JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 29, Number 7, 2022

Pp. 619–633
DOI: 10.1089/cmb.2021.0596

Research Articles

Open camera or QR reader and scan code to access this article and other resources online.

Translator: A *Trans*fer *L*earning Approach to Facilitate Single-Cell *AT*AC-Seq Data Analysis fr*o*m *R*eference Dataset

SIWEI XU,[1] MARIO SKARICA,[2] AHYEON HWANG,[4] YI DAI,[1] CHEYU LEE,[1] MATTHEW J. GIRGENTI,[3,5] and JING ZHANG[1]

ABSTRACT

Recent advances in single-cell sequencing assay for transposase-accessible chromatin (scATAC-seq) have allowed simultaneous epigenetic profiling over thousands of individual cells to dissect the cellular heterogeneity and elucidate regulatory mechanisms at the finest possible resolution. However, scATAC-seq is challenging to model computationally due to the ultra-high dimensionality, low signal-to-noise ratio, complex feature interactions, and high vulnerability to various confounding factors. In this study, we present Translator, an efficient transfer learning approach to capture generalizable chromatin interactions from high-quality (HQ) reference scATAC-seq data to obtain robust cell representations in low-to-moderate quality target scATAC-seq data. We applied Translator on various simulated and real scATAC-seq datasets and demonstrated that Translator could learn more biologically meaningful cell representations than other methods by incorporating information learned from the reference data, thus facilitating various downstream analyses such as clustering and motif enrichment measurements.

Moreover, Translator's block-wise deep learning framework can handle nonlinear relationships with restricted connections using fewer parameters to boost computational efficiency through Graphics Processing Unit (GPU) parallelism. Finally, we have implemented Translator as a free software package available for the community to leverage large-scale, HQ reference data to study target scATAC-seq data.

Keywords: deep generative model, single-cell ATAC-seq, transfer learning, variational autoencoder.

[1]Department of Computer Science, University of California, Irvine, California, USA.
Departments of [2]Neuroscience and [3]Psychiatry, School of Medicine, Yale University, New Haven, Connecticut, USA.
[4]Mathematical, Computational, and Systems Biology, University of California, Irvine, California, USA.
[5]Clinical Neurosciences Division, National Center for PTSD U.S. Department of Veterans Affairs, Washington, DC, USA.

1. INTRODUCTION

Recent advances in single-cell epigenetic sequencing technologies, especially assay for transposase-accessible chromatin using sequencing (ATAC-seq), have allowed scientists to probe accessible genome-wide chromatin in individual cells (Buenrostro et al., 2015; Cusanovich et al., 2015; Chen et al., 2018; Chen et al., 2019; Satpathy et al., 2019; Yan et al., 2020). Due to its high-throughput capacities, single-cell sequencing assay for transposase-accessible chromatin (scATAC-seq) has been widely used in many laboratories and pioneering consortia to characterize epigenetic heterogeneity and decipher cell type-specific regulatory grammar, especially because these open chromatin regions often host complex genomic interplay among numerous cis-regulatory elements (CREs), transcription factors (TFs), cofactors, and chromatin remodelers in the three-dimensional (3D) genome for precise spatiotemporal gene expression control (Boyle et al., 2008; Tsompana and Buck, 2014; Klemm et al., 2019; Zhou et al., 2021).

It is essential to develop robust, accurate, and scalable computational methods for scATAC-seq data analysis. However, scATAC-seq computational modeling is still in its infancy due to five major challenges: (1) ultra-high dimensionality and sparsity (i.e., zero entries in the data matrix); (2) complex feature interactions; (3) vulnerability to confounding factors (e.g., age, gender, condition, depth, batch); (4) scalability to millions of cells; and (5) low signal-to-noise ratio (SNR), especially in low-to-moderate quality datasets reflected by their higher fraction of background genomic reads. Researchers have developed numerous computational methods to address these challenges, each with their own advantages and disadvantages. For instance, ChromVAR (Schep et al., 2017) groups all peaks within a cell to calculate known motif enrichments, effectively handling data sparsity. However, it relies on incomplete motif patterns and ignores the impact of individual peaks, resulting in suboptimal clustering results. Later on, latent semantic indexing (LSI) was developed and used to project cells onto a lower dimensional space with improved clustering performance (Cusanovich et al., 2015).

However, its inherent linear peak interaction assumption might not be able to handle complex chromatin interactions in the 3D genome. SnapATAC uses Jaccard distance to measure robust cell similarities, but a hidden assumption is that peaks independently and equally contribute to the cell-to-cell similarities (Buenrostro et al., 2015; Fang et al., 2021). RA3 utilizes probabilistic Principal Component Analysis (PCA) (Tipping and Bishop, 1999) to compute a robust cell representation with the help of a reference data, but it was inherently a statistical model (Chen et al., 2021a,b). Recently, several deep learning-based models, such as SCALE (Xiong et al., 2019) and SAILER (Cao et al., 2021), have been developed using variational autoencoders (VAEs) to learn robust cell representations for various downstream analyses. However, their fully connected neural network architecture can contain up to a million parameters, resulting in tremendous difficulties when training low-to-moderate-quality scATAC-seq datasets with only a few cells. Moreover, some more recent approaches (Bravo Gonzalez-Blas et al., 2019; Chen et al., 2022; Huang et al., 2018; Ji et al., 2020; Liu et al., 2021) utilized deep generative model and statistical modeling for scATAC-seq data, but there was little discussion on the effect of confounding factors.

In this study, we developed an efficient scATAC-seq analysis method, Translator, with a straightforward intuition: cell type-specific feature interactions (e.g., peak-to-peak) are highly stable across different samples within the same cell type and can be learned in a generalizable fashion from a high-quality (HQ) reference dataset to facilitate cell representation learning in a different low-to-moderate-quality target dataset. Specifically, Translator uses a VAE to model nonlinear feature interactions in scATAC-seq, with restrictions and modifications to the fully connected neural networks for improved computational efficiency. Furthermore, using a reference dataset with clear signals and good read depths, we conducted joint training with the VAE deep neural network to create a better embedding for the target dataset using the information obtained from the reference dataset.

We tested Translator with various datasets, including simulated data from the SCAN-ATAC-sim simulator (Chen et al., 2021c) and real datasets from the peripheral blood mononuclear cells (PBMC) and prefrontal cortex (PFC) cells. We demonstrated that Translator outperformed methods without transfer learning in clustering performance, qualitatively by Uniform Manifold Approximation and Projection (UMAP) visualization and quantitatively by silhouette score and adjusted rand index (ARI). In addition, we showed its ability to capture important biological information by conducting cell type annotation using marker genes and cell type-specific motif enrichment analysis.

2. METHODS

2.1. Datasets

We evaluated the performance of our method Translator on (1) simulated data with ground-truth labels; (2) public PBMC scATAC-seq data from 10x Genomics; (3) prefrontal cortex scATAC-seq data from postmortem human brains; and (iv) mice skin tissue scATAC-seq data from the Share-seq dataset. We discuss more on data simulation and preprocessing in the following sections.

2.1.1. scATAC-seq data simulation with ground truth labels. Due to the challenge in obtaining gold-standard datasets, we first used a public software SCAN-ATAC-sim to generate labeled scATAC-seq data by efficiently down sampling bulk ATAC-seq profiles from ENCODE cell lines (Chen et al., 2021c). By tuning two key parameters: SNR and average fragment number per cell (μ), as suggested by the SCAN-ATAC-sim website, we simulated two kinds of scATAC-seq data to mimic the HQ reference and the low-quality (LQ) target data scenarios. We repeated this procedure for five different cell types, MONO, NEU, CMP, MEGA, and ERY, from the ENCODE data portal (Consortium, 2012; Davis et al., 2018) and generated the signal tracks using the SCAN-ATAC-sim preprocessor.

