A reference-guided approach for epigenetic characterization of single cells

Shengquan Chen^{1,2†}, Guanao Yan^{3,†}, Wenyu Zhang², Jinzhao Li², Rui Jiang^{1,*} & Zhixiang Lin^{2,*}

¹MOE Key Laboratory of Bioinformatics, Bioinformatics Division, Beijing National Research Center for Information Science and Technology, Department of Automation, Tsinghua University, Beijing 100084, China

²Department of Statistics, The Chinese University of Hong Kong, Sha Tin, Hong Kong SAR, China

³School of Mathematical Sciences, Zhejiang University, Hangzhou 310007, China

[†]These authors contributed equally: Shengquan Chen, Guanao Yan

*Correspondence should be addressed to: ruijiang@tsinghua.edu.cn, zhixianglin@cuhk.edu.hk

Abstract

The recent advancements in single-cell technologies, including single-cell chromatin accessibility sequencing (scCAS), have enabled profiling the epigenetic landscapes for thousands of individual cells. However, the characteristics of scCAS data, including high dimensionality, high degree of sparsity and high technical variation, make the computational analysis challenging. Reference-guided approach, which utilizes the information in existing datasets, may facilitate the analysis of scCAS data. We present RA3 (Reference-guided Approach for the Analysis of single-cell chromatin Accessibility data), which utilizes the information in massive existing bulk chromatin accessibility and annotated scCAS data. RA3 simultaneously models 1) the shared biological variation among scCAS data and the reference data, and 2) the unique biological variation in scCAS data that identifies distinct subpopulations. We show that RA3 achieves superior performance in many scCAS datasets. We also present several approaches to construct the reference data to demonstrate the wide applicability of RA3.

Chromatin accessibility is a measure of the physical access of nuclear macromolecules to DNA and is essential for understanding the regulatory mechanism^{1,2}. For rapid and sensitive probing of chromatin accessibility, assay for transposase-accessible chromatin using sequencing (ATAC-seq) directly inserts sequencing adaptors into accessible chromatin regions using hyperactive Tn5 transposase *in vitro*³. With the recent advancements in technology, single-cell chromatin accessibility sequencing (scCAS) further enables the investigation of epigenomic landscape in individual cells^{4,5}. However, the analysis of scCAS data is challenging because of its high dimensionality and high degree of sparsity, as the low copy number (two of a diploid-genome) of DNA leads to only 1-10% capture rate for the hundreds of thousands of possible accessible peaks⁶. The proposed approaches for the analysis of single-cell RNA-seq (scRNA-Seq) data thus present limitations due to the novelty and assay-specific challenges of extreme sparsity and tens of times higher dimensions⁶.

Several computational algorithms have been proposed to analyze scCAS data. chromVAR assesses the variation of chromatin accessibility using groups of peaks that share the same functional annotations⁷. scABC calculates weights of cells based on the number of distinct reads within the peak background and then uses weighted *k*-medoids to cluster the cells⁸. cisTopic applies latent Dirichlet allocation (LDA) model to explore *cis*-regulatory regions and characterizes cell heterogeneity from the generated regions-by-topics and topics-by-cells probability matrices⁹. Cusanovich et

al. proposed a method that performs the term frequency-inverse document frequency transformation (TF-IDF) and singular value decomposition (SVD) iteratively to get the final feature matrix^{5,10}. Scasat uses Jaccard distance to evaluate the dissimilarity of cells and performs multidimensional scaling (MDS) to generate the final feature matrix¹¹. SnapATAC segments the genome into fixed-size bins to build a bins-by-cells binary count matrix and uses principal component analysis (PCA) based on the Jaccard index similarity matrix to obtain the final feature matrix¹². SCALE combines a variational autoencoder and a Gaussian mixture model to learn latent features of scCAS data¹³.

Incorporating reference data in analyzing single-cell genomic data can better tackle the high level of noise and technical variation in single-cell genomic data. Most reference-guided methods are designed for single-cell transcriptome data, and they focus primarily on cell type annotation using reference data and marker genes, which limits their application to other downstream analyses, such as data visualization and trajectory inference^{14–24}. Massive amounts of bulk chromatin accessibility data has been generated from diverse tissues and cell lines^{25,26}. The amount of scCAS data is rapidly increasing^{4,5,10,27,28}. Meanwhile, computational tools that collect chromatin accessibility data and efficiently compute chromatin accessibility over the genomic regions facilitate the construction of reference data^{29,30}.

To utilize the information in existing chromatin accessibility datasets for the analysis of scCAS data, we propose a probabilistic generative model RA3, in short for Reference-guided Approach for the Analysis of single-cell chromatin Accessibility data. Incorporating reference data built from bulk ATAC-seq data, bulk DNase-seq data, and pseudo-bulk data by aggregating scCAS data, RA3 effectively extracts biological variation in single-cell data for downstream analyses, such as data visualization and clustering. RA3 not only captures the shared biological variation between single-cell chromatin accessibility data and reference data, but also captures the unique biological variation in single-cell data that is not represented in the reference data. RA3 can model known covariates, such as donor labels. Through comprehensive experiments, we show that RA3 consistently outperforms existing methods on datasets generated from different platforms, and of diverse sample sizes and dimensions. In addition, RA3 facilitates trajectory inference and motif enrichment analysis for more biological insight on the cell subpopulations.

Results

The RA3 model

RA3 is a generative model based on the framework of probabilistic principal component analysis³¹, and it decomposes the total variation in scCAS data into three components: the component that captures the shared biological variation with reference data, the component that captures the unique biological variation in single cells, and the component that captures other variations (Fig. 1a). More specifically, the first component utilizes the prior information of the projection vectors learned from reference data, and it captures the variation in single-cell data that is shared with the reference data. Choice of the reference data is flexible: it can be the chromatin accessibility profiles of bulk samples or pseudo-bulk samples by aggregating single cells. In practice, the reference data can be incomplete: novel cell types or novel directions of biological variation that are not captured in the reference data can be present in the single-cell data. The second component captures the unique biological variation in single-cell data that is not present in the reference data: it incorporates the spike-and-slab prior to capture the direction of variation that separates a small subset of cells from the other cells, since there may be rare cell types that are not captured in the reference data. The spike-and-slab prior also facilitates RA3 to distinguish biological variation from technical variation, assuming that the direction of variation that separates a small subset of cells more likely represents biological variation. The third component captures the other variations in single-cell data, and it likely represents the technical variation. Other than the three components, RA3 includes another term to model known covariates. The first and second components are used for downstream analyses, and we present results on data visualization, cell clustering, trajectory inference and motif enrichment analysis.

RA3 decomposes variation in single-cell data

We first use a simple example as a proof of concept to demonstrate RA3. We collected human hematopoietic cells with donor label BM0828 from a bone marrow scATAC-seq dataset²⁷ (referred as the “human bone marrow dataset”). To reduce the noise level, we first adopted a feature selection strategy similar to scABC and SCALE (Methods). We

Fig. 1 The reference-guided approach for the analysis of scCAS data. **a** A graphical illustration of the RA3 model. RA3 decomposes the variation in scCAS data into three components: the component that captures the shared biological variation with reference data, the component that captures the unique biological variation in single-cell data, and the component that captures other variations. **b** *t*-SNE visualization of the cells from donor BM0828 using latent features obtained from TF-IDF + PCA. **c** *t*-SNE visualization of the cells from donor BM0828 using latent features obtained from bulk projection. **d** We calculated the residuals after the bulk projection. PCA was performed on the residuals, followed by *t*-SNE visualization. **e** The learned second component with the spike-and-slab prior in RA3. **f** *t*-SNE visualization using the first two components learned by RA3.

