OXFORD

Briefings in Bioinformatics, 2023, 1–13
https://doi.org/10.1093/bib/bbad123
Problem Solving Protocol

SCREE: a comprehensive pipeline for single-cell multi-modal CRISPR screen data processing and analysis

Hailin Wei, Tong Han, Taiwen Li, Qiu Wu and Chenfei Wang

Corresponding authors: Chenfei Wang, Key Laboratory of Spine and Spinal Cord Injury Repair and Regeneration of Ministry of Education, Department of Orthopedics, Tongji Hospital, School of Life Science and Technology, Tongji University, 200092, China; Frontier Science Center for Stem Cells, School of Life Science and Technology, Tongji University, 1239 Siping Road, Shanghai 200092, China. Tel.: +86-21-65981195; Fax: +86-21-65981195; E-mail: 08chenfeiwang@tongji.edu.cn; Qiu Wu, Clinical and Translation Research Center of Shanghai First Maternity & Infant Hospital, School of Life Sciences and Technology, Tongji University, 1239 Siping Road, Shanghai 200092, China; Frontier Science Center for Stem Cells, School of Life Science and Technology, Tongji University, Shanghai 200092, China. Tel.: +86-21-65981195; Fax: +86-21-65981195; E-mail: qiu_wu@tongji.edu.cn

Abstract

Single-cell CRISPR screens have been widely used to investigate gene regulatory circuits in diverse biological systems. The recent development of single-cell CRISPR screens has enabled multimodal profiling of perturbed cells with both gene expression, chromatin accessibility and protein levels. However, current methods cannot meet the analysis requirements of different types of data and have limited functions. Here, we introduce Single-cell CRISPR screens data analysEs and perturbation modEling (SCREE) as a comprehensive and flexible pipeline to facilitate the analyses of various types of single-cell CRISPR screens data. SCREE performs read alignment, sgRNA assignment, quality control, clustering and visualization, perturbation enrichment evaluation, perturbation efficiency modeling, gene regulatory score calculation and functional analyses of perturbations for single-cell CRISPR screens with both RNA, ATAC and multimodal readout. SCREE is available at https://github.com/wanglabtongji/SCREE.

Keywords: analyses pipeline, single-cell CRISPR screens, perturbation efficiency, gene regulatory circuits

INTRODUCTION

Pooled CRISPR screens are widely used to screen candidate genes that contribute to specific phenotypes [1–3]. Traditional CRISPR screens use bulk DNA sequencing to quantify the changes in sgRNA count during the screen, and the readouts could measure phenotypes such as cell growth [4], differentiation [5], immune tolerance [6, 7], or drug resistance [8–10]. To further increase the readout for investigating gene regulatory circuits, recent technologies combined CRISPR screens with single-cell sequencing, such as single-cell RNA-sequencing (Perturb-seq [11, 12], CROP-seq [13], CRISP-seq [14] and other protocols [15–18]), single-cell ATAC-sequencing (Perturb-ATAC [19], CRISPR-sciATAC [20] and Spear-ATAC [21]), and multi-modal sequencing (Perturb-CITE-seq [22] and ECCITE-seq [23]). These new technologies generated rich profiles for understanding genotype–phenotype relationships; however, they also posed significant analytical challenges.

The increased amount of single-cell CRISPR screens data calls for a standardized analysis pipeline that could perform raw data processing, quality control, clustering, perturbation modeling and other downstream analyses. In addition, although several methods have been developed to model the targets of the perturbed genes, including MIMOSCA [12], scMAGeCK [24], SCEPTRE [25], GSFA [26] and MUSIC [27], none of them can be directly applied to scATAC-seq-based CRISPR screens. Finally, many current tools have poor computing efficiency and are incapable of handling genome-wide single-cell CRISPR screenings involving millions of cells. To fill these gaps, we introduce SCREE (Single-cell CRISPR screens data analyses and perturbation modeling), a comprehensive and flexible pipeline to streamline raw data processing, quality control, clustering, visualization, perturbation enrichment evaluation, perturbation efficiency modeling, inference of the gene regulatory relationships and functional interpretation of multi-modal single-cell CRISPR screens data.

Hailin Wei Key Laboratory of Spine and Spinal Cord Injury Repair and Regeneration of Ministry of Education, Department of Orthopedics, Tongji Hospital, School of Life Science and Technology, Tongji University, 200092, China. Frontier Science Center for Stem Cells, School of Life Science and Technology, Tongji University, Shanghai 200092, China.
Tong Han Key Laboratory of Spine and Spinal Cord Injury Repair and Regeneration of Ministry of Education, Department of Orthopedics, Tongji Hospital, School of Life Science and Technology, Tongji University, 200092, China. Frontier Science Center for Stem Cells, School of Life Science and Technology, Tongji University, Shanghai 200092, China.
Taiwen Li State Key Laboratory of Oral Diseases, National Clinical Research Center for Oral Diseases, Research Unit of Oral Carcinogenesis and Management, Chinese Academy of Medical Sciences, West China Hospital of Stomatology, Sichuan University, Chengdu, 610041, China. Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Collaborative Innovation Center for Cancer Medicine, Nanjing Medical University, Nanjing, 211166, China.
Qiu Wu Clinical and Translation Research Center of Shanghai First Maternity & Infant Hospital, School of Life Sciences and Technology, Tongji University, Shanghai 200092, China Frontier Science Center for Stem Cells, School of Life Science and Technology, Tongji University, Shanghai 200092, China.
Chenfei Wang Key Laboratory of Spine and Spinal Cord Injury Repair and Regeneration of Ministry of Education, Department of Orthopedics, Tongji Hospital, School of Life Science and Technology, Tongji University, 200092, China. Frontier Science Center for Stem Cells, School of Life Science and Technology, Tongji University, Shanghai 200092, China.
Received: October 17, 2022. **Revised:** February 17, 2023. **Accepted:** March 23, 2023

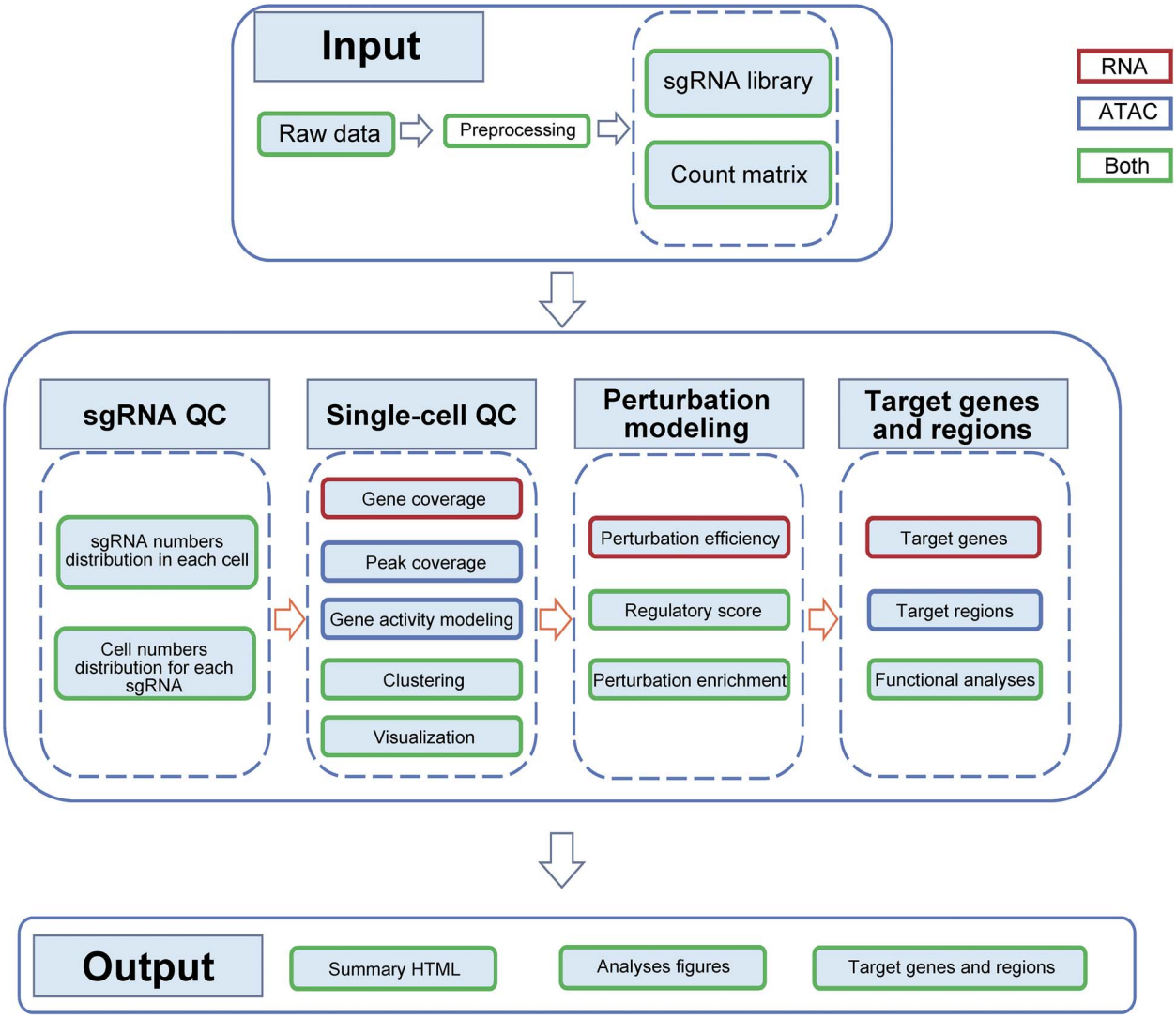

Figure 1. Workflow of SCREE. Schematic of SCREE workflow. SCREE takes either raw sequencing data or processed feature count matrix from single-cell CRISPR screens as input. If use raw data as input, SCREE will generate a gene/peak count matrix and a sgRNA count matrix first. With the feature count matrix and sgRNA information, SCREE performs sgRNA information visualization, single-cell quality visualization and quality control, perturbation enrichment calculation, perturbation efficiency evaluation, regulatory score estimation, potential targets identification and functional analysis. All figures and tables will be saved, and all results will be integrated into a summary HTML file.