To test Translator's robustness, we evaluated four possibilities: fixed depth, varying depth, mismatched cells with new cell types, and mismatched cells with missing cell types (Table 1). In addition, to evaluate the performance of eliminating confounding factors of Translator's invariant learning, we applied two experiments: one using one sample ($\mu = 5000,\ \sigma = 1.5,\ \rho = 0.4$) to test the effectiveness of removing sequencing depths; another using two samples ($\mu_1 = 2500,\ \sigma_1 = 1.5,\ \rho_1 = 0.4, \mu_2 = 5000,\ \sigma_2 = 1.5,\ \rho_2 = 0.5$) to test the performance under two samples.

2.1.2. PBMC scATAC-seq data preprocessing. To test Translator's performance on real single-cell datasets, we downloaded the fragment files of the PBMC multiome dataset publicly available from 10x Genomics. We used the ATAC modality for performance benchmarking and the RNA modality for more accurate labeling due to its higher quality. To mimic the scenario of datasets with different qualities, we took the PBMC multiome 10k dataset as the reference and the down-sampled PBMC multiome 3k dataset as the target dataset. The sequencing depth and other quality control (QC) parameters are provided in Supplementary Figure S1 and S2.

Specifically, we down-sampled the target dataset using the following algorithm. Given the sequencing depth of the raw target dataset d_r and the desired sequencing depth $d < d_r$, we calculated $r = \frac{(d_r - d)}{d_r}$, which is the proportion of the peaks needed to be deactivated. To avoid samples with extremely low sequencing depth, we set 1000 fragments as the lower bound depth. For each cell i, where n_i is the number of open chromatin peaks, we randomly sampled $\min(rn_i,\ n_i - 1000)$ peaks to be deactivated by setting their corresponding positions in the peak matrix as zero. Further iterations were conducted until the sequencing depth reached the desired sequencing depth $d_r \pm 50$.

We then used ArchR (Cao et al., 2021; Granja et al., 2021) (version 1.0.1) to preprocess the fragment file of the reference data using the default parameters. Specifically, we created an ArchR object and removed barcodes with a transcription starting site (TSS) enrichment score less than 4 or whose number of fragments

Table 1. Detailed Parameters Setting to Simulate Single-Cell Sequencing Assay for Transposase-Accessible Chromatin Data

Scenario	*Data*	*Depth*	*SNR*	*No. of cells*	*Cell proportion*
Fixed depth	HD	3k	0.8	20k	4k × 5
	LD	3k	0.2, 0.25, 0.3, 0.35	500	100 × 5
Varying depth	HD	$\mu = 10$k, $\sigma = 1.5$	0.8	20k	4k × 5
	LD	$\mu = 1.5$k, 2k, 2.5k, $\sigma = 1.5$	0.3, 0.45, 0.6	500	100 × 5
New cell type	HD	3k	0.8	20k	7k, 6k, 4k, 3k, 0
	LD	3k	0.3	500	100 × 5
Missing cell type	HD	3k	0.8	20k	4k × 5
	LD	3k	0.3	525	250, 100, 100, 75, 0

SNR, signal-to-noise ratio.

was less than 1000 (as suggested by ArchR Tutorial). We further removed doublets using ArchR's default filters, kept peaks within the autosome chromosomes, and then generated the reference data cell-by-peak matrix. Finally, we binarized the reference cell-by-peak matrix to set any value greater than 1 as 1. For the target data, Translator imported the reference peak files to create the target cell-by-peak matrix for the training process. To create benchmarking metrics for the target dataset, we used the latent semantic indexing (LSI) commonly utilized in natural language processing models as well as in scATAC-seq analyses (Cao et al., 2021; Stuart et al., 2021). Given the binarized target data, we utilized the RunLSI function from the R package Signac (Hao et al., 2021) (version 4.0.4) with default parameters and $n = 20$ to keep its latent dimension the same as Translator's.

2.1.3. Preprocessing scATAC-seq data in the human prefrontal cortex. In addition to blood cells, we also looked at the prefrontal cortex (PFC) to further evaluate our model (post-mortem tissues were used; no IRB applicable). Previously, we deeply sequenced scATAC-seq data from three frozen PFC tissues MS0169WW, MS0177EE, and MS0181II. We conducted the following protocol to generate scATAC-seq data from the prefrontal cortex tissues: single-nucleus suspensions were isolated from 25 mg of frozen human dorsolateral prefrontal cortex (Brodmann Area 9/46). Tissue was initially lysed and nuclei released by using a sucrose-based solution and NP-40 detergent (Sigma) in Dounce homogenizer followed by centrifugation at 1000 *g* for 10 minutes.

The nuclear pellet was resuspended and further purified using an Iodixanol (Optiprep, AxisShield) solution gradient by an additional centrifugation (3000 *g* for 30 minutes) and collection of nuclei in the interphase (30%/40%). Nuclei were washed and resuspended before running on the 10X Chromium Single Cell ATAC platform (#PN-1000110; 10x Genomics). Library quantification, quality checking, and sequencing (250 million reads) were done according to the manufacturer's recommendations by using the Illumina NovaSeq6000 (Illumina) flow cell S4 (Illumina) at the Yale Center for Genome Analysis.

Following the data generation, we used cell-ranger ATAC (version 2.0.0) to preprocess the raw reads with default parameters (Satpathy et al., 2019). Then, for quality control, cells with TSS enrichment score less than 4 or number of fragments less than 1000 were filtered out using ArchR. In addition, Harmony was used to integrate samples (Korsunsky et al., 2019). The sequencing depth and other QC parameters are provided in Supplementary Figures S3–S5. We performed dimensionality reduction, clustering, and cell type identification using default parameters in ArchR. Cells from these three samples were clearly clustered into biologically relevant cell types that matched known markers for neuronal and non-neuronal cell types, validating their high sequencing quality.

Additionally, a lower-quality PFC scATAC data was also collected in the same cohort, with moderate sequencing depth (2743.5), lower cell number (6285), and lower TSS enrichment than that of the reference dataset (4.35). Supplementary Figure S6 indicates the sample's moderate quality. Similar to the reference dataset, default parameters were used in cell-ranger (version 6.0.1) and ArchR to preprocess the lower-quality data.

For all PFC data, after the QC step, we added the reference peak set, which was the union of all three references' cell type-specific peak sets sorted by chromosomes. We generated the count matrix using the addPeakMatrix function from ArchR. Finally, we saved the matrix using getMatrixFromProject with useMatrix set to PeakMatrix and binarize set to True so that the exported matrix was binarized. Using the writeMM function in R allowed us to save the exported sparse matrix as a mtx file and then exported as a npz file usable in Python.

2.1.4. Data preprocessing of the Share-seq mouse skin scATAC-seq data. In addition, we further applied Translator on the mouse skin dataset from Share-seq dataset (Ma et al., 2020). We extracted the count matrix from its raw data using the default peak set provided with Signac. To construct the reference and the target data, we randomly sampled 1000 cells as the target data. The rest of them would be the reference data. We then conducted a feature selection workflow similar to SCALE (Xiong et al., 2019) by filtering out all peaks with a total signal less than 500 across the entire dataset. Then, we chose the top 80,000 peaks as inputs with the TF-IDF algorithm.

2.2. Model architecture and training

Translator uses a deep VAE to capture the complex and usually nonlinear feature interactions within the 3D genome. To better align multiple samples, it utilizes an invariant representation learning scheme for simultaneous bias removal by requiring the latent cell representations (z) to be independent of various

confounding factors (c), such as age, gender, batch, and sequencing depth. To overcome computational challenges of training fully connected VAEs with extremely large number of parameters, Translator uses a block-wise neural network architecture by only allowing interchromosome interactions in the first layer and then concatenates neurons in later layers with much smaller size, as shown in Figure 1. It is worth noting that due to the design of the Translator model, we require a consensus peak set across the reference and target data. We assume that the reference dataset has better quality than the target dataset, so we only used the reference data's peaks in the following analysis.