74 performed the term frequency-inverse document frequency (TF-IDF) transformation to normalize the scATAC-seq
75 data matrix, implemented principal component analysis (PCA), and then performed *t*-distributed stochastic neighbor

embedding (t -SNE)³² to reduce the dimension to two for visualization. This approach (TF-IDF + PCA) is similar to that in Cusanovich2018^{5,10}, which is among the top three methods suggested in a recent benchmark study⁶. It is hard to separate the majority of the cell types using TF-IDF + PCA, and only CLP and MEP cells are moderately separated from the other cells (Fig. 1b). We then collected a reference data: bulk ATAC-seq samples from four parent nodes in the hematopoietic differentiation tree²⁷, including samples that correspond to HSC, MPP, LMPP and CMP cells after fluorescent activated cell sorting (FACS). The genomic regions in the reference data are matched with that in the single-cell data. We first considered a simple approach, referred as the bulk projection approach, to utilize the information in the reference data: 1) We applied PCA on bulk ATAC-seq data; 2) We used the projection vectors learned from bulk data to project single-cell data after TF-IDF normalization; 3) For visualization, we applied t -SNE on the projected data to further reduce the dimension to two. This simple approach significantly improves the separation of the cell types (Fig. 1c). A method similar to the above approach was proposed in Buenrostro et al.²⁷, but the performance of the method was not as good as our simple approach (Supplementary Fig. 1a). The bulk projection approach implicitly assumes that all the biological variation is shared in single-cell and the reference data, as it only uses the projection vectors learned from the reference data and do not use single-cell data to learn the projection vectors. In this example, using the projection vectors from reference data alone cannot distinguish CLP from the other cells, since the variation of CLP cells is not captured in the reference data. We took residual after the bulk projection, and implemented PCA + t -SNE on the residual matrix: most cells other than MEP and CLP cells are mixed together, which indicates the presence of strong technical variation in the residual matrix (Fig. 1d). This observation suggests that including the direction of variation learned from single-cell data may help to separate CLP cells, but the direction needs to be carefully chosen because of the strong technical variation. A schematic plot for the desired direction of variation to learn from single-cell data is shown in Fig. 1d: it separates a small subset of cells from the other cells, and the technical variation is weaker along that direction.

To overcome the limitation of the bulk projection approach, RA3 not only incorporates prior information from the projection vectors learned from the reference data as the first component in the decomposition of variation, but also incorporates a second component to model the unique biological variation in single-cell data. The key for the second component in RA3 is a spike-and-slab prior^{33,34}, which facilitates the model to detect directions that lead to good separation of a small number of cells from the other cells, and not necessarily directions with the largest variation. Using the direction of the largest variation can be problematic given that the technical variation can be strong. Applying RA3 to the example, the second component with the spike-and-slab prior successfully distinguishes CLP cells (Fig. 1e). We note that the labels for CLP cells are not used in RA3 to detect the direction that separates CLP cells. In this example, RA3 effectively utilizes the prior information in reference data to separate CMP, GMP, HSC/MPP, LMPP, and MEP cells, and it also captured the unique variation in single-cell data, which separates CLP cells from GMP and LMPP cells (Fig. 1f). The loadings of the second component in RA3 provide functional insight on the cell subpopulations. Using the top 1000 peaks with largest magnitude in the loadings of the second component (we focus on peaks with negative loadings as the sign of H_2 for the identified cell subpopulation is mostly negative), we performed GREAT (Genomic Region Enrichment of Annotation Tool³⁵) analysis to identify significant pathways associated with the identified cell subpopulation in the second component (Methods). The top five pathways with smallest p -values from the binomial test are “regulation of lymphocyte activation”, “Fc receptor signaling pathway”, “immune response-regulating cell surface receptor signaling pathway”, “immune response-activating cell surface receptor signaling pathway”, and “positive regulation of lymphocyte activation” (see Supplementary Table 1 for the complete enrichment results). These enriched pathways are consistent with the function of CLP cells: CLP cells serve as the earliest lymphoid progenitor cells and give rise to T-lineage cells, B-lineage cells, and natural killer (NK) cells. To summarize, the second component in RA3 not only identifies the rare cell subpopulation, but also provides functional insight of the identified cell subpopulation.

RA3 builds effective reference from massive bulk data

The previous example of hematopoietic cells utilizes reference data constructed from manually curated bulk samples that have relevant biological context with the single-cell data. The cellular composition in single-cell data is generally unknown. It can be desirable to utilize bulk reference data generated from diverse biological contexts and cell types, such as all the bulk chromatin accessibility data generated in the ENCODE project^{25,26}. The implementation of

RA3 requires matched regions/features in the target single-cell data and the reference data. The web-based tool OPENANNO²⁹ provides a convenient way to construct the reference data: the input for OPENANNO is the peak information in single-cell data, and OPENANNO will calculate the accessibility of these peaks in 871 bulk DNase-seq samples of diverse biological context collected from ENCODE, which can be used as the reference data. Note that this approach requires the BAM files for the reference samples to calculate accessibility, an alternative approach that does not require BAM files will be discussed later. With the peak information in the human bone marrow dataset²⁷, we used OPENANNO to construct the reference data with samples of diverse biological context. The reference data constructed in this way achieved similar performance as the manually curated reference data using only the relevant cell types (Supplementary Fig. 1b).

We next applied RA3 to three other single-cell datasets: 1) A mixture of human GM12878 and HEK293T cells (referred as the “GM/HEK dataset”)⁵; 2) A mixture of human GM12878 and HL-60 cells (referred as the “GM/HL dataset”)⁵; 3) An InSilico mixture of H1, K562, GM12878, TF-1, HL-60 and BJ cells (referred as the “InSilico mixture dataset”)⁴. We first manually constructed the reference using BAM files of bulk DNase-seq samples from the relevant cell lines (Methods). The performance significantly improved over the approach not using reference data (Fig. 2a, Supplementary Fig. 1d-e), which indicates that the abundant BAM files of bulk chromatin accessibility profiles in the literature can be fully used to help the analysis of scCAS data. When we constructed the reference data with OPENANNO using all the 871 bulk samples (Methods), the performance of RA3 is comparable with the implementation using only the relevant cell lines for reference (Fig. 2a, Supplementary Fig. 1d-e).

We also considered an alternative approach to construct the reference using only the peak files (BED files) of the reference samples, to address the situation when BAM files are not available. The reference data can be constructed by counting the number of peaks in reference sample that overlap with every peak in single-cell data (Methods). This approach to construct the reference data also led to satisfactory performance (Fig. 2a, Supplementary Fig. 1c-e), which suggests that we can build useful reference with the peak files of bulk data when BAM files are not available. Therefore, databases that collect more comprehensive biological samples but only provide the peak information for each bulk sample, such as Cistrome DB³⁰, may further facilitate the usage of RA3.

RA3 incorporates pseudo-bulk data as reference

It can be difficult to obtain the bulk samples for certain cell populations, especially for the cells in frozen or fixed tissues, where cell sorting is challenging to implement. The recent efforts of cell atlas consortiums have generated massive amounts of single-cell transcriptome data for whole organisms³⁶⁻⁴⁴, and single-cell chromatin accessibility data is rapidly increasing^{4,5,10,27,28}. We can construct pseudo-bulk reference data by aggregating single cells of the same type/cluster to alleviate the high degree of sparsity in scCAS data. As a proof of concept, we first look at a single-nucleus ATAC-seq dataset generated from mouse forebrain (referred as the “mouse forebrain dataset”)²⁸, where the cell type labels were provided, including astrocyte (AC), three subtypes of excitatory neuron (EX1, EX2, and EX3), two subtypes of inhibitory neuron (IN1 and IN2), microglia (MG) and oligodendrocyte (OC). We randomly split the cells in this dataset into half: half of the cells were aggregated by the cell types to build the pseudo-bulk reference, and the other half of the cells were used as the single-cell data. It is hard to separate the subtypes of excitatory neurons using TF-IDF + PCA (Supplementary Fig. 2a). RA3 using the pseudo-bulk reference successfully identified all the cell types, with moderate separation in the three subtypes of excitatory neurons, EX1, EX2 and EX3 (Fig. 2b). To investigate the influence of incomplete reference data, we left out MG and OC cells in constructing the pseudo-bulk reference data. As expected, the bulk projection approach cannot distinguish MG and OC cells (Fig. 2b). The spike-and-slab prior in RA3 successfully detected the directions that lead to good separation of MG and OC cells (Fig. 2b, Supplementary Fig. 2b), and RA3 led to improved separation of all the cell types (Fig. 2b).