SCREE not only provides a standard workspace for single-cell CRISPR screens data but also facilitates access to future functional interfaces.

RESULTS

Overview

SCREE includes two components: preprocessing and analysis (Figure 1). In the preprocessing component, SCREE performs alignment and quantification for both sgRNA and mRNA/DNA using pair-end sequenced FASTQ files as input. For RNA-seq-based screens, SCREE generates gene-level expression count or UMI tables at the single-cell level. For ATAC-seq-based screens, users could choose to either output single-cell count tables based on genomic bins with fixed lengths or peaks called from the bulk level. Additionally, SCREE provides a function for allocating sgRNA to each cell based on the distribution of sgRNA counts. In the analysis component, SCREE performs a set of quality control, visualization and analysis steps using the count matrix as input. These steps include sgRNA information visualization, single-cell quality control, clustering and perturbation enrichment visualization, perturbation efficiency modeling, gene regulatory score estimation, target genes or enhancers identification and functional analysis of potential targets (Methods). All of the results were generated in a well-organized result directory and can be visualized in the summary HTML file (Supplementary Files S1 and S2 available online at http://bib.oxfordjournals.org/).

SCREE is a comprehensive pipeline for analyzing single-cell CRISPR screens data and offers several advantages. First of all, SCREE is flexible and can be applied to data from various experimental designs, including gene perturbations with scRNA-seq, scATAC-seq or multimodal readout, enhancer perturbations. SCREE is also supported for starting with either raw sequencing files or processed count matrices. Second, SCREE is highly integrated and can be easily installed via a simple Conda command, which is user-friendly and has eliminated obstacles in package distribution. Third, SCREE provides an HTML file that summarizes all of the outputs from single-cell CRISPR

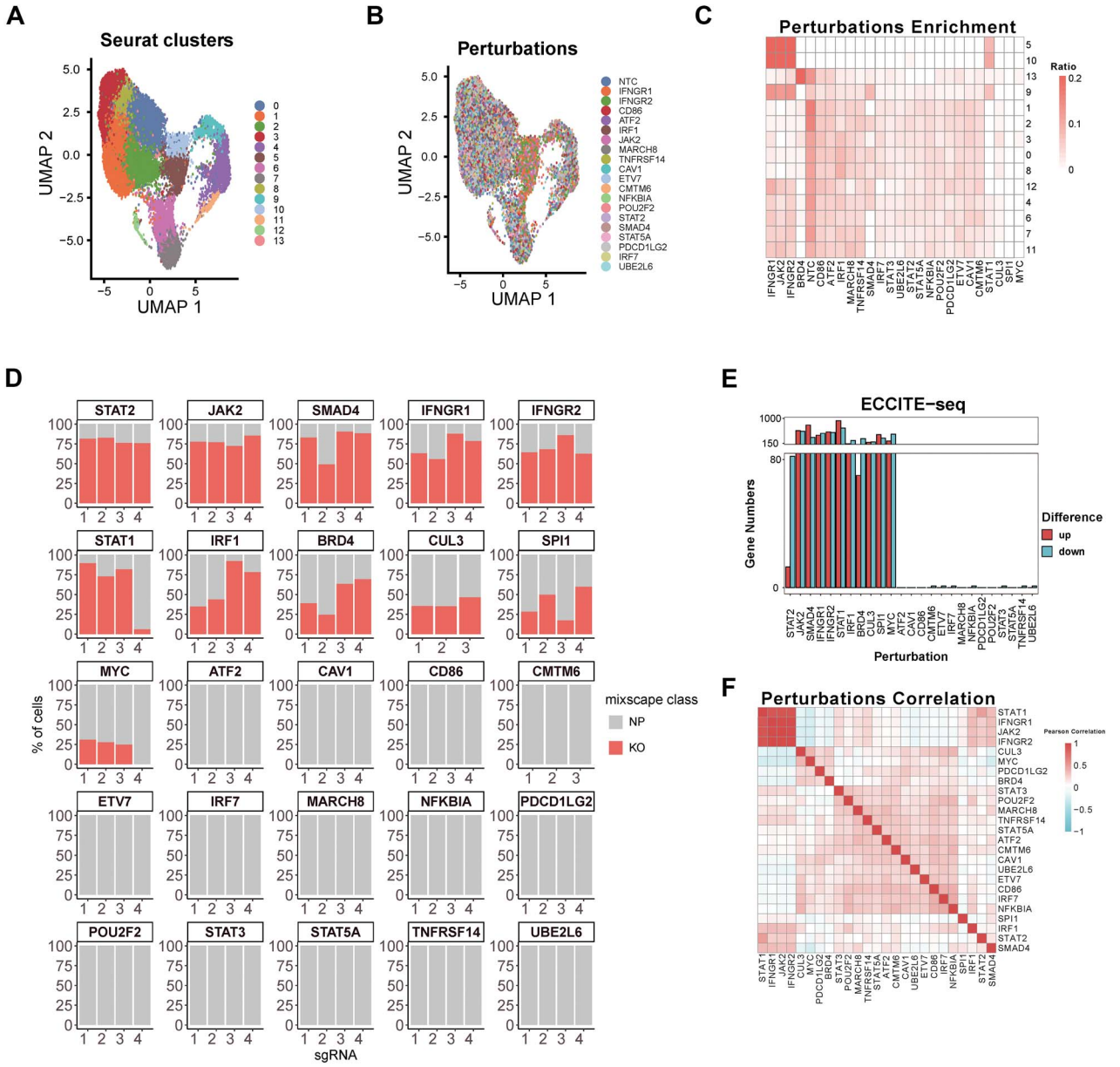

Figure 2. SCREE output of an ECCITE-seq dataset from human THP1 cell line. (**A**) UMAP visualization of clustering results, labeled by unsupervised clusters. (**B**) UMAP visualization of clustering results, labeled by top 20 perturbations with the most cell numbers. NTC: negative control. (**C**) Enrichment ratio of each perturbation in each cluster. NTC: negative control. (**D**) Perturbation efficiency of each perturbation calculated by Mixscape. KO: knockout; NP: non-perturbed. (**E**) Potential target gene numbers inferred from scMAGeCK results ($|score| > 0.2$, $P\_value < 0.05$). (**F**) Pearson correlation of each perturbation, using the regulatory score of a combination of all potential target genes for each perturbation ($|score| > 0.2$, $P\_value < 0.05$).

screen data. All of the functions could also be fine-tuned step-by-step following our detailed protocols, which will help the users better understand the data qualities. Finally, we have optimized several functions of the integrated packages including scMAGeCK, allowing SCREE to handle millions of cells with excellent computing efficiency.

Gene perturbation dataset

To demonstrate the function of SCREE, we first applied it to a gene perturbation dataset using the ECCITE-seq technology from the human THP1 cell line [23]. The sgRNAs in this dataset were designed to target putative regulators of PD-L1 in response to IFN-γ stimulation and this dataset used scRNA-seq as a readout. To visualize the sgRNA information and quality of this dataset, we first performed a quality control analysis (Figure S1A–C available online at http://bib.oxfordjournals.org/). Each cell was assigned with only one sgRNA, and the majority of sgRNAs were distributed throughout hundreds of cells (Figure S1A and B available online at http://bib.oxfordjournals.org/). We next performed a quality control to remove cells with less than 200 detected genes and genes that are detected in less than 1% of cells (Figure S1C available online at http://bib.oxfordjournals.org/, Methods).

