Privileged Communication

# scART: recognizing cell clusters and constructing trajectory from single-cell epigenomic data

Jingxin Guo1, Jingyu Li1, xxx1, Jiadong Chen and Li Shen1,2#

1 The MOE Key Laboratory of Biosystems Homeostasis & Protection, Life Sciences Institute, Zhejiang University, Hangzhou, Zhejiang 310058, China.

2 Stem Cell Institute, Zhejiang University, Hangzhou, Zhejiang 310058, China.

#To whom correspondence should be addressed

e-mail: li\_shen@zju.edu.cn,

Running Title: single-cell cluster recognition and trajectory construction

Manuscript information: XX pages, XX figures, XX supplemental figures, XX supplemental tables

SUMMARY

Development of genome-wide chromatin accessibility profile through ATAC-seq at single cell resolution (scATAC-seq) offers an opportunity to characterize cellular level epigenetic heterogeneity. However, the sparsity and high-level noise of scATAC-seq data poses unique computational challenges. Here, we introduce scART, a bioinformatics tool for scATAC-seq data analysis, which utilizes some highly stable techniques including KNN imputation, TF-IDF weighting scheme and cell-to-cell cosine similarity measurement to denoise scATAC data and recapitulate underlying heterogeneity. Thus, scART could recognize cell identities and construct the trajectory of cellular states accurately and robustly, especially for data with low sequencing depth. To validate the efficacy of scART, we reconstructed the trajectory of the development of embryonic mouse forebrain and uncovered the dynamic transition of layer neurons. scART is available at XXXXXXXXXXX.

Keywords: scATAC-seq, KNN imputation, TF-IDF, denoising, cellular heterogeneity, cosine similarity, trajectory (建议10个)

INTRODUCTION

A cell is considered the basic unit in a biological individual. Cells in each body share nearly identical genotypes, but the phenotype appeared to be different ([Huang, 2009](#_ENREF_36); [Li and Clevers, 2010](#_ENREF_46); [Shalek et al., 2014](#_ENREF_71)). Cellular heterogeneity is contributed to the variations in expression levels ([Harper et al., 2011](#_ENREF_32); [Rodriguez et al., 2019](#_ENREF_68)). And gene expression is determined by the binding of proteins such as transcription factors and initiators to genomic regulatory elements such as promoters and enhancers ([Cramer, 2019](#_ENREF_21); [Haberle and Stark, 2018](#_ENREF_31); [Lambert et al., 2018](#_ENREF_42)). Single cell RNA-seq (scRNA-seq) is a powerful technology to examine transcriptomes of individual cells to discovery the heterogeneity of complex tissues ([Hashimshony et al., 2012](#_ENREF_33); [Ma et al., 2017](#_ENREF_50); [Macosko et al., 2015](#_ENREF_52)) and reveal cell type specific functional genes for many important biological processes, such as cellular differentiation during development ([Biase et al., 2014](#_ENREF_8); [Goolam et al., 2016](#_ENREF_29); [Xue et al., 2013](#_ENREF_81)), reprogramming ([Biddy et al., 2018](#_ENREF_9); [Treutlein et al., 2016](#_ENREF_77)), tissue regeneration ([Di Talia and Poss, 2016](#_ENREF_26)) and tumor metastasis ([Lawson et al., 2018](#_ENREF_44); [Midde et al., 2018](#_ENREF_56); [Puram et al., 2017](#_ENREF_60)). While scRNA-seq provides the information of inter and intra-cellular variability, investigating through single cell epigenomic assays providing an opportunity to uncover the regulatory landscape in single cells. Recently, the emergence of single-cell assay for transposase-accessible chromatin (scATAC-seq) makes it possible to profile chromatin accessibility landscape to uncover regulatory logic of gene expression programs underlying cell identity ([Lake et al., 2018](#_ENREF_41); [Macaulay and Voet, 2014](#_ENREF_51)).

As the scATAC-seq technology provides an opportunity to understanding the regulatory landscape at single cell resolution, scATAC-seq data analysis comes new challenges. According to scRNA-seq analysis, scATAC-seq analysis can be separated into four parts, 1) data acquisition including alignment, de-barcoded and quantification; 2) data cleaning like removing low quality cells, data imputation, normalization et al.; 3) cell-level analysis like clustering, visualization, trajectory et al.; 4) cell type specific accessible regions identification and annotation. However, due to the low copy number of DNA molecules available for scATAC-seq profiling (two in a diploid organism), the scATAC-seq data is extreme sparsity and has high-level noise, presenting methodological challenges distinct from those of scRNA-seq data analysis.

The first difficulty for scATAC-seq data analysis is defining regions to count the aligned reads. Unlike well-annotated RNA transcripts, the set of accessible regions for most cell types is uncharacterized and has to be learned from data. For most published scATAC-seq methods like chromVAR ([Schep et al., 2017](#_ENREF_70)), Cicero ([Pliner et al., 2018](#_ENREF_58)), LSI ([Cusanovich et al., 2018](#_ENREF_23)), cisTopic ([Bravo Gonzalez-Blas et al., 2019](#_ENREF_10)) and APEC ([Li et al., 2020](#_ENREF_45)), define regions based on peak calling from either a reference bulk ATAC-seq profile or an aggregated single-cell ATAC-seq profile and reads overlapping peaks are counted to create cell-by-peak count matrix (Cell×Peak matrix). However, the use of predefined accessibility peaks asks sufficient number of cells to create robust peak reference and lack the ability to reveal rare populations. In addition to relying on peaks, snapATAC ([Fang et al., 2021](#_ENREF_28)) segments the genomes into uniform-sized bins (windows) and counts the reads within each bin to create cell-by-bin count matrix (Cell×Bin matrix). Since the set of bins covering with the whole genome, all of detected accessible regions are fully utilized to profile single cell chromatin accessibility, comparing the similarity between cells in an unbiased manner.

Another primary difficulty arises from the extremely sparsity of scATAC-seq data. Sparsity is the nature to scATAC-seq data, since signal at any open genomic locus is limited by DNA-copy number; only 0, 1 or 2 reads can be captured through the assay in a diploid genome ([McCarroll and Altshuler, 2007](#_ENREF_54)). Besides, the dropouts that represented as zero values also increase the sparsity of data. For scATAC-seq experiments, dropouts are caused by the small percentage of expected accessible signals (1%~15%) ([Preissl et al., 2018](#_ENREF_59)) captured and the failed amplification. In addition to the sparsity, dropouts also increase noise that largely affecting the downstream analyses. None of strategy have been developed to replace the missing data (dropouts) with substituted values for scATAC-seq data analysis.

Furthermore, dissecting the cellular heterogeneity from such sparse and high-dimensional data presents new challenge for scATAC-seq analysis. Since distance estimation in high dimensions are unreliable making the analysis difficult, a noise-tolerance dimensionality reduction is necessary for evaluating cell-to-cell similarity. Current methods to reduce the dimension can be divided in three classes (Supplement table 1). The first class methods like chromVAR ([Schep et al., 2017](#_ENREF_70)) and Cicero ([Pliner et al., 2018](#_ENREF_58)) selected a set of interested genomic regions, such as TSS, gene body et al. Cells were compared based on the accessibility of these pre-defined regions. While the noise is reduced through selection of regions, many informative regions are also filtered. The second class methods like LSI ([Cusanovich et al., 2018](#_ENREF_23)), cisTopic ([Bravo Gonzalez-Blas et al., 2019](#_ENREF_10)) and APEC ([Li et al., 2020](#_ENREF_45)) projected the raw matrix into a lower dimensional space by selecting some characteristics representing the set a regions. Since information is always lost during projection, the number of lower dimensions used is important. Besides, there are often more than ~106 accessible regions, calculating a set of lower number of features to represent such large number of regions is computational time consuming. The third-class method utilized in snapATAC which converted the binary matrix into a cell-pairwise Jaccard similarity matrix and adopt eigenvector decomposition for dimensionality reduction. However, Jaccard index is highly influenced by the size of the data like uneven sequencing depth or different droplet sizes among cells.

Single-cell sequencing technologies provide new opportunities for studying cellular dynamic processes, such as the cell cycle, cell differentiation and cell activation ([Etzrodt et al., 2014](#_ENREF_27); [Tanay and Regev, 2017](#_ENREF_75)). There are two approaches for trajectory inference in scRNA-seq analysis. The first approach is to use dimensionality reduction techniques to build a cell-cell graph and detect lineage branches like Monocle2 ([Qiu et al., 2017](#_ENREF_63)). The second approach is to use unsupervised clustering to group cells in a low-dimensional space and projecting individual cells onto the branches like Slingshot ([Street et al., 2018](#_ENREF_74)). In order to infer trajectory from extremely sparse and noisy scATAC-seq data ([Zamanighomi et al., 2018](#_ENREF_82)), .current methods like Cicero ([Pliner et al., 2018](#_ENREF_58)) and APEC ([Li et al., 2020](#_ENREF_45)) integrated Monocle2 to construct cellular trajectory. According to monocle2, it projected the cells into lower dimensional space through PCA. Here, the distances between cells are homogenous in cell-by-cell matrix. In order to distinguish the differences among cells, we introduce an approach that identifying low-dimensional density space that the cells lie upon using the whole genome-wide accessible profile of each cell, learning the principal tree and constructing a minimal spanning tree (MST) to project cell on the tree.

