Single-cell Gene Regulation Network Interference by Large-scale Data Integration

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

[1] Department of Urology, Tongji Hospital, Frontier Science Center for Stem Cells, School of Life Science and Technology, Tongji University, Shanghai 200092, China
[2] These authors contributed equally to this work

[*]Correspondence
Chenfei Wang, Ph.D., 1239 Siping Road, Shanghai 200092, China; Telephone: 86-21-65981195; Fax: 86-21-65981195; E-mail: 08chenfeiwang@tongji.edu.cn

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 various species, tissues, and samples in the past decades. Here, we developed SCRIP, an integrative method to infer single-cell TR activities and targets based on the integration of scATAC-seq and public bulk ChIP-seq datasets. Our method showed improved performance in evaluating TR binding activities with similar motif information and higher consistency with matched TR expressions. Besides, our method helps in identifying TR-regulated target genes as well as building gene regulation networks (GRN) at single-cell resolution. We demonstrate SCRIP's utility in accurate cell-type clustering, lineage tracing, and inferring cell-type-specific GRNs in multiple biological systems.

Keywords
gene regulation network; scATAC-seq; transcription factor; ChIP-seq; data integration

Introduction

Gene regulation is the fundamental basis of many biological processes, including cell development and differentiation, disease occurrence, and progression. Recently, many single-cell technologies have been developed to investigate gene regulation 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 different development and disease traits.

Although scATAC-seq has been widely used to tackle gene regulation and their association with phenotypes, many 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 activities in each cell. However, motif-based methods cannot discriminate against factors within the same TF family that have similar motifs, 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 each cell. Methods like Signac(8), EpiScanpy(9), MAESTRO(10), and SCALE(11) enhanced the signals by either using latent features or performing imputations, they still cannot identify the TR activities and binding at single-cell resolution. 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.

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. While the ChIP-seq method usually needs millions of cells as starting material, recently developed techniques such as scCUT&Run(14), scCUT&Tag(15), and scCUT&Tag-pro(16) could successfully generate TR binding profiles at the single-cell level. However, most of the data were generated for histone modifications (HMs) rather than TRs, as they have a better abundance on the genome. Although several attempts have been performed on specific TRs (17), they are highly dependent on the quality of the TR antibodies and usually have extremely fuzzy signals at the single-cell level.

In the past decade, numerous ChIP-seq data for TRs as well as HMs have been generated for different cell lines, tissues, and species in the public domain (18). These data provided a great resource for integrating with scATAC-seq data and predicting TR activities and their target genes at the single-cell level. Here, we present a computational method, SCRIP, which can i. generate a curated and comprehensive TR binding reference dataset for data re-use; ii. evaluate the TR activities in single-cell based on the integration of the scATAC-seq dataset and curated reference; iii. determine the target genes of TR at single-cell resolution; iv. construct the GRN in single-cell and identify cell-specific regulation. We applied SCRIP to diverse scATAC-seq datasets on various biological systems and proved the abilities of our method on investigating cell-type-specific regulation, performing lineage tracing, and identifying disease-associated GRNs.

Materials and Methods

Data collection and generation of SCRIP index

We downloaded the ChIP-seq datasets of unified processing from the Cistrome Data Browser(19–21) through the "batch download" function. In total, we obtained bed files of 11,348 human and 9,060 mouse TRs, and 11,079 human and 10,944 mouse histone modification ChIP-seq datasets. Since the auto-parsed metadata included mistakes, we manually 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 greater than 25, the percent of uniquely mapped reads was greater than 50%, PBC (PCR bottleneck coefficient) was greater than 0.8, the number of fold 10 peaks was greater than 100, FRiP (Fraction of Reads in Peaks) was greater than 0.01 and the number of top 5,000 peaks overlapping with union DHS is greater than 70%. We filtered histone modification datasets with the same criteria as above. Besides, we only retained the histone modifications with active functions, including H3K4me1, H3K4me2, H3K4me3, H3K9ac, and H3K27ac, to meet the nature of scATAC-seq. To acquire the high confidence peaks, we only kept the 5-fold enrichment peaks in each peak set. Then, we deleted the datasets with fewer than 1000 peaks. Finally, we obtained 2,322 human and 1,920 mouse TR ChIP-seq datasets, covering 674 and 479 TRs respectively. (Fig. S1)

To cover more transcription factors, we downloaded the motif information, including 7,704 human and 7,000 mouse TF PWMs, from the Cis-BP database(22). We combined PWMs from the same TF and converted the format to the HOMER(23) motif format. Then we scanned the motifs on the hg38 or mm10 genome with the HOMER and obtained the genome intervals where motifs appeared in each TF. We overlapped the scanning intervals with the ENCODE CCRE (24) list and Cistrome union DHS list and removed intervals with the intersection of the blacklist. Then, we sorted intervals by p-value and kept the first 25,000, which were in the magnitude of the average of filtered peaks in ChIP-seq. In total, we obtained 916 human and 866 mouse motif scanning pseudo peaks. (Fig. S1e, f)

Next, we combined the ChIP-seq peak sets and motif scanning 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(25), 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. Overall, the human TR index covered 1,252 TRs and the mouse covered 997 TRs. (Fig. S1e, f)

TR activity score calculation

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 peaks overlap between each cell and the ChIP-seq peaks set or motif scanning 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. To remove the bias from the dataset peaks number and cell peak length, SCRIP normalizes the matrix by:

$$M' = \frac{M}{D \times Q}$$

where D is a j length vector that records the number of peaks of each dataset, and Q is an i length vector that records the number of base-pair in each cell per 100 million. To center the score of each TR, SCRIP minuses the average score of TRs in each dataset:

$$M'' = M' - \mathrm{mean}(M'_j)$$

Since we have collected several datasets of the same TF from different tissue or cell type, SCRIP deduplicates the redundant dataset by getting the maximum score of each TR k:

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

In this step, we meanwhile determine the best match ChIP-seq dataset of each cell. To stabilize the TR enrich score and compress the outlier, SCRIP next applies the logistic sigmoid function to each TR:

$$z = \frac{Y_{i,k} - \mathrm{mean}(Y_k)}{\mathrm{std}(Y_k)}$$

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

Finally, SCRIP applies z-score normalization to each cell to ensure that they can be compared and achieve better clustering performance:

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

TR targets modeling

To acquire a scale and dependable targets of a TR, SCRIP first imputes the potential binding sites of each TR in each cell. In the last part, the best match ChIP-seq dataset of each cell was obtained from the deduplication 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. (Fig. 3a) Since some TR ChIP-seq datasets do not have a sufficient number of peaks and ChIP-seq peak sets were performed on bulk tissue, which may include the peaks from other cell-type, SCRIP furthermore provides a function that uses all the best match TR peak sets of this data found to acquire the 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 regulatory potential score (RP score) model. The RP score of a gene is its likelihood of being regulated by a TR. 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 d_0 is the half-decay distance, with the default value of 10 kb. n denotes the number of binding sites near the TSS of gene g. SCRIP takes genes within 15 d_0 into account. d_i is the distance between the i_{th} peak's center and TSS. With the RP score, we can rank the target genes for each TR, and obtain the most potential target genes to build the gene regulatory network.

Data processing on different scATAC-seq datasets

PBMC

Preprocessing and TRs activities

The PBMC multiome data was downloaded on the 10X genomics website. In the scRNA-seq data, cells with less than 200 genes and genes with less than 3 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. (Fig. S2c) Then, we transferred the cell type labels to scATAC-seq data by the matched cell barcodes. (Fig. S2e) We applied the SCRIP to the filtered scATAC-seq peak count matrix with the default parameters to evaluate the activities of TRs.

Clustering performance comparison
We compared the clustering performance between TR-based tools (SCRIP, chromVAR), peak-based tools (SCALE, 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 chromVARmotifs as the motif reference in the running of chromVAR. The Louvain algorithm in Seurat or scanpy was used to perform clustering with the tools without a built-in clustering function. Normalized mutual information score (NMI) was calculated with the python package sklearn.

