Single-cell ATAC-seq clustering and differential analysis by convolution-based approach

Li Lin[1*], Liye Zhang[1*]

[1]School of Life Science and Technology, ShanghaiTech University, Shanghai, China.

[*]Correspondence should be addressed to L.L. (linli@shanghaitech.edu.cn) or L.Z. (zhangly@shanghaitech.edu.cn).

Abstract

Single-cell ATAC-seq is a powerful tool to interrogate the epigenetic heterogeneity of cells. Here, we present a novel method to calculate the pairwise similarities between single cells by directly comparing their Tn5 insertion profiles instead of the binary accessibility matrix using a convolution-based approach. We demonstrate that our method retains the biological heterogeneity of single cells and is less affected by undesirable batch effects, which leads to more accurate results on downstream analyses such as dimension reduction and clustering. Based on the similarity matrix learned from epiConv, we develop an algorithm to infer differentially accessible peaks directly from heterogeneous cell population to overcome the limitations of conventional differential analysis through two-group comparisons.

Introduction

The expression of genes is regulated by a series of transcription factors (TFs) that bind to the regulatory elements of the genome. As the accessible chromatin covers more than 90% TF binding regions, many techniques, such as Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq), have been developed to detect the accessible states of chromatin[1, 2]. Recent technical advancements in ATAC-seq have made it possible to profile the chromatin states of single cells at a high-throughput manner[3-5]. However, both data processing and interpretation of single-cell ATAC-seq (scATAC-seq) data is more challenging than single-cell RNA-seq (scRNA-seq) data owing to low DNA copy number and complexity of chromatin states[1].

Up to now, most methods cluster single cells based on a peak by cell matrix (e.g. Buenrostro et al. 2015[6]). Unlike well-annotated RNA transcripts in the genome, the exact locus of regulatory elements is largely uncharacterized and must be learned from the data itself. However, learning cell type specific regulatory elements from cell mixtures is problematic. Given that there are no golden rules to define functional elements across the genome, the strategies to perform such task varied considerably in different studies[6, 7], and its effect on downstream analyses is largely unknown.

Detecting differentially expressed genes (or differentially accessible peaks for ATAC-seq, we call them DE peaks below) is another important task in single cell analysis. In a conventional pipeline, cells are first grouped into several clusters and subsequent differential analysis is performed by comparison between clusters. Thus, the resolution settings (e.g. number of clusters) may have strong effects on the identification of genes or locus accounting for the heterogeneity of cell population. Recently one method incorporated pseudotime as one predictor into the regression model to infer DE peaks, instead of performing two-group comparisons[8]. But it required cells to be properly embedded into one dimensional space (e.g. pseudotime through differentiation process), which greatly limits its application in complex cell population. Moreover, cells still need to be clustered into small groups (50~100 cells). Such processing step overcomes the sparsity of scATAC-seq data but reduces the sample size. In scRNA-seq, an alternative approach is to find highly variable genes instead of differentially expressed genes, which does not require the clustering of cell population to be defined. But this strategy cannot be applied to scATAC-seq as the chromatin state is always binarized. Despite that, several state-of-the-art tools designed for scATAC-seq merge individual peaks into meta features (regulomes, topics, principal components, k-mers, etc.) to overcome the sparsity of data[3, 9, 10]. Subsequent differential analysis is performed on meta features instead of individual peaks. Such strategy may help reveal the epigenetic programs that governs the cell identities but lacks sufficient resolution for the dynamic change of individual peaks.

Here, we introduce a novel tool, named epiConv, for scATAC-seq analysis. EpiConv addresses two important questions in scATAC-seq analysis, cell clustering and differential analysis. Unlike most of existing methods, epiConv learns the similarities (or distances) between single cells from their raw Tn5 insertion profiles by a convolution-based approach, instead of a binary accessibility matrix. We demonstrate that epiConv retains biological heterogeneity of single cells and is less sensitive to unwanted variations derived from multiple batches or sample preparing protocols. Utilizing the similarities learned by epiConv, we also develop an algorithm to infer DE peaks among single cells that can be directly applied to cell mixtures without resolving the intra population structure.

Results

Infer the similarity from Tn5 insertion profiles

First, we give an overview of the algorithm that calculates the similarity between cells from their Tn5 insertion profiles (**Fig. 1**). Given two cells, A with m insertions and B with n insertions in one genomic region, we collapse the insertions into a continuous distribution across the genome by Gaussian smoothing as follows:

$$f_{Ai}(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu_{Ai})^2}{2\sigma^2}\right),\ f_A(x) = \sum_i^m f_{Ai}(x)$$

$$f_{Bj}(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\left(x-\mu_{Bj}\right)^2}{2\sigma^2}\right),\ f_B(x) = \sum_j^n f_{Bj}(x)$$

Where μ_{Ai} is the locus of insertion *i* in cell A, μ_{Bj} is the locus of insertion *j* in cell B, $f_A(x)$ and $f_B(x)$ give the overall chromatin states of cell A and cell B in the given region. The similarity between A and B over the given region (S_{AB}) is calculated by the convolution of $f_A(x)$ and $f_B(x)$ and can be solved analytically as follows:

$$s_{AB} = \int f_A(x) f_B(x) dx = \mathrm{C} \cdot \sum_{i,j} \exp\left(-\frac{\left(\mu_{Ai} - \mu_{Bj}\right)^2}{4\sigma^2}\right)$$

Where C is an σ dependent constant. In this study, parameter σ is set to 100 bp. To save running time, long distance (> 4σ) is treated as infinity. Through weighted aggregation of the similarities from all informative regions across the genome and proper normalization with respect to sequencing depth, we can obtain the normalized similarity score between any two cells. Subsequent analyses such as dimension reduction or clustering can be performed on the similarity matrix. We also develop a simplified version of epiConv (epiConv-simp), which can be applied to binary accessibility matrix like existing methods. The simplified version does not perform as well as the full version but always generates similar results and runs much faster. In the benchmarking below, we show the results from both full and simplified versions. Other details of epiConv are provided in Methods section.

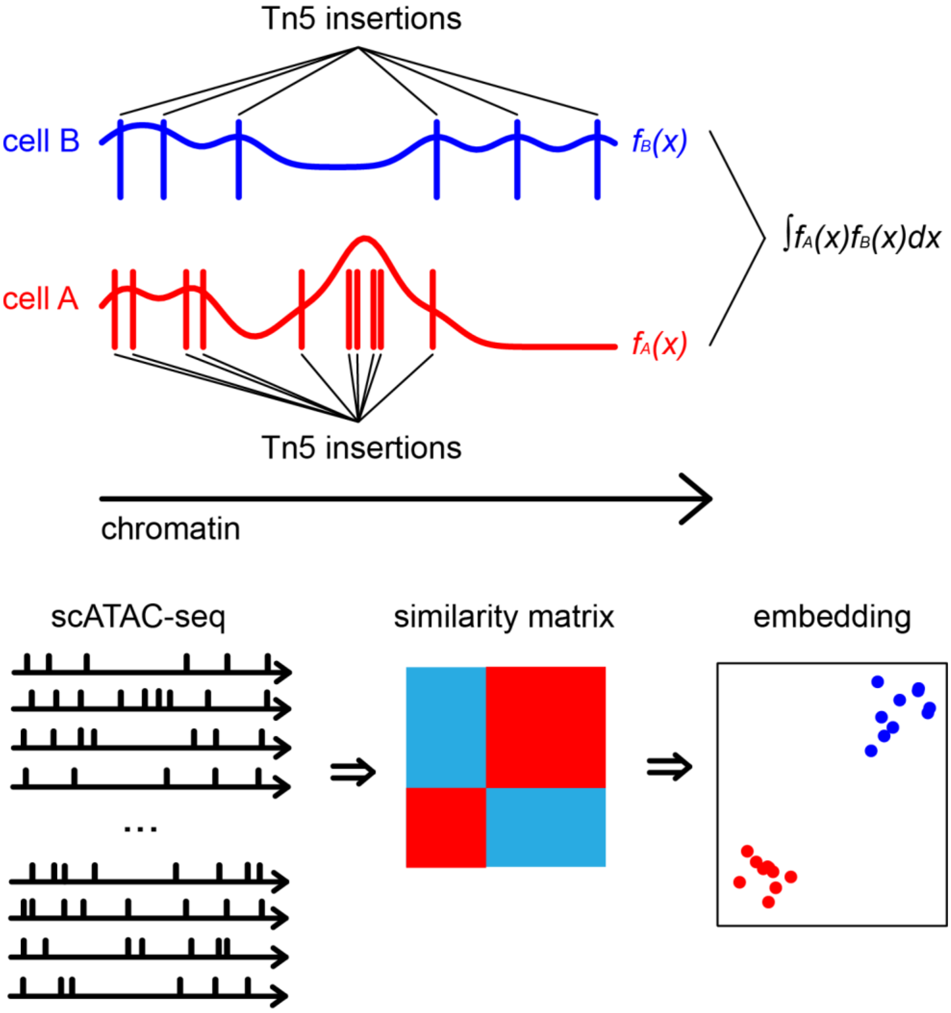


Figure 1. An overview of the epiConv algorithm.