2.2.1. Deep VAE with invariant representation learning. To capture complex feature interactions and remove various confounding factors (e.g., age, gender, sequencing platforms, and batch), Translator first used a deep VAE with an invariant representation learning scheme to learn robust cell embeddings free from bias (Kingma and Welling, 2013). Specifically, we encouraged cell representations $z \in \mathbb{R}^d$ to only reflect intrinsic biological states but be independent of confounding factors c. Specifically, a penalty term was added in the VAE to minimize the mutual information $I(z, c)$. Given a peak profile $x \in 0, 1^n$, we used a VAE to handle complex peak interactions and maximize evidence lower bound of the log-likelihood

$$L_{\text{VAE}} = \mathbb{E}_{x,\, c \sim q(x,\, c)}[-\mathbb{E}_{z \sim q_\theta(z|x)}[\log p_\phi(x|z,\ c)] + D_{\text{KL}}(q\theta(z|x) \parallel p(z))] \tag{1}$$

In Equation (1), the encoder probability, denoted by $q_\theta(z|x)$, is the probability of the latent embedding z after the reparameterization trick given the input data x and the encoder parameter θ. The decoder probability $p_\phi(x|z,\ c)$ is the likelihood of the reconstructed input data x given the latent embedding z and other confounding factors c (see Table 3 for detailed parameterizations). Since z could potentially depend on c, we added an upper bounded penalty $I(z, c)$ to encourage the independence of z and c, where

$$I(z, c) \leq \mathbb{E}_{x,\, c \sim q(x,\, c)}[D_{\text{KL}}(q_\theta(z|x) \parallel q_\phi(z)) - \mathbb{E}_{z \sim q_\phi(z|x)}[\log p_\theta(x|z,\ c)]] - H(x|c) \tag{2}$$

where the $H(x|c)$ is a constant and can be removed from Equation (2). Additionally, in VAE training, a common challenge was Kullback-Leibler (KL) annealing, lowering $D_K L(\cdot)$ with respect to training epochs. Therefore, we applied the cyclic annealing schedule (Fu et al., 2019) by adding a term $w = \min(1, \frac{2(n \bmod N)}{N})$, where n is the current epoch and N is the number of epochs in a cycle. Therefore, the final loss function we aimed to minimize was

$$\begin{aligned} L_{VAE} + \lambda I(z, c) = {} & w\mathbb{E}_{x \sim q(x)}[D_{KL}(q_\theta(z|x) \parallel p(z)) + \lambda D_{KL}(q_\theta(z|x) \parallel q_\theta(z))] \\ & - (1 + \lambda)\mathbb{E}_{x,\, c \sim q(x,\, c)[\mathbb{E}_{z \sim q_\theta(z|x)}[\log p_\phi(x|z,\, c)]]} \end{aligned} \tag{3}$$

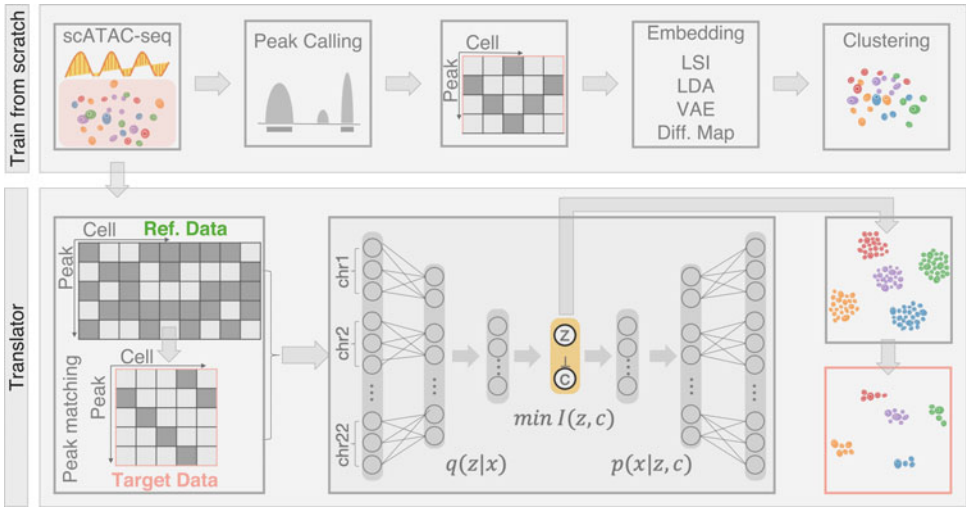

FIG. 1. Translator flowchart. scATAC-seq analysis is extremely challenging when sequencing depth, SNR, and cell number are limited. Different from traditional approaches of peak calling followed by dimension reduction from scratch, Translator uses peak features from a reference dataset and uses a jointed trained deep VAE to transfer complex feature interactions learnt from the high-quality data to facilitate robust cell embeddings. scATAC-seq, single-cell sequencing assay for transposase-accessible chromatin; SNR, signal-to-noise ratio; VAE, variational autoencoder.

2.2.2. Neural network architecture. A typical scATAC-seq data will generate at least 20k peaks as the input, so the first fully connected layer in the VAE model contributes the largest to the number of parameters used and usually requires a large number of cells for a stable training process. To reduce over-parameterization on data with a moderate cell number, we only allowed intrachromosomal peak-neuron connections in the first layer of the VAE with a straightforward intuition: CRE interactions usually happen within a single chromosome instead of across different chromosomes. As shown in Figure 1, each chromosome had its own fully connected network with 64–32 units. This way, the number of parameters in the first two neural network layers could be dramatically reduced to less than 10% of those in the fully connected layer, improving training efficiency and reducing the number of parameters trained to avoid potential overfitting. The 32-unit layers across all chromosomes were then concatenated, followed by a fully connected layer of 20 units, indicating the multivariate Gaussian mean and variance. The decoder was symmetric to the encoder.

2.2.3. Training procedure. Given a set of single-cell ATAC-seq data, we first sorted and grouped the peaks by their chromosome. Then, we fed the vertically stacked reference and target data into the encoder together in batches to obtain the learned mean and variance. The reparameterization step took the multivariate mean and variance from the encoder, generated a random sample, and fed them into the decoder. After the loss was computed, we conducted back propagation to update the parameters. Adam optimizer was used throughout the process. The main hyperparameters included learning rate *lr*, weight decay α, and the λ term in the loss function. To tune these parameters, for each training task, only the warmup procedure was done using parameters $lr \in \{1E-4, 5E{-}4, 1E{-}3, 2E{-}3\}$, $\alpha \in \{1E{-}5, 1E{-}4, 5E{-}4\}$, $\lambda \in \{0.1, 0.2, 0.5, 1, 2, 5, 10\}$. The set of hyperparameters that led to the best loss after the warmup was chosen to train the final model. All trainings were conducted using NVIDIA Tesla K80 GPUs.

2.2.4. Training from scratch using VAE. To evaluate the performance of Translator with and without the reference dataset, we adopted a baseline training scheme without the reference data, namely VAE-SCRATCH. In this protocol, only the target data were used in the training process. The training procedure was identical to the Translator training procedure. After training, we directly generated the latent embedding using the trained encoder with the target data.

