ARTICLES
https://doi.org/10.1038/s41592-022-01562-8

nature methods

scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural networks

Han Yuan ✉ and David R. Kelley ✉

Single-cell assay for transposase-accessible chromatin using sequencing (scATAC) shows great promise for studying cellular heterogeneity in epigenetic landscapes, but there remain important challenges in the analysis of scATAC data due to the inherent high dimensionality and sparsity. Here we introduce scBasset, a sequence-based convolutional neural network method to model scATAC data. We show that by leveraging the DNA sequence information underlying accessibility peaks and the expressiveness of a neural network model, scBasset achieves state-of-the-art performance across a variety of tasks on scATAC and single-cell multiome datasets, including cell clustering, scATAC profile denoising, data integration across assays and transcription factor activity inference.

Epigenetic landscapes at a single-cell resolution are revealed by scATAC[1]. The assay has been successfully applied to identify cell types and their specific regulatory elements, reveal cellular heterogeneity, map disease-associated distal elements and reconstruct differentiation trajectories[2–4].

However, there are still substantial challenges in the analysis of scATAC data, due to the inherent high dimensionality of accessible peaks and sparsity of sequencing reads per cell[5,6]. Multiple approaches have been proposed to address these challenges, which can be broadly categorized into two main classes: sequence-free and sequence-dependent methods. Starting from a sparse peak-by-cell matrix generated through aggregation of reads and peak calling, most methods represent these annotated peaks as genomic coordinates and ignore the underlying DNA sequences. Principal component analysis (PCA) and latent semantic indexing perform a linear transformation of the peak-by-cell matrix to project cells to a low-dimensional space[4,7]. SCALE and cisTopic model the data generation using Latent Dirichlet Allocation or a variational autoencoder[5,8]. These sequence-free methods are able to detect biologically meaningful covariance to effectively represent and cluster or classify cells; however, they ignore sequence information and rely on post hoc motif-matching tools to relate accessibility to transcription factors (TFs). In contrast, sequence-dependent methods such as chromVAR and BROCKMAN represent peaks by their TF motif or *k*-mer content and aggregate these features across peaks or other regions of interest to learn cell representations[9,10]. While chromVAR directly associates peaks to TFs, emphasizing interpretability, it tends to perform worse at learning cell representations, potentially due to the loss of information from its simple implicit model relating sequence to accessibility through position weight matrices[6].

We propose a more expressive sequence-dependent model for scATAC based on deep convolutional neural networks (CNNs) applied to DNA sequences. In these models, the initial convolution layer learns TF motifs and other sequence factors. Subsequent layers compute nonlinear combinations of these features, to produce an explicit embedding of the sequence. When trained on multiple tasks, the final linear layer transforms the sequence embedding to predict accessibility for each task (sequencing experiments). Its parameters implicitly embed the multiple tasks based on how they make use of the latent variables in the sequence embedding.

Here, we extend the Basset deep CNN architecture to predict single-cell chromatin accessibility from a DNA sequence. In this arrangement, the multiple tasks represent single cells and the model's final layer learns cell embeddings. We show that these cell embeddings outperform state-of-the-art methods for clustering and cell-state representation in multiome data. By making use of sequence information in a deep learning framework, we also achieve improved scATAC denoising, integration with single-cell RNA-seq (scRNA) and TF activity inference over alternative methods.

Results

scBasset predicts single-cell chromatin accessibility on held-out peaks. scBasset is a deep CNN to predict chromatin accessibility from sequence. CNNs have demonstrated state-of-the-art performance for predicting epigenetic profiles in bulk data and have been successfully used for genetic variant effect prediction and TF motif grammar inference[11–14]. Here, we move the focus away from maximizing accuracy on held-out sequences and view the model as a representation learning machine. When trained on multiple tasks, the final layer of these models involves a sequence embedded by the convolutional layers and a linear transformation to predict the data in each separate task (Fig. 1a). The linear transformation matrix comprises a vector representation of each task (each single cell), which specifies how to make use of each of the sequence-embedding latent variables to predict cell-specific accessibility. In a simple ideal scenario, one can imagine each latent variable representing various regulatory factors such as TF binding or nucleotide composition and the final transformation specifying how much each cell depends on that factor. We propose that these single-cell vectors serve as representations of the cells for downstream tasks such as visualization and clustering.

We recommend that users first apply standard processing techniques, such as the 10x CellRanger scATAC pipeline, to bring the raw data to a peak-by-cell binary count matrix. scBasset takes as input a 1,344-bp DNA sequence from each peak's center and one-hot encodes it as a $4 \times 1{,}344$ matrix. The input DNA sequence goes through eight convolution blocks, where each block is composed of one-dimensional (1D) convolution, batch normalization, max pooling and Gaussian error linear unit (GELU) activation layers. Unlike most previous architectures, we follow these by a

Calico Life Sciences, South San Francisco, CA, USA. ✉e-mail: yuanh@calicolabs.com; drk@calicolabs.com

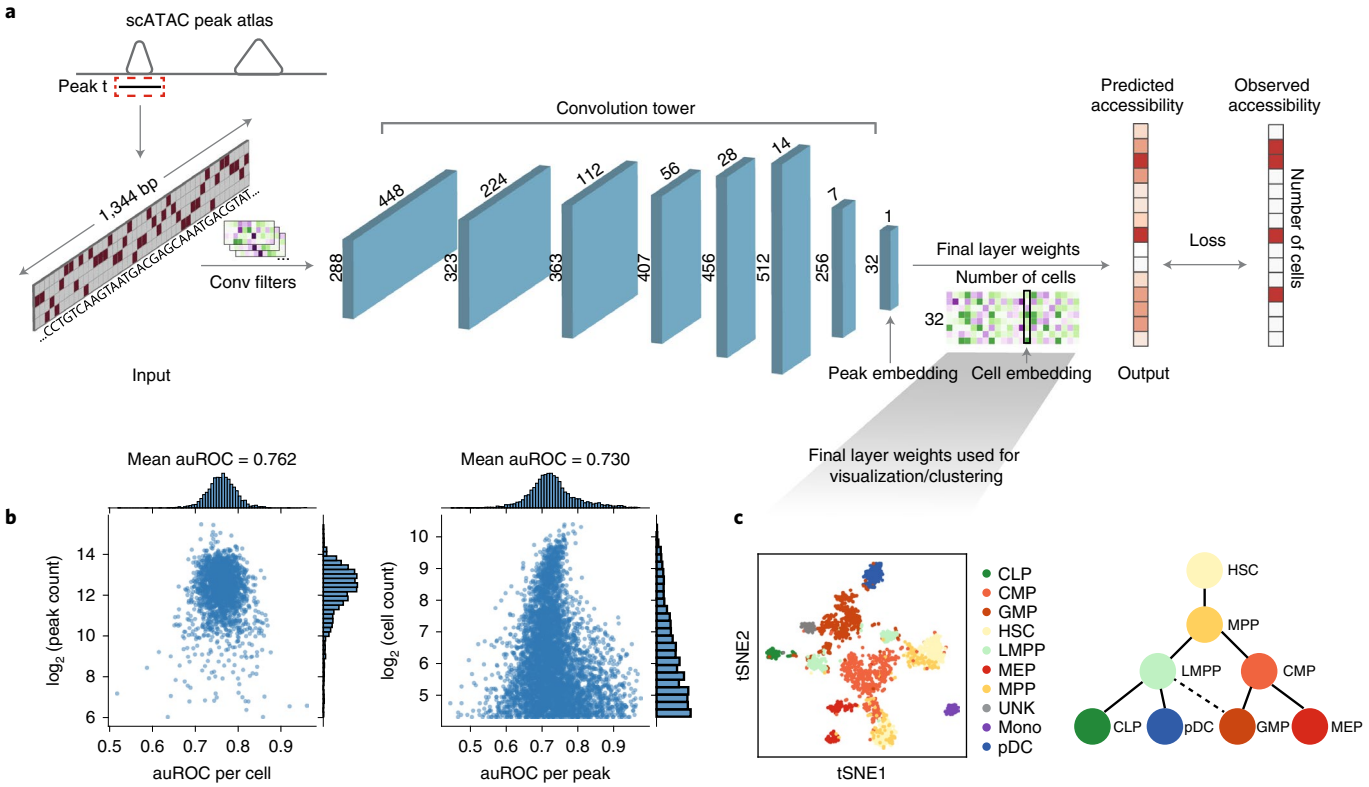


Fig. 1 | scBasset architecture. a, scBasset is a deep CNN to predict single-cell chromatin accessibility from the DNA sequence underlying peak calls. The input to the model is a 1,344-bp DNA sequence from each peak's center and the output is accessibility per cell (corresponding to one row of the peak × cell matrix). Conv, convolution. **b**, scBasset prediction performance on held-out peaks evaluated by auROC per cell (left) and auROC per peak (right) for the Buenrostro2018 dataset. **c**, *t*-SNE visualization of cell embeddings learned by scBasset as the weights of the final dense layer, colored by cell type (left). Hematopoietic stem cell differentiation lineage diagram in the Buenrostro2018 study (right). The cell type labels refer to hematopoietic stem cell (HSC), multipotent progenitor (MPP), lymphoid primed MPP (LMPP), common lymphoid progenitor (CLP), plasmacytoid dendritic cell (pDC), common myeloid progenitor (CMP), granulocyte macrophage progenitor (GMP), megakaryocyte-erythroid progenitor (MEP) cell, monocyte (Mono) and unkonwn (UNK). Source data for this figure are provided.

bottleneck layer (with size fixed to 32 in all analyses) intended to learn a low-dimensional representation of the peak via the layer output and the cells via the parameters of the following layer. Finally, a dense linear transformation connects the bottleneck sequence embeddings to predict binary accessibility in each cell (Fig. 1a). We apply the standard binary cross-entropy loss function and optimize model parameters with stochastic gradient descent (Methods).

To benchmark our approach, we applied scBasset to three public datasets: scATAC of FACS-sorted hematopoietic differentiation (referred to as Buenrostro2018) with 2,000 cells[15], 10x Multiome RNA + ATAC peripheral blood mononuclear cells (PBMCs) with 3,000 cells and 10x Multiome RNA + ATAC mouse brain with 5,000 cells. The first dataset provides ground-truth cell type labels from flow cytometry. We consider the multiome datasets to be a valuable resource to validate scATAC methods as they provide independent measurements of gene expression and chromatin accessibility in the same cells. Although these assays deliver different data, previous work demonstrates that they have substantial mutual information[16,17].

First, we asked how well scBasset can predict accessibility across cells for held-out peak sequences to ensure that the model has learned a meaningful relationship between DNA sequence and accessibility despite the sparse noisy labels. For held-out peaks, we computed the area under the receiver operating characteristic curve (auROC) and area under the precision recall curve (auPR) across peaks for each cell (referred to as 'per cell'). To evaluate cell-type specificity, we also computed auROC and auPR across cells for each peak (referred to as 'per peak') (Supplementary Fig. 1). scBasset achieved compelling accuracy levels that indicate successful learning: auROC of 0.762 per cell and 0.730 per peak for the Buenrostro2018 dataset (Fig. 1b), 0.640 per cell and 0.662 per peak for the 10x multiome PBMCs and 0.701 per cell and 0.734 per peak for the 10x multiome mouse brain dataset. Randomly shuffling the active peaks within each cell led to mean 0.5 auROC per cell and decreased auROC per peak (although >0.5 due to the influence of different sequencing depths across cells; Supplementary Fig. 2).

Although these statistics are slightly below the 0.75–0.95 range achieved for bulk DNase samples in the original Basset publication, this is inevitable due to the substantially increased measurement noise due to sparse sequencing for the single cell assay. In support of this claim, we observed that in the 10x multiome PBMC and mouse brain datasets, peaks with very high read coverage are easier to predict (Supplementary Fig. 1). Given that ubiquitous accessible peaks are known to exist, these peaks are likely truly accessible in all cells and represent a rough upper bound on the achievable accuracy. To further assess the influence of sequencing depth, we downsampled the 10x multiome PBMCs at various levels and trained scBasset. As expected, validation auROC and cell-embedding metrics decrease with decreasing depth, but scBasset performance is still better than random even when the dataset contains only 1% nonzero entries (Supplementary Fig. 3).

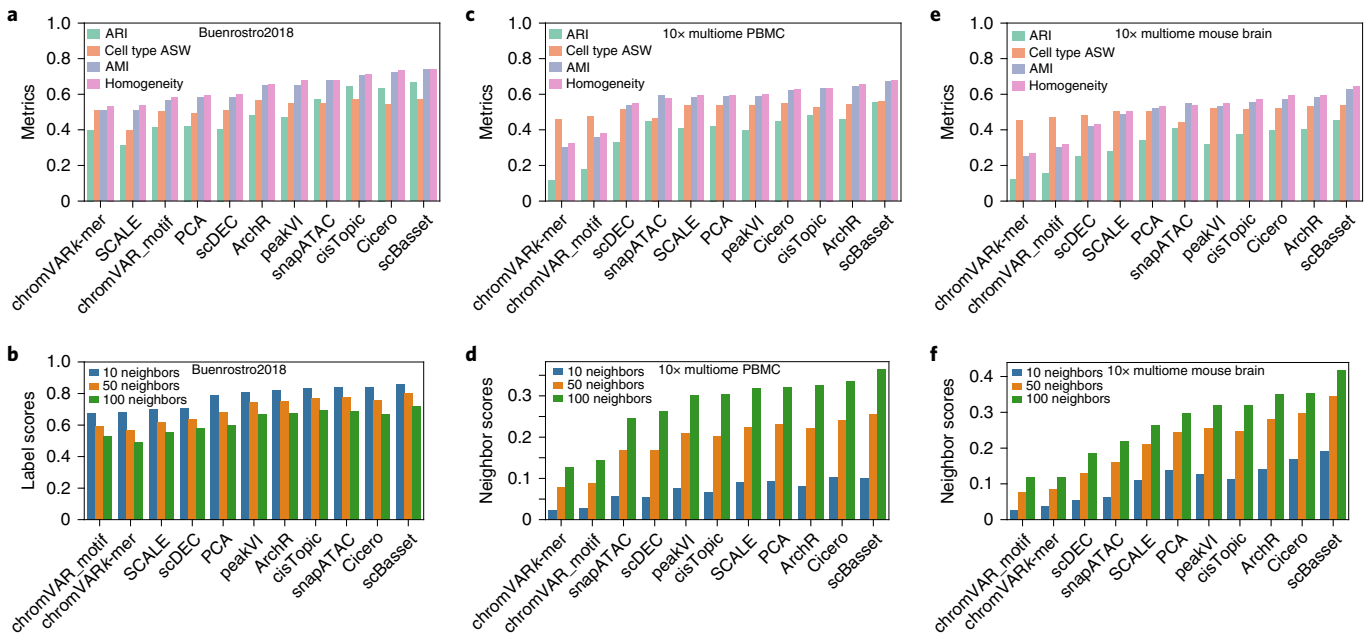


Fig. 2 | scBasset cell representation performance. a, Performance comparison of different cell-embedding methods evaluated by clustering metrics (ARI, cell type ASW and AMI) on the Buenrostro2018 dataset. **b**, Performance comparison of different cell-embedding methods evaluated by label score—the proportion of cells' nearest neighbors that share its cell type label (Methods)—on Buenrostro2018 dataset. **c**, Performance comparison of different cell-embedding methods evaluated by clustering metrics on 10x multiome PBMC dataset. **d**, Performance comparison of different cell-embedding methods evaluated by neighbor score—the proportion of cells' nearest neighbors that are also nearest neighbors in an independent scRNA analysis (Methods)—on 10x multiome PBMC dataset. **e**, Performance comparison of different cell-embedding methods evaluated by clustering metrics on 10x multiome mouse brain dataset. **f**, Performance comparison of different cell-embedding methods evaluated by neighbor score (Methods) on 10x multiome mouse brain dataset. Source data for this figure are provided.

scBasset final layer learns cell representations. We propose that the weight matrix learned in the final dense layer, which connects the bottleneck to the predictions, be used as a low-dimensional representation of the single cells. One requirement for an effective cell representation is removal of the influence of sequencing depth. Thus, we first verified that the intercept vector in the model's final layer almost perfectly correlates with cell sequencing depth for all datasets (Supplementary Fig. 4), suggesting that depth has been normalized out from the representations. Next, we compared the cell representations learned by scBasset with other methods both qualitatively and quantitatively. For the Buenrostro2018 dataset, we visualized the cell embeddings in two-dimensions (2D) using *t*-distributed stochastic neighbor embedding (*t*-SNE) (Fig. 1c) and observed that cells of the same type clustered together in the embedding space.

