Genes to Cells

WILEY

Genes to Cells

ORIGINAL ARTICLE OPEN ACCESS

An Attention-Based Deep Neural Network Model to Detect Cis-Regulatory Elements at the Single-Cell Level From Multi-Omics Data

Ken Murakami[1,2] | Keita Iida[1] | Mariko Okada[1]

[1]Laboratory for Cell Systems, Institute for Protein Research, Osaka University, Suita, Japan | [2]Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives (OTRI), Osaka University, Suita, Japan

Correspondence: Mariko Okada (mokada@protein.osaka-u.ac.jp)

Received: 9 October 2024 | **Revised:** 17 December 2024 | **Accepted:** 12 January 2025

Transmitting Editor: Shumpei Ishikawa

Funding: This work was supported by the Japan Society for the Promotion of Science KAKENHI [Grant Number 18H04031], the Japan Science and Technology Agency CREST [Program Number JPMJCR21N3], and the Uehara Memorial Foundation to M.O. K.M. was supported by Grant-in-Aid for JSPS Fellows [Grant Number 23KJ1476], Osaka University Institute for Open and Transdisciplinary Research Initiatives, and the JST CREST AIP challenge program 2022–2023. K.I. was supported by JST Moonshot R&D [Grant Number JPMJMS2021].

Keywords: attention-based neural network | cis-regulatory elements | deep learning | enhancers | explainable artificial intelligence | gene regulation | intra-tumor heterogeneity | single-cell analysis | single-cell ATAC-seq | single-cell multiome

ABSTRACT

Cis-regulatory elements (cREs) play a crucial role in regulating gene expression and determining cell differentiation and state transitions. To capture the heterogeneous transitions of cell states associated with these processes, detecting cRE activity at the single-cell level is essential. However, current analytical methods can only capture the average behavior of cREs in cell populations, thereby obscuring cell-specific variations. To address this limitation, we proposed an attention-based deep neural network framework that integrates DNA sequences, genomic distances, and single-cell multi-omics data to detect cREs and their activities in individual cells. Our model shows higher accuracy in identifying cREs within single-cell multi-omics data from healthy human peripheral blood mononuclear cells than other existing methods. Furthermore, it clusters cells more precisely based on predicted cRE activities, enabling a finer differentiation of cell states. When applied to publicly available single-cell data from patients with glioma, the model successfully identified tumor-specific SOX2 activity. Additionally, it revealed the heterogeneous activation of the ZEB1 transcription factor, a regulator of epithelial-to-mesenchymal transition-related genes, which conventional methods struggle to detect. Overall, our model is a powerful tool for detecting cRE regulation at the single-cell level, which may contribute to revealing drug resistance mechanisms in cell sub-populations.

1 | Introduction

The human genome contains approximately 20,000 protein-coding genes and over 900,000 cis-regulatory elements (cREs) that regulate gene expression (Moore et al. 2020). Typical cREs, known as enhancers, contain transcription factor (TF) binding sites and are located at distances ranging from a few kilobase pairs to several megabase pairs from the transcription start site (TSS) (Panigrahi and O'Malley 2021). Enhancers make physical contact with promoters via TFs and are known to activate gene expression, regardless of their orientation (Yang and Hansen 2024). Enhancers act as on/off switches for gene expression, and variations in active enhancer regions contribute to differences in cell types and species (Villar

et al. 2015). Although the coding regions follow clear rules, as determined by the DNA codon table, the regulation of cREs is highly complex. Active cREs are typically present in an open chromatin state (Thurman et al. 2012) and regulate gene expression by allowing TFs to bind to specific motifs (Spitz and Furlong 2012). However, the cell controls when and where the genome becomes accessible, as well as the number of TFs that can bind cREs, depending on the cell state. In addition, stochastic changes in chromatin accessibility are known to exist within heterogeneous in-cell populations (Bohrer and Larson 2021). This heterogeneity is thought to be a driving force for cell state transition, as shown in the Waddington Landscape (Waddington 1957), and contributes to the plasticity and drug resistance of cancer cells in pathological contexts (Marusyk, Janiszewska, and Polyak 2020). Therefore, to elucidate transcriptional regulation and the control of multicellular processes, detecting cREs and their effects on gene expression levels in individual cells is essential.

A single-cell assay for transposase-accessible chromatin sequencing (scATAC-seq) (Buenrostro et al. 2015) has been widely used to detect the open chromatin regions at the single-cell level. Several computational methods have been developed to annotate functions and detect cREs in scATAC-seq-derived open chromatin regions. Early approaches to detect cRE-gene relationships from single-cell data were based on the correlation between the chromatin accessibility of cRE candidates and the gene expression level (or chromatin accessibility at the promoter) of target genes (Pliner et al. 2018; Granja et al. 2021). Despite being significantly faster at detecting cREs compared to traditional reporter assays, these correlation-based methods also detect false-positive pairs, especially when the dataset does not contain a sufficient variety of cell states or cell types. To address this issue, curation rules such as penalization were introduced based on the genomic distance from the gene's TSS (Pliner et al. 2018; Granja et al. 2021). Recent studies (Zhang, Zhang, and Nie 2022; Bravo González-Blas et al. 2023) used machine-learning methods, such as XGBoost (Chen and Guestrin 2016), utilizing chromatin accessibility counts from multiple cRE candidates around the target gene as input to predict gene expression levels. These models identify cREs via contribution scores and, by incorporating multiple cREs, reduce false positives and improve precision. However, models based solely on chromatin accessibility require separate training to predict the expression level of each target gene, increasing the risk of overfitting due to limited gene expression diversity relative to model parameters and limiting their ability to learn transcriptional regulation rules that apply across multiple genes.

Another machine-learning approach focuses on the relationship between DNA sequences and the characteristics of cREs, such as quantitative TF binding (Avsec et al. 2021b), histone modification (Kelley et al. 2018; Avsec et al. 2021a; Yuan and Kelley 2022), and chromatin accessibility (Kelley et al. 2018; Avsec et al. 2021a; Yuan and Kelley 2022) or activity in reporter assays (de Almeida et al. 2022; de Almeida et al. 2024). Although these models cannot be directly applied to single-cell-level cRE-gene relationships because they only accept DNA sequences as input, they suggest that the DNA sequence of cRE regions has sufficient information to predict cRE characteristics regardless of the interacting target gene. Thus, incorporating the DNA sequence of cRE candidates into the prediction of gene expression levels might enable the training of comprehensive models that can learn gene-cRE relationships with a single parameter set and improve the prediction performance. However, such architecture is lacking.

Here, we propose an attention-based deep learning model that integrates chromatin accessibility, DNA sequence information, and genomic distance to learn a comprehensive genetic regulatory code. Recently, deep neural networks have offered a promising avenue for learning complex regulatory rules without prior knowledge. Among these, an attention method that allows flexible learning independent of the order or size of the input data has achieved remarkable results (Vaswani et al. 2023), particularly in natural language processing. We propose this attention-based deep neural network framework for detecting cRE activity at single-cell resolution.

2 | Results

2.1 | Overview of the Framework

Our framework uses a single sample of single-cell Multiome ATAC-seq + GEX (scMultiome) data to simultaneously obtain single-cell RNA sequencing (scRNA-seq) and scATAC-seq information from the same cells. First, we trained deep neural networks to predict gene expression levels at the single-cell level from scATAC-seq counts, DNA sequences, and the genomic distance of the gene's neighboring ATAC-seq peaks (Figure 1A, left). Then, our framework calculated the contribution score of the cRE candidates to the expression of the target gene (Figure 1A, right), which reflects the potential contribution of each peak to gene expression. Previous studies (Zhang, Yang, and Zhang 2022; Bravo González-Blas et al. 2023) have reported that high contribution scores in regression models are associated with high cRE activity. Therefore, we used the contribution score as an indicator of cRE activity at the single-cell level.

Our model assumes that gene expression levels are determined by the combination of gene promoter and cRE activities, based on previous knowledge indicating that the combination of promoter and enhancer activities can explain more than 60% of the variance in gene expression levels (Bergman et al. 2022). For each gene, we defined promoter peaks as ATAC peaks within ± 500 bp of the gene's TSS and considered all ATAC peaks within ± 300,000 bp of the gene's TSS as cRE candidates. The input data for the models comprised the DNA sequence, genomic distance, and scATAC-seq counts of the promoter and cRE candidates of each gene. The output was the normalized scRNA-seq count.

Our neural network models consist of the DNA-seq encoder, which extracts DNA sequence features from raw sequence data, and the Attention block, which combines ATAC-seq, genomic distance, and the output of the DNA-seq Encoder (Figure 1B). In our framework, the DNA-seq encoder first compressed all ATAC peaks one by one and extracted important motif information from the 1344 bp DNA sequence of all input peaks (Figure 1B, upper left). This 8-layer convolutional neural network (CNN) compressed each peak to a 32-dimensional peak embedding.

The attention block had a structure similar to that of a general cross-attention-based neural network (Rombach et al. 2022) (Figure 1C). Initially, it extracts features from peak embeddings and the genomic distance between the promoter and candidate cREs to compute the cross-attention between these elements (Figure 1C, Attention Matrix). This attention matrix highlighted important cRE candidates based on DNA-seq data and genomic distances. In a typical cross-attention-based

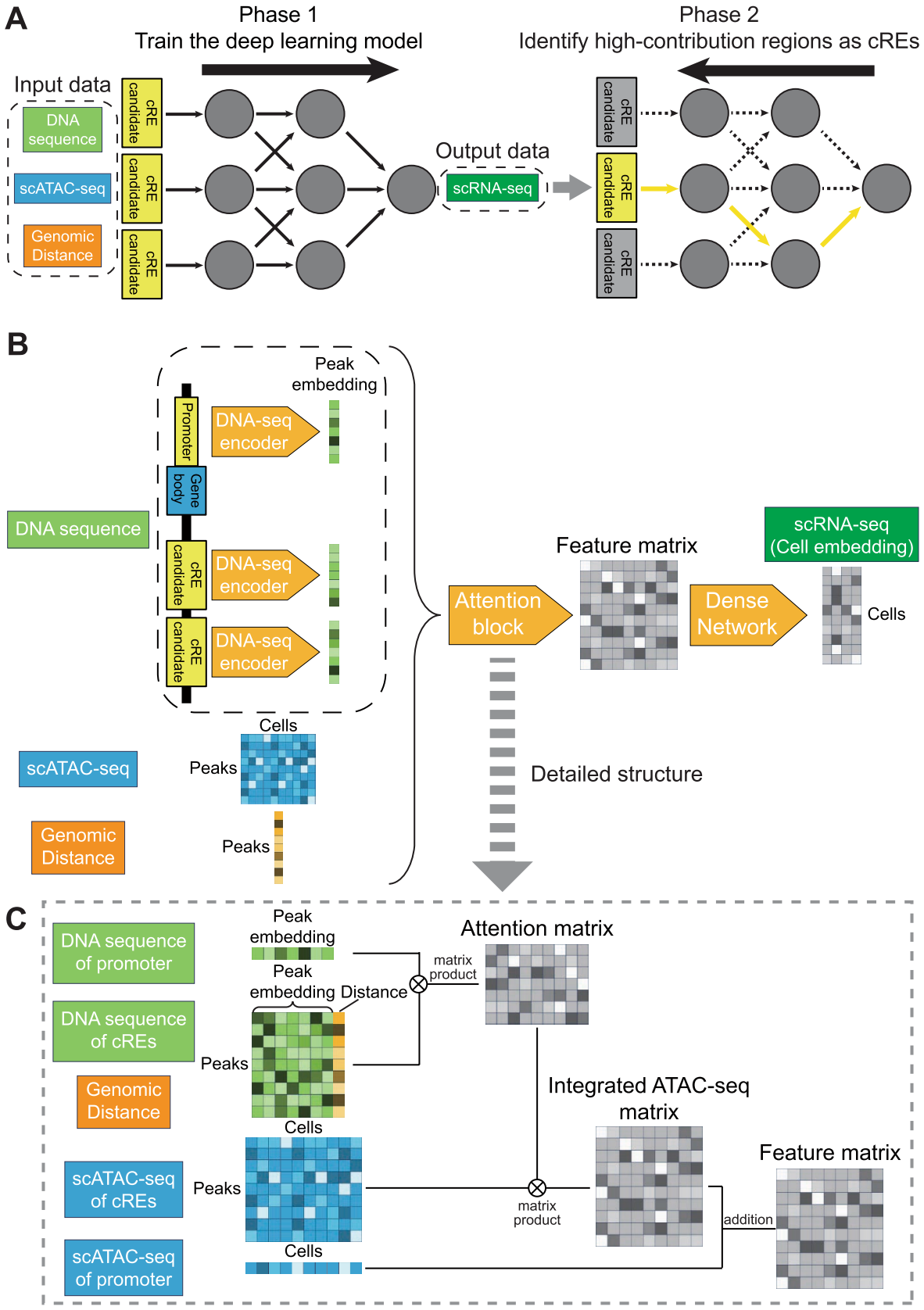

FIGURE 1 | Legend on next page.

neural network, the next step is to apply the attention matrix to the same feature matrix used to calculate the cross-attention (in this case, the matrix of peak embeddings and genomic distance) to extract the relevant features. However, our model uniquely applied the attention matrix to the scATAC-seq count matrix, extracting the scATAC-seq features of important regions identified by DNA-seq features and genomic distance (Figure 1C, integrated scATAC-seq matrix). This cross-modal attention mechanism enabled the integration of DNA sequences, genomic distances, and scATAC-seq information using a simple cross-attention structure. Finally, the scATAC-seq counts from the promoter regions were integrated to produce a final feature matrix (Figure 1C).

Finally, the dense neural network converted the feature matrix from the attention block to the gene expression level (Figure 1B, right). To stabilize the learning, we used principal component analysis (PCA) coordinates derived from scRNA-seq data instead of normalized gene expression. After learning the neural network model, our framework used DeepLIFT, which was developed as a tool to calculate the contribution score of a neural network (Shrikumar, Greenside, and Kundaje 2019) to evaluate the contribution of cRE candidates and promoters to gene expression levels.

Model training consisted of two steps. First, the DNA-seq encoder was trained following a methodology similar to scBasset (Yuan and Kelley 2022). Specifically, a DNA-seq encoder was trained to predict chromatin accessibility at the single-cell level using DNA sequences. This process enabled the DNA-seq encoder to learn the critical sequence patterns for chromatin accessibility. Training of the DNA-seq encoder ran for 1000 epochs. Second, the attention block and dense network were trained using scRNA-seq and scATAC-seq. 10% of the ATAC peaks and 20% of the cells were reserved as test data, and the remaining data were used for training.

2.2 | Benchmarking of cREs Prediction Performance

First, we benchmarked our framework using the public scMultiome dataset of human peripheral blood mononuclear cells (PBMC) from healthy donors (10x Genomics 2021). The data includes 11,898 cells, 36,601 genes, and 143,887 peaks as raw data. After filtering, 11,754 cells, 6853 genes, and 36,071 peaks were retained. The maximum number of peaks associated with a single gene was 79. After training the model and defining cRE activity as 60% cRE from the higher end, 83,519 cRE-gene pairs were detected. The prediction of PCA coordinates for cells derived from scRNA-seq demonstrated high accuracy for the primary PC axes, particularly for PC1 to PC8 (Figure S1, 2). After annotating cell types based on their gene expression (Figure 2A), our framework detected cell type-specific cRE activities, such as $CD4^+$ T cell-specific (Figure 2B, left) and $CD8^+$ T cell-specific (Figure 2B, right) cRE activities in the CD3D gene region, despite similar gene expression levels in these cell types (Figure 2C).

To assess how well our model performs at predicting cREs, we decided to benchmark it against existing tools. Cicero (Pliner et al. 2018) and ArchR (Granja et al. 2021) use scATAC-seq counts at candidate regions to detect cREs by correlating them either with scATAC-seq counts at promoter regions (Cicero) or with scRNA-seq gene counts (ArchR). On the other hand, DIRECT-NET (Zhang, Zhang, and Nie 2022) and SCENIC+ (Bravo González-Blas et al. 2023) rely on XGBoost, a machine-learning method that predicts gene expression levels from scATAC-seq (Chen and Guestrin 2016). To compare these different models, we then measured how well they performed at predicting cREs from the FANTOM5 database (Lizio et al. 2019) or detected using Promoter Capture Hi-C (PCHiC; Javierre et al. 2016). Importantly, our framework showed a significantly higher area under the receiver operating characteristic curve (AUROC) for both datasets (Figure 2D,E), indicating higher accuracy.

The likelihood of cRE-gene interactions decreases with increasing genomic distances, and the risk of false positives is generally higher for distant cRE-gene pairs. Although our model directly integrates distance information, Cicero and ArchR simply apply uniform penalties to distant cRE-gene pairs (Pliner et al. 2018; Granja et al. 2021), which might hinder the detection of bona fide long-range interactions. To assess how genomic distance affects our model's predictions, we stratified gene-cRE pairs into 20 bins of increasing genomic distance. Importantly, our framework showed a higher prediction performance at every distance using CAGE-seq data (Figure 2F, left) and at most distances using PCHiC data (Figure 2F, right).

Next, we wanted to probe whether the cREs detected by our model were enriched for expression quantitative trait loci (eQTL), which would indicate that they correspond

FIGURE 1 | Workflow to detect cREs at the single-cell level. (A) The strategy to detect cRE regions using the deep learning method. As a first step, our framework trains a deep neural network to predict scRNA-seq counts from scATAC-seq counts, DNA sequence, and genomic distance. Next, the contribution of each cRE candidate to the gene expression level is calculated, and the regions with high contribution are determined as cREs. (B) The model structure of the attention-based neural network to predict gene expression level from DNA sequence, ATAC-seq, and genomic distance. First, the DNA-seq encoder extracts important features from the DNA sequence. Next, the features derived from the DNA sequence, scATAC-seq counts, and genomic distance are input into the attention block. The attention block integrates these input data and converts it into a feature matrix. Finally, the dense network predicts gene expression level. (C) The detailed structure of the attention block. First, the attention matrix is calculated between the promoter and cRE candidates based on the peak embedding derived from the DNA sequence and genomic distance. This attention matrix stores information on which cRE candidates have important features. By applying this attention matrix to scATAC-seq counts, an integrated scATAC-seq matrix is generated that only contains scATAC-seq counts for important regions. Finally, the scATAC-seq counts for promoters are added to generate a feature matrix. cRE, cis-regulatory elements; scRNA-seq, single-cell RNA sequencing; scATAC-seq, single-cell assay for transposase-accessible chromatin sequencing.

to bona fide cREs. We therefore used the eQTL database from the GTEx portal (GTEx Consortium 2020) to show that our predicted set of cREs showed significantly higher eQTL enrichment ratios compared to cREs inferred using CAGE-seq or PCHiC contacts (Figure 2G, p-value = $1.89*10^{-3}$ for FANTOM5, p-value = $8.04*10^{-6}$ for PCHiC), suggesting that our model outperforms existing ones at detecting functional cRE-gene pairs.

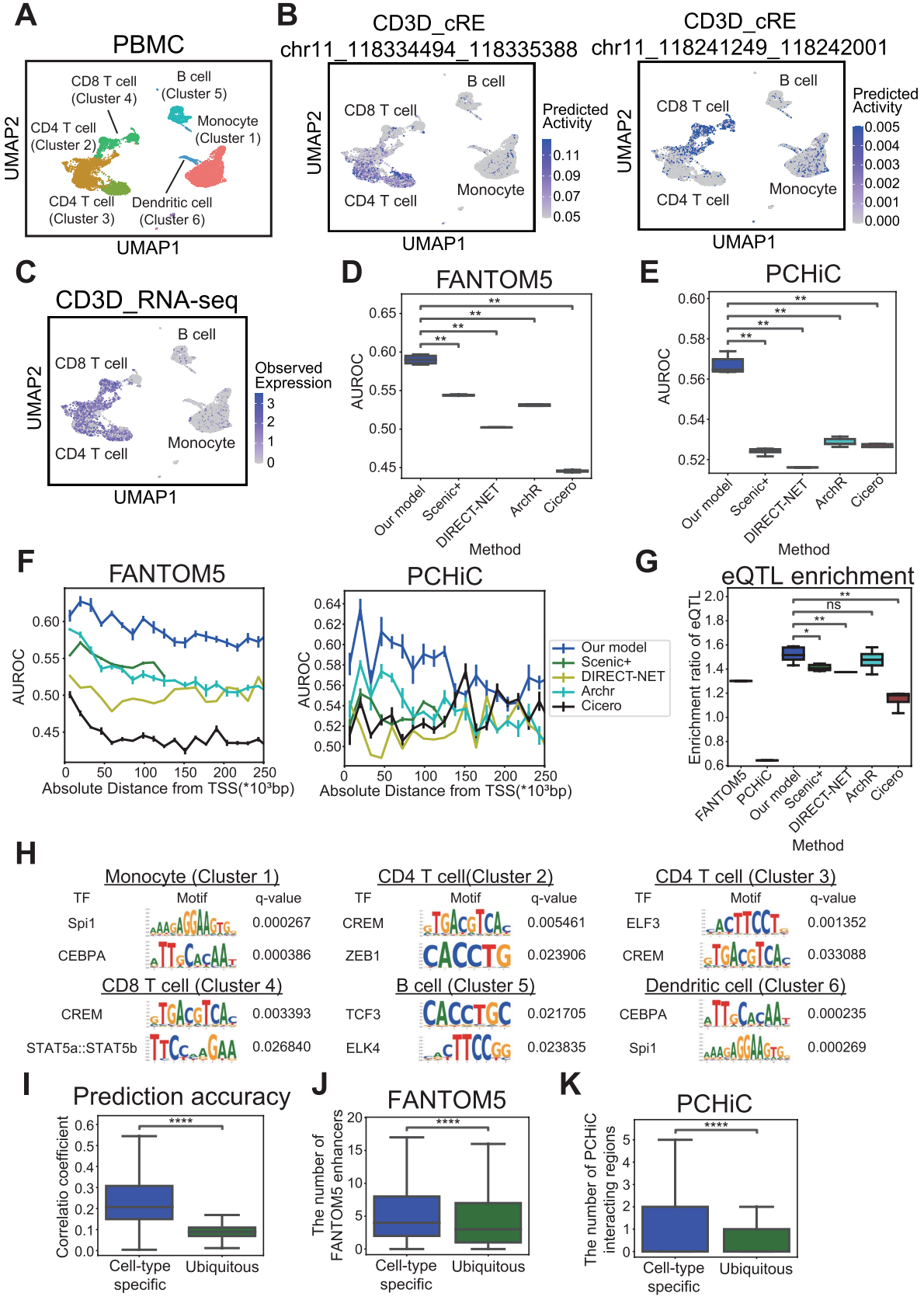

FIGURE 2 | Legend on next page.

Finally, we evaluated the robustness of our model to changes in dataset size by varying the number of cells in the same PBMC dataset to 10,000, 5000, and 2500 cells, and compared the prediction accuracy of cREs. The results showed that while the prediction accuracy of cREs tended to decrease as the number of cells decreased, our model maintained robust performance (Figure S3). These findings suggest that our model can function effectively even with smaller datasets.

2.3 | Functional Characteristics of Predicted cREs

Besides detecting cRE-gene pairs, our framework evaluates DNA sequence motifs that are crucial for gene expression, allowing us to identify key TFs and infer transcriptional regulatory gene networks. After computing the contribution score of the input DNA sequence using Integrated Gradients (Sundararajan, Taly, and Yan 2017) with single-base pair resolution, we used TF-MoDISCo (Shrikumar et al. 2020) to identify the enriched DNA motif patterns contributing to gene expression levels. In the PBMC dataset, our framework successfully detected established cell-type-specific TFs, such as CEBPA in monocytes (Scott et al. 1992), STAT5 in $CD8^+$ T cells (Tripathi et al. 2010), and ELK4 in B cells (Yasuda et al. 2008), among the top five highly contributing DNA motifs (Figure 2H).

In parallel, we interestingly observed that our model was generally more accurate at predicting the expression levels of cell type-specific genes compared to ubiquitously expressed housekeeping genes (Figure 2I). A previous study (Zabidi et al. 2015) reported that cell type-specific genes are more dependent on distal cREs for their expression compared to housekeeping genes, which led us to hypothesize that differences in prediction accuracy might depend on the extent to which gene expression depends on cRE control. Consistently, we found that cell type-specific genes had a significantly larger number of cREs compared to housekeeping genes, using both the FANTOM5 and the PCHiC dataset (Figure 2J,K).

2.4 | Analysis of Tumor-Specific cRE Regulation

Next, we wanted to evaluate the performance of our model in predicting differences between cell types by measuring their clustering performance using healthy human PBMC data (10× Genomics 2021). The cells were clustered using k-means, either on the predicted cRE activity from our model or using eRegulon activity inferred using Scenic+ (Bravo González-Blas et al. 2023). Importantly, the cRE activities predicted by our model outperformed eRegulon predictions, as they showed a significantly higher Adjusted Rand Index (ARI, see Section 4).

To further assess the capacity of each model at identifying functionally relevant cREs, we then used the top 15% cREs with the highest contribution scores detected by each model and used overlapping scATAC-seq counts to cluster the different cell types present in the data. The cREs identified by our model clustered cells with a higher ARI compared to other tested tools (Figure 3A, right panel), suggesting that it is more efficient at identifying relevant cREs from scATAC-seq data. However, ARI scores were consistently lower than those obtained using our model's predicted cRE activity or the eRegulons from Scenic+ (Figure 3A), indicating that scATAC-seq counts do not perfectly reflect the gene expression status of cells. However, our model can use this data to identify differences in cRE regulation between cell types that previous methods overlooked.

Next, to verify whether our model can detect tumor-specific cRE regulation, we applied our model to public single-cell data from patients with pediatric glioma (Jessa et al. 2022; 66,070 cREs detected in total, see Section 4). First, cell clusters identified using scRNA-seq and cRE activity corresponded to the cell types identified by scRNA-seq (Figure 3B). To determine the key DNA binding motifs associated with cell type-specific gene expression changes, we then performed motif enrichment at cell type-specific cREs using TF-MoDISCo (Shrikumar et al. 2020). This way, we identified sequence signatures for the NHLH2 and NFIC neuronal factors (Frazel et al. 2023; Wilczynska et al. 2009) but also for SOX2 (Figure 3C), which has been associated with poor prognosis in gliomas (Garros-Regulez et al. 2016) and was the most enriched motif in glioma-specific cREs. Consistent with motif analysis, predicted SOX2 activity in cREs was significantly higher in glioma cells (Figure 3D, $p = 0.000$). However, SOX2 was more highly expressed in oligodendrocyte precursor cells (OPCs) than in glioma cells (Figure 3E), highlighting the limitation of relying solely on TF gene expression levels to predict TF activity.

To further validate SOX2 binding to tumor-specific cREs, we used publicly available SOX2 CUT&Tag data from patients

FIGURE 2 | Model performance to predict cREs. (A) Cell type annotation of PBMC scMultiome data mapping was performed using scRNA-seq counts. (B, C) cREs (B) and observed gene expression levels of (C) CD3D. UMAP mapping was performed by scRNA-seq. Blue indicates gene expression level or predicted activity. (D, E) Prediction accuracy of cREs in the FANTOM5 (D) and PCHiC datasets (E). Each tool was trained 5 times. Boxplot represents quartile points. $^{**}p \leq 1.0^{*}10^{-2}$ (vs. our model, Mann–Whitney *U* test). (F) Prediction accuracy of cREs by distance from the TSS in the FANTOM5 dataset (left) and PCHiC data (right). Accuracy was measured using AUROC. The length of each bin was 1500 bp. Error bars indicate standard deviations. (G) eQTL enrichment ratio of cREs predicted using computational tools, including our model and experimental methods. The box plot shows the result of 5-fold learning. Boxplot represents quartile points. ns, $5.0^{*}10^{-2} < p$; $^{*}p <= 5.0^{*}10^{-2}$; $^{**}p \leq 1.0^{*}10^{-2}$ (vs. our model, Mann–Whitney *U* test). (H) Enriched transcription factor-binding motifs in cell type-specific cREs. The enriched motifs and q-values were calculated using TF-MoDISCo (Shrikumar et al. 2020). (I) Comparison of the accuracy of the prediction of gene expression levels between genes with cell type-specific expression and those with ubiquitous expression. The y-axis shows Spearman's correlation coefficient. *p*-value = 0.000. Student's *t*-test was used after Fisher's Z-transformation to statistically test the correlation coefficients. (J, K) Comparison of the number of cRE regions linked to genes with cell type-specific and ubiquitous expression. (J) In the case of the FANTOM5 dataset, $p = 6.201^{*}10^{-18}$, and (K) in the case of PCHiC data, $p = 1.770^{*}10^{-9}$. Unpaired Student's *t*-test was used for statistical analysis. AUROC, area under the receiver operating characteristic curve; cRE, cis-regulatory elements; PBMC, peripheral blood mononuclear cells; PCHiC, Promoter Capture Hi-C; scMultiome, single-cell Multiome ATAC-seq + GEX; scRNA-seq, single-cell RNA sequencing; UMAP, Uniform Manifold Approximation and Projection.

with glioma (Benedetti et al. 2022). As predicted by our model, tumor-specific cREs with SOX2 binding motifs showed significantly higher CUT&Tag signals than non-cRE regions with the SOX2 motif (Figure 3F, $p = 2.4*10^{-10}$). This result demonstrates that our model can effectively integrate DNA sequence, epigenetic, and transcriptomic information to predict TF-binding events, which is difficult to achieve using only DNA sequence data.

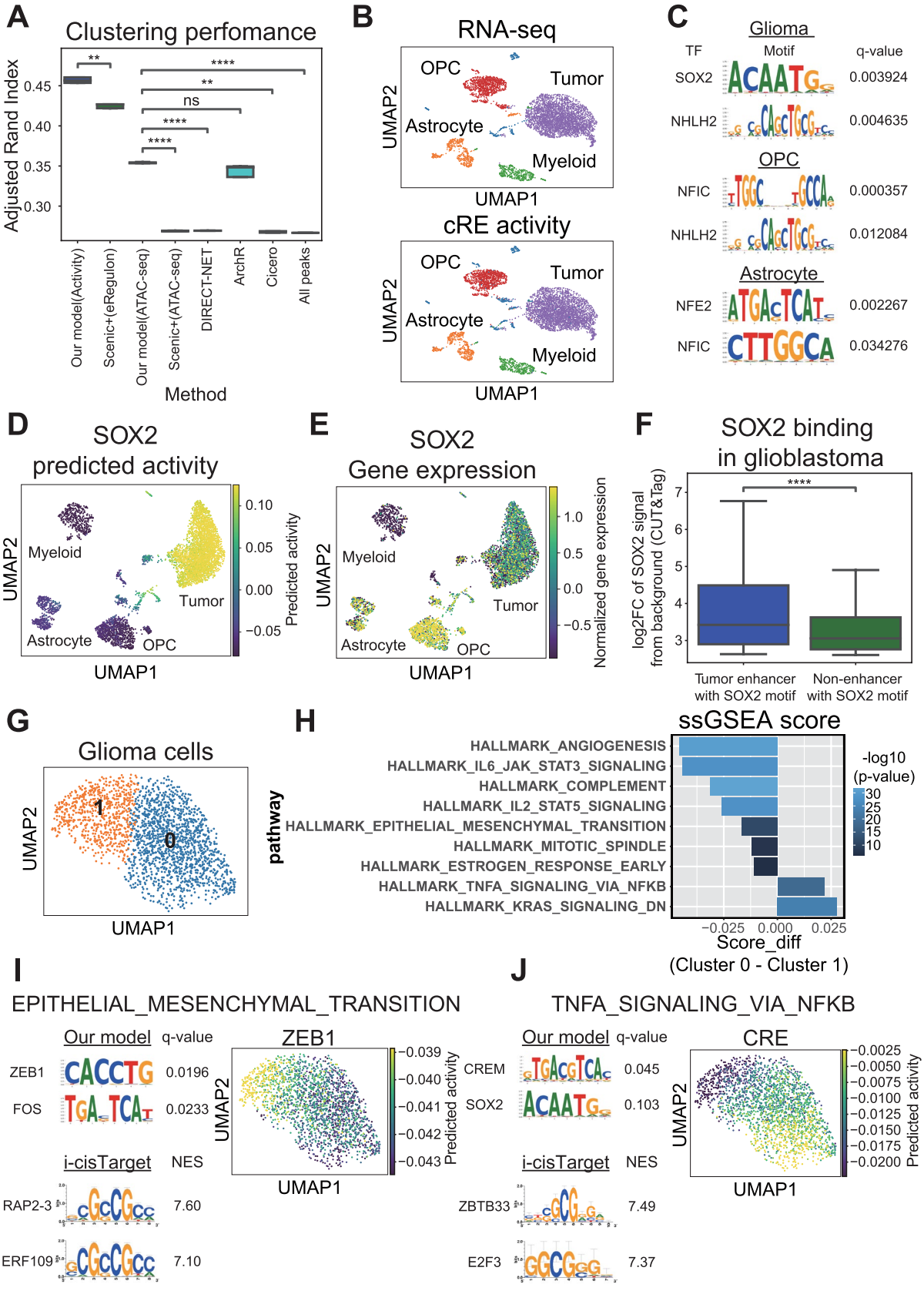

FIGURE 3 | Legend on next page.

Finally, we analyzed the intra-tumor heterogeneity of cRE regulation. After clustering tumor cells into two groups using k-means on predicted cRE activity (Figure 3G), we examined their differences in gene expression levels using ssGSEA (Subramanian et al. 2005). Cluster 1 showed elevated expression of angiogenesis- and epithelial-mesenchymal transition (EMT)-related genes (Figure 3H). Focusing on EMT-related genes, we identify TFs regulating the cREs associated with these genes by our model with TF-MoDISCo (Shrikumar et al. 2020) (Figure 3I, left). ZEB1, a well-known EMT-related TF (Zhang, Sun, and Ma 2015), emerged as a promising candidate. Consistent with this, we observed a gradual increase in ZEB1-predicted activity from Cluster 0 to Cluster 1 (Figure 3I, right), whereas existing statistical methods (i-cisTarget; Herrmann et al. 2012) failed to identify significant motifs in the EMT-related cREs (Figure 3I, bottom-left). Thus, our model can detect intra-tumor heterogeneity in EMT-related cell states and identify key TFs that regulate EMT-related genes.

On the other hand, genes involved in TNF-α and NFκB signaling pathways were more expressed in cluster 0. However, when analyzing the cREs associated with these genes using TF-MoDISCo, the binding motif for NFκB was not enriched. Instead, CREM, a TF belonging to the cAMP response element (CRE) family, was identified as the top motif (Figure 3J, left). CREB1, a member of the CRE family, has been reported to co-regulate NFκB target genes with RelA (Nakayama 2013). Consistent with this finding, the predicted CRE activity was specific to cluster 0 (Figure 3J, right). In conclusion, our model was able to pinpoint subpopulation-specific TF signatures within tumors using single-cell multi-omics data.

3 | Discussion

This study proposed a new method for detecting cRE activity at the single-cell level using deep neural networks equipped with attention mechanisms. Our model integrates DNA sequence information, chromatin accessibility, and genomic distance within the model architecture, leading to a more accurate detection of cRE activity compared to previous methods. Notably, our model learns transcriptional regulation rules in a data-driven manner by incorporating genomic distance from the TSS, which is manually included in other models (Pliner et al. 2018; Granja et al. 2021). This data-driven approach allows the model to handle distance information more effectively than rule-based methods, such as Cicero and ArchR (Pliner et al. 2018; Granja et al. 2021).

Additionally, our model could detect differences in cRE regulation between cell types more accurately than previous tools. In our analysis of glioma samples, the model accurately identified the SOX2 transcriptional activity specific to gliomas. Additionally, intra-tumor analysis revealed heterogeneity in the regulation of EMT-related genes, which are key factors in metastasis. The model also identified the critical TFs involved in the regulation of these cREs. When tumors acquire metastatic abilities or become resistant to treatment, only a subset of tumor cells develops these characteristics (Dagogo-Jack and Shaw 2018). Additionally, in certain cancers, such as acute myeloid leukemia, only a fraction of cells, such as leukemia stem cells, possess the capacity to regenerate large numbers of tumor cells (Shlush et al. 2014). Therefore, to assess tumor pathology and identify precise therapeutic targets, it is crucial to measure transcriptional regulation in specific cell subpopulations. Our model enables high-resolution analysis of specific gene sets within defined cell populations, making it a valuable tool for multi-omics single-cell data analysis.

In addition, our model was designed to analyze a smaller size of clinical samples of any disease in a hospital setting without requiring pre-training on large datasets. This allowed us to assess cRE activity at the single-cell level, providing insights into the heterogeneity of cRE regulation in rare diseases. However, a limitation of our method is that deep learning-based regression models capture only the correlations between inputs and outputs. Thus, a high contribution score to the gene expression level does not guarantee a causal relationship between the detected cREs and target genes. To address this, one potential solution involves training the model on a large dataset and using causal analysis approaches such as a causal transformer (Melnychuk, Frauen, and Feuerriegel 2022). Another approach involves experimental validation. Nevertheless, our model prioritizes versatility and applicability to single

FIGURE 3 | The analysis of tumor-specific cRE regulation. (A) Clustering performance in PBMC data. The Y-axis represents the adjusted Rand index. The first two plots on the left show the clustering performance of predicted cRE activity from our model and eRegulon activity from Scenic+ (Bravo González-Blas et al. 2023). The remaining plots compare clustering performance based on scATAC-seq counts, including the top 15% of high-activity peaks predicted by each tool and the total scATAC-seq counts. ns, $5.0*10^{-2} < p$; $^{**}p \leq 1.0*10^{-2}$; $^{****}p \leq 1.0*10^{-4}$ (vs. our model, Student's *t*-test, unpaired). (B) UMAP of glioma samples colored by the scRNA-seq counts and predicted cRE activities. Cell embeddings were generated from scRNA-seq counts of all genes. Colors represent cell type annotations based on scRNA-seq counts of all genes or the top 15% of cREs with the highest activity. (C) The cell type-specific transcription factors detected by our model and TF-MoDISCo. Significant motifs, defined by a q-value of < 0.05, were plotted for gliomas, OPC, and astrocytes. The top two motifs were highlighted in each case. (D) Predicted SOX2 activity at the single-cell level. The color indicates the predicted SOX2 activity. (E) SOX2 gene expression levels in scRNA-seq counts. Colors indicate log-normalized and scaled gene expression levels. (F) SOX2 CUT&Tag signaling between tumor-specific cREs and non-cRE regions with a SOX2 binding motif. The Y-axis represents the $\log_2$ fold change in the SOX2 CUT&Tag signal from glioma cell samples relative to the background signal. $p = 2.4*10^{-10}$ (Student's *t*-test, unpaired). (G) Clustering of glioma cells based on predicted cRE activity. The UMAP is based on the cRE activities. (H) Enrichment of gene sets in intra-tumor clusters by ssGSEA analysis. Statistical analysis was performed using the Mann–Whitney U test. (I, J) Gene sets and transcription factors are involved in the significant enrichment of each intra-tumor cluster. The motif enrichment analysis was performed using TF-MoDISCo (Shrikumar et al. 2020) and i-cisTarget (Herrmann et al. 2012). (I) Results for "HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION" and (J) "HALLMARK_TNFA_SIGNALING_VIA_NFKB." cRE, cis-regulatory elements; OPCs, oligodendrocyte precursor cells; PBMC, peripheral blood mononuclear cells; scATAC-seq, single-cell assay for transposase-accessible chromatin sequencing; scRNA-seq, single-cell RNA sequencing; UMAP, Uniform Manifold Approximation and Projection.

samples by not relying on extensive pre-training on large datasets. Another limitation is that it may be difficult to detect cRE activity that is consistently high in all cells. As this model relies on differences in activity and gene expression levels between cells, it is expected that it will not be possible to learn about cREs that do not differ between cells. To address this issue, analysts should carefully select appropriate reference cell types based on the analysis goal and include these cells in the samples.

Another limitation of this study is that the performance evaluation was conducted exclusively using PBMC data. As highlighted in previous research (Zhang, Yang, and Zhang 2022), the states of scATAC-seq and scRNA-seq may not be fully compatible in samples where cellular states change continuously, making learning more challenging. To stabilize learning, it is important to include differentiated cell types in the dataset, such as normal cells along tumor cells.

As a future direction, developing a batch correction method capable of learning from multiple batches simultaneously is worth considering. Previous studies (Yuan and Kelley 2022) have proposed using L1 regularization for scATAC-seq correction in the context of batch correction and learning across multiple batches using deep learning methods. Similarly, MultiVI (Ashuach et al. 2023) presents an approach for integrating multiple batches of scMultiome data by learning latent representations of cells through deep learning. However, these methods mainly primarily focus on removing batch effects at the cellular level, and it remains unclear whether such corrections are sufficient for high-resolution analyses, such as gene-cRE pair prediction. Incorporating batch correction into the training process for gene-cRE pair prediction in this model could facilitate the learning of cRE regulation from a large, integrated dataset spanning multiple batches.

4 | Experimental Procedures

4.1 | Preparation of scMultiome

The PBMC dataset comprised "Peripheral Blood Mononuclear Cells from a Healthy Donor, with Granulocytes Removed Through Cell Sorting (10k)," sourced from the 10× website (10× Genomics 2021). We analyzed the respective available count data, namely "pbmc_granulocyte_sorted_10k_filtered_feature_bc_matrix.h5" for PBMC datasets. Subsequently, scRNA-seq and scATAC-seq data were processed using Scanpy (version 1.9.3) (Wolf, Angerer, and Theis 2018). Only genes expressed in 5% or more of the cells and cells expressing 500 or more genes were included for downstream analysis.

4.2 | Annotation of Cells for PBMCs

Cell clustering was conducted using Seurat (version 4.3.0) (Hao et al. 2021) with the parameter "resolution = 0.1." The cells were annotated using marker genes. We used S100A8 and CD14 for monocytes, CD3D and CD4 for $CD4^+$ T cells, PRF1 for $CD8^+$ T cells, CD19 and EBF1 for B cells, and CD1C for dendritic cells.

4.3 | Prediction Model for Gene Expression Levels

Our framework was designed to use single sample scMultiome data as the input. For each gene, the TSS position in the GRCh38.p13 genome was obtained from the Ensembl database using BiomaRt (version 2.50.3) (Durinck et al. 2005). First, ATAC-seq peaks located within ± 300,000 bp of the TSS were considered cRE candidates for the respective gene. The peak located within ± 500 bp of the TSS of the genes was identified as the promoter peak. In cases where multiple candidate promoter peaks fell within this range, the closest peak was defined as the promoter peak. The promoter peak of each gene was excluded from the cRE candidates. Genes lacking promoter peaks or cRE candidates were excluded from the downstream analysis. The model was trained on a per-gene basis, with each gene constituting a single dataset. Specifically, the input consisted of chromatin accessibility counts for the cRE candidates/promoter, DNA sequence, and distance from the TSS, whereas the output was normalized to gene expression. Our gene expression prediction model comprised two components: a DNA sequence encoder and an attention-based block. The DNA sequence encoder was an 8-layer CNN that accepted a one-hot vector representation of a 1344 bp nucleotide sequence as input and yielded a compressed 32-dimensional representation. The attention-based block initially processed the 32-dimensional DNA sequence features of cRE candidates and promoters and 1-dimensional distance information through a layer of a Feed Forward Neural Network (FFNN), followed by the computation of an attention matrix via the matrix product. Subsequently, the product of the attention matrix and ATAC-seq counts was used to integrate the ATAC-seq information with the DNA sequence information. The resulting matrix was normalized using Instance Normalization. The ATAC-seq information of the promoter was subsequently added and a feature matrix was computed. Finally, the gene expression levels were derived from the feature matrix using a two-layer FFNN. The model structure is summarized in Figure S4.

4.4 | Model Training

First, the dataset was divided into training and testing datasets. Initially, 10% of the input ATAC-seq peaks were earmarked as test data, whereas the remaining 90% constituted the training set. Only genes with all the cRE candidates and promoters included in the training peaks were designated as training genes; the remainder were allocated to the test set. In terms of cells, 80% were designated as training cells, and the remaining 20% were assigned to the test set. The initial step involved training the DNA sequence encoder using a method similar to that used in the scBasset (Yuan and Kelley 2022). Specifically, we extracted a 32-dimensional peak embedding from the DNA sequence and trained the model to predict ATAC-seq counts at the single-cell level based on these features. Pre-training was conducted for 1000 epochs. The outputs from the DNA-seq encoder, such as 32-dimensional feature vectors and predicted ATAC-seq counts, were used for analyzing the peak embedding and noise-suppressed ATAC-seq counts, respectively. Subsequently, the scRNA-seq count matrix was compressed using PCA with the PCA transition matrix derived solely from the training cells. This compression reduced the scRNA-seq count matrix to 50 dimensions.

Attention-based block training was then performed using a 32-dimensional compressed sequence representation, ATAC-seq, and distance information as input. The model was trained to predict 50-dimensional cellular coordinates compressed using PCA at the single-cell level. The model training was run for 500 epochs, with early stopping if the test loss did not improve after 10 consecutive epochs. After training, the predicted gene expression levels were calculated by inverting the PCA coordinates. Notably, to evaluate the prediction accuracy, a 10-fold cross-validation was conducted.

4.5 | Acquisition of cRE Activity

cRE activity was determined by assessing the contribution of the input ATAC-seq count to gene expression levels. Contribution scores were calculated using the DeepLift function in the Captum package (version 0.6.0) (Kokhlikyan et al. 2020). For analyses other than benchmarking of cRE prediction, we only used regions with a positive correlation between ATAC-seq counts and cRE activity to unify the interpretability of the results.

4.6 | Detection of Cell Type-Specific cREs

The cell-type specificity of each cRE was determined by comparing its activity in a given cell type to its activity across all other cell types. Statistical significance was assessed using the Mann–Whitney U test. Peaks that met the criteria of adjusted p-value <0.0001 and a mean activity difference >0.01 were classified as cell-type-specific cREs. The cRE region was defined as the region with the top 60% of activity among the cRE candidates. The cutoff value is based on the observation that, empirically, the mode of cRE activity typically falls within the top 60%–70%.

4.7 | Motif Analysis Using the Proposed Model

To detect cell type-specific TF-binding motifs, we first identified cell type-specific genes. This was performed using Seurat (version 4.3.0) (Hao et al. 2021), where differentially expressed genes (DEGs) were defined as those with a $\log_2$ fold change greater than 0 and p-value <0.05 when comparing the target clusters to other clusters. Cell type-specific cREs related to these DEGs were used for motif analysis. In order to suppress the noise in the contribution score, only cells in which the cREs linked to that gene are constantly active were used in the analysis, and cells in which cRE activity was predicted to be high by chance were excluded. Then, the cells having the top 25% cRE activities in over 70% of cREs related to the target gene are included in the analysis. This analysis was performed using TF-MoDISCo (lite version 1.0.0) (Shrikumar et al. 2020), which utilized the contribution scores calculated using our model. The position weight matrices (PWMs) identified by TF-MoDISCo were then compared with the JASPAR CORE 2024 vertebrate (non-redundant) database (Rauluseviciute et al. 2024). TF motifs with a q-value <0.05 were considered statistically significant. When searching for motifs related to a certain gene set, we used the target gene set instead of DEGs.

4.8 | Evaluation of cRE Prediction Results

The cRE prediction results were evaluated using data from the FANTOM5 database (Lizio et al. 2019) and PCHiC data (Javierre et al. 2016) for PBMCs. The enhancer list of the FANTOM5 database was derived from "F5.hg38.enhancers.bed.gz", which is the curated list of enhancers in the FANTOM5 database. PCHiC data were obtained from "PCHiC_peak_matrix_cutoff5.tsv." The FANTOM5 data includes enhancer regions curated from enhancer data across all human cell types registered in the FANTOM5 database, whereas the PCHiC data comprises data derived from PBMCs. Among the list of ATAC-seq peaks used as model inputs, regions that overlapped by more than a single-base pair with FANTOM5 enhancers were considered FANTOM5 enhancer regions. As with the analysis of the FANTOM5 dataset, we calculated the overlap between the ATAC-seq peaks used as input for the model and PCHiC peaks. The cRE candidates in which the promoter region of PCHiC overlapped with the model promoter region and the other end of PCHiC overlapped with the cRE candidates were identified as cRE regions. Since the PCHiC data was mapped to the hg19 genome, it was converted to the hg38 genome using UCSC liftOver (ver 469) (Hinrichs et al. 2006).

The predictive performance of our model was compared with that of Cicero, ArchR, DIRECT-NET, and Scenic+ (Pliner et al. 2018; Granja et al. 2021; Zhang, Zhang, and Nie 2022; Bravo González-Blas et al. 2023). Each detected cRE-gene interaction was identified according to the respective tutorial of the tool. To compare performance, we calculated the overlap of cRE-gene pairs detected by each tool with the ATAC-seq peak regions that served as the input to our model. The overlapping ATAC-seq peak gene pairs were considered as the cRE-gene pairs detected by each tool. Prediction performance was quantified as the AUROC using the roc_auc_score function in sklearn (version 1.4.0) (Pedregosa et al. 2011). To calculate the AUROC per distance, the distance from the TSS was divided into 20 bins ranging from 0 to 28,000 bp. AUROC was then calculated for the cRE-gene pairs within each bin.

4.9 | Processing of eQTL Data

The eQTL data were obtained from the GTEx v8 database (GTEx Consortium 2020). Only variants with PIP >0.5 were defined as causal variants for those detected in whole blood cells. The enrichment ratio was calculated as follows (Sakaue et al. 2024):

$$\text{Enrichment ratio} = \frac{\text{The number of causal variants overlapping cREs}}{\text{The number of common variants overlapping cREs}} \div \frac{\text{The number of causal variants overlapping all input peaks}}{\text{The number of common variants overlapping all input peaks}}$$

4.10 | Calculating the Distance From the TSS

The TSS positions of all genes in the GRCh38.p13 genome were obtained from the Ensembl database using the biomaRt package (version 2.50.3) (Durinck et al. 2005). Distances between all candidate cREs and the TSS of the closest gene were calculated.

4.11 | Analysis of Cell Type-Specific Genes and Genes With Ubiquitous Expression

Genes with cell type-specific expression were selected using Seurat (version 4.3.0) (Hao et al. 2021), with the criteria of $\log_2$ FC > 0 and p-value < 0.05 in each cluster. Genes not included in the set of genes with cell-type-specific expression were defined as those with ubiquitous expression. The accuracy of the gene expression prediction was evaluated using Spearman's correlation coefficient. Fisher's Z-transformation was used for the statistical test of the correlation coefficient, and Student's t-test was used.

4.12 | Benchmarking of Downsampling Analysis

The PBMC data described above was used for the downsampling analysis. Cells were randomly selected from the original data (10k), and the data was downsampled to 100%, 50%, and 25%. The cRE regions were then predicted for each downsampled dataset. Prediction accuracy was evaluated using the same methodology described earlier.

4.13 | Benchmarking of Clustering Performance

Clustering performance was benchmarked using ARI, with scRNA-seq-based clustering serving as the ground truth. The standard Seurat pipeline (version 4.3.0) (Hao et al. 2021) was used for scRNA-seq clustering. Specifically, following normalization, the top 2000 variable features were selected for PCA. Clustering was then performed using PCA coordinates with the resolution parameter set to 0.1, resulting in the identification of nine distinct clusters. For clustering based on the predicted cRE activity from our model, the top 15% of the regions with the highest predicted cRE activities were selected, and k-means clustering was performed to group the cells into nine clusters based on these predicted activities. In the case of eRegulon clustering from Scenic+ (Bravo González-Blas et al. 2023), the "eRegulon_AUC" values generated by Scenic+ were used for clustering. For clustering based on scATAC-seq counts, the top 15% of regions showing the highest activity, as identified by each computational tool, were used. Cells were grouped into nine clusters using k-means clustering. For the clustering based on "all peaks," all scATAC-seq peaks were included in the analysis. ARI was calculated using the "adjusted_rand_score" function, and k-means clustering was conducted using the "K-Means" function from the sklearn package (version 1.4.0) (Pedregosa et al. 2011). Statistical comparisons were made using the Student's t-test, and differences with p-value < 0.05 were considered statistically significant.

4.14 | Analysis of the Sample From Patients With Glioma

We obtained the scMultiome data for a glioma sample from NCBI GSE210568 (Jessa et al. 2022). Specifically, we analyzed the sample labeled "P-1694_S-1694_multiome." Data were processed using the same methodology that was applied to PBMCs, including the prediction of cRE activity. The raw data included 5530 cells, 60,658 genes, and 107,873 peaks. After filtering, the dataset for analysis comprised 5304 cells, 7805 genes, and 33,266 peaks. The maximum number of peaks associated with a single gene was 69. Uniform Manifold Approximation and Projection (UMAP) visualization was generated based on the predicted activity of the top 15% of the most active cREs using Scanpy (version 1.9.3) (Wolf, Angerer, and Theis 2018).

4.15 | Cell Annotation for the Glioma Sample

Cell annotation was performed based on the marker genes reported in the existing literature. Specifically, PDGFA and FGFR2 for gliomas (Verhaak et al. 2010; Jimenez-Pascual and Siebzehnrubl 2019), PDGFRA for OPCs (Pringle et al. 1992), GFAP for astrocytes (Hol and Pekny 2015), and S100A8 for myeloid cells (Odink et al. 1987) were used.

4.16 | Detection of TF Binding Regions

The cREs bound by each TF were identified using the FIMO function from the MEME Suite (version 5.5.5) (Bailey et al. 2015), with PWM files sourced from the JASPAR CORE 2024 vertebrate (non-redundant) database. FIMO was run with a threshold of 0.001 ('—threshold 0.001'). The top 5000 highest-scoring matches were selected and defined as TF-binding regions.

4.17 | Calculation of TF Activity at the Single-Cell Level

The predicted activity of TFs at the single-cell level was calculated as the average activity of cREs containing TF-binding motifs in each cell.

4.18 | SOX2 CUT&tag Analysis

The SOX2 CUT&Tag data were obtained from NCBI GSE200062 (Benedetti et al. 2022). Specifically, the file "GSM6008250_SMNB19_SOX2_3_broadPeak.bed.gz" was used for the analysis. From the results of our model, we identified tumor-specific cREs and non-cRE regions containing SOX2 binding sites for subsequent analysis. Overlapping SOX2 CUT&Tag broad peaks were detected for each group. To detect these overlapping peaks, we used the intersect function of Bedtools (version 2.30.0) (Quinlan and Hall 2010) with the parameters "-wa -u -F 1 -a cRE regions from models -b SOX2 CUT&Tag peaks files". The average $\log_2$ fold change from the background signal was calculated for each SOX2 CUT&Tag peak group. Statistical significance was evaluated using Student's t-test, with p-value < 0.05.

4.19 | Clustering of Tumor Cells

The top 15% of cREs with high predicted activity were selected to map glioma cells using UMAP. The glioma cells were then clustered using the sc.tl.leiden function in Scanpy (version 1.9.3)

(Wolf, Angerer, and Theis 2018), with a resolution parameter of 0.2, resulting in two distinct clusters.

4.20 | ssGSEA Analysis

First, to identify the gene groups regulated by cREs, we performed gene ontology analysis on the target genes of the top 15% of the most active predicted cREs using the H collection from the Molecular Signatures Database (MSigDB) (Liberzon et al. 2015). The "msigdbr" package (version 7.5.1) (Dolgalev 2022) in R was used to obtain the gene sets. Gene sets with an adjusted p-value <0.05 were considered to be regulated by cREs. Next, we used the ssGSEA package to assess the enrichment of these cRE-regulated gene sets in glioma cells and calculated the enrichment score at the single-cell level. We then used the Mann–Whitney U test to identify cluster-specific gene sets, with those having a p-value <0.05 deemed significant. Finally, we calculated the differences in the average enrichment scores between the clusters.

4.21 | TF Analysis of Gene Sets

First, cREs associated with the gene sets "HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION" and "HALLMARK_TNFA_SIGNALING_VIA_NFKB" were identified. The contribution scores for these cREs were calculated at the single-nucleotide level, followed by motif analysis using TF-MoDISCo (lite version 1.0.0) (Shrikumar et al. 2020). TF motifs were then identified by comparing them with the JASPAR CORE 2024 vertebrate (non-redundant) motif database using a significance threshold of q-value <0.05.

To complement this, conventional motif enrichment analysis was performed on the cRE regions corresponding to each gene set using i-cisTarget (Herrmann et al. 2012). Finally, the predicted TF activities of ZEB1 and CREM in the glioma cells were calculated and visualized as cRE contribution scores.

Author Contributions

Ken Murakami: conceptualization, investigation, writing – original draft, writing – review and editing, methodology, software, formal analysis, data curation, visualization. **Keita Iida:** investigation, funding acquisition, writing – review and editing, validation, supervision, data curation, project administration, methodology. **Mariko Okada:** conceptualization, investigation, funding acquisition, writing – original draft, writing – review and editing, supervision, project administration, data curation.

Acknowledgments

We would like to thank Dr. Hidetoshi Shimodaira (Kyoto Univ.), Dr. Makoto Taiji (RIKEN), Dr. Masako Iwasaki (Osaka Metropolitan Univ. & Osaka Univ.), Dr. Hajime Nagahara (Osaka Univ.), Dr. Yuta Nakashima (Osaka Univ.), Dr. Hideaki Hayashi (Osaka Univ.), Dr. Naoki Hosen (Osaka Univ.), Dr. Michiko Ichii (Osaka Univ.) and all members of JST CREST Bio-DX for their meaningful suggestions and discussion. We would like to thank Dr. Vincent Loubiere (Research Institute of Molecular Pathology) for helpful discussions and advice regarding the manuscript.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The PBMC scMultiome dataset, "Peripheral Blood Mononuclear Cells from a Healthy Donor, with Granulocytes Removed Through Cell Sorting (10k)" was obtained from the 10× website (https://www.10xgenomics.com/datasets/pbmc-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-10-k-1-standard-2-0-0). The enhancer list from the FANTOM5 database, "F5.hg38.enhancers.bed.gz," was obtained from FANTOM5 website (https://fantom.gsc.riken.jp/5/). The promoter capture HiC (PCHiC) data, "PCHiC_peak_matrix_cutoff5.tsv," was obtained from the Supplemental Information of Javierre et al. (2016) ("Data S1"; https://www.cell.com/cms/10.1016/j.cell.2016.09.037/attachment/5bc79f6f-1b69-4192-8cb8-4247cc2e0f39/mmc4.zip). The scMultiome data for a glioma sample "P-1694_S-1694_multiome" was obtained from NCBI GSE210568 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE210568). SOX2 CUT&Tag data, "GSM6008250_SMNB19_SOX2_3_broadPeak.bed.gz" was obtained from NCBI GSE200062 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE200062). The code used in this study is available at https://github.com/okadalabipr/cREscENDO.

References

"10x Genomics. 2021. PBMC from a Healthy Donor – Granulocytes Removed Through Cell Sorting (10k), Single Cell Multiome ATAC + Gene Expression Dataset by Cell Ranger ARC v2.0.0." https://www.10xgenomics.com/datasets/pbmc-from-a-healthy-donor-granulocytes-removed-through-cell-sorting-10-k-1-standard-2-0-0.

Ashuach, T., M. I. Gabitto, R. V. Koodli, G. A. Saldi, M. I. Jordan, and N. Yosef. 2023. "MultiVI: Deep Generative Model for the Integration of Multimodal Data." *Nature Methods* 20, no. 8: 1222–1231. https://doi.org/10.1038/s41592-023-01909-9.

Avsec, Ž., V. Agarwal, D. Visentin, et al. 2021a. "Effective Gene Expression Prediction From Sequence by Integrating Long-Range Interactions." *Nature Methods* 18, no. 10: 1196–1203. https://doi.org/10.1038/s41592-021-01252-x.

Avsec, Ž., M. Weilert, A. Shrikumar, et al. 2021b. "Base-Resolution Models of Transcription-Factor Binding Reveal Soft Motif Syntax." *Nature Genetics* 53, no. 3: 354–366. https://doi.org/10.1038/s41588-021-00782-6.

Bailey, T. L., J. Johnson, C. E. Grant, and W. S. Noble. 2015. "The MEME Suite." *Nucleic Acids Research* 43, no. W1: W39–W49. https://doi.org/10.1093/nar/gkv416.

Benedetti, V., F. Banfi, M. Zaghi, et al. 2022. "A SOX2-Engineered Epigenetic Silencer Factor Represses the Glioblastoma Genetic Program and Restrains Tumor Development. Science." *Advances* 8, no. 31: eabn3986. https://doi.org/10.1126/sciadv.abn3986.

Bergman, D. T., T. R. Jones, V. Liu, et al. 2022. "Compatibility Rules of Human Enhancer and Promoter Sequences." *Nature* 607, no. 7917: 176–184. https://doi.org/10.1038/s41586-022-04877-w.

Bohrer, C. H., and D. R. Larson. 2021. "The Stochastic Genome and Its Role in Gene Expression." *Cold Spring Harbor Perspectives in Biology* 13, no. 10: a040386. https://doi.org/10.1101/cshperspect.a040386.

Bravo González-Blas, C., S. de Winter, G. Hulselmans, et al. 2023. "SCENIC+: Single-Cell Multiomic Inference of Enhancers and Gene Regulatory Networks." *Nature Methods* 20, no. 9: 1355–1367. https://doi.org/10.1038/s41592-023-01938-4.

Buenrostro, J. D., B. Wu, U. M. Litzenburger, et al. 2015. "Single-Cell Chromatin Accessibility Reveals Principles of Regulatory Variation." *Nature* 523, no. 7561: 486–490. https://doi.org/10.1038/nature14590.

Chen, T., and C. Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." arXiv. https://doi.org/10.1145/2939672.2939785.

Dagogo-Jack, I., and A. T. Shaw. 2018. "Tumour Heterogeneity and Resistance to Cancer Therapies." *Nature Reviews. Clinical Oncology* 15, no. 2: 81–94. https://doi.org/10.1038/nrclinonc.2017.166.

de Almeida, B. P., F. Reiter, M. Pagani, and A. Stark. 2022. "DeepSTARR Predicts Enhancer Activity From DNA Sequence and Enables the de Novo Design of Synthetic Enhancers." *Nature Genetics* 54, no. 5: 613–624. https://doi.org/10.1038/s41588-022-01048-5.

de Almeida, B. P., C. Schaub, M. Pagani, S. Secchia, E. E. M. Furlong, and A. Stark. 2024. "Targeted Design of Synthetic Enhancers for Selected Tissues in the Drosophila Embryo." *Nature* 626, no. 7997: 207–211. https://doi.org/10.1038/s41586-023-06905-9.

Dolgalev, I. 2022. "msigdbr: MSigDB Gene Sets for Multiple Organisms in a Tidy Data Format." https://igordot.github.io/msigdbr/.

Durinck, S., Y. Moreau, A. Kasprzyk, et al. 2005. "BioMart and Bioconductor: A Powerful Link Between Biological Databases and Microarray Data Analysis." *Bioinformatics* 21, no. 16: 3439–3440. https://doi.org/10.1093/bioinformatics/bti525.

Frazel, P. W., D. Labib, T. Fisher, et al. 2023. "Longitudinal scRNA-Seq Analysis in Mouse and Human Informs Optimization of Rapid Mouse Astrocyte Differentiation Protocols." *Nature Neuroscience* 26, no. 10: 1726–1738. https://doi.org/10.1038/s41593-023-01424-2.

Garros-Regulez, L., I. Garcia, E. Carrasco-Garcia, et al. 2016. "Targeting SOX2 as a Therapeutic Strategy in Glioblastoma." *Frontiers in Oncology* 6: 222. https://doi.org/10.3389/fonc.2016.00222.

Granja, J. M., M. R. Corces, S. E. Pierce, et al. 2021. "ArchR Is a Scalable Software Package for Integrative Single-Cell Chromatin Accessibility Analysis." *Nature Genetics* 53, no. 3: 403–411. https://doi.org/10.1038/s41588-021-00790-6.

GTEx Consortium. 2020. "The GTEx Consortium Atlas of Genetic Regulatory Effects Across Human Tissues." *Science* 369, no. 6509: 1318–1330. https://doi.org/10.1126/science.aaz1776.

Hao, Y., S. Hao, E. Andersen-Nissen, et al. 2021. "Integrated Analysis of Multimodal Single-Cell Data." *Cell* 184, no. 13: 3573–3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.

Herrmann, C., B. Van de Sande, D. Potier, and S. Aerts. 2012. "I-cisTarget: An Integrative Genomics Method for the Prediction of Regulatory Features and Cis-Regulatory Modules." *Nucleic Acids Research* 40, no. 15: e114. https://doi.org/10.1093/nar/gks543.

Hinrichs, A. S., D. Karolchik, R. Baertsch, et al. 2006. "The UCSC Genome Browser Database: Update 2006." *Nucleic Acids Research* 34: D590–D598. https://doi.org/10.1093/nar/gkj144.

Hol, E. M., and M. Pekny. 2015. "Glial Fibrillary Acidic Protein (GFAP) and the Astrocyte Intermediate Filament System in Diseases of the Central Nervous System." *Current Opinion in Cell Biology* 32: 121–130. https://doi.org/10.1016/j.ceb.2015.02.004.

Javierre, B. M., O. S. Burren, S. P. Wilder, et al. 2016. "Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters." *Cell* 167, no. 5: 1369–1384.e19. https://doi.org/10.1016/j.cell.2016.09.037.

Jessa, S., A. Mohammadnia, A. S. Harutyunyan, et al. 2022. "K27M in Canonical and Noncanonical H3 Variants Occurs in Distinct Oligodendroglial Cell Lineages in Brain Midline Gliomas." *Nature Genetics* 54, no. 12: 1865–1880. https://doi.org/10.1038/s41588-022-01205-w.

Jimenez-Pascual, A., and F. A. Siebzehnrubl. 2019. "Fibroblast Growth Factor Receptor Functions in Glioblastoma." *Cells* 8, no. 7: 715. https://doi.org/10.3390/cells8070715.

Kelley, D. R., Y. A. Reshef, M. Bileschi, D. Belanger, C. Y. McLean, and J. Snoek. 2018. "Sequential Regulatory Activity Prediction Across Chromosomes With Convolutional Neural Networks." *Genome Research* 28, no. 5: 739–750. https://doi.org/10.1101/gr.227819.117.

Kokhlikyan, N., V. Miglani, M. Martin, et al. 2020. "Captum: A unified and generic model interpretability library for PyTorch." arXiv. https://arxiv.org/abs/2009.07896.

Liberzon, A., C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, and P. Tamayo. 2015. "The Molecular Signatures Database (MSigDB) Hallmark Gene Set Collection." *Cell Systems* 1, no. 6: 417–425. https://doi.org/10.1016/j.cels.2015.12.004.

Lizio, M., I. Abugessaisa, S. Noguchi, et al. 2019. "Update of the FANTOM Web Resource: Expansion to Provide Additional Transcriptome Atlases." *Nucleic Acids Research* 47, no. D1: D752–D758. https://doi.org/10.1093/nar/gky1099.

Marusyk, A., M. Janiszewska, and K. Polyak. 2020. "Intratumor Heterogeneity: The Rosetta Stone of Therapy Resistance." *Cancer Cell* 37, no. 4: 471–484. https://doi.org/10.1016/j.ccell.2020.03.007.

Melnychuk, V., D. Frauen, and S. Feuerriegel. 2022. "Causal Transformer for Estimating Counterfactual Outcomes." arXiv. https://arxiv.org/abs/2204.07258.

Moore, J. E., M. J. Purcaro, H. E. Pratt, et al. 2020. "Expanded Encyclopaedias of DNA Elements in the Human and Mouse Genomes." *Nature* 583, no. 7818: 699–710. https://doi.org/10.1038/s41586-020-2493-4.

Nakayama, K. 2013. "cAMP-Response Element-Binding Protein (CREB) and NF-κB Transcription Factors Are Activated During Prolonged Hypoxia and Cooperatively Regulate the Induction of Matrix Metalloproteinase MMP1." *Journal of Biological Chemistry* 288, no. 31: 22584–22595. https://doi.org/10.1074/jbc.M112.421636.

Odink, K., N. Cerletti, J. Brüggen, et al. 1987. "Two Calcium-Binding Proteins in Infiltrate Macrophages of Rheumatoid Arthritis." *Nature* 330, no. 6143: 80–82. https://doi.org/10.1038/330080a0.

Panigrahi, A., and B. W. O'Malley. 2021. "Mechanisms of Enhancer Action: The Known and the Unknown." *Genome Biology* 22, no. 1: 108. https://doi.org/10.1186/s13059-021-02322-1.

Pedregosa, F., G. Varoquaux, A. Gramfort, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–2830.

Pliner, H. A., J. S. Packer, J. L. McFaline-Figueroa, et al. 2018. "Cicero Predicts Cis-Regulatory DNA Interactions From Single-Cell Chromatin Accessibility Data." *Molecular Cell* 71, no. 5: 858–871.e8. https://doi.org/10.1016/j.molcel.2018.06.044.

Pringle, N. P., H. S. Mudhar, E. J. Collarini, and W. D. Richardson. 1992. "PDGF Receptors in the Rat CNS: During Late Neurogenesis, PDGF Alpha-Receptor Expression Appears to Be Restricted to Glial Cells of the Oligodendrocyte Lineage." *Development* 115, no. 2: 535–551. https://doi.org/10.1242/dev.115.2.535.

Quinlan, A. R., and I. M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26, no. 6: 841–842. https://doi.org/10.1093/bioinformatics/btq033.

Rauluseviciute, I., R. Riudavets-Puig, R. Blanc-Mathieu, et al. 2024. "JASPAR 2024: 20th Anniversary of the Open-Access Database of Transcription Factor Binding Profiles." *Nucleic Acids Research* 52, no. D1: D174–D182. https://doi.org/10.1093/nar/gkad1059.

Rombach, R., A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. 2022. "High-Resolution Image Synthesis with Latent Diffusion Models." arXiv. https://arxiv.org/abs/2112.10752.

Sakaue, S., K. Weinand, S. Isaac, et al. 2024. "Tissue-specific enhancer-gene maps from multimodal single-cell data identify causal disease alleles." *Nature Genetics* 56, no. 4: 615–626. https://doi.org/10.1038/s41588-024-01682-1.

Scott, L. M., C. I. Civin, P. Rorth, and A. D. Friedman. 1992. "A Novel Temporal Expression Pattern of Three C/EBP Family Members in Differentiating Myelomonocytic Cells." *Blood* 80, no. 7: 1725–1735.

Shlush, L. I., S. Zandi, A. Mitchell, et al. 2014. "Identification of Pre-Leukaemic Haematopoietic Stem Cells in Acute Leukaemia." *Nature* 506, no. 7488: 328–333. https://doi.org/10.1038/nature13038.

Shrikumar, A., P. Greenside, and A. Kundaje. 2019. "Learning Important Features Through Propagating Activation Differences." arXiv. https://arxiv.org/abs/1704.02685.

Shrikumar, A., K. Tian, Ž. Avsec, et al. 2020. "Technical Note on Transcription Factor Motif Discovery from Importance Scores (TF-MoDISco) version 0.5.6.5." arXiv. https://arxiv.org/abs/1811.00416.

Spitz, F., and E. E. Furlong. 2012. "Transcription Factors: From Enhancer Binding to Developmental Control." *Nature Reviews Genetics* 13, no. 9: 613–626. https://doi.org/10.1038/nrg3207.

Subramanian, A., P. Tamayo, V. K. Mootha, et al. 2005. "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles." *Proceedings of the National Academy of Sciences of the United States of America* 102, no. 43: 15545–15550. https://doi.org/10.1073/pnas.0506580102.

Sundararajan, M., A. Taly, and Q. Yan. 2017. "Axiomatic Attribution for Deep Networks." arXiv. https://arxiv.org/abs/1703.01365.

Thurman, R. E., E. Rynes, R. Humbert, et al. 2012. "The Accessible Chromatin Landscape of the Human Genome." *Nature* 489, no. 7414: 75–82. https://doi.org/10.1038/nature11232.

Tripathi, P., S. Kurtulus, S. Wojciechowski, et al. 2010. "STAT5 Is Critical to Maintain Effector CD8+ T Cell Responses." *Journal of Immunology* 185, no. 4: 2116–2124. https://doi.org/10.4049/jimmunol.1000842.

Vaswani, A., N. Shazeer, N. Parmar, et al. 2023. "Attention Is All You Need." arXiv. https://arxiv.org/abs/1706.03762.

Verhaak, R. G., K. A. Hoadley, E. Purdom, et al. 2010. "Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1." *Cancer Cell* 17, no. 1: 98–110. https://doi.org/10.1016/j.ccr.2009.12.020.

Villar, D., C. Berthelot, S. Aldridge, et al. 2015. "Enhancer Evolution Across 20 Mammalian Species." *Cell* 160, no. 3: 554–566. https://doi.org/10.1016/j.cell.2015.01.006.

Waddington, C. H. 1957. *The Strategy of the Genes; a Discussion of Some Aspects of Theoretical Biology*. London: Routledge.

Wilczynska, K. M., S. K. Singh, B. Adams, et al. 2009. "Nuclear Factor I Isoforms Regulate Gene Expression During the Differentiation of Human Neural Progenitors to Astrocytes." *Stem Cells* 27, no. 5: 1173–1181. https://doi.org/10.1002/stem.35.

Wolf, F. A., P. Angerer, and F. J. Theis. 2018. "SCANPY: large-scale single-cell gene expression data analysis." *Genome Biology* 19, no. 1: 15. https://doi.org/10.1186/s13059-017-1382-0.

Yang, J. H., and A. S. Hansen. 2024. "Enhancer Selectivity in Space and Time: From Enhancer-Promoter Interactions to Promoter Activation." *Nature Reviews Molecular Cell Biology* 25, no. 7: 574–591. https://doi.org/10.1038/s41580-024-00710-6.

Yasuda, T., H. Sanjo, G. Pagès, et al. 2008. "Erk Kinases Link Pre-B Cell Receptor Signaling to Transcriptional Events Required for Early B Cell Expansion." *Immunity* 28, no. 4: 499–508. https://doi.org/10.1016/j.immuni.2008.02.015.

Yuan, H., and D. R. Kelley. 2022. "scBasset: Sequence-Based Modeling of Single-Cell ATAC-Seq Using Convolutional Neural Networks." *Nature Methods* 19, no. 9: 1088–1096. https://doi.org/10.1038/s41592-022-01562-8.

Zabidi, M. A., C. D. Arnold, K. Schernhuber, et al. 2015. "Enhancer-Core-Promoter Specificity Separates Developmental and Housekeeping Gene Regulation." *Nature* 518, no. 7540: 556–559. https://doi.org/10.1038/nature13994.

Zhang, L., J. Zhang, and Q. Nie. 2022. "DIRECT-NET: An Efficient Method to Discover Cis-Regulatory Elements and Construct Regulatory Networks From Single-Cell Multiomics Data." *Science Advances* 8, no. 22: eabl7393. https://doi.org/10.1126/sciadv.abl7393.

Zhang, P., Y. Sun, and L. Ma. 2015. "ZEB1: At the Crossroads of Epithelial-Mesenchymal Transition, Metastasis and Therapy Resistance." *Cell Cycle* 14, no. 4: 481–487. https://doi.org/10.1080/15384101.2015.1006048.

Zhang, Z., C. Yang, and X. Zhang. 2022. "scDART: Integrating Unmatched scRNA-Seq and scATAC-Seq Data and Learning Cross-Modality Relationship Simultaneously." *Genome Biology* 23, no. 1: 139. https://doi.org/10.1186/s13059-022-02706-x.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.