TR activity and expression correlation
SCRIP TR activities score and chromVAR z-score were used to calculate the Spearman correlation with gene expression. Only the 505 TRs that appear in SCRIP, chromVAR, and gene expression were used to calculate the correlation. The absolute values of the correlation were used to draw the boxplot. (Fig. S4a)

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 histone modification reference to impute the loci of H3K27ac modifications. The genome track was built by merging the same cell type and normalizing by 0-1 normalization. (Fig. 3b) 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 original scATAC-seq dataset. We merged the T cells and monocytes with the maximum RP score. 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 1,000 RP score targets of each dataset. (Fig. 3c)

HSC
Preprocessing
The HSC scATAC-seq peak count matrix was obtained from the GEO. LiftOver was used to convert genome build from hg19 to hg38. We annotated the cell type with the labels from the original study. To evaluate the activities 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 activities score and calculated the NMI with the cell type annotations. (Table S1) The TRs' activities score matrix was used to do the following analysis.

Trajectory analysis
We applied the R packages destiny(26) 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 activities score at TR level instead of dataset level. The top 600 most variable TRs were used to reconstruct the differential path. The k was set to 4 for the 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 projected the HOXA9, GATA1, CEBPB, TCF4, and other TRs activity scores to the diffusion map to show the TRs activities on each branch. (Fig. 4d, Fig. S7b-f) To visualize the dynamic changes of the TRs' activities in the differential lineages, we showed the TR's activity score with the cell's DPT. (Fig. S7g-v)

Triangle Plot

The activity score of each TR of each cell type was calculated by averaging the TR's activities 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' activities on different lineages were represented by the TR activities score of the terminally differentiated cell type. For example, we used the CLP to represent the lymphoid branch, the monocytes to represent the myeloid branch, and the MEP to represent the erythroid branch. We displayed the positions of TRs on each branch with the ggtern package. (Fig. 4c)

Human Fetal Organ

Preprocessing and Clustering

The different human organs scATAC-seq datasets were obtained from the GEO, which provided the filtered peak count matrix and cell labels with the Seurat object format. 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. (Fig. 5b-d, Fig. S8a-c, Table S2) We performed the unsupervised clustering with the TR activities score and calculated the NMI with the cell type annotations. (Fig. 5a, 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 targets, we selected the top 1000 target genes according to the RP score to do the GO enrichment analysis with the R packages ClusterProfiler. (Fig. 5e-h, Table S3)

BCC tumor microenvironment

Preprocessing

The scATAC-seq dataset of tumor cells and T cells in BCC was obtained from GEO. 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 activities in each cell type were used to plot the heatmap. 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(27). (Fig. S10)

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. (Fig. S9a, b) The JUNB target analysis was done on both naïve CD8 T cells and terminal TEx cells. (Fig. 6f, g, Table S3) Genes with low RP scores, which are considered to have few peaks of this TR, were removed. We normalized RP with natural logarithm and scaled for each cell. The FindMarkers function in the R package Seurat was applied to identify the differential target genes of naïve CD8 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. (Fig. S9d) 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 (Fig. 1). We first built a comprehensive reference dataset to help evaluate the TR enrichment in the cells from scATAC-seq data. 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) (19, 20). We have carefully curated metadata such as factor information, tissue types, and cell types (Fig. 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 674 human TRs and 479 mouse TRs (Fig. 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 (Fig. S1e-f, See Methods).

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 might 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 (see Methods). After identifying the best-matched ChIP-seq dataset for a TR, we could combine the ChIP-seq peaks with scATAC-seq peaks, and apply the regulatory potential (RP) model (28) to quantitatively evaluate the TR enrichment on its target genes for each cell. 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 activities, candidate TR targets, and TR GRNs both at single-cell resolution.

TR Activity Performance Evaluation using PBMC Multiome dataset

To systematically evaluate the performance of SCRIP, we applied it on 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 (Fig. S2a-e). The 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 (Fig. 2a, Fig. S3). Also, the cells can be well clustered to their cell type lineages using TR activities alone. We also compared the consistency of the clustering results with scRNA-seq transferred labels and benchmarked them using normalized mutation information (NMI). 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(29), CisTopic(30), ArchR(31), and SnapATAC(32) (Fig. 2b, Table S1). This result suggests that SCRIP could accurately predict TR activities at the single-cell level, which should show superior performance in determining cell-type lineages.

Next, we compared TR activity with its gene expression. The key TRs in each cell type should be highly or moderately expressed in that cell type. Therefore, we performed correlation analyses between TR activities and gene expressions. Compared to chromVAR, our method

has significantly higher correlations with gene expression, both by averaging all cell types (Fig. S4a) or by individual cell types (Fig. S4b-i). Also, SCRIP correctly estimated the activities of factors with similar motifs (Fig. 2c, 132 vs 67). For example, previous studies have suggested that BCL11A is required for the generation of B progenitor cells (33), while BCL11B activates the transcription of interleukin-2 during T cell activation (34). These two factors share similar motifs but are expressed in distinct lineages (Fig. 2d, Figure S2c). Consistently, SCRIP predicts BCL11A to be enriched in B-cells and myeloid lineages, while BCL11B was enriched in T and NK-cells (Fig. 2e-f). On the contrary, the chromVAR score shows no significant difference between BCL11A and BCL11B, which is biased by the similar motif sequences (Fig. 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. POLR2A is the TR with the most abundant ChIP-seq data in various tissue types (Fig. S1c-d). We tested the ability of SCRIP to find the correct POLR2A ChIP-seq dataset for different single cells. As we expected, SCRIP has high accuracy in finding corresponding TR ChIP-seq datasets for both T cells, B cells, and monocytes (Fig. S5a-c). For each TR dataset, they can also match the correct single cells with correct cell-type lineages (Fig. S5d-f). Taken together, these results suggest that SCRIP could accurately predict TR activities at the single-cell level, identify tissue-specific regulations, and find the correct TR dataset for each cell.

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 regulatory potential (RP) model to infer the putative targets of TR on each cell (Fig. 3a, see Methods). While single-cell TR CUT&Tag is of low quality, we benchmarked our method using several published histone modification scCUT&Tag-pro datasets.

H3K27ac modification is an active enhancer marker that has been profiled using scCUT&Tag-pro in PBMC (16). We built a reference dataset with active histone modifications including H3K27ac (Fig. S6) and imputed the H3K27ac signal using the PBMC scATAC-seq dataset. 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 (Fig. 3b). Besides, when we calculated correlations between the real scCUT&Tag-pro RP, SCRIP imputed TP, 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 in identifying H3K27ac regulated genes (Fig. 3c). More specifically, when comparing the top 1,000 H3K27ac regulated genes in T-cells and monocytes, SCRIP imputed RP could identify more target genes in scCUT&Tag-pro data than using bulk H3K27ac data directly (Fig. 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 activities 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 (35). The HSC differentiation is a well-characterized system, with HSCs differentiating into three different major lineages (Fig. 4a). After identifying TR activities in different HSC subpopulations, we performed pseudo-time analysis and reconstructed the differentiation trajectory of HSC using TR activities (Fig. 4b and Fig. S7a). The diffusion map suggests that HSC was differentiated into three major directions, CLP, Monocytes, and MEP, with a little spike towards pDC (Fig. 4b). These directions are perfectly aligned with the known differentiation path of HSCs.

Next, we sought to identify the driven TRs for the three major differentiation lineages. We use the average TR activities 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 (Fig. 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 (36, 37), which is highly consistent in the SCRIP results (Fig. 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 (Fig. 4d, Fig. S7b-f). Besides, the dynamic changes of the TRs' activities in the differential lineages indicate their potential role in lineage differentiation (Fig. S7g-v). These results prove that SCRIP enables the trajectory analyses of scATAC-seq with known driver TR activities.