To determine whether the perturbations will trigger cell state transition, we performed unsupervised clustering and UMAP visualization (Figure 2A and B). Although there were several distinct clusters (Figure 2A), we were unable to observe perturbations that were enriched in any of the clusters from the UMAP, which might be due to a large number of perturbations in comparison to the clusters (Figure 2B). We therefore introduced an enrichment ratio for each perturbation to visualize the perturbation

enrichment in each cluster (Methods). Using the enrichment ratio, IFNGR1, JAK2, IFNGR2 and STAT1 showed comparable enrichment patterns, and BRD4 was exclusively enriched in cluster 13 (Figure 2C). The perturbation signature for each cell was then normalized, which enhanced the outcomes of the clustering and enrichment visualization (Methods, Figure S1D–F available online at http://bib.oxfordjournals.org/). We employed Mixscape [23] to predict the effectiveness of perturbation for each sgRNA using the perturb signature (Figure 2D). Only 11 of the 25 perturbations had non-zero perturbation efficiency. These perturbations included IFNGR1, JAK2, IFNGR2, STAT1 and BRD4. Notably, different sgRNA of the same perturbation exhibited distinct perturbation efficacy, such as STAT1, IRF1 and BRD4. This information will be helpful for future sgRNA design and selection (Figure 2D).

To further identify the potential targets for each perturbation, we calculated the regulatory score for each perturbation of each gene using scMageck (see section Methods). As expected, perturbations that showed distinct perturbation enrichment also had greater numbers of potential targets than other perturbations (Figure 2C–E and Figure S1F available online at http://bib.oxfordjournals.org/). To investigate whether the perturbations exhibited a co-regulation pattern, we calculated the correlation between each perturbation using the regulatory score of all potential targets (Methods). Consistent with perturbation enrichment results, IFNGR1, JAK2, IFNGR2 and STAT1 displayed a strong positive correlation, suggesting that these factors were co-regulated in response to IFN-γ stimulation (Figure 2C, F and Figure S1F available online at http://bib.oxfordjournals.org/). To illustrate the accuracy of potential target genes, we performed functional analysis on IFNGR1 targets (Figure S1G and H available online at http://bib.oxfordjournals.org/). As expected, genes associated with IFN-γ response, inflammation and antigen processing were down-regulated after IFNGR1 perturbation, which was compatible with the function of IFNGR1 (Figure S1G and H available online at http://bib.oxfordjournals.org/). In addition, we also applied SCREE to another paired gene perturbation dataset using the CROP-seq technology on the human K562 cell line [28] and return similar results (Figure S2 available online at http://bib.oxfordjournals.org/). In summary, our analyses suggest that SCREE can be applied to datasets of gene perturbation in conjunction with scRNA-seq and generate accurate and reasonable results.

Gene overexpression dataset

Besides gene perturbation using CRISPR-based editing, there are also approaches that attempt to overexpress known genes or coding variants via exogenetic vectors [29, 30]. Compare to CRISPR-based editing, which might suffer from the unpredictable off-target effect, these strategies are more stable in modulating gene expressions. To evaluate whether the sgRNA assignment strategy and analytic functions in SCREE can be applied to this type of dataset, we applied SCREE to an OverCITE-seq gene overexpression dataset from human primary T-cells [30]. The sgRNA library of this dataset was composed of around 30 ORFs, which were transduced into CD8+ T cells from a healthy donor for overexpression. Notably, this dataset included information on the cell state (resting T-cells or stimulated T-cells) for each cell. After the ORF assignment, we performed the quality control analysis to display the ORF information and scRNA-seq data quality (Figure S3A–C available online at http://bib.oxfordjournals.org/). A portion of cells was allocated with multiple ORFs, with RAN and CICL1 being assigned to substantially more cells than other ORFs, which might conceivably impact the outcomes of subsequent investigations (Figure S3A and B available online at http://bib.oxfordjournals.org/).

We first investigate whether the perturbation enrichment in SCREE could display a similar trend as the residual of the standardized chi-square test used in the original study. We performed clustering, perturbation enrichment calculation and visualization in SCREE (Figure 3A–C). Although there is no distinct perturbation enrichment from UMAP visualization, LTBR revealed a substantial and unique enrichment in Cluster 8 similar to the original study (Figure 3B and C). Interestingly, several clusters including Cluster 8 were mostly contributed by stimulated T-cells (C0, C3, C6, C8), whereas the majority of other clusters contained resting T-cells (Figure 3D). These analyses confirmed that LTBR might be an important regulator for T-cell activation.

To investigate the downstream targets for LTBR to induce the transition of T-cells from resting to stimulated, we estimated the regulatory score for each ORF relative to each gene and identified thousands of potential targets for each ORF (Figure S3D and E available online at http://bib.oxfordjournals.org/). We then performed functional analysis for LTBR. The top genes with a positive regulatory score in LTBR-overexpressed cells were consistent with the marker genes of Cluster 8 in SCREE results and markers of LTBR-specific cluster in the original study (Figure 3E). Under stimulation, the up-regulated LTBR targets were enriched in the NF-κB signaling pathway, MHC II protein assembly and antigen processing, which might highlight the role of LTBR on CD8+ T cells activation through NF-κB signaling pathway (Figure 3F). In summary, these results suggest that SCREE could also generate reliable results for gene exogenous overexpression data paired with scRNA-seq.

Enhancer perturbation dataset

To evaluate whether SCREE had the potential to investigate enhancer-gene regulations using single-cell CRISPR screens data, we applied SCREE to an enhancer perturbation dataset [31]. To identify prospective enhancer-gene pairs, the sgRNAs in this dataset were intended to target potential enhancer regions but not gene coding regions of the K562 cells. We first depicted the sgRNA information and scRNA-seq quality (Figure S4A–C available online at http://bib.oxfordjournals.org/). Since almost half of the cells were assigned with more than one sgRNA and there were over 1000 sgRNAs designed in this dataset (Figure S4A and B available online at http://bib.oxfordjournals.org/), we did not show the clustering and perturbation enrichment analyses here.

We assessed the regulatory score for each enhancer and performed the same downstream analyses using SCREE (Figure S4D and E available online at http://bib.oxfordjournals.org/). We obtained hundreds of potential targets for a variety of enhancers (Figure 4A). Given the hypothesis that enhancers are more likely to target neighboring genes, we classified potential target genes near the enhancer region (±2 Mb) as direct targets and discovered that almost one-third of enhancers lacked direct targets (Figure 4B). The number of perturbed direct targets, along with the number of total targets from scMageck results, could potentially be used as functional metrics to evaluate the efficiency of enhancer targeting sgRNAs.

To demonstrate SCREE could identify reliable enhancer-gene pairs, we exploited enhancer chr1_26259847_26259870 as a case study and displayed its potential direct targets using Cicero (Figure 4C and Figure S4D, E available online at http://bib.oxfordjournals.org/). The gene expressions of these direct targets were consistent with their regulatory score. For