2.3. Evaluation metric

To evaluate model performance, we employed qualitative and quantitative measurements. Qualitative analysis included visualizations of Translator-generated latent embeddings and comparisons with their counterparts, particularly in subsets of cells where traditional matrix decomposition methods were not possible. To extensively test our model, we compare Translator with: Signac (LSI), ArchR, and SCALE (Xiong et al., 2019; Cao et al., 2021; Stuart et al., 2021). From the learned cell representations, we generated the UMAP using the RunUMAP function from the Seurat package (version 4.0.4) with default parameters. Then, we visualized the UMAP projection with a scatterplot. On the other hand, quantitative metrics such as the ARI and the silhouette score were obtained to measure the clustering performance. The LEIDEN algorithm in Seurat was used to perform clustering, leading to the ARI. We use the default parameters for LEIDEN as described in the FindClusters function in Seurat package across all experiments. Also, the annotated clustering assignments were used to calculate the silhouette score. Finally, downstream analysis was conducted using the learned embeddings with biological annotations.

2.3.1. Motif analysis for PBMC dataset. To better understand the cells and different clusters of the PBMC dataset, we generated the motifs for all cell types using Translator's transferred latent embeddings. Since we borrowed the cell type labels from the RNA modality, we could compare and contrast generated motifs and potentially find new subtypes that enables further downstream analysis. Using the R package Signac (Stuart et al., 2020) (version 1.3.0), we first created a ChromatinAssay object, and then used the getMatrixSet function to get a set of motifs from the JASPAR2020 dataset (Fornes et al., 2020) with collection being CORE and tax group being vertebrates. Then, we used the AddMotifs function to add both the reference genome, UCSC hg38 (Schneider et al., 2017), and the motif to the object. Using the function runChromVAR in signac, we received a matrix of motif enrichment scores for each motif in the set and for each cell (Schep et al., 2017). Using the LEIDEN clusters calculated based on latent embeddings, we plotted and generated statistical metrics of the distribution of certain cell type-specific motifs for each cluster, allowing qualitative and quantitative comparisons.

2.3.2. Gene enrichment analysis and cell type annotation for the postmortem brain dataset. We also conducted gene enrichment analysis for the PFC scATAC-seq dataset to evaluate the clustering accuracy of our model. In the analysis, we clustered Translator's latent embedding with LEIDEN. Then, using the list of cell-type marker genes identified in previous work (Lake et al., 2016; Zonouski et al., 2019), we curated a gene score matrix with a score for each cell and each gene. More specifically, we used the following marker genes to determine the cell-type annotations—astrocytes: *GLUL*, *SOX9*, *AQP4*, *GJA1*, *NDRG2*, and *GFAP*; microglia: *MRC1*, *TMEM119*, and *CX3CR1*; endothelial: *CLDN5*; *OPCs*: *PDGFRA*, *PCDH15*, and *OLIG2*; oligodendrocytes: *PLP1*, *MAG*, and *MOG*; excitatory: *SATB2* and *SLC17A7*; and inhibitory: *GAD1* and *SLC32A1*. Finally, for each gene in every cluster, we inspected the gene enrichment score distribution to assign cell types. As a result, qualitative evaluation was done using overlapped UMAP visualization between the reference and target datasets. In addition, to evaluate the performance of Translator under similar or distinct data, we train the PFC target data with PBMC reference and conduct a UMAP and ARI analysis as well as comparison.

2.3.3. Reference and target analysis of the Share-seq dataset. We conducted multiple experiments to evaluate our model on Share-seq. First, we train Translator, Signac, ArchR, and SCALE (Xiong et al., 2019; Cao et al., 2021; Stuart et al., 2021) on the same preprocessed data as aforementioned, all with 20 dimensions. Note that because we have done feature selection, we disabled the feature selection function on SCALE. As a quantitative measurement, UMAP was generated using the same methods for both reference and target data. Quantitative measurements include silhouette scores and ARI, calculated consistent with the other datasets.

3. RESULTS

Recent advances in single-cell epigenetic sequencing, especially the scATAC-seq technology (Buenrostro et al., 2015; Pott and Lieb, 2015; Chen et al., 2018; Xu et al., 2021), has allowed for parallel epigenetic profiling over thousands to millions of cells, providing an unprecedented opportunity to dissect the cellular heterogeneity of complex tissues at a single-cell resolution. Similar to scRNA-seq data, scATAC-seq data analysis suffers from ultra-high dimensionality, extremely high missingness, and complex feature interactions. Additionally, its computational modeling usually first involves the peak calling (or bin selection) process to select open chromatin candidates to generate the peak-by-cell matrix for downstream analyses such as dimension reduction and clustering, which usually require a sufficient number of cells for accurate region selection. This unique characteristic is distinct from the RNA analysis starting from gene sets with fixed genomic coordinates, introducing additional computational challenges in moderately sized scATAC-seq datasets.

Finally, the quality of the scATAC-seq experiments varies dramatically in terms of average sequencing depth, TSS enrichment, and SNR. As a result, accurate dimension reduction and clustering is sometimes difficult on moderate-quality scATAC-seq data (e.g., low sequencing depth or low TSS enrichment with limited cell numbers).

In this study, we propose a robust transfer-learning scheme, named Translator, to facilitate moderate-quality scATAC-seq data analyses using deeply sequenced, HQ reference dataset (Fig. 1). The motivation of our Translator framework is straightforward and widely accepted—open chromatin regions are faithfully conserved within the same or similar tissues and our VAE model will learn robust peak interactions in the 3D genome that are transferrable to similar datasets.

3.1. Translator significantly improves clustering performance on simulated datasets

Due to the lack of gold-standard datasets for performance benchmarking, we first simulated scATAC-seq data with ground truth labels by efficiently down-sampling bulk ATAC-seq data using the SCAN-ATAC-seq simulator (Chen et al., 2021c). First, we simulated cells from five different cell types with equal proportions and an average fragment number of 3k per cell. We used SNR at 0.8 for the HQ reference dataset and 0.2 as well as 0.35 for the LQ target datasets (scenario 1 in Table 1).

We found that although VAE slightly outperformed linear methods, such as LSI, neither of these models, when trained from scratch using the Translator VAE architecture, worked well on the target dataset, as reflected by the low ARI ranging from 0.134 to 0.219 (Fig. 2A, B). This was mainly due to the relatively

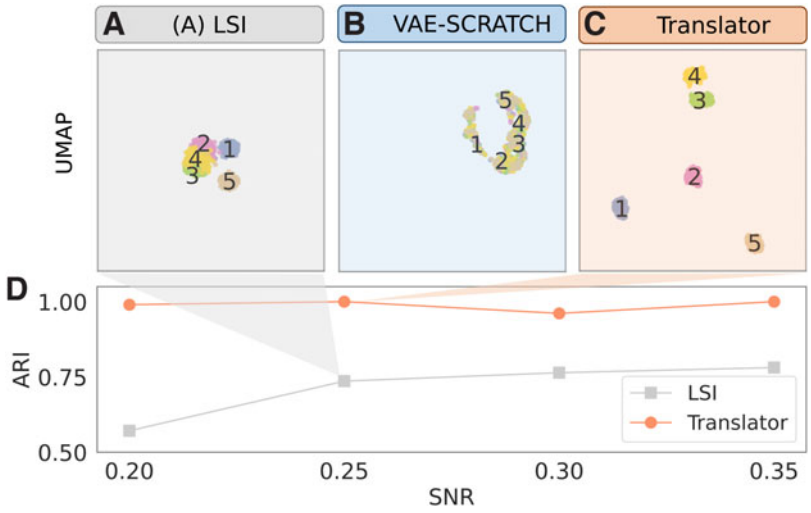

FIG. 2. Translator improves clustering results on simulation data with fixed depth. Top: UMAP visualizations using embeddings from: **(A)** latent semantic indexing (LSI), **(B)** training from scratch with VAE, and **(C)** Translator. **(D)** Bottom: Clustering performance (ARI) of simulated data with fixed depth for LSI and Translator. ARI, adjusted rand index; UMAP, Uniform Manifold Approximation and Projection.

low SNR (0.2 0.35) and the relatively small cell number (500) with limited mappable reads. To combat these problems, Translator first directly adopted the peaks from the reference scATAC-seq of the same tissue for more robust feature selection. Then, it combined both datasets for joint training in the VAE so that the reference-learned cell type-specific peak interaction could be directly used in the target data for robust cell embeddings. As a result, Translator clearly separated all five cell types into distinct groups (Fig. 2C) and demonstrated significantly higher ARIs using ground truth labels (0.571 0.781 vs. 0.990 1.0; Fig. 2D). In addition to these quantitative metrics, visualization results from other SNR simulations also show consistent improvement with Translator versus LSI or Signac (Supplementary Fig. S7).

