRN of the four TRs, we retained the 500 cells with the highest RP in each cell type. The GRNs were built by the R package ggraph. To know the functions of TR's tar-

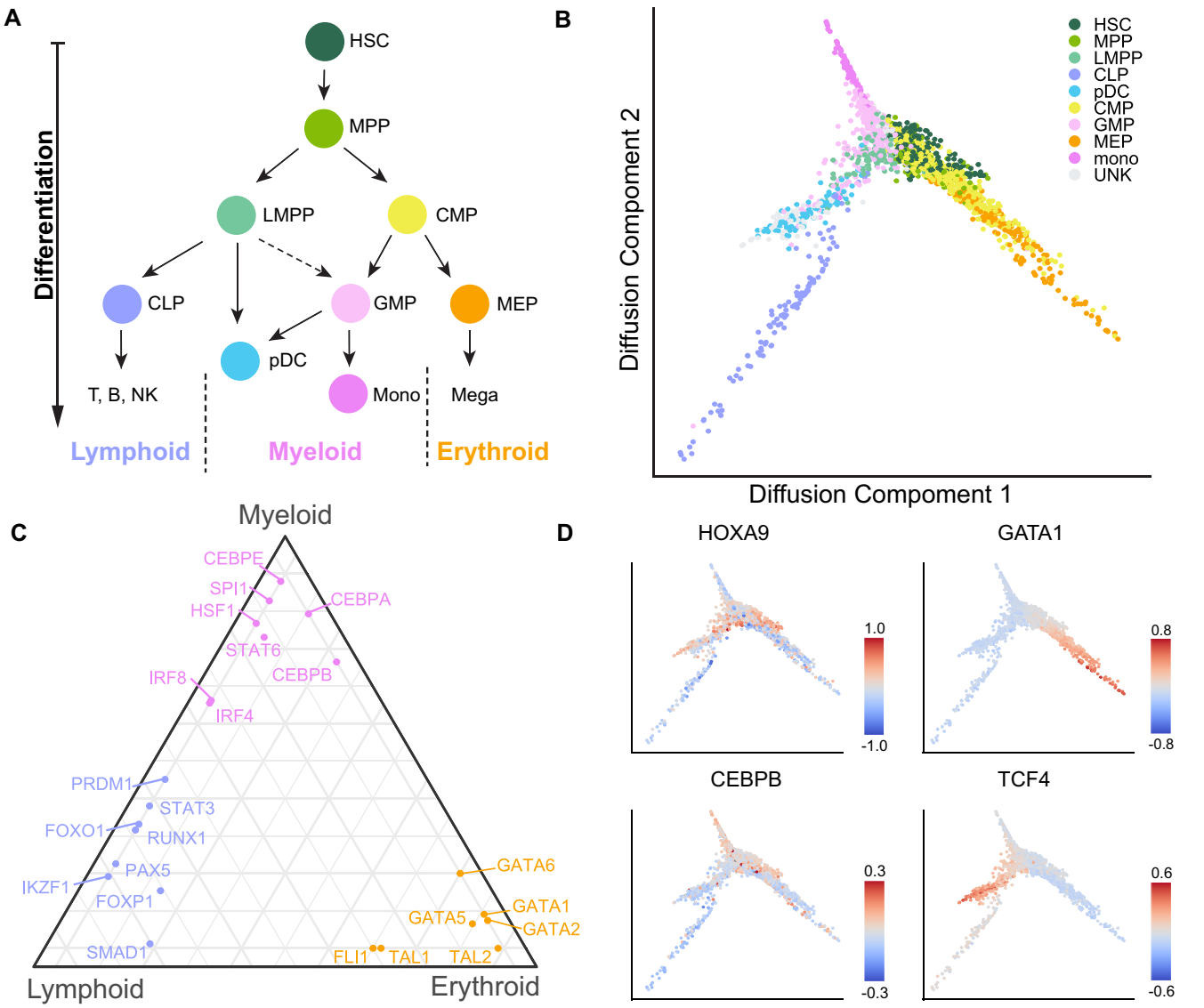

Figure 4. SCRIP reconstructs the path of differentiation of HSCs based on the TR activity. (**A**) Schematic of HSC differentiation. (**B**) Diffusion map of HSC with the cell-type annotations. MPP: multipotent progenitor; LMPP: lymphomyeloid-primed progenitor; CMP: common myeloid progenitor; GMP: granulocyte-macrophage progenitors; UNK: unknown (original study annotation). (**C**) Triangle plot of TRs that regulate HSC differentiation towards three main lineages. (**D**) Projecting HOXA9, GATA1, CEBPB and TCF4 activity onto the diffusion map.

gets, we selected the top 1000 target genes according to the RP score to do the gene ontology (GO) enrichment analysis with the R packages ClusterProfiler (38) (Figure 5E-H, Supplementary Table S3).

BCC tumor microenvironment Datasets

Preprocessing. The scATAC-seq dataset of tumor cells and T cells in BCC was obtained from GEO (GSE129785). Cells were first clustered using a 2.5 kb bin-based method, then the cells from each cluster (cell type) were merged as a pseudo bulk and the peaks were called using MACS2. LiftOver was used to convert genome build from hg19 to hg38. We applied the SCRIP enrich function to the provided scATAC-seq peak count matrix with the default parameters. The averages of TR activity in each cell type were used to plot the heatmap (Figure 6A). The pseudo-time analysis was conducted by the custom scripts from the original study of this dataset. The UMAPs and violin plots of gene expression were obtained from the TISCH database (39).

Target analysis. We applied the SCRIP impute and target functions with default parameters to determine the BATF target genes in terminal TEx cells and the IRF4 target genes in CD4 Tfh cells (Supplementary Figure S13a, b). The JUNB target analysis was done on both naïve CD8 T cells and terminal TEx cells (Figure 6F, G, Supplementary Table S3). Genes with low RP scores, which are considered to have few peaks of this TR, were removed. We normalized RP with the natural logarithm and scaled it for each cell. The FindMarkers function in the R package Seurat was applied to identify the differential target genes of naïve CD8