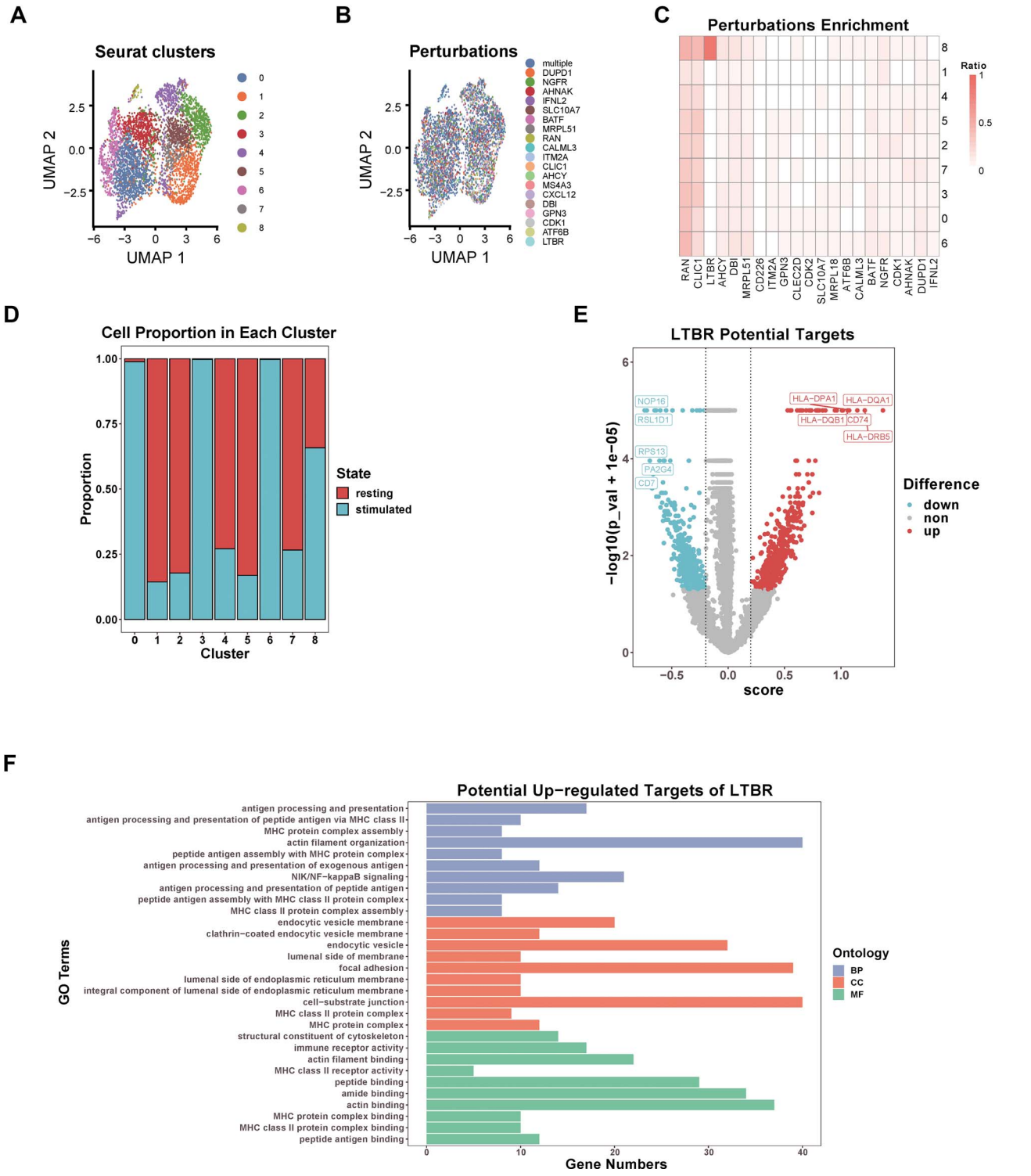

Figure 3. SCREE output of an OverCITE-seq dataset from human primary T-cells. (**A**) UMAP visualization of clustering results, labeled by unsupervised clusters. (**B**) UMAP visualization of clustering results, labeled by perturbations. Multiple: cells with more than one sgRNA assigned. (**C**) Enrichment ratio of each perturbation in each cluster. (**D**) Cell proportion of different cell states in each cluster. Resting: 'resting T-cells'; stimulated: 'stimulated T-cells'. (**E**) Regulatory score and P_value distribution of LTBR. The top 5 up-regulated and down-regulated genes with the highest score are labeled (|score| > 0.2, P_value < 0.05). (**F**) GO enrichment results, which take potential up-regulated targets of LTBR as input (score > 0, P_value < 0.05). BP: biological process; MF: molecular function; CC: cell component.

example, TMEM50A was down-regulated after the enhancer perturbation (Figure 4D). In addition, the expression level of TMEM50A was higher in cells allocated with several enhancers than in cells only assigned one enhancer. In contrast, for LIN28A, both disrupt a single enhancer and multiple enhancers could downregulate its gene expression. These results indicate the complex effect of synergistic enhancers in the same cells. In summary, our analyses suggest that the SCREE could be employed to study enhancer-gene pairs in enhancer perturbation datasets.

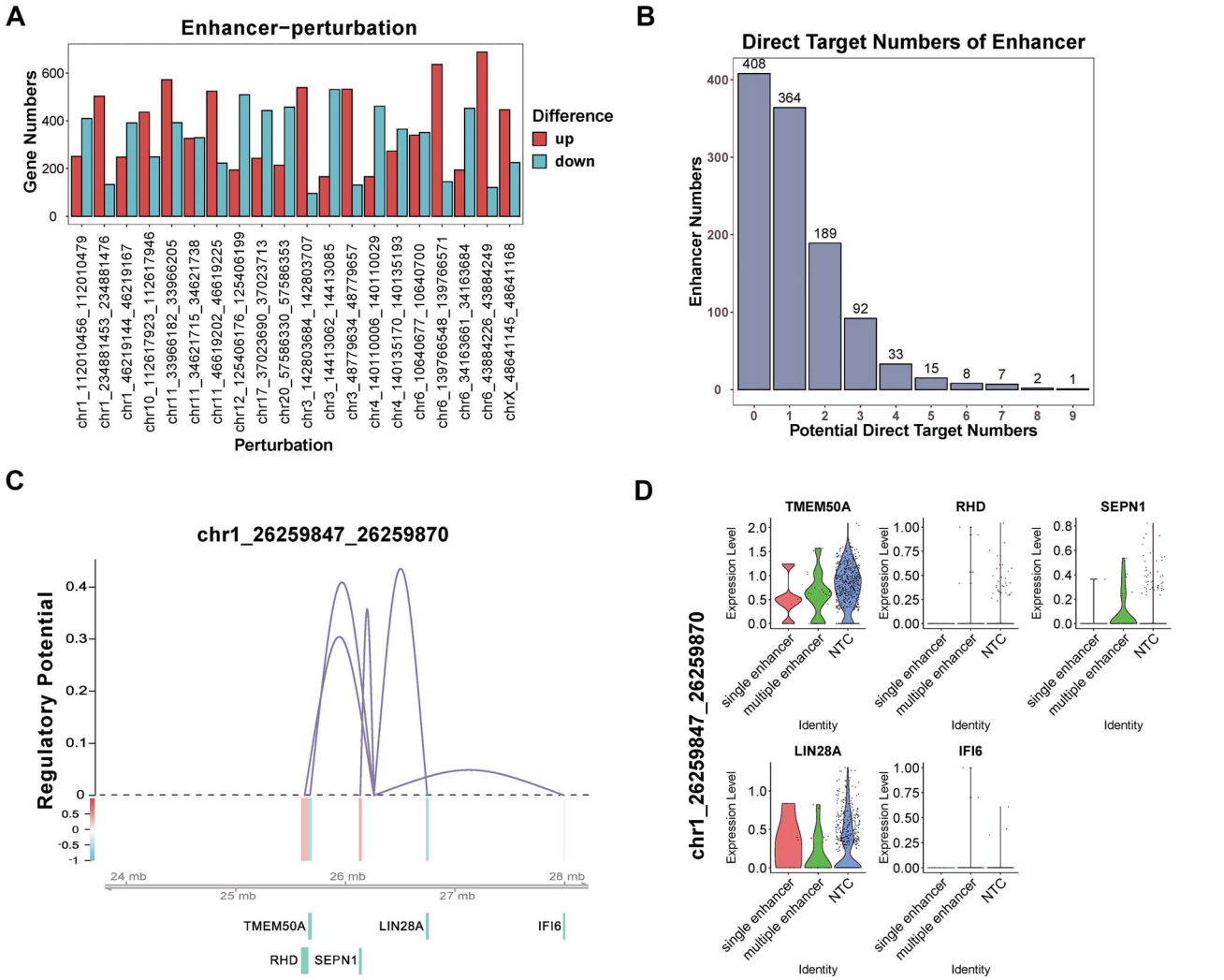

Figure 4. SCREE output of an enhancer-perturbation dataset from human K562 cell line. (**A**) Potential target gene numbers inferred from scMAGeCK results (|score| > 0.2, P_value < 0.05). (**B**) Potential direct targets numbers of each enhancer (|score| > 0.2, P_value < 0.05). (**C**) Regulatory potential of enhancer to surrounding genes ($P < 0.05$). (**D**) Surrounding gene expression level under the condition of enhancer perturbation and negative control. Single enhancer: cells only assigned with specific enhancer; multiple enhancers: cells assigned with more than one sgRNA including specific enhancer; NTC: negative control.

Gene perturbation dataset with scATAC-seq readout

Recently studies combined CRISPR-screen with scATAC-seq to investigate the gene regulatory circuit changes after perturbation [19–21]. To demonstrate the function of SCREE on these datasets, we applied SCREE on a Spear-ATAC dataset with perturbations on 6 transcription factors (TFs) in K562 cells [21]. The fragments size distribution of this dataset exhibited two narrow peaks, one around 50 bp and another around a mono-nucleosomal 180 bp, which is consistent with conventional scATAC-seq (Figure 5A). After sgRNA assignment, we visualized the sgRNA information and scATAC-seq quality (Figure S5A–C available online at http://bib.oxfordjournals.org/). All cells were assigned with a single sgRNA or no sgRNA, and the majority of sgRNA was assigned to less than 200 cells (Figure S5A and B available online at http://bib.oxfordjournals.org/).

To better evaluate the perturbation effect on genes, we converted the peak count matrix to the gene activity matrix to investigate gene regulatory circuits. We performed clustering before and after normalizing the perturb signature (Figure 5B, C and Figure S5D, E available online at http://bib.oxfordjournals.org/). There are no distinguishable clusters or enrichment for both before and after normalizing perturb signatures, which may be caused by the similar function of these TFs on gene regulation. We then identified the potential targets on different TFs using scMageck based on the gene activity matrix. Among these TFs, GATA1 and KLF1 had hundreds of potential targets, while others only had dozens (Figure 5D). Interestingly, all TFs showed a positive correlation with each other and were distinct from the negative control, which confirmed that these TFs might have similar gene regulatory circuits in K562 cells (Figure 5E). Functional analyses on the GATA1 targets revealed its potential functions on regulating endocytosis (Figure S5F and G available online at http://bib.oxfordjournals.org/). Taken together, our analyses suggest that SCREE could robustly identify perturbed genes in scATAC-seq data with CRISPR perturbations.