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 (38). The TR activity score showed a better performance in clustering compared to the motif-based method chromVAR in almost all these tissues (Fig. 5a, 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 (Fig. 5b-d, Fig. S8a-c, 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 (39) (Fig. 5b, Bronchiolar and Alveolar Epithelial Cells, BAEC). GATA6 regulates the development of primitive intestinal cells (40) (Fig. 5c, Intestinal epithelial cells, IEC). HNF1A, HNF4A, and GATA4 are well-known hepatocyte TFs in liver tissues (41) (Fig. 5d, Hepatoblasts, HB). These results proved that SCRIP can not only cluster the same cell type with TRs activities 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 to establish the cell-type-specific GRN of SCRIP, we inferred the potential target genes of TRs and built the cell-type-specific GRNs (Fig. 5e-h, Table S3). In the lung, we identified the target genes of GATA3, which is mainly enriched in the lymphoid cells (LC) (Fig. 5b). The target genes of GATA3 mainly contribute to the immune functions through responding to interleukin-7 (IL-7) and negatively regulating differentiation of myeloid cells, which is in line with previous studies (42) (Fig. 5e). In the intestine, MYOD1 controls the differentiation of smooth muscle cells (SMC) by regulating its downstream genes and function (43) (Fig. 5c, f). Finally, we built a co-regulatory

GRN of GATA4 and its downstream targets EPAS1(44) in liver hepatoblasts (HB) (Fig. 5g). Although their downstream target genes show a great difference (Fig. S8d-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 (45, 46) (Fig. 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 (47). We applied SCRIP to a Basal Cell Carcinoma (BCC) tumor microenvironment (TME) dataset (48) to investigate how TRs and their target genes were changed in different cell states under disease status. First, we confirmed the TR activities are accurately predicted in the corresponding cell types (Fig. 6a). For instance, CEBPG, a TF that promotes cancer development by enhancing the PI3K-Akt signaling pathway (49), was found to be robustly more active in tumor cells than in other cells. In addition, the activities of TFs such as EOMES and TBX21 were higher in immune cells than in tumor cells (Fig. 6a), which is consistent with the role these TFs play in driving lymphocyte differentiation (50, 51).

T cells are the major cytotoxic cells responsible for anti-tumor 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 activities 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 (Fig. 6b). The IRF4 activities were significantly increased in Tfh, the function of its target genes also enriched in lymphocyte activation and differentiation (Fig. S9a). These analyses are consistent with the IRF4 function in Tfh cell expansion (52). Another path is from naïve CD8 T cells to terminal T exhaustion (TEx) cells. BATF, a key regulator of T cell exhaustion(53), has higher activities in terminal TEx (Fig. 6c). Consistently, the BATF target genes tend to have an immunosuppressive effect on terminal TEx cells (Fig. S9b). These analyses suggest that the TR activities and targets inferred by SCRIP could be used to track cell state changes under the disease condition.

Interestingly, we found that the activities of JUNB are both higher in naïve CD8 T and terminal TEx (Fig. S9c). 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 (Fig. 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 (Fig. S9d, Fig. 6f, g, Fig. S10). We found that PRMT5, which is critical for the transition of naïve T cells to the effector or memory phenotype (54), is presented in the JUNB GRNs only in Naïve CD8 T cells (Fig. 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 (55, 56), has a high RP score in JUNB GRNs in terminal TEx (Fig. 6e, g). The function enrichment results suggest that JUNB mainly tends to function as a positive regulator of T cell activation and to migrate T cells to lymphoid organs, while negatively modulating the immune system process in terminal TEx cells. These results 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 interference 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 data. Based on the reference, we developed a method that can evaluate TR activities 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 three different biological systems, including 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 activities 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, only 3k ChIP-seq datasets are 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 DC includes 13,366 human and 99,53 mouse epigenome datasets, while the number almost doubled to 25,000 human and 22,000 mouse epigenome datasets after only 2 years (19, 20). Large consortiums like ENCODE, epigenome-roadmap will also generate a great number of TR datasets with high quality. Second, machine learning algorithms, such as generative adversarial networks (GAN) could be used to generate more TR ChIP-seq datasets *in silico*. Thirdly, with the development of scCUT&RUN and scCUT&Tag, we could also integrate the single-cell TR dataset into our reference for annotating TRs for scATAC-seq in other cell types. The expanded reference will improve the performance of our method for predicting TR activities. Finally, although we demonstrated that SCRIP is powerful in predicting TR activities and GRNs for scATAC-seq, we could potentially extend its applications to scRNA-seq. The Cistrome DC 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 activities 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.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

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).

Code availability

SCRIP is an open-source python package with source code freely available at: https://github.com/wanglabtongji/SCRIP.

Competing interests

The authors declare no competing financial interests.

Authors' 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

This work was supported by the National Natural Science Foundation of China [32170660], Shanghai Rising Star Program [21QA1408200], Natural Science Foundation of Shanghai [21ZR1467600]. The authors thank the Bioinformatics Supercomputer Center of Tongji University for offering computing resources.

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. *Science Advances*, **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.

14. Skene,P.J. and Henikoff,S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife*, **6**, e21856.

15. 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.

16. Zhang,B., Srivastava,A., Mimitou,E., Stuart,T., Raimondi,I., Hao,Y., Smibert,P. and Satija,R. (2021) Characterizing cellular heterogeneity in chromatin state with scCUT&Tag-pro. *bioRxiv*, 10.1101/2021.09.13.460120.

17. 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.

18. Ernst,J. and Kellis,M. (2015) Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat Biotechnol*, **33**, 364–376.

19. 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.

20. 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.

21. Cistrome Data Browser and Toolkit: analyzing human and mouse genomic data using compendia of ChIP-seq and chromatin accessibility data (2020) *Quant. Biol.*, **8**, 267–276.

22. 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.

23. 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.

24. 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.

25. 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.

26. 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.

27. 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.

28. Elnitski,L., Hardison,R.C., Li,J., Yang,S., Kolbe,D., Eswara,P., O'Connor,M.J., Schwartz,S., Miller,W. and Chiaromonte,F. (2003) Distinguishing Regulatory DNA From Neutral Sites. *Genome Res*, **13**, 64–72.

29. 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.

30. 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.

31. 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.

32. 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.

33. 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.

34. 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.

35. 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.e16.

36. 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.

37. Huang,S., Guo,Y.-P., May,G. and Enver,T. (2007) Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. *Developmental Biology*, **305**, 695–713.

38. 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.

39. 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.

40. 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.

41. 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.

42. 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.

43. 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.

44. 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.* GATA4 induces liver fibrosis regression by deactivating hepatic stellate cells. *JCI Insight*, **6**, e150059.

45. Paquot,N. (2019) [The metabolism of alcohol]. *Rev Med Liege*, **74**, 265–267.

46. Haase,V.H. (2013) Regulation of erythropoiesis by hypoxia-inducible factors. *Blood Rev*, **27**, 41–53.

47. 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.

48. 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.

49. 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.

50. 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.e10.

51. 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. *European Journal of Immunology*, **48**, 738–750.

52. 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. *PNAS*, **109**, 8664–8669.

53. 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.e5.

54. 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.

55. 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.

56. Zarour,H.M. (2016) Reversing T-cell Dysfunction and Exhaustion in Cancer. *Clin Cancer Res*, **22**, 1856–1864.

Figure Legend

Figure 1 Workflow of SCRIP

Schematic of SCRIP workflow. SCRIP takes the feature count matrix of scATAC-seq as input. The ChIP-seq and motif reference datasets were built with careful curation. The normalized enrichment score was estimated between every single cell and every reference dataset. The best-matched reference dataset was combined with single-cell chromatin accessibility to determine the target gene with the RP model.