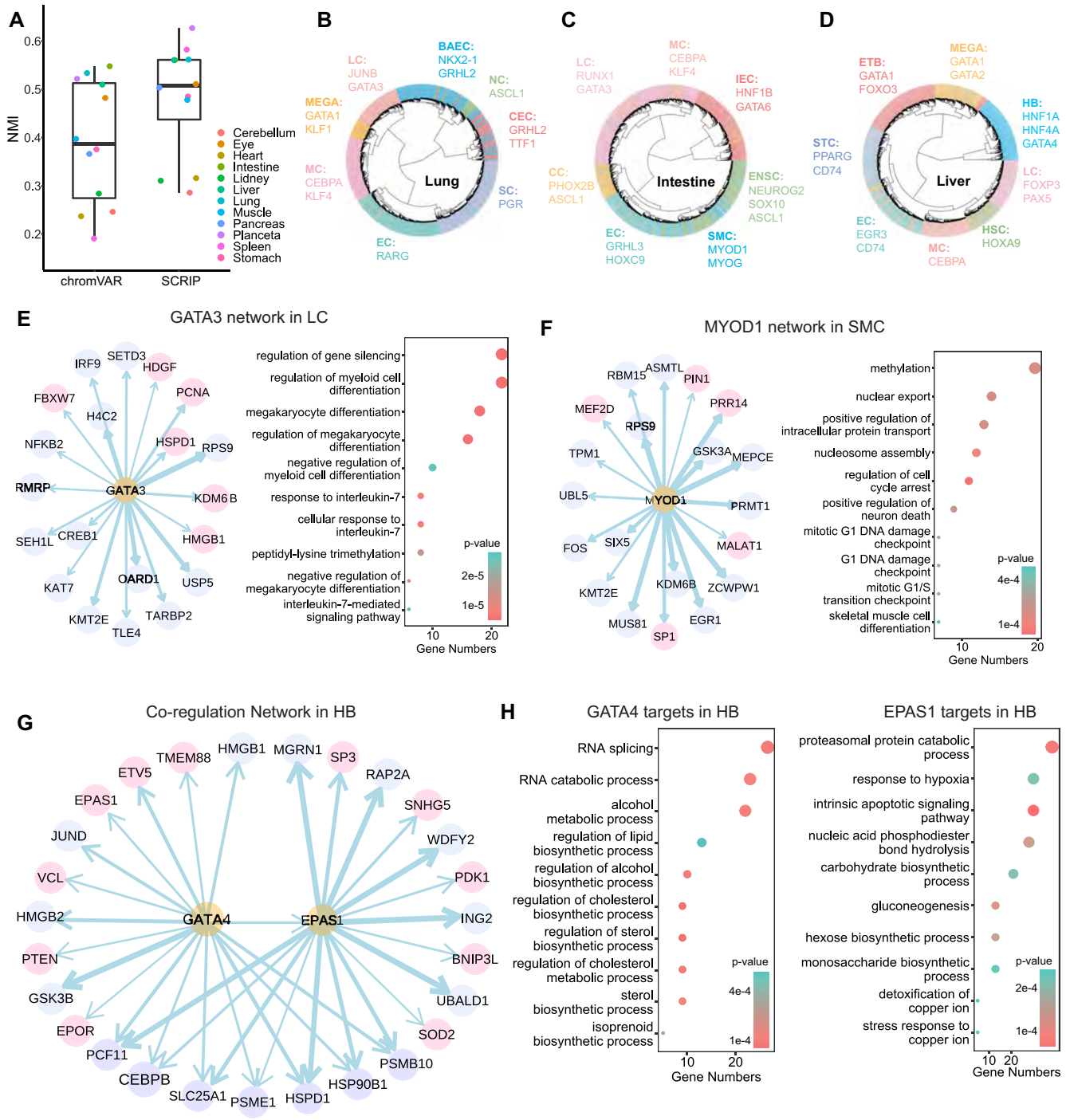

Figure 5. SCRIP finds the key regulators and builds GRNs in human fetal organs. (**A**) Clustering performance of 12 different organs. Only NMI plots here. ARI can be checked in Supplementary Table S1. (**B–D**) Clustering of the human lung (B), intestine (C) and liver (D) scATAC-seq data based on TR activity. Marked TRs of specific cell types are identified by SCRIP with literature support. LC: lymphoid cells; MEGA: megakaryocytes; MC: myeloid cells; EC: endothelial cells; SC: stromal cells; CEC: ciliated epithelial cells; NC: neuroendocrine cells; BAEC: bronchiolar and alveolar epithelial cells; ETB: erythroblasts; HB: hepatoblasts; HSC: hematopoietic stem cells; STC: stellate cells; IEC: intestinal epithelial cells; ENSC: enteric neural stem cells; SMC: smooth muscle cells. (**E**). (left) Inferred GATA3 GRN with human lung's LC scATAC-seq dataset. Pink circles denote the target genes that are supported by previous studies. (right) GO results showed the terms enriched of GATA3 target genes. (**F**) Inferred MYOD1 GRN with human intestine's SMC scATAC-seq dataset. Pink circles denote the target genes that are supported by previous studies. (Right) GO results showed the terms enriched of MYOD1 target genes. (**G**) Inferred GATA4 and EPAS1 co-regulation network with HB cells of human liver scATAC-seq dataset. Pink circles denote the target genes that are supported by previous studies. (**H**) GO results showed the terms enriched of GATA4 target genes and EPAS1 target genes in HB.