We next investigate whether SCREE could identify potential perturbed enhancers in the scATAC-seq data. We performed differential accessible peak (DA) analyses on the clustering result using scATAC-seq peaks directly and identified 30 DA peaks, mostly enriched in GATA1 and KLF1 perturbation. The number of DA peaks is consistent with the perturbed genes from

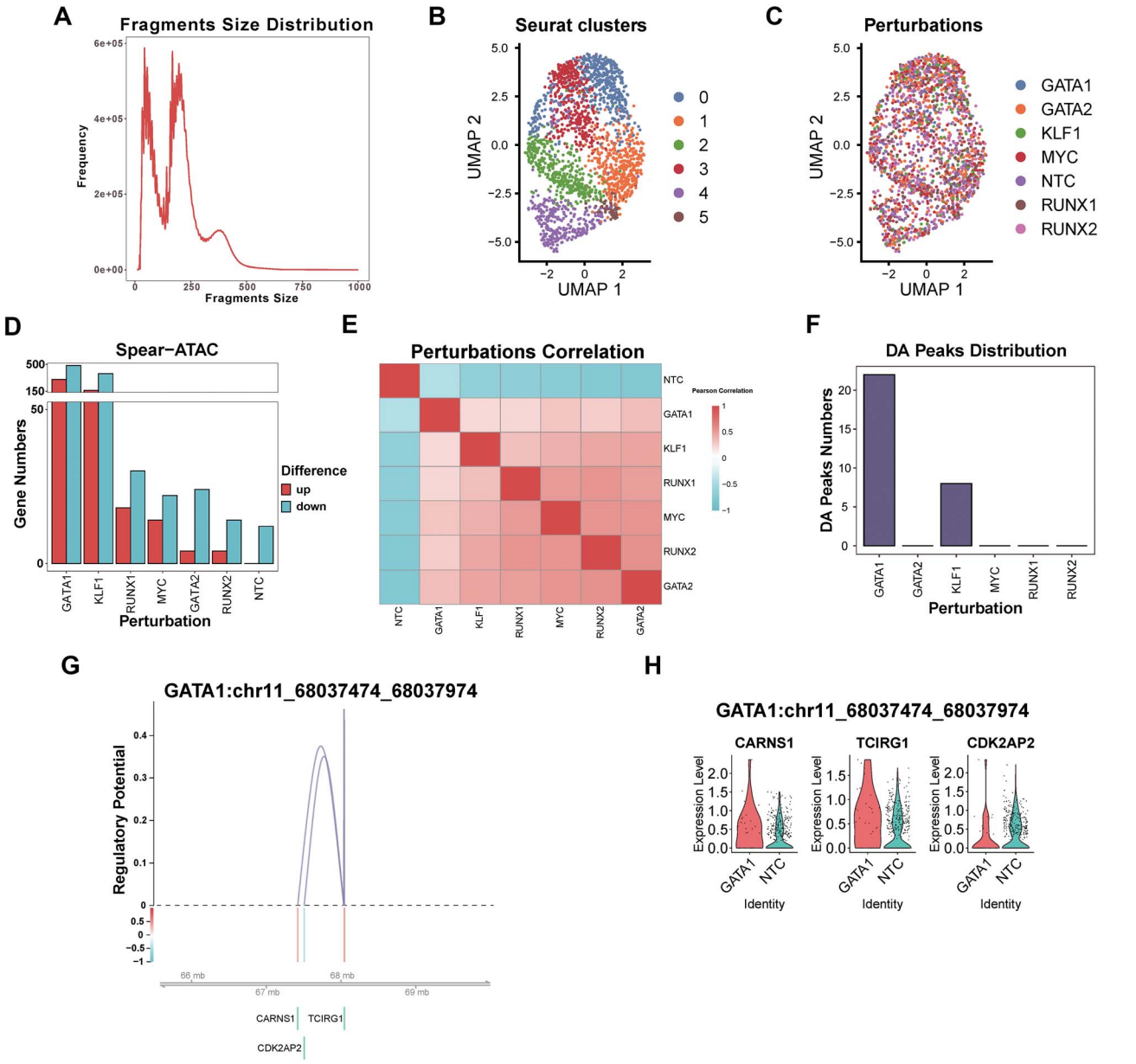

Figure 5. SCREE output of a Spear-ATAC dataset from human K562 cell line. (**A**) Fragments size distribution. (**B**) UMAP visualization of clustering results, labeled by unsupervised clusters, using gene activity matrix. (**C**) UMAP visualization of clustering results, labeled by perturbations, using gene activity matrix. NTC: negative control. (**D**) Potential target gene numbers inferred from scMAGeCK results (|score| > 0.2, *P*_value < 0.05). NTC: negative control. (**E**) Pearson correlation of each perturbation, using the regulatory score of a combination of all potential target genes for each perturbation (|score| > 0.2, *P*_value < 0.05). NTC: negative control. (**F**) DA peaks numbers of each TF. (**G**) Regulatory potential of DApeaks to surrounding genes (*P*_value < 0.05). (**H**) Expression of genes around potential enhancer regions, targeted by GATA1. NTC: negative control.

previous analyses (Figure 5D and F). We then visualized a potential enhancer region in GATA1 perturbation (chr11:68037474–68037974), which showed that the disruption of GATA1 binding on this enhancer significantly upregulates the CARNS1 and TCIRG1 gene while downregulating the CDK2AP2 gene nearby (Figure 5G and H). These results collectively suggest SCREE could be used to identify potential enhancers in gene perturbation datasets paired with scATAC-seq.

Computational efficiency

Several tools, including MIMOSCA [12], MUSIC [27], scMAGeCK [24], GSFA [26] and SCEPTRE [25], have been developed to predict gene regulatory relationships and model gene perturbations using single-cell CRISPR screens data. Among them, scMAGeCK, GSFA and SCEPTRE are R-based workflows that are more convenient for customizing the parameter and analyses. We compared the performance and memory consumption of SCREE with other R-based tools. SCREE demonstrates less time and memory usage compared to other tools (Figure 6, Table S1 available online at http://bib.oxfordjournals.org/). In addition, both GSAF and SCEPTRE failed to generate output for datasets larger than 30 K cells due to memory overflow. In summary, SCREE is a multi-function, multi-modal analysis pipeline for single-cell CRISPR screen data with high computational efficiency.

DISCUSSION

In this study, we introduce SCREE, a comprehensive pipeline to analyze single-cell CRISPR screens data. SCREE is able to

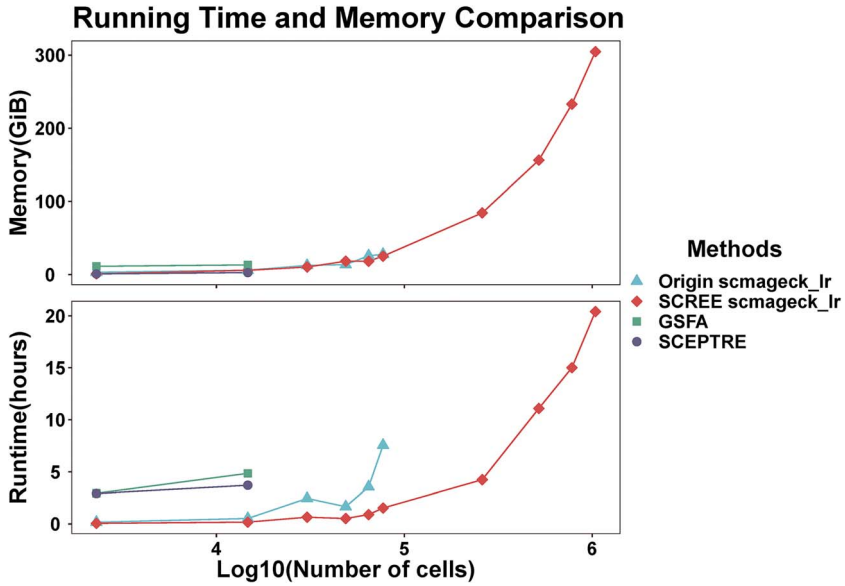

Figure 6. Comparison of SCREE with existing R-based packages. Time usage and memory usage of scmageck_lr function in SCREE, compared to the other two existing packages and the original scmageck_lr function.

perform preprocessing, quality control, clustering, perturbation enrichment visualization, perturbation efficiency modeling, gene regulatory score calculation, enhancer identification and functional analyses for both scRNA-seq data in conjunction with gene perturbations, gene overexpression, enhancer perturbations and scATAC-seq data with gene perturbations (Figure S6). SCREE can flexibly analyze these data from either raw sequencing data or processed count matrix and provides end-to-end solutions for single-cell CRISPR screen data analyses scaling to millions of cells. The comprehensive functions provided in SCREE and its high computational efficiency will facilitate future investigation of the gene regulatory circuits using large genome-wide scale single-cell CRISPR screens.