Figure 2 SCRIP achieves better performance on the PBMC dataset

a. Heatmap of TR activities with different cell types. The X-axis denotes the unsupervised clustering results with the TR activities score.
b. The clustering performance among SCRIP, chromVAR, SCALE, CisTopic, ArchR, and SnapATAC. Y-axis: NMI score.
c. TR activity and its gene expression spearman correlation. X-axis: correlation of chromVAR TF enrichment z-score and TF expression; Y-axis: correlation of SCRIP TR activity and TR expression. The number of each corner denotes the number of TR in each quadrant.
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. RP scores of Spearman correlation with scCUT&Tag-pro among SCRIP imputed, original scATAC-seq and bulk dataset 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.

Figure 4 SCRIP reconstructs the path of differentiation of HSCs based on the TR activities

a. Schematic of HSC differentiation.
b. Diffusion map of HSC with the cell-type annotations.
c. Triangle plot of TRs that regulate HSC differentiation towards three main lineages.
d. Projecting HOXA9, GATA1, CEBPB, and TCF4 activities onto the diffusion map.

Figure 5 SCRIP finds the key regulators and builds GRNs in human fetal organs

a. Clustering performance of 12 different organs.
b-d. Clustering of the human lung (b), intestine (c), and liver (d) scATAC-seq data based on TR activities. Marked TRs of specific cell types are identified by SCRIP with literature supporting. 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 activities change during pseudo-time from naive CD4 T cells to CD4 Tfh cells.
c. TR activities 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.

Supplementary figure 1 Overview of TR reference datasets of human and mouse
a-b. Tissue distribution of human and mouse TR reference datasets. Cell lines and minor tissue types are not fully marked.
c-d. Histogram of TR of human and mouse reference datasets. Only the top 60 TRs with datasets are shown.
e-f. The number of covered TRs of human and mouse reference datasets in SCRIP and chromVAR.

Supplementary figure 2 PBMC scRNA-seq dataset and cell-type annotation
a. UMAP which is based on PBMC scRNA-seq gene expression shows Louvain clustering.
b. UMAP which is based on PBMC scATAC-seq peak count shows Louvain clustering.
c. Gene markers were used to annotate the PBMC scRNA-seq dataset.
d. UMAP which is based on PBMC scRNA-seq gene expression shows the cell-type annotation.
f. UMAP which is based on PBMC scATAC-seq peak count shows the cell-type annotations that were transferred from scRNA-seq by barcodes.

Supplementary figure 3 Well-known TR activities in the diverse cell types of PBMC
a-c. UMAP shows the YY1, RELA, and FOXP3 activities, which are important TRs in T cells.
d-f. UMAP shows the SPI1, CEBPA, and CEBPB activities, which are important TRs in monocytes.
g-i. UMAP shows the EBF1, PAX5, and FOXO1 activities, which are important TRs in B cells.
j-l. UMAP shows the GATA2, GATA3, and GATA4 activities, which share similar motifs (not shown).

Supplementary figure 4 Comparison of the TR activities with gene expression between SCRIP and chromVAR
a. Absolute values of TR activities spearman correlation with gene expression between two methods. The P-value was calculated using the Wilcoxon test.
b-l. Scatter plot of average TR activities of two methods and average gene expression in different cell types of PBMC. TRs in the top 1% either SCRIP or chromVAR were marked. The color of each dot denotes the average gene expression in the cell type.

Supplementary figure 5 SCRIP found the matched cell type between single-cell data and bulk data
a-c. Histogram showing the majority of found reference datasets in T cells, B cells, and monocytes.
d-f. Histogram to show what cells matched the top 3 found on reference datasets.

Supplementary figure 6 Overview of histone modifications reference datasets of human and mouse
a-b. Tissue distribution of human and mouse histone modifications reference datasets. Cell lines and minor tissue types are not fully marked.
c-d. Histogram of histone modifications of human and mouse reference datasets.

Supplementary figure 7 Pseudo time of HSC and changes of important TR activities during differentiation
a. Diffusion map of HSC pseudo-time. Dark purple denotes the early stages, and yellow denotes the late stages.
b-f. Projecting FOXO1, GATA2, IRF4, IKZF1, and TAL2 activities onto the diffusion map.
g-v. TR activity changes during cell differentiation. **g.** HSC to pDC; **h-l.** HSC to CLP; **m-p.** HSC to monocytes; **q-v**. HSC to MEP.

Supplementary figure 8 TR activity clustering of three human organ datasets and differences in co-regulation
a-c. Heatmap of TR activities with different cell types in the lung, intestine, and liver.
d. Overlap of top 1,000 target genes of GATA4 and EPAS1 in HB of human liver.
c. Overlap of enriched GO terms of GATA4 and EPAS1 target genes in HB of the human liver.

Supplementary figure 9 Key TR GRNs and differential target gene determination
a. (left) Inferred IRF4 GRN in CD4 Tfh cells. Pink circles denote the target genes that are supported by previous studies. (right) GO results showed the terms enriched of IRF4 target genes in CD4 Tfh cells.
b. (left) Inferred BATF GRN in terminal TEx cells. Pink circles denote the target genes that are supported by previous studies. (right) GO results showed the terms enriched of BATF target genes in terminal TEx cells.
c. Line plot to show the changes of 6 TRs activities during the differentiation from naive CD8 T cells to terminal TEx cells. Y-axis: average TR activities of each cell type.
d. Volcano plot to show the different target genes of JUNB between naive CD8 T cells and terminal TEx cells.

Supplementary figure 10 Gene expression of example TRs in BCC dataset
a. UMAP of cell identities of BCC datasets and the gene expression of JUNB, BATF, and IRF4 obtained from the TISCH database.
b. Violin plot to show the gene expression of JUNB, BATF, and IRF4 in different cell types of BCC datasets.

Supplementary Tables

Supplementary Table 1

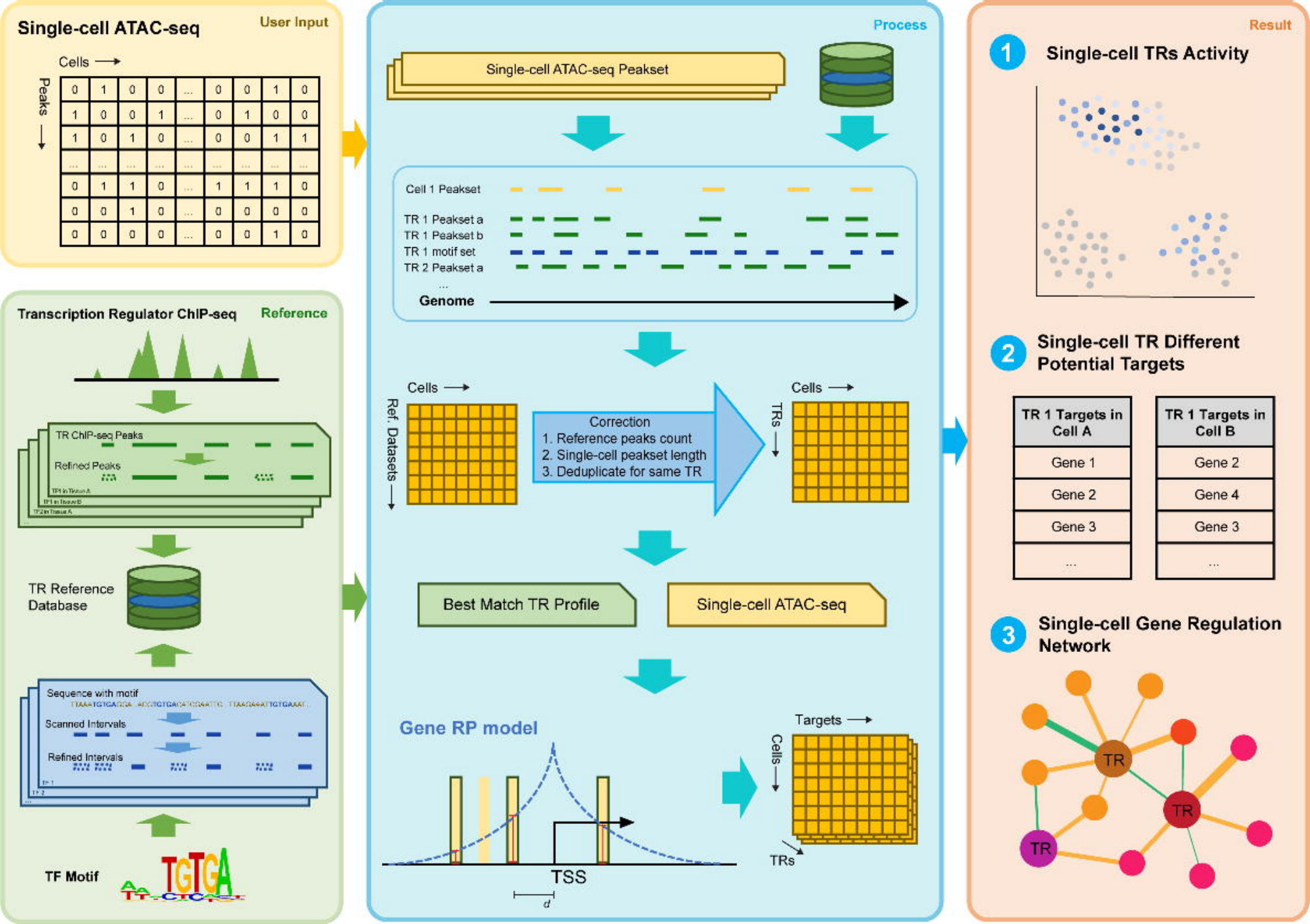
Clustering performance of different methods on different tissues. The numbers suggest the NMI score. In the table, "x" means that fails to run because of the size of data, blank means that we have not run this tool on the dataset.