In this paper, we present Single-cell Chromatin Accessibility-based cluster Recognition and Trajectory reconstruction (scART) algorithm for scATAC-seq data analysis. Specifically, scART fully utilizes the whole genome-wide regions to count sequencing reads from each cell to minimize the loss of information. In order to reduce the impact of high-level noise, scART adopts k-nearest neighbor (KNN) imputation to replace the missing values ([Beretta and Santaniello, 2016](#_ENREF_7); [Liu et al., 2016](#_ENREF_47)), weights feature regions of each cell by term frequency-inverse document frequency (TF-IDF) normalization weighting scheme, and evaluates the cell-pairwise similarity by cosine value between two cell vectors, which is robustness to noise in cells with low coverage ([Cai et al., 2018](#_ENREF_13)). Compared with published methods, the performance of scART shows its superiority in identifying cellular heterogeneity and constructing cellular lineage with high accuracy and sensitivity, especially for datasets with low sequencing depth. As a demonstration, we applied scART to reconstruct the dynamic changes of cell identities during embryo forebrain development. For annotating the clustering results, scART incorporated other existing tools, like single cell motif variability evaluation, gene accessibility analysis to interpret the biological information from scATAC-seq data. Together, scART can be used to construct a global and high-resolution view of the chromatin accessibility at single-cell resolution.

RESULTS

scART overview and workflow.

The workflow of scART is displayed in Figure 1 and the detailed strategies are shown in Figure S1. Briefly, for scATAC-seq data quantification, scART segments the genome into uniformed-size bins as the set of regions to count reads and creates the single cell cell-by-bin matrix (Figures 1A). Next, scART converts the raw count matrix into binary matrix, in which 0 or 1 indicates the absence or presence of mapped reads fall within that bin (Figure 1B, STAR Methods). To reduce the noise caused by dropouts, scART adopts KNN imputation to replace the missing values. After data imputation, unwanted bins are also filtered to remove noise (Figure 1C, STAR Methods). Then, scART transforms the binary matrix by Term Frequency-Inverse Document Frequency (TF-IDF) for normalization (Figure 1D, STAR Methods). In detail, the differences among cells like sequencing depth can be removed through TF process and cell specific regions can be up-weighted through IDF process. In order to fully utilize the genome-wide accessibility profile of each cell for dissecting cellular heterogeneity, scART directly convert the cell-by-bin matrix into cell-to-cell similarity matrix by cosine similarity (Figure 1E, STAR Methods). Cosine similarity measures the angular cosine value between two cell vectors according to their co-accessible region set and is more robust to noise for sparse and high-dimensional data analysis ([Andrews and Hemberg, 2018](#_ENREF_4); [Sohangir and Wang, 2017](#_ENREF_73)). Using the cosine similarity matrix, truncated SVD is utilized to generate a low-rank matrix approximation for dimensionality reduction (Figure 1E, STAR Methods). Clustering and trajectory analysis are performed based on the lower dimensional space (Figure 1F). For clustering, scART takes the advantage of density-based clustering method, a clustering algorithm does not assume clusters of a particular shape or size ([Rodriguez and Laio, 2014](#_ENREF_67)) (STAR Methods). Besides, trajectory analysis is performed using DDRTree algorithm to learn the principal tree and project each cell onto its nearest location on the tree, and the pseudotime for each cell is assigned by minimal spanning tree (MST) algorithm ([Mao et al., 2015](#_ENREF_53); [Qi et al., 2017](#_ENREF_61); [Qi Ma, 2015](#_ENREF_62)) (STAR Methods). Furthermore, cells within each cluster are pooled together for identification of cluster specific accessible regions through differentially chromatin accessibility analysis (Figure 1F, STAR Methods). Motif analysis is also carried out in each cell or group to infer cluster specific transcript factors, which also provides insights of cell type identification (Figure 1F, STAR Methods).

scART identifies cell types accurately, robustly and sensitively in simulated and real datasets

To benchmark scART, we considered other six recently published algorithms for scATAC-seq analysis: chromVAR, Cicero, Cusanovich2018, cisTopic, snapATAC and APEC. In order to evaluate the clustering performance of scART, we generated simulated datasets of scATAC-seq profiles by down sampling from 13 previously published bulk ATAC-seq datasets of human hematopoietic lineage cell lines ([Corces et al., 2016](#_ENREF_20)) (STAR Methods). The performance of each method in identifying the original cell types is evaluated according to the Adjusted Rand Index (ARI) ([Ailon, 2008](#_ENREF_2)) (STAR Methods). Comparing with published methods, scART clustered cells into their corresponding identities accurately (Figure 2A). On average, more than 90% cells were correctly classified by scART with ARI=0.96 (Figure 2B). Besides, we also compared the accuracy of cell clustering between scART with other methods using two published scATAC-seq datasets. For the dataset with six cultured human cell lines ([Buenrostro et al., 2015](#_ENREF_12)), most of methods were capable to clearly separating the distinct cell line cells (Figure S2A and S2C). The unexpected low ARI values of APEC were resulted from the affection of batch effects (Figure S2E). And for hematopoietic lineage cell dataset ([Buenrostro et al., 2018](#_ENREF_11)), scART and APEC performed better than other methods (Figure S2B and S2D, scART: ARI=0.518; APEC: 0.435). However, the clustering results of APEC also showed the bias of batch effect (Figure S2E).

Based on a recent reports of single-cell ATAC-seq methods, selecting predefined accessibility peaks based on the aggregate signals as the set of regions to count sequencing reads lack the ability to uncover the rare cell populations. Among published method, only snapATAC clustering cells in an unbiased manner and is able to reveal rare populations. Therefore, we compared scART with snapATAC in distinguishing rare population, we generated two simulated datasets from bulk ATAC-seq datasets， in which MPP cells or CD4 T cells made up lower than 1% of the total cells (STAR Methods). According to the visualization of clustering results, scART successfully dissected the rare population of MPP from HSC cells, as well as the rare population of CD4 T cell from CD8 T cell (Figures 2C). Taken together, these results indicated that scART is a sensitive and accurate method for cellular heterogeneity identification.

Furthermore, scART shows to be in high computational speed and scalability. We benchmarked scART and all the other tools by randomly sampled the single-cell chromatin accessibility dataset of human bone marrow and blood cells which contains 63882 cells ([Lareau et al., 2019](#_ENREF_43)). For most of methods, processing of dimensionality reduction is the most complex and time-consuming step (Figures 2D). However, the number of cells within a cluster will increased along with the expansion of scale of total sequenced cells, which will improve the sensitivity of cell type specific peak detection for bulk-level peak calling and increased dimensionality of cell-by-peak matrix. Thus, we took account of all the 571,400 peaks as reported, and randomly sampled different number of cells as the only factor for comparison (STAR Methods). As the number of cells increases, the CPU-time of scART scales at significantly lower slope than other methods. When clustering scATAC-seq dataset with 571400×60000 size, scART only took 570 minutes with 1 CPU thread.

scART outperforms in dissecting cellular heterogeneity in low sequencing depth datasets

In massive single cell sequencing studies, sequencing depth is required to generate high complexity, unbiased and interpretable datasets ([Rizzetto et al., 2017](#_ENREF_66); [Sims et al., 2014](#_ENREF_72)). Unlike scRNA-seq, it’s hard to buffer the impact of complexity by partial sampling through the analysis of high-copy-number molecules ([Tanay and Regev, 2017](#_ENREF_75)). Due to the intrinsic sparsity of scATAC-seq data, scATAC-seq analysis method should be robust to clustering cells from datasets with varying coverage. We generate a series of simulated datasets of scATAC-seq profiles from the bulk hematopoietic cell line ATAC-seq datasets, from 10000 reads per cell (high coverage) to 2500 reads per cell (low coverage) (STAR Methods). As the sequencing depth is reduced lower than 5k reads per cell, all of peak dependent methods lost the ability to identify cell types as the ARI values are lower than 0.5 (Figure 3A). Comparatively, scART and snapATAC outperformed robustly and accurately across all ranges of data sparsity (Figure S3), and scART was more sensitive to uncover differences among cell type for dataset with low sequencing depth (Figure 3B). We also benchmarked the sensitivity of scART, snapATAC and cisTopic on a published scATAC-seq dataset with an average sequencing depth, 3000 reads per cell. It is obviously that scART performed better on identifying the complex cellular heterogeneity in adult mouse brain (Figure 3C and S4). Collectively, scART outperforms the other methods in identifying cell types of different coverages and noise level in both simulated and real datasets.

scART reconstructs the trajectory of embryo mouse forebrain development.