It is well known that sequencing depth varies considerably across different studies and even within cells from the same experiment, producing artifacts in dimension reduction and clustering steps (Cao et al., 2021). Therefore, we mimicked this situation by simulating both deeply and shallowly sequenced scATAC-seq datasets (see details in Table 2). We then tested Translator's performance on various target scATAC-seq simulation settings, with average sequencing depth from 1500 to 3000 and SNR from 0.3 to 0.6 (see details in methods). In all scenarios, Translator's joint training scheme that incorporated the HQ reference dataset significantly improved the clustering results, as compared with the training from scratch scheme using Signac (ARI 0.930 1.0 in Translator vs. 0.548 0.742 in Signac, Table 2). We saw a larger improvement in extremely low sequencing depth cases. For instance, with extremely low SNR (0.3), Signac barely separated the different cell types, and even if the sequencing depth was improved (3000), Signac separated only one cluster (Supplementary Fig. S8). On the other hand, for all low SNR cases, Translator separated almost all cell types well (Supplementary Fig. S9).

3.2. Translator can robustly handle missing and novel cell types in the transfer learning process

We further tested the robustness of Translator's joint training scheme to learn biologically relevant cell representation on similar tissues with slightly unmatched cell types. Specifically, we simulated the reference and target dataset with unmatched cell population by removing existing or introducing novel cell types

Table 2. Adjusted Rand Indices of the Simulated Varying Depth Dataset

SNR	*Depth*	*LSI*	*Translator*	*SNR*	*Depth*	*LSI*	*Translator*	*SNR*	*Depth*	*LSI*	*Translator*
0.3	1500	0.090	0.975	0.45	1500	0.548	0.930	0.6	1500	0.742	0.991
	2000	0.250	0.970		2000	0.656	0.985		2000	0.735	0.978
	2500	0.406	0.802		2500	0.714	0.822		2500	0.768	0.981

LSI, representation learned using latent semantic indexing (LSI).

Table 3. Parameterization in Translator

Parameter	*Definition*
$x \in 0, 1^n$	Input scATAC-seq binarized peak profile
d	Dimension for latent cell embedding
N	Number of epochs in a cyclic annealing cycle
w	Cyclic annealing term
$z \in \mathbb{R}^d$	Cell representation for biological states
c	Observed confounding factor
$I(z, c)$	Mutual information
$q_\theta(z\|x)$	Encoder probability with parameter θ
$p_\phi(x\|z, c)$	Decoder probability with parameter ϕ

scATAC-seq, single-cell sequencing assay for transposase-accessible chromatin.

(Fig. 3). In the first case, the reference scATAC-seq data were simulated with five cell types with an average sequencing depth of 10,000, standard deviation of 1.5, and SNR of 0.8 (Table 1). To mimic the missing cell-type scenario, we only included four cell types in the target dataset (depth at 3000 and SNR at 0.3). We found that Translator clearly separated the remaining four cell types even when the fifth missing cell type-specific peaks were included to construct the cell-by-peak matrix (ARI = 1.000, Fig. 3A). Moreover, we also tested the novel cell type situation where there were only four cell types in the reference dataset and a new fifth cell type that was unique to the target data. Optimistically, Translator also accurately detected the novel cell type and clearly separated it from the others (ARI at 0.945; Fig. 3B).

3.3. *Translator improves PBMC scATAC-seq data analysis*

In addition to the simulated dataset, we applied Translator to public PBMC scATAC-seq data for performance evaluation. Due to the lack of gold standard cell labeling, we first downloaded the PBMC 10K sc-multiome data as the reference dataset, which has both RNA and ATAC profiling in the same cells. We used the PBMC 3k multiome dataset as the target data, but down-sampled the ATAC-seq modality to

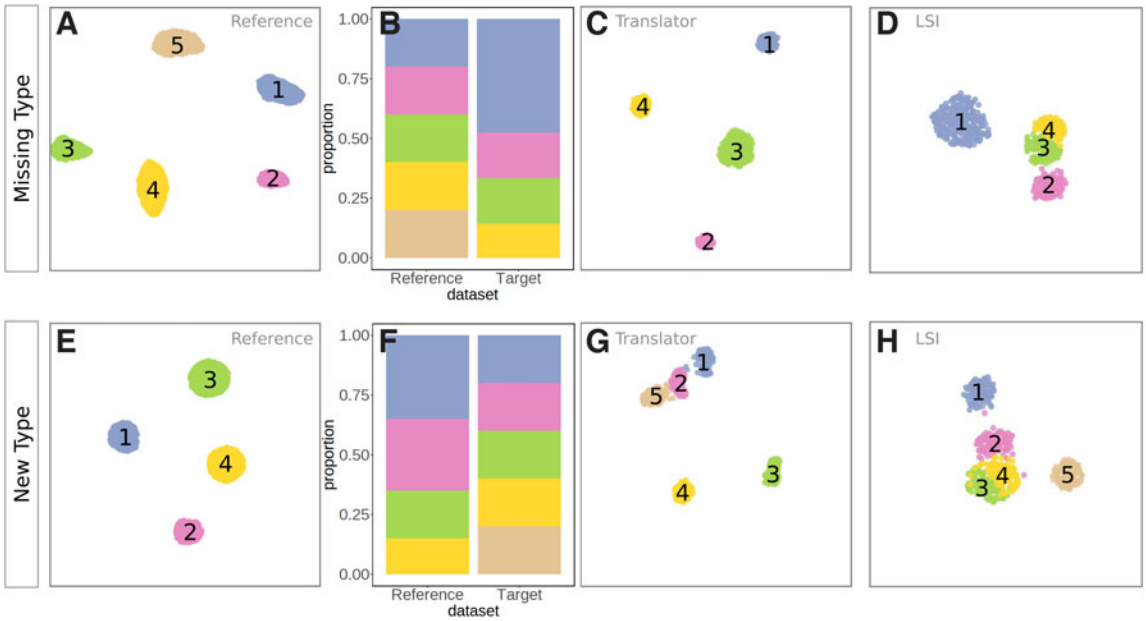

FIG. 3. Clustering results of the two special cases of the simulated data: **(A–D)** the missing type case and **(E–H)** the new type case. **(A, E)** UMAP of reference dataset (depth = 3k, cell_number = 4k × 5 for missing type and depth = 3k, cell_number = 7k:6k:4k:3k:0); **(B, F)** distribution of cell type ratios between reference and target datasets; **(C, G)** UMAP of Translator embedding of the target dataset (depth = 3k, cell_number = 250:100:100:75:0 for missing type and depth = 3k, cell_number = 100:100:100:100:100 for new type); and **(D, H)** UMAP of LSI embedding of the target dataset.

mimic the moderate-quality data (see Section 2.1.2). On both datasets, Translator was applied to the scATAC-seq modality and inferred the ground truth labels from the higher-quality scRNA-seq modality for performance benchmarking.

We found that Translator demonstrated improved UMAP with better clustering for all scenarios with sequencing depths 1000 3500 (Fig. 4A and Supplementary Figs. S10 and S11). For example, at a sequencing depth of 2500, Translator's ARI was 0.706 using cell labels from RNA, compared with the trained-from-scratch Signac's ARI of 0.502. Specifically, Translator was able to better distinguish CD4 groups, CD14/16 groups, and the B cell groups (Fig. 4C), while Signac generated mixed clusters for these main cell groups. Also, ArchR and SCALE failed the competition, with an ARI of 0.225 and 0.505, respectively (Table 4).