Following previous work, we quantified the correctness of cell embeddings in Buenrostro2018 by comparing Louvain clustering results with ground-truth cell-type labels using the adjusted rand index (ARI), adjusted mutual information (AMI) and homogeneity[6] or by directly evaluating the distance between cells of the same label by cell type average silhouette width (ASW)[18]. scBasset outperforms the other methods across these metrics (Fig. 2a). As Louvain clustering depends on hyperparameter choice and initialization, we proposed an alternative cluster-free method for evaluating cell embeddings. We computed a 'label score' by building a nearest-neighbor graph based on cell embeddings and asked what percentage of each cell's neighbors share its same label. For each embedding method, we computed label scores across a range of neighborhoods and observed that scBasset consistently outperforms the competitors at learning cell representations that embed cells of the same type near each other (Fig. 2b). We also evaluated label scores for each cell type individually and observed that plasmacytoid dendritic cells (pDCs) are learned best, whereas multipotent progenitor (MPP) cells are most difficult to distinguish (Supplementary Fig. 5). Visualizing the scATAC cell embeddings generated by different approaches by *t*-SNE, we observed that chromVAR, PCA and scDEC struggle to distinguish common lymphoid progenitor (CLP) cells from lymphoid primed MPP (LMPP) cells, whereas scDEC and SCALE struggle to distinguish megakaryocyte–erythroid progenitor (MEP) cells from common myeloid progenitor (CMP) cells (Extended Data Fig. 1).

For the multiome PBMC and mouse brain datasets, we computed an analog to the label scores for cell embeddings. As the ground-truth cell types for the multiome datasets are unknown, we used cluster identifiers from scRNA-seq Leiden clustering as cell type labels. Again, scBasset outperforms the competitors by this metric across a range of neighborhoods using label score or conventional clustering metrics (Fig. 2c,e and Supplementary Fig. 6). For these multiome datasets, we also computed a 'neighbor score', in which we built independent nearest-neighbor graphs from the scRNA and scATAC and asked what percentage of each cell's neighbors are shared between the two graphs. scBasset outperforms the competitors on both multiome PBMC and multiome mouse brain datasets when evaluated with neighbor scores across a range of neighborhoods (Fig. 2d,f). We annotated multiome PBMC cells based on expression of marker genes (Methods) and visualized the cell embeddings from different methods using Uniform Manifold Approximation and Projection (UMAP; Extended Data Fig. 2). We observed that chromVAR, cisTopic, scDEC and peakVI struggle to distinguish FCGR3A$^+$ monocytes from CD14$^+$ monocytes, whereas scBasset clearly separates the two.

Batch-conditioned scBasset corrects cell embeddings for batch. In the Buenrostro2018 dataset, hematopoietic stem cells (HSCs) cluster into two populations, regardless of which cell-embedding

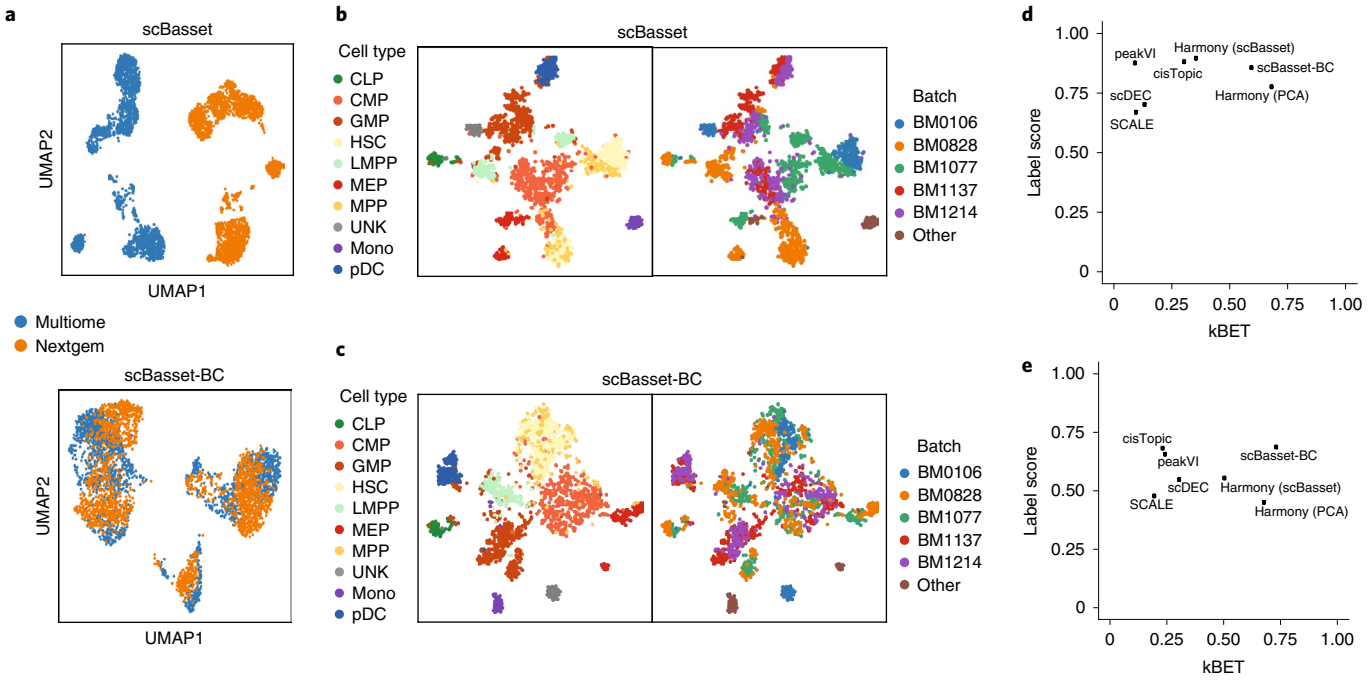


Fig. 3 | scBasset batch correction. a, Cell embeddings learned by scBasset without batch correction on a mixture of PBMC scATAC from 10x multiome and 10x next GEM chemistries (top). Cells are colored by chemistry. Cell embeddings learned by scBasset with batch correction (scBasset-BC) on the same data (bottom). **b**, Buenrostro2018 cell embeddings learned by scBasset, colored by cell type (left) or batch (right). **c**, Buenrostro2018 cell embeddings learned by scBasset-BC, colored by cell type (left) or batch (right). **d**, Performance comparison of different batch correction methods on chemistry-mixed PBMC data. Harmony is applied on either PCA, named Harmony (PCA), or scBasset embeddings, named Harmony (scBasset), and performance was evaluated by kBET and label score (with a neighborhood of 100). **e**, Similar performance comparison of different batch correction methods on Buenrostro2018 data. Source data for this figure are provided.

method we apply (Extended Data Fig. 1). As noted in previous studies, this is caused by batch effects due to different donors (Fig. 3b)[5,15]. To correct for this, and batch effects more generally, we explored modifications to the scBasset architecture. Specifically, after the bottleneck layer, we added a second fully connected layer to predict the batch-specific contribution to accessibility (Methods; Extended Data Fig. 3). We added the output of the batch layer and cell-specific layer before computing the final sigmoid. Intuitively, we expect the batch-specific variation will be captured in this path, whereas the original weight matrix will focus on the remainder of biologically relevant variation. By introducing an L2 normalization regularization term on the cell-specific layer, we can control the information flow and degree of batch mixing.

We first implemented scBasset-BC to correct for batch effects in a mixture of PBMC scATAC from 10x multiome and 10x next GEM chemistries (Fig. 3a). We quantified the mixing performance using integration local inverse Simpson's index (iLISI) and *k*-nearest-neighbor batch-effect test (kBET) acceptance rate (Methods) and quantified the conservation of biological variation using label score[19–21]. We observed that training an scBasset-BC model with increasing L2 regularization results in better batch mixing at the expense of losing biological variation (Extended Data Fig. 3). The optimal balance was achieved at $L2 = 1 \times 10^{-6}$ for the chemistry-mixed PBMC dataset.

We comprehensively evaluated the batch correction performance of scBasset compared to alternative methods such as cisTopic, Harmony, peakVI, SCALE and scDEC (Methods). Harmony is a general batch correction algorithm that can be implemented on top of any cell embeddings. We implemented two versions of Harmony, applying it to embeddings from (1) PCA or (2) scBasset. We observed that in the chemistry-mixed PBMC dataset, with a balanced batch design where every cell type is represented by roughly the same number of cells in each batch, SCALE and scDEC result in loss of biological structure, whereas peakVI and cisTopic result in poor batch mixing. Harmony and scBasset-BC achieve the best balance between batch mixing and preserving structure in the data (Fig. 3d and Extended Data Fig. 3). Harmony (PCA) achieved slightly better mixing than scBasset-BC at the expense of worse cell type label maintenance. Harmony (scBasset) achieves the highest clustering and batch-mixing performance evaluated by label score and iLISI. Overall, Harmony and scBasset-BC performed similarly for correcting batches from this balanced design.

Next, we trained scBasset-BC on the Buenrostro2018 dataset. This dataset has an unbalanced batch design and represents a more practical case for batch correction application. Again, we evaluated scBasset-BC performance as a function of L2 regularization and observed that $L2 = 1 \times 10^{-8}$ achieved the best balance between mixing and conserving biological variation (Extended Data Fig. 4). On these data, Harmony over-mixes and results in loss of biological variation. cisTopic, peakVI, scDEC and SCALE all tend to under-correct and maintain separation of batches. scBasset-BC achieves the best balance between mixing and preserving biological variability (Extended Data Fig. 4 and Fig. 3e). The two HSC batches (BM0106 and BM0828) merge into one cluster. In addition, pDC cells from BM1137 and BM1214 batches previously fell into two distinct subclusters, but are mixed together after batch correction (Fig. 3b,c).

scBasset denoises single cell accessibility profiles. Due to the sparsity of scATAC, the binary accessibility indicator for any given cell and peak contains frequent false negatives, such that the data cannot be studied with true single cell resolution and is usually aggregated across cells; however, numerous methods deliver denoised (or imputed) numeric values to represent the accessibility status at

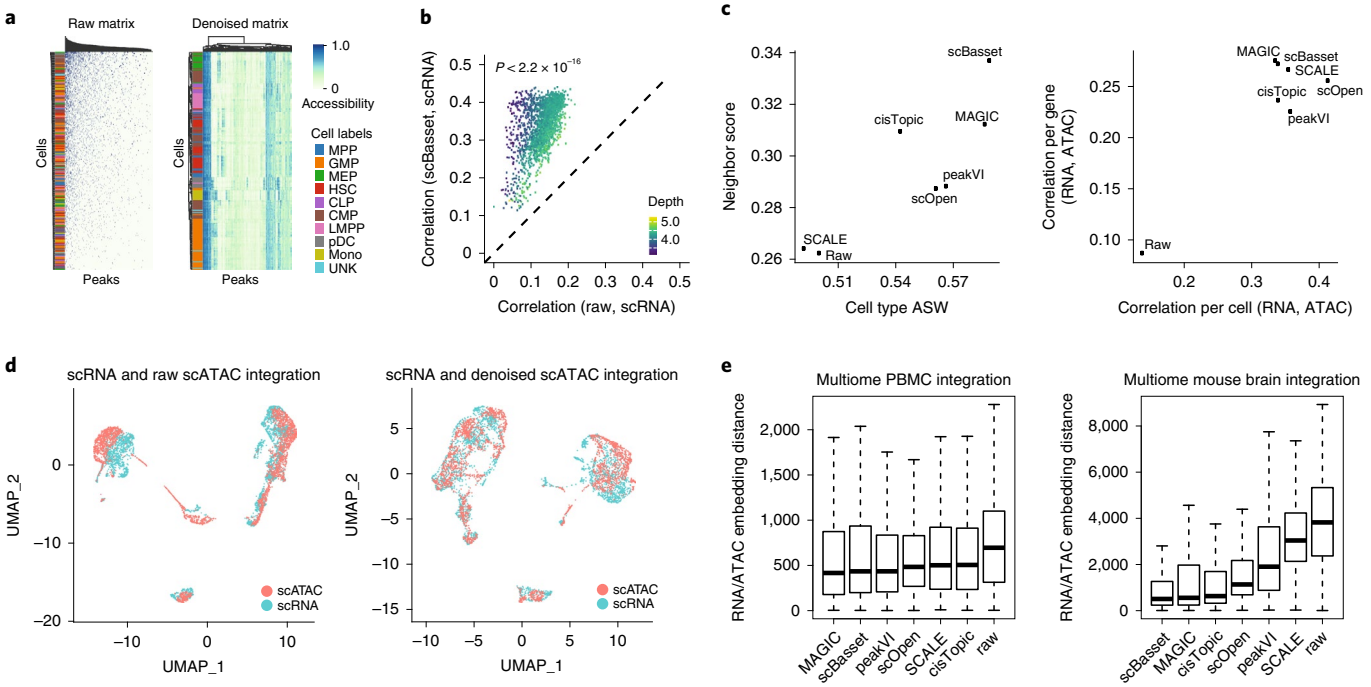


Fig. 4 | scBasset denoising performance evaluation. a, Binary count matrix of 200 cells and 500 peaks sampled from Buenrostro2018 dataset, hierarchically clustered by both cells and peaks (left). Cell type labels annotate the rows. The same matrix and procedure after scBasset denoising (right). **b**, Correlation between gene accessibility score and gene expression across genes for each cell before (*x* axis) and after scBasset denoising (*y* axis) for the multiome PBMC dataset. A one-sided Wilcoxon signed-rank test was performed. Cells are colored by sequencing depth. **c**, Comparison of different denoising methods in multiome PBMC dataset as evaluated by label score and cell type ASW (left). Comparison of different denoising methods in multiome PBMC dataset as evaluated by correlation between scVI-denoised RNA and denoised ATAC profiles across genes per cell (correlation per cell), and correlation between scVI-denoised RNA and denoised ATAC profiles across cells per gene (correlation per gene) (right). **d**, UMAPs of RNA and ATAC co-embeddings after integration for multiome PBMC dataset. Integration performed on scVI-denoised RNA (blue) and raw ATAC (red) (left). Integration performed on scVI-denoised RNA (blue) and scBasset-denoised ATAC (red) (right). **e**, Comparison of integration performance on multiome PBMC dataset. Performance is measured by the relative distances between each cell's RNA and ATAC embeddings (Methods) when integrating the scVI-denoised RNA profiles with ATAC profiles denoised with different methods; $n = 2{,}714$ cells for each box plot on the left, and $n = 4{,}881$ cells for each box plot on the right. The box plot shows min and max as whiskers (excluding outliers), first and third quartiles as boxes and median in the center. Outliers (>1.5× interquartile range away from the box) are not shown. Source data for this figure are provided.

every cell/peak combination. scBasset computes such values in its sequence-based predictions.