Despite the advantages, SCREE has a few limitations. First, only sequencing data based on 10× Genomics Inc. platforms can be used for alignment and quantification. Since the alignment component of SCREE is dependent on the cellranger and cellrangerATAC, sequencing data with different UMI/barcode designs need a custom designed preprocess step, and could still use SCREE for the downstream analysis with cell by gene and sgRNA by cell matrix. Second, the current approach for sgRNA assignment is simple and could be prone to errors in sgRNA alignment and affected by the sequencing depth of sgRNAs. Future assignment strategies based on probability modeling or coverage imputation might improve the assignment of sgRNAs [25]. Third, currently, the perturbation efficiency modeling only supports single perturbations as it depends on Mixscape, we hope to develop computational methods that could fit the distribution of cells with multiple sgRNAs and model its perturbation efficiency compare to the cells with single sgRNA. Nevertheless, SCREE offers a platform for incorporating future development of efficiency modeling, perturbation targeting identification and gene regulatory circuits investigation. With these functions updated, we anticipate that SCREE could be widely used and benefit the biomedical community in understanding gene regulatory mechanisms based on single-cell CRISPR screens.

METHODS

Detailed supplementation of SCREE

Alignment and quantification

SCREE performs sequence alignment and expression quantification for scRNA-seq-based datasets using cellranger. For datasets with sgRNA and mRNA reads in separate FASTQ files, SCREE will use the 'multi' subcommand in cellranger, no sgRNA reference file is needed, instead, a table containing the sequence of sgRNA will be used as input. For datasets including both sgRNA and mRNA reads in the same FASTQ files, SCREE will employ the 'count' subcommand of cellranger, utilizing the reference file which combined the genome and sgRNA sequence information. All the gene expressions were quantified by cellranger.

For scATAC-seq-based datasets, SCREE will align sgRNA reads using STAR in cellranger and DNA reads using bwa in cellranger-atac separately. By default, a peak matrix and corresponding fragments file will be generated by cellranger-atac. SCREE also provides a function based on GenomeBinMatrix in Signac [32] to generate a bin-based matrix in addition to the default peak-based matrix, using fragments file from cellranger-atac.

SgRNA assignment for each cell

After quantification of expression and peaks, sgRNA will be assigned to each cell. SCREE adopts an assignment strategy according to the raw read counts and reads proportion in each cell for every sgRNA. By default, sgRNA with read counts >20 and a proportion of total read counts >80% will be assigned to each cell. After assigning sgRNA, SCREE will generate a table including 3 columns: cell, sgRNA and the corresponding gene.

Visualization for sgRNA distribution

After the sgRNA assignment, SCREE can visualize the distribution of cell counts for each sgRNA and the number of sgRNAs in each cell.

Add meta info for each cell

SCREE adopts Seurat to create a SeuratObject based on a gene, peak, or bin count matrix. To facilitate cell filtering, SCREE appends the sgRNA assignment, sgRNA numbers and replicates information to the object. Cells without an assigned sgRNA will be labeled 'blank', whereas cells with more than one assigned sgRNAs will be labeled 'multiple'. Furthermore, cells with only one sgRNA assigned will be tagged with the corresponding gene of sgRNA. Besides the sgRNA and replicate information, SCREE will utilize functions in Seurat and Signac to calculate mitochondrial gene percentage for scRNA-seq based data and fractions of reads in peaks (FRiP) for scATAC-seq based data, separately.

ScRNA-seq quality control and visualization

For scRNA-seq based datasets, SCREE will exclude cells with low overall UMI counts (<1000), expressed gene numbers (<200), or a large fraction of mitochondrial genes (>10%). Since the scMAGeCK [24] algorithm takes the negative control as the baseline for all cells, including 'blank' cells, SCREE may also remove cells without assigned sgRNA optionally. Besides quality control on cells, SCREE will also remove genes that are expressed in less than 1% of cells. Before and after quality control, SCREE can visualize nFeature_RNA, nCount_RNA and the fraction of mitochondrial genes.

ScATAC-seq quality control and visualization

For scATAC-seq-based datasets, SCREE will also remove cells with low total UMI counts (<1000), peak/bin numbers (<200), or FRiP (<10%), as well as optionally remove 'blank' cells. In addition to removing cells, SCREE will also remove peaks or bins that occurred in a small number of cells (default 1%). SCREE enables the visualization of nFeature_peak, nCount_peak and FRiP prior to and after quality control. In addition, using the fragments file generated by sequence alignment, SCREE can also visualize the distribution of fragment size.

Gene activity matrix generation

Gene activity calculation involves identifying the promoter and gene body of each gene and generating a feature matrix based on fragments file, promoter and gene body regions. SCREE adopts Signac to calculate gene activity based on the peak count matrix of scATAC-seq-based datasets and provides a function 'CalculateGeneActivity' to perform the calculation.

Normalization and scale

SCREE adopts Seurat for the normalization of scRNA-seq-based datasets and gene activity matrix. By default, SCREE employs the global scaling normalization method in Seurat to scale the expression in each cell to 10 000, followed by a log transformation. After normalization, SCREE scales the matrix using all genes and regresses the matrix by nCount_RNA and the fraction of mitochondrial genes. For scATAC-seq-based datasets, SCREE uses RunTFIDF in Signac to perform normalization and omits the scaling step.

Feature selection

After normalization, SCREE adopts FindVariableFeatures in Seurat to identify variable features for scRNA-seq-based datasets and gene activity matrices. By default, SCREE adjusts the variance using variance-stabilizing transformation (vst) and returns the top 2000 genes with the highest standardized variance. For scATAC-seq-based datasets, SCREE only adopts FindTopFeatures in Signac to compute the feature metadata, and all input peaks were utilized in the dimension reduction analysis.

Dimension reduction

To minimize the dimension of scRNA-seq-based datasets and gene activity matrices, SCREE performs principal component analysis (PCA) on the top variable features. For scATAC-seq-based datasets, SCREE adopts RunSVD in Signac to reduce the dimension to 50 via latent semantic indexing (LSI). In addition, SCREE computes the correlation between total counts and each reduced dimension component in order to select the dimension for downstream analysis.

Clustering and visualization

After dimension reduction, SCREE performs clustering based on the original Louvain algorithm and shared nearest-neighbor (SNN) graph construction for all kinds of datasets. To visualize the clustering results, SCREE performs Uniform Manifold Approximation and Projection (UMAP) to further compress the dataset to two dimensions. To evaluate the impact of perturbation on clustering, UMAP visualization results will be labeled separately by cluster and perturbations. By default, SCREE uses the top 40 dimensions for scRNA-seq-based datasets and dimensions with an absolute correlation value of less than 0.5 for scATAC-seq-based datasets.

Perturbation enrichment

The perturbation enrichment ratio is defined as the proportion of cell numbers assigned with each perturbation in each cluster. If we define the numbers of cells involving perturbation i in cluster j as m_{ij}, and define the total cell numbers of cluster j as n_j, the ratio can be expressed as:

$$R_{ij} = m_{ij}/n_j$$

SCREE calculates the perturbation enrichment ratio for each perturbation to each cluster and visualizes it by heatmap.

Perturbation signature calculation

For scRNA-seq-based datasets and gene activity matrices, SCREE calculates the perturbation signature for each cell using CalcPerturbSig in Seurat [23]. By default, the 20 nearest negative control cells in the same replicate condition of each cell will be used to calculate the perturbation signature, along with the top 40 dimensions. SCREE will not calculate perturb signature and model perturbation efficiency if too few negative control cells are present in any of the replicates.

Perturbation efficiency modeling

Based on the assay of perturbation signatures, SCREE adopts RunMixscape in Seurat to model perturbation efficiency for each sgRNA. Since this function only accepts cells with one sgRNA as input, only cells with a single sgRNA are included in the perturb signature assay. The default minimum number of differentially expressed genes is 5, and the number of iterations is 10.

Regulatory score estimation

There are three existing R-based methods including scmageck_lr, GSFA and SCEPTRE, which are designed to estimate the regulatory relationship from single-cell CRISPR screens data. GSFA is based on the combination of Bayesian factor analysis and linear regression while SCEPTRE is based on the combination of logistic regression, conditional resampling and negative binomial regression. Compared to the other two similar R-based methods, scmageck_lr

utilizes a simple linear regression model, which will make it easier to process datasets with a large number of cells and will take less memory and running time. Therefore, for scRNA-seq-based datasets and gene activity matrices, SCREE adopts scmageck_lr in scMAGeCK to estimate the regulatory score and *P*-value for each perturbation to each gene. The original steps for selecting highly expressed genes in scmageck_lr will result in memory error when processing datasets with a large number of cells. SCREE performed the selection of highly expressed genes during the scRNA-seq QC step and skipped the parameter for selecting the highly expressed genes in scmageck_lr. In addition, SCREE removed several redundant matrix transposes in the original scmageck_lr function and modifies all the source codes to matrix operations. The final outputs are the same as the original function but with much-improved efficiency.