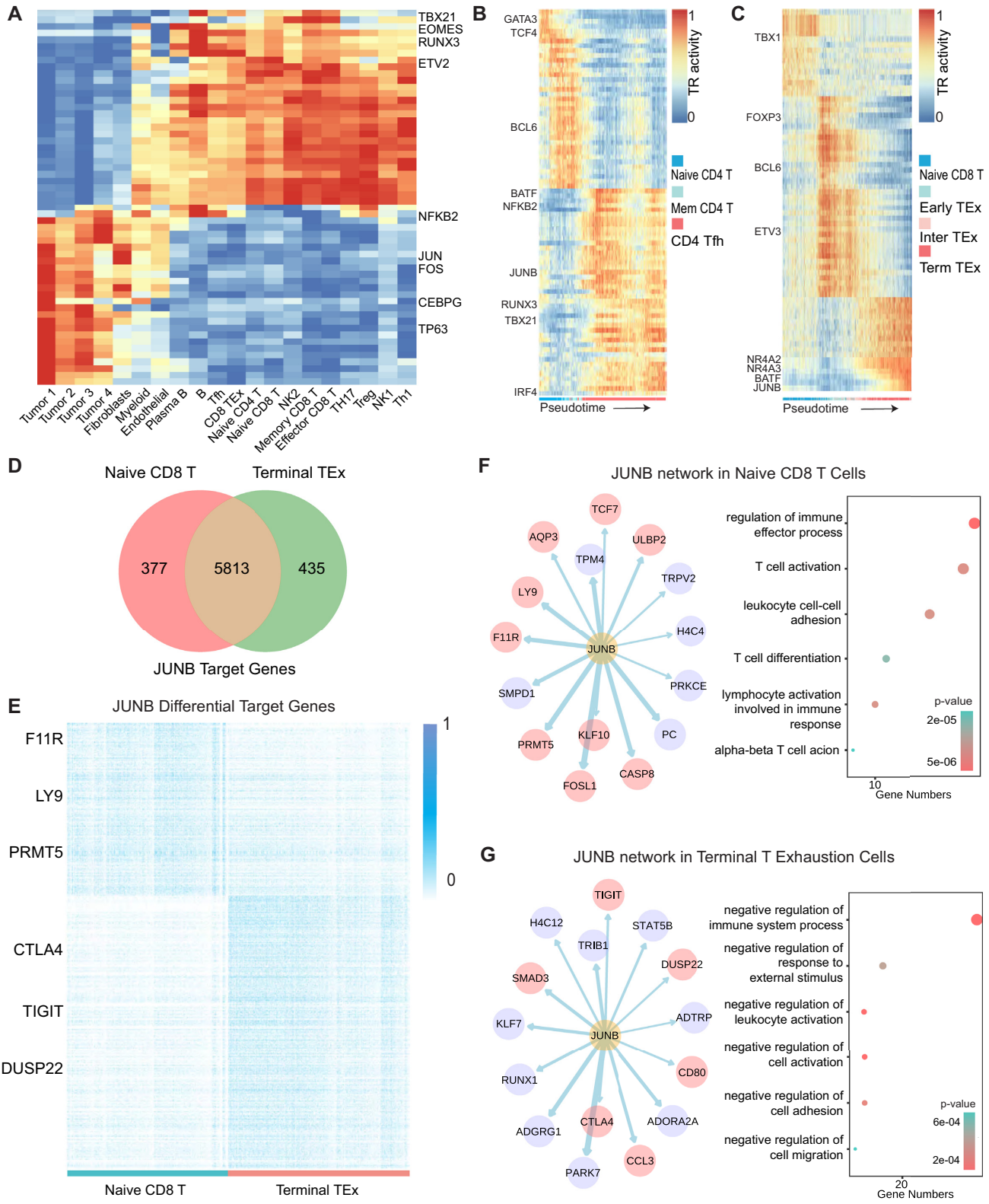

Figure 6. SCRIP uncovers differential targets in the tumor microenvironment. (**A**) Heatmap of cell-type-specific TRs in BCC tumor microenvironment. (**B**) TR activity change during pseudo-time from naive CD4 T cells to CD4 Tfh cells. (**C**) TR activity change during pseudo-time from naive CD8 T cells to terminal TEx cells. (**D**) Overlap of specific target genes of JUNB of naive CD8 T cells and terminal TEx cells. (**E**) Heatmap of RP scores of specific target genes of JUNB of naive CD8 T cells and terminal TEx cells. RP scores were normalized to 0 to 1 with min-max normalization. (**F, G**) (left) Inferred JUNB GRN in naive CD8T cells or terminal TEx cells. Pink circles denote the target genes that are supported by previous studies. (Right) GO results showed the terms enriched of JUNB target genes in naive CD8T cells or terminal TEx cells.

T cells and terminal TEx cells according to the normalized RP score. Different target genes were obtained by 0.25 log fold change and 0.01 *P*-value (Supplementary Figure S13d). The ClusterProfiler was used for GO analysis of these target genes. The GRNs were built using the R package ggraph.

RESULTS

Workflow of the SCRIP

The SCRIP workflow takes the peak count or bin count matrix of scATAC-seq data as input and outputs the TR activity score and their target genes in each cell (Figure 1). We first built a comprehensive reference dataset to help evaluate the TR enrichment in the cells from scATAC-seq data (Supplementary Figures S1 and S2). The reference dataset includes two components. The first one is a TR ChIP-seq reference based on a large collection of 20k ChIP-seq datasets from the Cistrome Data Collection (11k human TRs and 9k mouse TRs) (16,26). We have carefully curated metadata such as factor information, tissue types, and cell types (Supplementary Figure S1a, b). Then we filtered out the ChIP-seq datasets of bad quality and removed low confidence peaks from retained datasets to generate high confidence TR peak sets covering 671 human TRs and 440 mouse TRs (Supplementary Figure S1c, d). Considering there are also TRs without ChIP-seq datasets, we also scanned motifs on the whole genome and obtained the refined intervals with high confidence. These two references were combined to generate the TR reference database containing 1,252 human TRs and 997 mouse TRs in different tissues (Supplementary Figure S1e, f).

Next, we evaluated the TR enrichment in each cell by modeling the peak overlaps between scATAC-seq peaks and TR reference. While scATAC-seq peaks are usually sparse and noisy, we first implemented an imputation step using nearest neighbor cells. Then, we calculated the intersections of each TR dataset or motifs in every single cell. This score was further normalized by the number of reference peaks and length of scATAC-seq peaks in each cell. For each TR, there may be ChIP-seq datasets from different tissues or cell lines, we deduplicate the TR score matrix and keep the TRs with the largest score as the best-matched tissues or dataset for this cell. This generated a normalized TR activity score-by-cell matrix and can be further used to perform clustering, lineage tracing, and other downstream analyses (Supplementary Figures S3 and S4). After identifying the best-matched ChIP-seq dataset for a TR, we can combine the ChIP-seq peaks with scATAC-seq peaks, and apply the regulatory potential (RP) model (10,23,25) to quantitatively evaluate the TR enrichment on its target genes for each cell (Supplementary Figure S5). The RP scores reflect the TR regulation ability of its target genes and can be used to construct single-cell GRNs for that TR. Overall, SCRIP will output the TR activity, candidate TR targets, and TR GRNs at single-cell resolution.

TR activity performance evaluation using PBMC multiome dataset

To systematically evaluate the performance of SCRIP, we applied it to a peripheral blood mononuclear cell (PBMC) dataset that was produced using the 10X Genomics Multiome platform, which generates scRNA-seq and scATAC-seq in the same cell. We annotated the dataset with cell-type markers from the scRNA-seq dataset and transferred the cell-type labels to the scATAC-seq dataset (Supplementary Figure S6a–e). SCRIP successfully finds the key TRs in the corresponding cell types, for example, CEBPA and CEBPB are enriched in monocytes, and PAX5 and BCL6 are enriched in B cells (Figure 2A, Supplementary Figure S7). Also, the cells can be well clustered to their cell type lineages using TR activity alone (Supplementary Figure S3). We also compared the consistency of the clustering results with scRNA-seq transferred labels and benchmarked them using normalized mutation information (NMI) and Adjusted Rand index (ARI). Interestingly, clustering using SCRIP TR activity scores shows better consistency with scRNA-seq transferred cell-type compared to existing motif-based methods such as chromVAR, and peak or bin-based methods such as SCALE (11), Signac (8), CisTopic (40), ArchR (41) and SnapATAC (42) (Figure 2B, Supplementary Table S1). This result suggests that SCRIP could accurately predict TR activity at the single-cell level, which should show superior performance in determining cell-type lineages.