From the Buenrostro2018 dataset, we sampled 500 peaks and 200 cells and directly visualized the raw cell-by-peak matrix versus the denoised matrix (Fig. 4a). In the raw binary matrix, we observed that cells and peaks clustered by sequencing depth, showing no biologically relevant patterns. However, we observed that after scBasset denoising, cells of the same cell type share similar accessibility profiles and hierarchical clustering of cells matched well with ground-truth labels.

Following a previous study, we evaluated the denoising performance of scBasset on Buenrostro2018 by its impact on cell–cell distance estimation and cell embeddings[22]. (1) We evaluated the cell–cell distance matrix calculated from the denoised cell-by-peak matrix and asked whether cells of the same label are closer together, using the cell type ASW. (2) We performed PCA embedding (with 50 components) on the denoised cell-by-peak matrix and asked whether cells of the same type embed closer together, as evaluated by our label score on nearest neighbor sets. Comparing to cisTopic, peakVI, MAGIC, SCALE and scOpen, we observed that scBasset denoising outperformed these alternative approaches on the two metrics (Extended Data Fig. 5). scBasset denoising also results in more robust differential accessibility results (Supplementary Fig. 7).

Several published strategies aggregate scATAC counts in the region around a gene's transcription start site to estimate its expression[7,23]. We propose that effective denoising would improve the correlation between these gene accessibility estimates and the gene's measured RNA expression in multiome experiments. Thus, we computed accessibility scores for each gene by averaging the predicted accessibility values at all promoter peaks before and after denoising (Methods). For both the 10x multiome PBMC and mouse brain datasets, we observed that scBasset denoising improves the consistency between gene accessibility and expression ($P < 2.2 \times 10^{-16}$, Wilcoxon signed-rank test). As one would expect, the improvement is greater for cells with fewer scATAC unique molecular identifiers (Fig. 4b and Extended Data Fig. 5).

For 10x multiome PBMC and 10x multiome mouse brain datasets, we quantified denoising performance by adopting metrics from[22] and our additional multiome-specific metrics. We quantified the consistency between gene accessibility and expression either as correlation across genes for each cell (correlation per cell) or as correlation across cells for each gene (correlation per gene).

We observed that scBasset outperformed all alternative methods on the cell type ASW and neighbor score metrics in the multiome datasets; however, there was not a clear winner for the RNA/ATAC consistency metrics (Fig. 4c and Extended Data Fig. 5). Methods such as scBasset and MAGIC achieve better correlation per gene, whereas methods such as peakVI and scOpen achieve greater correlation per cell. Overall, Spearman's correlation between 'correlation by cell' and 'correlation by gene' metrics for different methods

(excluding the raw count matrix) is −0.71 for multiome PBMCs and −0.89 for multiome mouse brain datasets. This suggests a tradeoff between smoothing across cells and preserving cell–cell variability.

Integration of cells independently profiled by scRNA and scATAC into a shared latent space is a key step for many scATAC annotation and analysis methods[24]. We hypothesized that scATAC denoising would improve scRNA and scATAC integration performance. To evaluate integration performance, we treated the 10x multiome scRNA and scATAC profiles as having originated from two independent experiments and quantitatively measured the rank distance between the RNA and ATAC embeddings for each matching cell (Methods). We observed that denoising either the scRNA or the scATAC profiles improves integration performance and optimal performance is achieved when both profiles are denoised (Fig. 4d and Extended Data Fig. 5). Comparing scBasset to alternative scATAC denoising methods, we observed that scBasset and MAGIC outperformed alternative methods for data integration (Fig. 4e).

scBasset infers transcription factor activity at single-cell resolution. TF binding is a major driver of chromatin accessibility[25]. As scBasset learns to predict accessibility from sequence, we expect the model to capture sequence information predictive of TF binding. To query the single cell TF activity, we leveraged the flexibility of the scBasset model to predict arbitrary sequences. More specifically, we fed synthetic DNA sequences (dinucleotide shuffled peaks) with and without a particular TF motif of interest to a trained scBasset model and evaluated the activity of the motif in each cell based on changes in predicted accessibility (Methods)[11]. If a TF is playing an activating role in a particular cell, we expect to see increased accessibility after the TF motif is inserted.

TF regulation in the hematopoietic lineage profiled in the Buenrostro2018 dataset has been studied in detail. We performed motif insertion for all 733 human CIS-BP motifs using the Buenrostro2018-trained model and recapitulated known trajectories of motif activity. For example, CEBPB, a known regulator of monocyte development, shows the highest activity in monocytes; GATA1, a key regulator of the erythroid lineage, is predicted to be most active in MEPs; and HOXA9, a known master regulator of HSC differentiation, has the highest predicted activity in HSCs (Extended Data Fig. 6)[15].

Previous sequence-based methods such as chromVAR are also able to quantify TF motif activity. To systematically compare scBasset and chromVAR on this task, we analyzed the 10x PBMC multiome dataset, in which TF expression measured in the RNA-seq can serve as a proxy for its motif's activity. We inferred motif activity for all 733 human CIS-BP motifs using both scBasset and chromVAR. For the 203 TFs that are significantly differentially expressed between cell type clusters, we asked how well the inferred TF activity per cell correlates with its expression. We observed that overall scBasset TF activities correlate significantly better with expression than chromVAR TF activities ($P < 3.38 \times 10^{-2}$, Wilcoxon signed-rank test) (Fig. 5b). This one-sided test is an underestimate of scBasset's performance advantage over chromVAR, as we would expect TF expression and inferred activity to be negatively correlated for repressors. Thus, we evaluated scBasset and chromVAR on activating and repressive TFs separately. For 74 TFs that both methods agreed on a positive TF expression–activity correlation, scBasset-predicted TF activities have significantly greater correlation with expression than chromVAR-predicted activity ($P < 7.38 \times 10^{-12}$, Wilcoxon signed-rank test; Extended Data Fig. 7). For 41 TFs that both methods agreed on a negative TF expression–activity correlation, scBasset-predicted TF activities have a significantly lesser correlation (more negative) with expression than chromVAR-predicted activity ($P < 1.62 \times 10^{-8}$, Wilcoxon signed-rank test). This is also true for the 10x multiome mouse brain dataset (Extended Data Fig. 7).

Examining some of the key regulators of PBMC cell types, we observed that scBasset TF activities have greater cell type specificity and correlate better with TF expression than chromVAR (Fig. 5c and Extended Data Fig. 8). We observed this for B-cell-specific activity of PAX5, T-cell-specific activity of TCF7, natural killer (NK)-cell-specific activity of RUNX3 and monocyte-specific activity of CEBPB. Notably, while scRNA-seq shows monocyte-specific expression of RXRA, scBasset and chromVAR strongly disagree, making opposite predictions for RXRA activity; scBasset predicts RXRA as a repressor ($r = -0.70$), whereas chromVAR suggests an activating role ($r = 0.56$). A literature review revealed stronger evidence that RXRA plays a repressive role in the myeloid lineage through direct DNA binding, which is more consistent with the scBasset prediction[26].

Beyond TF activity correlation with TF expression, one can validate TF activity inference by studying data in which the TF has been perturbed. Pierce et al. introduced a technique called spear-ATAC, in which they targeted specific TFs with CRISPRi in single cells and read out the perturbed ATAC-seq profile[27]. To further validate scBasset TF activity inferences, we studied a dataset from this manuscript, consisting of a pool of nine CRISPRi single guide RNAs (sgRNAs) targeting GATA1 (sgGATA1) and GATA2 (sgGATA2) and nine inert sgRNA controls (sgNT) that were introduced into K562 cells expressing a dCas9-KRAB cassette. We trained a scBasset model on the preprocessed peak-by-cell matrix from the original paper. UMAP on the scBasset cell embeddings showed that cells with sgGATA1 can be clearly distinguished from cells with sgGATA2 and sgNT (Extended Data Fig. 9), which is consistent with the original publication. We compared scBasset and chromVAR's single cell TF activity inference scores by their ability to distinguish sgGATA1 cells from sgNT and sgGATA2 cells from sgNT using their inferred GATA1 and GATA2 scores, respectively (Fig. 5d and Extended Data Fig. 9). While both scBasset and chromVAR inferred GATA1 activity very well, scBasset achieved better auPR and auROC. chromVAR prediction begins to lose precision at recall >0.8, whereas scBasset maintained near-perfect precision even at 0.95 recall (Fig. 5d). Both methods struggle to distinguish sgGATA2 from sgNT cells in this experiment.

Unlike chromVAR, scBasset makes use of an accurate quantitative model that predicts cell-specific accessibility from the DNA. Not only are we able to query scBasset for TF activity on a per-cell level, we can also infer TF activity at per-cell per-nucleotide resolution. As a proof of principle, we examined a known enhancer for the β-globin gene that regulates erythroid-specific β-globin expression[28,29]. We performed in silico saturation mutagenesis (ISM) for this 100-bp sequence, in which we predicted the change in accessibility in every cell after mutating each position to its three alternative nucleotides. We aggregated to a single score for each position by taking the normalized ISM score for each reference nucleotide (Methods). Figure 5e,f shows the average ISM score for each cell type in the erythroid lineage. Using a procedure based on mapping position weight matrices (PWMs) and computing a Pearson correlation between the PWM and ISM scores, we observed that the most influential nucleotides correspond to GATA1 and KLF1 motifs, which are known to bind to this enhancer region and regulate β-globin expression[30].

Examining the per-cell ISM scores, we observed the GATA1 and KLF1 motifs contribute more to accessibility as the cells differentiate in the erythroid lineage (Fig. 5f,g). In comparison, these two motifs' nucleotides have low scores in cell types outside of the erythroid lineage (Supplementary Fig. 8). This experiment suggests that scBasset learns the accessibility regulatory grammar at a single-cell resolution and could be used to identify the TFs regulating specific enhancers in individual cells and lineages.

scBasset scales to million cell datasets. As single-cell datasets continue to grow in size, scalable and efficient computational methods become critical. scBasset trains on batches of sequences,

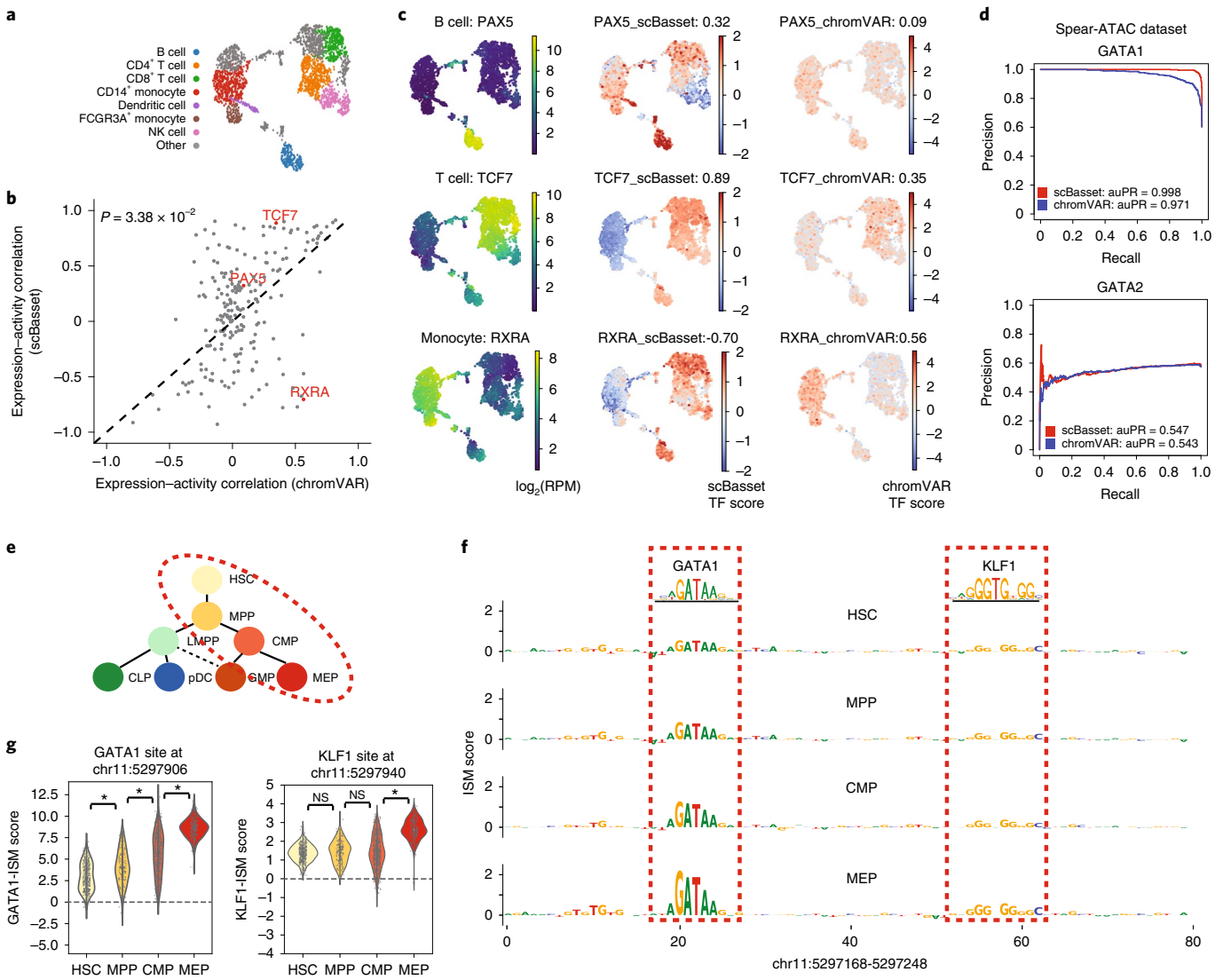


Fig. 5 | scBasset infers single cell TF activity. a, UMAP showing annotated PBMC cell types. **b**, Pearson correlation between TF expression and scBasset or chromVAR-predicted TF activity for 203 differentially expressed TFs. A one-sided Wilcoxon signed-rank test was performed. The example TFs that we examined in **c** are highlighted in red. **c**, UMAP visualization of TF expression (left), scBasset TF activity (middle) and chromVAR TF activity (right) for key PBMC regulators. Pearson correlation between inferred TF activity and expression are shown in the title. **d**, Precision-recall (PR) curves of scBasset and chromVAR for distinguishing sgGATA1 cells from sgNT cells in the spear-ATAC dataset (top). PR curves of scBasset and chromVAR for distinguishing sgGATA2 cells from sgNT cells (bottom). **e**, HSC differentiation lineage diagram in the Buenrostro2018 study. **f**, ISM scores for β-globin enhancer at chr11:5297158-5297258 for HSC, MPP, CMP and MEP cell types. Sequences that match GATA1 and KLF1 motifs are highlighted in red boxes. **g**, Distributions of per-cell TF PWM-ISM scores for GATA1 and KLF1 for cells in HSC, MPP, CMP and MEP cell types. $n = 502, 344, 142, 138$ cells for each of CMP, HSC, MPP and MEP. The PWM-ISM score is the dot product of the PWM and ISM measurements at sites of motif matches (GATA1 at chr11:5297906 and KLF1 at chr11:5297940). A one-sided Wilcoxon rank-sum test was performed to test for significance. $^*P < 0.01$; NS, not significant. Exact P values are $P = 2.06 \times 10^{-9}$ for MPP versus HSC, $P = 2.46 \times 10^{-11}$ for CMP versus MPP, and $P = 3.83 \times 10^{-39}$ for MEP versus CMP for GATA1; $P = 0.10$ for MPP versus HSC, $P = 0.38$ for CMP versus MPP and $P = 4.95 \times 10^{-41}$ for MEP versus CMP for KLF1. Source data for this figure are provided.

but predicts all cells in every batch. Thus, the complexity of each batch step depends on the number of cells but not peaks.