Potential targets identification

Given the regulatory score and corresponding *P*-value, SCREE identifies potential target genes for each perturbation. By default, for each perturbation, genes with regulatory score >0.2 and *P*-value <0.05 are defined as potential target genes. For scATAC-seq-based datasets, SCREE adopts FindMarkers in Seurat to identify differentially accessible (DA) peaks for each perturbation, compared to negative control cells. By default, the Wilcoxon rank sum test will be used to identify DA peaks, and peaks with log2FC > 0.25 and adjusted *P*-value < 0.05 will be considered as DA peaks.

Visualization for the regulatory score of genes in enhancer regions

For enhancer perturbation datasets, SCREE will visualize the location of potential target genes with a *P*-value < 0.05 as default and its regulatory score for each enhancer. By default, SCREE visualizes the extended region for each enhancer (±2 Mb). For scATAC-seq-based datasets, SCREE identifies DA peaks that do not overlap with the promoter region as potential enhancer regions. By default, the promoter regions for each gene are defined as 2 kb upstream of the transcription start sites (TSS). With the potential enhancer regions and regulatory score from gene activity matrices, SCREE can visualize the regulatory score for the genes in potential enhancer regions. In addition, the visualization function is derived from the plot function of Cicero [33].

Expression of genes surrounding enhancers

To provide further evidence for identifying potential enhancer-gene pairs, SCREE depicts scaled gene expression of all surrounding genes for each enhancer in the enhancer perturbation dataset. For each enhancer, SCREE chooses for visualization those cells assigned to it and the negative control. SCREE illustrates single, multiple and negative controls as distinct conditions due to the fact that cells assigned with multiple sgRNAs may have different surrounding gene expression than cells assigned with a single sgRNA.

Functional analysis

SCREE performs gene ontology (GO) enrichment analysis for each perturbation using the putative target genes of each perturbation. By default, SCREE conducts functional analysis on all potential targets and visualizes the top 10 findings for biological processes, molecular functions and cellular components.

HTML summary

SCREE incorporates all output figures and tables with the parameter information into a summary HTML file to facilitate visualization. In addition, the summary HTML file contains fundamental information about the input dataset, including genome reference version, perturbation numbers, data type, etc.

Implementation of SCREE

The preprocessing functions of SCREE were achieved using python with command-line interfaces for parameter parsers. The analysis functions were packed into an independent R package for repeated analyses. Finally, all of the components of SCREE are provided under the Conda environment for an easy and streamlined installation with a single command.

Data analysis

Gene perturbation data analysis

The ECCITE-seq dataset was obtained from GEO (GSE153056). We used the raw UMI count matrix of RNA assay to create a new SeuratObject and used the sgRNA information to create a table with three columns of information: cells, sgRNA and the corresponding genes. We also retained the replicate information in meta data, as Mixscape would return more accurate results with this information. With the new SeuratObject and the sgRNA table, we first used the Add_meta_data function in SCREE to add metadata of mitochondrial gene percentage, replicate information and sgRNA assignment. To a uniform format, we renamed the negative control in this dataset to 'NTC' and each sgRNA to 'gene_sgRNA1, gene_sgRNA2'.

After adding metadata, we filtered out genes expressed in less than 1% of cells and filter out cells with mitochondrial gene percentage > 10, nFeature_RNA < 200, or nCount_RNA < 1000. After filtering, we normalized and scaled the count matrix using the normalize_scale function with default parameters in SCREE. To simply perform clustering and evaluate perturbation efficiency, we executed IntegratedMixscape in SCREE with default parameters, and the UMAP results labeled by perturbations only displayed the top 20 perturbed genes with the greatest number of cells. To visualize perturbation enrichment of all perturbed genes; however, CalculatePerturbEnrichment was re-executed for both "RNA" and "PRTB" assays.

To estimate the regulatory score, we ran the improved_scmageck_lr in SCREE. For each perturbed gene, we termed genes with |score| > 0.2 and *P*_value < 0.05 as potential targets to perform downstream analyses and visualization (Figure 2E, F and Figure S1G, H available online at http://bib.oxfordjournals.org/).

The paired gene perturbation dataset was obtained from GEO (GSE133344). We only retained sgRNA with good coverage and UMI counts >20 in the cell identities file. We then removed cells with sgRNA of 'NegCtrl10_NegCtrl0', 'NegCtrl1_NegCtrl0', 'NegCtrl11_NegCtrl0' and only retained one negative control sgRNA, 'NegCtrl0_NegCtrl0' to retain fewer cells. We only performed single-cell quality control, sgRNA information visualization, regulatory score calculation, and downstream analyses and visualization. The parameters used in this dataset were the same as the parameters used in the ECCITE-seq dataset (Table S2).

OverCITE-seq data analysis

The OverCITE-seq dataset was obtained from GEO (GSE193736). We converted the ORF count matrix to the same format as the sgRNA table via the sgRNAassign function (freq_cut = 0, freq_percent = 0) in SCREE, without the 'sgRNA' suffix. Using the GEX count matrix, we also created a Seurat object and included meta data. The HTO count matrix was utilized to produce labels

for 'resting T-cells' and 'stimulated T-cells'. We set the replicate to '1' for all cells for simplicity, indicating that there was no replicate. Additionally, we calculated the cell cycle score for each cell using the CellCycleScoring function, as described in the original paper.

We normalize the count matrix using the centered log ratio (CLR) transformation and then scale it by the fraction of mitochondrial genes, nCount_RNA, and an additional cell cycle score. We next perform the same clustering and perturbation enrichment calculation as in ECCITE-seq data processing. We selected 'resting T-cells' to estimate the regulatory score. In contrast, we termed all genes with P_value < 0.05 as potential targets for each ORF to perform downstream analysis and visualization (Figure 3E, F and Figure S2D, E available online at http://bib.oxfordjournals.org/). As in the original paper, we performed GO enrichment analysis on putative LTBR targets (Figure 3F).

Enhancer-perturbation data analysis

The enhancer-perturbation dataset was obtained from GEO (GSE120861). After creating SeuratObject and adding metadata, we performed quality control and normalize the scale using the same cutoffs and parameters as ECCITE-seq data analysis. Mixscape evaluated perturbation efficiency based on the change in transcript profile, however, the majority of the potential enhancer may have no effect on the transcript profile and only affect a small number of genes; thus, we also do not perform Mixscape to this enhancer-perturbation dataset. The regulatory score was then computed for downstream analysis. We termed genes with |score| > 0.2 and P_value < 0.05 as potential targets for each perturbation to perform downstream analysis and visualization (Figure 4 and Figure S3E, F available online at http://bib.oxfordjournals.org/). In addition, we visualize the regulatory of potential direct targets for each enhancer, which means all genes with regulatory P_value < 0.05 in the extended region (±2 Mb) of the corresponding enhancer (Figure 4B and C). For visualization of gene expression, we utilize the EnhancerGeneExpression function in SCREE to visualize the expression of all genes in the extended region for each enhancer (Figure 4D).

Spear-ATAC data analysis

The Spear-ATAC dataset was obtained from GEO (GSE168851). We first used tabix to generate a tbi file for fragments file, and then visualized fragments size distribution via the fragmentsSize function (maxSize = 1000) in SCREE. Since the cell barcode sequences in the sgRNA file were the reverse complement of the cell barcode sequence in the count matrix, we regenerated the sgRNA file according to the correct cell barcode first. The sgRNA table was generated using the sgRNAassign function (freq_cut = 20, freq_percent = 0.8). With the fragments file and peak count matrix, we calculated FRiP for each cell instead of mitochondrial gene percentage as a quality control metric. We filtered out genes expressed in less than 1% of cells and filtered out cells with FRiP < 0.1, nFeature_peak < 200, or nCount_peak < 1000.

To perform common analysis in SCREE, we converted the peak count matrix into the gene activity matrix via the CalculateGeneActivity function (version = "v86", pro_up = 2000, pro_down = 0) in SCREE. Subsequent analysis steps and parameters were identical to those used for ECCITE-seq data analysis. In addition, for each perturbed TF, we identified potential enhancer regions from DA peaks and visualized the regulation relationship in the extended peak region (±2 Mb) via the ATACciceroPlot function in SCREE. DA peaks were identified using the FindMarkers function in Seurat (min.pct = 0.1, logfc.threshold = 0.2).

Running time and memory analysis with the existing packages

To compare the improved scmageck lr function in SCREE with other existing R-based packages, we use 10 different datasets with varying cell counts following quality control. We do not test MUSIC since it is based on the topic model and does not produce a score table for each perturbation to each gene. The 10 datasets we used are from 10× Genomics Inc. (https://www.10xgenomics.com/cn/resources/datasets/5-k-a-549-lung-carcinoma-cells-no-treatment-transduced-with-a-crispr-pool-3-1-standard-6-0-0), GSE90546, single cell portal (https://singlecell.broadinstitute.org/single_cell/study/SCP1064/multi-modal-pooled-perturbcite-seq-screens-in-patient-models-define-novel-mechanismsof-cancer-immune-evasion.), and GSE-90063, GSE168620. Notably, to generate large datasets and a dataset containing millions of cells, we duplicated the matrix of GSE168620 2 times, 3 times and 4 times, which are the datasets containing almost 500 k, 750 k and 1 M cells. We perform quality control, normalize and scale for these datasets with the same criteria as previously.