To further demonstrate the advantage of RA3 for the analysis of single-cell epigenetic profiles utilizing existing single-cell data, we collected single cells from a sciATAC-seq dataset of the adult mouse brain (referred as the “MCA mouse brain dataset”)¹⁰. We used the provided cell type labels to evaluate different methods¹⁰. We first look at cells from the mouse prefrontal cortex. It has been suggested that there is heterogeneity within excitatory neurons in mouse prefrontal cortex¹⁰. TF-IDF + PCA can hardly distinguish the subtypes of excitatory neurons and inhibitory neurons (Fig. 2c). We used the complete mouse forebrain dataset²⁸ to construct a pseudo-bulk reference for RA3 (Methods). RA3 achieved much better separation for the subtypes of excitatory neurons (Fig. 2c). RA3 also achieved better

Fig. 2 RA3 incorporates reference data constructed from different sources. **a** *t*-SNE visualizations of the cells in the GM/HEK dataset using latent features obtained from TF-IDF + PCA and from RA3 using reference data constructed from different samples, including BAM files of bulk GM12878 and HEK293T DNase-seq samples, BAM files of all the bulk samples in OPENANNO, and BED files of all the bulk samples in OPENANNO. **b** We split the cells in the mouse forebrain dataset into half: half of the cells were used to construct pseudo-bulk reference, and the other half were treated as single-cell data. *t*-SNE visualization using the latent features learned by RA3 with the complete reference are shown. We also constructed an incomplete pseudo-bulk reference by leaving out MG and OC cells. *t*-SNE visualization using the latent features obtained by bulk projection with incomplete reference is shown. We implemented RA3 with the incomplete reference: the learned second component with spike-and-slab prior and *t*-SNE visualization of the learned latent features are shown. **c** *t*-SNE visualizations of cells in the mouse prefrontal cortex dataset, using the latent features obtained from TF-IDF + PCA and the latent features obtained from RA3 with pseudo-bulk reference constructed from the mouse forebrain dataset. **d** *t*-SNE visualizations of cells in the 10X PBMC dataset using the latent features obtained from RA3 with pseudo-bulk reference constructed from another PBMC dataset. Chromatin accessibility of S100A12 (a marker gene of monocytes) and MS4A1 (a marker gene of B cells) are projected onto the visualizations, respectively.

performance on three other brain tissues in the MCA mouse brain dataset, including cerebellum and two replicates of the whole brain (Supplementary Fig. 2d-f).

Next, we collected peripheral blood mononuclear cells (PBMCs) from 10X Genomics (referred as the “10X PBMC dataset”). The labels of cell type are not provided in this dataset, and 8 cell populations inferred by cell markers are suggested in recent studies^{6,9,45}, including CD34+ cells, natural killer cells, dendritic cells, monocytes, lymphocyte B cells, lymphocyte T cells, and terminally differentiated CD4 and CD8 cells. For the implementation of RA3, we

used a previously published PBMC dataset⁴⁶ to construct the pseudo-bulk reference (Methods). We first performed dimension reduction with RA3 and then performed clustering on the low-dimensional representation. To evaluate the performance, we adopted the Residual Average Gini Index (RAGI) score⁶, which calculates the difference between a) the variability of marker gene accessibility across clusters and b) the variability of housekeeping gene accessibility across clusters, and a larger RAGI score indicates a better separation of the clusters (Methods). RA3 achieved a RAGI score of 0.152, while TF-IDF + PCA achieved a RAGI score of 0.110. Aside from RAGI score, RA3 also led to more compact patterns for marker gene activity (Fig. 2d, Supplementary Fig. 2c), compared with TF-IDF + PCA (Supplementary Fig. 2g). In the previous implementations of RA3, we used the peaks in the target single-cell data to calculate accessibility for the reference data. When we use the peaks in the reference data to calculate accessibility for the target single-cell data, the RAGI score for RA3 is 0.150. This implementation demonstrates that RA3 can be implemented when only the count matrix of the reference data is available.

To summarize, the examples implementing RA3 with pseudo-bulk reference data demonstrate the potential that RA3 can utilize existing scCAS data to facilitate the analysis of newly-generated scCAS data.

Comparison with other methods

RA3 was benchmarked against six baseline methods, including scABC⁸, Cusanovich2018^{5,10}, Scasat¹¹, cisTopic⁹, SCALE¹³, and SnapATAC¹² (Methods). We evaluated the methods by dimension reduction and clustering with the provided cell labels (except for the 10X PBMC dataset, where the cell labels are not provided): We implemented *t*-SNE and uniform manifold approximation and projection (UMAP)⁴⁷ to further reduce the low-dimensional representation provided by each method to two for visualization; To evaluate the clustering performance, we implemented Louvain clustering on the low-dimensional representation provided by each method as suggested by Chen et al.⁶, and we assessed the clustering performance by Adjusted Mutual Information (AMI), Adjusted Rand Index (ARI), Homogeneity score (Homogeneity), and Normalized Mutual Information (NMI). For the 10X PBMC dataset, we used RAGI score to evaluate the clustering performance since the cell labels are not available. We only evaluated the clustering performance for scABC because it does not perform dimension reduction.

Human bone marrow dataset²⁷. We used all the bulk chromatin accessibility data provided by Buenrostro et al.²⁷ as the reference data in RA3. To evaluate different methods, We used both the subset of cells and all the cells in this dataset.

1) CLP, LMPP and MPP cells. These cells come from two donors, BM0828 and BM1077. Scasat and Cusanovich2018 cannot distinguish MPP and LMPP (Fig. 3a). Although SCALE, cisTopic and SnapATAC can separate the three cell types, the effect of the two donors is obvious (Fig. 3a, Supplementary Fig. 3), which leads to poor clustering performance for the three cell types (Fig. 4a, Supplementary Fig. 4). When the variation of different donors is not of interest, it is necessary to incorporate donor labels as covariates⁴⁸, while all the baseline methods do not include such a component. RA3 benefited from using reference data and from including donor labels as the covariates, and outperformed the baseline methods in both visualization (Fig. 3a) and clustering (Fig. 4a, Supplementary Fig. 4).

2) Cells from donor BM0828. Scasat and Cusanovich2018 cannot separate most cell types except CLP and MEP (Fig. 3b). We encountered an error message with output "Nan" when implementing SCALE with the default parameters in this dataset. So we did not include SCALE in this comparison. cisTopic and SnapATAC performed reasonably well except for separating CMP cells (Fig. 3b). RA3 performed the best in both visualization (Fig. 3b) and clustering (Fig. 4a, Supplementary Fig. 4).

3) The full dataset with all the cells and donors. Although cisTopic, Cusanovich2018 and SnapATAC achieved reasonable separation of the cells, their performance was influenced by the effect of donors, especially for HSC, MPP and LMPP cells (Fig. 3c). RA3 achieved the best clustering performance (Fig. 4a, Supplementary Fig. 4).

Three datasets with mixture of cell lines^{4,5}. We constructed the reference in RA3 using bulk DNase-seq samples with relevant biological context for the target single-cell data. Compared to the baseline methods, RA3 also achieved satisfactory performance on the GM/HEK, GM/HL and InSilico mixture datasets (Supplementary Fig. 5).

Mouse forebrain dataset²⁸. We randomly split the cells in this dataset into half: half of the cells were aggregated by the provided cell types to build the pseudo-bulk reference, and the other cells were used as the single-cell data. All the baseline methods can hardly distinguish the three subtypes of excitatory neuron (EX1, EX2 and EX3), while RA3 can separate EX1 and it achieved moderate separation of EX2 and EX3 (Fig. 3d, Fig. 4a, Supplementary Fig. 4).

Fig. 3 Evaluation of the visualization of scCAS data. **a** The dataset of CLP/LMPP/MPP cells. **b** The dataset of donor BM0828. **c** The human bone marrow dataset. **d** The mouse forebrain dataset (half). **e** The mouse forebrain dataset (half) with 25% dropout rate. **f** The dataset of mouse prefrontal cortex. For all the datasets, we obtained the latent features from SCALE, Scasat, cisTopic, Cusanovich2018, SnapATAC and RA3, and then implemented *t*-SNE for visualization.

Fig. 4 Assessment of the clustering results. We implemented Louvain clustering on the low-dimensional representation provided by each method to get the cluster assignments. The cluster assignments for scABC were obtained directly from the model output. **a** The clustering performance using different methods evaluated by Adjusted Mutual Information (AMI). **b** The clustering performance using different methods on the mouse forebrain dataset (half) at different dropout rates evaluated by AMI. **c** The clustering performance using different methods on the 10X PBMC dataset evaluated by RAGI score.

4). To mimic platforms that generate sparser scCAS data, we downsampled the reads in single-cell data (Methods). RA3 consistently outperformed other methods when the dropout rate varies from 5% to 50% (Fig. 4b, Supplementary Fig. 6). Compared with the baseline methods, RA3 was less affected when the dropout rate increases, and RA3 still achieved reasonable separation of the cells when dropout rate = 25% (Fig. 3e). This observation suggests the increased benefit of utilizing reference data when the single-cell dataset has higher degree of sparsity.

MCA mouse brain dataset¹⁰. We used the complete mouse forebrain dataset²⁸ to construct a pseudo-bulk reference. We first look at cells from the mouse prefrontal cortex. Cusanovich2018 cannot distinguish excitatory neurons as previously reported¹⁰ (Fig. 3f). We encountered an error when implementing SCALE on this dataset. Scasat achieved a slight improvement over Cusanovich2018 in separating excitatory neurons. cisTopic, SnapATAC, and RA3 all provided better separation of the inhibitory neurons and the subtypes of excitatory neurons (Fig. 3f). RA3 achieved better separation of astrocytes and oligodendrocytes (Fig. 3f), and the best overall clustering performance (Fig. 4a). RA3 also outperformed the baseline methods on three other datasets, including mouse cerebellum and two samples of the mouse whole brain (Supplementary Fig. 5).