We assessed scBasset's scaling properties by training on one of the largest available scATAC datasets, in which the sci-ATAC method was applied to many human tissues[31], to map 1.3 million cells and more than 200 cell types. After filtering, we trained scBasset with 1,114,621 cells and 118,043 peaks (Extended Data Fig. 10). scBasset takes 273 s per epoch on this dataset using an Nvidia A100 GPU, with a peak CPU memory usage of 59.5 GB and peak GPU memory usage of 19.2 GB. We trained for 1,000 epochs, which required 76 h. However, the results change minimally in the later epochs, so some users might choose to train for fewer epochs or begin examining intermediate results during training.

To study the influence of cell number on runtime and memory usage, we trained scBasset on downsampled human sci-ATAC data with 10,000, 20,000, 50,000, 100,000, 200,000, 400,000, 600,000, 800,000, 1 million and all cells, and measured the runtime, peak CPU memory usage, and peak GPU memory usage (Extended Data Fig. 10). We observed that runtime, CPU memory and GPU memory all scale linearly with cell number but with a small slope. When we increase the number of cells from 10,000 to 1 million (100×), runtime per epoch goes from 49 s to 293 s (6×), CPU memory goes

from 8 G to 60 G (8×) and GPU memory goes from 1.5 G to 19 G (13×). This result suggests that scBasset is suitable for analysis of very large scATAC compendium.

Discussion

In this study we present scBasset, a sequence-based deep-learning framework for modeling scATAC data. scBasset is trained to predict individual cell accessibility from the DNA sequence underlying ATAC peaks, learning a vector embedding to represent the single cells in the process. A trained scBasset model can strengthen multiple lines of scATAC, and we demonstrate state-of-the-art performance on several tasks. Clustering the model's cell embeddings achieves greater alignment with ground-truth cell type labels. scBasset can be adapted to achieve state-of-the-art performance in batch correction tasks. The model outputs can be used as denoised accessibility profiles, which improve concordance with RNA measurements. The model learns to recognize TF motifs and their influence on accessibility, and we designed an in silico experiment to insert motifs into background sequences to query for TF motif activity in single cells. The model can also be applied to predict the influence of mutations, enabling in silico saturation mutagenesis of regulatory sequences of interest at a single-cell resolution. Compared to previous sequence-based approaches for scATAC such as chromVAR, scBasset achieves better performance at learning cell embeddings and inferring TF activity because scBasset benefits from a more expressive CNN model that learns more sophisticated sequence features, including nonlinear relationships. Compared to previous sequence-free approaches such as cisTopic, peakVI or SCALE, scBasset achieves better performance on benchmarking tasks and delivers a more interpretable model that can be directly queried for TF activity or identifying regulatory sequences.

Sequence-based approaches have several limitations. First, we make use of the reference genome, but many samples will have variant versions, including copy number variations that could lead our models astray. Second, we assume that the regulatory motifs and their interactions generalize across the genome. This assumption may not be entirely true at some genomic loci for which evolution led to bespoke regulatory solutions, such as for X chromosome inactivation in females. However, scBasset takes a completely independent approach to covariance-based methods, which handle this better, and researchers may appreciate running both on their data for multiple perspectives.

The foundational work with DNA CNNs has primarily focused on modeling bulk datasets[11,12]. scATAC, analyzed with existing workflows to clusters or cell type labels, can be aggregated into pseudo-bulk profiles representing those clusters or cell types. Previous work has demonstrated the validity and utility of training DNA CNNs on these single-cell-derived profiles to infer cell-type-specific TF regulators and predict cell-type-specific genetic variant effects[4,32,33]. scBasset also achieves these research objectives, but we focus here on the contributions of the method to single-cell embeddings for clustering and visualization, denoising and TF activity inference. Working at a single-cell resolution may be ideal for applications like continuous trajectory of cell states and other cases where discrete clusters may lose information, but working at cluster resolution may be a fine alternative for many other datasets and analyses.

In addition, we foresee several paths to further improve our method. To enhance scBasset memory efficiency to scale to extremely large datasets far beyond one million cells, one could sample mini-batches of both sequences and cells rather than only sequences in our current implementation. Methods such as TF-MoDISco could be applied to scBasset ISM scores for de novo motif discovery[14,34]. All approaches to scATAC depend on accurate peak calls, and predictive modeling frameworks have been proposed to help identify highly specific regulatory elements[35]. We expect a neural network model would further improve scATAC peak-calling by taking into account sequence information (and accounting for Tn5 transposition bias). Finally, we plan to explore transfer learning approaches in which models are pre-trained on large data compendia before fine-tune training on specific single cell datasets.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests and statements of data and code availability are available at https://doi.org/10.1038/s41592-022-01562-8.

Received: 8 September 2021; Accepted: 27 June 2022;
Published online: 8 August 2022

References

1. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
2. 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).
3. Miao, Z. et al. Single cell regulatory landscape of the mouse kidney highlights cellular differentiation programs and renal disease targets. *Nat. Commun.* **12**, 2277 (2021).
4. Cusanovich, D. A. et al. A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell* **174**, 1309–1324 (2018).
5. Bravo González-Blas, C. et al. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat. Methods* **16**, 397–400 (2019).
6. Chen, H. et al. Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biol.* **20**, 241 (2019).
7. Pliner, H. A. et al. Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol. Cell* **71**, 858–871 (2018).
8. Xiong, L. et al. SCALE method for single-cell ATAC-seq analysis via latent feature extraction. *Nat. Commun.* **10**, 4576 (2019).
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. de Boer, C. G. & Regev, A. BROCKMAN: deciphering variance in epigenomic regulators by k-mer factorization. *BMC Bioinf.* **19**, 253 (2018).
11. Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* **26**, 990–999 (2016).
12. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep-learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
13. Kelley, D. R. et al. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res.* **28**, 739–750 (2018).
14. Avsec, Ž. et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.* **53**, 354–366 (2021).
15. Buenrostro, J. D. et al. Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. *Cell* **173**, 1535–1548.e16 (2018).
16. Qin, Q. et al. Lisa: inferring transcriptional regulators through integrative modeling of public chromatin accessibility and ChIP-seq data. *Genome Biol.* **21**, 32 (2020).
17. Wu, K. E., Yost, K. E., Chang, H. Y. & Zou, J. BABEL enables cross-modality translation between multiomic profiles at single-cell resolution. *Proc. Natl Acad. Sci. USA* **118**, e2023070118 (2021).
18. Lotfollahi, M. et al. Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol.* **40**, 121–130 (2022).
19. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
20. Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A. & Theis, F. J. A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods* **16**, 43–49 (2019).
21. Luecken, M. D. et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
22. Li, Z. et al. Chromatin-accessibility estimation from single-cell ATAC-seq data with scOpen. *Nat. Commun.* **12**, 6386 (2021).
23. Granja, J. M. et al. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).
24. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).

25. Thurman, R. E. et al. The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
26. Kiss, M. et al. Retinoid X receptor suppresses a metastasis-promoting transcriptional program in myeloid cells via a ligand-insensitive mechanism. *Proc. Natl Acad. Sci. USA* **114**, 10725–10730 (2017).
27. Pierce, S. E., Granja, J. M. & Greenleaf, W. J. High-throughput single-cell chromatin accessibility CRISPR screens enable unbiased identification of regulatory networks in cancer. *Nat. Commun.* **12**, 2969 (2021).
28. Tuan, D., Solomon, W., Li, Q. & London, I. M. The 'β-like-globin' gene domain in human erythroid cells. *Proc. Natl Acad. Sci. USA* **82**, 6384–6388 (1985).
29. Li, Q., Peterson, K. R., Fang, X. & Stamatoyannopoulos, G. Locus control regions. *Blood* **100**, 3077–3086 (2002).
30. Tallack, M. R. et al. A global role for KLF1 in erythropoiesis revealed by ChIP-seq in primary erythroid cells. *Genome Res.* **20**, 1052–1063 (2010).
31. Zhang, K. et al. A single-cell atlas of chromatin accessibility in the human genome. *Cell* **184**, 5985–6001 (2021).
32. Kelley, D. R. Cross-species regulatory sequence activity prediction. *PLoS Comput. Biol.* **16**, e1008050 (2020).
33. Janssens, J. et al. Decoding gene regulation in the fly brain. *Nature* **601**, 630–636 (2022).
34. Shrikumar, A. et al. Technical note on transcription factor motif discovery from importance scores (TF-MoDISco) version 0.5.6.5. *arXiv.* https://arxiv.org/abs/1811.00416 (2018).
35. Lal, A. et al. Deep learning-based enhancement of epigenomics data with AtacWorks. *Nat. Commun.* **12**, 1507 (2021).

Methods

scATAC-seq preprocessing. We downloaded the processed peak set for Buenrostro2018 generated by Chen et al. at https://github.com/pinellolab/scATAC-benchmarking/blob/master/Real_Data/Buenrostro_2018/input/combined.sorted.merged.bed, which involved calling peaks on the aggregated profile of each cell type and merging them into a single atlas. We downloaded the aligned bam files from https://github.com/pinellolab/scATAC-benchmarking/tree/master/Real_Data/Buenrostro_2018/input/sc-bams_nodup, also provided by Chen et al.[6]. Peaks accessible in fewer than 1% cells were filtered out. The final dataset contains 103,151 peaks and 2,034 cells.

We downloaded the 10x multiome datasets from 10x Genomics: https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/2.0.0/pbmc_granulocyte_sorted_3k for the PBMC dataset and https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/2.0.0/e18_mouse_brain_fresh_5k for the mouse brain dataset. Genes expressed in fewer than 5% cells were filtered out. Peaks accessible in fewer than 5% cells were filtered out.

scRNA-seq preprocessing. For the 10x multiome datasets, we processed the expression data with scVI v.0.6.5 with n_layers, 1; n_hidden, 768; latent, 64 and a dropout rate of 0.2 (ref. [36]). We trained scVI for 1,000 epochs with a learning rate of 0.001, using the option to reduce the learning rate upon plateau using options lr_patience of 20 and lr_factor of 0.1. We enabled early stopping when there was no improvement on the evidence lower bound loss for 40 epochs. To generate denoised expression profiles, we used the get_sample_scale() function to sample from the generative model ten times and took the average.

Briefly, scVI performs denoising by modeling single-cell gene counts by negative binomial distributions and infers the parameters of these distributions with a variational autoencoder[36]. We used scVI-denoised expression profiles to benchmark scATAC denoising and integration performance as previous work has demonstrated that denoised expression values reflect the true values in the cell more accurately than the observed counts[37], and we observed better integration performance when both RNA and ATAC profiles were denoised (Extended Data Fig. 5). We used the learned latent cell representations to build nearest-neighbor graphs and perform cell clustering.

PBMC cell annotations. For multiome PBMC datasets, we performed a simple cell-type annotation based on gene expression data following a scanpy tutorial (https://scanpy-tutorials.readthedocs.io/en/latest/pbmc3k.html). Briefly, we first clustered the cells based on scVI latent cell embeddings using the Leiden algorithm. Then we normalized a cell-by-gene expression matrix by log(reads per 10,000). We ran rank_genes_groups() on the normalized gene expression matrix and plotted the top 25 enriched genes in each Leiden cluster. We compared the top enriched genes in each cluster with PBMC marker genes provided in the tutorial to assign cell type annotation to each cluster. Clusters where no marker genes were found in the top 25 enriched genes were assigned to 'other'.

Model architecture. scBasset is a neural network architecture that predicts binary accessibility vectors for each peak based on its DNA sequence. scBasset takes as input a 1,344-bp DNA sequence from each peak's center and one-hot encodes it as a $1{,}344 \times 4$ matrix. The neural network architecture includes the following blocks:

- 1D convolution layer with 288 filters of size 17×4, followed by batch normalization, GELU and width 3 max pooling layers, which generates a 488×288 output matrix.
- Convolution tower of six convolution blocks each consisting of convolution, batch normalization, max pooling and GELU layers. The convolution layers have increasing numbers of filters (288, 323, 363, 407, 456 and 512) and kernel width 5. The output of the convolution tower is a 7×512 matrix.
- 1D convolution layer with 256 filters of width 1, followed by batch normalization and GELU. The output is a 7×256 matrix, which is then flattened into a $1 \times 1{,}792$ vector.
- Dense bottleneck layer with 32 units, followed by batch normalization, dropout with rate 0.2, and GELU. The output is a compact peak representation vector of size 1×32.
- Final dense layer predicting continuous accessibility logits for the peaks in every cell.
- (Optional) to perform batch correction, we attach a second parallel dense layer to the bottleneck layer predicting batch-specific accessibility. This batch-specific accessibility is multiplied by the batch-by-cell matrix to compute the batch contribution to accessibility in every cell. This vector is then added to the previous continuous accessibility logits per cell (Extended Data Fig. 3). L2 regularization can be optionally applied to the cell-embedding path (with hyperparameter λ_1) or the batch-specific path (with hyperparameter λ_2) to tune the contribution of the batch covariate to the predictions.
- Final sigmoid activation to [0,1] accessibility probability.

The total number of trainable parameters in the model is a function of the number of cells (n) in the dataset. Specifically, the model will have $4{,}513{,}960 + 33 \times n$ trainable parameters. Due to extensive previous work establishing high-performing model architecture hyperparameter ranges[11–13], we only performed hyperparameter searches for the size of the bottleneck layer and optimization parameters, including batch size, learning rate, β1 and β2. For the optimization parameters, we chose the values that minimized training loss. For the bottleneck layer, we also examined cell-embedding metrics.

Training approach. We used a binary cross-entropy loss and monitored the training auROC after every epoch. We stopped training when the maximum training auROC improved by less than 1×10^{-6} in 50 epochs. This stopping criterion led to training for around 600 epochs for the Buenrostro2018 dataset, 1,100 epochs for the 10x multiome PBMC dataset and 1,200 epochs for the 10x multiome mouse brain dataset.

We focused on training auROC instead of validation auROC for model selection because we observed that the model continues to improve cell embeddings even after the point where the validation auROC has plateaued (Supplementary Fig. 9). Stopping criteria based on training set loss are typical for optimization of many statistical models but atypical for overparameterized deep-learning models that are prone to overfitting. The primary overfitting risk is reduced performance on held-out data, which we do not observe; validation auROC during the later stages of training is stable. Our hyperparameter analyses indicate that the 32-unit bottleneck layer is a major impediment to true overfitting. Thus, although the convolution towers may learn sequence factors that do not generalize well during the later training phase, the final layer weights (which serve as cell embeddings) are constrained and continue to learn from the cell–cell accessibility correlations in the training data.