EpiConv outperforms other methods in cell lines data

We evaluated the performance of epiConv on several datasets and compared it with cisTopic[10], one stat-of-the-art method showing better performance than most existing clustering methods and Latent Semantic Indexing (LSI)[3], which had been widely used in many studies. We first applied epiConv to the data from Buenrostro et al. 2015[6]. Specifically, we mixed the data of four cell lines from hematopoietic lineages (K562, GM12878, HL-60 and TF-1) together and tested whether epiConv could cluster single cells correctly based on their biological identities. Given the apparent difference among cell lines, each method performed well in clustering single cells from the same cell line together (**Fig. 2**). However, we found that LSI could not clearly segregate drug-treated and untreated K562 cells. CisTopic segregated treated and untreated K562 cells into two clusters but cells treated by different drugs were still mixed together. Both epiConv-full and epiConv-simp grouped K562 cells treated by different drugs into distinct clusters, yielding the best results. Notably, untreated K562 cells from four replicates were grouped into one cluster without obvious batch effects. Thus, the segregation of cells treated by different drugs were more likely to be attributed to their biological variations rather than batch effects. The simplified version of epiConv performed slightly worse than the full version for K562 cells but was still capable of segregating cells according to their treatment (**Fig. 2b**).

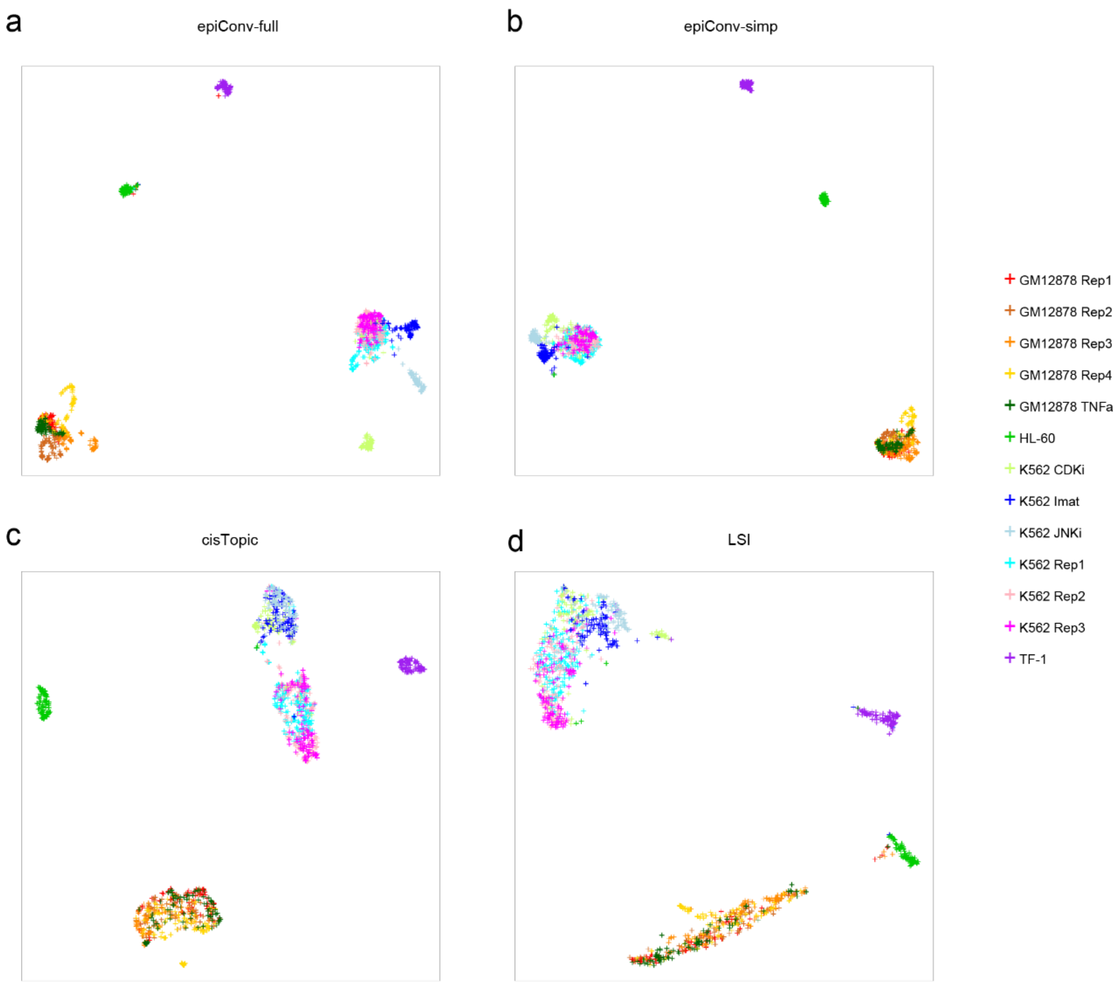


Figure 2. EpiConv performs better than other methods on cell lines data. (**a**) Embedding by epiConv full version. (**b**) Embedding by epiConv simplified version. (**c**) Embedding by cisTopic. (**d**) Embedding by LSI.

EpiConv is less sensitive to batch effects

Next, we applied epiConv to the data generated by droplet-based protocol from Satpathy et al. 2019[4]. The authors reported detectable batch effects from LSI method that confounded downstream analyses. Here we asked whether epiConv could perform better. We tested the performance of epiConv on two datasets, one dataset containing cells from two batches of unsorted peripheral blood mononuclear cells (PBMCs), two batches of sorted CD4+CD45RA+ naïve CD4 T cells and two batches of sorted CD4+CD45RA- memory CD4 T cells (PBMC dataset), and the other dataset containing two batches of sorted CD34+ hematopoietic progenitors (CD34+ dataset).

In PBMC dataset, the majority of cells from two replicates of memory CD4 T cells were clustered into one tightly related group by epiConv and were close to a small fraction of unsorted PBMCs. Two replicates of naive CD4 T cells also showed similar results. Other unsorted PBMCs formed several groups without strong batch effects (**Fig. 3a**). On the contrary, cells were mostly clustered by batches for both cisTopic and LSI (**Fig. 3b,c**). These results showed that epiConv was less sensitive to the technical variations between multiple replicates even without dedicated steps to remove batch effects. To verify whether epiConv clustered single cells based on their biological identities, we marked single cells according to their annotations from Satpathy et al. 2019[4]. The results of epiConv were also largely consistent with the annotations and revealed all major lineages of PBMCs (T cells, NK cells, B cells and Monocytes) and several subpopulation of T cells (**Fig. S1a-c**). In CD34+ dataset, epiConv was still less sensitive to batch effects compared to cisTopic and LSI (**Fig. 3d-f**). We only found obvious batch effect for the HSC-MPP-LMPP cluster but cells from two replicates were still closer to each other than to other cell types (**Fig. 3d**). Based on the annotations from Satpathy et al. 2019[4], the results of epiConv were also consistent with our knowledge on hematopoietic differentiation (**Fig. S1d-f**). However, unlike most methods, epiConv grouped multipotent progenitors (HSC, MPP and LMPP) and other lineage restricted progenitors into several distinct clusters instead of a continuous differentiation trajectory, highlighting the difference of chromatin states between multipotent progenitors and lineage restricted progenitors.

To demonstrate that the power of epiConv was not restricted to specific cell lineages or sample-preparing protocols, we combined scATAC-seq data of adult mouse brain from three experimental protocols, mouse cortex from 10x Genomics, whole mouse brain from droplet single-cell assay for transposase-accessible chromatin using sequencing (dscATAC-seq)[5] and sci-protocols for chromatin accessibility (sci-ATAC-seq)[7]. The dataset contained single cells from 5 batches, one from 10x Genomics, two from dscATAC-seq and two from sci-ATAC-seq. Consistent with previous results, epiConv performed better than cisTopic and LSI in removing batch effects (**Fig. 3g-i**) and agreed with the annotations from Cusanovich et al. 2018[7] and Lareau et al. 2019[5] by clustering cells with the same identity together (**Fig. S1g,j**). CisTopic also largely agreed with the annotations from original articles (**Fig. S1h,k**) while LSI did not agreed with the annotations on excitatory neuron cells (**Fig. S1i,l**). As described from Lareau et al. 2019, the annotations were based on k-mer deviation scores (7-mers) using the chromVAR algorithm but the embedding of

LSI was also consistent with the annotations[9]. Thus, LSI might require a larger sample size to resolve the relationships between highly similar cells. Although we lacked direct evidence to evaluate which method performed best in clustering cells according to their cell identities, the results of epiConv could always be supported by the annotations from original article. Besides that, only epiConv was capable of clustering cells in a batch-independent manner.

Finally, we compared the results between full and simplified versions of epiConv. The results of simplified version were highly consistent with full version and were also less sensitive to batch effects on the three datasets described above (**Fig. S2**). However, for CD34+ cells, epiConv-simp failed to reveal the intra-structure of CLP, Pro-B and Pre-B cluster (compare Fig. S2d with Fig. S1d). In conclusion, the performances of full version and simplified version are similar but sometimes the resolution of simplified version might be slightly lower.