10X PBMC dataset. We used the previously published dataset of PBMC cells⁴⁶ to construct a pseudo-bulk reference based on the provided cell type labels. RA3 consistently outperforms the baseline methods according to the RAGI score (Fig. 4c). In addition, RA3 led to more compact patterns for marker gene activity (Supplementary Fig. 7).

In all the examples, the results of UMAP visualization are similar to that of *t*-SNE visualization (Supplementary Fig. 3).

RA3 facilitates trajectory inference and motif analysis

Other than data visualization and cell clustering, the output of RA3 can also be implemented in downstream analyses including trajectory inference and motif enrichment analysis. We use the cells from donor BM0828 for illustration. We implemented Slingshot⁴⁹ for the trajectory inference, which is suggested in a benchmark study⁵⁰. The inputs

Fig. 5 Trajectory inference and motif enrichment analysis. **a** *t*-SNE visualization of the cells from donor BM0828 and the inferred trajectory with Slingshot using the output of RA3 and Louvain clustering. The hematopoietic differentiation tree²⁷ is shown on the bottomleft. **b** The top 50 most variable TF binding motifs within the cluster-specific peaks for the cells of donor BM0828. The deviations calculated by chromVAR are shown.

for Slingshot include the low-dimensional representation provided by RA3 and the cluster labels obtained from RA3 + Louvain clustering, and the output for Slingshot are the smooth curves representing the estimated cell lineages (Methods). RA3 with Slingshot revealed the differentiation lineage and the inferred trajectory greatly mimics the hematopoietic differentiation tree²⁷ (Fig. 5a). For the motif enrichment analysis, we first obtained the cluster labels by RA3 + Louvain clustering, and then identified the cluster-specific peaks by the hypothesis testing procedure in scABC⁸. We selected the top 1000 peaks with the smallest *p*-values for each cluster, and then applied chromVAR⁷ to infer the enriched transcription factor (TF) binding motifs within these peaks (Methods). Visualization of the top 50 most variable TF binding motifs is shown (Fig. 5b). 39 of the top 50 most variable TF binding motifs have been implicated in hematopoietic development (Supplementary Table 2). Among them, some TF binding motifs are specific to one or two clusters (Fig. 5b), and previous literature further corroborates the role of these TFs/motifs in the same cell types that the clusters represent: EBF1, TCF3, and TCF4 are specific to cluster 7, which corresponds to CLP cells^{27,51,52}; JDP2, CEBPB, and CEBPE are specific to cluster 6, which corresponds to GMP cells^{27,53,54}; and GATA1::TAL1, GATA2, and GATA3 are specific to cluster 4, which corresponds to MEP cells^{55,56}.

Discussion

In this work, we have developed RA3 for the analysis of high-dimensional and sparse single-cell epigenetic data using a reference-guided approach. RA3 simultaneously models the shared biological variation with reference data and the unique variation in single-cell data that identifies distinct cell subpopulations by incorporating the spike-and-slab prior. We have shown that the reference data in RA3 can be constructed from various sources, including bulk DNase-seq, bulk ATAC-seq, and pseudo-bulk chromatin accessibility data, which will facilitate the usage of RA3. We demonstrated that RA3 outperforms baseline methods in effectively extracting latent features of single cells for downstream analyses, including data visualization and cell clustering. RA3 also facilitates trajectory inference and motif enrichment analysis for scCAS data. In addition, RA3 is robust and scalable to datasets generated with different profiling technologies, and of diverse sample sizes and dimensions.

Finally, our modeling framework is flexible and can be extended easily. We can incorporate other types of single-cell profiles as reference data to avoid the situation that the cell types in scCAS data have not been adequately studied from the epigenetic landscape. Besides, we can also extend our approach to non-linear projection of scCAS data by incorporating deep neural networks to capture higher level features.

Methods

The model of RA3

We first apply term frequency-inverse document frequency (TF-IDF) transformation to the read count matrix in scCAS data. Our proposed method RA3 decomposes the variation in the scCAS data after TF-IDF transformation into three components: the shared biological variation with the reference data, the unique biological variation in scCAS data, and other variations. Aside from the three components, RA3 also includes a term for known covariates.

Consider the random vector $\mathbf{y}_j \in \mathbb{R}^{p \times 1}$, which represents the observed p features/regions for the j -th cell among n cells in total. Borrowing the general framework of probabilistic PCA³¹, RA3 models $\mathbf{y}_j$ as the following:

$$\mathbf{y}_j | \boldsymbol{\lambda}_j \sim \mathcal{N}_p(\boldsymbol{\lambda}_j, \sigma^2 \mathbf{I}_p), \quad (1)$$

$$\boldsymbol{\lambda}_j = \beta \mathbf{x}_j + \mathbf{W} \mathbf{h}_j, \boldsymbol{\lambda}_j \in \mathbb{R}^{p \times 1}, \quad (2)$$

where $\mathbf{x}_j \in \mathbb{R}^{q \times 1}$ represents q known covariates, including the intercept and other variables, such as the labels of donors or batches, and $\beta \in \mathbb{R}^{p \times q}$ represents the unknown coefficients for $\mathbf{x}_j$. The variables $\mathbf{W} \in \mathbb{R}^{p \times K}$ and $\mathbf{h}_j \in \mathbb{R}^{K \times 1}$ are latent variables: the columns in $\mathbf{W}$ have similar interpretation as the projection vectors in PCA, and $\mathbf{h}_j$ can be interpreted as the low-dimensional representation of $\mathbf{y}_j$. We further decomposes the term $\mathbf{W} \mathbf{h}_j$ into three components:

$$\mathbf{W} \mathbf{h}_j = \mathbf{W}_1 \mathbf{h}_{j_1} + \mathbf{W}_2 \mathbf{h}_{j_2} + \mathbf{W}_3 \mathbf{h}_{j_3}, \quad (3)$$

where the dimensions are $\mathbf{W}_1 \in \mathbb{R}^{p \times K_1}$, $\mathbf{W}_2 \in \mathbb{R}^{p \times K_2}$, $\mathbf{W}_3 \in \mathbb{R}^{p \times K_3}$, $\mathbf{h}_{j_1} \in \mathbb{R}^{K_1 \times 1}$, $\mathbf{h}_{j_2} \in \mathbb{R}^{K_2 \times 1}$, $\mathbf{h}_{j_3} \in \mathbb{R}^{K_3 \times 1}$. So we have $\mathbf{W} = [\mathbf{W}_1 \quad \mathbf{W}_2 \quad \mathbf{W}_3]$ and $\mathbf{h}_j = [\mathbf{h}_{j_1}^T \quad \mathbf{h}_{j_2}^T \quad \mathbf{h}_{j_3}^T]^T = [\mathbf{h}_{j_1}^T \quad \mathbf{h}_j^{*T}]^T$ with $\mathbf{h}_j^* \triangleq [\mathbf{h}_{j_2}^T \quad \mathbf{h}_{j_3}^T]^T$.

We set $\mathbf{W}_1$ to be equal to the K_1 projection vectors learned from PCA on the reference data, so the first component $\mathbf{W}_1 \mathbf{h}_{j_1}$ utilizes prior information from the reference data and capture the shared biological variation among scCAS data and the reference data.

For the columns in $\mathbf{W}_2$ and $\mathbf{W}_3$, the prior specification is similar to that in Bayesian PCA⁵⁷:

$$\mathbf{w}_k \sim \mathcal{N}_p(0, \alpha_k^{-1} \mathbf{I}_p), \text{ for } k = K_1 + 1, \dots, K,$$

where hyper-parameters $\alpha = \{\alpha_{K_1+1}, \dots, \alpha_K\}$ are precision parameters that controls the inverse variance of the corresponding $\mathbf{w}_k$, for $k = K_1 + 1, \dots, K$.

Let $\boldsymbol{\gamma}_j = [\gamma_{1j}, \dots, \gamma_{Kj}]^T \in \mathbb{R}^{K \times 1}$ denote a K -dimensional binary latent vector for cell j . We assume the following priors for $\boldsymbol{\gamma}_j$ and $\mathbf{h}_j$:

$$\begin{cases} \gamma_{kj} = 1, h_{kj} | \gamma_{kj} \sim \gamma_{kj} \mathcal{N}(0, 1), & k = 1, \dots, K_1, K_1 + K_2 + 1, \dots, K \\ \gamma_{kj} \sim \text{Bernoulli}(\theta), h_{kj} | \gamma_{kj} \sim (1 - \gamma_{kj}) \mathcal{N}(0, \tau_0^2) + \gamma_{kj} \mathcal{N}(0, \tau_1^2), & k = K_1 + 1, \dots, K_1 + K_2 \end{cases}$$

Note that for the entries in $\mathbf{h}_{j_1}$ and $\mathbf{h}_{j_3}$, the prior specification is the same as that in Bayesian PCA, while for the entries in $\mathbf{h}_{j_2}$, we assume the spike-and-slab^{33,34} prior with pre-specified parameters $\tau_0 < 1$ and $\tau_1 > 1$. The spike-and-slab prior is the key that encourages the model to find directions in $\mathbf{W}_2$ that separate rare and distinct cell types from the other cells, and it will help us to distinguish biological variation from technical variation. Therefore the second component $\mathbf{W}_2 \mathbf{h}_{j_2}$ captures the variation unique in scCAS data that separates distinct and rare cell types from the other cells. We note that the sparsity specification on $\mathbf{h}_j$ is different from that in sparse PCA⁵⁸, where $\mathbf{W}$ is assumed to be sparse for variable selection. The third component $\mathbf{W}_3 \mathbf{h}_{j_3}$ models other variations such as technical variation. We believe that $\mathbf{h}_{j_1}$ and $\mathbf{h}_{j_2}$ more likely represent the biological variation, and thus we use the learned $\mathbf{h}_{j_1}$ and $\mathbf{h}_{j_2}$ as the low-dimensional representation for cell j . The parameter θ is pre-specified to determine the size of the rare cell types. In practice, we set the parameters $\tau_0 = 0.9$, $\tau_1 = 5$, and $\theta = 0.1$. We set K_1 equal to the number of reference bulk/pseudo-bulk samples, as we used all the principal components (PCs) in the reference data. When more than 30 bulk samples are used to construct the reference, we retained the loading matrix of the first 30 PCs learned from the reference data when implementing RA3. We set both K_2 and K_3 equal to 5. The performance of RA3 is robust to the number of K_2 (Supplementary Fig. 8).

Model fitting and parameter estimation

Given the observed scCAS matrix $\mathbf{Y} \in \mathbb{R}^{p \times n}$ after TF-IDF transformation and the matrix of covariates $\mathbf{X} \in \mathbb{R}^{q \times n}$, we treat the latent variable matrix $\mathbf{H} = [\mathbf{h}_1, \dots, \mathbf{h}_n] \in \mathbb{R}^{K \times n}$ as missing data, and use the expectation-maximization (EM) algorithm to estimate the model parameters $\Theta = (\Gamma, \mathbf{A}, \mathbf{W}^*, \beta, \sigma)$, where $\Gamma = [\gamma_1, \dots, \gamma_n] \in \mathbb{R}^{K \times n}$, $\mathbf{A} = \text{diag}(\alpha_{K_1+1}, \dots, \alpha_K) \in \mathbb{R}^{(K_2+K_3) \times (K_2+K_3)}$ and $\mathbf{W}^* = [\mathbf{W}_2 \ \mathbf{W}_3] = [\mathbf{w}_{K_1+1}, \dots, \mathbf{w}_K] \in \mathbb{R}^{p \times (K_2+K_3)}$. For simplicity of the derivation that will follow, we also write $\mathbf{H} = [\mathbf{H}_1^T \ \mathbf{H}_2^T \ \mathbf{H}_3^T]^T$, where $\mathbf{H}_1 \in \mathbb{R}^{K_1 \times n}$, $\mathbf{H}_2 \in \mathbb{R}^{K_2 \times n}$, and $\mathbf{H}_3 \in \mathbb{R}^{K_3 \times n}$, corresponding to the decomposition of variation introduced previously, and $\mathbf{h}_j^* = [\mathbf{h}_{j2}^T \ \mathbf{h}_{j3}^T]^T$, $\mathbf{H}^* = [\mathbf{H}_2^T \ \mathbf{H}_3^T]^T$.

In the expectation step (E-step), given the parameters estimated in the previous iteration Θ_{t-1} , the posterior distribution of $\mathbf{h}_j$ is

$$\mathbf{h}_j | \mathbf{y}_j, \mathbf{x}_j, \Theta_{t-1} \sim \mathcal{N}_K(\hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j),$$

where $\hat{\boldsymbol{\Sigma}}_j = (\sigma^{-2} \mathbf{W}^T \mathbf{W} + \mathbf{D}_j^{-1})^{-1}$, $\hat{\boldsymbol{\mu}}_j = \sigma^{-2} (\mathbf{y}_j - \beta \mathbf{x}_j)^T \mathbf{W} \hat{\boldsymbol{\Sigma}}_j$, and

$$\mathbf{D}_j = \begin{bmatrix} \mathbf{I}_{K_1} & & & & 0 \\ & (1 - \gamma_{(K_1+1)j})\tau_0^2 + \gamma_{(K_1+1)j}\tau_1^2 & \dots & & \\ \vdots & \vdots & \ddots & & \vdots \\ & 0 & \dots & (1 - \gamma_{(K_1+K_2)j})\tau_0^2 + \gamma_{(K_1+K_2)j}\tau_1^2 & \\ 0 & & \dots & & \mathbf{I}_{K_3} \end{bmatrix} \in \mathbb{R}^{K \times K}.$$

Based on the posterior distribution, we compute the following expectations:

$$\begin{cases} \mathbb{E}(\mathbf{h}_j) = \hat{\boldsymbol{\mu}}_j^T, \\ \mathbb{E}(\mathbf{h}_j \mathbf{h}_j^T) = \mathbb{E}(\mathbf{h}_j) \mathbb{E}(\mathbf{h}_j)^T + \hat{\boldsymbol{\Sigma}}_j. \end{cases}$$

In the maximization step (M-step), we maximize the expected value of the complete log-likelihood with respect to Θ . The corresponding objective function is

$$\begin{aligned} Q &= \mathbb{E}_{\mathbf{H}|\mathbf{Y}, \Theta} \left(\log P(\mathbf{Y}|\mathbf{W}, \mathbf{H}, \mathbf{X}, \mathbf{A}, \beta) + \log P(\mathbf{W}|\mathbf{A}) + \log P(\mathbf{H}|\Gamma) + \log P(\Gamma) \right) \\ &= -\frac{np}{2} \log \sigma^2 - \frac{1}{2} \mathbb{E}_{\mathbf{H}|\mathbf{Y}, \Theta} \left(\text{trace} [\sigma^{-2} (\mathbf{Y} - \beta \mathbf{X})^T (\mathbf{Y} - \beta \mathbf{X}) - 2\sigma^{-2} (\mathbf{Y} - \beta \mathbf{X})^T \mathbf{W} \mathbf{H} + \sigma^{-2} (\mathbf{W} \mathbf{H})^T \mathbf{W} \mathbf{H}] \right) \\ &\quad + \sum_{k=K_1+1}^K \left[-\frac{1}{2} \log |\alpha_k^{-1} \mathbf{I}_p| - \frac{1}{2} \alpha_k \mathbf{w}_k^T \mathbf{w}_k \right] \\ &\quad - \frac{1}{2} \sum_{j=1}^n \sum_{k=1}^K \left[\mathbb{E}(\mathbf{h}_j \mathbf{h}_j^T)_{kk}^2 \varphi(\gamma_{kj}) - \log \varphi(\gamma_{kj}) \right] \\ &\quad + \sum_{j=1}^n \sum_{k=K_1+1, \dots, K_1+K_2} [\gamma_{kj} \log \theta + (1 - \gamma_{kj}) \log(1 - \theta)], \end{aligned}$$

where $\varphi(\gamma_{kj}) = \begin{cases} (1 - \gamma_{kj}) \tau_0^{-2} + \gamma_{kj} \tau_1^{-2}, & k = K_1 + 1, \dots, K_1 + K_2 \\ 1, & k = 1, \dots, K_1, K_1 + K_2 + 1, \dots, K \end{cases}$, and $\mathbb{E}(\mathbf{h}_j \mathbf{h}_j^T)_{kk}$ is the k -th diagonal element of matrix $\mathbb{E}(\mathbf{h}_j \mathbf{h}_j^T)$.