Cellular processes such as proliferation, differentiation and reprogramming are governed by complex epigenetic and transcriptional regulatory programs. Progress through these processes is not clear distinctions between cellular states, but instead a smooth transition, where individual cells represent points along a continuum or lineage([Wagner et al., 2016](#_ENREF_79)). Developmental decisions during lineage commitment are driven by genomic regulatory programs including the concerted action of TFs, chromatin modifiers and other regulatory factors ([Long et al., 2016](#_ENREF_49)). According to the published research, the chromatin accessibility may foreshadow the future expression of lineage-determining genes and can be used to predict lineage choice differentiation of cell states([Clyde, 2021](#_ENREF_19); [Llorens-Bobadilla et al., 2020](#_ENREF_48)). Therefore, we supposed that cellular transitions during development can be inferred according to single cell chromatin accessibility profiles.

In scART, we integrates MST and DDRTree algorithms used in RGE, a population graph-based pseudotime analysis algorithm in scRNA-seq analysis([Mao et al., 2015](#_ENREF_53); [Qi et al., 2017](#_ENREF_61); [Qi Ma, 2015](#_ENREF_62)), to predicted the trajectory based on the lower dimensional space that the cells lie upon and use a cell-cell graph to describe the structure among cell (STAR Methods). By default, scART identifies branch points that describe significant divergences in cellular states automatically. To assess the accuracy of trajectory outputs, we applied scART to reconstruct the differentiated pathways of human hematopoietic cells (Figure S5A). According to the trajectory, scART built a developmental tree with one branch point that led to three outcomes (Figure S5B). Along with the predicted developmental tree, we defined HSC group cells as the root, and CLP, GMP, and MEP group cells were distributed over the remainder of the tree (Figure S5B). The trajectory successfully reflected three main biological programs during HSC differentiation, including lymphoid transcriptional program (path1), erythroid transcriptional program (path2) and myeloid transcriptional program (path3) ([Corces et al., 2016](#_ENREF_20)). By evaluating the deviation of main TFs along the trajectory for HSC differentiation by chromVAR (STAR Methods), the accessible regions of HSC and MPP cells are enriched with HOX motif, as the myeloid, erythroid, and lymphoid differentiation pathways are enriched with GATA1, CEBPB, and EBF1 motif (Figure S5C and S5D), respectively ([Buenrostro et al., 2018](#_ENREF_11); [Schep et al., 2017](#_ENREF_70); [Vilagos et al., 2012](#_ENREF_78)) .

To further demonstrate the utility of scART, we applied scART to predict the development trajectory along neurogenesis of embryonic forebrain development. We utilized a dataset containing 12,733 high-quality snATAC-seq profiles that derived from fetal mouse forebrains at seven developmental stages. scART identified 12 distinct cell populations including three radial glia like cell groups (RG1-3), three excitatory neurons (EX1-3), four inhibitory neurons (IN1-4), astrocytes (AC) and erythro-myeloid progenitors (EMP) and predicted three developmental branches: inhibitory neuron development branch (IN branch), excited neuron development branch (EX branch) and glia cell development branch (Glia branch) (Figure 4A, 4B and 4C). The distinct population exhibited changes in abundance through development (Figure 4D and 4E). Furthermore, scART dissected excited neurons into 3 distinct cell identities which were mixed-clustered in snATAC ([Preissl et al., 2018](#_ENREF_59)). Among excited neurons clusters, EX1 and EX2 appeared at early development stage and showed the characterizations of lower layer neuron, in which the gene accessible score of deep layer markers Foxp2，Bcl11b were higher (Figure 4B) ([Chen et al., 2008](#_ENREF_15); [Chiu et al., 2014](#_ENREF_18)). And EX3 was a group of lower layer neurons with the high accessible score of lower layer markers like Calb1, Cux1 ([Cubelos et al., 2010](#_ENREF_22); [Gray et al., 2017](#_ENREF_30)) (Figure 4B). Besides, EX3 cluster cells expanded at stage E14.5-E15.5, as expected given the timing of cortical layer formation ([Jabaudon, 2017](#_ENREF_38); [Molyneaux et al., 2007](#_ENREF_57)) (Figure 4D and 4E). Thus, scART clustered embryonic forebrain cells sensitively and construct the trajectory of neurogenesis and gliogenesis programs, precisely.

DISCUSSION

In summary, we propose scART, a powerful bioinformatics tool for clustering and trajectory construction for interpreting single cell epigenetic data. Without any prior information, scART could fully utilize the genome-wide accessibility profiles to separate cell populations. In order to reduce the noise and make it easier to identify the underlying structure from sparse, high dimensional scATAC-seq data, scART imputed the dropouts by KNN imputation and ameliorated the technical and biological biases through TF-IDF transformation. Furthermore, scART evaluated cell-to-cell similarity using cosine similarity, a noise-tolerant similarity metric for sparse and high dimensional data analysis. The dimensions of similarity matrix were reduced by SVD for downstream analysis including clustering, visualization and trajectory construction. Based on the clustering results, gene accessible score was calculated by aggregating accessible regions around TSS of each gene and cell types were identified based on the gene accessible score of well-known cell type markers. scART also incorporated chromVAR to estimate the motif variability. Furthermore, scART is applicable to most single-cell chromatin detection techniques including sci-ATAC-seq, snATAC-seq, and SHARE-seq.

Through extensive benchmark, we have demonstrated that scART outperforms existing methods substantially in clustering accuracy, robustness and sensitivity. The superiority of scART could be attributed to the strategies for noise reduction performed before downstream analysis. Differing from previous methods, scART first adopted data imputation strategy to replacing zero values within scATAC-seq dataset. Removing zeroes could make it easier to identify the underlying structure of the data like cell cluster, cellular trajectories and et al. Besides, unlike snapATAC, scART normalized binary values in matrix before evaluating cell-to-cell similarity by TF-IDF transformation. After transformation, bias owing to uneven sequencing depth could be ameliorated and cell-specific accessible regions could be extracted since the values of these regions in matrix were up-weighted.

Although, snapATAC is capable to uncover rare populations, we find that scART is more sensitive to dissect rare populations that sharing similar accessibility profiles with the abundant cell types. One of the reasons for the undesirable performance of snapATAC might be subject to the clustering algorithm it used. Since snapATAC clustered cells using a graph-based clustering method, Louvain algorithm, which identified highly interconnected cells in the graph-structure by representing cells as nodes connected by edges to their k nearest neighbors, assuming equally sized cell populations. Thus, using Louvain algorithm to identify clusters based on KNN graph lacks the sensitivity to detect rare populations with an extremely small percentage. The other reason is that snapATAC hasn’t imputed dropouts which can largely affect the calculation of Jaccard similarity matrix, since Jaccard Index is sensitive to noise.

With the improvement of scATAC-seq technologies, the scale of cells increased in the exponential. In a massive high throughput scATAC-seq experiment, processing millions of cells increases the computational cost quadratically. Especially, with the increase of cell numbers, calculating cell-pairwise similarity matrix becomes computationally infeasible for large-scale dataset. In order to solve the computational requirement of massive datasets, snapATAC applied Nyström method to calculate the low-dimensional embedding for large-scale dataset. snapATAC used the sampled subset of K cells to break the symmetric kernel matrix calculated using normalized Jaccard coefficient. Since the cosine similarity matrix is also a symmetric matrix that can replace the Jaccard coefficient matrix in snapATAC, the Nyström method is also suitable for scART. Because cosine similarity matrix is more robustness to noise, using the cosine similarity matrix might improve the clustering accuracy for large-scale dataset analysis.

To benchmark the size of bins, we compared the clustering accuracy of scART across different bin size resolution 5kb, 10kb, 15kb and 20kb (Figure S1B), measured by the adjusted Rand Index (ARI), Normalized Mutual Information (NMI) and homogeneity score (Methods). We found that scART performed better with 5kb resolution. SnapATAC reported that bin size in the range from 1 to 10 kb appeared to work well on all of the benchmarks. In consideration of the computational cost and resolution, we choose 5 kb as the default bin size for all the analysis in this work.

Despite these advantages of scART, there are still challenges unaddressed. scART asks users to define the parameters varied with datasets and small value of parameters were required to identify rare population cells. In the future, we could adopt automatic clustering algorithms to make it more user-friendly. And for big scale dataset, users can adopt other more scalable clustering methods based on the low-dimensional matrix for clustering. In general, there are still lots of improvement needed to be performed.

METHODS

KEY RESOURCES TABLE

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Li Shen ([li\_shen@zju.edu.cn)](mailto:li_shen@zju.edu.cn)).