Next, we compared TR activity with its gene expression. We compared the Spearman correlation coefficients (SCC) distribution of TRs activity scores and its gene expression for both SCRIP and chromVAR. Interestingly, SCRIP has a larger dynamic range for both positive and negative correlations (Figure 2C, S8a). The majority of chromVAR correlations were around 0, indicating that the motif information might not be able to capture real TR activity. Also, SCRIP identifies more high-confidence TRs for both positive and negative regulators (Supplementary Figure S8b). Compared to chromVAR, SCRIP also has generally higher correlations with gene expression in individual cell types (Supplementary Figure S8c–k). Also, SCRIP correctly estimated the activity of factors with similar motifs. For example, previous studies have suggested that BCL11A is required for the generation of B progenitor cells (43), while BCL11B activates the transcription of interleukin-2 during T cell activation (44). These two factors share similar motifs but are expressed in distinct lineages (Figure 2D, Supplementary Figure S6c). Consistently, SCRIP predicts BCL11A to be enriched in B-cells and myeloid lineages, while BCL11B is enriched in T and NK-cells (Figure 2E, F). On the contrary, the chromVAR score shows no significant difference between BCL11A and BCL11B, which is biased by the similar motif sequences (Figure 2E, F). These results suggest that SCRIP can identify tissue-specific regulations even for factors with similar motifs, which cannot be achieved by motif-based methods.

Finally, we evaluated whether SCRIP can identify cell-type-specific regulations for the same TR. Due to the relative sparsity of TR ChIP-seq datasets, there are slightly more TRs that were only covered by motifs than ChIP-seq (Supplementary Figure S3a). We next compared the performance of TR ChIP-seq datasets and motif datasets on the human PBMC dataset. For each TR, we calculated the percentage of cells that selected the motif dataset as the best-matched dataset. Interestingly, for the 335 shared TRs, most of them tend to find ChIP-seq datasets than mo-

tif datasets (Supplementary Figure S3b). These results suggest although the motifs could serve as a complementary reference to fulfill the TR reference, the ChIP-seq dataset still carries more information than motif datasets for the TRs with both ChIP and motif information. POLR2A is the TR with the most abundant ChIP-seq data in various tissue types (Supplementary Figure S1c, d). We tested the ability of SCRIP to find the correct POLR2A ChIP-seq dataset for different single cells. As we expected, for most of the T cells, B cells, and monocytes in PBMC scATAC-seq datasets, SCRIP could successfully identify the corresponding TR ChIP-seq datasets (Supplementary Figure S9a–c). If we focused on specific datasets, they were also assigned to the cells with matched cell types (Supplementary Figure S9d–f). These results suggest that SCRIP could accurately find the TR datasets with matched cell type information for each cell. In summary, our analyses suggest that SCRIP could accurately predict TR activity at the single-cell level, identify tissue-specific regulations, and find the correct TR dataset for different single-cells.

TR targets evaluation using PBMC scCUT&Tag-pro datasets

The main purpose of ChIP-seq experiments is to find target genes for TR, which is crucial for constructing GRNs. As SCRIP can correctly match the TR ChIP-seq dataset for single cells from different lineages, we asked whether integrating bulk ChIP-seq data and single-cell accessibility could impute the ChIP-seq signals and further identify TR targets at the single-cell level. We thus predicted the TR peaks of each cell with the best match bulk ChIP-seq dataset and applied a modified RP model to infer the putative targets of TR on each cell (Figure 3A, see Materials and Methods). Several single-cell ChIP-seq profiles are available for HMs and a few TRs using scCUT&Tag (20,22) and scCUT&Tag-pro. While the TR scCUT&Tag data is of low quality, we benchmarked our method using several published HM scCUT&Tag-pro datasets (21).

H3K27ac modification is an active enhancer marker that has been profiled using scCUT&Tag-pro in PBMC (21). We built a reference dataset with active histone modifications including H3K27ac and imputed the H3K27ac signal using the PBMC scATAC-seq dataset (Supplementary Figure S2). While the scATAC-seq cells and scCUT&Tag-pro cells are not from the same populations, we cannot compare the performance at the single-cell level. However, when we piled up the H3K27ac scCUT&Tag-pro signal and the SCRIP imputed signal for different cell types, we found that SCRIP could accurately identify the T-cell-specific peaks around STAT4, a TF that plays an important role in T cells (Figure 3B). Many regions with only scATAC-seq peaks were removed from SCRIP. The CRAMP1 and FAM22A loci show both scATAC-seq signals for monocytes and T cells. However, these two loci do not have H3K27ac signals from the bulk data and they are not output by SCRIP, which were also not observed in the scCUT&Tag-pro data. For the locus of SCL24A42, SCRIP could accurately predict the T-cell-specific H3K27ac signal based on the combination of bulk H3K27ac signal and scATAC-seq peak (Supplementary Figure S10). Besides, when we calculated correlations between the real scCUT&Tag-pro RP, SCRIP imputed RP, scATAC-seq RP, and bulk H3K27ac RP, the SCRIP imputed RP shows the highest consistency with the scCUT&Tag-pro RP, indicating its better ability and accuracy in identifying H3K27ac regulated genes (Figure 3C). More specifically, when comparing the top 1000 H3K27ac regulated genes in T-cells and monocytes, SCRIP imputed RP could identify more common target genes in scCUT&Tag-pro data than using bulk H3K27ac data directly (Figure 3D). These results collectively suggest that integrating scATAC-seq data with bulk TR or HM ChIP-seq data could accurately identify their target genes.

SCRIP underlies differentiation paths for human HSC differentiation

TRs are often the driving source of cellular differentiation. To prove that SCRIP can infer TR activity in a complex system and could be potentially used to track cell differentiation, we applied SCRIP on a human hematopoietic stem cell (HSC) differentiation scATAC-seq dataset (45). The HSC differentiation is a well-characterized system, with HSCs differentiating into three different major lineages (Figure 4A). SCRIP also achieved the second-best result in all methods and shows the best performance in the TR-based method in the clustering performance (Supplementary Figure S11e). After identifying TR activity in different HSC subpopulations, we performed a pseudo-time analysis and reconstructed the differentiation trajectory of HSC using TR activity (Figure 4B and Supplementary Figure S11a–c, e). The diffusion map of SCRIP suggests that HSC was differentiated into three major directions, CLP (common lymphoid progenitor), monocytes, and MEP (megakaryocyte-erythroid progenitor), with a little spike towards pDC (plasmacytoid dendritic cells) (Figure 4B). These directions are perfectly aligned with the known differentiation path of HSCs. By contrast, the diffusion map generated using the original peak count matrix showed a relatively vague separation for different lineages (Supplementary Figure S11a). In addition, we have calculated the averaged distance of terminally differentiated cells (monocytes, MEP, CLP and pDC) versus HSC, SCRIP showed the largest distance compared to using peak-count matrix and chromVAR results, indicating a better lineage separation result (Supplementary Figure S11a–d).

