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

Monalika Padma Reddy

December 2022

1 Introduction

ATAC-seq (Assay for Transposase-Accessible Chromatin using Sequencing) is a method used in molecular biology used to evaluate the accessibility of chromatin across the entire genome. Single cell ATAC-seq (scATAC) holds enormous promise for examining cellular heterogeneity in epigenetic landscapes, but because to its high dimensionality and sparsity, it still presents substantial processing hurdles. Epigenetic landscapes at the single cell level are revealed using single cell ATAC-seq (scATAC). The assay has been effectively used to decipher cellular heterogeneity, map disease-associated distal elements, and reconstruct differentiation trajectories in addition to identifying different cell types and their distinct regulatory components.

Due to the intrinsic high dimensionality of accessible peaks and the sparsity of sequencing reads per cell, there are still substantial obstacles in the processing of scATAC data. To overcome these issues, numerous strategies have been put forth, which can be broadly divided into two classes: sequence-free methods and sequence-dependent methods. The majority of approaches express these identified peaks as genomic coordinates without taking into account the underlying DNA sequence, starting from a sparse peak-by-cell matrix produced by read aggregation and peak calling in exposed chromatin.

The peak-by-cell matrix is linearly transformed by latent semantic indexing (LSI) and principal component analysis (PCA) to project the cells to a low-dimensional space. SCALE and cisTopic use latent dirichlet allocation or a variational autoencoder to simulate the generative process of the data distribution. These sequence-free techniques can identify biologically significant covariance to accurately characterize, cluster, or categorize cells. To connect accessibility to transcription factors, they neglect sequence information and rely on post-hoc motif matching methods (TFs). The TF motif or k-mer content of peaks is represented by sequence-dependent algorithms like chromVAR and BROCKMAN, which aggregate these properties across peaks or other regions of interest to develop cell representations. While chromVAR emphasizes interpretability by

directly associating peaks to TFs, it tends to perform poorly when learning cell representations, possibly because information from its straightforward implicit model connecting sequence to accessibility through position weight matrices was lost.

Deep convolutional neural networks (CNNs)-based expressive model that is sequence dependent is proposed. DeepSEA and Basset serve as examples of how CNNs outperform k-mer or TF motif models in peak prediction from bulk chromatin profiling tests. Through the use of convolutional layers, these models compute explicit embeddings of the sequences beneath the peaks as well as implicit embeddings of the various "tasks" in the parameters of the final linear transformation. In order to predict single cell chromatin accessibility from sequences, Basset architecture is enhanced. A bottleneck layer is used to learn low dimensional representations of the individual cells. The model outperforms state-of-the art methods for cell representation learning, single cell accessibility denoising, scATAC integration with scRNA, and transcription factor activity inference by utilizing sequence information in a deep learning framework.

There are various drawbacks to sequence-based techniques. The reference genome is utilized, but many samples will have variable versions, including copy number variations, which may cause our models to fail. Second, the regulatory motifs and their interactions are believed to be generalizable across the genome. This assumption may not be totally correct at some genomic loci where evolution has resulted in specialized regulatory solutions, such as X chromosome inactivation in females. However, because scBasset approaches covariance-based methods fully independently, these two approaches can be coupled to better their analysis.

2 Methods

The scBasset, which is a sequence-based deep learning framework for modeling scATAC data, is proposed in this work.

2.1 Dataset

i) scATAC-seq preprocessing - The Buenrostro2018 dataset's count matrix and peak atlas files were downloaded from GEO(Accession GSE96769). Peaks accessible less than 1 percent are eliminated. The total dataset has 2,034 cells and 126,719 peaks. For the PBMC dataset and mouse brain dataset, the 10x multiome files were retrieved from 10x Genomics. Genes that were expressed in fewer than 5 percent of cells were eliminated. Peaks that could be accessed by fewer than 5 percent of cells were removed.

ii) scRNA-seq preprocessing - The expression data for the 10x multiome datasets was preprocessed with scVI version 0.6.5 with n layers=1, n hidden=768, latent=64, and a dropout rate of 0.2. scVI was trained for 1000 epochs using a learning rate of 0.001 and the options lr patience of 20 and lr factor of 0.1 to reduce the learning rate upon plateau. When the ELBO loss

did not improve after 40 epochs, early halting was enabled. The get sample scale() method was used to sample from the generative model ten times and take the average to obtain denoised expression profiles. The learned latent cell representations are utilized to construct nearest neighbor graphs and cluster cells.