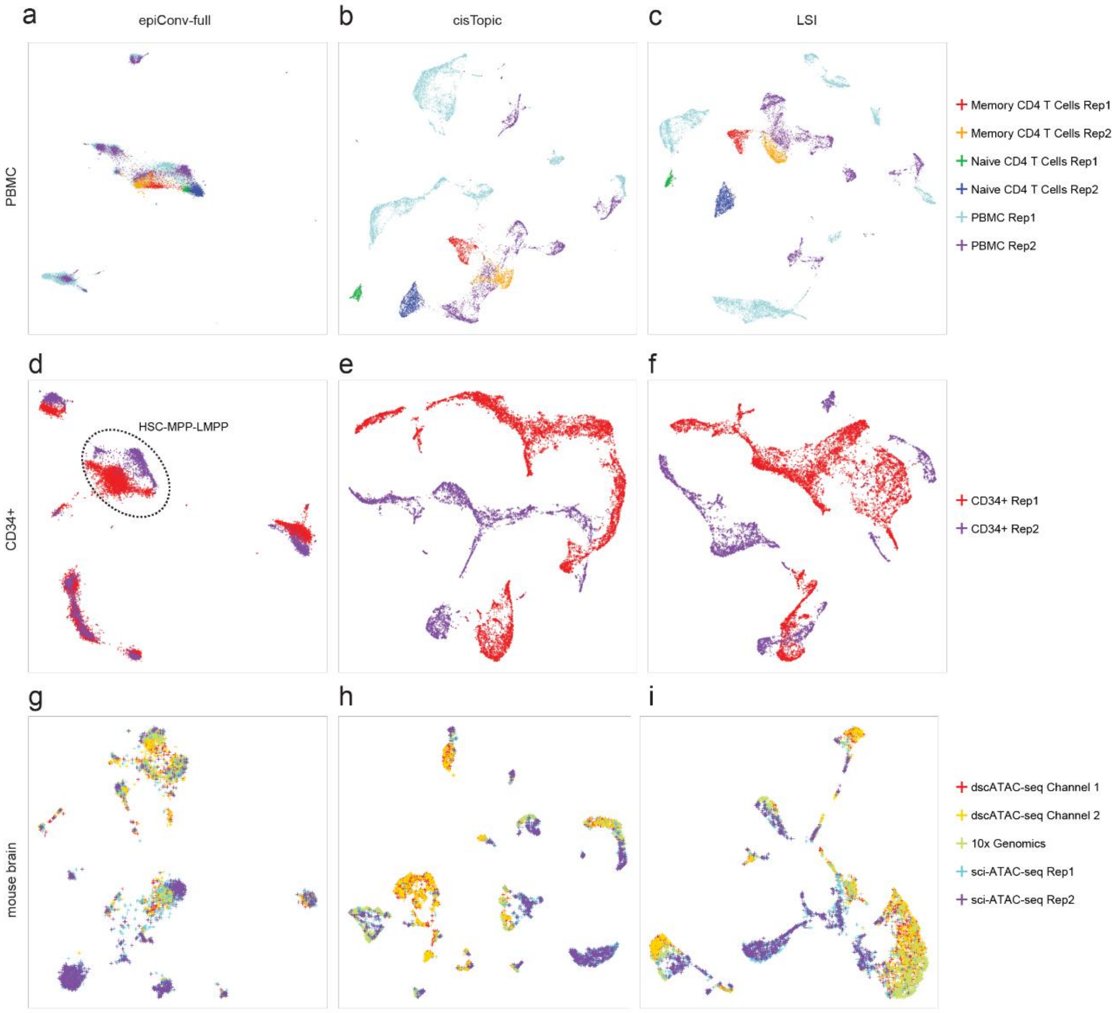


Figure 3. EpiConv is less sensitive to batch effects. (**a-c**) Embedding by epiConv full version, cisTopic and LSI for PBMC dataset. (**d-f**) Embedding by epiConv full version, cisTopic and LSI for

CD34+ dataset. The HSC-MPP-LMPP cluster in (**d**) is circled. (**g-i**) Embedding by epiConv full version, cisTopic and LSI for the integration of mouse brain data from dscATAC-seq, 10x Genomics and sci-ATAC-seq.

EpiConv is scalable with large datasets

As the full version of epiConv do pairwise comparisons between single cells, the step of insertions counting is slower than other methods but can be split into small jobs and run in parallel. Based on our tests, it requires 75 CPU hours for 50 million fragments from 5,000 cells (after removing low quality cells and fragments outside informative regions) and 2,400 CPU hours for 270 million fragments from 20,000 cells. The simplified version runs much faster and can be applied to large datasets. Based on our tests, the simplified version requires 17 hours and 520 GB RAM for the Mouse Cell Atlas dataset[7] (81,173 cells and 436,206 peaks) with single thread, faster than cisTopic (48 hours) but slower than LSI (1 hour). The results of Mouse Cell Atlas dataset by epiConv-simp also largely agreed with the annotations from Cusanovich et al. 2018[7] (**Fig. S3**).

Notably, a large proportion of cells were marked as unknown in the Mouse Cell Atlas dataset (**Fig. 4a-c**). In the results of cisTopic and LSI, these cells formed a large cluster of their own, showed close relationships with several clusters with known identities but did not overlap with them (**Fig. 4a-b**). However, unknown cells did not form a single cluster but were mixed with other known cell types in the results of epiConv-simp (mainly associated with 7 clusters with more than 10% cells marked as unknown, **Fig. 4c**). This might suggest a large improvement of epiConv over cisTopic and LSI. In order to validate our findings, we aggregated the cells with known and unknown cell identities respectively for each cluster. Then we calculated the spearman correlation between the 14 aggregated samples over a set of highly accessible peaks (accessible in at least 1% cells from these 7 clusters). We found that 6 out of 7 unknown samples showed highest correlations with corresponding known samples within the same clusters (**Fig. 4d**), suggesting that epiConv assigned “unknown” cells to correct clusters. The only exception was the cluster that contained collecting duct, distal convoluted tubule and loop of henle. Unknown cells from this cluster did not show higher correlation (> 0.6) with any other samples. We thought that this might be due to the high level of heterogeneity between tubule cells. By these results, we confirmed that epiConv showed significant improvements over current methods on the Mouse

Cell Atlas dataset.

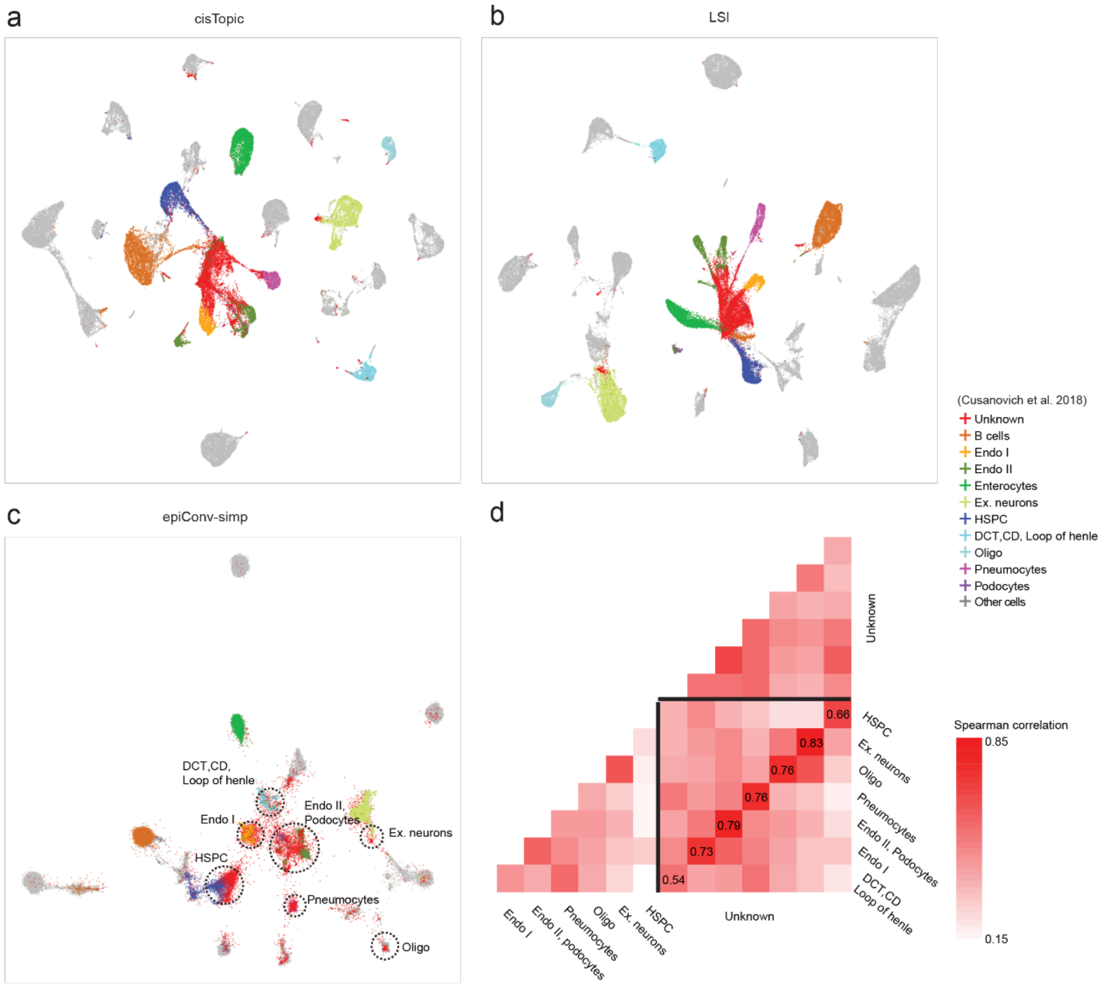


Figure 4. EpiConv reveals the identities of unknown cells in Mouse Cell Atlas dataset. (**a**) Embedding by cisTopic. (**b**) Embedding by LSI. (**c**) Embedding by epiConv-simp. In (**a-c**), unknown cells and cells showing close relationships with them are colored according to the annotations from Cusanovich et al. 2018. Other irrelevant cells are colored in grey. Seven major clusters in (**c**) that contain high proportion of unknown cells are circled. (**d**) Spearman correlations between aggregated samples with known and unknown identities from 7 major clusters marked in (**c**). Unknown samples are sorted in the same order as corresponding known samples belonging to the same cluster. Numbers in the diagonal elements show the correlations between unknown samples and corresponding known samples. Endo I, endothelial I cells; Endo II, endothelial II cells; Ex. neurons, excitatory neurons; HSPC, hematopoietic progenitors; DCT, distal convoluted tubule; CD, collecting duct; Oligo, oligodendrocytes.