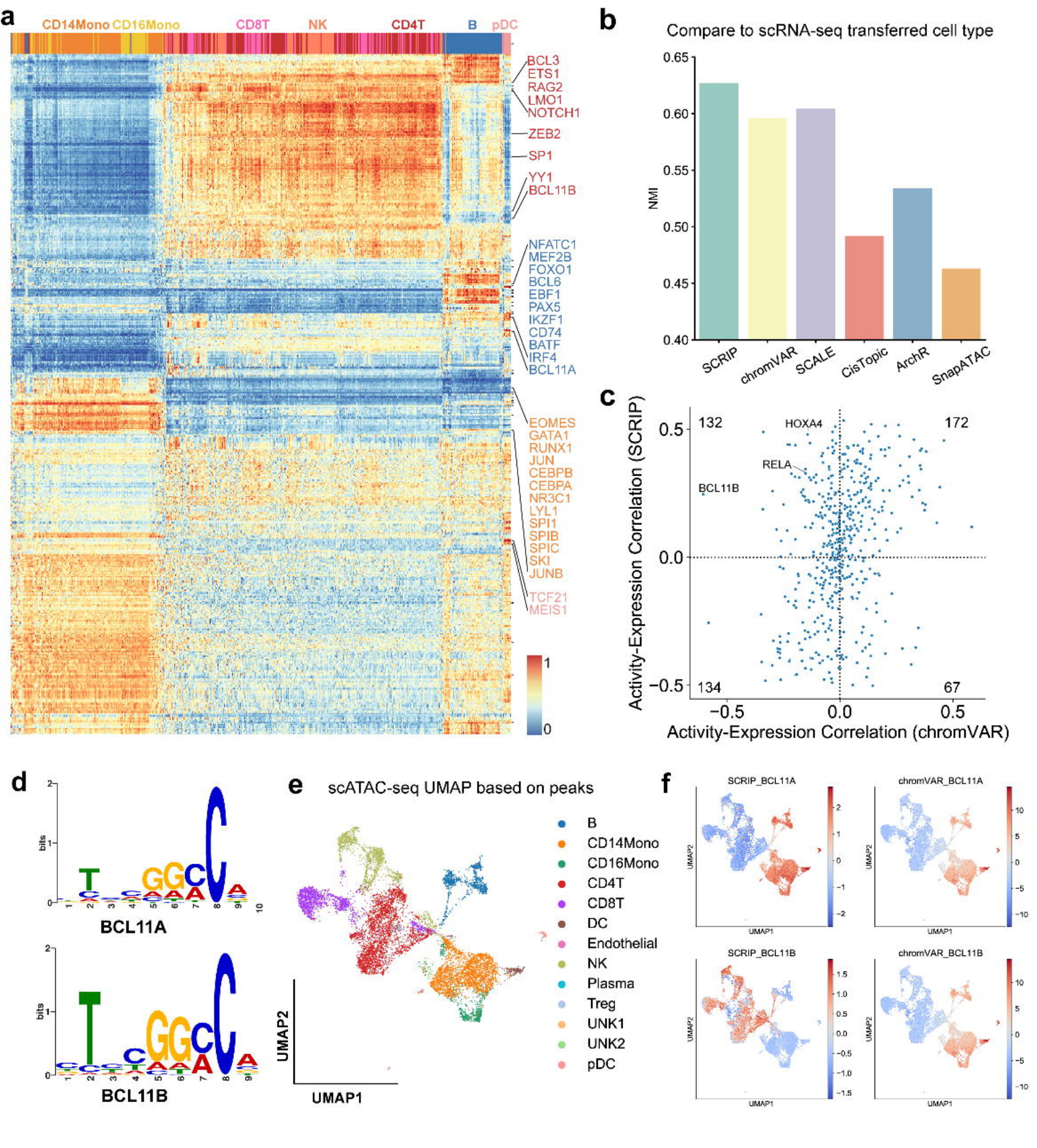
Supplementary Table 2

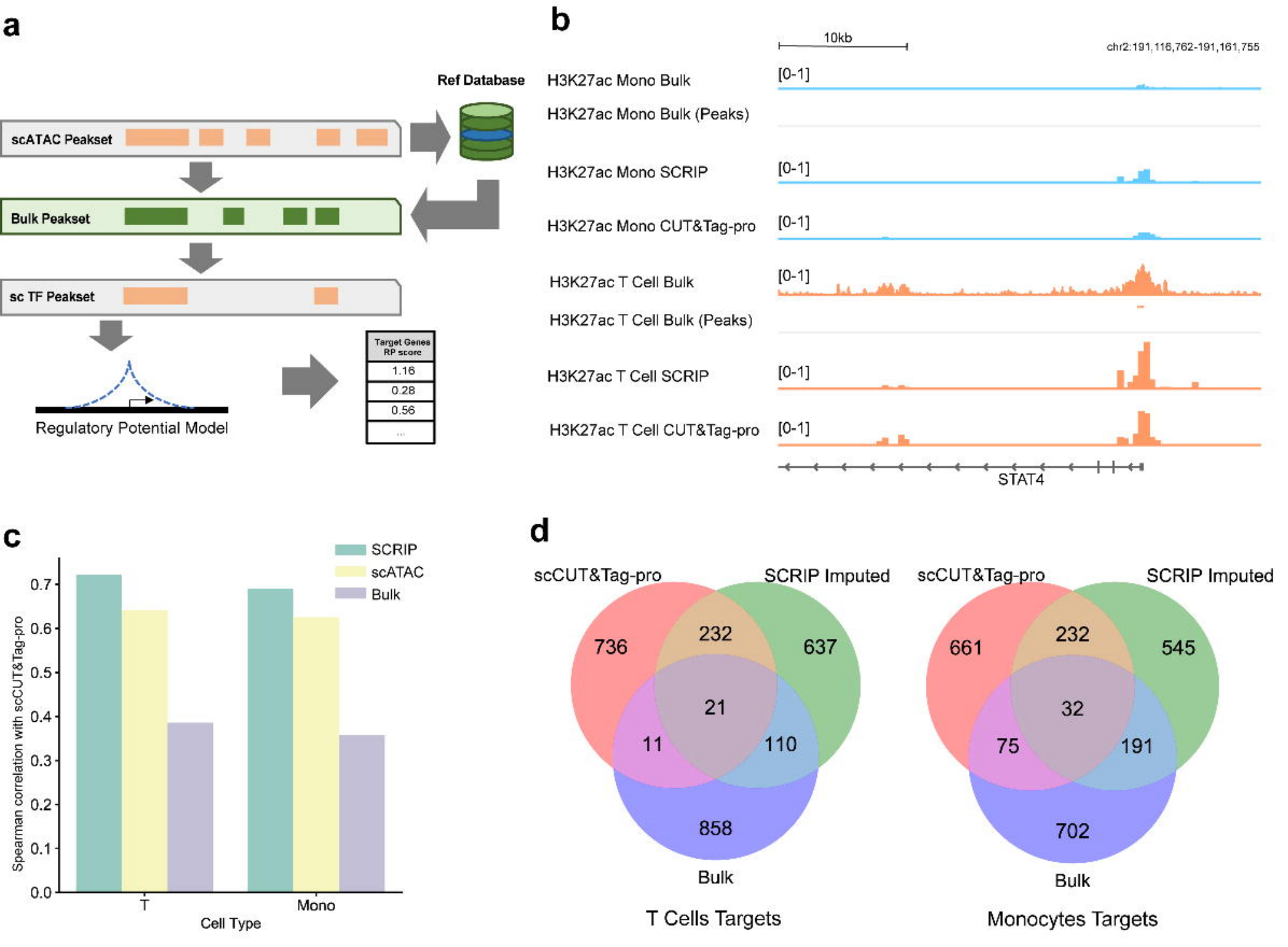
Specific-cell-type TRs with literature supporting. The numbers in the brackets denote the PubMed ID of the paper that supports the TR playing roles in this cell type.