The optimization problem can be solved by the iterative conditional mode algorithm: in each step, we search for the mode of one group of variables, fixing the other variables. To update $\mathbf{A}$, $\mathbf{W}$, β and σ , we use the following

formulas derived by setting the gradients to zero:

$$\begin{aligned}\alpha_k &= \frac{p}{\mathbf{w}_k^T \mathbf{w}_k}, \text{ for } k = K_1 + 1, \dots, K \\ \mathbf{W}^* &= (\mathbf{Y} - \beta \mathbf{X} - \mathbf{W}_1 \mathbf{H}_1) \mathbf{E}(\mathbf{H}^*)^T \left(\sum_{j=1}^n \mathbf{E}(\mathbf{h}_j \mathbf{h}_j^T) + \sigma^2 \mathbf{A} \right)^{-1}, \\ \beta &= (\mathbf{Y} - \mathbf{W} \mathbf{H}) \mathbf{X}^T (\mathbf{X} \mathbf{X}^T)^{-1}, \\ \sigma^2 &= \frac{\sum_{j=1}^n [(\mathbf{y}_j - \beta \mathbf{x}_j)^T (\mathbf{y}_j - \beta \mathbf{x}_j) - 2(\mathbf{y}_j - \beta \mathbf{x}_j)^T \mathbf{W} \mathbf{E}(\mathbf{h}_j) + \text{trace}(\mathbf{E}(\mathbf{h}_j \mathbf{h}_j^T) \mathbf{W}^T \mathbf{W})]}{pn}.\end{aligned}$$

The variable Γ is updated element-wise by choosing $\gamma_{kj} \in \{0, 1\}$ that maximizes

$$f_{\gamma_{kj}}(\gamma_{kj}) = -\frac{1}{2} E(\mathbf{h}_j \mathbf{h}_j^T)_{kk}^2 \varphi(\gamma_{kj}) + \frac{1}{2} \log \varphi(\gamma_{kj}) + \gamma_{kj} \log \theta + (1 - \gamma_{kj}) \log(1 - \theta),$$

where $\varphi(\gamma_{kj})$ is defined as

$$\begin{cases} \varphi(\gamma_{kj}) = 1, & k = 1, \dots, K_1, K_1 + K_2 + 1, \dots, K \\ \varphi(\gamma_{kj}) = (1 - \gamma_{kj}) \tau_0^{-2} + \gamma_{kj} \tau_1^{-2}, & k = K_1 + 1, \dots, K_1 + K_2. \end{cases}$$

We iteratively repeat the above E-step and M-step until convergence. $\mathbf{E}(\mathbf{H}_1)$ and $\mathbf{E}(\mathbf{H}_2)$ obtained from the last iteration of the EM algorithm are used for downstream analyses, including data visualization, cell clustering, trajectory inference and motif analysis. We implement an optional post-processing step for $\mathbf{E}(\mathbf{H}_2)$: We first implement t -test on each row of $\mathbf{E}(\mathbf{H}_2)$ and discard the rows with p -value > 0.05 ; We then truncate the values in each remaining row of $\mathbf{E}(\mathbf{H}_2)$ by the 5th and 95th percentile. The purpose of the post-processing step is to alleviate the effect of noisy cells with low sequencing depth. In practice, some components in $\mathbf{H}_2$ may capture very few cells (< 5) with extreme values, and these cells are usually noisy cells with low sequencing depth. The optional post-processing step has minimal effect on data visualization but it can improve the clustering result, especially for methods based on Euclidean space, such as k -means clustering.

Initialization

The initial values for $\Theta = (\Gamma, \mathbf{A}, \mathbf{W}^*, \beta, \sigma)$ are needed for the EM algorithm. We consider a warm start for $\mathbf{W}_2$. More specifically, we calculate the residual matrix $r_{\mathbf{W}_1}$, by subtracting the projection onto the column space of $\mathbf{W}_1$ from $\mathbf{Y}$: $r_{\mathbf{W}_1} = \mathbf{Y} - \mathbf{W}_1 \mathbf{W}_1^T \mathbf{Y}$. We then apply PCA to the residual matrix $r_{\mathbf{W}_1}$, and obtain the loading matrix $\mathbf{V}$ and the principal component score matrix. Varimax rotation is then implemented on the principal component score matrix, and rotation matrix $\mathbf{R}$ is obtained. We then initialize $\mathbf{W}_2$ as $\mathbf{V} \mathbf{R}$. The intuition for varimax rotation is that we are looking for directions within the principal subspace that separates a small number of cells from the rest, mimicking the sparsity prior on $\mathbf{W}_2$. We also consider a warm start for σ . To initialize σ , we first obtain the residual by subtracting from $\mathbf{Y}$ the projection onto the column space of $\mathbf{W}_1$ and $\mathbf{V} \mathbf{R}$, and we then initialize σ as the standard deviation of the residual. We initialize $\mathbf{A}$ with the identity matrix, and β with zeros. We initialize γ_{kj} with 0 for $k = K_1 + 1, \dots, K_1 + K_2$ and 1 for other k . $\mathbf{W}_3$ is initialized by a random matrix with elements distributed as standard Gaussian distribution.

Datasets and data processing

Datasets: The human bone marrow dataset contains single-cell chromatin accessibility profiles across 10 populations of immunophenotypically defined human hematopoietic cell types from 7 donors²⁷. The GM/HEK and GM/HL datasets are mixtures of the cell lines GM12878/HEK293T and GM12878/HL-60, correspondingly⁵. The InSilico mixture dataset was constructed by computationally combining scCAS datasets from H1, K562, GM12878, TF-1, HL-60 and BJ cell lines⁴. The mouse forebrain dataset was generated from the forebrain tissue from an 8-week-old

adult mouse (postnatal day 56) by single-nucleus ATAC-seq²⁸. The MCA mouse brain dataset was generated from the prefrontal cortex, cerebellum and two samples from the whole brain of 8-week-old mice using sciATAC-seq, which is a combinatorial indexing assay¹⁰. The 10X PBMC dataset produced by 10X Chromium Single Cell ATAC-seq was generated from peripheral blood mononuclear cells (PBMCs) from a healthy donor.

Data preprocessing: To reduce the noise level, we selected peaks/regions that have at least one read count in at least 3% of the cells in the scCAS count matrix. Similar to Cusanovich et al.¹⁰, we performed TF-IDF transformation to normalize the scCAS count matrix before implementing our model: we first weighted all the regions in individual cells by the “term frequency”, which is the total number of accessible regions in that cell, and then multiplied these weighted matrix by the logarithm of “inverse document frequency”, which is the inverse frequency of each region to be accessible across all cells.

Downsampling procedure: We randomly dropped out the non-zero entries in the data matrix to zero with probability equal to the dropout rate.

Construction of reference

Manually curated bulk reference: For the human bone marrow dataset, we used bulk ATAC-seq data of HSC, MPP, LMPP, CMP, GMP, MEP, Mono, CD4, CD8, NK, NKT, B, CLP, Ery, UNK, pDC and Mega cells provided by Buenroostro et al.²⁷ as the reference data in RA3. To construct the reference for the GM/HEK, GM/HL and InSilico mixture datasets, we first downloaded BAM files of bulk DNase-seq samples with relevant biological context from ENCODE^{25,26}: GM12878 and HEK293T cell lines for the GM/HEK dataset, GM12878 and HL-60 cell lines for the GM/HL dataset, and H1, K562, GM12878, HL-60 and BJ cell lines for the InSilico mixture dataset. We then counted the reads that fall into the regions of scCAS data for the bulk samples to form a count matrix that has same features/regions as the target single-cell data, and we finally obtained the reference data through scaling the count matrix by total mapped reads of each bulk sample.

OPENANNO: The webserver OPENANNO²⁹ provides a convenient and straightforward approach to construct the reference. OPENANNO can annotate chromatin accessibility of arbitrary genomic regions by the normalized number of reads that fall into the regions using BAM files, or the normalized number of peaks that overlap with the regions using BED files. After submitting the peak file in scCAS data to the webserver, OPENANNO computes the accessibility of the single-cell peaks across 199 cell lines, 48 tissues, and 11 systems based on 871 DNase-seq samples from ENCODE. When more than 30 bulk samples are used to construct the reference, we retained the loading matrix of the first 30 principal components learned from the reference data when implementing RA3.

Pseudo-bulk reference by aggregating single cells: Given that it can be difficult to obtain the bulk samples for certain cell populations, we can alternatively construct pseudo-bulk reference data by aggregating single cells of the same type/cluster. To address the bias of the difference in cell type abundance in constructing the reference, we took average (instead of taking sum) for each peak over single cells of the same type/cluster. When the peaks do not match between the target single-cell data and the reference single-cell data, we mapped the reads in the reference single-cell data to the peaks in the target single-cell data to obtain the count matrix for the reference single-cell data. The pseudo-bulk reference constructed in this way requires the BAM files for the reference single-cell data. When the BAM files for the reference single-cell data are not available, we can map the reads in the target single-cell data to the peaks in the reference single-cell data instead. We demonstrated this second way of constructing pseudo-bulk reference through the 10X PBMC dataset, and the performance was similar.