We updated model parameters using stochastic gradient descent using the Adam update algorithm. We performed a random search for optimal hyperparameters including batch size, learning rate and β1 and β2 for the Adam optimizer. The best performance was achieved with a batch size of 128, learning rate of 0.01, β1 of 0.95 and β2 of 0.9995.

We focused on the Buenrostro2018 dataset to select the optimal bottleneck layer size. We trained models with bottleneck sizes of 8, 16, 32, 64 and 128 and observed that bottleneck size 32 gave the best performance (Supplementary Fig. 9).

scBasset trained on shuffled labels. To establish baseline performance, for each of the datasets, we trained scBasset on a training set with labels shuffled. For each cell in the training set, we first binarized the accessibility vector and then randomly shuffled the positives (accessibility regions), while the total number of positives (coverage) was not affected, and re-trained the scBasset model.

Performance evaluation on data dropout. To benchmark model performance as a function of data sparsity, we choose a scATAC dataset with relatively high sequencing depth, the 10x multiome PBMC dataset. The original scATAC peak-by-cell matrix contains 21.2% nonzero entries. We downsampled reads from this matrix and generated datasets of the same size but increasing sparsity. The sampled datasets contain 16.9%, 12.7%, 8.45%, 4.22%, 2.11% and 1.06% nonzero entries, which is 80%, 60%, 40%, 20%, 10% and 5% of the original data. Then we trained scBasset models on each of these dropout datasets and evaluated the training area under the curve and validation area under the curve, as well as clustering performance (neighbor score), as a function of sparsity.

Benchmarking existing methods. For evaluation of cell embeddings, we compared scBasset to principal component analysis (PCA) implemented in scikit-learn[38], latent sematic indexing (LSI) implemented in cicero[7], cisTopic[5], SCALE[8], chromVAR with motifs or k-mer features[9], ArchR[23], snapATAC[39], peakVI[40], and scDEC[41].

For evaluation of batch correction performance, we compared scBasset to Harmony[19], peakVI[40], scDEC[41], cisTopic[5] and SCALE[8].

For evaluation of scATAC denoising performance, we compare scBasset to cisTopic[5], peakVI[40], MAGIC[42], SCALE[8] and scOpen[22].

Cell-embedding evaluation. For implementation details of embedding methods, see Supplementary Notes.

Clustering-based metrics. We evaluated learned cell embeddings by comparing the clustering to the ground-truth labels (FACS-sorted cell-type labels for Buenrostro2018, RNA-based cell cluster labels for multiome data). We first built a nearest-neighbor graph using scanpy with default n neighbors of 15. Then we followed a previous study to tune for a resolution that outputs 10 clusters for Buenrostro2018, 18 clusters for multiome PBMC and 21 clusters for multiome mouse brain so that they match the number of ground-truth labels[6]. Finally, we compared the clustering outcome to the ground-truth cell type labels using ARI, AMI and homogeneity as implemented in sklearn.metrics.

Cell type average silhouette width. Silhouette width evaluates whether cells of the same label are embedded close together by quantifying the distance of a cell to other cells of the same label, as compared to distance to cells of different labels. We evaluated cell embeddings by cell type ASW as proposed in previous single-cell studies[18], which is the silhouette score average across all cells and re-normalized to 0 and 1.

Label score. We evaluated the learned cell embeddings using label score for all three datasets. For a given nearest-neighbor graph, label score quantifies what percentage of each cell's neighbors share its same label in a given neighborhood. For each cell-embedding method, we computed the label score across a neighborhood of 10, 50 and 100. As the ground-truth cell types for the multiome datasets are unknown, we used cluster identifiers from scRNA-seq Leiden clustering as cell-type labels.

Neighbor score. We evaluated the learned cell embeddings using neighbor score for the 10x multiome datasets. For a 10x multiome dataset, we built independent nearest-neighbor graphs from the scRNA (using scVI) and scATAC (using the cell-embedding method we wanted to evaluate) and quantified the percentage of each cell's neighbors that were shared between the two graphs across neighborhoods of size 10, 50 and 100.

Batch correction evaluation. For implementation details of batch correction methods, see Supplementary Notes.

Chemistry-mixed PBMC dataset. We first evaluated batch correction performance on a dataset with perfect batch design. We mixed PBMC populations from 10x PBMC multiome chemistry (https://cf.10xgenomics.com/samples/cell-arc/1.0.0/pbmc_granulocyte_sorted_10k/) and 10x PBMC next GEM chemistry (https://cf.10xgenomics.com/samples/cell-atac/2.0.0/atac_pbmc_10k_nextgem/). We generated a shared atlas of 21,017 peaks from the two datasets by resizing the 10x peak calls from the two datasets to 1,000 bp and took the intersection. We subsampled 2,000 cells from each dataset and merged them over the shared atlas.

Buenrostro2018 dataset. We compared the batch correction performance of different methods on the Buenrostro2018 dataset. This dataset has an unbalanced batch design and represents a more practical case for batch correction application. Since popular metrics for batch correction such as kBET and iLISI assume all batches are present in a local neighborhood in a batch-corrected population[21,43], we sampled the Buenrostro2018 dataset to contain only cells from batch 'BM0828' and 'BM1077' to compute kBET and iLISI metrics.

k-nearest-neighbor batch-effect test acceptance rate. kBET acceptance rate measures batch mixing by the concordance of local batch distribution with the global batch distribution[20]. Higher acceptance rate indicates better mixing. We implemented the kBET R package (v.0.99.6) to compute kBET acceptance rate.

Integration local inverse Simpson's index. iLISI measures batch mixing by the effective number of batch labels in a local neighborhood[19]. Higher iLISI score indicates better mixing. We implemented the lisi R package (v.1.0) to compute iLISI scores.

Label score. We quantified the conservation of biological variation after batch correction by evaluating the cell embeddings with label score. Ground-truth cell-type labels for Buenrostro2018 are provided by FACS-sorting. Ground-truth cell-type labels for multiome PBMCs are generated by annotating the matched RNA profiles as described previously.

Denoising evaluation. For implementation details of denoising methods, see Supplementary Notes.

To compute denoised and normalized accessibility across cells for a query peak with scBasset, we ran a forward pass on the input DNA sequence to compute the latent embedding for the peak. Then we generated the normalized accessibility across all cells through dot product of the peak, embedding with the weight matrix of the final layer. As sequencing depth information is entirely captured by the intercept vector of the final layer, we excluded the intercept term so that scBasset generates denoised profiles normalized for sequencing depth.

Following a previous study, we evaluated the denoising performance of scBasset for cell–cell distance estimation and cell embedding[22].

- Cell type ASW: we computed a cell–cell distance matrix from the denoised cell-by-peak matrix using 1 – PearsonR as the distance metric and asked whether cells of the same label are closer together, using the cell type ASW.
- Label score or neighbor score: we performed PCA embedding (PC = 50) on the denoised cell-by-peak matrix and asked whether cells of the same type embed closer together, as evaluated by our label score for Buenrostro2018 dataset and neighbor score for multiome datasets.

Then we evaluated additional multiome-specific metrics for 10x multiome datasets. Our evaluation is based on the hypothesis that effective denoising would improve the correlation between accessibility at genes' promoters and the genes' expression in multiome measurements[7,23]. For each gene, we computed a gene accessibility score by averaging accessibility values for peaks at the gene's promoter (2 kb from transcription start site). We evaluated denoising performance by:

- Correlation per cell: computing the Pearson correlation between the gene accessibility score and gene expression (after scVI denoising) across all genes for each individual cell.
- Correlation per gene: computing the Pearson correlation between the gene accessibility score and gene expression (after scVI denoising) across all cells for each gene.

Integration evaluation. To evaluate integration performance, we treated the 10x multiome scRNA and scATAC profiles as originated from two independent experiments. We summarized the accessibility profile to the gene level by computing the gene accessibility score as described above and integrated the scRNA and scATAC data by embedding them into a shared space using Seurat FindTransferAnchors() and TransferData() functions[24].

To quantify the integration performance, we measured a 'RNA/ATAC embedding distance' R_c between the RNA embedding and the ATAC embedding of each cell c in the co-embedding space. We use R_{rna} to represent the ranking of the Euclidean distance between RNA embedding and ATAC embedding of cell c among all neighbors of *c*'s RNA embedding and R_{atac} to represent the ranking of the same distance among all neighbors of *c*'s ATAC embedding. R_c is computed as the average of R_{rna} and R_{atac}. A smaller R_c indicates better integration, whereas a higher R_c indicates worse integration.

spear-ATAC analysis. spear-ATAC preprocessed count matrix 'K562-Pilot-scATAC-Peak-Matrix-SE.rds' was downloaded from the Gene Expression Omnibus (accession code GSE168851)[27]. This dataset contains a pool of nine CRISPRi sgRNAs targeting GATA1 (sgGATA1) and GATA2 (sgGATA2) and inert sgRNA controls (sgNT) that were introduced into K562 cells expressing a dCas9-KRAB367 cassette. Cells with unknown sgRNAs were filtered out (sgAssignFinal, 'UNK'). We kept cells with at least 5% peaks accessible and peaks accessible in at least 5% cells for training the scBasset model.

We used the cell embeddings generated by scBasset for visualization using UMAP. We scored GATA1 and GATA2 motif activity using either an scBasset motif insertion approach or using chromVAR. We compared scBasset and chromVAR in distinguishing sgGATA1 cells from sgNT using the predicted GATA1 scores and distinguishing sgGATA2 cells from sgNT cells using the predicted GATA2 scores. Prediction performance was evaluated by auPR and auROC.

sci-ATAC human atlas analysis. We downloaded the processed peak-by-cell matrix from the sci-ATAC human atlas stored at http://renlab.sdsc.edu/kai/Key_Processed_Data/Cell_by_cCRE/[31]. We kept peaks accessible in more than 0.5% cells, and cells with at least 500 peaks accessible. The filtered matrix contains 1,114,621 cells and 118,043 peaks. Storing such a matrix in a dense format would take more than 1 terabyte of disk space. The data are thus stored in h5ad and sequences used for training are also stored in h5 format. scBasset can easily be trained on a dataset of this size because it takes sparse data as input and interacts with batches of input at training time.

We trained scBasset on the whole sci-ATAC atlas as well as a sampled dataset with 10,000, 20,000, 50,000, 100,000, 200,000, 400,000, 600,000, 800,000 and 1,000,000 cells. We measured CPU memory, GPU memory and runtime when training scBasset on each dataset. CPU memory is monitored by psutil.Process.memory info() command after reading or creating matrices and peak memory usage is reported. GPU memory is monitored using Tensorboard Profiler. Runtime per epoch is reported by Tensorflow during training.

Motif insertion. We performed motif insertion on scBasset to compute a TF activity score for each TF for each cell. Specifically, we first generated 1,000 genomic background sequences by performing dinucleotide shuffling of 1,000 randomly sampled peaks from the atlas using fasta ushuffle[44]. For each TF in the motif database, we sampled a motif sequence from the PWM and inserted it into the center of each of the genomic background sequences. We ran forward passes through the model for both the motif-inserted sequences and background sequences to predict normalized accessibility across all cells. We took the difference in predicted accessibility between the motif-inserted sequences and background sequences as the motif influence for each sequence. We averaged this influence score across all 1,000 sequences for each cell to generate a cell-level prediction of raw TF activity. Finally, we *z* score-normalized the raw TF activities to generate the final TF activity predictions across all cells.

We used CIS-BP 1.0 single species DNA database motifs downloaded from https://meme-suite.org/meme/db/motifs for our motif analysis[45].

In silico saturation mutagenesis. We performed ISM to compute the importance scores of all single nucleotides on a sequence of interest. For each position, we ran three scBasset forward passes, each time mutating the reference nucleotide to an alternative. For each mutation, we compared the alternative accessibility prediction to that of the reference to compute the change in accessibility for each cell. We normalized the ISM scores for the four nucleotides at each position such that they summed to zero. We then took the normalized ISM score at the reference nucleotide as the importance score for that position.

In the β-globin enhancer ISM analysis, we labeled TF motifs using the following procedure. First, we scanned the DNA sequence for candidate motif matches using FIMO with a permissive *P* value threshold of 1×10^{-3} (ref. [46]). For any motif match, we assigned a score using a Pearson correlation or dot product

between the PWM and ISM. Finally, we performed a statistical test on the match score by comparing the observed correlation with a null distribution computed from shuffled input.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

We used only public datasets in this study. We downloaded the processed peak set for Buenrostro2018 generated by Chen et al. at https://github.com/pinellolab/scATAC-benchmarking/blob/master/Real_Data/Buenrostro_2018/input/combined.sorted.merged.bed. We downloaded the aligned bam files from https://github.com/pinellolab/scATAC-benchmarking/tree/master/Real_Data/Buenrostro_2018/input/sc-bams_nodup. The original datasets are from the Gene Expression Omnibus (GEO) under accession code GSE96769. We downloaded the 10x multiome datasets from 10x Genomics at https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/2.0.0/pbmc_granulocyte_sorted_3k for the PBMC dataset and https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/2.0.0/e18_mouse_brain_fresh_5k for the mouse brain dataset. We downloaded the processed peak-by-cell matrix from sci-ATAC human atlas (GEO accession code GSE184461) stored at http://renlab.sdsc.edu/kai/Key_Processed_Data/Cell_by_cCRE/. spear-ATAC preprocessed count matrix 'K562-Pilot-scATAC-Peak-Matrix-SE.rds' was downloaded from GEO (accession code GSE168851). Source data are provided with this paper.

Code availability

Code for training and using the scBasset model can be found at https://github.com/calico/scBasset. Instructions and tutorials are provided at the GitHub repository for how to train scBasset models from anndata and to compute cell embeddings, denoise accessibility profiles, perform TF activity inference and ISM from a trained scBasset model. A trained scBasset model for the Buenrostro2018 dataset is available in the kipoi model zoo (https://github.com/kipoi/models/tree/master/scbasset).

References

36. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
37. Hou, W., Ji, Z., Ji, H. & Hicks, S. C. A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biol.* **21**, 218 (2020).
38. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
39. Fang, R. et al. Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nat. Commun.* https://doi.org/10.1038/s41467-021-21583-9 (2021).
40. Ashuach, T., Reidenbach, D. A., Gayoso, A. & Yosef, N. PeakVI: A deep generative model for single-cell chromatin accessibility analysis. *Cell Rep. Methods* **2**, 100182 (2022).
41. Liu, Q., Chen, S., Jiang, R. & Wong, W. H. Simultaneous deep generative modelling and clustering of single-cell genomic data. *Nat. Mach. Intell.* https://doi.org/10.1038/s42256-021-00333-y (2021).
42. van Dijk, D. et al. Recovering gene interactions from single-cell data using data diffusion. *Cell* https://doi.org/10.1016/j.cell.2018.05.061 (2018).
43. Tran, H. T. N. et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* **21**, 12 (2020).
44. Jiang, M., Anderson, J., Gillespie, J. & Mayne, M. uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. *BMC Bioinf.* **9**, 192 (2008).
45. Weirauch, M. T. et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).
46. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).