2.2 Model architecture

Based on each peak's DNA sequence, the neural network architecture known as scBasset predicts binary accessibility vectors for each peak. A 1344 bp DNA sequence from the center of each peak is provided as input to scBasset, which one-hot encodes it as a 13444 matrix. The scBasset architecture includes the following blocks -

i) A 1D convolution layer with 288 filters of size 17x4 is followed by layers of batch normalization, Gaussian error linear unit (GELU), and width 3 max pooling, producing an output matrix of 488x288.

ii) 6 convolution blocks make up the convolution tower, each with layers for convolution, batch normalization, maximum pooling, and GELU. The convolution layers contain a kernel width of 5 and an increasing number of filters (288, 323, 363, 407, and 512). A 7x512 matrix is the convolution tower's output.

iii) A 1D convolution layer with 256 filters of kernel width 1, is followed by batch normalization and GELU is applied. The result is a 7x256 matrix that is flattened into a 11792 vector.

iv) Following batch normalization, dropout (rate=0.2), and GELU, there is a dense bottleneck layer of 32 units. The result is a 128-bit compressed peak representation vector. The final dense layer predicts continuous accessibility logits for each cell's peaks.

v) Batch correction is optional and is accomplished by attaching a second parallel dense layer to the bottleneck layer and anticipating batch-specific accessibility. To compute the batch contribution to accessibility in each cell, this batch-specific accessibility is multiplied by the batch-by-cell matrix. This vector is then added to the preceding per-cell continuous accessibility logits. To tailor the contribution of the batch covariate to the predictions, L2 regularization can be performed to the cell-embedding path or the batch-specific path (with hyperparameter).

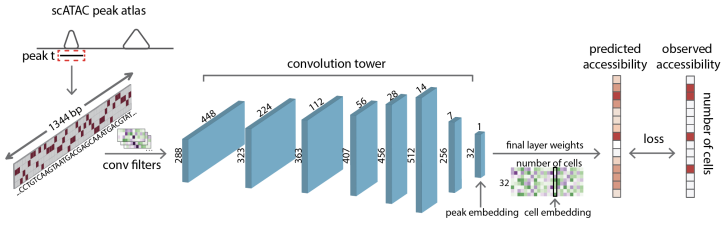

Figure 1: Architecture of scBasset

The model's total number of trainable parameters is proportional to the number of cells in the dataset. The model's trainable parameters is 4513960+33n. The model parameters were updated using the Adam update approach and stochastic gradient descent. For the Adam optimizer, a random search was done for optimal hyperparameters such as batch size, learning rate, beta1, and beta2. The best results were obtained with a batch size of 128, a learning rate of 0.01, a beta 1 of 0.95, and a beta 2 of 0.9995.

3 Discussion

The scBasset is trained to predict individual cell accessibility from the DNA sequence underlying ATAC peaks, learning a vector embedding in the process to represent the single cells. A trained scBasset model can improve numerous lines of scATAC analysis, and we show cutting-edge performance on several tasks. The model's cell embeddings are better aligned with ground-truth cell type labels when they are clustered. The model outputs can be used to generate denoised accessibility profiles, which increase RNA measurement concordance. The model gains knowledge of TF motifs and their impact on accessibility.

The scBasset model can also anticipate the effect of mutations, allowing for in silico saturation mutagenesis of regulatory sequences of interest at single cell resolution. scBasset outperforms earlier sequence-based techniques for scATAC analysis, such as chromVAR, in learning cell embeddings and inferring TF activity because it uses a more expressive CNN model that learns more nuanced sequence properties, including as non-linear connections. scBasset outperforms earlier sequence-free approaches such as cisTopic and SCALE in benchmarking tasks and produces a more interpretable model that can be directly queried for TF activity or finding regulatory sequences.

4 Results

4.1 scBasset predicts single-cell chromatin accessibility based on held-out peaks

scBasset takes a 1344 bp DNA sequence from the center of each peak as input and one-hot encodes it into a 4x1344 matrix. The input DNA sequence goes through eight convolution blocks, each with a 1D convolution, batch normalization, max pooling, and GELU activation layer. These are followed by a bottleneck layer, unlike most previous systems. This is designed to learn a low-dimensional representation of the peak from the layer output and the cells from the subsequent layer's parameters. Finally, a dense linear transformation connects the bottleneck sequence embeddings to predict binary accessibility in each cell.