GREAT analysis, trajectory inference and motif analysis

GREAT analysis: We first selected the top 1000 peaks with largest magnitude in the loadings of the second component: we focus on peaks with negative loadings because the sign of H_2 for the identified cell subpopulation (mainly CLP cells) is mostly negative, and the peaks with negative loadings tend to be more accessible in the identified cell subpopulation. We then submitted these peaks to the GREAT³⁵ server with a whole genome background and the default parameter settings to identify significant pathways associated with the peaks and thus obtain functional insight on the identified cell subpopulation. We note that our analysis with GREAT does not require knowing the cell labels.

Trajectory inference: We adopted Slingshot⁴⁹ for trajectory inference, which is the best method of single-cell trajectory inference for tree-shaped trajectories as suggested in a benchmark study⁵⁰. With the default parameter

settings, we used the ‘*getLineages*’ function to learn cluster relationships using the low-dimensional representation provided by RA3 and the cluster labels obtained from RA3 + Louvain clustering, and then constructed smooth curves representing the estimated cell lineages using the ‘*getCurves*’ function. We then used the ‘*embedCurves*’ function to map the curves to the two-dimensional *t*-SNE space for visualization.

Motif enrichment analysis: With the cluster labels obtained from RA3 + Louvain clustering, we identified the cluster-specific peaks by the hypothesis testing procedure in scABC⁸. We selected the top 1000 peaks with smallest *p*-values for each cluster, and then performed TF binding motif enrichment within these peaks using chromVAR⁷.

Baseline methods

We compared the performance of RA3 with six baseline methods (using their default parameters), including scABC⁸, SCALE¹³, Scasat¹¹, cisTopic⁹, Cusanovich2018^{5,10}, and SnapATAC¹². Source code for implementing the baseline methods was obtained from a benchmark study⁶. We set uniform random seeds in all experiments to ensure the reproducibility of results. Similar to SCALE, we applied all the baseline methods to reduce the scATAC-seq data to 10 dimensions for downstream visualization and clustering. To ensure data consistency, we disabled the parameters for filtering peaks and cells in SCALE, and we used the peaks in single-cell data instead of binning the genome into fixed-size windows when implementing SnapATAC.

Visualization and Clustering

Visualization: We first obtained the low-dimensional representation provided by each method. To further reduce the dimension to two, we then implemented *t*-SNE using the R package Rtsne³², and UMAP using the R package UMAP⁴⁷.

Clustering: We first obtained the low-dimensional representation provided by each method except for scABC, and then implemented Louvain clustering⁵⁹⁻⁶¹, which is a community detection-based clustering method. The cluster assignments for scABC were directly obtained from the model output. We implemented binary search to tune the resolution parameter in Louvain clustering to make the number of clusters and the number of cell types as close as possible⁶. We used 8 as the number of expected cell populations for the 10X PBMC dataset.

Metrics for evaluation of clustering results

We evaluated the clustering methods based on adjusted Rand index (ARI), normalized mutual information (NMI), adjusted mutual information (AMI), and homogeneity (H) scores. Suppose T is the known ground-truth labels of cells, P is the predicted clustering assignments, N is the total number of single cells, x_i is the number of cells assigned to the i -th cluster of P , y_j is the number of cells belong to the j -th unique label of T , and n_{ij} is the number of overlapping cells between the i -th cluster and the j -th unique label. Then the ARI score, a corrected-for-chance version of the Rand index, is computed as described in the following formula:

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{x_i}{2} \sum_j \binom{y_j}{2} \right] / \binom{N}{2}}{\frac{1}{2} \left[\sum_i \binom{x_i}{2} + \sum_j \binom{y_j}{2} \right] - \left[\sum_i \binom{x_i}{2} \sum_j \binom{y_j}{2} \right] / \binom{N}{2}}.$$

NMI, a normalization of the mutual information score to scale the results between 0 and 1, is computed as:

$$NMI = \frac{MI(P, T)}{\sqrt{H(P)H(T)}},$$

where H is the entropy, and MI is the mutual entropy.

AMI is calculated as:

$$AMI = \frac{MI(P, T) - E[MI(P, T)]}{avg[H(P), H(T)] - E[MI(P, T)]},$$

where E is the expectation.

To assess if all of the resulting clusters contain only cells which are members of a single class, the homogeneity score is computed as follows:

$$\text{Homo} = 1 - \frac{H(T|P)}{H(T)},$$

where $H(T|P)$ indicates the uncertainty of true labels based on the knowledge of predicted assignments.

We adopted the Residual Average Gini Index (RAGI) score⁶ to evaluate the clustering performance for the 10X PBMC dataset, since no cell type labels are provided in this dataset. The RAGI score calculates the difference between a) the variability of marker gene accessibility across clusters and b) the variability of housekeeping gene accessibility across clusters. More specifically, we first used the Gene Scoring method⁶² to quantify the accessibility of each gene in each cell. We then computed the mean of the accessibility values in all cells for each cluster, and computed the Gini index⁶³ for each marker gene⁴⁵ based on the mean accessibility values. The Gini index for a set of annotated housekeeping genes reported in https://m.tau.ac.il/~elieis/HKG/HK_genes.txt was also calculated to obtain the RAGI score by measuring the difference of mean Gini index for the two groups. A larger RAGI score indicates a better separation of the clusters.

Data availability

The human bone marrow dataset and the corresponding reference ATAC-seq data were retrieved from NCBI Gene Expression Omnibus (GEO) with the accession number GSE96772. Single-cell data of the GM/HEK and GM/HL datasets are available at GEO GSE68103. The InSilico mixture dataset was collected from GEO with the accession number GSE65360. BAM files of the reference data for GM/HEK, GM/HL and InSilico mixture datasets were obtained from ENCODE with the following accession number: ENCFF774HUB, ENCFF775ZJX, and ENCFF783ZLL for the GM/HEK dataset; ENCFF783ZLL, ENCFF775ZJX, ENCFF746RUG, and ENCFF328JJT for the GM/HL dataset; and ENCFF328JJT, ENCFF826DJP, ENCFF923SKV, ENCFF775ZJX, and ENCFF949CIK for the InSilico mixture dataset. The mouse forebrain dataset can be accessed in GEO with the accession number GSE1000333. The MCA mouse brain dataset is available at <http://atlas.gs.washington.edu/mouse-atac>. The 10X PBMC dataset is available at <https://support.10xgenomics.com/single-cell-atac/datasets>. The reference data of fresh PBMCs was downloaded from GEO with the accession number GSE129785. A reporting summary for this Article is available as a Supplementary Information file.

Code availability

The RA3 software with a detailed tutorial is freely available at <https://github.com/cuhklinlab/RA3>.

References

1. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet* **20**, 207–220 (2019).
2. Tsompana, M. & Buck, M. J. Chromatin accessibility: a window into the genome. *Epigenetics Chromatin* **7**, 33 (2014).
3. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nat Methods* **10**, 1213–8 (2013).
4. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–90 (2015).

- 473 5. Cusanovich, D. A. et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular
474 indexing. *Science* **348**, 910–4 (2015).
- 475 6. Chen, H. et al. Assessment of computational methods for the analysis of single-cell atac-seq data. *Genome Biol*
476 **20**, 241 (2019).
- 477 7. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromvar: inferring transcription-factor-associated
478 accessibility from single-cell epigenomic data. *Nature methods* **14**, 975–978 (2017).
- 479 8. Zamanighomi, M. et al. Unsupervised clustering and epigenetic classification of single cells. *Nat Commun* **9**,
480 2410 (2018).
- 481 9. Bravo Gonzalez-Blas, C. et al. cistopic: cis-regulatory topic modeling on single-cell atac-seq data. *Nat Methods*
482 **16**, 397–400 (2019).
- 483 10. Cusanovich, D. A. et al. A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell* **174**, 1309–1324
484 e18 (2018).
- 485 11. Baker, S. M., Rogerson, C., Hayes, A., Sharrocks, A. D. & Rattray, M. Classifying cells with scasat, a single-cell
486 atac-seq analysis tool. *Nucleic acids research* **47**, e10–e10 (2019).
- 487 12. Fang, R. et al. Fast and accurate clustering of single cell epigenomes reveals cis-regulatory elements in rare cell
488 types. *bioRxiv* 615179 (2019).
- 489 13. Xiong, L. et al. Scale method for single-cell atac-seq analysis via latent feature extraction. *Nat Commun* **10**, 4576
490 (2019).
- 491 14. Li, H. et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human
492 colorectal tumors. *Nat Genet* **49**, 708–718 (2017).
- 493 15. Miao, Z. et al. Putative cell type discovery from single-cell gene expression data. *Nat Methods* (2020).
- 494 16. de Kanter, J. K., Lijnzaad, P., Candelli, T., Margaritis, T. & Holstege, F. C. P. Chetah: a selective, hierarchical cell
495 type identification method for single-cell rna sequencing. *Nucleic Acids Res* **47**, e95 (2019).
- 496 17. Aran, D. et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic
497 macrophage. *Nat Immunol* **20**, 163–172 (2019).
- 498 18. Xie, P. et al. Superct: a supervised-learning framework for enhanced characterization of single-cell transcriptomic
499 profiles. *Nucleic Acids Res* **47**, e48 (2019).
- 500 19. Zhang, A. W. et al. Probabilistic cell-type assignment of single-cell rna-seq for tumor microenvironment profiling.
501 *Nat Methods* **16**, 1007–1015 (2019).
- 502 20. Pliner, H. A., Shendure, J. & Trapnell, C. Supervised classification enables rapid annotation of cell atlases. *Nat*
503 *Methods* **16**, 983–986 (2019).
- 504 21. Tan, Y. & Cahan, P. Singlecellnet: A computational tool to classify single cell rna-seq data across platforms and
505 across species. *Cell Syst* **9**, 207–213 e2 (2019).
- 506 22. Ma, F. & Pellegrini, M. Actinn: automated identification of cell types in single cell rna sequencing. *Bioinformatics*
507 **36**, 533–538 (2020).
- 508 23. Lin, Y. et al. scclassify: hierarchical classification of cells. *bioRxiv* 776948 (2019).
- 509 24. Wagner, F. & Yanai, I. Moana: A robust and scalable cell type classification framework for single-cell rna-seq
510 data. *BioRxiv* 456129 (2018).
- 511 25. Consortium, E. P. An integrated encyclopedia of dna elements in the human genome. *Nature* **489**, 57–74 (2012).