To test how different clustering results can impact further downstream analysis, we conducted motif enrichment analysis for different cell types based on the clustering results from Translator and Signac. Specifically, we first used the LEIDEN algorithm to cluster cells using embeddings generated from Signac and Translator, and then calculated a motif enrichment score for all cells in each cluster using ChromVar (Schep et al., 2017). Then, we compared the distribution of motif enrichment scores for cells within the clusters. In CD8 cell clusters, we selected the *GRHL2* gene, encoding for grainyhead like TFs 2 and *FOXP3* encoding for Forkhead box P3 (Fig. 5). Lines of literatures have shown that both TFs are significantly enriched in T cells, especially in CD8 cells (Uhlen et al., 2019; Bai et al., 2021). We found that the clustering from Translator reported significantly higher GRHL2 and FOXP3 motif enrichments (p values at 8.47E-5 and 9.79E-5, two-side t-test), validating its ability to produce more biologically relevant cell clustering.

3.4. Translator allows better clustering on PFC scATAC-seq data from human postmortem brains

We further tested Translator's performance on scATAC-seq data with more discrete cell types from postmortem brains. Specifically, we first combined three deeply sequenced scATAC-seq PFC samples as the reference dataset. The three reference samples have around 21k cells with an average sequencing depth around 6198 and average TSS enrichment around 6.5 (Fig. 6A). We ran ArchR using the default settings and annotated different clusters using marker genes from Lake et al. (2016). Nine distinct groups were discovered with clear marker gene patterns (Supplementary Fig. S12). On the other hand, the target dataset showed fewer cells that had passed QC (6285) and had a substantially lower sequencing depth (2473).

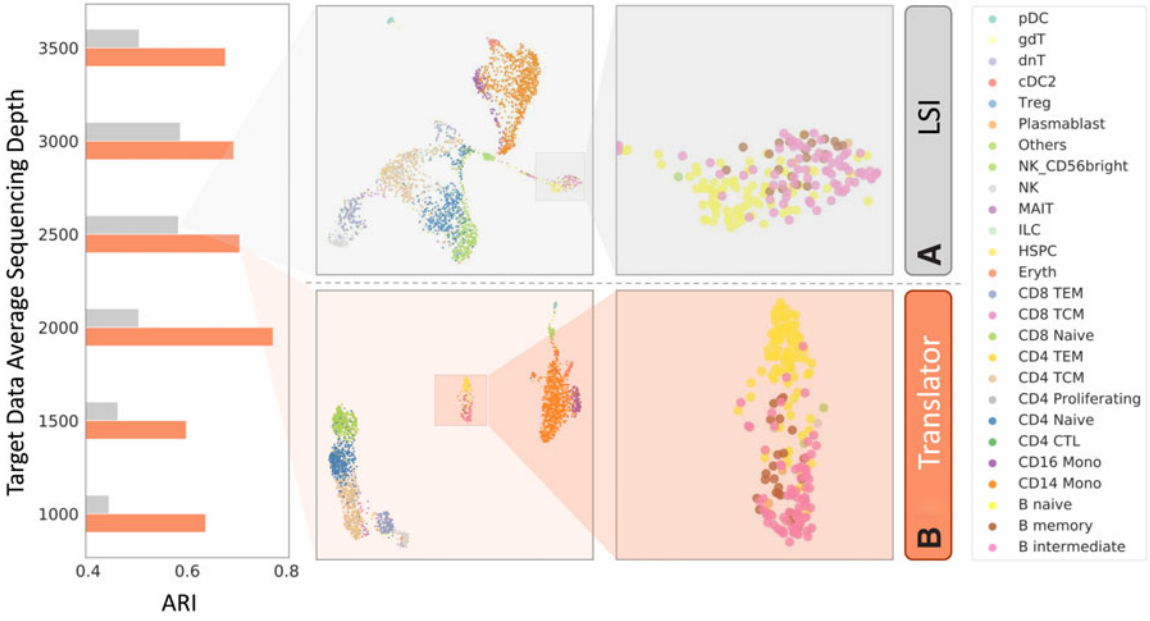

FIG. 4. Clustering results of the PBMC dataset: (left) the distribution of ARI with Translator and LSI with respect to sequencing depth, (middle) UMAP of full PBMC dataset at sequencing depth of 2500 using **(A)** LSI and **(B)** Translator, and (right) enlarged portion from the full UMAPs showing B cells. PBMC, peripheral blood mononuclear cells.

Table 4. Quantitative Evaluations of Real Datasets with Models

	Silhouette score				*ARI (calculated with LEIDEN)*			
	Signac	*ArchR*	*SCALE*	*Translator*	*Signac*	*ArchR*	*SCALE*	*Translator*
PBMC								
1500	–0.163	–0.565	–0.397	0.032	0.46	0.242	0.459	0.598
2000	–0.139	–0.587	–0.424	0.06	0.502	0.212	0.478	0.773
2500	–0.114	–0.567	–0.369	0.074	0.582	0.225	0.505	0.706
3000	–0.105	–0.584	–0.329	0.075	0.586	0.243	0.509	0.694
PFC	0.060	0.063	–0.074	0.143	0.417	0.256	0.497	0.539
Share-seq	–0.300	–0.379	–0.484	–0.028	0.064	0.000	0.128	0.272

ARI, adjusted rand index; PBMC, peripheral blood mononuclear cells; PFC, prefrontal cortex.

The trained-from-scratch Signac resulted in more mixed cell types and long shared clusters confounded by fragment counts (Fig. 6B), and the gene enrichment visualizations showed less-than-ideal separations (Supplementary Fig. S13). In contrast, using the three HQ samples as the reference dataset and Translator for joint training led to noticeably improved clustering results by generating more separatable cell groups, as reflected by the higher Silhouette scores (0.060 vs. 0.143) and ARI (0.539 vs. 0.417). Interestingly, clusters concentrating in the middle area, including VIP, microglia, OPC, and oligodendrocyte cells, were separated by Translator (Fig. 6C), demonstrating the effectiveness of Translator's joint training scheme to learn robust feature interactions and facilitate target data analysis. Plus, Translator also outperformed ArchR and Signac in similar tasks (Table 4, ARI = 0.539 vs. 0.256 and 0.497).

Plus, Translator also outperformed ArchR and Signac in similar tasks (Table 4, ARI = 0.539 vs. 0.256 and 0.497). We further tested the impact of different reference data selections to the final clustering result. We showed that Translator with cross-training using the PBMC reference has a worse clustering performance on PFC target dataset, especially in excitatory and inhibitory neurons than the original Translator results (Supplementary Fig. S16D–F, ARI = 0.539 with matched reference vs. 0.441 with unmatched reference). Hence, it is important to select HQ reference data from similar tissues.