The scATAC-seq approach is applied to 3 datasets - Buenrostro2018 with 2k cells, 10x Multiome RNA+ATAC PBMC dataset with 3k cells, and 10x Multiome RNA+ATAC mouse brain.

The auROC across peaks is computed for each cell and averaged across cells for held out peaks. auROC across cells is computed for each peak and averaged across peaks to assess cell type specificity. scBasset obtained impressive accuracy values indicative of successful learning: 0.734 per peak and 0.740 per cell for the Buenrostro2018 dataset, 0.734 per peak and 0.701 per cell for the 10x multiome mouse brain dataset and 0.662 per peak and 0.640 per cell for the 10x multiome PBMC dataset. Although these data are marginally below the 0.75-0.95 range observed for bulk DNase samples in the original Basset publication, this is predictable given to the substantially higher measurement noise due to sparse sequencing for the single cell assay.

Dataset	Per Peak	Per Cell
Buenrostro2018	0.734	0.740
10x multiome PBMC	0.662	0.640
10x multiome mouse brain	0.734	0.701

Figure 2: Accuracy of scBasset model

4.2 scBasset final layer learns cell representations

The cell embeddings for the Buenrostro2018 dataset are shown in 2D using t-distributed stochastic neighbor embedding (t-SNE) and observed differentiation trajectories in the t-SNE space. When compared to other prominent approaches for scATAC embedding, chromVAR and PCA have trouble discriminating CLP from LMPP, whereas Cicero, SCALE, cisTopic, and scBasset do. Following earlier work, the correctness of cell embeddings is evaluated by comparing Louvain clustering results with ground-truth cell type labels using the modified rank index (ARI).

According to this criterion, scBasset outperforms the other approaches. Label scores are generated across a range of neighborhoods for each embedding approach, and observed scBasset regularly outperforms competitors in learning cell representations that embed cells of the same type near one other. The label scores for cell embeddings are derived for the multiome PBMC and mouse brain datasets. Because the ground-truth cell types for the multiome datasets are unknown, scRNAseq Leiden clustering cluster identifiers are utilized as cell type labels. scBasset once again leads the competition in this statistic across a variety of neighborhoods. When tested with neighbor scores across a range of neighborhoods, scBasset surpasses the competition on both multiome PBMC and multiome mouse brain datasets.

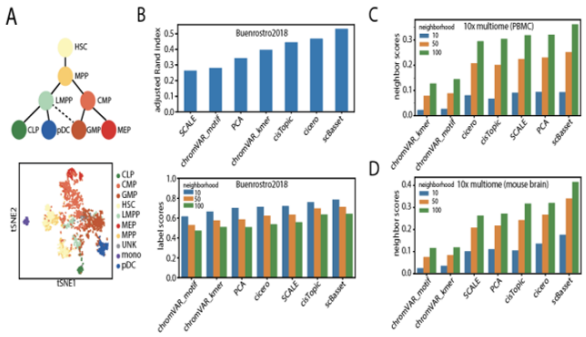

Figure 2: scBasset performance at learning cell representations. A) Top, hematopoietic stem cell differentiation lineage diagram in the Buenrostro2018 study; bottom, t-SNE visualization of cell embeddings learned by scBasset, colored by cell types. B) Top, performance comparison of different cell embedding methods evaluated by adjusted Rand index; bottom, performance comparison of different cell embedding methods evaluated by label score (Methods). C) Performance comparison of different cell embedding methods evaluated by neighbor scores for the 10x multiome PBMC dataset. D) Performance comparison of different cell embedding methods evaluated by neighbor scores for the 10x multiome mouse brain dataset.

4.3 scBasset denoises single cell accessibility profiles

Because of the sparsity of scATAC, the binary accessibility indication for each given cell and peak contains a large number of false negatives, hence the data cannot be investigated with real single cell precision and must be aggregated across cells. Several techniques, however, provide denoised numeric values to indicate the accessibility status at each cell/peak combination. In its sequence-based predictions, scBasset computes such values. The raw cell-by-peak matrix against the denoised matrix were directly shown using 500 peaks and 200 cells from the Buenrostro2018 dataset. In the raw count matrix, it was discovered that cells and peaks clustered by sequencing depth, indicating no significant patterns. However, after scBasset denoising, it was shown that cells of the same cell type have comparable accessibility profiles, and hierarchical cell clustering correlated well with ground-truth labels.