METHOD DETAILS

Sequencing reads alignment and counting

Single cell ATAC-seq datasets are aligned to the corresponding reference genome (i.e., mm10 or hg19) using STAR (v.2.4.2) in pair-end mode applying the parameters --alignIntronMax 1, --alignIntronMin 2 and --alignMatesGapMax 2,000. Mapped reads were filtered and sorted with SAMtools (v.0.1.19) by default. Using filtered and sorted bam file, the open chromatin fragments overlapping with uniform-sized (default 5kb) genome bins in each cell were counted to generate raw cell-by-bin count matrix.

For evaluation of computational methods rely on pre-defined accessibility peaks, candidate features were defined by peak calling with MACS2 in each bulk profile merged per cell type (v.2.1.2, with q < 0.001 and nomodel parameters). All of the peaks called by each dataset or cell type were merged to generate a catalog of accessible chromatin regions. The open chromatin fragments overlapping with peaks in each cell were counted to generate raw cell-by-peak count matrix

Matrix binarization

In the cell-by-bin count matrix, we found that the vast majority of the values is “0”, indicating either the real profile of inaccessible (closed chromatin) or missing data. Due to the low copy number of DNA molecules, the count matrix was converted into a binary matrix, in which all of non-zero elements were replaced by “1”.

Data imputation

Due to the limitation of scATAC-seq technologies, scATAC-seq data generally contains many missing values or dropouts. The dropout events can largely increase the noise in scATAC-seq data. Thus, we adopt KNN imputation. KNN imputation fills in the missing data determined by the nearest neighbors with different weights, which is reducing by the distance to the missing value. According to the distribution of accessible regions in the whole genome, the open state of a region often correlated with other open regions, for example, most gene promoters were correlated with at least one accessible peaks within 500 kilobases of TSS ([Dao et al., 2017](#_ENREF_24); [Thurman et al., 2012](#_ENREF_76)). And Many of these peak-to-gene links occur in a continuous regulator units like super enhancers where enriched with multiple nearby peaks ([Hnisz et al., 2013](#_ENREF_35)). Thus, we supposed that the open state of a region can be predicted by the accessibility of its correlated regions within one regulator units and KNN imputation is performed to replace the missing values. In detail, for each binary vector representing the chromatin accessibility profile of a single cell, a bin with value “0” is filled if its k nearest bins with the value “1”, default k=2.

Bin filtering

At first, bins overlapping with the ENCODE blacklist (http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/) are removed to prevent from any potential artifacts. Next, we filter out exceedingly high coverage bins such as bins overlapped with housekeeping gene promoters (default: 90% of cells) to remove uninformative features, as well as bins just appeared in extremely low number of cells (default: 5% of cells) or bins overlapped with mitochondrial DNA to remove noise. The threshold depends on the experimental designed, we suggest users to carefully define the thresholds for bin filtering based on the coverage distribution plots. More specifically, sequenced results of ATAC-seq can be suffer from biases introduced by GC content, which describes the dependence between count (read coverage) and GC content of each mapped fragment dominating the signal of interest for analyses ([Benjamini and Speed, 2012](#_ENREF_6); [Reske et al., 2020](#_ENREF_65)). To remove this technological noise, we estimated the GC content bias of each bin and fitted GC distribution. The distribution of GC bias of each sample showed to be unimodality and bins with high GC bias (default: 0.55) were filtered.

TF-IDF weighting

We apply the term frequency-inverse document frequency (“TF-IDF”) to evaluate the importance of each accessible bin to a cell. According to TF-IDF, the importance increases proportionally to the number of times an accessible region appears in the cell but is offset by the frequency of the region in all cells ([Inbal Yahav, 2019](#_ENREF_37); [Wu et al., 2008](#_ENREF_80)). In detail, we calculate the frequency of each accessible bin within a cell. The frequency is calculated as follows:

Where represents the number of times the peak p appears in cell c, as well as is the set of all the peaks detected in cell c. Since cells with higher chromatin accessibility are likely to have more accessible regions and biased to have more sequenced reads, the importance of an accessible region for cell type identification should be down-weighted. TF process can remove these cell-to-cell variances. We then calculate the inverse frequency of each region across all cells, the ‘‘inverse document frequency’’ as followed:

Here is the number of cells that opened at peak p. IDF process is used for up-weighted the cell-specific bins.

Finally, we multiplied these two weighted values as following:

Since the value of the element in binary matrix just reflect the open state of a region within a cell, the weight value calculated by TF-IDF can be used to extract cell specific open regions.

Cosine similarity evaluation and dimensionality reduction

In order to reduce the dimensionality, we first converted the cell-by-bin matrix to cell-by-cell similarity matrix. The value of each element in similarity matrix was calculated by the cosine value of every two cell vector using vector space model (VSM) ([Aggarwal et al., 2001](#_ENREF_1); [Sohangir and Wang, 2017](#_ENREF_73)). Instead of measuring variation in total accessibility between two cells like Euclidean distance, the cosine similarity measures whether two cells we accessible at the same set of co-accessible regions using angular cosine distance and it is more robust to noise for sparse and high-dimensional data analysis ([Andrews and Hemberg, 2018](#_ENREF_4); [Sohangir and Wang, 2017](#_ENREF_73)). More specifically, where Ci and Cj are cell vectors for cell i and cell j, the similarity between cell Ci and cell Cj is calculated as follows:

We converted the cell-by-bin matrix to cell-by-cell similarity matrix S by cosine similarity between every pair of cells.

We then performed truncated singular value decomposition (SVD) to learn the low-rank matrix approximation of the similarity matrix and projected it into lower dimensional space.

Briefly, SVD attempts to decompose the given matrix into three separate matrices as:

Where is the M×M real or complex unitary matrix, is the M×N rectangular diagonal matrix with non-negative real numbers on the diagonal, and V is the N×N real or complex unitary matrix. The diagonal entries ofare known as the [singular values](https://en.wikipedia.org/wiki/Singular_value) of S. The columns of  are called the left-singular vectors of S, while V is called as the right-singular vector of S. Truncated SVD was performed by selecting r significant singular values and calculating the lower-dimensional matrix S’ as the low-rank matrix approximation, where . At last, matrix X was projected into r-rank lower dimensional space through

where is the number of dimensions in the low-dimensional space.

The primary idea of SVD is to lower the dimension of input data with maximum information gain, we recommend users to look at the plot of the standard deviations of the singular values and draw your cutoff if there is a clear elbow in the graph, default dim=10.

Visualization and Clustering

After dimensionality reduction, we applied t-SNE to map the lower dimensional data to a three dimension (3D) or two dimension (2D) visualization ([Hinton, 2008](#_ENREF_34)). Since t-SNE has a nonconvex objective function, it is possible for different runs to yield different solutions. Thus, we ran t-SNE several times with different initiations and adopted the result with the lowest Kullback-Leibler divergence and best visualization. For clustering, density-based clustering method is applied to partition cells into groups based on the embedded 3D space ([Rodriguez and Laio, 2014](#_ENREF_67)), since the lower embedded latent space remained the coherent structure among cells in high dimension and cells with similar accessibility profiles would group into “clouds” with high density ([Amini et al., 2015](#_ENREF_3)). The algorithm assumes that the centroid of a cluster is surrounded by neighbors with lower local density and that they are at a relatively large distance from any points with a higher local density. Here we identify cluster centers that were characterized by two properties: (i) high local density and (ii) large distance from points of higher density, which are centers of the clusters. Both these quantities depend only on the distances  between cells, which are assumed to satisfy the triangular inequality. The local density  of cell   is defined as

where  if and otherwise, and is a cutoff distance. Basically,  is equal to the number of points that are closer than to point . The algorithm is sensitive only to the relative magnitude of  in different points, implying that, for large data sets, the results of the analysis are robust with respect to the choice of . is measured by computing the minimum distance between the point  and any other point with higher density:

For the point with highest density, we conventionally take . Note that  is much larger than the typical nearest neighbor distance only for points that are local or global maxima in the density. Thus, cells that recognized as points for which the value of  is anomalously large are defined thresholds were considered as centers of cluster. Next, the rest of cells were assigned to the center as described here ([Rodriguez and Laio, 2014](#_ENREF_67)). Clearly, different thresholds will generate different number of clusters.

Differentially accessible region (DA) analysis

To identify regions were specifically accessible in individual cell types, we used a logistic regression framework to test whether cells of a given cluster k were more likely to have Tn5 insertions at a given region relative to all the other cells. We built a generalized additive model (GAM) relate to two predictor variables to a response variable as followed:

For cells in cluster , where p is the probability that the ith region is accessible in the cell, μ is the total proportion of cells that are accessible at the region, α indicates the membership of the cell in the cluster being tested, and ε is an error term for the region. In cluster k, the averaged accessible level of region i among cells is depends on a latent variable:

Where λ is a detection threshold.