3.5. Translator outperforms existing models in the Share-seq mouse skin tissue dataset

We further tested our model on the Share-seq dataset by separating the ATAC modality into reference and target dataset, while keeping labels obtained from the RNA modality for performance benchmarking. We demonstrated that on the reference dataset, Translator outperformed Signac and SCALE, especially on

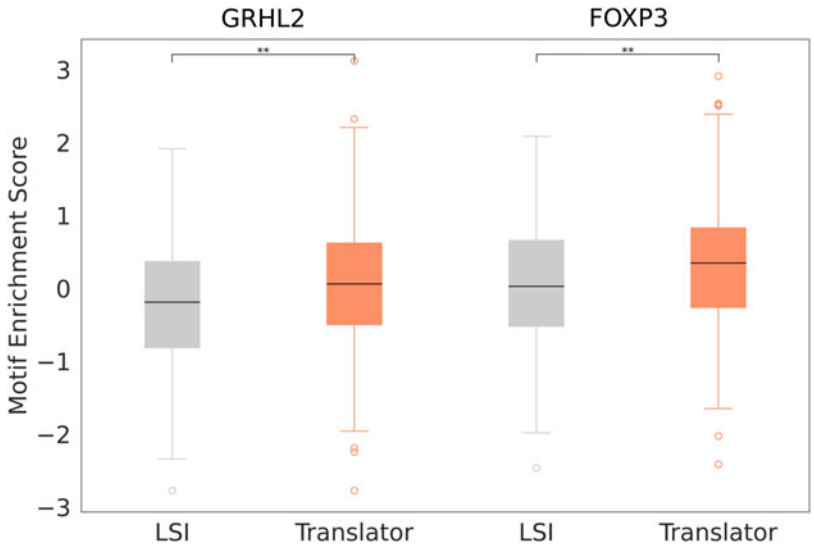

FIG. 5. Distribution of motif enrichment scores using LSI and Translator LEIDEN clusters for CD8 cell type-specific genes. **p-value = less than 0.01.

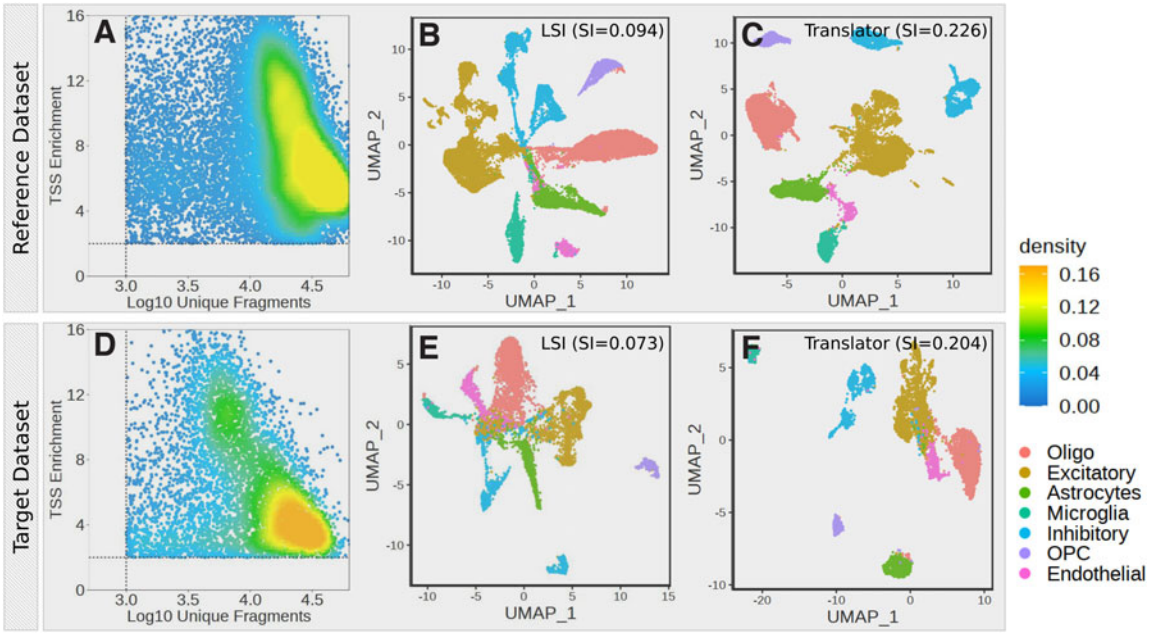

FIG. 6. QC and clustering results of the PFC data. **(A)** QC results of the reference dataset; **(B)** UMAP of LSI embeddings of the reference dataset; **(C)** UMAP of translator embeddings of the reference dataset; **(D)** QC results of the target dataset; **(E)** UMAP of the LSI embeddings of the target dataset; and **(F)** UMAP of the translator embeddings of the target dataset. QC, quality control.

certain cell types (e.g., Dermal Papilla vs. Dermal Fibroblast, Supplementary Fig. S14). We also found that melanocyte cells stand out as their own cluster on the top. Moreover, Translator also outperforms existing methods on the target data with transfer learning, as reflected by its improved ARI (0.211 for Translator vs. 0.064 for Signac and 0.128 for SCALE, Supplementary Fig. S15)

4. DISCUSSION

In this study, we present Translator, a VAE-based transfer learning model to facilitate single-cell ATAC-seq data analysis using a HQ reference dataset. Translator can learn robust feature interaction patterns from reference data and facilitate the analysis of scATAC-seq data of moderate quality. Moreover, Translator has several characteristics that boost its performance: (1) it increases model efficiency by reducing the number of chromosomes in the fully connected network, utilizing the biological assumption that intrachromosomal interactions are much more frequent than interchromosomal interactions; and (2) it is easily scalable using Graphics Processing Unit (GPU) parallelism. We applied Translator to various datasets to test its performance, including simulated datasets with various scenarios, a PBMC scATAC-seq dataset, and a PFC single-nuclei ATAC-seq dataset. In addition, we benchmarked our model with other methods, including the commonly used LSI and the VAE models trained from scratch. We showed that Translator can facilitate downstream analyses, such as clustering and motif enrichment, by learning transferrable feature interactions from the HQ reference dataset.

In addition, it simultaneously removes bias by penalizing dependencies between cell embeddings and known confounding factors, such as age, gender, and sequencing depth. Plus, our utilization of the block-wise neural network provided a fast and efficient way to conduct cell representation learning (Supplementary Fig. S17). In terms of running time, with regard to both number of cells and number of peaks, Translator achieved a linear increase. Also, Translator used 15 times less number of parameters on average than its fully connected counterparts, resulting in significantly reduced GPU memory usage during the training process.

One important note about our transfer learning approach is that chromatin interaction patterns should be highly conserved between the reference and target dataset so that features learned from our deep generative learning model can be directly applied to the target datasets. In other words, Translator will achieve its best

performance when the reference and target datasets are from similar tissues. On the other hand, strongly mismatched peak set might even negatively impact the training process. We also demonstrated that as long as the major cell populations are consistent, Translator can robustly handle different cell abundances or even slightly mismatched cases with new or missing cell types.

5. CONCLUSION

In summary, we developed an efficient deep generative model, Translator, to improve cell representation learning in moderate-quality single-cell ATAC-seq data by incorporating reference-learned information. With recent initiatives from the scientific community and funding agencies to encourage transparent data sharing, we anticipate that the number of HQ reference datasets will exponentially increase across various tissues and conditions. Therefore, our transfer learning approach will play a pivotal role in facilitating scATAC-seq data analysis.

AUTHORS' CONTRIBUTIONS

S.X.: Conceptualization, methodology, and software. M.S.: Resource, writing-original draft. A.H.: Data curation, writing-review and editing. Y.D.: Validation. C.L.: Writing-review and editing. M.G.: Supervision. J.Z.: Overall project design, conceptualization, supervision, and writing-review and editing.

DISCLAIMER

The views and opinions expressed are those of the authors.

ACKNOWLEDGMENTS

The authors would like to thank Laiyi Fu and Yingxin Cao for the constructive discussions during this project. They thank the Yale Center for Research Computing and UCI ICS Computing Support for guidance and use of the research computing infrastructure.

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no competing financial interests.

FUNDING INFORMATION

This work was funded in part by the U.S. Department of Veterans Affairs and the State of Connecticut, Department of Mental Health and Addiction Services, but this publication does not express the views of the Department of Mental Health and Addiction Services, the State of Connecticut, or the U.S. government. This work was supported by NIH grant K01MH123896 and the U.S. Department of Veterans Affairs National Center for PTSD.

SUPPLEMENTARY MATERIAL

Supplementary Figure S1
Supplementary Figure S2
Supplementary Figure S3
Supplementary Figure S4
Supplementary Figure S5
Supplementary Figure S6

Supplementary Figure S7
Supplementary Figure S8
Supplementary Figure S9
Supplementary Figure S10
Supplementary Figure S11
Supplementary Figure S12
Supplementary Figure S13
Supplementary Figure S14
Supplementary Figure S15
Supplementary Figure S16
Supplementary Figure S17

REFERENCES

Bai, X., Li, Y., Li, Y., et al. 2021. GRHL2 is a candidate prognostic and immunotherapy biomarker in breast cancer. *Res. Square*. DOI: 10.21203/rs.3.rs-999774/v1.

Boyle, A.P., Davis, S., Shulha, H.P., et al. 2008. High-resolution mapping and characterization of open chromatin across the genome. *Cell* 132, 311–322.

Bravo Gonzalez-Blas, C., Minnoye, L., Papasokrati, D., et al. 2019. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat. Methods* 16, 397–400.

Buenrostro, J.D., Wu, B., Litzenburger, U.M., et al. 2015. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 523, 486–490.

Cao, Y., Fu, L., Wu, J., et al. 2021. Sailer: Scalable and accurate invariant representation learning for single-cell ATAC-seq processing and integration. *Bioinformatics* 37(Suppl_1), i317–i326.

Chen, H., Lareau, C., Andreani, T., et al. 2019. Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biol.* 20, 241.

Chen, S., Liu, Q., Cui, X., et al. 2021a. Openannotate: A web server to annotate the chromatin accessibility of genomic regions. *Nucleic Acids Res.* 49(W1), W483–W490.

Chen, S., Yan, G., Zhang, W., et al. 2021b. Ra3 is a reference-guided approach for epigenetic characterization of single cells. *Nat Commun.* 12, 2177.

Chen, X., Chen, S., Song, S., et al. 2022. Cell type annotation of single-cell chromatin accessibility data via supervised Bayesian embedding. *Nat. Mach. Intell.* 4, 116–126.

Chen, X., Miragaia, R.J., Natarajan, K.N., et al. 2018. A rapid and robust method for single cell chromatin accessibility profiling. *Nat. Commun.* 9, 5345.

Chen, Z., Zhang, J., Liu, J., et al. 2021c. Scan-ATAC-sim: A scalable and efficient method for simulating single-cell ATAC-seq data from bulk-tissue experiments. *Bioinformatics.* DOI: 10.1093/bioinformatics/btaa1039.

Consortium, E.P. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74.

Cusanovich, D.A., Daza, R., Adey, A., et al. 2015. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* 348, 910–914.

Davis, C.A., Hitz, B.C., Sloan, C.A., et al. 2018. The encyclopedia of DNA elements (encode): Data portal update. *Nucleic Acids Res.* 46(D1), D794–D801.

Fang, R., Preissl, S., Li, Y., et al. 2021. Comprehensive analysis of single cell ATAC-seq data with snapatac. *Nat. Commun.* 12, 1337.

Fornes, O., Castro-Mondragon, J.A., Khan, A., et al. 2020. Jaspar 2020: Update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 48(D1), D87–D92.

Fu, H., Li, C., Liu, X., et al. 2019. Cyclical annealing schedule: A simple approach to mitigating KL vanishing. *arXiv* arXiv:190310145. Doi: 10.48550/ARXIV.1903.10145.

Granja, J.M., Corces, M.R., Pierce, S.E., et al. 2021. Archr is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* 53, 403–411.

Hao, Y., Hao, S., Andersen-Nissen, E., et al. 2021. Integrated analysis of multimodal single-cell data. *Cell* 184, 3573.e29–3587.e29.

Huang, M., Wang, J., Torre, E., et al. 2018. Saver: Gene expression recovery for single-cell RNA sequencing. *Nat. Methods* 15, 539–542.

Ji, Z., Zhou, W., Hou, W., et al. 2020. Single-cell ATAC-seq signal extraction and enhancement with scate. *Genome Biol.* 21, 161.

Kingma, D.P., and Welling, M. 2013. Auto-encoding variational bayes. *arXiv* arXiv:13126114.

Klemm, S.L., Shipony, Z., and Greenleaf, W.J. 2019. Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.* 20, 207–220.

Korsunsky, I., Millard, N., Fan, J., et al. 2019. Fast, sensitive and accurate integration of single-cell data with harmony. *Nat. Methods* 16, 1289–1296.

Lake, B.B., Ai, R., Kaeser, G.E., et al. 2016. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science* 352, 1586–1590.

Liu, Q., Chen, S., Jiang, R., et al. 2021. Simultaneous deep generative modelling and clustering of single-cell genomic data. *Nat. Mach. Intell.* 3, 536–544.

Ma, S., Zhang, B., LaFave, L.M., et al. 2020. Chromatin potential identified by shared single-cell profiling of rna and chromatin. *Cell* 183, 1103.e20–1116.e20.

Pott, S., and Lieb, J.D. 2015. Single-cell ATAC-seq: Strength in numbers. *Genome Biol.* 16, 172.

Satpathy, A.T., Granja, J.M., Yost, K.E., et al. 2019. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* 37, 925–936.

Schep, A.N., Wu, B., Buenrostro, J.D., et al. 2017. chromvar: Inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* 14, 975–978.

Schneider, V.A., Graves-Lindsay, T., Howe, K., et al. 2017. Evaluation of grch38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* 27, 849–864.

Stuart, T., Srivastava, A., Lareau, C., et al. 2020. Multimodal single-cell chromatin analysis with Signac. *bioRxiv*. DOI: 10.1101/2020.11.09.373613.

Stuart, T., Srivastava, A., Madad, S., et al. 2021. Single-cell chromatin state analysis with Signac. *Nat. Methods* 18, 1333–1341.

Tipping, M.E., and Bishop, C.M. 1999. Probabilistic principal component analysis. *J. R. Stat. Soc. B (Stat. Methodol.)* 61, 611–622.

Tsompana, M., and Buck, M.J. 2014. Chromatin accessibility: A window into the genome. *Epigenet. Chromatin* 7. DOI: 10.1186/1756-8935-7-33.

Uhlen, M., Karlsson, M.J., Zhong, W., et al. 2019. A genome-wide transcriptomic analysis of protein-coding genes in human blood cells. *Science* 366. DOI: 10.1126/science.aax9198.

Xiong, L., Xu, K., Tian, K., et al. 2019. Scale method for single-cell atac-seq analysis via latent feature extraction. *Nat. Commun.* 10, 4576.

Xu, W., Wen, Y., Liang, Y., et al. 2021. A plate-based single-cell ATAC-seq workflow for fast and robust profiling of chromatin accessibility. *Nat. Protoc.* 16, 4084–4107.

Yan, F., Powell, D.R., Curtis, D.J., et al. 2020. From reads to insight: A hitchhiker's guide to ATAC-seq data analysis. *Genome Biol.* 21. doi: 10.1186/s13059-020-1929-3.

Zhou, C., Yuan, Z., Ma, X., et al. 2021. Accessible chromatin regions and their functional interrelations with gene transcription and epigenetic modifications in sorghum genome. *Plant Commun.* 2, 100140.

Zonouzi, M., Berger, D., Jokhi, V., et al. 2019. Individual oligodendrocytes show bias for inhibitory axons in the neocortex. *Cell. Rep.* 27, 2799–2808e3.

Address correspondence to:
Dr. Matthew J. Girgenti
Department of Psychiatry
School of Medicine
Yale University
New Haven, CT 06519
USA

E-mail: matthew.girgenti@yale.edu

Dr. Jing Zhang
Department of Computer Science
University of California, Irvine
Irvine, CA 92697
USA

E-mail: zhang.jing@uci.edu