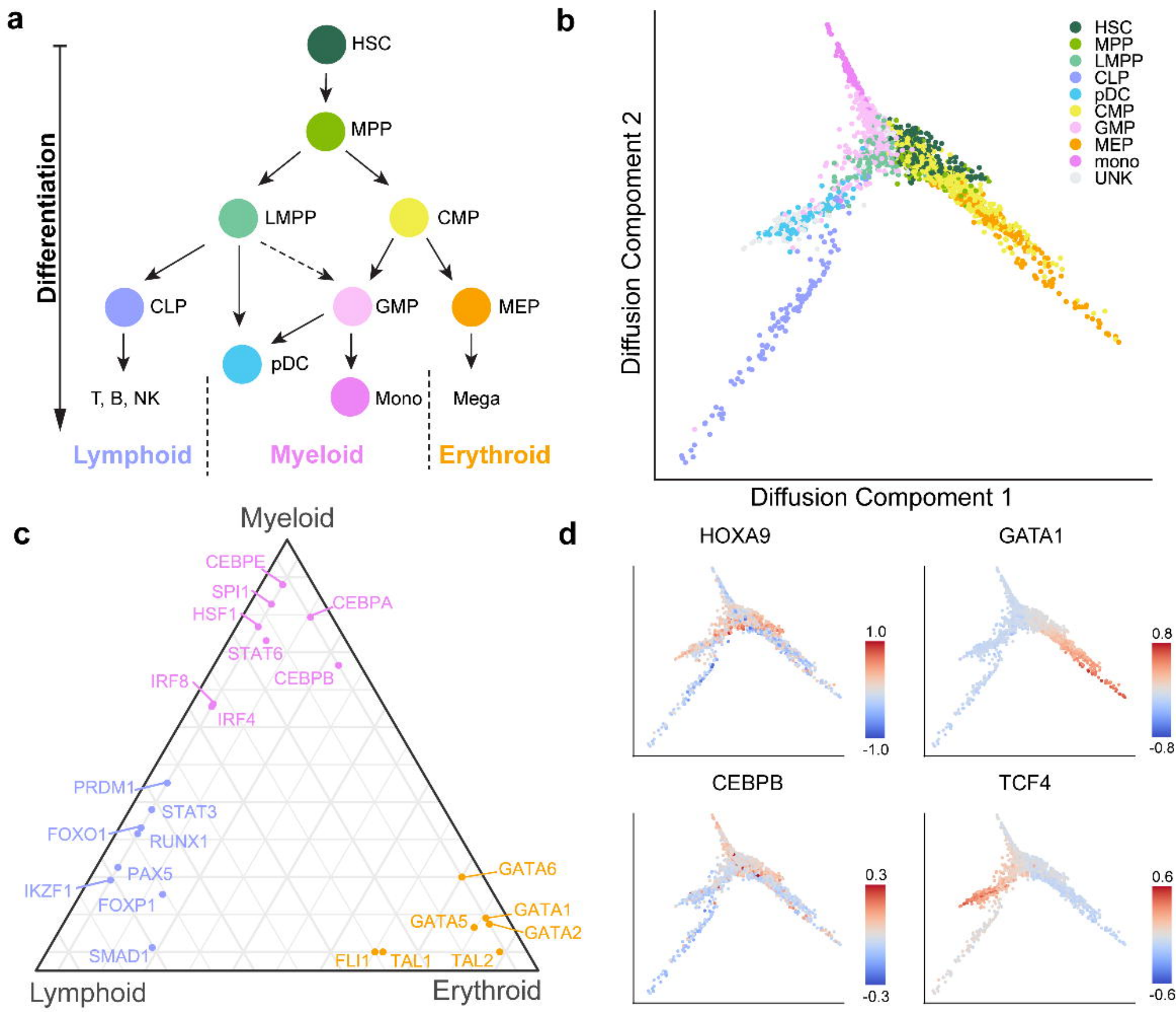
Supplementary Table 3

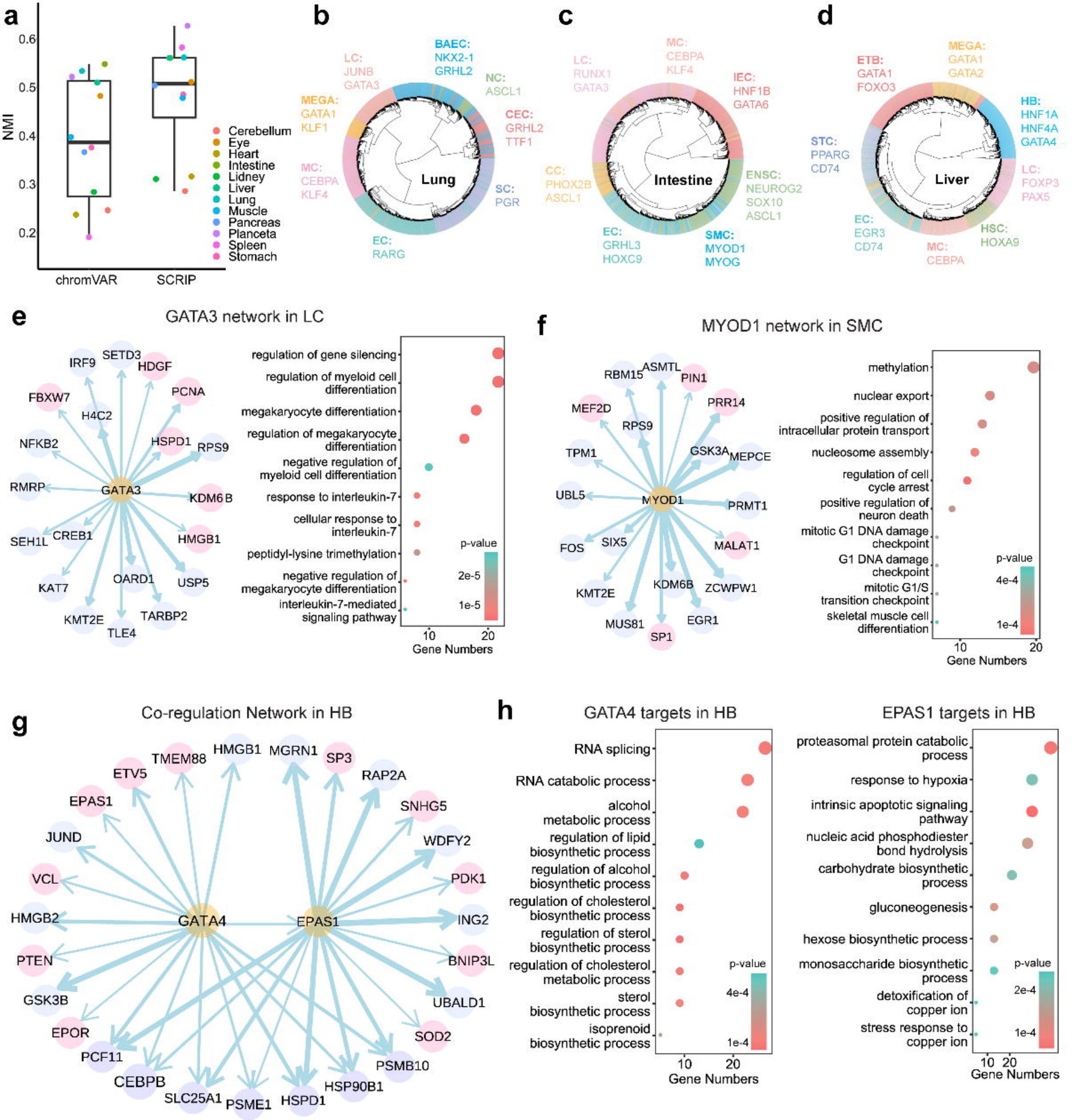
TF-target pairs with literature supporting. The numbers in the brackets denote the PubMed ID of the paper that supports the gene that is the target of this TR.

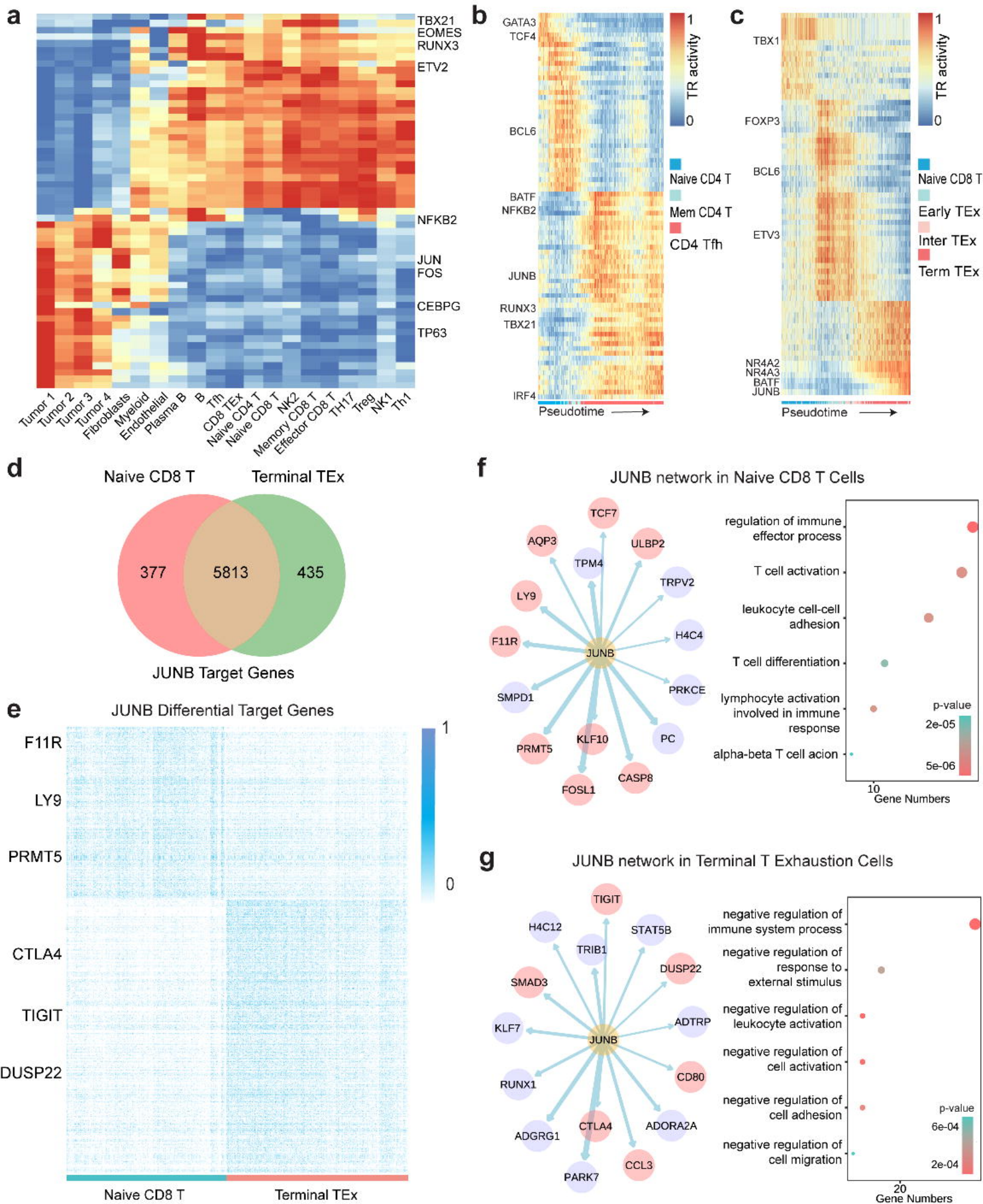

Single-cell ATAC-seq
User Input
Cells
Peaks
Transcription Regulator ChIP-seq
Reference
TR ChIP-seq Peaks
Refined Peaks
TR Reference Database
Sequence with motif
Scanned Intervals
Refined Intervals
TF Motif
Process
Single-cell ATAC-seq Peakset
Cell 1 Peakset
TR 1 Peakset a
TR 1 Peakset b
TR 1 motif set
TR 2 Peakset a
Genome
Ref. Datasets
Correction
1. Reference peaks count
2. Single-cell peakset length
3. Deduplicate for same TR
TRs
Best Match TR Profile
Single-cell ATAC-seq
Gene RP model
TSS
d
Targets
Result
1 Single-cell TRs Activity
2 Single-cell TR Different Potential Targets
TR 1 Targets in Cell A
Gene 1
Gene 2
Gene 3
TR 1 Targets in Cell B
Gene 2
Gene 4
Gene 3
3 Single-cell Gene Regulation Network
TR


a
CD14Mono
CD16Mono
CD8T
NK
CD4T
B
pDC
BCL3
ETS1
RAG2
LMO1
NOTCH1
ZEB2
SP1
YY1
BCL11B
NFATC1
MEF2B
FOXO1
BCL6
EBF1
PAX5
IKZF1
CD74
BATF
IRF4
BCL11A
EOMES
GATA1
RUNX1
JUN
CEBPB
CEBPA
NR3C1
LYL1
SPI1
SPIB
SPIC
SKI
JUNB
TCF21
MEIS1
1
0
b
Compare to scRNA-seq transferred cell type
NMI
0.65
0.60
0.55
0.50
0.45
0.40
SCRIP
chromVAR
SCALE
CisTopic
ArchR
SnapATAC
c
Activity-Expression Correlation (SCRIP)
Activity-Expression Correlation (chromVAR)
132
172
134
67
HOXA4
RELA
BCL11B
0.5
0.0
−0.5
d
bits
BCL11A
BCL11B
e
scATAC-seq UMAP based on peaks
UMAP1
UMAP2
B
CD14Mono
CD16Mono
CD4T
CD8T
DC
Endothelial
NK
Plasma
Treg
UNK1
UNK2
pDC
f
SCRIP_BCL11A
chromVAR_BCL11A
SCRIP_BCL11B
chromVAR_BCL11B
UMAP1
UMAP2


a
Ref Database
scATAC Peakset
Bulk Peakset
sc TF Peakset
Regulatory Potential Model
Target Genes
RP score
1.16
0.28
0.56
...
b
10kb
chr2:191,116,762-191,161,755
H3K27ac Mono Bulk
H3K27ac Mono Bulk (Peaks)
H3K27ac Mono SCRIP
H3K27ac Mono CUT&Tag-pro
H3K27ac T Cell Bulk
H3K27ac T Cell Bulk (Peaks)
H3K27ac T Cell SCRIP
H3K27ac T Cell CUT&Tag-pro
[0-1]
STAT4
c
SCRIP
scATAC
Bulk
Spearman correlation with scCUT&Tag-pro
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
T
Mono
Cell Type
d
scCUT&Tag-pro
SCRIP Imputed
Bulk
736
232
637
21
11
110
858
T Cells Targets
scCUT&Tag-pro
SCRIP Imputed
Bulk
661
232
545
32
75
191
702
Monocytes Targets


a
Differentiation
HSC
MPP
LMPP
CMP
CLP
GMP
MEP
pDC
Mono
T, B, NK
Mega
Lymphoid
Myeloid
Erythroid
b
HSC
MPP
LMPP
CLP
pDC
CMP
GMP
MEP
mono
UNK
Diffusion Compoment 2
Diffusion Compoment 1
c
Myeloid
CEBPE
SPI1
HSF1
CEBPA
STAT6
CEBPB
IRF8
IRF4
PRDM1
STAT3
FOXO1
RUNX1
PAX5
IKZF1
FOXP1
SMAD1
GATA6
GATA5
GATA1
GATA2
FLI1
TAL1
TAL2
Lymphoid
Erythroid
d
HOXA9
1.0
-1.0
GATA1
0.8
-0.8
CEBPB
0.3
-0.3
TCF4
0.6
-0.6


a
NMI
0.6
0.5
0.4
0.3
0.2
chromVAR
SCRIP
Cerebellum
Eye
Heart
Intestine
Lidney
Liver
Lung
Muscle
Pancreas
Planceta
Spleen
Stomach
b
Lung
BAEC: NKX2-1 GRHL2
NC: ASCL1
CEC: GRHL2 TTF1
SC: PGR
EC: RARG
MC: CEBPA KLF4
MEGA: GATA1 KLF1
LC: JUNB GATA3
c
Intestine
MC: CEBPA KLF4
IEC: HNF1B GATA6
ENSC: NEUROG2 SOX10 ASCL1
SMC: MYOD1 MYOG
EC: GRHL3 HOXC9
CC: PHOX2B ASCL1
LC: RUNX1 GATA3
d
Liver
MEGA: GATA1 GATA2
HB: HNF1A HNF4A GATA4
LC: FOXP3 PAX5
HSC: HOXA9
MC: CEBPA
EC: EGR3 CD74
STC: PPARG CD74
ETB: GATA1 FOXO3
e
GATA3 network in LC
GATA3
IRF9
SETD3
HDGF
PCNA
FBXW7
H4C2
NFKB2
HSPD1
RPS9
RMRP
KDM6B
SEH1L
CREB1
HMGB1
KAT7
OARD1
USP5
KMT2E
TARBP2
TLE4
regulation of gene silencing
regulation of myeloid cell differentiation
megakaryocyte differentiation
regulation of megakaryocyte differentiation
negative regulation of myeloid cell differentiation
response to interleukin-7
cellular response to interleukin-7
peptidyl-lysine trimethylation
negative regulation of megakaryocyte differentiation
interleukin-7-mediated signaling pathway
p-value
2e-5
1e-5
10
20
Gene Numbers
f
MYOD1 network in SMC
MYOD1
RBM15
ASMTL
PIN1
PRR14
MEF2D
RPS9
TPM1
GSK3A
MEPCE
UBL5
PRMT1
FOS
SIX5
MALAT1
KMT2E
KDM6B
ZCWPW1
MUS81
EGR1
SP1
methylation
nuclear export
positive regulation of intracellular protein transport
nucleosome assembly
regulation of cell cycle arrest
positive regulation of neuron death
mitotic G1 DNA damage checkpoint
G1 DNA damage checkpoint
mitotic G1/S transition checkpoint
skeletal muscle cell differentiation
p-value
4e-4
1e-4
10
20
Gene Numbers
g
Co-regulation Network in HB
GATA4
EPAS1
HMGB1
MGRN1
TMEM88
SP3
ETV5
RAP2A
EPAS1
SNHG5
JUND
WDFY2
VCL
PDK1
HMGB2
ING2
PTEN
BNIP3L
GSK3B
UBALD1
EPOR
SOD2
PCF11
PSMB10
CEBPB
HSP90B1
SLC25A1
PSME1
HSPD1
h
GATA4 targets in HB
RNA splicing
RNA catabolic process
alcohol metabolic process
regulation of lipid biosynthetic process
regulation of alcohol biosynthetic process
regulation of cholesterol biosynthetic process
regulation of sterol biosynthetic process
regulation of cholesterol metabolic process
sterol biosynthetic process
isoprenoid biosynthetic process
p-value
4e-4
1e-4
10
20
Gene Numbers
EPAS1 targets in HB
proteasomal protein catabolic process
response to hypoxia
intrinsic apoptotic signaling pathway
nucleic acid phosphodiester bond hydrolysis
carbohydrate biosynthetic process
gluconeogenesis
hexose biosynthetic process
monosaccharide biosynthetic process
detoxification of copper ion
stress response to copper ion
p-value
2e-4
1e-4
10
20
Gene Numbers


a
TBX21
EOMES
RUNX3
ETV2
NFKB2
JUN
FOS
CEBPG
TP63
Tumor 1
Tumor 2
Tumor 3
Tumor 4
Fibroblasts
Myeloid
Endothelial
Plasma B
B
Tfh
CD8 TEx
Naive CD4 T
Naive CD8 T
NK2
Memory CD8 T
Effector CD8 T
TH17
Treg
NK1
Th1
b
GATA3
TCF4
BCL6
BATF
NFKB2
JUNB
RUNX3
TBX21
IRF4
TR activity
1
0
Naive CD4 T
Mem CD4 T
CD4 Tfh
Pseudotime
c
TBX1
FOXP3
BCL6
ETV3
NR4A2
NR4A3
BATF
JUNB
TR activity
1
0
Naive CD8 T
Early TEx
Inter TEx
Term TEx
Pseudotime
d
Naive CD8 T
Terminal TEx
377
5813
435
JUNB Target Genes
e
JUNB Differential Target Genes
F11R
LY9
PRMT5
CTLA4
TIGIT
DUSP22
1
0
Naive CD8 T
Terminal TEx
f
JUNB network in Naive CD8 T Cells
TCF7
AQP3
ULBP2
TPM4
TRPV2
LY9
H4C4
F11R
JUNB
PRKCE
SMPD1
PC
PRMT5
KLF10
CASP8
FOSL1
regulation of immune effector process
T cell activation
leukocyte cell-cell adhesion
T cell differentiation
lymphocyte activation involved in immune response
alpha-beta T cell acion
p-value
2e-05
5e-06
10
Gene Numbers
g
JUNB network in Terminal T Exhaustion Cells
TIGIT
H4C12
STAT5B
TRIB1
DUSP22
SMAD3
KLF7
ADTRP
JUNB
CD80
RUNX1
ADGRG1
CTLA4
ADORA2A
CCL3
PARK7
negative regulation of immune system process
negative regulation of response to external stimulus
negative regulation of leukocyte activation
negative regulation of cell activation
negative regulation of cell adhesion
negative regulation of cell migration
p-value
6e-04
2e-04
20
Gene Numbers