Testing for differential accessible regions is performed with a chi-square goodness of fit test. Chi-squared test uses a measure of goodness of fit which is the sum of differences between observed and expected outcome frequencies (that is, counts of observations), each squared and divided by the expectation. The hypotheses take the following form:

The null hypothesis H0: The region is accessible in cluster ,

The alternative hypothesis H1: The region is accessible in cluster ,

Typically, the null hypothesis (H0) specifies the proportion of observations at each level of the categorical variable. The alternative hypothesis (H1) is that at least one of the specified proportions is not true. Thus, we can acquire the statistics:

Where is the observed averaged count for cluster and is the expected count for cluster . The expected frequency is calculated by: . We set a 1% FDR threshold (Benjamin-Hochberg method) to determine whether regions were significantly differentially accessible for each of cell clusters. For this analysis, only regions observed in at least one cell in a given cluster were tested.

Trajectory analysis

We next developed a computational method that integrated reserved graph embedding (RGE) ([Mao et al., 2015](#_ENREF_53); [Qi et al., 2017](#_ENREF_61)) to predict the cell trajectory and encode the position of each cell within the branching structure of the trajectory. RGE simultaneously learns a principal graph that represents the cell trajectory, as well as a function that maps points on the trajectory (which is embedded in low dimensions) back to the original high-dimensional space. Here, we used RGE based on the lower dimensional space. In detail, we first project the cells into density space, in which cells sharing similar chromatin profiles were grouped into local density clouds. Second, we applied DDRTree, the default RGE technique ([Mao et al., 2015](#_ENREF_53); [Qiu et al., 2017](#_ENREF_63); [Reid and Wernisch, 2016](#_ENREF_64)), to learn a latent point for each cell, along with a linear projection function:

,

where , N is the number of cells and Zi is the coordinate vector of cell i in trajectory space. is a matrix with columns that forms an orthogonal basis (D is the dimension of density space calculated by tSNE, D=5; and d is the dimension of latent space to represent structure of trajectory, d=3), that maps the intrinsic space to the 5D density space. DDRTree simultaneously learns a graph on a second set of latent points . These points are treated as the centroids of , where and the principal graph is the spanning tree of those centroids. This scheme works by optimizing

s.t. is a spanning tree

In effect, the algorithm acts as soft K-means clustering on points , and it jointly learns a graph on the K-cluster centers. The matrix with the th element as transforms the hard assignments used in K-means into soft assignments with as a regularization parameter for soft clustering. Due to the distribution of cells in the lower density space, we can find more stable clustering results from soft K-means. We then used the centroids of cell clusters (principal points) to build a principal tree in the trajectory space. Next, we embedded cells to the principal tree Y. For each latent point not near ‘tip’ principal points (i.e., end nodes of the principal tree), we found the nearest line segment on the principal tree and then projects it to the nearest point on that segment. More formally, we defined a vector of a cell , where denote the coordinates of the cell in the latent space T, and the nearest principal point A by . The line segment formed by the two nearest principal points ( and ) is . We then calculate t as

The projection can be calculated as:

For latent points near the tip principal points, we orthogonally project the latent point to the line segment formed by extending the tip principal point and its nearest neighbor principal point in the graph to obtain the projection point, i.e., . In order to encode the position of each cell within the branching structure of the trajectory, we performed a depth-first traversal of the principal tree learned during embedding. Finally, after embedding all of cells to the principal tree, we then calculated the distance between all of the projection points and construct a minimal spanning tree (MST) on the projection points to assign pseudo-time for each cell. To avoid zero values of distance between cells projected to the same principal points, which prevents the calculation of a MST, the smallest positive distance between all cell pairs is added to all distance values.

Methods for the comparison of clustering solutions

To evaluate the similarity of different clustering solutions, we rely on two different metrics. We use the ARI, a metrics routinely applied in the field of clustering, to assess the similarity of clusters to the known truth. ARI takes values from 0 to 1, with 0 signifying no overlap between two groupings and 1 signifying complete overlap and is also applicable in the absence of known cluster labels.

ARI

Let X be a finite set of size n. A clustering solution is a set of non-empty disjoint subsets of X such that their union equals X. Let be a second clustering solution or the supervised labeling solution with the same properties. The contingency table M of the pair of sets , is a matrix whose ith, jth entry equals the number of elements in the intersection of clusters and :

Where ， and . For ease of notation, this is referred to as ARI in the text, dropping the reference to specific pairs of sets.

We calculated stability of clustering solutions by selecting 95% of the samples for each of the datasets randomly and running each method 10 times of cluster identification for all of the compared clustering methods. In each iteration, we calculated the adjusted rand index (ARI) using the cluster assignments in each dataset as the reference for accuracy.

We also calculated normalized mutual information (NMI) and homogeneity score to compare the clustering homogeneity of scART with different resolution.

NMI

where is the entropy of C. Note that

, where and ，is the mutual information of and .

Homogeneity.

Now let us assume C’ is the pre-defined and cluster identity of the cells. Then,

Parameter settings for other algorithms

To evaluate the clustering performance of scART with other scATAC-seq algorithms on the same datasets, we compared scART with chromVAR, Cusanovich2018, Cicero, cisTopic, snapATAC and APEC. The default settings of different tools are used for the most of the comparisons in this paper, except for some parameters have to be defined manually, such as the number of top components in snapATAC.

Preparing the bin-cell count matrix from the raw data

scART adopted the general mapping, alignment, peak calling, and motif searching procedures to process the scATAC-seq data. The paired-end reads were mapped to the human genome (hg19) or mouse genome (mm10) using STAR (v.2.4.2) applying the parameters --alignIntronMax 1, --alignIntronMin 2 and --alignMatesGapMax 2,000. Mapped reads were filtered and sorted with SAMtools (v.0.1.19) by default. Mapped reads were filtered and sorted with SAMtools (v.0.1.19) by default. We next created cell-by-bin count matrices at the resolution of 5kb per bin. Users can apply count matrices preparation steps of snapATAC to create cell-by-bin count matrices of different resolution. For the methods relied on the predefined regions, candidate features were defined by peak calling with MACS2 from the merged single-cell profiles of the same type cells by applying the parameters --nomodel --qval 1e-2 -B --SPMR and --call-summits. All of the peaks called by each cell type are merge together to generate a catalog of accessible chromatin regions and created cell-by-peak matrix.

Downsampling of published single cell ATAC-seq datasets

We calculated stability of clustering solutions by selecting 95% of the samples for each of the datasets randomly and running each method 10 times of cluster identification for all of the compared clustering methods. In each iteration, we calculated the adjusted rand index (ARI) using the cluster assignments in each of the published studies as the reference for accuracy.

Analysis of simulated scATAC-seq profiles from FACS-sorted bulk ATAC-seq datasets from the hematopoietic system.

We used a publicly available FACS-sorted bulk ATAC-seq profiles, which contained 13 different cell types from the hematopoietic lineage The downloaded paired-end reads were mapped to the human genome (hg19) using STAR (v.2.24.12) applying the parameters --alignIntronMax 1, --alignIntronMin 2 and --alignMatesGapMax 2,000. Mapped reads were filtered and sorted with SAMtools (v.0.1.19) by default. Then, we merged the aligned BAM files per cell type using SAMtools merge function (v.0.1.19), and simulated 1300 single-cell epigenomes (100 cells per bulk) by randomly sampling a given number of reads (10,000, 7,500, 5,00 and 2,000 reads per cell in each experiment). For rare population detection, we simulated 6500 single-cell epigenomes (500 cells per bulk) by randomly sampling 10,000 reads per cell and subsampled MPP or CD4 T cell populations to 50 cells, which accounted for around 1% of all cells. Using simulated single cell ATAC-seq datasets, we created a cell-by-bin matrix with 5kb bin size for scART and snapATAC clustering. For the other method, candidate features were defined by peak calling with MACS2 in each bulk profile merged per cell type (v.2.1.21, with q < 0.001 and nomodel parameters). All of the peaks called by each cell type are merge together to generate a catalog of accessible chromatin regions.

Gene annotation and chromatin accessible score.

To access the accessibility of genes, we merged aligned fragments of each cluster to called peaks by MACS2 (v.2.1.1). In each cell, read counts of all peaks at the promoter (20kb upstream of TSS) are aggregated as its raw score (for cell and gene ), then the gene accessible score were defined by normalizing the raw score by , making it in a range comparable to the gene expression from scRNA-seq data.

Motif variability and transcription factor analysis.

scART incorporated chromVAR ([Schep et al., 2017](#_ENREF_70)) for motif variability and transcription factor analysis. To identify potential regulators of chromatin accessibility, we calculated motif variability of known motifs from the JASPAR database (JASPAR\_CORE\_2018\_vertebrates.meme) ([Khan et al., 2018](#_ENREF_40)) in each cell and found the main regulated transcription factors according to the motif analysis results.