EpiConv detects differentially accessible peaks in cell mixtures

In the section below, we aim to develop an algorithm to infer DE peaks directly from cell mixtures. Our algorithm compares the number of accessible cells among each cell's neighbors with the frequency of accessible cells in cell mixture for each peak and turns the binary chromatin states into normalized z-scores, which show the enrichment of accessible cells among neighbors (we call it z-score below). If the number of cells showing high z-scores for one peak exceeds the threshold, we then consider the peak to be differentially accessible. Notably, the way of normalization may strongly affect the results of differential analysis. Although most studies adopt library size normalization (scaling the library size of single cells to be the same), few studies may use other strategies (e.g. scale the number of Tn5 insertions falling in promoters to be the same[5]). In this study, we do not want to address this question but modify our algorithm to be compatible with user-specified scaling factors in normalization. In this study, we try two normalization strategies: 1) set the scaling factors of all cells to be 1 (cell-norm); 2) set the scaling factors equal to the total number of insertions falling into peaks (lib-norm). In cell-norm strategy, the number of neighbors for each cell remains constant while the total library size of neighbors may vary. This strategy reflects the change of raw binary accessibility. In lib-norm strategy, the total library size of neighbors for each cell remains constant while the number of neighbors may vary. This strategy reflects the change of relative abundance of accessibility and can be considered as conventional library size normalization. When the library sizes do not vary between neighbors and non-neighbors for most cells, two strategies should give similar results.

In order to test whether the algorithm could detect DE peaks in cell mixture, we first applied our method to one dataset of myoblast differentiation[8]. We found that although epiConv could reconstruct the differentiation process of myoblasts, where cells were roughly ordered by harvesting times (**Fig. 5a**; the results were similar for cisTopic, LSI and epiConv-simp, see **Fig. S4a-c**), it was difficult to cluster cells. Using our algorithm, we detected 7,219 peaks to be differentially accessible (cell-norm strategy) during the differentiation process. To show the regulation pattern of DE peaks, we plotted heatmap of z-scores, where cells and DE peaks were embedded into one-dimensional (1D) space based on the similarity matrix and the spearman correlation of z-scores between peaks (**Fig. 5b**). The results showed approximately half peaks to be more accessible in the early stage of differentiation and others to be more accessible in the later stage. The dynamic changes of z-scores along differentiation was consistent with merged scATAC-seq profiles by harvesting times, demonstrating the reliability of our algorithm (**Fig. 5c**). The results from cell-norm and lib-norm strategies showed some difference for the intermediate cell types as these cells had lower library sizes but they still agreed with each other on the global regulation patterns of peaks (up- or down-regulated through differentiation**, Fig. S4d,e**). As mentioned above, the binary accessibility profiles agreed better with the z-scores from cell-norm strategy (**Fig. S4f**).

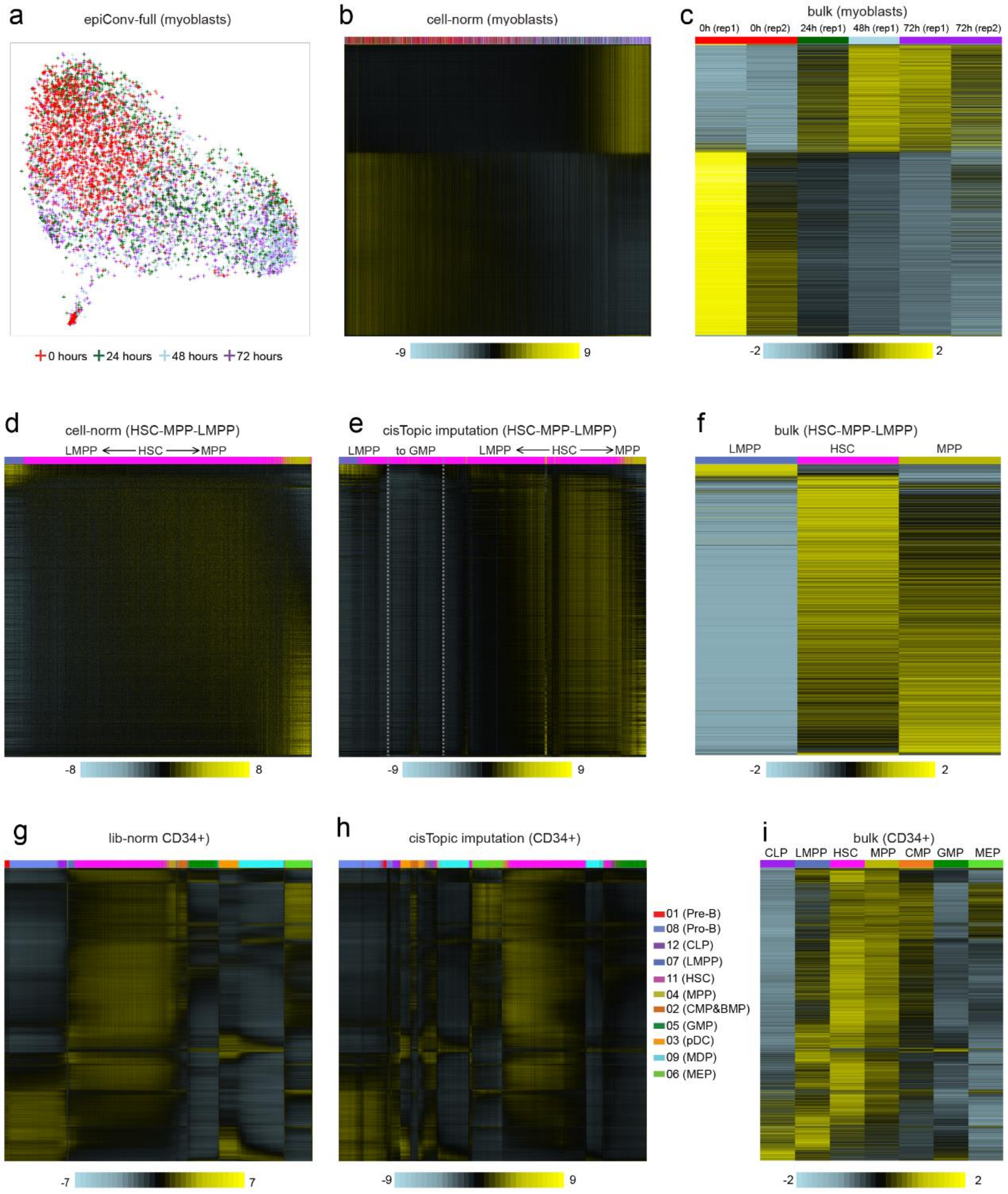


Figure 5. EpiConv detects differentially accessible peaks in cell mixtures. (**a**) Embedding of

myoblast single cells by epiConv-full. (**b**) Accessibility z-scores of myoblast single cells inferred by epiConv. (**c**) Accessibility profiles of aggregated myoblast bulk samples by harvesting times. Cells or aggregated samples in (**b,c**) are colored by harvesting times according to (**a**). (**d**) Accessibility z-scores of HSC-MPP-LMPP single cells inferred by epiConv. (**e**) Accessibility Imputations of HSC-MPP-LMPP single cells inferred by cisTopic. (**f**) Accessibility profiles of HSC, MPP and LMPP bulk samples. (**g**) Accessibility z-scores of CD34+ single cells inferred by epiConv. (**h**) Accessibility Imputations of CD34+ single cells inferred by cisTopic. (**i**) Accessibility profiles of CD34+ bulk samples. Peaks (y-axis) in (**b,c**), (**d-f**) and (**g-i**) are ordered according to 1D embedding by z-scores, respectively. Cells (x-axis) in (**d,g**) are ordered according to 1D embedding by epiConv and cells in (**e,h**) are ordered according to 1D embedding by cisTopic. HSC, hematopoietic stem cells; MPP, multipotent progenitors; LMPP, lymphoid-primed multipotent progenitors; CMP, common myeloid progenitors; BMP, basophil-mast cell progenitors; GMP, granulocyte-macrophage progenitors; MDP, monocyte-dendritic cell progenitors; pDC, plasmacytoid dendritic cells; MEP, megakaryocyte-erythroid progenitors; CLP, common lymphoid progenitors.

Next, we want to test the sensitivity of our algorithm. We first clustered cells by density cluster algorithm[11] and then applied our algorithm to the HSC-MPP-LMPP cluster in the CD34+ dataset described above (**Fig. S5a**). In order to prevent detecting differentially accessible peaks between replicates, we did not perform cross-batch analysis and applied our algorithm to cells in replicate 1. To our knowledge, few tools could detect DE peaks without known cell identities or differentiation trajectory but some methods were capable of revealing the dynamics of accessibility in single-cell resolution by imputation approach (e.g. cisTopic). So, we also included the imputations of cisTopic in our benchmarking (the cells from HSC-MPP-LMPP cluster also formed a cluster in the results of cisTopic, see **Fig. S5b**). Through our algorithm, we detected 1,513 DE peaks (cell-norm strategy) within the HSC-MPP-LMPP cluster and compared the z-scores of them with the imputations of cisTopic and bulk ATAC-seq profiles (**Fig. 5d-f**). The dynamic changes of z-scores were highly consistent with the bulk ATAC-seq profiles of FACS-sorted HSCs, MPPs and LMPPs[12]. As there were no obvious local enrichment of cells with high or low library size in the HSC-MPP-LMPP cluster (**Fig. S5c**), the results from two strategies and binary accessibility profiles did not show conflictions with each other (**Fig. S5d-f**). All DE peaks were properly ordered through the 1D embedding and agreed with their accessibility dynamics in both single-cell and bulk samples, suggesting that the co-accessible pattern between peaks could be revealed by z-scores. (**Fig. 5d-f**). Moreover, our results also showed gradual gain or loss of accessibility in a wide range of peaks through the continuous transition of cell states. Notably, the dynamic changes of accessibility did not completely match the clustering. These results demonstrated that inferring DE peaks directly from cell mixtures helped reveal proper clusters and intermediate cell states in a signature-driven manner instead of statistical ways. The imputation from cisTopic did not show strong confliction with bulk profiles but the pattern of gradual gain or loss of accessibility through x-axis (cells) and y-axis (peaks) was not obvious (**Fig. 5e**). Some cells were highly accessible in LMPP unique peaks but also moderately accessible in HSC or MPP unique peaks, which did not agree with the bulk profiles. Moreover, the chromatin states of some cells were inaccessible for almost all peaks. We found that these cells might be intermediate cell types under the differentiation to GMP, as suggested by cisTopic (**Fig. S5b**). Thus, we concluded that cisTopic lacked sufficient resolution for the dynamic changes of individual peaks and some conflictions between cisTopic and our algorithm could be inherited from the results of clustering.

We also applied our algorithm to all cells in replicate 1 from CD34+ dataset to test the scalability of our algorithm. Z-scores from lib-norm strategy agreed with cisTopic imputations and bulk samples (11,126 DE peaks, **Fig. 5g-i**). Similar with previous results on HSC-MPP-LMPP cluster, we also found a series of peaks that gradually gained or lost accessibility through differentiation (e.g. MDPs to cDCs, **Fig. 5g**). The z-scores did not fully capture the chromatin states of bulk samples for a few peaks (**Fig. 5i**). We found that it was derived from the difference between single-cell and bulk samples (data not shown), probably because there might be some batch effects between them. CisTopic showed similar imputations but was less likely to arrange cells with similar accessibility profiles together when embedding cells to 1D space (**Fig. 5h**). Interestingly, if cells were ordered according to the 1D embedding of epiConv, the cisTopic imputations of single cells were better revealed and almost identical to the results of z-scores (**Fig. S5g**). We suspected that the distance matrix inferred by cisTopic might be nosier than epiConv, which makes cisTopic perform worse than epiConv when embedding cells to 1D space. As the library size of single cells varied considerably between clusters (HSC-MPP-LMPP cluster had a smaller library size, see **Fig. S5c**), cell-norm strategy selected another group of DE peaks (2,358 DE peaks, **Fig. S5h,i**). And as expected, the z-scores from these peaks agreed with binary accessibility profiles. These results demonstrated that different normalization strategies had strong effects on differential analysis when the library size of single cells varied considerably between major clusters. In fact, it was not difficult to find DE peaks between clearly segregated clusters and there were many existing methods that could perform such task. But we demonstrated that our algorithm was flexible enough to detect DE peaks at different scales and compatible with various normalization strategies.

Discussions

In this study, we developed a novel clustering algorithm for scATAC-seq data and compared it with two other methods, cisTopic and LSI. The most significant difference between our algorithm and others is that we calculated the distance between single cells using a convolution-based approach instead of commonly used Euclidean-distance. The Euclidean-distance must be calculated from a matrix and easily suffers from data sparsity, which is the most remarkable feature of scATAC-seq data. However, as researchers have already gained a lot of experience on Euclidean-distance based algorithms through analyzing scRNA-seq data, most methods put their efforts on merging individual peaks into meta features to make Euclidean-distance applicable. Here, we demonstrated several advantages of convolution-based approach (performing better in integrated data from multiple sources and showing higher accuracy in some datasets). However, Euclidean-distance based approaches still have their advantages (e.g. much faster running speed with reasonable accuracy). Importantly, each method benchmarked in this study showed some unique patterns that other methods did not capture (see **Fig. 4** and **Fig. S1**). Given that it is difficult to benchmark the accuracy of different methods in most datasets, we think that it would be better to compare results from multiple methods rather than relying on single method and our method proves to be one of the best candidates for scATAC-seq analyses.

Methods

Informative region calling for epiConv. EpiConv takes processed fragments as input file. To call informative regions for epiConv, we first extended Tn5 insertions from both directions using the pileup command in MACS2[13] (-B --extsize 100). Then, we sorted all sites of the genome by their density in decreasing order and selected regions with cumulative density less than 70% of total insertions. These regions were extended from both directions by 100 bp and merged together if having any overlap. Tn5 insertions overlapping with these informative regions (~70% of total reads) were used for downstream analysis. We used such strategy instead of MACS2 because the proportion of reads used in downstream analyses could be easily specified through the threshold of cumulative density. Moreover, this strategy can always obtain some peaks, while MACS2 may fail when the number of cells is low (e.g. < 200, reported by Satpathy et al. 2019[4]). The threshold of cumulative density is determined by the distribution of insertion length. Based on our preliminary analysis, fragments spanning one or more nucleosomes are nosier than fragments from nucleosome-free regions. Thus, the threshold should be close to the proportion of fragments from nucleosome-free regions. For the myoblast and mouse brain datasets, we set the threshold to 50% as they had higher proportion of fragments spanning one or more nucleosomes (data not shown).

epiConv algorithm. In the results section, we described the algorithm to calculate the similarity between two cells over one region. Here assume that we have N cells and K regions, with the similarities between any two cells *i* and *j* over region *k* (s_{ijk}) being known. First, we weight each region as follows:

$$freq_k = \sqrt{\frac{2}{N(N-1)}\sum_{ij} s_{ijk}}$$

$$w_k = log10(1 + freq_k{}^{-1})$$

The form of weight is similar to that used in LSI but the frequency is replaced by a pseudo-frequency estimated from our convolution-based approach. We use such form of weight to increase the contribution of low-density regions to the similarity score. The similarity between cell *i* and *j* is calculated using a bootstrap approach. Assuming we perform L replicates (L = 30 in this study) and in each replicate we randomly sample some regions (12.5% of total informative regions in this study). The similarity of s_{ij} is calculated as follows:

$$s_{ij} = \frac{\sum_l log10\left(\sum_{k\in rep_l} s_{ijk} \cdot {w_k}^2\right)}{L} - \log10\left(lib_i \cdot lib_j\right)$$

where *lib*$_i$ and *lib*$_j$ is the library size of cell *i* and *j*. We normalize the aggregated similarity by $lib_i \cdot lib_j$ because $\sum_{k\in rep_l} s_{ijk} \cdot {w_k}^2$ can be considered as the sum of $lib_i \cdot lib_j$ random variables with identical distribution given the analytical form of similarity described above. Averaging the similarities from replicates helps reduce the noise compared to simple aggregation of similarities from all regions.

In the simplified version, matrix is first binarized and TF-IDF transformed like LSI[3] (In epiConv-simp, normalization with respect to sequencing depth and peak weighting are identical as LSI). Given TF-IDF matrix *M* and L bootstrap matrices M_{rep_l} by randomly sampling peaks from *M*, the similarity matrix S can be calculated as follows:

$$S = \frac{\sum_l log10\left({M_{rep_l}}^T \cdot M_{rep_l}\right)}{L}$$

Where ${M_{rep_l}}^T \cdot M_{rep_l}$ is the matrix product. Unlike LSI implemented in Cusanovich et al. 2015[3] and Cusanovich et al. 2018[7], we do not filter any peaks. By adopting the formula above, the distance between two insertions $\mu_{Ai} - \mu_{Bj}$ is considered as zero if they are in the same peak or infinite otherwise. Further steps are identical for full and simplified versions.

Next, we denoise the similarities between cells by borrowing the information from their neighbors, which is called similarity blur. Given N cells and their similarity matrix S where s_{ij} is the similarity between cell *i* and *j*, we first transform S to a weight matrix W as follows:

$$w_{ij} = \begin{cases} 10^{s_{ij}} \cdot \log10(lib_i), i \in j's\ neighbors \\ 0, \qquad\qquad\qquad\quad i \notin j's\ neighbors \end{cases}$$

Where *j*'s neighbors are the top 20 cells with highest similarities to *j*. For each column j, we scale the sum of column (excluding the diagonal elements) to a fraction parameter θ between 0 and 1 and the diagonal elements of W are set to $1-\theta$. Then the sum of each column is equal to 1. The matrix W defines how to mix the information from the cell itself and its neighbors, where θ proportion of information comes from its neighbors and the weight of each neighbor is determined by its similarity to cell *j* multiplied by its log10 library size, and $1-\theta$ proportion of information comes from cell *j* itself. In this study we set θ to 0.25. We create a similarity matrix S' where its elements are equal to S except for the diagonal elements (the similarity of each cell to itself, which is not defined for S). The diagonal element s'_{jj} is set to the 99th percentile of column *j*, which can be used to approximate the similarity of cell *j* to itself. The blurred similarity matrix $S_{blurred}$ is calculated by matrix product of S' and W as follows:

$$S_{blurred} = \frac{S' \cdot W + (S' \cdot W)^T}{2}$$

Given $S' \cdot W$ is not a symmetrical matrix, we average $S' \cdot W$ and $(S' \cdot W)^T$ to obtain the similarity matrix. As a proof of the reliability of our algorithm, the upper triangle and lower triangle of $S' \cdot W$ are always close to each other. The distance matrix D is calculated by $\mathrm{D} = -S_{blurred}$, which can be used for downstream analysis such as dimension reduction and clustering.

Pre-processing of ATAC-seq data. We took the processed fragment file or peak by cell matrix as inputs if available. For the unprocessed data from Buenrostro et al. 2015[6] and bulk samples from Corces et al. 2016[12], we aligned raw reads to the hg19 genome using Bowtie2[14] (-X 2000 --no-mixed --no-discordant) and removed reads with mapping quality <10 and duplicates using Picard tools. The start and end of the fragments were adjusted (+5 for forward strand and −4 for reverse strand). We called peaks using MACS2[13] (--nomodel --nolambda --keep-dup all --shift -200 --extsize 400) and generated the count matrix by counting the number of Tn5 insertions falling in peaks.

For the mouse brain dataset, we randomly sampled 2,000 cells from Channel 1 and Channel 2 in Lareau et al. 2019 (dscATAC-seq)[5], 1,000 cells from the mouse cortex data from 10x Genomics and 2,000 cells from two replicates of whole mouse brain in Cusanovich et al. 2018 (sciATAC-seq)[7]. The dataset contains 5,000 cells in total. Data from Cusanovich et al. 2018 were converted from mm9 to mm10 using liftOver[15]. Data from 10x Genomics and Cusanovich et al. 2018 were re-counted against the peaks called by Lareau et al. 2019 for data integration.

For the myoblast dataset, we perform differential analysis on replicate 1 but validate our results by aggregated samples from both replicate 1 and replicate 2. Few outlier cells in replicate 1 that did not cluster together with the majority of cells were excluded in differential analysis.

Implement of cisTopic and LSI. In cisTopic, the number of topics is set to 20, 30, 40 and 50 and automatically decided by cisTopic. For the analysis of cell lines data from Buenrostro et al. 2015[6], in order to explore whether increased number of topics could provide higher resolution for K562 cells, we increase the number of topics from 20 to 100 with a step of 10 but the optimal number of topics is still decided by cisTopic. The imputation from cisTopic is obtained using the function predictiveDistribution(). In LSI, we use the scripts from Cusanovich et al. 2018[7], filter out peaks with frequency < 0.01 and use the top 50 components of singular value decomposition for dimension reduction.

Differential analysis algorithm. The input data is a binarized peak by cell matrix and a distance matrix between cells. Here we use the peak by cell matrix from previous steps. For each single cell, we define *k* cells with highest similarities as its neighbors (including itself). Then for each peak, we test whether it is more likely to be accessible in the cell's neighbors. This problem can be resolved using hypergeometric test, with cells accessible as black balls, cells inaccessible as white balls. The sampling times ($\hat{k}$, the adjusted number of neighbors) is calculated by the total scaling factor of all neighbors divided by the average scaling factors of all cells. By such definition, $\hat{k}$ remains constant ($\hat{k} = k$) in cell-norm strategy while the total library size of $\hat{k}$ neighbors (average library size multiplied by $\hat{k}$) remains constant in lib-norm strategy. The z-scores are calculated by the number of cells accessible among neighbors and z-normalized by corresponding mean and variance of the null distribution.

In differential analyses in this study, the number of neighbors *k* is set to 5% of total cells. The number of neighbors *k* defines the size of potential clusters, which serves similar function as the number of clusters in conventional pipeline. However, the results demonstrated that our algorithm with fixed *k* could still detect DE peaks in clusters with a wide range of size. Here, *k* is set to 5% in order to make our algorithm more sensitive to DE peaks of small clusters. After obtaining the z-scores, we select peaks with z-score > 2 in at least 10% cells as DE peaks. For all cells from replicate 1 of CD34 dataset, we select peaks with z-score > 2 in at least 30% cells as we only want to detect DE peaks between major clusters and the criterion of 10% cells suggested most peaks to be differentially accessible, which was reasonable but not desired. All DE peaks are selected by z-scores from cell-norm strategy except for the CD34+ cells. As the results from two normalization strategies differs from each other for the CD34+ cells, we selected DE peaks based on z-scores from cell-norm and lib-norm strategies, respectively.

In fact, it is not straightforward to choose a proper threshold for z-score. We find that peaks that do not satisfy the threshold described above may also show weak DE pattern. Here, we use the threshold of 10% cells with z-score >2 because selected peaks can be easily validated by bulk samples. For general purpose, users can set the threshold manually to obtain appropriate number of DE peaks.

Dimension reduction. We perform dimension reduction of single cells using the uniform manifold projection (UMAP) algorithm[16] by feeding umap with the distance matrix learned by epiConv, cisTopic and LSI using default settings. The number of reduced components was set to 1 for heatmaps and 2 for scatterplot of cells. We also embed DE peaks into 1D space by feeding umap with the distance matrix that is calculated by one minus spearman correlation of z-scores between peaks.

Density clustering. We use the density clustering algorithm[11] in R package densityClust to cluster single cells for CD34+ single cells. The thresholds of ρ and δ are manually adjusted to match the annotations from Satpathy et al. 2019[5]. As differential analysis does not rely on the results of clustering, the thresholds of ρ and δ won't affect downstream analyses.

Bulk sample processing. For bulk samples of hematopoietic cells from Corces et al. 2016[12], we count the Tn5 insertions against the peaks called from Satpathy et al. 2019[5], normalize the counts by library size and average the normalized counts across all replicates for each cell type. For the myoblast dataset, we de-multiplex the reads, count the Tn5 insertions and normalize the counts by harvesting times.

Data availability. The cell lines data of Buenrostro et al. 2015[6] is obtained from Gene Expression Omnibus (GEO) accession GSE65360. The data of Satpathy et al. 2019[4] is obtained from GEO accession GSE129785. The data of Lareau et al. 2019[5] is obtained from GEO accession GSE123581. The data of Cusanovich et al. 2018[7] is obtained from Mouse Cell Atlas (http://atlas.gs.washington.edu/mouse-atac/). The data of adult mouse cortex is obtained from 10X Genomics website (https://support.10xgenomics.com/single-cell-atac/datasets/1.2.0/atac_v1_adult_brain_fresh_5k). Myoblasts data[8] is obtained from GEO accession GSE109828. EpiConv is available at Github (https://github.com/LiLin-biosoft/epiConv).

Acknowledgements

This project was funded by the National Key Research and Development Program of China (2018YFC1004602), National Natural Science Foundation of China (NSF 31871332) and a startup fund to L.Z. from ShanghaiTech University. We would like to thank Xiaojing Zhao for testing the scripts. We would like to thank Yingdong Zhang on his technical support on the HPC platform of ShanghaiTech University.

Author contributions

L.L. conceived the study, developed the methods and performed analyses. L.L. and L.Z. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Reference:

1. 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**, 1213-1218 (2013).
2. Klemm, S.L., Shipony, Z. & Greenleaf, W.J. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet* **20**, 207-220 (2019).
3. Cusanovich, D.A. et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910-914 (2015).
4. Satpathy, A.T. et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat Biotechnol* **37**, 925-936 (2019).
5. Lareau, C.A. et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat Biotechnol* **37**, 916-924 (2019).
6. Buenrostro, J.D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486-490 (2015).
7. Cusanovich, D.A. et al. A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell* **174**, 1309-1324 e1318 (2018).
8. Pliner, H.A. et al. Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol Cell* **71**, 858-871 e858 (2018).
9. Schep, A.N., Wu, B., Buenrostro, J.D. & Greenleaf, W.J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat Methods* **14**, 975-978 (2017).
10. Bravo Gonzalez-Blas, C. et al. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat Methods* **16**, 397-400 (2019).
11. Rodriguez, A. & Laio, A. Machine learning. Clustering by fast search and find of density peaks. *Science* **344**, 1492-1496 (2014).
12. Corces, M.R. et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat Genet* **48**, 1193-1203 (2016).
13. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).
14. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**,

357-359 (2012).

15. Kent, W.J. et al. The human genome browser at UCSC. *Genome Res* **12**, 996-1006 (2002).

16. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* **3** (2018).

Supplementary materials

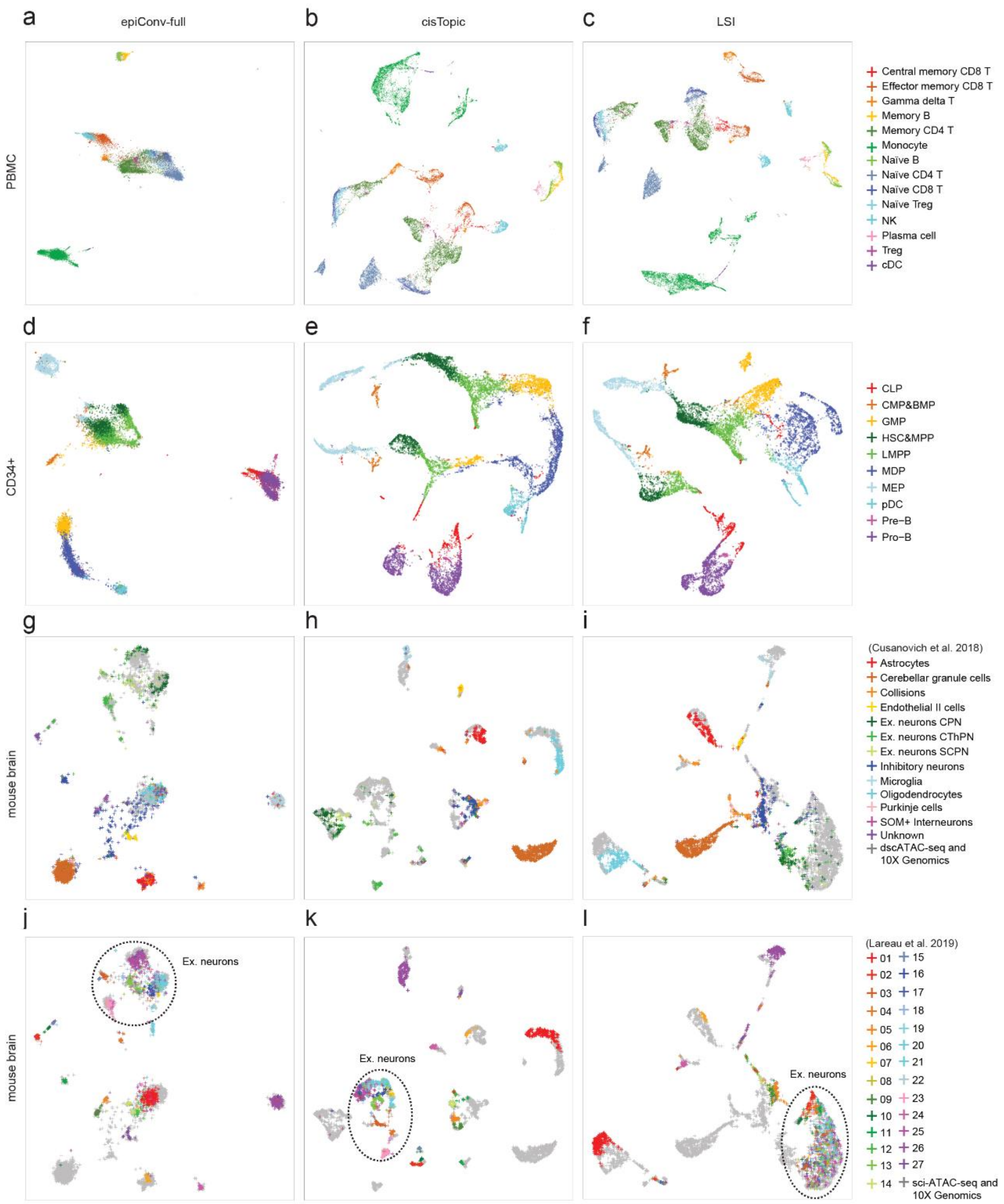


Figure S1. Comparisons of embedding by epiConv, cisTopic and LSI with cell annotations from original articles. (**a-c**) Embedding by epiConv-full, cisTopic and LSI for PBMC dataset, annotated by Satpathy et al. 2019. (**d-f**) Embedding by epiConv-full, cisTopic and LSI for CD34+ dataset, annotated by Satpathy et al. 2019. (**g-l**) Embedding by epiConv-full, cisTopic and LSI for the integration of mouse brain data from dscATAC-seq, 10x Genomics and sci-ATAC-seq, annotated by Cusanovich et al. 2018 (**g-i**) and Lareau et al. 2019 (**j-l**).

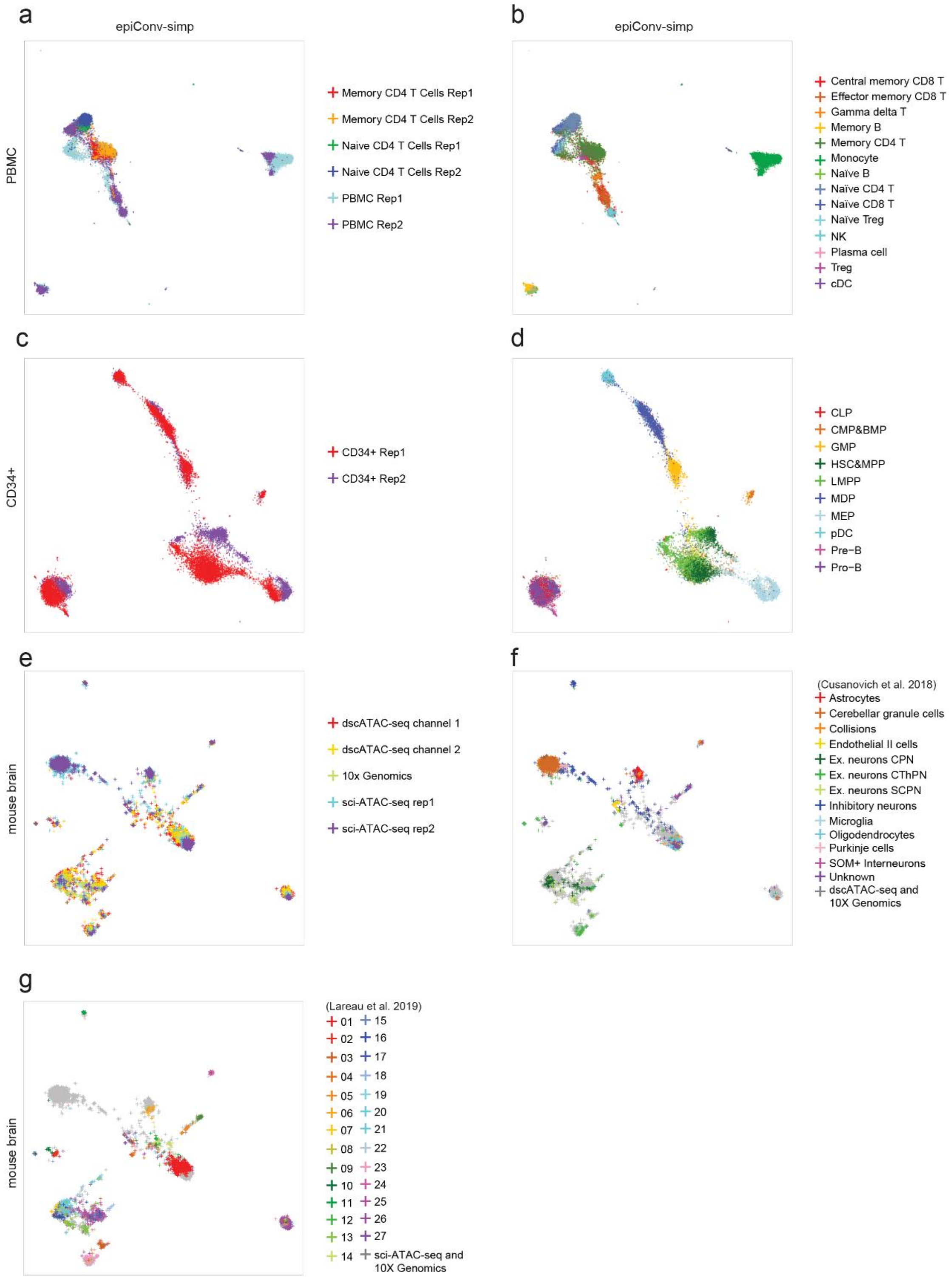
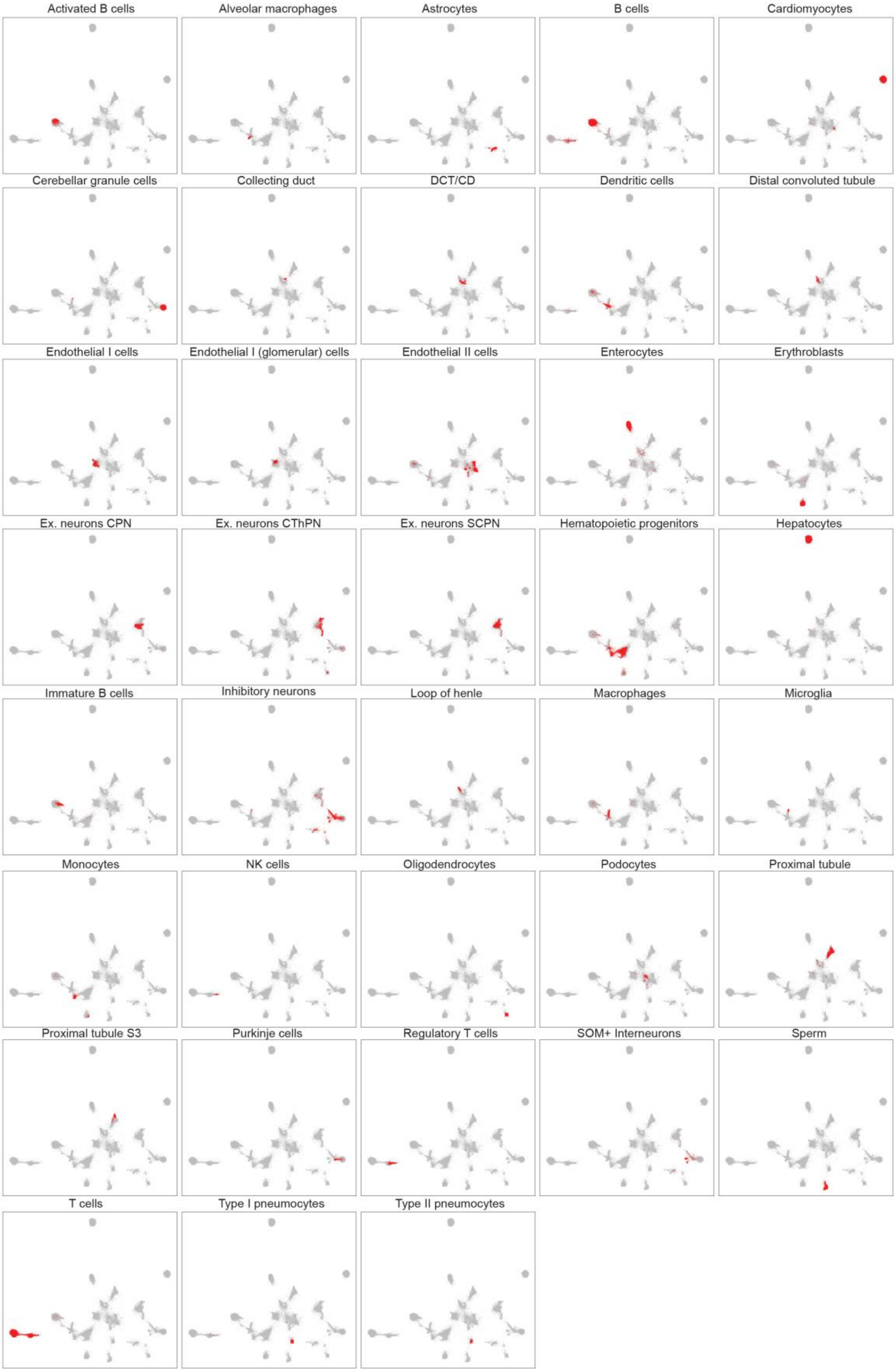


Figure S2. Embeddings of PBMC, CD34+ and mouse brain datasets by epiConv-simp. (**a,b**) Embedding by epiConv-simp for PBMC dataset, colored by batch (**a**) and annotations from Satpathy et al. 2019 (**b**). (**c,d**) Embedding by epiConv-simp for CD34+ dataset, colored by batch (**c**) and annotations from Satpathy et al. 2019 (**d**). (**e-g**) Embedding by epiConv-simp for mouse brain

dataset, colored by batch (**e**), annotations from Cusanovich et al. 2018 (**f**) and annotations from Lareau et al. 2019 (**g**).



543

544 **Figure S3.** Embedding of Mouse Cell Atlas dataset by epiConv-simp. The corresponding cell types

545 are colored in red and other cells are colored in grey.

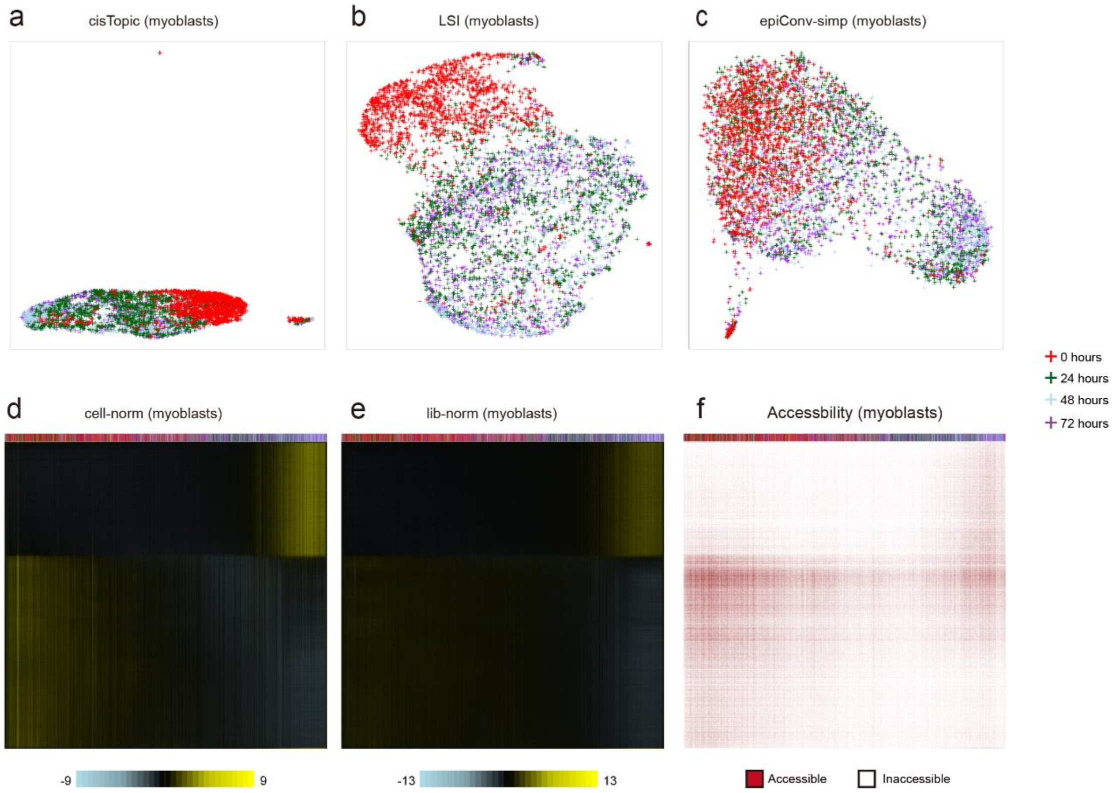


Figure S4. Embedding of myoblasts by different methods and results of differential analysis. (**a**) Embedding by cisTopic. (**b**) Embedding by LSI. (**c**) Embedding by epiConv-simp. (**d**) Accessibility z-scores by cell-norm strategy, identical as **Fig. 4b**. (**e**) Accessibility z-scores by lib-norm strategy. (**f**) Binary accessibility profiles.

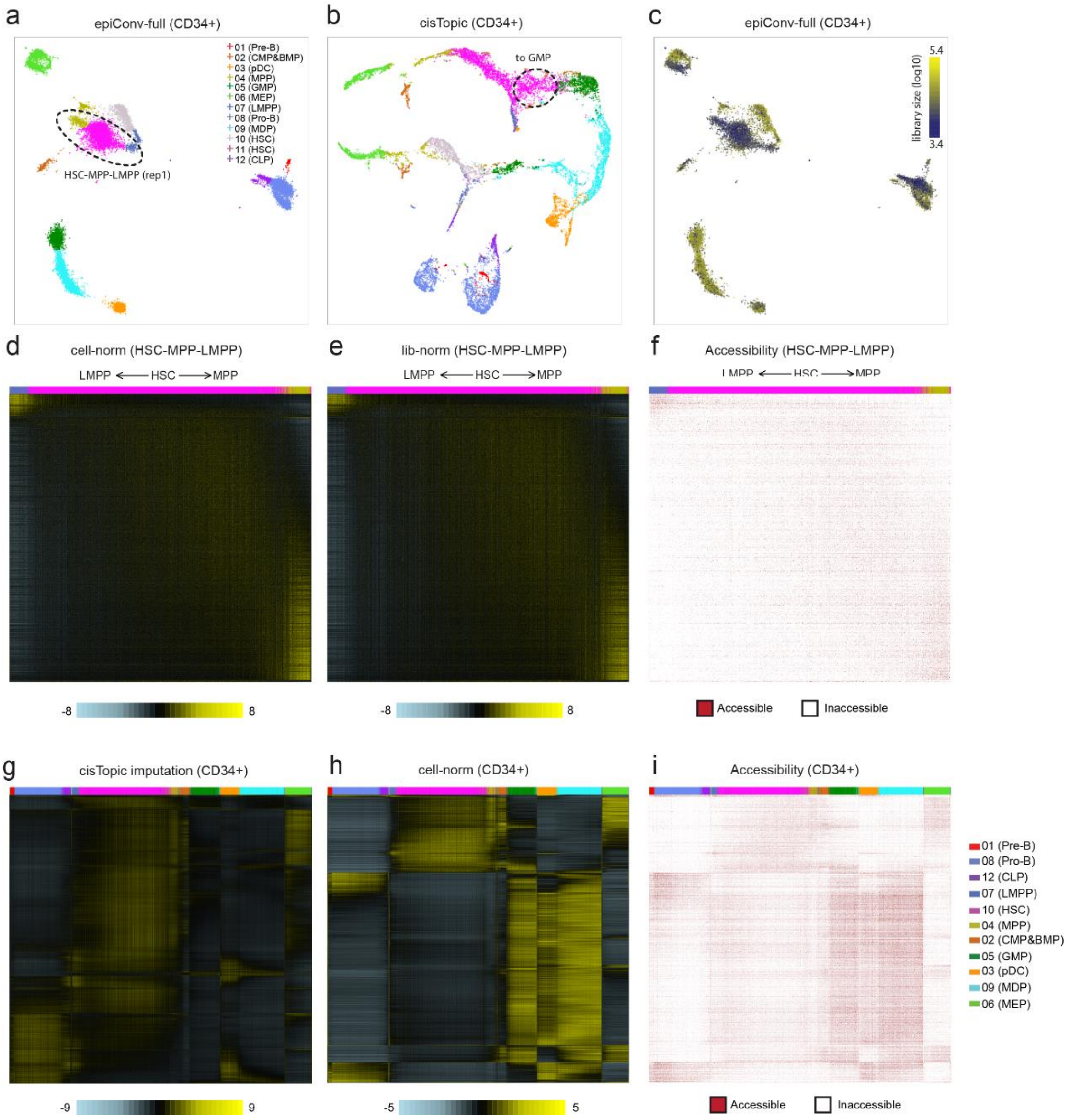


Figure S5. Density clustering of CD34+ single cells and results of differential analysis. (**a**) Embedding by epiConv-full. Cells are colored by the results of density clustering. HSC-MPP-LMPP cluster examined in differential analysis is circled. (**b**) Embedding by cisTopic. Cells are colored according to (**a**). Cells under the differentiation to GMP that are marked in **Fig. 5e** are circled. (**c**) Embedding by epiConv-full, colored by library size. (**d**) Accessibility z-scores by cell-norm strategy for HSC-MPP-LMPP cluster, identical as **Fig. 4d**. (**e**) Accessibility z-scores by lib-norm strategy for HSC-MPP-LMPP cluster. (**f**) Binary accessibility profiles for HSC-MPP-LMPP cluster. (**g**) Accessibility imputations of HSC-MPP-LMPP single cells inferred by cisTopic, identical as **Fig. 5e** but cells (x-axis) are ordered according to 1D embedding by epiConv. (**h**) Accessibility z-scores by cell-norm strategy for CD34+ single cells. (**i**) Binary accessibility profiles for CD34+ single cells. Peaks (y-axis)

in (**h,i**) are NOT the same as **Fig. 5g-i** and are selected by cell-norm strategy, independently.