Next, we sought to identify the driven TRs for the three major differentiation lineages. We use the average TR activity of CLP, monocytes, and MEP to denote the lineage lymphoid, myeloid and erythroid respectively. Our results correctly distinguish and locate the key TRs into different lineages (Figure 4C). For example, GATA1 and SPI1 are well-known mutually inhibiting TFs acting as fate-determining regulators in the hematopoietic system. GATA1 specifies the erythroid lineages while SPI1 specifies the myeloid lineage (46,47), which is highly consistent with the SCRIP results (Figure 4C). We also found other well-known regulators show high activity in their corresponding lineages, such as HOXA9 for HSCs, CEBPB for myeloid lineages, and TCF4 for lymphoid lineages (Figure 4D, Supplemen-

tary Figure S11f–i). Besides, the dynamic changes in the TRs' activity of the differential lineages indicate their potential role in lineage differentiation (Supplementary Figure S11j–y). These results prove that SCRIP enables the trajectory analyses of scATAC-seq with known driver TR activity.

SCRIP constructs GRNs in human fetal organ development

To prove the ability that SCRIP can be applied to diverse tissue types and infer the target genes of TRs, we applied SCRIP to a scATAC-seq dataset of human fetal organs that covers 14 different tissues (48). The TR activity score showed a better performance in clustering compared to the motif-based method chromVAR in almost all these tissues (Figure 5A, Supplementary Table S1). To check whether SCRIP could identify the TRs that are involved in the production and maintenance of specific cell types in different organs, we focused on the lung, intestine, and liver datasets. Again, SCRIP could correctly identify the cell-type-specific TRs in these three different organs (Figure 5B–D, Supplementary Figure S12a–c, Supplementary Table S2). For example, GRHL2 and its downstream direct target gene NKX2-1 form a positive feedback loop to connect lung epithelial cell identity, migration, and lung morphogenesis (49) (Figure 5B, bronchiolar and alveolar epithelial cells, BAEC). GATA6 regulates the development of primitive intestinal cells (50) (Figure 5C, intestinal epithelial cells, IEC). HNF1A, HNF4A and GATA4 are well-known hepatocyte TFs in liver tissues (51) (Figure 5D, hepatoblasts, HB). These results proved that SCRIP can not only cluster the same cell type with TRs activity but also identify crucial TRs in different cell types using chromatin accessibility data.

Master TRs and their cofactors regulate each other or co-regulate downstream target genes, forming a potential GRN that could modulate cell fate and identities. To validate the ability of SCRIP to establish cell-type-specific GRNs, we inferred the potential target genes of TRs and built the cell-type-specific GRNs for different organs (Figure 5E–H, Supplementary Table S3). In the lung, we identified the target genes of GATA3, which is mainly enriched in the lymphoid cells (LC) (Figure 5B). The target genes of GATA3 mainly contribute to immune functions through responding to interleukin-7 (IL-7) and negatively regulating the differentiation of myeloid cells, which is in line with previous studies (52) (Figure 5E). In the intestine, MYOD1 controls the differentiation of smooth muscle cells (SMC) by regulating its downstream genes and function (53) (Figure 5C, F). Finally, we built a co-regulatory GRN of GATA4 and its downstream targets EPAS1 (54) in liver hepatoblasts (HB) (Figure 5G). Although their downstream target genes show a great difference (Supplementary Figure S12d, e), the GO analysis suggests that the functions are both enriched in the biosynthetic process. In addition, GATA4 tends to regulate alcohol metabolism, while EPAS1 targets are enriched in response to hypoxia (55,56) (Figure 5H). These results show that SCRIP allows identifying the targets of different TRs in diverse cell types and constructing GRNs of multiple TRs in the same cell.

Disease-specific GRNs identified by SCRIP in the tumor microenvironment

The target genes of TR can be changed due to different cooperation of co-regulators, especially under disease status (57). We applied SCRIP to a basal cell carcinoma (BCC) tumor microenvironment (TME) dataset (58) to investigate how TRs and their target genes were changed in different cell states under disease status. First, we confirmed the TRs activity is accurately predicted in the corresponding cell types (Figure 6A). For instance, CEBPG, a TF that promotes cancer development by enhancing the PI3K–Akt signaling pathway (59), was found to be robustly more active in tumor cells than in other cells. In addition, the activity of TFs such as EOMES and TBX21 were higher in immune cells than in tumor cells (Figure 6A), which is consistent with the role of these TFs in driving lymphocyte differentiation (60,61).

T cells are the major cytotoxic cells responsible for antitumor immunity. Diverse T cell differentiation paths and phenotypes drive the immune response in TME. We performed the pseudo-time analysis of T cells in TME using the TR activity from SCRIP, which uncovered two distinct paths. The first differentiation path is from naïve CD4 T cells to T follicular helper (Tfh) cells, for which IRF4 is gradually activated in Tfhs (Figure 6B). The IRF4 activity was significantly increased in Tfh, and the function of its target genes was also enriched in lymphocyte activation and differentiation (Supplementary Figure S13a). These analyses are consistent with the IRF4 function in Tfh cell expansion (62). Another path is from naïve CD8 T cells to terminal T exhaustion (TEx) cells. BATF, a key regulator of T cell exhaustion (63), has higher activity in terminal TEx (Figure 6C). Consistently, the BATF target genes tend to have an immunosuppressive effect on terminal TEx cells (Supplementary Figure S13b). These analyses suggest that the TR activity and targets inferred by SCRIP could be used to track cell state changes under the disease condition.

Interestingly, we found that the activity of JUNB is both higher in naïve CD8 T and terminal TEx (Supplementary Figure S13c). We then checked the target genes of JUNB between these two cell types. Although most targets were shared, there are a considerable number of differential targets between these two stages (Figure 6D, E). We asked whether JUNB has different functions in naïve CD8 T cells and terminal TEx cells, then we examined their differential target genes and built cell-type-specific GRNs for JUNB (Supplementary Figure S13d, Figure 6F, G, Supplementary Figure S14). We found that PRMT5, which is critical for the transition of naïve T cells to the effector or memory phenotype (64), is presented in the JUNB GRNs only in naïve CD8 T cells (Figure 6E, F). In contrast, CTLA4, which could encode a protein that transmits an inhibitory signal to T cells and its upregulation has been described as a marker of T cell exhaustion in chronic infections and cancer (65,66), has a high RP score in JUNB GRNs in terminal TEx (Figure 6E, G). The function enrichment results suggest that JUNB mainly tends to function as a positive regulator of T cell activation and migration to lymphoid organs, while negatively modulating the immune system process in terminal TEx cells. In summary, our analyses suggest that SCRIP can

identify cell-type-specific GRNs as well as uncover disease-specific GRNs in complex biological systems.