- 512 26. Davis, C. A. et al. The encyclopedia of dna elements (encode): data portal update. *Nucleic acids research* **46**,
513 D794–D801 (2018).
- 514 27. Buenrostro, J. D. et al. Integrated single-cell analysis maps the continuous regulatory landscape of human
515 hematopoietic differentiation. *Cell* **173**, 1535–1548 e16 (2018).
- 516 28. Preissl, S. et al. Single-nucleus analysis of accessible chromatin in developing mouse forebrain reveals cell-type-
517 specific transcriptional regulation. *Nat Neurosci* **21**, 432–439 (2018).
- 518 29. Chen, S., Wang, Y. & Jiang, R. Openanno: annotating genomic regions with chromatin accessibility. *bioRxiv*
519 596627 (2019).
- 520 30. Zheng, R. et al. Cistrome data browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic*
521 *Acids Res* **47**, D729–D735 (2019).
- 522 31. Tipping, M. E. & Bishop, C. M. Probabilistic principal component analysis. *Journal of the Royal Statistical*
523 *Society: Series B (Statistical Methodology)* **61**, 611–622 (1999).
- 524 32. Maaten, L. v. d. & Hinton, G. Visualizing data using t-sne. *Journal of machine learning research* **9**, 2579–2605
525 (2008).
- 526 33. George, E. I. & McCulloch, R. E. Variable selection via gibbs sampling. *Journal of the American Statistical*
527 *Association* **88**, 881–889 (1993).
- 528 34. George, E. I. & McCulloch, R. E. Approaches for bayesian variable selection. *Statistica sinica* 339–373 (1997).
- 529 35. McLean, C. Y. et al. Great improves functional interpretation of cis-regulatory regions. *Nature biotechnology* **28**,
530 495 (2010).
- 531 36. Tabula Muris, C. et al. Single-cell transcriptomics of 20 mouse organs creates a tabula muris. *Nature* **562**, 367–372
532 (2018).
- 533 37. Zeisel, A. et al. Molecular architecture of the mouse nervous system. *Cell* **174**, 999–1014 e22 (2018).
- 534 38. Rozenblatt-Rosen, O., Stubbington, M. J. T., Regev, A. & Teichmann, S. A. The human cell atlas: from vision to
535 reality. *Nature* **550**, 451–453 (2017).
- 536 39. Han, X. et al. Mapping the mouse cell atlas by microwell-seq. *Cell* **173**, 1307 (2018).
- 537 40. Zheng, G. X. et al. Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, 14049
538 (2017).
- 539 41. Muraro, M. J. et al. A single-cell transcriptome atlas of the human pancreas. *Cell Syst* **3**, 385–394 e3 (2016).
- 540 42. Davie, K. et al. A single-cell transcriptome atlas of the aging drosophila brain. *Cell* **174**, 982–998 e20 (2018).
- 541 43. Karaïskos, N. et al. A single-cell transcriptome atlas of the mouse glomerulus. *J Am Soc Nephrol* **29**, 2060–2068
542 (2018).
- 543 44. Reid, A. J. et al. Single-cell rna-seq reveals hidden transcriptional variation in malaria parasites. *Elife* **7** (2018).
- 544 45. Pliner, H. A., Shendure, J. & Trapnell, C. Supervised classification enables rapid annotation of cell atlases. *Nat*
545 *Methods* **16**, 983–986 (2019).
- 546 46. Satpathy, A. T. et al. Massively parallel single-cell chromatin landscapes of human immune cell development and
547 intratumoral t cell exhaustion. *Nat Biotechnol* **37**, 925–936 (2019).
- 548 47. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension
549 reduction. *arXiv preprint arXiv:1802.03426* (2018).

- 550 48. Lin, Z. et al. Simultaneous dimension reduction and adjustment for confounding variation. *Proc Natl Acad Sci U*
551 *S A* **113**, 14662–14667 (2016).
- 552 49. Street, K. et al. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics*
553 **19**, 477 (2018).
- 554 50. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods.
555 *Nature biotechnology* **37**, 547–554 (2019).
- 556 51. Zandi, S. et al. Ebf1 is essential for b-lineage priming and establishment of a transcription factor network in
557 common lymphoid progenitors. *The Journal of Immunology* **181**, 3364–3372 (2008).
- 558 52. Somasundaram, R., Prasad, M. A., Ungerback, J. & Sigvardsson, M. Transcription factor networks in b-cell
559 differentiation link development to acute lymphoid leukemia. *Blood, The Journal of the American Society of*
560 *Hematology* **126**, 144–152 (2015).
- 561 53. Ji, H. et al. Comprehensive methylome map of lineage commitment from haematopoietic progenitors. *Nature*
562 **467**, 338–342 (2010).
- 563 54. Satoh, T. et al. Identification of an atypical monocyte and committed progenitor involved in fibrosis. *Nature* **541**,
564 96–101 (2017).
- 565 55. Iwasaki, H. et al. Gata-1 converts lymphoid and myelomonocytic progenitors into the megakaryocyte/erythrocyte
566 lineages. *Immunity* **19**, 451–462 (2003).
- 567 56. Pimkin, M. et al. Divergent functions of hematopoietic transcription factors in lineage priming and differentiation
568 during erythro-megakaryopoiesis. *Genome research* **24**, 1932–1944 (2014).
- 569 57. Bishop, C. M. Bayesian pca. In *Advances in neural information processing systems*, 382–388.
- 570 58. Zou, H., Hastie, T. & Tibshirani, R. Sparse principal component analysis. *Journal of computational and graphical*
571 *statistics* **15**, 265–286 (2006).
- 572 59. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks.
573 *Journal of statistical mechanics: theory and experiment* **2008**, P10008 (2008).
- 574 60. Levine, J. H. et al. Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with
575 prognosis. *Cell* **162**, 184–197 (2015).
- 576 61. Wolf, F. A., Angerer, P. & Theis, F. J. Scanpy: large-scale single-cell gene expression data analysis. *Genome*
577 *biology* **19**, 15 (2018).
- 578 62. Lareau, C. A. et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility.
579 *Nat Biotechnol* **37**, 916–924 (2019).
- 580 63. Gini, C. Concentration and dependency ratios. *Rivista di politica economica* **87**, 769–792 (1997).

581 Acknowledgments

582 This work is supported by the Chinese University of Hong Kong direct grant 2018/2019 No. 4053360, direct grant
583 2019/2020 No. 4053423, Hong Kong Research Grant Council Grant ECS 24301419, the National Key Research and
584 Development Program of China (No. 2018YFC0910404), the National Natural Science Foundation of China (Nos.
585 61873141, 61721003, 61573207), and the Tsinghua-Fuzhou Institute for Data Technology. We thank Kui Hua for
586 helpful suggestions. We thank Lei Xiong for the cell-type labels of the mouse forebrain dataset.

Author contributions

Z.L. and R.J. conceived and supervised the project. Z.L., S.C., and G.Y. designed, implemented, and validated RA3. W.Z. and J.L. helped analyzing the results. S.C., Z.L., R.J., and G.Y. wrote the manuscript with inputs from all the authors.