For both the 10x multiome PBMC and mouse brain datasets, it was observed that scBasset denoising enhances the consistency between gene accessibility and expression. For consistency between differential expression and differential accessibility, scBasset and SCALE were examined for accessibility denoising. Differential expression and accessibility studies were done versus the rest of the cells for each cell type cluster defined by scRNA in the 10x PBMC dataset. Surprisingly, while SCALE exceeds scBasset in terms of baseline accessibility/expression, scBasset greatly outperforms SCALE in terms of differential accessibility/expression. Because each peak is only considered during its sequence, scBasset will be less prone to over-smoothing. As a result,

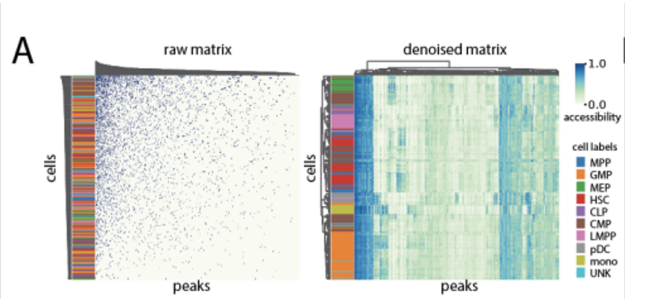

Figure 4: scBasset denoising performance. A) Left, binary count matrix of 200 cells and 500 peaks sampled from Buenrostro2018 dataset, hierarchically clustered by both cells and peaks. Cell type labels annotate the rows. Right, the same matrix and procedure after scBasset denoising. B) Correlation between

SCALE outperforms scBasset in terms of baseline accessibility, whereas scBasset outperforms SCALE in terms of differential accessibility, which emphasizes cell identity.

4.4 scBasset infers transcription factor activity at single cell resolution

Transcription factor binding is a significant determinant of chromatin accessibility. The scBasset model is anticipated to capture sequence information predictive of TF binding, which learns to predict accessibility from sequence. A trained scBasset model is given synthetic DNA sequences with and without a certain TF motif of interest. The motif's activity in each cell is assessed based on changes in anticipated accessibility. The 10x PBMC multiome dataset is used to compare scBasset and chromVAR. TF expression in RNA can be used to predict the activity of its motif. Using scBasset and chromVAR, motif activity for all 733 human CIS-BP motifs is inferred. scBasset TF activities correlate with expression considerably better than chromVAR. scBasset and chromVAR were tested individually on activating and repressive TFs.

scBasset predicted TF activities with considerably better correlation with expression than chromVAR predicted activity for positive TF expression-activity correlation. In terms of negative TF expression-activity correlation, scBasset projected TF activities have a much lower (more negative) correlation than chromVAR predicted activity.

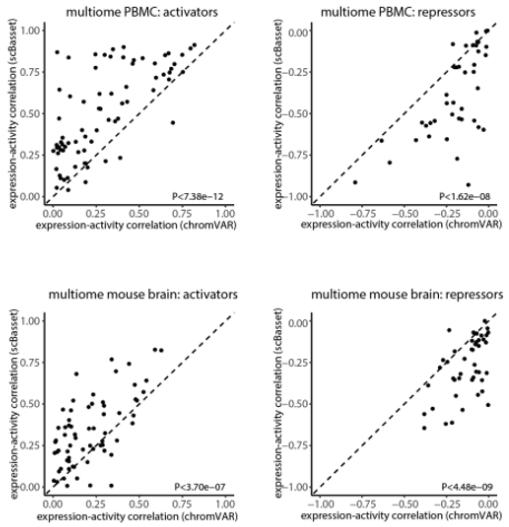

Figure S10: TF expression and TF activity correlation for the 10x mulitome datasets. Scatterplots of correlations between chromVAR-inferred activity and expression (x-axis) versus correlations of scBasset-inferred TF activity and expression (y-axis) for activating TFs (left) and repressive TFs (right) in the 10x multiome PBMC (top) and 10x multiome mouse brain (bottom). Activating TFs are TFs which both scBasset and chromVAR agree on a positive correlation between TF expression and activity. Repressive TFs are TFs which both scBasset and chromVAR agree on a negative correlation between TF expression and activity.