Acknowledgements

We thank V. Agarwal, J. Kimmel and M. Mohamed for feedback on the manuscript. We thank S. Spock for feedback on the code. We also thank N. Bernstein and A. Odak for helpful discussions.

Author contributions

D.R.K. conceived the project. H.Y. and D.R.K. developed the model. H.Y. performed the analysis. H.Y. and D.R.K prepared the manuscript.

Competing interests

H.Y. and D.R.K. are paid employees of Calico Life Sciences.

Additional information

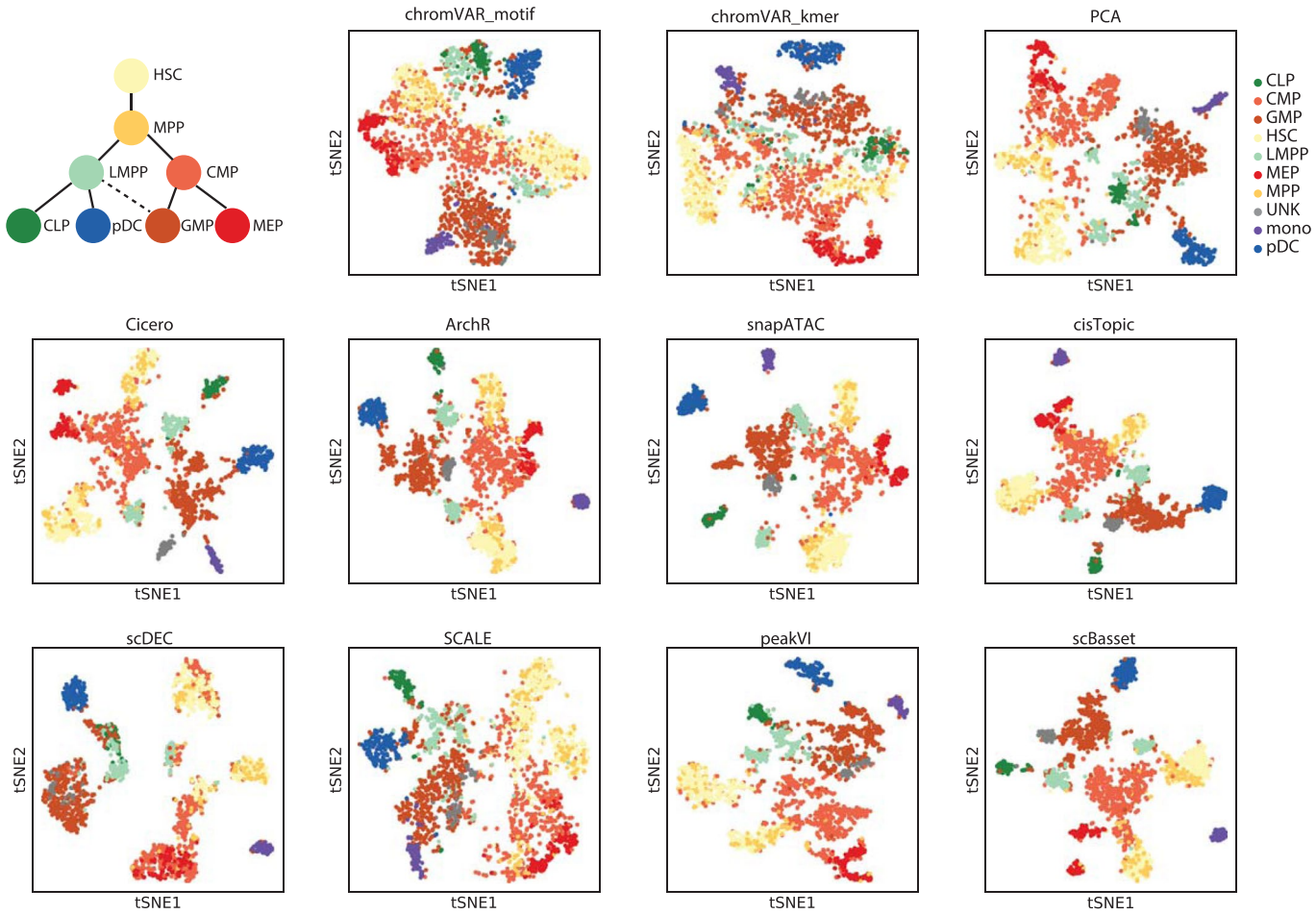
Extended data is available for this paper at https://doi.org/10.1038/s41592-022-01562-8.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-022-01562-8.

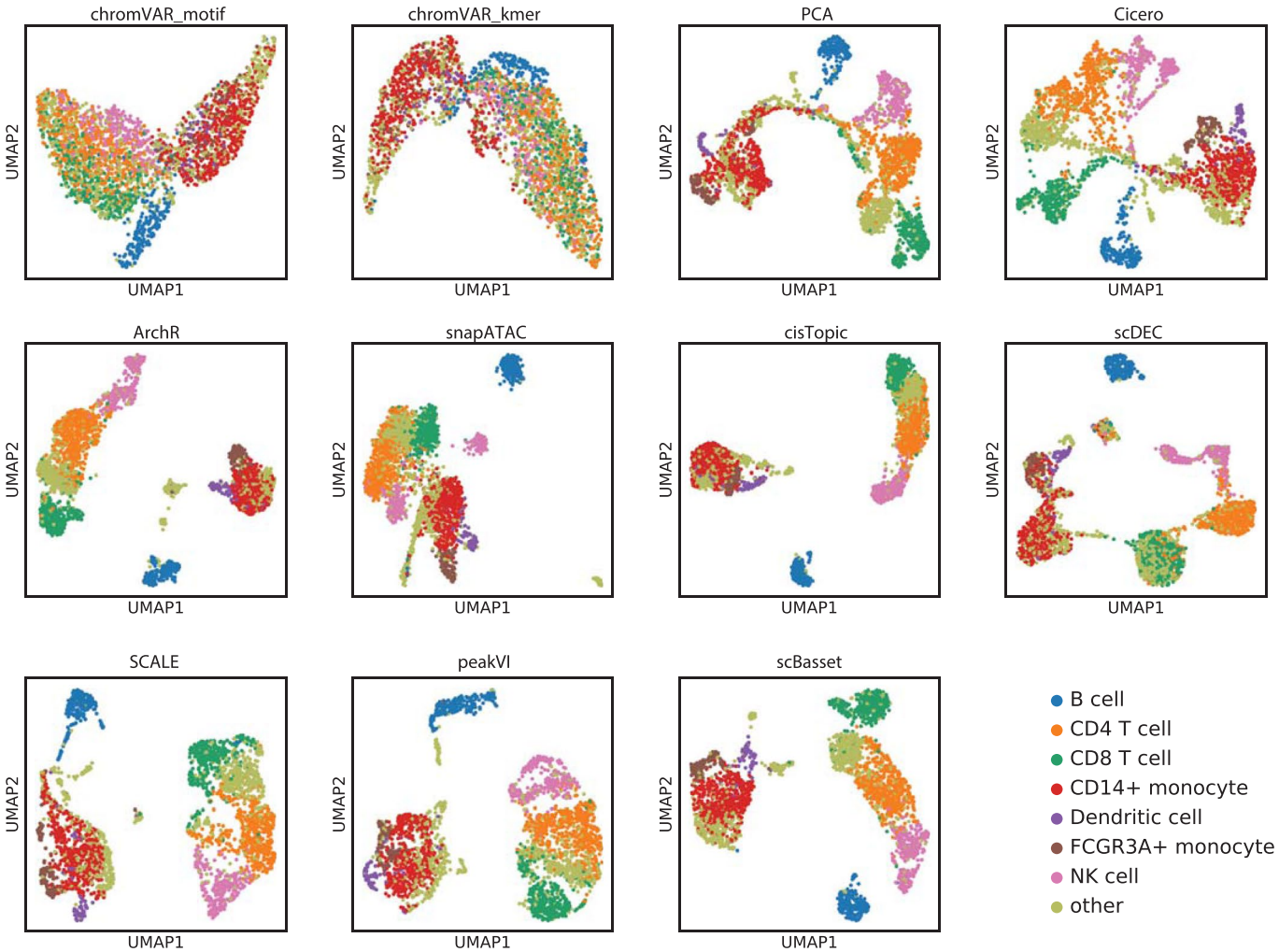
Correspondence and requests for materials should be addressed to Han Yuan or David R. Kelley.

Peer review information *Nature Methods* thanks Luca Pinello, Qiangfeng Cliff Zhang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling editor: Lin Tang, in collaboration with the *Nature Methods* team.

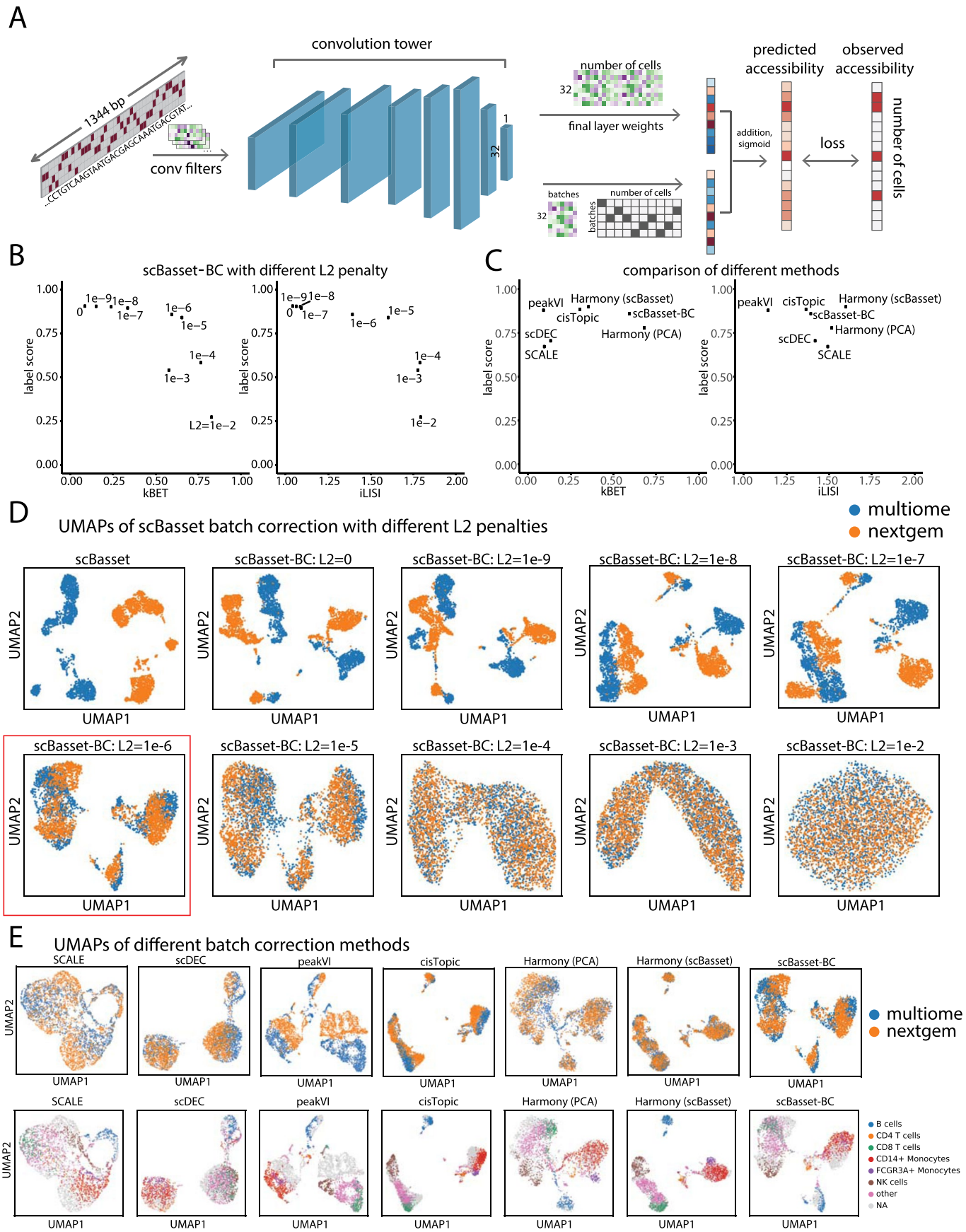
Reprints and permissions information is available at www.nature.com/reprints.



Extended Data Fig. 1 | Buenrostro2018 cell embeddings. t-SNE visualization of different cell embedding methods on Buenrostro2018, including: chromVAR motif, chromVAR kmer (k = 6), PCA, cicero (LSI), ArchR, snapATAC, cisTopic, scDEC, SCALE, peakVI and scBasset.

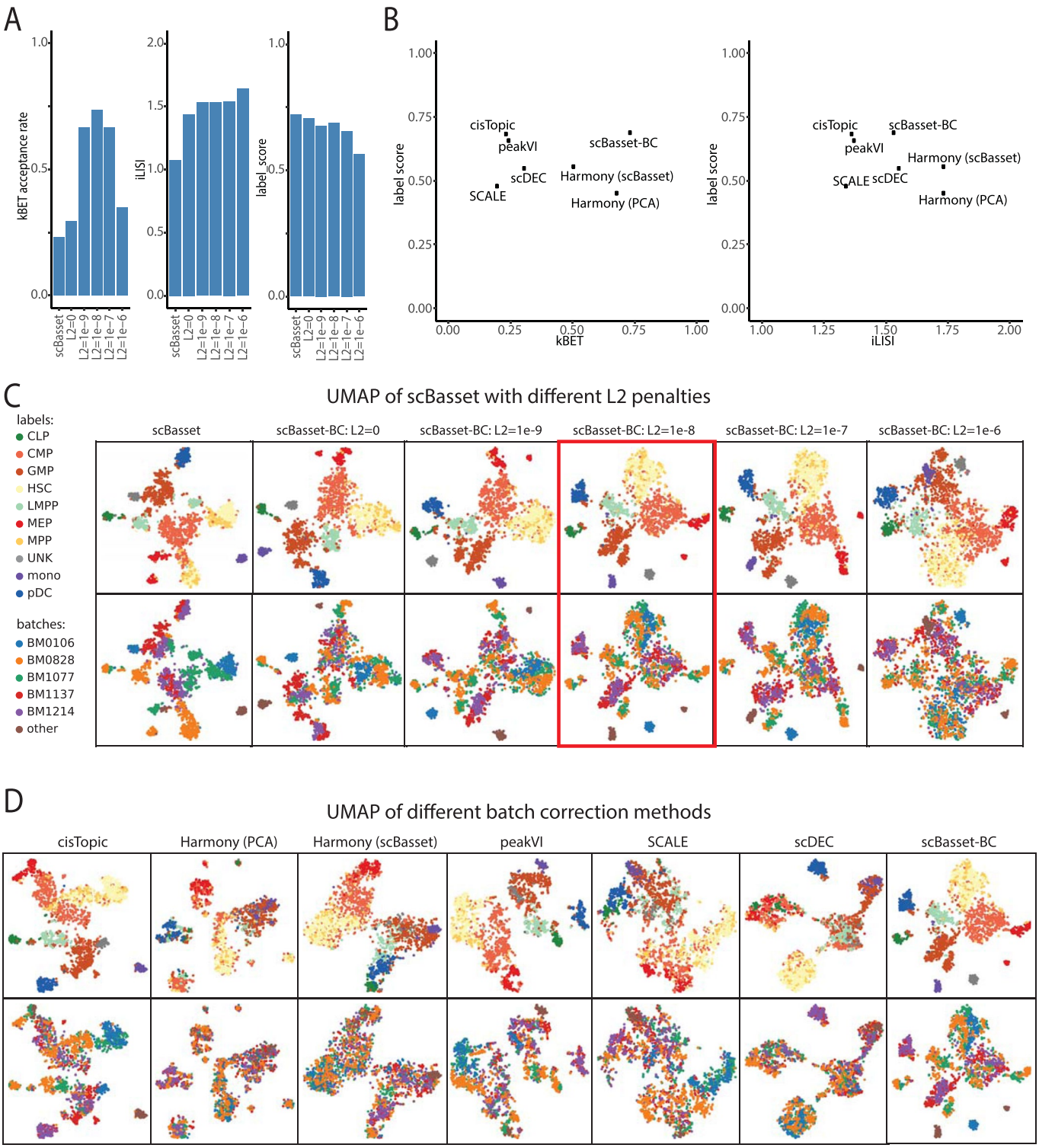


Extended Data Fig. 2 | 10x multiome PBMC cell embeddings. UMAP visualization of different cell embedding methods on the 10x multiome PBMC dataset, including: chromVAR_motif, chromVAR_kmer (k = 6), PCA, cicero (LSI), ArchR, snapATAC, cisTopic, scDEC, SCALE, peakVI and scBasset.