Data and software availability

ScATAC-seq datasets used in this study are available from NIH GEO with accession numbers GSE74310, GSE96769, GSE100033, GSE6536 and GSE140203. scART pipeline can be downloaded from the GitHub repository ( ). A Life Sciences Reporting Summary for this paper is available.

Statistics.

Clustering of single nuclei based on chromatin accessibility was performed in an unbiased manner. Cell types were assigned afterwards. To identify differentially accessible regions within a given cell type between developmental stages, a negative binomial test was used and the resulting P value was corrected using the Bonferroni method([Qiu et al., 2017](#_ENREF_63)).

ACKNOWLEDGEMENTS

REFERENCES

Aggarwal, C.C., Hinneburg, A., and Keim, D.A. (2001). On the Surprising Behavior of Distance Metrics in High Dimensional Space. Paper presented at: Database Theory — ICDT 2001 (Berlin, Heidelberg: Springer Berlin Heidelberg).

Ailon, N. (2008). Aggregation of Partial Rankings, p-Ratings and Top-m Lists. Algorithmica *57*, 284-300.

Amini, R., Lisetti, C., and Ruiz, G. (2015). HapFACS 3.0: FACS-Based Facial Expression Generator for 3D Speaking Virtual Characters. Ieee T Affect Comput *6*, 348-360.

Andrews, T.S., and Hemberg, M. (2018). Identifying cell populations with scRNASeq. Molecular aspects of medicine *59*, 114-122.

Baker, S.M., Rogerson, C., Hayes, A., Sharrocks, A.D., and Rattray, M. (2019). Classifying cells with Scasat, a single-cell ATAC-seq analysis tool. Nucleic acids research *47*, e10.

Benjamini, Y., and Speed, T.P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic acids research *40*, e72-e72.

Beretta, L., and Santaniello, A. (2016). Nearest neighbor imputation algorithms: a critical evaluation. BMC Medical Informatics and Decision Making *16*, 74.

Biase, F.H., Cao, X., and Zhong, S. (2014). Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing. Genome research *24*, 1787-1796.

Biddy, B.A., Kong, W., Kamimoto, K., Guo, C., Waye, S.E., Sun, T., and Morris, S.A. (2018). Single-cell mapping of lineage and identity in direct reprogramming. Nature *564*, 219-224.

Bravo Gonzalez-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. Nature methods *16*, 397-400.

Buenrostro, J.D., Corces, M.R., Lareau, C.A., Wu, B., Schep, A.N., Aryee, M.J., 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 e1516.

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.

Cai, S., Georgakilas, G.K., Johnson, J.L., and Vahedi, G. (2018). A Cosine Similarity-Based Method to Infer Variability of Chromatin Accessibility at the Single-Cell Level. Frontiers in genetics *9*.

Cao, J., Cusanovich, D.A., Ramani, V., Aghamirzaie, D., Pliner, H.A., Hill, A.J., Daza, R.M., McFaline-Figueroa, J.L., Packer, J.S., Christiansen, L.*, et al.* (2018). Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science *361*, 1380-1385.

Chen, B., Wang, S.S., Hattox, A.M., Rayburn, H., Nelson, S.B., and McConnell, S.K. (2008). The &lt;em&gt;Fezf2–Ctip2&lt;/em&gt; genetic pathway regulates the fate choice of subcortical projection neurons in the developing cerebral cortex. Proceedings of the National Academy of Sciences *105*, 11382.

Chen, S., Lake, B.B., and Zhang, K. (2019). High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. Nature biotechnology *37*, 1452-1457.

Chen, X., Miragaia, R.J., Natarajan, K.N., and Teichmann, S.A. (2018). A rapid and robust method for single cell chromatin accessibility profiling. Nature communications *9*, 5345.

Chiu, Y.C., Li, M.Y., Liu, Y.H., Ding, J.Y., Yu, J.Y., and Wang, T.W. (2014). Foxp2 regulates neuronal differentiation and neuronal subtype specification. Developmental neurobiology *74*, 723-738.

Clyde, D. (2021). SHARE-seq reveals chromatin potential. Nature reviews Genetics *22*, 2.

Corces, M.R., Buenrostro, J.D., Wu, B., Greenside, P.G., Chan, S.M., Koenig, J.L., Snyder, M.P., Pritchard, J.K., Kundaje, A., Greenleaf, W.J.*, et al.* (2016). Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. Nature genetics *48*, 1193-1203.

Cramer, P. (2019). Organization and regulation of gene transcription. Nature *573*, 45-54.

Cubelos, B., Sebastian-Serrano, A., Beccari, L., Calcagnotto, M.E., Cisneros, E., Kim, S., Dopazo, A., Alvarez-Dolado, M., Redondo, J.M., Bovolenta, P.*, et al.* (2010). Cux1 and Cux2 regulate dendritic branching, spine morphology, and synapses of the upper layer neurons of the cortex. Neuron *66*, 523-535.

Cusanovich, D.A., Hill, A.J., Aghamirzaie, D., Daza, R.M., Pliner, H.A., Berletch, J.B., Filippova, G.N., Huang, X., Christiansen, L., DeWitt, W.S.*, et al.* (2018). A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. Cell *174*, 1309-1324 e1318.

Dao, L.T.M., Galindo-Albarran, A.O., Castro-Mondragon, J.A., Andrieu-Soler, C., Medina-Rivera, A., Souaid, C., Charbonnier, G., Griffon, A., Vanhille, L., Stephen, T.*, et al.* (2017). Genome-wide characterization of mammalian promoters with distal enhancer functions. Nature genetics *49*, 1073-1081.

de Boer, C.G., and Regev, A. (2018). BROCKMAN: deciphering variance in epigenomic regulators by k-mer factorization. BMC bioinformatics *19*, 253.

Di Talia, S., and Poss, K.D. (2016). Monitoring Tissue Regeneration at Single-Cell Resolution. Cell stem cell *19*, 428-431.

Etzrodt, M., Endele, M., and Schroeder, T. (2014). Quantitative single-cell approaches to stem cell research. Cell stem cell *15*, 546-558.

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. Nature communications *12*, 1337.

Goolam, M., Scialdone, A., Graham, S.J.L., Macaulay, I.C., Jedrusik, A., Hupalowska, A., Voet, T., Marioni, J.C., and Zernicka-Goetz, M. (2016). Heterogeneity in Oct4 and Sox2 Targets Biases Cell Fate in 4-Cell Mouse Embryos. Cell *165*, 61-74.

Gray, L.T., Yao, Z., Nguyen, T.N., Kim, T.K., Zeng, H., and Tasic, B. (2017). Layer-specific chromatin accessibility landscapes reveal regulatory networks in adult mouse visual cortex. eLife *6*.

Haberle, V., and Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. Nat Rev Mol Cell Biol *19*, 621-637.

Harper, C.V., Finkenstädt, B., Woodcock, D.J., Friedrichsen, S., Semprini, S., Ashall, L., Spiller, D.G., Mullins, J.J., Rand, D.A., Davis, J.R.*, et al.* (2011). Dynamic analysis of stochastic transcription cycles. PLoS Biol *9*, e1000607.

Hashimshony, T., Wagner, F., Sher, N., and Yanai, I. (2012). CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell reports *2*, 666-673.

Hinton, L.v.d.M.G. (2008). Visualizing Data using t-SNE. Journal of Machine Learning Research.

Hnisz, D., Abraham, Brian J., Lee, Tong I., Lau, A., Saint-André, V., Sigova, Alla A., Hoke, Heather A., and Young, Richard A. (2013). Super-Enhancers in the Control of Cell Identity and Disease. Cell *155*, 934-947.

Huang, S. (2009). Non-genetic heterogeneity of cells in development: more than just noise. Development *136*, 3853-3862.

Inbal Yahav, O.S., and David Schwartz (2019). Comments Mining With TF-IDF: The Inherent Bias and Its Removal. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING *31*.

Jabaudon, D. (2017). Fate and freedom in developing neocortical circuits. Nature communications *8*, 16042.

Ji, Z., Zhou, W., and Ji, H. (2017). Single-cell regulome data analysis by SCRAT. Bioinformatics *33*, 2930-2932.

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J.A., van der Lee, R., Bessy, A., Cheneby, J., Kulkarni, S.R., Tan, G.*, et al.* (2018). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic acids research *46*, D1284.

Lake, B.B., Chen, S., Sos, B.C., Fan, J., Kaeser, G.E., Yung, Y.C., Duong, T.E., Gao, D., Chun, J., Kharchenko, P.V.*, et al.* (2018). Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. Nature biotechnology *36*, 70-80.

Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., and Weirauch, M.T. (2018). The Human Transcription Factors. Cell *172*, 650-665.

Lareau, C.A., Duarte, F.M., Chew, J.G., Kartha, V.K., Burkett, Z.D., Kohlway, A.S., Pokholok, D., Aryee, M.J., Steemers, F.J., Lebofsky, R.*, et al.* (2019). Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. Nat Biotechnol *37*, 916-924.

Lawson, D.A., Kessenbrock, K., Davis, R.T., Pervolarakis, N., and Werb, Z. (2018). Tumour heterogeneity and metastasis at single-cell resolution. Nature cell biology *20*, 1349-1360.

Li, B., Li, Y., Li, K., Zhu, L., Yu, Q., Cai, P., Fang, J., Zhang, W., Du, P., Jiang, C.*, et al.* (2020). APEC: an accesson-based method for single-cell chromatin accessibility analysis. Genome biology *21*, 116.

Li, L., and Clevers, H. (2010). Coexistence of quiescent and active adult stem cells in mammals. Science *327*, 542-545.

Liu, Z.-g., Pan, Q., Dezert, J., and Martin, A. (2016). Adaptive imputation of missing values for incomplete pattern classification. Pattern Recognition *52*, 85-95.

Llorens-Bobadilla, E., Chell, J.M., Le Merre, P., Wu, Y., Zamboni, M., Bergenstrahle, J., Stenudd, M., Sopova, E., Lundeberg, J., Shupliakov, O.*, et al.* (2020). A latent lineage potential in resident neural stem cells enables spinal cord repair. Science *370*.

Long, H.K., Prescott, S.L., and Wysocka, J. (2016). Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. Cell *167*, 1170-1187.

Ma, S., Murphy, T.W., and Lu, C. (2017). Microfluidics for genome-wide studies involving next generation sequencing. Biomicrofluidics *11*, 021501.

Macaulay, I.C., and Voet, T. (2014). Single cell genomics: advances and future perspectives. PLoS genetics *10*, e1004126.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M.*, et al.* (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202-1214.

Mao, Q., Wang, L., Goodison, S., and Sun, Y. (2015). Dimensionality Reduction Via Graph Structure Learning. 765-774.

McCarroll, S.A., and Altshuler, D.M. (2007). Copy-number variation and association studies of human disease. Nature genetics *39*, S37-42.

Mezger, A., Klemm, S., Mann, I., Brower, K., Mir, A., Bostick, M., Farmer, A., Fordyce, P., Linnarsson, S., and Greenleaf, W. (2018). High-throughput chromatin accessibility profiling at single-cell resolution. Nature communications *9*, 3647.

Midde, K., Sun, N., Rohena, C., Joosen, L., Dhillon, H., and Ghosh, P. (2018). Single-Cell Imaging of Metastatic Potential of Cancer Cells. iScience *10*, 53-65.

Molyneaux, B.J., Arlotta, P., Menezes, J.R., and Macklis, J.D. (2007). Neuronal subtype specification in the cerebral cortex. Nature reviews Neuroscience *8*, 427-437.

Pliner, H.A., Packer, J.S., McFaline-Figueroa, J.L., Cusanovich, D.A., Daza, R.M., Aghamirzaie, D., Srivatsan, S., Qiu, X., Jackson, D., Minkina, A.*, et al.* (2018). Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. Molecular cell *71*, 858-871 e858.

Preissl, S., Fang, R., Huang, H., Zhao, Y., Raviram, R., Gorkin, D.U., Zhang, Y., Sos, B.C., Afzal, V., Dickel, D.E.*, et al.* (2018). Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-specific transcriptional regulation. Nature neuroscience *21*, 432-439.

Puram, S.V., Tirosh, I., Parikh, A.S., Patel, A.P., Yizhak, K., Gillespie, S., Rodman, C., Luo, C.L., Mroz, E.A., Emerick, K.S.*, et al.* (2017). Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. Cell *171*, 1611-1624 e1624.

Qi, M., Li, W., Tsang, I.W., and Yijun, S. (2017). Principal Graph and Structure Learning Based on Reversed Graph Embedding. IEEE transactions on pattern analysis and machine intelligence *39*, 2227-2241.

Qi Ma, L.Y., Li Wang, Steve Goodison, Yijun Sun (2015). SimplePPT: A simple principal tree algorithm. In SDM.

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H.A., and Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. Nature methods *14*, 979-982.

Reid, J.E., and Wernisch, L. (2016). Pseudotime estimation: deconfounding single cell time series. Bioinformatics *32*, 2973-2980.

Reske, J.J., Wilson, M.R., and Chandler, R.L. (2020). ATAC-seq normalization method can significantly affect differential accessibility analysis and interpretation. Epigenetics Chromatin *13*, 22.

Rizzetto, S., Eltahla, A.A., Lin, P., Bull, R., Lloyd, A.R., Ho, J.W.K., Venturi, V., and Luciani, F. (2017). Impact of sequencing depth and read length on single cell RNA sequencing data of T cells. Scientific reports *7*, 12781.

Rodriguez, A., and Laio, A. (2014). Machine learning. Clustering by fast search and find of density peaks. Science *344*, 1492-1496.

Rodriguez, J., Ren, G., Day, C.R., Zhao, K., Chow, C.C., and Larson, D.R. (2019). Intrinsic Dynamics of a Human Gene Reveal the Basis of Expression Heterogeneity. Cell *176*, 213-226.e218.

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. Nature biotechnology *37*, 925-936.

Schep, A.N., Wu, B., Buenrostro, J.D., and Greenleaf, W.J. (2017). chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nature methods *14*, 975-978.

Shalek, A.K., Satija, R., Shuga, J., Trombetta, J.J., Gennert, D., Lu, D., Chen, P., Gertner, R.S., Gaublomme, J.T., Yosef, N.*, et al.* (2014). Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. Nature *510*, 363-369.

Sims, D., Sudbery, I., Ilott, N.E., Heger, A., and Ponting, C.P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. Nature reviews Genetics *15*, 121-132.

Sohangir, S., and Wang, D. (2017). Improved sqrt-cosine similarity measurement. Journal of Big Data *4*.

Street, K., Risso, D., Fletcher, R.B., Das, D., Ngai, J., Yosef, N., Purdom, E., and Dudoit, S. (2018). Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC genomics *19*.

Tanay, A., and Regev, A. (2017). Scaling single-cell genomics from phenomenology to mechanism. Nature *541*, 331-338.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B.*, et al.* (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75-82.

Treutlein, B., Lee, Q.Y., Camp, J.G., Mall, M., Koh, W., Shariati, S.A., Sim, S., Neff, N.F., Skotheim, J.M., Wernig, M.*, et al.* (2016). Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq. Nature *534*, 391-395.

Vilagos, B., Hoffmann, M., Souabni, A., Sun, Q., Werner, B., Medvedovic, J., Bilic, I., Minnich, M., Axelsson, E., Jaritz, M.*, et al.* (2012). Essential role of EBF1 in the generation and function of distinct mature B cell types. The Journal of experimental medicine *209*, 775-792.

Wagner, A., Regev, A., and Yosef, N. (2016). Revealing the vectors of cellular identity with single-cell genomics. Nature biotechnology *34*, 1145-1160.

Wu, H.C., Luk, R., Wong, K., and Kwok, K. (2008). Interpreting TF-IDF term weights as making relevance decisions. ACM Trans Inf Syst *26*, 13:11-13:37.

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C.Y., Feng, Y., Liu, Z., Zeng, Q., Cheng, L., Sun, Y.E.*, et al.* (2013). Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature *500*, 593-597.

Zamanighomi, M., Lin, Z., Daley, T., Chen, X., Duren, Z., Schep, A., Greenleaf, W.J., and Wong, W.H. (2018). Unsupervised clustering and epigenetic classification of single cells. Nature communications *9*, 2410.

FIGURE LEGENDS

Figure 1. Overview of scART workflow for scATAC-seq data analysis.

(A) Single cell accessibility quantification: scART segmented the whole genome as uniform-sized bins and built a chromatin accessibility matrix (Cell x Bin matrix) by counting the number of sequenced reads per cell in each of bins.

(B) Dropout imputation: the raw input Cell x Bin matrix is converted to binary count matrix (with “1” indicating a specific bin is accessible in a given cell and “0” denoting inaccessible chromatin or missing data) and KNN imputation is performed to learn and fill missing values.

(C) Bin filtering: undesirable bins such as low or extremely high coverage bins, bins overlapped high GC contents or black list peaks are removed.

(D) TF-IDF weighting: the importance of a bin to a cell is weighted by TF-IDF weighting scheme.

(E) Dimensionality reduction: converting the weighting matrix into cell-pairwise cosine similarity matrix and projecting it into lower-dimensional space by truncated SVD.

(F) Downstream analysis: scART visualizes scATAC-seq data in 2-dimensional (2D) plots, builds the 3-dimensional (3D) trajectory plots and annotates clusters by cell type specific markers and motifs.

Figure 2. Evaluation the performance of scART.

(A) Accuracy. 2D t-Distributed Stochastic Neighbor Embedding (tSNE) diagrams of the simulated HSC scATAC-seq datasets showed the clustering results of different methods, cells are colored by cell type. The human hematopoietic cell lines contain 13 cell types including HSCs (hematopoietic stem cell), MPP (multipotent progenitor cells), CMP (common myeloid progenitor), GMP (granulocyte-macrophage progenitor), LMPP (lymphoid-primed multipotential progenitors), CLP (common lymphoid progenitor), MEP (megakaryocytic-erythroid progenitor), Mono (monocyte), Ery (erythroid), B (B cells), CD4 (CD4+ T-cells), CD8 (CD8+ T cells), and NK (natural killer cells).

(B) Robustness. Adjusted Rand index (ARI) showed the clustering accuracy of each method on the simulated HSC scATAC-seq datasets. Center line, median; box limits, upper and lower quartiles; whisker, 1.5 interquartile range; outliers were removed.

(C) Scalability. Visualization of scART and snapATAC on identifying rare population cells that account for less than 0.1% of the total population by simulated scATAC-seq datasets from bulk ATAC-seq data of 13 human hematopoietic cell lines.

(D)Sensitivity. The computing time required for different algorithms to cluster simulated scATAC-seq with cell number from 5,000 to 100,000, sampled from the 13 previously published bulk ATAC-seq datasets.

Figure 3. The superior performance of scART on dataset with large-scale sparsity.

(A) Visualization of scART and snapATAC on distinguishing minor chromatin accessibility variation among cells by simulated scATAC-seq dataset from bulk ATAC-seq data of 13 human hematopoietic cell lines with varying coverages.

(B) 2D uniform manifold approximation and projection (UMAP) plots of adult mouse brain dataset from different methods. Cells were colored based on the reported clusters. As reported, the clusters were assigned based on both marker genes and scATAC-seq signals.

Figure 4. scART revealed dynamic changes along with the embryonic mouse forebrain development.

(A) 3D visualization showed the pseudotime trajectory of embryonic mouse forebrain development. scART clustered 12377 cells from 7 different developmental stage into 12 cell types including 3 group of radial glia cells (RG1, RG2 and RG3), 4 group of inhibitory neuron cells (IN1-4), 3 group of excitatory neuron cells (EX1-2), astrocyte cell (AC) and erythromyeloid progenitors cell (EMP).

(B) Average scores of the marker genes for each cell cluster identified by scART. The score values were normalized by the standard score (z-score). The bottom row showed the cell types annotated according to the gene score of cell type specific markers.

(C) Hierarchical clustering of the cluster-cluster correlation matrix.

(D) Quantification of cellular composition was calculated at each development stage.

(E) The 3D trajectory diagrams at different stage visualized the dynamic changes of annotated cell types.

Figure S1. Strategies of scART

(A) Complete workflow of the scART.

(B) Visualization of clusters identified across different bin size selection of human hematopoietic scATAC-seq data. The Adjusted Rand Index (ARI), the Normalized Mutual Information (NMI), and the homogeneity scores are shown on the top.

Figure S2. Validation of scART performance relative to alternative methods on published scATAC-seq datasets

(A) The 2D tSNE diagrams showed the clustering accuracy of different algorithms on real human hematopoietic cell line (HSC) scATAC-seq datasets with 8 cell types. The table blow the diagrams showed the ARI value of the clustering results for each algorithm.

(B) The 2D tSNE diagrams showed the clustering accuracy of different algorithms on human cell line scATAC-seq datasets with 6 cell types. The table blow the diagrams showed the ARI of the cell clustering results for each algorithm.

(C) Boxplots showed the ARI values for clustering stability of scART and published algorithm on HSC cell line datasets. We subsampled 90% of the cells from each dataset 10 times and re-run different analysis algorithm with default parameters to evaluate the clustering robustness. Center line, median; box limits, upper and lower quartiles; whisker, 1.5 interquartile range; outliers were removed.

(D) Boxplots showed the ARI values for clustering stability of scART and published algorithm on human cell line scATAC-seq datasets with 6 cell types. We subsampled 90% of the cells from each dataset 10 times and re-run different analysis algorithm with default parameters to evaluate the clustering robustness. Center line, median; box limits, upper and lower quartiles; whisker, 1.5 interquartile range; outliers were removed.

(E) Visualization of scATAC-seq datasets with different annotation showed scART is capable to identify potential batch effects than APEC. Cells were colored by clustering results, published cell types and batches of different donors.

Figure S3. Validation scART performance relative to alternative methods on simulated scATAC-seq datasets with different sequencing depth

Method comparison on 1,300 simulated scATAC-seq data down sampled from 13 bulk ATAC-seq datasets with varying coverages.

Figure S4. scART accurately identified cell subtypes for SHARE-ATAC-seq dataset.

Heatmaps showed the proportion of cells in the reported clusters that overlapped in ATAC clusters in the mouse brain from scART, snapATAC and cisTopic. ARI value shows the clustering performance of each method.

Figure S5. scART constructed differentiation pathways from scATAC-seq data from human hematopoietic cells.

(A) The differentiation hierarchy of human hematopoietic cells.

(B) Visualization of scATAC-seq data from human hematopoietic cells. Left, 2D plot showed the visualization of HSC lineage cell, cells are colored by cell type. Right, 3D visualization of the pseudotime trajectory. scART predicted three different developmental pathways of human hematopoietic cells from scATAC-seq data. Path1 represents for lymphoid differentiation pathway, path2 for erythroid differentiation pathway and path3 for differentiation transcriptional pathway.

(C) The deviations of significant differential motifs (HOXD11, HOXB, STAT1, CEBPB, GATA1 and EBF1) were plotted on the pseudotime trajectory.

(D) Variability associated with trans-factor motifs across 8 cell types. Each row was normalized to the maximum variability for that motif across cell types.

SUPPLEMENT TABLE 1

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Method | Define regions | Regions used for analysis | Imputation | Binary | Filtering | Transformation/Normalization | Matrix conversion | Trajectory | Availability | Ref |
| chromVAR | Peak | Peaks that overlapped with genomic regions like TSS, gene body and TF motifs and et al. |  |  | √ | √ |  |  | R package | ([Schep et al., 2017](#_ENREF_70)) |
| SCRAT | Peak | Peaks that overlapped with features like TF motifs, DHSs defined by ENCODE, regions of genes in genome annotation and GSEA database. |  |  |  |  |  |  | GUI | ([Ji et al., 2017](#_ENREF_39)) |
| scABC | Peak | Pre-defined peaks |  |  | √ | √ |  |  | R package | ([Zamanighomi et al., 2018](#_ENREF_82)) |
| Cicero | Peak | Pre-defined peaks |  |  |  | √ | Cell×CCAN | √ | R package | ([Pliner et al., 2018](#_ENREF_58)) |
| LSI | Peak | Pre-defined peaks |  | √ | √ |  |  |  | - | ([Cusanovich et al., 2018](#_ENREF_23)) |
| Scasat | Peak | Pre-defined peaks |  | √ | √ | √ |  |  | Jupyter notebook | ([Baker et al., 2019](#_ENREF_5)) |
| BROCKMAN | Peak | Pre-defined peaks |  |  |  |  |  |  | R package | ([de Boer and Regev, 2018](#_ENREF_25)) |
| cisTopic | Peak | Pre-defined peaks |  | √ |  |  | Cell×topic |  | R package | ([Bravo Gonzalez-Blas et al., 2019](#_ENREF_10)) |
| SCALE | Peak | Pre-defined peaks |  | √ | √ |  |  |  | Python | ([Satpathy et al., 2019](#_ENREF_69)) |
| APEC | Peak | Pre-defined peaks |  |  | √ |  | Cell×accesson | √ | Python | ([Li et al., 2020](#_ENREF_45)) |
| snapATAC | Bin | Uniform-sized bins |  | √ | √ | √ | Cell×Cell  (Jaccard similarity) | √ | R package | (Fang et al., 2021) |
| scART | Bin | Uniform-sized bins | √ | √ | √ | √ | Cell×Cell  (Cosine similarity) | √ | R package |  |

Supplementary Table 1. Comparison of current bioinformatics methods for analyzing single-cell ATAC-seq data. **Define regions** indicates the strategy to define regions to count the sequencing reads of each method. Peak means the method defined regions based on peak calling from either a reference bulk ATAC-seq profile or an aggregated single-cell ATAC-seq profile. Bin means the method defined regions by the set of uniform-sized bins that segmented the whole genome. **Regions used** means the genomic regions that the method used for analysis. **Imputation** indicates that whether these methods replaced the missing values caused by dropouts. **Binary** indicates that whether represent count matrix as a binary accessibility matrix. **Filtering** indicates that these methods selected high quality cells and regions for analysis. **Transformation/Normalization** indicates that whether these methods transformed the raw accessibility matrix to a new representation. **Matrix conversion** indicates that whether these methods convert the input Cell×Peak/Cell×Bin matrix into new matrix for dimensionality reduction. **Trajectory** indicates that these methods can perform cellular trajectory.