DISCUSSION

In this study, we present SCRIP, a computational workflow for single-cell gene regulation inference by large-scale data integration. We first built a manually curated and comprehensive epigenome reference dataset including 11k human and 9k mouse TR ChIP-seq datasets. Based on the reference, we developed a method that can evaluate TR activity and build GRNs at the single-cell resolution using scATAC-seq. Our method achieves better performance compared to the previous motif-based methods in terms of clustering accuracy, consistency with gene expression, and ability to discriminate factors within the same family. We applied SCRIP to four different biological systems, including PBMC, HSC differentiation, human fetal organ development, and the BCC tumor microenvironment. SCRIP does not merely identify the key TRs in different cell types under diverse biological settings. In addition, the TR activity predicted by SCRIP could be used to trace the cell lineages and identify lineage-specific regulators. The single-cell GRNs constructed by SCRIP enable the identification of the co-regulation relationship between different TRs and reveal the disease-associated GRNs in the terminal exhausted T-cells from the tumor microenvironment.

Although in SCRIP, the ChIP-seq-based method outperforms motif-based methods in many aspects, there are still several limitations. First, our method significantly relies on the data quality of the ChIP-seq datasets. After filtering the 20k human and mouse datasets, there are only 4k ChIP-seq datasets with good data quality. This significantly reduced the number of TRs as well as different types of tissues covered by our reference. To compensate for this, we also integrate the motif scanning results into the TR reference. Second, there might be potential batch effects between the TR ChIP-seq data with the scATAC-seq data. To avoid this, for each cell we score TR enrichment using multiple TR ChIP-seq from different tissues, and only keep the one with the highest TR enrichment score, which is usually from the same cell type. This could partially solve the batch effect for TRs with a large number of datasets, but may not be appropriate for TRs with few numbers of matched datasets. Third, the experiment of the public TR ChIP-seq may have been performed with different perturbations, which may alter the TR's binding sites and introduce biases to our results. Finally, the bulk-level ChIP-seq datasets have the probability of losing the signals on rare populations, which also impacts the results of binding sites and target genes for our method, especially for some minority populations.

We foresee several ways to further improve our method. First, there will be an increasing number of TR ChIP-seq, CUT&RUN, and CUT&Tag datasets in the future. The first version of Cistrome DB includes 13 366 human and 9953 mouse epigenome datasets, while the number almost doubled to 25 000 human and 22 000 mouse epigenome datasets after only 2 years (16,26). Large consortiums like ENCODE, and Epigenome Roadmap will also generate a great number of TR datasets with high quality. With the development of scCUT&RUN and scCUT&Tag-pro, we could also integrate the single-cell TR dataset into our reference for annotating TRs from scATAC-seq in other cell types. These expanded references will improve the performance of our method for predicting TR activity. Second, machine learning algorithms, such as generative adversarial networks (GAN) could be used to generate more TR ChIP-seq datasets *in silico*. Finally, although we demonstrated that SCRIP is powerful in predicting TR activity and GRNs for scATAC-seq, we could potentially extend its applications to scRNA-seq. The Cistrome DB also has a decent collection of public ATAC-seq and DNase-seq datasets, which could be used to infer the chromatin accessibility for scRNA-seq, then infer the TR activity using the predicted accessibilities. In addition, gene expression correlation could be considered to increase the accuracy of constructing single-cell GRNs. With the implementation of those features, we anticipate SCRIP to help researchers identify driver TRs and interpret single-cell GRNs in different biological areas.

DATA AVAILABILITY

PBMC multiome dataset is available on the 10X genomic website (https://www.10xgenomics.com/resources/datasets/pbmc-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-10-k-1-standard-2-0-0). scCUT&Tag-pro H3K27ac dataset was obtained from their original studies: https://zenodo.org/record/5504061. Other datasets analyzed during the current study are available in the GEO with the following accession: HSC (GSE96769), human fetal organ (GSE149683), BCC tumor microenvironment (GSE129785). SCRIP is an open-source python package with source code freely available at: https://github.com/wanglabtongji/SCRIP. The analysis codes and reference processing codes in this paper are available at https://github.com/wanglabtongji/SCRIP_notebook.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank the Bioinformatics Supercomputer Center of Tongji University for offering computing resources.
Author contributions: C.W. conceived and supervised the project. X.D. designed and implemented the SCRIP algorithm. X.D. collected and preprocessed the ChIP-seq and motif datasets, and built the SCRIP index. X.D., K.T. and Y.X. evaluated the performance. X.D. performed the analysis of PBMC. K.T. performed the analysis of HSC and human fetal organs. Y.X. performed the analysis of T cells in the tumor microenvironment. X.D., K.T., Y.X. and C.W. wrote the manuscript with the help of other authors. All authors read and approved the final manuscript.

FUNDING

National Natural Science Foundation of China [32170660]; Shanghai Rising Star Program [21QA1408200]; Natural Science Foundation of Shanghai [21ZR1467600]; Fundamental Research Funds for the Central Universities

[20002150073]. Funding for open access charge: National Natural Science Foundation of China.
Conflict of interest statement. None declared.

REFERENCES

1. Mortazavi,A., Williams,B.A., McCue,K., Schaeffer,L. and Wold,B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.
2. Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J., Ziller,M.J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
3. Ramani,V., Deng,X., Qiu,R., Gunderson,K.L., Steemers,F.J., Disteche,C.M., Noble,W.S., Duan,Z. and Shendure,J. (2017) Massively multiplex single-cell Hi-C. *Nat. Methods*, **14**, 263–266.
4. 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.
5. Schep,A.N., Wu,B., Buenrostro,J.D. and Greenleaf,W.J. (2017) chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods*, **14**, 975–978.
6. Fu,L., Zhang,L., Dollinger,E., Peng,Q., Nie,Q. and Xie,X. (2020) Predicting transcription factor binding in single cells through deep learning. *Sci. Adv.*, **6**, eaba9031.
7. 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.
8. 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.
9. Danese,A., Richter,M.L., Chaichoompu,K., Fischer,D.S., Theis,F.J. and Colomé-Tatché,M. (2021) EpiScanpy: integrated single-cell epigenomic analysis. *Nat. Commun.*, **12**, 5228.
10. Wang,C., Sun,D., Huang,X., Wan,C., Li,Z., Han,Y., Qin,Q., Fan,J., Qiu,X., Xie,Y. *et al.* (2020) Integrative analyses of single-cell transcriptome and regulome using MAESTRO. *Genome Biol.*, **21**, 198.
11. Xiong,L., Xu,K., Tian,K., Shao,Y., Tang,L., Gao,G., Zhang,M., Jiang,T. and Zhang,Q.C. (2019) SCALE method for single-cell ATAC-seq analysis via latent feature extraction. *Nat. Commun.*, **10**, 4576.
12. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.
13. Park,P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, **10**, 669–680.
14. Ernst,J. and Kellis,M. (2015) Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat. Biotechnol.*, **33**, 364–376.
15. Consortium,E.P., Dunham,I., Kundaje,A., Aldred,S.F., Collins,P.J., Davis,C., Doyle,F., Epstein,C.B., Frietze,S., Harrow,J. *et al.* (2013) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
16. Zheng,R., Wan,C., Mei,S., Qin,Q., Wu,Q., Sun,H., Chen,C.-H., Brown,M., Zhang,X., Meyer,C.A. *et al.* (2019) Cistrome data browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Res.*, **47**, D729–D735.
17. Consortium,RoadmapEpigenomics, Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–329.
18. Layer,R.M., Pedersen,B.S., DiSera,T., Marth,G.T., Gertz,J. and Quinlan,A.R. (2018) GIGGLE: a search engine for large-scale integrated genome analysis. *Nat. Methods*, **15**, 123–126.
19. Hainer,S.J., Bošković,A., McCannell,K.N., Rando,O.J. and Fazzio,T.G. (2019) Profiling of pluripotency factors in single cells and early embryos. *Cell*, **177**, 1319–1329.
20. Wu,S.J., Furlan,S.N., Mihalas,A.B., Kaya-Okur,H.S., Feroze,A.H., Emerson,S.N., Zheng,Y., Carson,K., Cimino,P.J., Keene,C.D. *et al.* (2021) Single-cell CUT&Tag analysis of chromatin modifications in differentiation and tumor progression. *Nat. Biotechnol.*, **39**, 819–824.
21. Zhang,B., Srivastava,A., Mimitou,E., Stuart,T., Raimondi,I., Hao,Y., Smibert,P. and Satija,R. (2022) Characterizing cellular heterogeneity in chromatin state with scCUT&Tag-pro. *Nat. Biotechnol.*, **40**, 1220–1230.
22. Bartosovic,M., Kabbe,M. and Castelo-Branco,G. (2021) Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. *Nat. Biotechnol.*, **39**, 825–835.
23. Wang,S., Sun,H., Ma,J., Zang,C., Wang,C., Wang,J., Tang,Q., Meyer,C.A., Zhang,Y. and Liu,X.S. (2013) Target analysis by integration of transcriptome and chip-seq data with BETA. *Nat. Protoc.*, **8**, 2502–2515.
24. Wang,S., Zang,C., Xiao,T., Fan,J., Mei,S., Qin,Q., Wu,Q., Li,X., Xu,K., He,H.H. *et al.* (2016) Modeling cis-regulation with a compendium of genome-wide histone H3K27ac profiles. *Genome Res.*, **26**, 1417–1429.
25. Li,S., Wan,C., Zheng,R., Fan,J., Dong,X., Meyer,C.A. and Liu,X.S. (2019) Cistrome-GO: a web server for functional enrichment analysis of transcription factor chip-seq peaks. *Nucleic Acids Res.*, **47**, W206–W211.
26. Mei,S., Qin,Q., Wu,Q., Sun,H., Zheng,R., Zang,C., Zhu,M., Wu,J., Shi,X., Taing,L. *et al.* (2017) Cistrome data browser: a data portal for chip-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res.*, **45**, D658–D662.
27. Zheng,R., Dong,X., Wan,C., Shi,X., Zhang,X. and Meyer,C.A. (2020) Cistrome Data Browser and Toolkit: analyzing human and mouse genomic data using compendia of ChIP-seq and chromatin accessibility data. *Quant. Biol.*, **8**,267–276.
28. Bell,O., Tiwari,V.K., Thomä,N.H. and Schübeler,D. (2011) Determinants and dynamics of genome accessibility. *Nat. Rev. Genet.*, **12**, 554–564.
29. Klemm,S.L., Shipony,Z. and Greenleaf,W.J. (2019) Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.*, **20**, 207–220.
30. Zhang,T., Zhang,Z., Dong,Q., Xiong,J. and Zhu,B. (2020) Histone H3K27 acetylation is dispensable for enhancer activity in mouse embryonic stem cells. *Genome Biol.*, **21**, 45.
31. Weirauch,M.T., Yang,A., Albu,M., Cote,A.G., Montenegro-Montero,A., Drewe,P., Najafabadi,H.S., Lambert,S.A., Mann,I., Cook,K. *et al.* (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**, 1431–1443.
32. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Mol. Cell*, **38**, 576–589.
33. Dunham,I., Kundaje,A., Aldred,S.F., Collins,P.J., Davis,C.A., Doyle,F., Epstein,C.B., Frietze,S., Harrow,J., Kaul,R. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
34. Quinlan,A.R. (2014) BEDTools: the swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinformatics*, **47**, 11.12.1–11.12.34.
35. Kuhn,R.M., Haussler,D. and Kent,W.J. (2013) The UCSC genome browser and associated tools. *Brief. Bioinform*, **14**, 144–161.
36. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nussbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of chip-Seq (MACS). *Genome Biol.*, **9**, R137.
37. Angerer,P., Haghverdi,L., Büttner,M., Theis,F.J., Marr,C. and Buettner,F. (2016) destiny: diffusion maps for large-scale single-cell data in R. *Bioinformatics*, **32**, 1241–1243.
38. Wu,T., Hu,E., Xu,S., Chen,M., Guo,P., Dai,Z., Feng,T., Zhou,L., Tang,W., Zhan,L. *et al.* (2021) clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *The Innovation*, **2**, 100141.
39. Sun,D., Wang,J., Han,Y., Dong,X., Ge,J., Zheng,R., Shi,X., Wang,B., Li,Z., Ren,P. *et al.* (2020) TISCH: a comprehensive web resource enabling interactive single-cell transcriptome visualization of tumor microenvironment. *Nucleic Acids Res.*, **49**, D1420–D1430.
40. Bravo González-Blas,C., Minnoye,L., Papasokrati,D., Aibar,S., Hulselmans,G., Christiaens,V., Davie,K., Wouters,J. and Aerts,S.

(2019) cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat. Methods*, **16**, 397–400.
41. 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. *Nat. Genet.*, **53**, 403–411.
42. Fang,R., Preissl,S., Li,Y., Hou,X., Lucero,J., Wang,X., Motamedi,A., Shiau,A.K., Zhou,X., Xie,F. *et al.* (2021) Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nat. Commun.*, **12**, 1337.
43. Singh,H., Medina,K.L. and Pongubala,J.M.R. (2005) Contingent gene regulatory networks and b cell fate specification. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 4949–4953.
44. Cismasiu,V.B., Ghanta,S., Duque,J., Albu,D.I., Chen,H.-M., Kasturi,R. and Avram,D. (2006) BCL11B participates in the activation of IL2 gene expression in CD4+ t lymphocytes. *Blood*, **108**, 2695–2702.
45. Buenrostro,J.D., Corces,M.R., Lareau,C.A., Wu,B., Schep,A.N., Aryee,M., Majeti,R., Chang,H.Y. and Greenleaf,W.J. (2018) Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. *Cell*, **173**, 1535–1548.
46. Heinäniemi,M., Nykter,M., Kramer,R., Wienecke-Baldacchino,A., Sinkkonen,L., Zhou,J.X., Kreisberg,R., Kauffman,S.A., Huang,S. and Shmulevich,I. (2013) Gene pair signatures in cell type transcriptomes reveal lineage control. *Nat. Methods*, **10**, 577–583.
47. Huang,S., Guo,Y.-P., May,G. and Enver,T. (2007) Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. *Dev. Biol.*, **305**, 695–713.
48. Domcke,S., Hill,A.J., Daza,R.M., Cao,J., O'Day,D.R., Pliner,H.A., Aldinger,K.A., Pokholok,D., Zhang,F., Milbank,J.H. *et al.* (2020) A human cell atlas of fetal chromatin accessibility. *Science*, **370**, eaba7612.
49. Varma,S., Cao,Y., Tagne,J.-B., Lakshminarayanan,M., Li,J., Friedman,T.B., Morell,R.J., Warburton,D., Kotton,D.N. and Ramirez,M.I. (2012) The transcription factors Grainyhead-like 2 and NK2-Homeobox 1 form a regulatory loop that coordinates lung epithelial cell morphogenesis and differentiation. *J. Biol. Chem.*, **287**, 37282–37295.
50. Rogerson,C., Britton,E., Withey,S., Hanley,N., Ang,Y.S. and Sharrocks,A.D. (2019) Identification of a primitive intestinal transcription factor network shared between esophageal adenocarcinoma and its precancerous precursor state. *Genome Res.*, **29**, 723–736.
51. Strick-Marchand,H. and Weiss,M.C. (2002) Inducible differentiation and morphogenesis of bipotential liver cell lines from wild-type mouse embryos. *Hepatology*, **36**, 794–804.
52. Zhong,C., Cui,K., Wilhelm,C., Hu,G., Mao,K., Belkaid,Y., Zhao,K. and Zhu,J. (2016) Group 3 innate lymphoid cells continuously require the transcription factor GATA-3 after commitment. *Nat. Immunol.*, **17**, 169–178.
53. Long,X., Creemers,E.E., Wang,D.-Z., Olson,E.N. and Miano,J.M. (2007) Myocardin is a bifunctional switch for smooth versus skeletal muscle differentiation. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 16570–16575.
54. Arroyo,N., Villamayor,L., Díaz,I., Carmona,R., Ramos-Rodríguez,M., Muñoz-Chápuli,R., Pasquali,L., Toscano,M.G., Martín,F., Cano,D.A. *et al.* (2021) GATA4 induces liver fibrosis regression by deactivating hepatic stellate cells. *JCI Insight*, **6**, e150059.
55. Paquot,N. (2019) (The metabolism of alcohol). *Rev. Med. Liege*, **74**, 265–267.
56. Haase,V.H. (2013) Regulation of erythropoiesis by hypoxia-inducible factors. *Blood Rev.*, **27**, 41–53.
57. Stallcup,M.R. and Poulard,C. (2020) Gene-Specific actions of transcriptional coregulators facilitate physiological plasticity: evidence for a physiological coregulator code. *Trends Biochem. Sci*, **45**, 497–510.
58. Satpathy,A.T., Granja,J.M., Yost,K.E., Qi,Y., Meschi,F., McDermott,G.P., Olsen,B.N., Mumbach,M.R., Pierce,S.E., Corces,M.R. *et al.* (2019) Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral t cell exhaustion. *Nat. Biotechnol.*, **37**, 925–936.
59. Huang,Y., Lin,L., Shen,Z., Li,Y., Cao,H., Peng,L., Qiu,Y., Cheng,X., Meng,M., Lu,D. *et al.* (2020) CEBPG promotes esophageal squamous cell carcinoma progression by enhancing PI3K-AKT signaling. *Am. J. Cancer Res.*, **10**, 3328–3344.
60. Weulersse,M., Asrir,A., Pichler,A.C., Lemaitre,L., Braun,M., Carrié,N., Joubert,M.-V., Le Moine,M., Do Souto,L., Gaud,G. *et al.* (2020) Eomes-dependent loss of the Co-activating receptor CD226 restrains CD8+ t cell anti-tumor functions and limits the efficacy of cancer immunotherapy. *Immunity*, **53**, 824–839.
61. Zhang,J., Marotel,M., Fauteux-Daniel,S., Mathieu,A.-L., Viel,S., Marçais,A. and Walzer,T. (2018) T-bet and eomes govern differentiation and function of mouse and human NK cells and ILC1. *Eur. J. Immunol.*, **48**, 738–750.
62. Bollig,N., Brüstle,A., Kellner,K., Ackermann,W., Abass,E., Raifer,H., Camara,B., Brendel,C., Giel,G., Bothur,E. *et al.* (2012) Transcription factor IRF4 determines germinal center formation through follicular T-helper cell differentiation. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 8664–8669.
63. Man,K., Gabriel,S.S., Liao,Y., Gloury,R., Preston,S., Henstridge,D.C., Pellegrini,M., Zehn,D., Berberich-Siebelt,F., Febbraio,M.A. *et al.* (2017) Transcription factor IRF4 promotes CD8+ t cell exhaustion and limits the development of memory-like t cells during chronic infection. *Immunity*, **47**, 1129–1141.
64. Tanaka,Y., Nagai,Y., Okumura,M., Greene,M.I. and Kambayashi,T. (2020) PRMT5 is required for t cell survival and proliferation by maintaining cytokine signaling. *Front. Immunol.*, **11**, 621.
65. Saka,D., Gökalp,M., Piyade,B., Cevik,N.C., Arik Sever,E., Unutmaz,D., Ceyhan,G.O., Demir,I.E. and Asimgil,H. (2020) Mechanisms of T-Cell exhaustion in pancreatic cancer. *Cancers (Basel)*, **12**, 2274.
66. Zarour,H.M. (2016) Reversing T-cell dysfunction and exhaustion in cancer. *Clin. Cancer Res.*, **22**, 1856–1864.