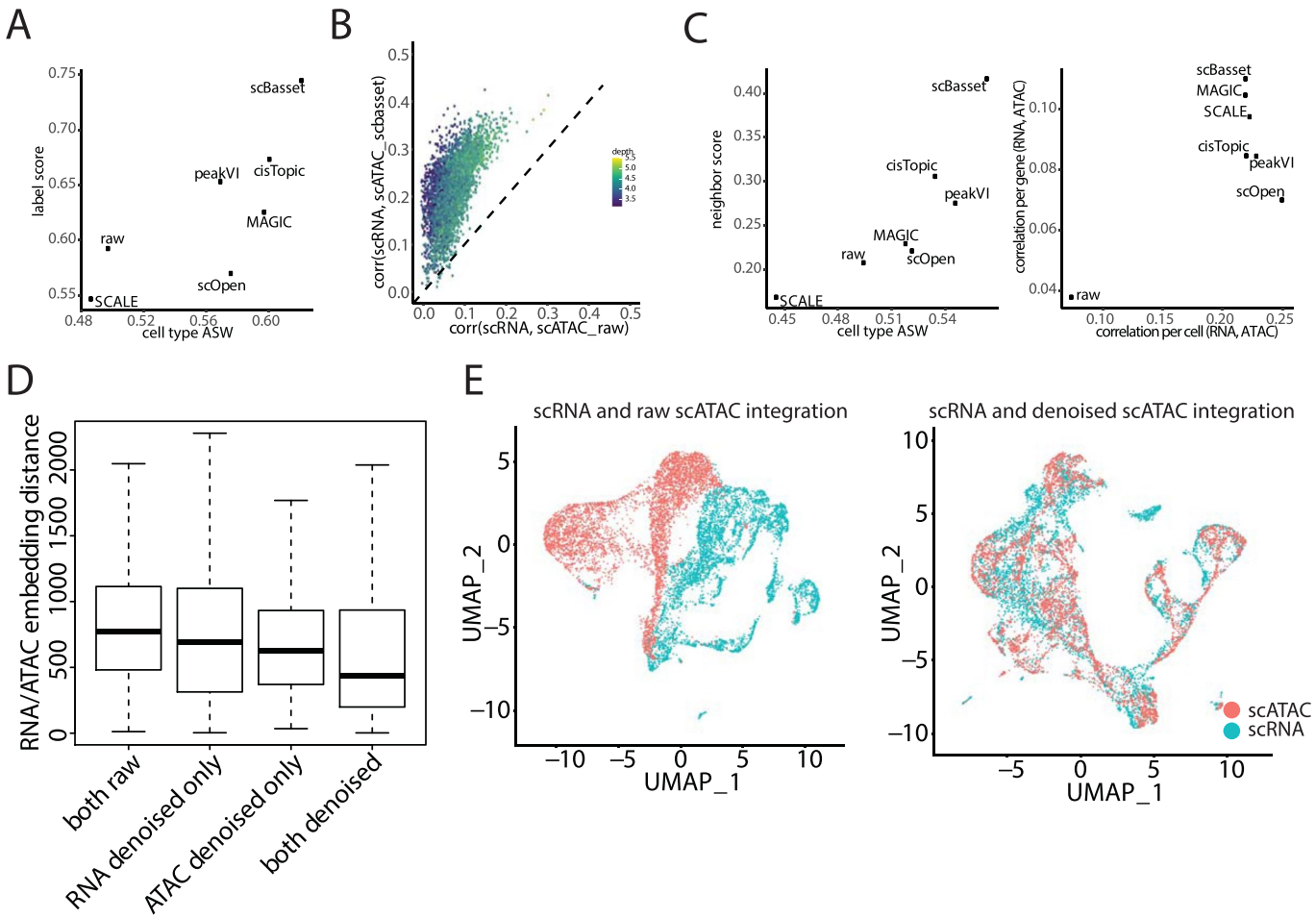


Extended Data Fig. 3 | See next page for caption.

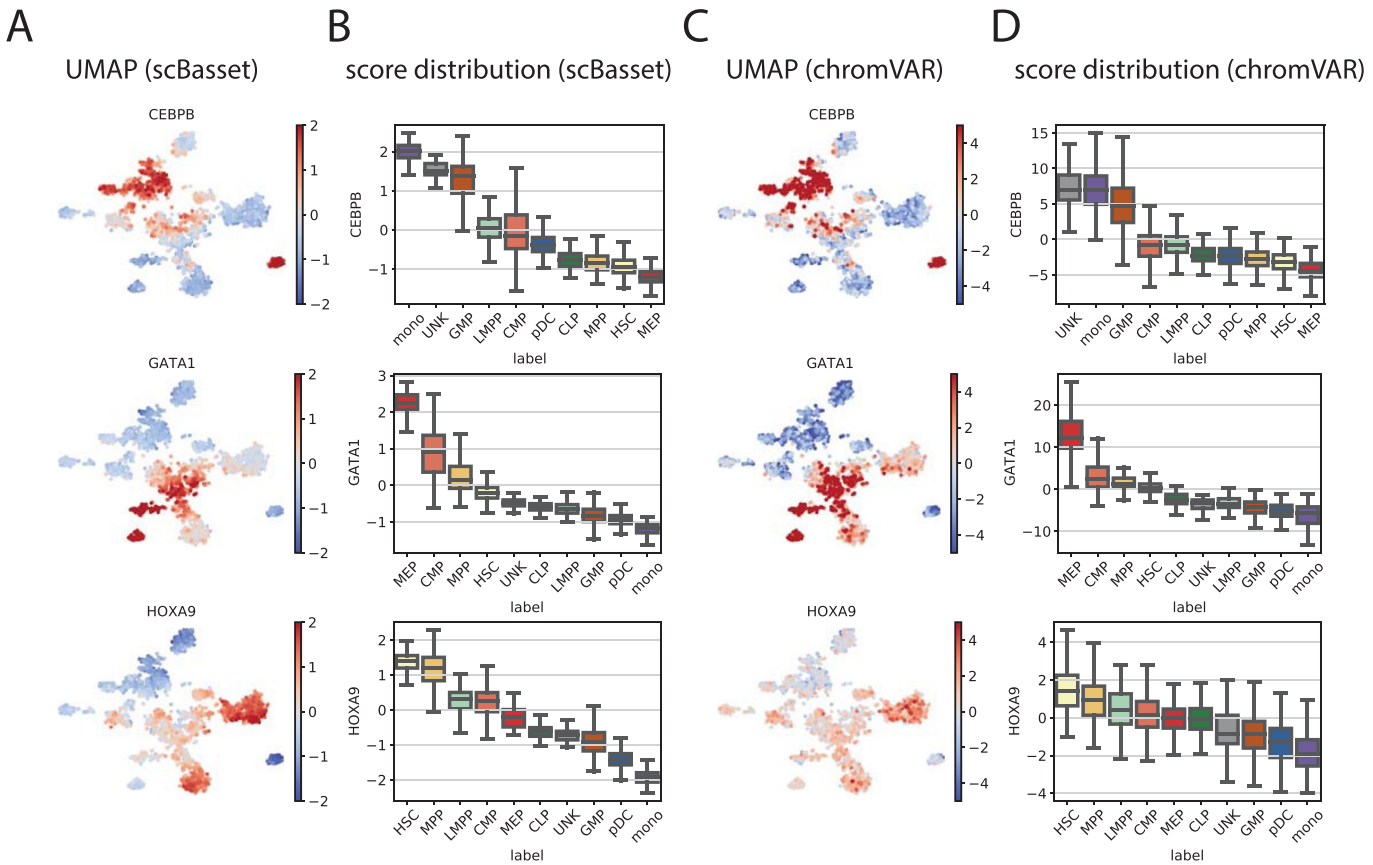
Extended Data Fig. 3 | scBasset batch correction on chemistry-mixed PBMC. a) Model architecture of scBasset-BC. **b**) Performance of scBasset batch correction models trained on the chemistry-mixed PBMC data. We trained scBasset-BC models with increasing L2 penalty (from 0 to 1e-2). Batch-mixing is measured with kBET and iLISI, and conservation of biological variation is measured with label score. **c**) Performance comparison of different batch correction methods on the chemistry-mixed PBMC data. Harmony is applied on either PCA, Harmony(PCA), or scBasset embeddings, Harmony(scBasset). Performance is evaluated by kBET, iLISI and label score. **d**) UMAPs of scBasset batch correction with different L2 penalties on the chemistry-mixed PBMC data. We selected L2 = 1e-6 as the fin42al scBasset-BC model. **e**) UMAPs of different batch correction methods on the chemistry-mixed PBMC data.



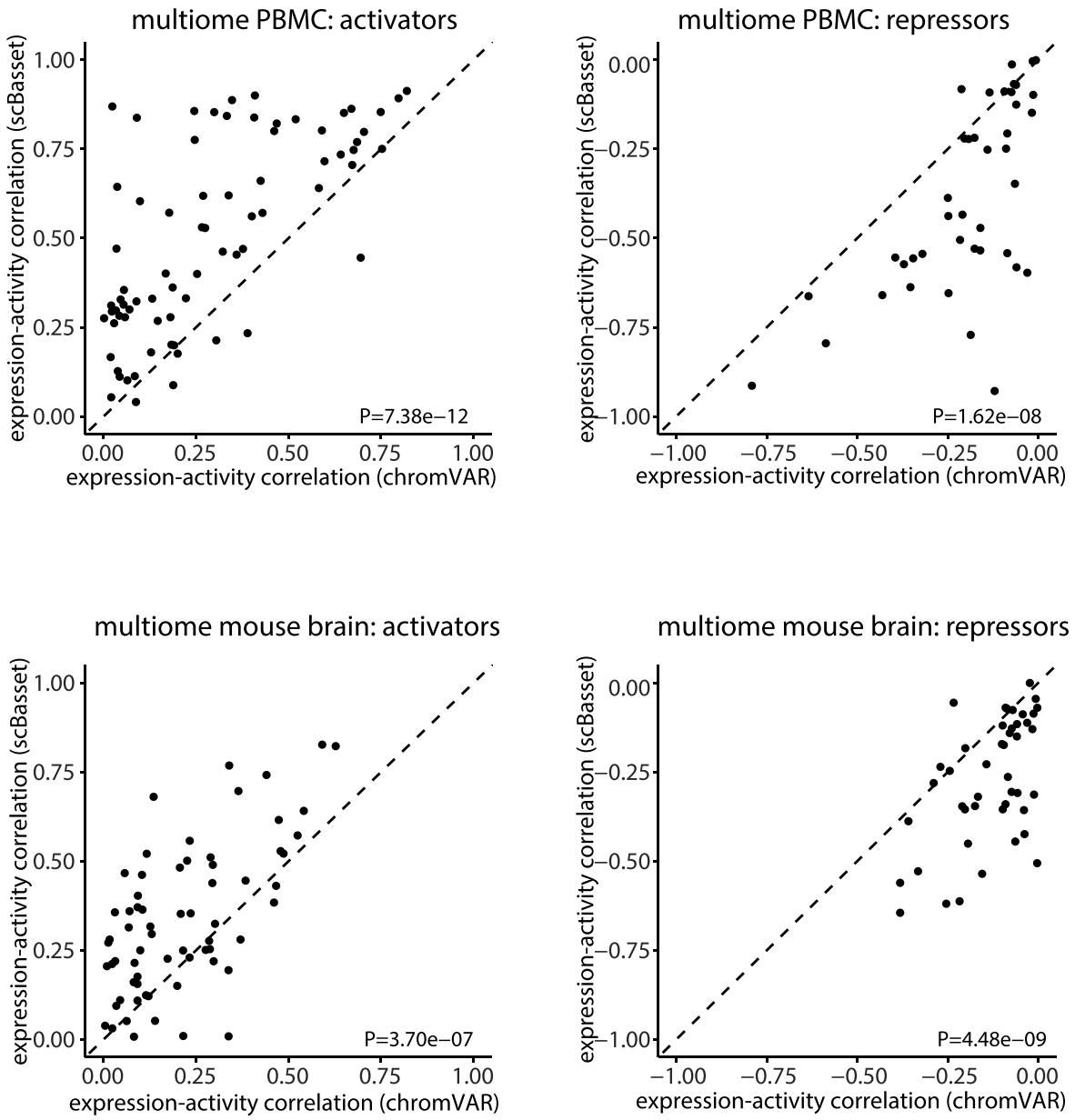
Extended Data Fig. 4 | scBasset batch correction on Buenrostro2018. a) scBasset batch correction performance as a function of L2 penalty on Buenrostro2018 dataset. Performance is evaluated by kBET, iLISI and label score. **b**) Performance comparison of different batch correction methods on Buenrostro2018 data. Harmony is applied on either PCA, Harmony(PCA), or scBasset embeddings, Harmony(scBasset). Performance is evaluated by kBET, iLISI and label score. **c**) UMAPs of scBasset batch correction with different L2 penalties on Buenrostro2018 data. We selected L2 = 1e-8 as the final scBasset-BC model. **d**) UMAPs of different batch correction methods on Buenrostro2018 data.