GSFA (v0.2.8), SCEPTRE (v0.1.0) and scMAGeCK (v1.9.2) are all installed from GitHub. We test GSFA with 1 k iterations in a single script twice, SCEPTRE with 2 k permutations in parallel, scMAGeCK and our function with 2 k permutations. In addition, SCEPTRE removes cells and genes as part of its function, which is unfair to other tools in this comparison, we thus do not include the quality control phase of SCEPTRE.

Key Points

- A comprehensive single-cell CRISPR screen data analyses workflow that supports multi-modal data analyses.
- Rich data pre-processing, analysis and visualization functions that enable one-stop data analyses, quality control, and identifying functional downstream targets and regulatory circuits.
- Easy installation, user-friendly analysis reports, and fast and memory-efficient processing steps.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

FUNDING

National Key R&D Program of China (2022YFA1106000 to C.W.), the National Natural Science Foundation of China (32222026, 32170660 to C.W., 32000561 to Q.W.), Shanghai Rising Star Program (21QA1408200 to C.W.), Natural Science Foundation of Shanghai (21ZR1467600 to C.W.), the Fundamental Research Funds for the Central Universities (20002150073 to C.W.). The authors thank the Bioinformatics Supercomputer Center of Tongji University for offering computing resources.

CODE AVAILABILITY

SCREE is an open-source python and R package with source code freely available at: https://github.com/wanglabtongji/SCREE. The tutorial of SCREE is available at https://hailinwei98.github.io/SCREE.html. We have provided step-by-step command line examples for different processing steps.

Alignment step: https://hailinwei98.github.io/alignment.html
Gene perturbation dataset: https://hailinwei98.github.io/RNA.html
ORF overexpression dataset: https://hailinwei98.github.io/ORF.html
Enhancer perturbation dataset: https://hailinwei98.github.io/enhancer.html
Gene perturbation scATAC-seq dataset: https://hailinwei98.github.io/ATAC.html

DATA AVAILABILITY

ECCITE-seq, SCIFI-seq, OverCITE-seq, enhancer-perturbation and Spear-ATAC datasets are all available in the GEO with the following accession number: ECCITE-seq (GSE153056), SCIFI-seq (GSE133344), OverCITE-seq (GSE193736), enhancer-perturbation (GSE120861), Spear-ATAC (GSE168851).

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

CONSENT FOR PUBLICATION

Not applicable.

AUTHORS' CONTRIBUTIONS

C.W. and Q.W. conceived the project. H.W. developed the algorithm, wrote the workflow and benchmarked the performance. T.H. helped in the data processing. T.L. facilitated testing the large-scale datasets. H.W., Q.W. and C.W. wrote the manuscript. C.W. supervised the whole project. All authors read and approved the final manuscript.

REFERENCES

1. Wang T, Wei JJ, Sabatini DM, *et al.* Genetic screens in human cells using the CRISPR-Cas9 system. *Science* 2014;**343**(6166):80–4.
2. Shalem O, Sanjana NE, Hartenian E, *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* 2014;**343**(6166):84–7.
3. Zhou Y, Zhu S, Cai C, *et al.* High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. *Nature* 2014;**509**(7501):487–91.
4. Hart T, Chandrashekhar M, Aregger M, *et al.* High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. *Cell* 2015;**163**(6):1515–26.
5. Wang T, Birsoy K, Hughes NW, *et al.* Identification and characterization of essential genes in the human genome. *Science* 2015;**350**(6264):1096–101.
6. Burr M, Sparbier C, Chan YC, *et al.* CMTM6 maintains the expression of PD-L1 and regulates anti-tumour immunity. *Nature* 2017;**549**:101–5.
7. Manguso R, Pope H, Zimmer M, *et al.* In vivo CRISPR screening identifies Ptpn2 as a cancer immunotherapy target. *Nature* 2017;**547**:413–8.
8. Kurata M, Rathe S, Bailey N, *et al.* Using genome-wide CRISPR library screening with library resistant DCK to find new sources of Ara-C drug resistance in AML. *Sci Rep* 2016;**6**:36199.
9. Han K, Jeng E, Hess G, *et al.* Synergistic drug combinations for cancer identified in a CRISPR screen for pairwise genetic interactions. *Nat Biotechnol* 2017;**35**:463–74.
10. Shi J, Wang E, Milazzo J, *et al.* Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. *Nat Biotechnol* 2015;**33**:661–7.
11. Adamson B, Norman TM, Jost M, *et al.* A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* 2016;**167**(7):1867–82.
12. Dixit A, Parnas O, Li B, *et al.* Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* 2016;**167**(7):1853–66.
13. Datlinger P, Rendeiro AF, Schmidl C, *et al.* Pooled CRISPR screening with single-cell transcriptome readout. *Nat Methods* 2017;**14**(3):297–301.
14. Jaitin DA, Weiner A, Yofe I, *et al.* Dissecting immune circuits by linking CRISPR-pooled screens with single cell RNA-Seq. *Cell* 2016;**167**(7):1883–96.
15. Xie S, Duan J, Li B, *et al.* Multiplexed engineering and analysis of combinatorial enhancer activity in single cells. *Mol Cell* 2017;**66**(2):285–5.
16. Schraivogel D, Gschwind AR, Milbank JH, *et al.* Targeted perturb-seq enables genome-scale genetic screens in single cells. *Nat Methods* 2020;**17**:629–35.
17. Replogle JM, Norman TM, Xu A, *et al.* Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. *Nat Biotechnol* 2020;**38**:954–61.
18. Jin X, Simmons SK, Guo A, *et al.* In vivo perturb-Seq reveals neuronal and glial abnormalities associated with autism risk genes. *Science* 2020;**370**(6520):eaaz6063. https://doi.org/10.1126/science.aaz6063.
19. Rubin AJ, Parker KR, Satpathy AT, *et al.* Coupled single-cell CRISPR screening and Epigenomic profiling reveals causal gene regulatory networks. *Cell* 2019;**176**(1–2):361–376.e17.
20. Liscovitch-Brauer N, Montalbano A, Deng J, *et al.* Profiling the genetic determinants of chromatin accessibility with scalable single-cell CRISPR screens. *Nat Biotechnol* 2021;**39**(10):1270–7.
21. Pierce SE, Granja JM, Greenleaf WJ. High-throughput single-cell chromatin accessibility CRISPR screens enable unbiased identification of regulatory networks in cancer. *Nat Commun* 2021;**12**(1):2969.
22. Frangieh CJ, Melms JC, Thakore PI, *et al.* Multimodal pooled perturb-CITE-seq screens in patient models define mechanisms of cancer immune evasion. *Nat Genet* 2021;**53**:332–41.
23. Papalexi E, Mimitou EP, Butler AW, *et al.* Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. *Nat Genet* 2021;**53**:322–31.
24. Yang L, Zhu Y, Yu H, *et al.* scMAGeCK links genotypes with multiple phenotypes in single-cell CRISPR screens. *Genome Biol* 2020;**21**(1):19.
25. Barry T, Wang X, Morris JA, *et al.* SCEPTRE improves calibration and sensitivity in single-cell CRISPR screen analysis. *Genome Biol* 2021;**22**(1):344.
26. Zhou Y, Luo K, Chen M, He X. A novel Bayesian factor analysis method improves detection of genes and biological processes affected by perturbations in single-cell CRISPR screening. bioRxiv. https://doi.org/10.1101/2022.02.13.480282.
27. Duan B, Zhou C, Zhu C, *et al.* Model-based understanding of single-cell CRISPR screening. *Nat Commun* 2019;**10**:2233.

28. Norman TM, Horlbeck MA, Replogle JM, *et al.* Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science* 2019;**365**(6455):786–93.
29. Ursu O, Neal JT, Shea E, *et al.* Massively parallel phenotyping of coding variants in cancer with perturb-seq. *Nat Biotechnol* 2022;**40**:896–905.
30. Legut M, Gajic Z, Guarino M, *et al.* A genome-scale screen for synthetic drivers of T cell proliferation. *Nature* 2022;**603**:728–35.
31. Gasperini M, Hill AJ, McFaline-Figueroa JL, *et al.* A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell* 2019;**176**(1–2):377–390.e19.
32. Stuart T, Srivastava A, Madad S, *et al.* Single-cell chromatin state analysis with Signac. *Nat Methods* 2021;**18**:1333–41.
33. Pliner HA, Packer JS, McFaline-Figueroa JL, *et al.* Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol Cell* 2018;**71**(5):858–871.e8.