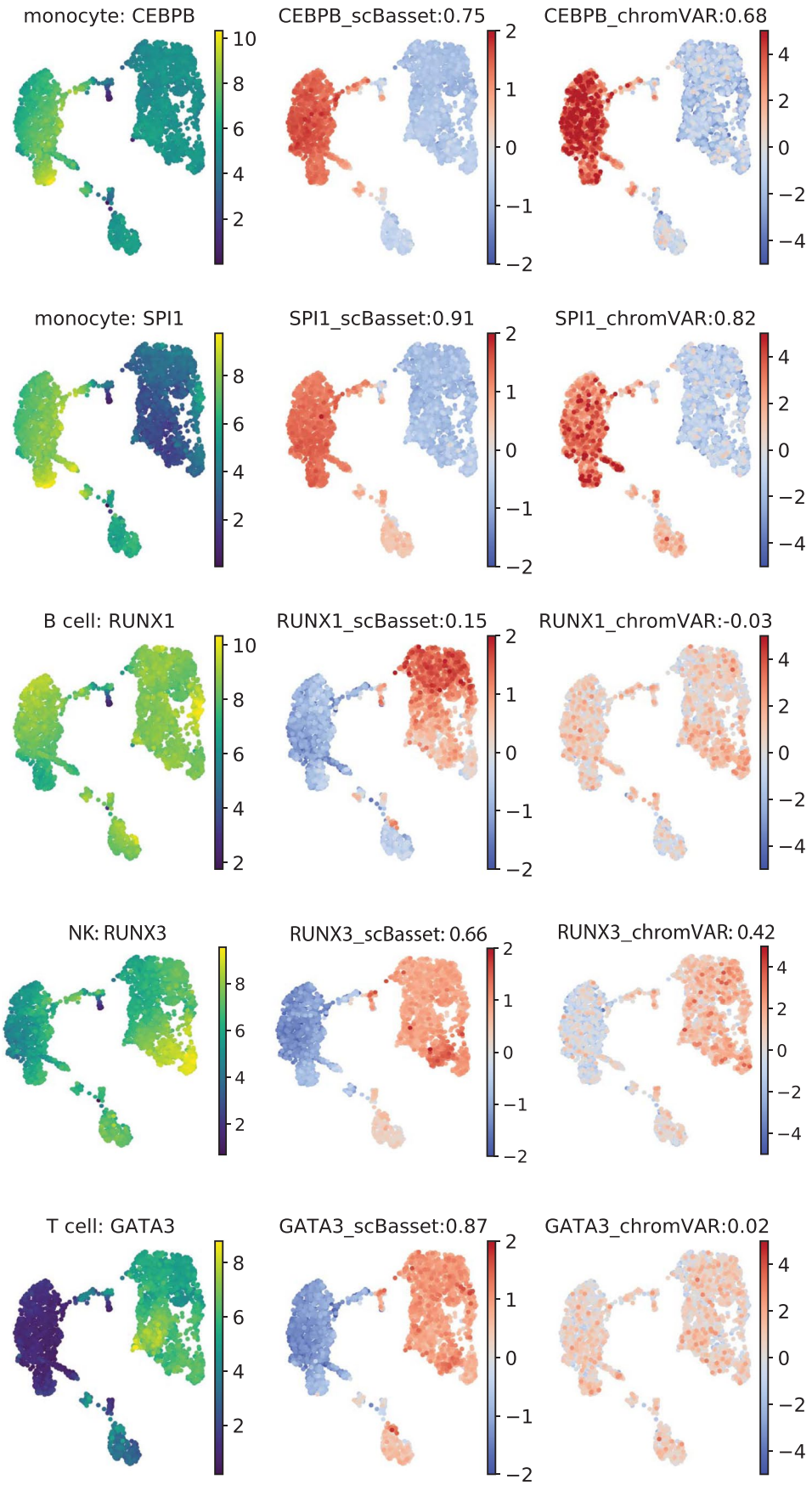
Extended Data Fig. 5 | Additional scBasset denoising and integration results. a) Comparison of different denoising methods in Buenrostro2018 dataset as evaluated by label score and cell type average Silhouette width (cell type ASW). **b**) Correlation between gene accessibility score and gene expression across genes for each cell before (x-axis) and after scBasset denoising (y-axis) for the multiome mouse brain dataset. Cells are colored by sequencing depth. **c**) Left, Comparison of different denoising methods in multiome mouse brain dataset as evaluated by label score and cell type ASW. Right, comparison of different denoising methods in multiome mouse brain dataset as evaluated by correlation between scVI-denoised RNA and denoised ATAC profiles across genes per cell (correlation per cell (RNA,ATAC)), and correlation between scVI-denoised RNA and denoised ATAC profiles across cells per gene (correlation per gene (RNA,ATAC)). **d**) Integration performance comparison in multiome 10x PBMC dataset when (i) both RNA and ATAC profiles are raw; (ii) only RNA profile is denoised with scVI; (iii) only ATAC profile is denoised with scBasset; and (iv) both RNA and ATAC profiles are denoised. n = 2714 cells for each boxplot. The boxplot shows min and max as whiskers (excluding outliers), 1st and 3rd quartiles as boxes and median in the center. Outliers (> 1.5x interquartile range away from the box) are not shown. **e**) UMAPs of RNA and ATAC co-embedding after integration for 10x multiome mouse brain dataset. Left, integration performed on RNA (blue) and raw ATAC (red) profile embeddings. Right, integration performed on RNA (blue) and scBasset-denoised ATAC (red).



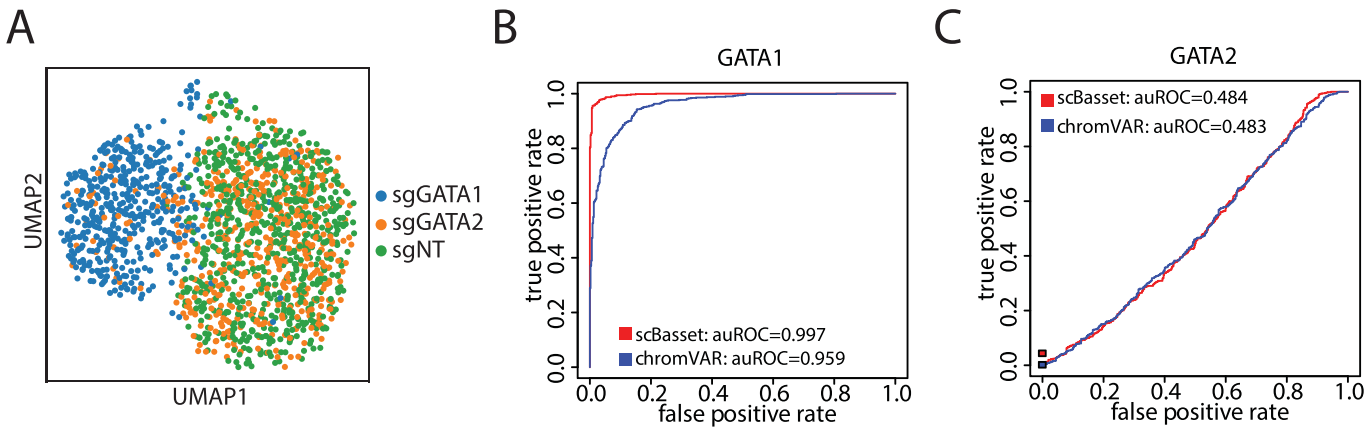
Extended Data Fig. 6 | Motif activity inference using scBasset and chromVAR on Buenrostro2018 dataset. a) UMAPs showing scBasset-predicted TF activity. **b**) Boxplots showing scBasset-predicted TF activity by cell type. **c**) UMAPs showing chromVAR-predicted TF activity. **d**) Boxplots showing chromVAR-predicted TF activity per cell type. For boxplots in B) and D), n = 502, 402, 344, 160, 142, 141, 138, 78, 64, 60 cells for each of CMP, GMP, HSC, LMPP, MPP, pDC, MEP, CLP, mono, UNK cell types. The box extends from the first quartile (Q1) to the third quartile (Q3) of the data, with a line at the median. The whiskers extend from the box by 1.5x the interquartile range (IQR). Outliers are not shown.



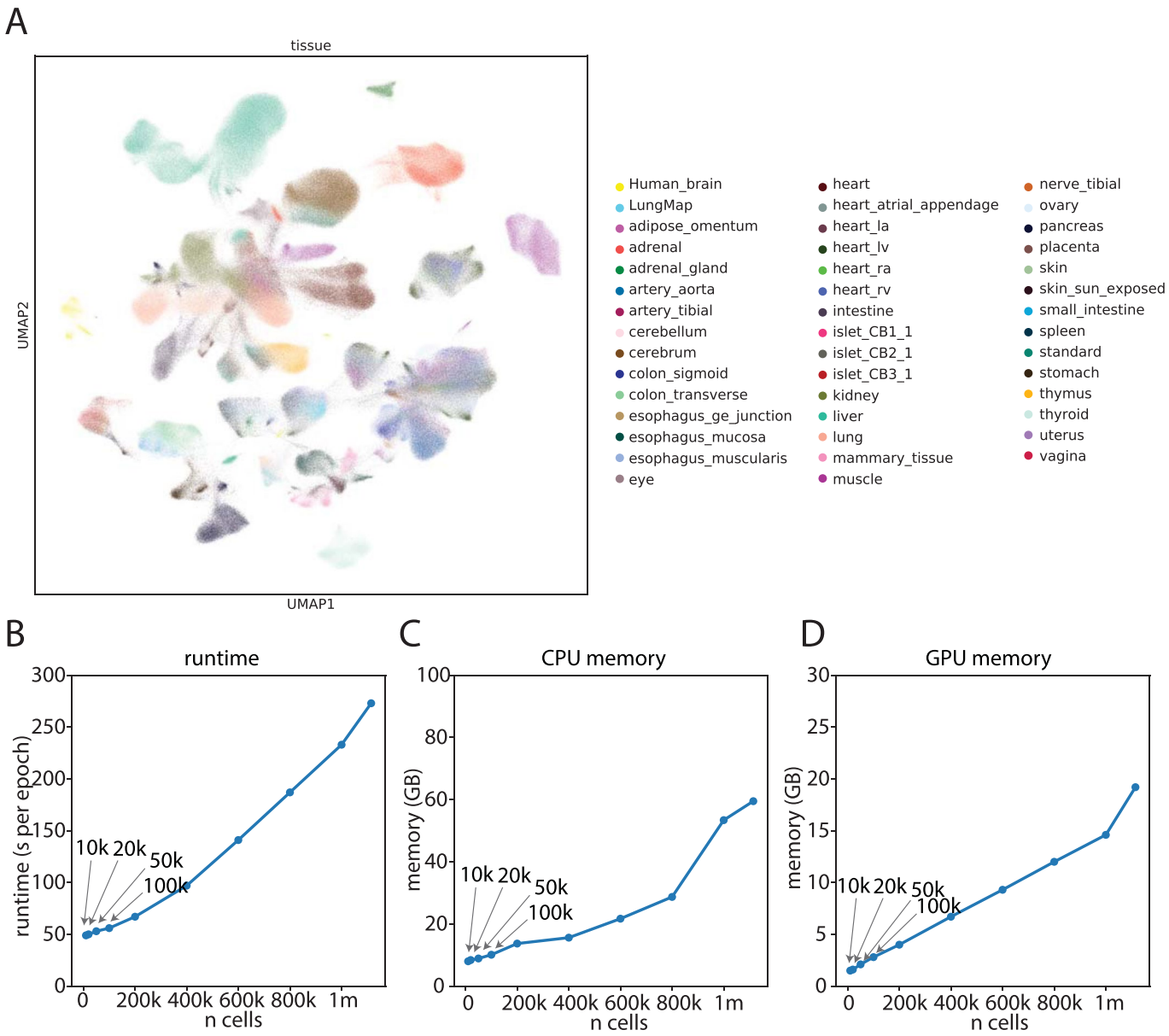
Extended Data Fig. 7 | TF expression and activity correlation for the 10x multiome datasets. Scatterplots of correlations between chromVAR-inferred activity and expression (x-axis) versus correlations of scBasset-inferred TF activity and expression (y-axis) for activating TFs (left) and repressive TFs (right) in the 10x multiome PBMC (top) and 10x multiome mouse brain (bottom). Activating TFs are TFs for which both scBasset and chromVAR agree on a positive correlation between TF expression and activity. Repressive TFs are TFs for which both scBasset and chromVAR agree on a negative correlation between TF expression and activity. A one-sided Wilcoxon signed rank test was performed.



Extended Data Fig. 8 | Motif activity inference using scBasset and chromVAR on 10x multiome PBMC dataset. UMAP visualization of TF expression (left), scBasset TF activity (middle), and chromVAR TF activity (right) for additional known PBMC regulators. Pearson correlation between inferred TF activity and expression are shown in the titles.



Extended Data Fig. 9 | Comparison of scBasset and chromVAR in TF perturbation experiments. a) Embeddings of K562 cells transfected by a pool of 9 CRISPRi sgRNAs targeting GATA1 (sgGATA1) and GATA2 (sgGATA2) and 9 inert sgRNA controls (sgNT). **b**) Performance comparison of scBasset and chromVAR in distinguishing sgGATA1 cells from sgNT cells in ROC curves. **c**) Performance comparison of scBasset and chromVAR in distinguishing sgGATA2 cells from sgNT cells in ROC curves.



Extended Data Fig. 10 | scBasset result on sci-ATAC human cell atlas. a) UMAP of sci-ATAC human cell atlas. Cells colored by tissue of origin. **b-d**) Runtime, peak CPU memory and GPU memory usage of scBasset as a function of the number of cells in the dataset.

nature research

Corresponding author(s): Han Yuan, David R Kelley

Last updated by author(s): May 19, 2022

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	Trained model, additional documentation, and code for training and predicting with scBasset available at: https://github.com/calico/scBasset For data analysis, we used the following softwares: scikit-learn python package (v1.0), Cicero (v1.4.4), cisTopic (v0.3.0), SCALE command line tool, ChromVAR (v1.8.0), ArchR (v1.0.1), snaptools (v1.4.8), snapATAC (v1.0.0) , scvi-tools (v0.14.5), scDEC, magic (v3.0.0), scopen (v1.0.1), harmony (v0.1.0), kBET R package (v0.99.6) , lisi R package (v1.0), and FIMO.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

We used only public data sets in this study. We downloaded the processed peak set for Buenrostro2018 generated by Chen et. al at https://github.com/pinellolab/scATAC-benchmarking/blob/master/Real_Data/Buenrostro_2018/input/combined.sorted.merged.bed. We downloaded the aligned bam files from https://github.com/pinellolab/scATAC-benchmarking/tree/master/Real_Data/Buenrostro_2018/input/sc-bams_nodup. The original datasets is from accession: GSE96769.

We downloaded the 10x multiome datasets from 10x Genomics: https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/2.0.0/pbmc_granulocyte_sorted_3k for PBMC dataset, and https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/2.0.0/e18_mouse_brain_fresh_5k for mouse brain dataset.

We downloaded the processed peak by cell matrix from sci-ATAC human atlas (accession: GSE184461) stored here http://renlab.sdsc.edu/kai/Key_Processed_Data/Cell_by_cCRE/.

spear-ATAC preprocessed count matrix "K562-Pilot-scATAC-Peak-Matrix-SE.rds" is downloaded from GEO (accession: GSE168851).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical method was used to predetermine sample size. Sample size is the number of cells in each public scATAC dataset. Buenrostro2018 dataset: 2034 cells. 10x multiome PBMC dataset: 2714 cells. 10x multiome mouse brain dataset: 4881 cells. sci-ATAC dataset: 1,114,621 cells.
Data exclusions	Low quality peaks and cells were filtered out. For Buenrostro2018 dataset, Peaks accessible in less than 1% cells were filtered out. For multiome dataset, genes expressed in less than 5% cell were filtered out, and peaks accessible in less than 5% cells were filtered out.
Replication	We compared scBasset to other methods on three independent datasets. We also showed that scBasset can be trained with scATAC data of different level of sparsity, and consistently achieve high performance.
Randomization	Randomization is not relevant. Each analysis is independently performed on a different dataset.
Blinding	Blinding is not relevant. scBasset model is trained using only sequence information. Any phenotypic information (labels for the cells) are not used during training, and only used for validation.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging