Refining sequence-to-expression modelling with chromatin accessibility

Orsolya Lapohos[1,2], Gregory J. Fonseca[1,2*], and Amin Emad[1,2,3,4,5*]

[1]Department of Quantitative Life Sciences, McGill University, Montreal, QC, Canada
[2]Meakins-Christie Laboratories, Research Institute of the McGill University Health Centre, Montreal, QC, Canada
[3]Department of Electrical and Computer Engineering, McGill University, Montreal, QC, Canada
[4]Mila, Quebec AI Institute, Montreal, QC, Canada
[5]The Rosalind and Morris Goodman Cancer Institute, Montreal, QC, Canada
*Corresponding authors: amin.emad@mcgill.ca, gregory.fonseca@mcgill.ca

Abstract

Sequence-to-expression models have gained popularity in the past decade, enabling prediction of gene expression from genomic sequence alone. However, these models typically do not take into account chromatin accessibility, a major factor limiting gene regulation. We hypothesized that supplying accessibility as an input feature would allow a sequence-to-expression model to focus on important open regions of the genome. Using single-nucleus multiome RNA- and ATAC-sequencing data, we found that the predictive performance of such an augmented model was significantly greater than that of sequence-only or accessibility-only models with similar architectures. Specifically, its ability to predict the expression of highly variable genes and gene expression in other cell types improved, which we attribute to a reduction in bias originating from lowly variable genes. Moreover, post-hoc analyses revealed that higher attribution scores in the input DNA sequences of the augmented model conformed to accessibility, whereas those in the sequence-only model were scattered. Additionally, we show that fine-tuning a pre-trained sequence-only model with both sequence and accessibility can boost performance even further. We also highlight the importance of sequencing depth in sequence-to-expression prediction. Altogether, the results of this study provide a compelling argument for the inclusion of chromatin accessibility in sequence-to-expression models, as this strategy can be implemented easily and may improve downstream applications.

1 Introduction

Metazoan gene expression is a complex process that is temporally and spatially orchestrated by a multitude of factors. At the lowest level, the blueprint for gene expression is ciphered into non-coding regulatory regions, defining the universe of DNA-binding proteins that can direct the regulation of each gene [1]. Although distal non-coding regions such as enhancers play important roles in controlling gene expression [2, 3], the promoter is arguably the master regulator. At the next level, the set of possible regulatory interactions is limited by context—chromatin accessibility and the set of available DNA-binding proteins both depend on cell type and environment [4]. Available regulatory proteins will compete and cooperate to bind accessible regions and ultimately activate or pause transcription initiation [5, 6]. Gene expression is further regulated during transcription and translation [1].

Despite the myriad of interactions regulating gene expression, sequence-to-expression models have been successful in predicting transcriptional activity from only promoter sequences [7, 8, 9, 10, 11, 12, 13]. The 1D convolutional neural network (CNN) is naturally suited to the task of learning local sequence patterns and is the most frequently used base architecture for sequence-to-expression modelling. Recently, attention mechanism was also leveraged to capture long-range interactions in Enformer, a state-of-the-art model that predicts several genome tracks from long input sequences [10]. However, existing models do not make use of rate-limiting gene-regulatory factors, such as chromatin accessibility, in order to predict gene expression. Assay for transposase-accessible chromatin sequencing (ATAC-seq) detects open chromatin [14] and this data is increasingly available at single-cell resolution in parallel with RNA-seq [15], providing both cell type-specific gene expression and regulatory context. Thus, we hypothesized that providing ATAC-seq as an additional input feature would focus a sequence-to-expression model on open regions of chromatin and thereby improve prediction of gene expression.

The contributions of this study are as follows. We used an architecture based on Xpresso [9] and included an additional input feature representing chromatin accessibility and trained it on 12 different cell types from three independent human multiome RNA- and ATAC-seq datasets from 10x Genomics. We showed that this augmented model predicted the expression of held-out

genes more accurately than models with similar architectures that only used sequence information or accessibility. This strategy also resulted in a greater improvement on the difficult task of highly variable gene expression prediction, while being able to generalize to other cell types. Additionally, models that used scrambled inputs performed significantly worse compared to our augmented model, revealing that all input data modalities contribute to the prediction performance. Importantly, we also showed that the magnitude of attribution scores obtained via the Shapley Additive Explanations DeepExplainer [16] in all DNA input channels increased in regions where the underlying chromatin was accessible, compared to the unfocused scores obtained for the sequence-only model. Additionally, we found that we could further boost performance by initializing weights of the augmented model using the weights of a sequence-only model and then fine-tuning it with both sequence and accessibility. Finally, we demonstrated a strong relationship between data quality and model performance across all models and datasets used in our study. The strategies tested and proposed here can be adapted to other sequence-to-expression models with ease, potentially benefiting various downstream tasks.

2 Methods

2.1 Promoter sequences

The genome sequence FASTA and comprehensive gene annotation GTF files for the GRCh38.p14 reference genome (release 44) were downloaded from GENCODE [17]. Then, the annotation file was filtered to obtain the coordinates of "Ensembl canonical" transcripts for protein-coding genes, which were used to extract the coding strand sequence 1kb upstream and 1kb downstream of the start of the first exon (transcription start site, or TSS).

2.2 Datasets

Human single-nucleus multiome ATAC and gene expression datasets for peripheral blood mononuclear cells (PBMC), brain, and jejunum, as well as a human PBMC single-cell gene expression dataset with cell surface marker antibody-derived tags (ADTs) were sourced from 10x Genomics (Data availability). Dataset details are listed in Supplementary Table S1. Preprocessing and annotation of multiome and single-cell datasets are described in Supplementary Notes 1 and 2, respectively. Four major cell types were used from each dataset, shown in Supplementary Figure S1A-C. Cell type marker genes and numbers are indicated in Supplementary Tables S2 and S3, respectively.

2.3 Representing chromatin accessibility and gene expression for sequence-to-expression modelling

To obtain pooled cell type-specific ATAC tracks for each 2kb input sequence around the TSS, raw ATAC-seq fragments mapped to the cell type of interest were aligned to input sequence coordinates. The pooled cell type-specific ATAC tracks were smoothed using the 1D Gaussian filter function from the SciPy package (with standard deviation 20) [18], and track values were then min-max normalized by cell type. For other analyses, a summary metric termed auATAC was also defined to represent the accessibility of a gene promoter as the area under its normalized 2kb ATAC track.

To obtain a measure of gene expression by cell type, the proportion of cells of the cell type of interest with a nonzero unique molecular identifier (UMI) count was calculated for each gene. This metric, termed "GEx" in this study, represents the probability that a gene is expressed while ensuring that the output values range between 0 and 1.

To analyze the similarity of gene expression or accessibility between cell types, the Pearson correlation coefficient between all GEx or auATAC values was calculated (Supplementary Figure S1D-F). Spearman correlation was used to determine concordance between GEx and auATAC (Supplementary Figure S1G-I).

In one of our analyses, we categorized genes based on gene expression and chromatin accessibility. The details of this method are described in Supplementary Note 3.

2.4 Model architecture

Figure 1 shows an overview of the model. One-hot-encoded DNA sequences with pooled and normalized ATAC-seq tracks serve as the 5 input channels to the neural network. The model outputs a single value that represents the expression of the input gene. The architecture of the neural network is based on Xpresso [9], but with a shorter input sequence (2kb instead of 10kb), an additional input channel, fixed depth and layer widths, and an additional dense layer (details in Supplementary Table S4 and Supplementary Figure S2). Our model was constructed with PyTorch v2.2 [19].

For ablation experiments, model architecture remained unchanged, other than the input channels, which could consist of the cell type-specific ATAC track (1 channel), the one-hot-encoded DNA sequence of the promoter (4 channels), or both (5 channels).

2.5 Model training

Each model was trained and tested using 5-fold cross-validation (CV) on the entire set of human autosomal protein-coding genes (19,095), where each test set consisted of unique genes. Training for each fold was repeated 5 times, each with different random seeds and

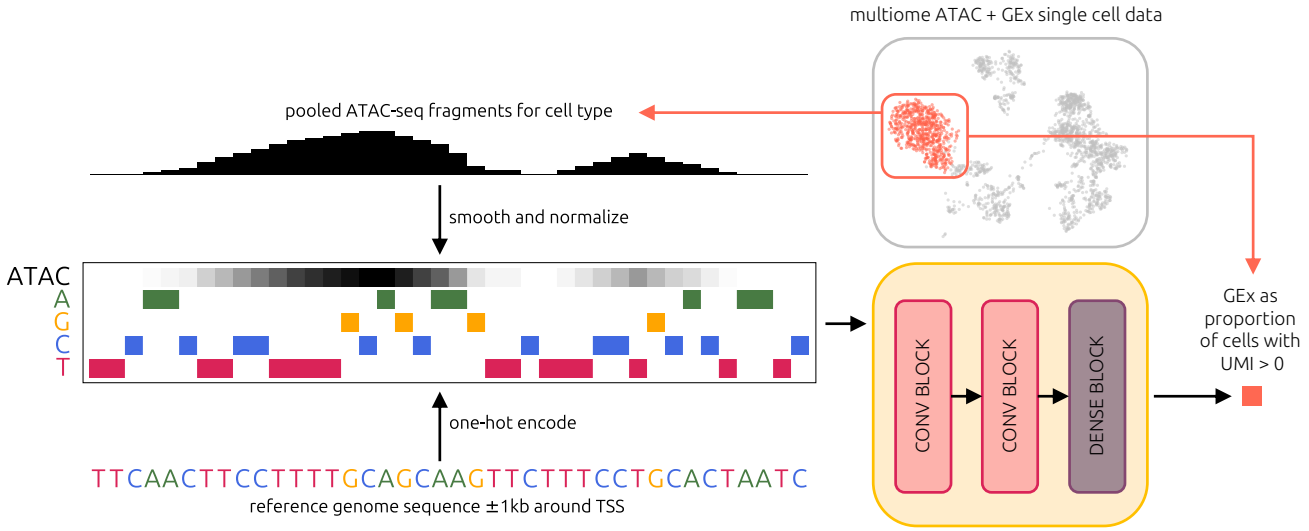

Figure 1: Overview of the ATAC-augmented sequence-to-expression model. Each input example consists of 5 vectors of length 2000, representing the promoter of a gene in a specific cell type. A pooled ATAC-seq track accompanies the one-hot encoded DNA nucleotides of the promoter sequence. For each input example, the model outputs a single value that represents the proportion of cells expressing the input gene.

inner validation sets. Gene order was preserved across all datasets for each seed in each CV fold. The model was trained with mean squared error (MSE) loss, defined as

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1}$$

where n is the number of genes, y_i is the ground truth GEx of gene i, and $\hat{y}_i$ is the predicted GEx for gene i. Training was guided by the Adam optimizer [20], with learning rate 5×10^{-5} and weight decay set to 10^{-3}. Batch size was set to 512, and the training procedure was run for 500 epochs. Validation loss was computed every 4 epochs, and the model with the lowest validation loss was saved.

2.6 Fine-tuning of pre-trained models

An additional model training strategy involved a two-step procedure. First, a sequence-to-expression model was trained on sequence only, as described in the previous subsection. Then, these optimized weights were used to initialize a new full model, with input ATAC channel weights initialized randomly. This full model was then fine-tuned by training on both DNA and ATAC features. The fine-tuning procedure was performed for a maximum of 200 epochs, with reduced weight decay (10^{-4}), a lower learning rate (10^{-5}), and using the same 5-fold CV and random seeds. The model with the lowest validation loss was saved.

2.7 Evaluation

For each experiment, performance was assessed across 5 folds. For each fold, the average prediction across 5 models trained with different random seeds was evaluated. Reported means and standard deviations were calculated across the 5 folds. Four different metrics were calculated: Pearson correlation coefficient, Spearman correlation coefficient, MSE, and the coefficient of determination (R^2).

2.8 Naïve prediction and evaluation

For each input gene promoter, the naïve predictor outputs its auATAC. Evaluation was performed using the test sets from the same 5-fold CV, calculating the Spearman correlation coefficient between the prediction (auATAC) and the ground truth GEx.

2.9 Post-hoc model explanations

We obtained attribution scores for each held-out test sample using the SHapley Additive exPlanations (SHAP) DeepExplainer [16]. As background for each sample, we supplied 100 shuffled versions that preserved its dinucleotide frequencies, using the dinucleotide shuffler from the DeepLIFT package [21]. For background samples used to explain the DNA+ATAC model, values in the ATAC input channel were randomly shuffled as well. To summarize positional attribution scores by input channel, the mean SHAP score for each sequence position in each channel was calculated across 5 random seeds, for each CV fold. Then, means from all held-out test genes of the 5-fold CV were concatenated. Finally, the mean SHAP score was calculated at each position across all held-out test genes, for each input channel.

2.10 Statistical testing

The sets of held-out test genes used for evaluation across all 5 folds were preserved across all models. Thus, we used the one-sided Wilcoxon signed-rank test (from SciPy [18]) to test for statistical significance, when comparing performance of different models.

3 Results

3.1 ATAC-seq refined sequence-to-expression prediction

In order to determine whether providing chromatin accessibility as an additional input feature could improve sequence-to-expression modelling, we augmented a CNN-based model architecture similar to Xpresso [9] by including an extra input channel that represents a pooled ATAC-seq track for a specific cell type, in addition to the one-hot encoding of the input sequence (Figure 1). This architecture could be ablated easily by modifying only the set of input channels. We trained such models on sequence (DNA-only), accessibility (ATAC-only), and both (DNA+ATAC), to predict the proportion of cells within a cell type that express a gene (GEx).

First, we focused on DNA-only as our baseline sequence-to-expression model. This model achieved an average Pearson correlation of 0.3658 across the four major cell types in the PBMC dataset (Table 1, DNA only). Spearman correlations were higher for all four cell types, with an average of 0.5319. Performance was better on the brain and jejunum datasets overall, with average Pearson correlations of 0.4966 and 0.4573, and Spearman correlations of 0.5970 and 0.5620, respectively (Supplementary Table S5). The same trend was observed with MSE and R^2 metrics.

As expected, ATAC alone was a better predictor of gene expression than sequence alone (Table 1, ATAC only). This was already apparent in a naïve predictor that predicts GEx based on its auATAC (Methods). The average Spearman correlation of the naïve predictions with ground truth GEx for PBMCs was 0.694 (Supplementary Figure S3). The ATAC-only model achieved an average Pearson correlation of 0.4930 and Spearman correlation of 0.7145 on the PBMC dataset, improving on the naïve predictor.

Combining DNA with ATAC significantly and consistently improved model performance across all cell types (Table 1, DNA+ATAC; Supplementary Table S5; $p < 0.05$ based on one-sided Wilcoxon signed-rank tests, comparing DNA+ATAC to ablated models). The average Pearson and Spearman correlations for PBMCs were 0.5339 and 0.7241, for the brain dataset, they were 0.6621 and 0.7579, and for the jejunum dataset 0.5522 and 0.6925.

Since ATAC alone is a relatively strong predictor of GEx, we performed scrambling experiments to determine the contribution of DNA sequence to the DNA+ATAC model output. We scrambled one or both input features in the training sets with respect to the target genes, trained new models, and evaluated their performance on the unscrambled test sets. We also re-evaluated predictions of models trained on unscrambled training sets using scrambled test sets as ground truth. In all scrambling experiments, performance was significantly worse than that of the unscrambled DNA+ATAC model (Figure 2; $p < 0.05$ based on one-sided Wilcoxon signed-rank tests).

When the DNA inputs in the training set were scrambled, the DNA+ATAC model achieved approximately the same performance on the unscrambled test set as the unscrambled ablated ATAC-only model (mean Pearson $r \approx 0.47$), indicating that unscrambled DNA is required to achieve the full performance of the DNA+ATAC model. Likewise, when the ATAC inputs in the training set were scrambled, similar performance was achieved on the unscrambled test set as the ablated DNA-only model (mean Pearson $r \approx 0.36$). We also experimented with scrambling both inputs for training, to determine whether combinatorial effects between sequence patterns and accessibility could be learned by the augmented model. When inputs were scrambled pairwise (i.e.: DNA and ATAC of the same gene promoter were kept together) the model performed poorly, with mean Pearson $r = 0.0492$. When inputs were scrambled separately (independently with respect to the target gene), performance was significantly worse ($p = 0.0339$, one-sided Wilcoxon signed-rank test), with mean Pearson $r = -0.0317$. This suggests that, despite training on scrambled data, some predictive patterns can be learned when DNA and ATAC are complementary. Our findings were corroborated by a similar experiment where only the test set was scrambled (Figure 2). Altogether, results of the ablation and scrambling experiments showed that the input DNA sequences contributed meaningfully to the DNA+ATAC model output.

3.2 ATAC-seq feature improved cross-cell type and highly variable gene expression prediction

We evaluated generalizability by testing each trained cell type-restricted model on other cell types within the same dataset. In all three datasets, Pearson correlations on the cross-cell type prediction tasks were similar to those obtained when predicting on the same cell type, for both the ablated and augmented models (Figure 3A). These results indicate that the ATAC-seq input refined sequence-to-expression prediction from a generalization perspective as well. However, it should be noted that ground truth GEx values correlate highly between different cell types. The Pearson correlation coefficients between ground truth GEx of different cell types within the same dataset ranged from 0.699 to 0.981 (Supple-

Table 1: Results of ablation experiments with the PBMC dataset. Performance metrics on the held-out test sets are shown. MSE: mean squared error. All comparisons within cell type are significantly different by one-sided Wilcoxon signed-rank test ($p < 0.05$), except ATAC only versus DNA+ATAC Spearman r for CD4 T cell.

cell type	model	Pearson r	Spearman r	MSE	R^2
B cell	DNA only	0.3620 ± 0.0196	0.5344 ± 0.0063	0.0151 ± 0.0006	0.1162 ± 0.0094
	ATAC only	0.4741 ± 0.0160	0.7042 ± 0.0058	0.0138 ± 0.0007	0.1927 ± 0.0120
	DNA+ATAC	**0.5215 ± 0.0163**	**0.7200 ± 0.0059**	**0.0126 ± 0.0005**	**0.2649 ± 0.0154**
$CD14^+$ monocyte	DNA only	0.3839 ± 0.0133	0.5237 ± 0.0106	0.0247 ± 0.0018	0.1364 ± 0.0062
	ATAC only	0.5466 ± 0.0094	0.7338 ± 0.0079	0.0207 ± 0.0017	0.2770 ± 0.0105
	DNA+ATAC	**0.5801 ± 0.0096**	**0.7422 ± 0.0080**	**0.0193 ± 0.0015**	**0.3259 ± 0.0117**
CD4 T cell	DNA only	0.3536 ± 0.0170	0.5304 ± 0.0084	0.0137 ± 0.0007	0.1037 ± 0.0069
	ATAC only	0.4749 ± 0.0084	0.7127 ± 0.0079	0.0123 ± 0.0006	0.1958 ± 0.0027
	DNA+ATAC	**0.5116 ± 0.0090**	**0.7169 ± 0.0052**	**0.0114 ± 0.0005**	**0.2543 ± 0.0085**
CD8 T cell	DNA only	0.3635 ± 0.0191	0.5391 ± 0.0084	0.0160 ± 0.0010	0.1124 ± 0.0077
	ATAC only	0.4762 ± 0.0101	0.7074 ± 0.0064	0.0145 ± 0.0009	0.1931 ± 0.0043
	DNA+ATAC	**0.5222 ± 0.0119**	**0.7171 ± 0.0053**	**0.0132 ± 0.0007**	**0.2657 ± 0.0108**

Figure 2: Scrambling experiments with B cell models. Pearson correlations of predicted versus ground truth GEx across 5-fold CV for each indicated ablated or scrambled model. Pairwise scrambling means the promoter sequence and ATAC track were kept together. Separate scrambling means promoter sequences, and ATAC tracks were scrambled independently. All vertically adjacent pairs of boxplots, except for the bottom pair, are significantly different by one-sided Wilcoxon signed-rank test ($p < 0.05$).

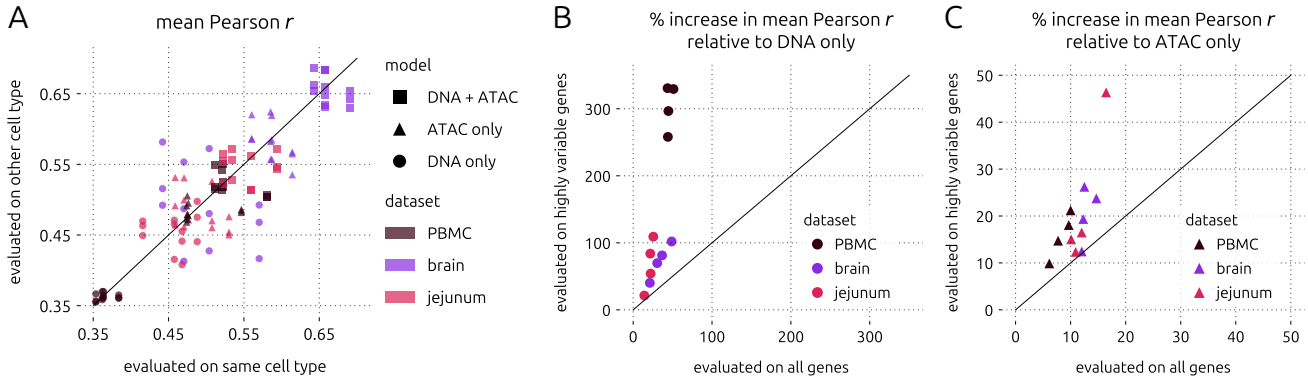

Figure 3: Cross-cell type and highly variable gene expression prediction results. A. Mean Pearson r of models evaluated on held-out test sequences using another cell type from the same dataset (tissue) versus on the cell type used for training. B. Percent increase in mean Pearson correlation from DNA only to DNA+ATAC for all major cell types in all three multiome datasets, evaluating on highly variable genes versus all genes. D. Same as C, but showing percent increase in mean Pearson correlation from ATAC only to DNA+ATAC. Solid diagonal lines indicate $y = x$.

mentary Figure S1D-F, top). Cross cell-type Pearson correlation of auATAC was higher, with values between 0.817 and 0.992 (Supplementary Figure S1D-F, bottom). With such similarities, it is not surprising that sequence-to-expression models can generalize between cell types.

In order to assess whether this observed generalizability is due to lowly variable (likely housekeeping) genes, we also evaluated the same models using subsets of the held-out test genes that were highly variable (Supplementary Note 1). As expected, all models had difficulty predicting the expression of highly variable genes, but models using ATAC were more resilient. When considering models trained and tested on the same cell type, the relative increase in mean Pearson correlation from DNA-only to DNA+ATAC was greater when evaluating on highly variable genes compared to all test genes (Figure 3B). We observed the same trend in the relative performance improvement from ATAC-only to DNA+ATAC (Figure 3C). Interestingly, we also observed these patterns on the cross-cell type prediction tasks (Supplementary Figures S4-7, Supplementary Data), although GEx of some cell types seemed to be inherently easier to predict. Taken together, these results suggest that the ATAC-augmented model is less biased by lowly variable genes while being capable of predicting on other cell types.

3.3 Positional attribution scores of input channels aligned with chromatin accessibility

To better understand our ATAC-augmented sequence-to-expression model, we examined attribution scores corresponding to each input as determined by the SHAP DeepExplainer [16], starting with the ATAC channel. Given that ATAC alone is a strong predictor of GEx—considering auATAC as a naïve predictor of GEx, as well as the performance of the ATAC-only model—we wanted to determine its contribution to model outputs for different categories of genes. We sorted genes into low, intermediate, and high bins, separately by GEx and auATAC (Figure 4A, Methods). Then, we intersected these bins to obtain 9 categories. For 97.9% of genes, GEx tended to increase with auATAC in CD4 T cells. A few genes (348) had high auATAC and low GEx, while a smaller number of genes had low auATAC and high GEx (60). We suspected that temporary silencing/pausing or long mRNA half-lives may explain these rare discrepancies. However, there was no concordance between an aggregated general mRNA half-life feature from a published meta-analysis [22] and the high GEx $\cap$ low auATAC category (Supplementary Figure S8A). Analyzing pausing indices calculated from a published PBMC CD4 T cell precision run-on sequencing (PRO-seq) dataset [23] (Supplementary Note 4) revealed that in the low GEx $\cap$ high auATAC category, only 25.2% of genes had a pausing index greater than 2 (Supplementary Figure S8B). Altogether, these observations suggest that neither mRNA half-life nor promoter-proximal pausing can explain all discrepancies between GEx and auATAC—these rare occurrences may be better explained by technical limitations.

Nevertheless, we computed positional attribution scores and averaged them for each gene category (Figure 4B). We observed that the ATAC input heavily influenced our model's output, and that the largest mean ATAC attribution scores aligned with the mean ATAC track peak, slightly upstream of the TSS. Interestingly, relatively large values in the ATAC input track ($\sim$0.25) did not contribute positively to the model's output in a small window (400-450bp) upstream of the TSS and in a wider window (100-400bp) downstream of the TSS. For genes with high auATAC, ATAC attribution scores were highly variable 500-1000bp downstream the TSS, indicating that accessibility in this section of the 5' untranslated region was deemed relevant to the model. Overall, these results show that our DNA+ATAC model learned TSS-relative position-specific patterns in chromatin accessibility.

Next, we compared positional attribution scores in all input channels between the DNA-only and the DNA+ATAC models (Figure 5A). The SHAP scores of the DNA-only model were more distributed across the positions of the input sequence, while the DNA channels of the DNA+ATAC model had more variable SHAP scores that mirrored variability in the ATAC channel. We also calculated Spearman correlation coefficients of mean SHAP scores with the ATAC input track values at each position, across all test genes, for each channel (Figure 5B). Correlation coefficients were generally close to zero in the DNA-only model, taking on values between -0.461 and 0.372. In the DNA+ATAC model, SHAP scores in the ATAC channel correlated highly with the input ATAC track, with a minimum of 0.224 and maximum of 0.960, and coefficients in the DNA channels followed a similar pattern to those in the ATAC channel, with values between -0.303 and 0.830. Hence, we surmise that the ATAC input feature guided the training of model weights in the DNA channels.

3.4 Replacing shallow single-nucleus GEx with deeper single-cell GEx improved performance

As the gene expression data obtained from multiome datasets is derived from single nuclei, we questioned whether model performance could be improved by replacing gene expression values with those from a matched more deeply sequenced single-cell dataset. When comparing the gene expression values between the multiome and single-cell RNA-seq PBMC datasets, we observed that the single-cell dataset was less sparse (Supplementary Figure S9A,B). This difference could be explained by the isolation of nuclei, but also by the median UMI counts per cell being approximately dou-

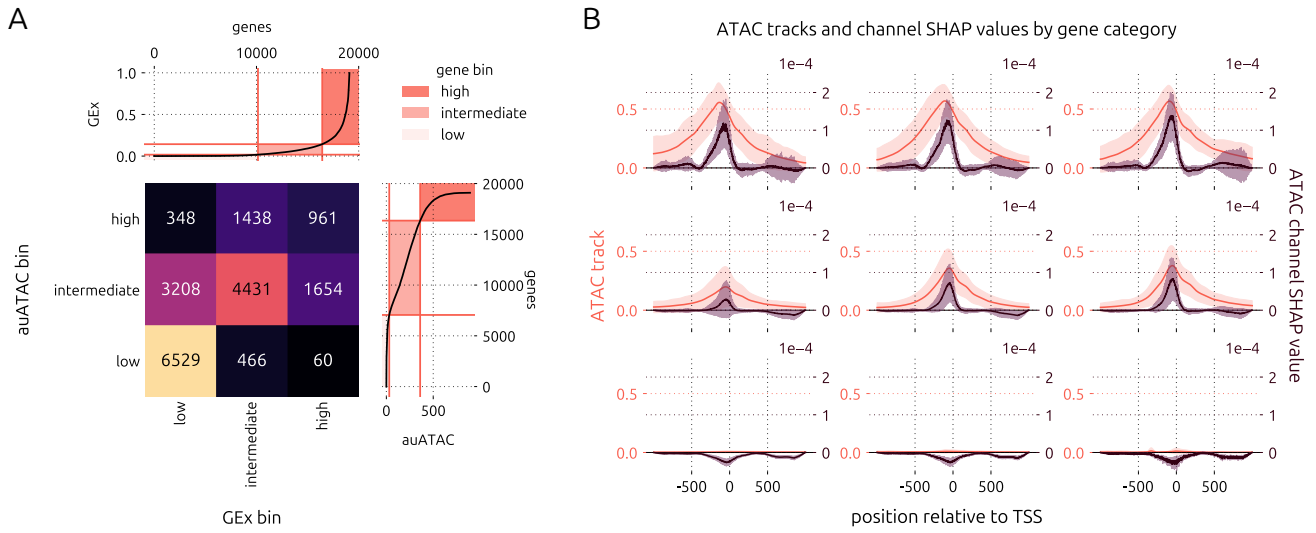

Figure 4: Chromatin accessibility and corresponding attribution scores in CD4 T cells, broken down by gene category. A. Gene contingency table generated by double knee point thresholding on GEx and area under ATAC (auATAC) for CD4 T cells. B. ATAC SHAP scores and ATAC input tracks for each category of the gene contingency table shown in A (DNA+ATAC model). Solid lines show means and shaded regions show standard deviations across all genes (after taking mean positional SHAP values across all random seeds).

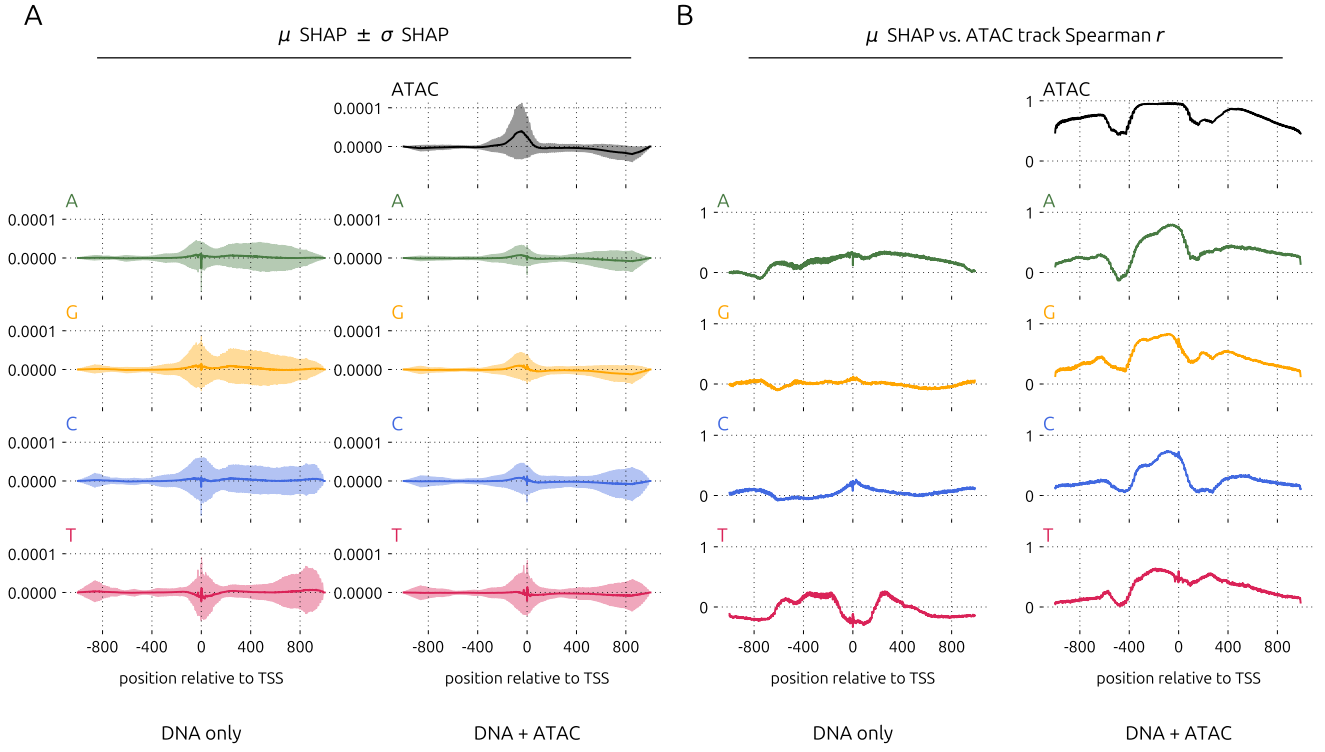

Figure 5: Positional attribution score distribution and correlation with accessibility across all input channels of the DNA-only and DNA+ATAC CD4 T cell models. A. μ and σ SHAP represent the mean and standard deviation of the positional SHAP scores in the indicated channel across 5 random seeds for all test genes (concatenated from all CV folds). Solid lines show μ and shaded regions indicate σ. B. μ SHAP vs. ATAC track Spearman r represents the positional Spearman correlation coefficient between μ SHAP and the input ATAC track across all test genes (concatenated from all CV folds).

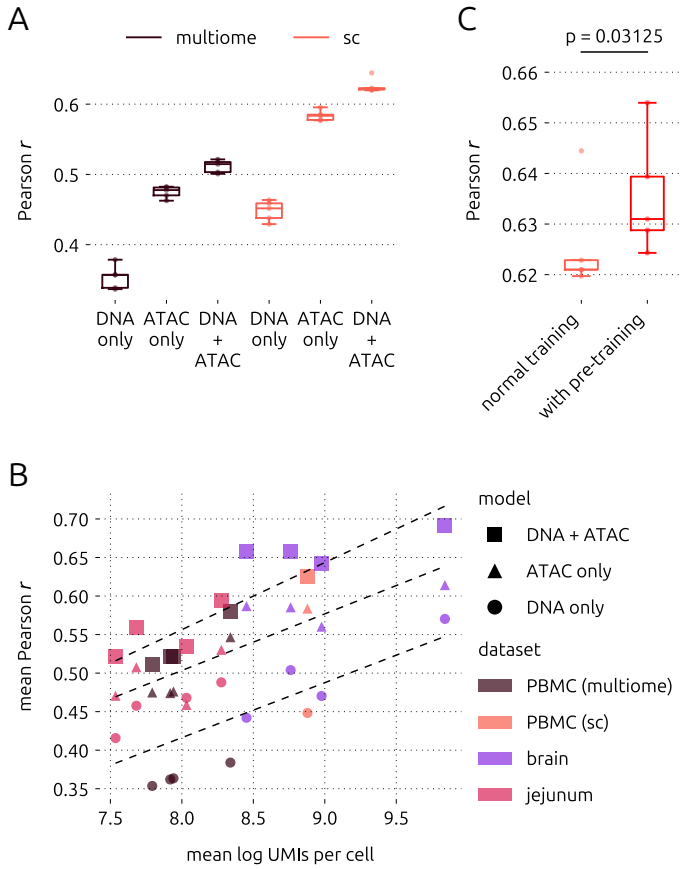

Figure 6: Performance improvement with single-cell GEx, sequencing depth, and pre-training. A. Performance improvement with single-cell versus multiome CD4 T cell data. Each comparison is significantly different ($p < 0.05$) by one-sided Wilcoxon signed-rank test. B. Relationship between model performance and mean log UMI counts. Each point represents the mean Pearson correlation achieved by a model trained and tested on one cell type, colour coded according to its dataset of origin. Square, triangle, and circle markers represent DNA+ATAC, ATAC only, and DNA only models, respectively. Dashed lines of best fit are shown for each model (top: DNA+ATAC, bottom: DNA only). C. Performance improvement with pre-training, using CD4 T cells (single-cell). "Normal training" refers to the DNA+ATAC model being trained starting from randomly initialized weights. "With pretraining" refers to the DNA+ATAC model being trained starting from weights copied from the DNA-only model. Boxplots show test results from 5-fold CV. Significance was determined by one-sided Wilcoxon signed-rank test.

bled in the single-cell dataset, despite both datasets having ∼70,000 mean raw read pairs per cell prior to preprocessing (Supplementary Table S1). Notably, GEx values derived from the single-cell dataset were more highly correlated with auATAC values (Supplementary Figure S9C), than those from the multiome dataset (Supplementary Figure S1G). When we re-trained CD4 T cell models on gene expression data from single-cells, we observed a significant improvement in the performance of the DNA+ATAC model, as well as the ablated models (Figure 6A). Further, we compared the correlation between positional attribution scores and chromatin accessibility obtained from models trained on CD4 T cells from the multiome and single-cell PBMC datasets (Supplementary Figure S10). While these correlational patterns were similar between the two datasets, the single-cell versions had more variability, which may be indicative of more nuanced attribution.

3.5 UMI count affected model performance

After having trained sequence-to-expression models on several datasets, we suspected that differences in underlying statistics may have affected model performance. We investigated whether the number of reads per cell (Supplementary Figure S11), the number of pooled cells (Supplementary Table S3), or the variation in GEx values impacted model training. Plotting the mean Pearson correlation of each cell type-restricted model against the corresponding mean log UMIs per pooled cell, we observed that performance increased linearly with read depth (Figure 6B). This trend occurred in the DNA-only, ATAC-only, and DNA+ATAC models, with R^2 values of 0.52, 0.75, and 0.79, respectively (Supplementary Figure S12, top). The number of pooled cells (Supplementary Figure S12, bottom left) did not seem to have an effect on performance, implying that the minimum number of pooled cells (185) was sufficient to obtain accurate gene expression probabilities. Finally, we found a trend between mean Pearson correlation and the variance-to-mean ratio (VMR) of GEx values for the ATAC-only and DNA+ATAC models ($R^2 = 0.57$ and 0.46, respectively), but not the DNA-only model (Supplementary Figure S12, bottom right). However, it should be noted that GEx VMR correlated highly with mean log UMIs per cell (Spearman $r = 0.819$). Altogether, these results demonstrate the importance of sequencing depth in sequence-to-expression modelling.

3.6 Pre-training strategy improved model performance

To explore additional strategies that could improve the learning of sequence patterns predictive of gene ex-

pression, we tested whether fine-tuning the pre-trained single-cell CD4 T cell DNA-only model with DNA and ATAC would help. We hypothesized that letting model weights reach an optimum based only on sequence prior to introducing the ATAC feature would improve performance. To test this hypothesis, we trained a new DNA+ATAC model, initializing it with weights copied from the corresponding DNA-only model (details in Methods). We found that there was indeed a statistically significant improvement in performance with the pre-training strategy, compared to the DNA+ATAC model trained from randomly intialized weights (Figure 6C; $p = 0.0313$, one-sided Wilcoxon signed-rank test). Additionally, we observed a slight change in the positional correlation of channel attribution scores with the input ATAC tracks (Supplementary Figure S10). With the pre-training strategy, Spearman correlation coefficients were generally similar to or lower than those obtained with the DNA+ATAC model. With the improvement in performance, these observations suggest that the pre-training strategy helps to balance information learned from sequence and accessibility. Thus, fine-tuning a pre-trained model with both sequence and accessibility is a viable strategy for boosting performance and it may allow the learning of more nuanced sequence patterns as determined by post-hoc explainers.

4 Discussion

Sequence-to-expression models are not expected to capture all variation in gene expression on their own. Many layers of regulation restrict the realm of possibilities defined by a DNA sequence, and these layers cannot be known without measuring the cellular context to a sufficient degree. For instance, when Xpresso [9] and scXpresso [13] incorporated mRNA half-life features, gene expression prediction improved. Although general half-life features are not context-specific, they contribute an additional set of "rules" to define transcriptional dynamics. As this field moves towards building more complex models of gene expression, more variation will be explained by accounting for every possible regulatory layer, both sequence-inherent, and context-specific. Here, we demonstrated that augmenting input DNA sequences with chromatin accessibility captured in multiome single-cell datasets refined sequence-to-expression modelling. In all datasets we used, performance consistently and significantly improved not only with respect to the baseline DNA-only model, but also with respect to the ATAC-only model.

A major concern we had with the concept of feeding accessibility as an additional input rather than output to the model, was that accessibility alone may be too strong of a predictor of gene expression. As a result, it could be too easy for model weights to converge at a local minimum that lacks sequence-related knowledge during training. However, our concerns were alleviated after confirming, via ablation and scrambling experiments, that the input DNA sequences had contributed a significant amount of information. Moreover, the finding that the relative improvement from both DNA-only and ATAC-only to DNA+ATAC increased for highly variable genes (Figure 3C) implies that our strategy lessened biases by lowly variable (likely housekeeping) genes.

To the best of our knowledge, the generalizability of sequence-to-expression models trained on one cell type had never been evaluated on other cell types in the same tissue. Here, we found that the performance of each model type on other cell types within the same dataset was similar to its performance on the same cell type. Further, these trends were not affected when evaluating on only highly variable genes. Thus, we suspect that either (1) the learned sequence patterns within accessible promoters were general, and/or (2) variability within tissue is too low to capture cell-type specific patterns. Follow-up studies are needed to further investigate these hypotheses. Nevertheless, improved modelling of highly variable genes was a positive outcome of our approach.

Taking our analysis further, attribution scores in the DNA input channels of the augmented sequence-to-expression model increased at accessible sequence positions. This result contrasted with the attribution scores for the DNA-only model which were more widely distributed across input sequence positions. This "focusing" of the sequence attribution scores is a promising result of our strategy, as it may result in improved inference of motif grammar and prediction of single nucleotide variant effects in future work.

Lastly, we found two extra steps that improved model performance. The first of these steps involved the replacement of GEx values from relatively shallow single-nucleus multiome data with those from more deeply sequenced single-cell data. We also showed that model performance improves linearly with mean log UMIs per pooled cell, but it is worth mentioning that differences between single-nucleus and single-cell RNA-seq (see [24, 25]) may have considerable effects on ground truth and therefore model accuracy. The second step was the fine-tuning of a pre-trained DNA-only model using both DNA and ATAC inputs. We speculate that the reason for this improvement is that delaying the use of the strongly predictive accessibility feature might prevent distraction from learning subtle sequence patterns. However, we did not observe major changes in the positional correlation of channel attribution scores with the ATAC input track using the pre-training strategy, compared to normal training. Small decreases in correlation were observed in short segments of the promoter sequences, taking on a seemingly more balanced profile in between the DNA-only model and the DNA+ATAC model. This result implies that applying the fine-tuning approach to existing sequence-to-

expression models could yield refined results with minimal effort.

Our study has the following limitations. Firstly, techniques that measure chromatin accessibility each have biases, so using alternative (or combined) accessibility data measured by DNase-seq may lead to different learned patterns [26]. Another limitation of our input data is that sequencing depth has a major impact on performance, and obtaining deeply sequenced single-cell data is costly. However, the concepts demonstrated in this study can be applied to bulk data as well, provided that both RNA-seq and ATAC-seq are available for a homogeneous biological sample. We also note that our input sequences are relatively short, and most sequence-to-expression models use much longer input sequences. On the other hand, models using long input sequences have been shown to ignore effects of enhancers [27]. Since enhancers affect highly variable genes in particular [28], exploring new architectures and strategies capable of learning long-range effects is still important. Finally, although we used chromatin accessibility as input to the first layer of the CNN, alternative approaches incorporating similar data may lead to further advances.

In conclusion, we added context to a plain sequence-to-expression model by including chromatin accessibility as an input feature. By modelling gene expression as a function of both sequence and accessibility, learning sequence patterns predictive of highly variable gene expression became easier and attribution scores of DNA input channels became aligned with accessibility. We are convinced that this sequence-to-expression augmentation strategy would benefit larger models and result in the inference of more nuanced cell type-specific regulatory grammar and improved single nucleotide variant effect prediction.

5 Data availability

Multiome human PBMC data were obtained from https://www.10xgenomics.com/resources/datasets/pbmc-from-a-healthy-donor-no-cell-sorting-10-k-1-standard-2-0-0.
Multiome human brain data were obtained from https://www.10xgenomics.com/resources/datasets/frozen-human-healthy-brain-tissue-3-k-1-standard-2-0-0.
Multiome human jejunum data were obtained from https://www.10xgenomics.com/resources/datasets/human-jejunum-nuclei-isolated-with-chromium-nuclei-isolation-kit-saltyez-protocol-and-10x-complex-tissue-dp-ct-sorted-and-ct-unsorted-1-standard. The single-cell PBMC dataset with ADTs was obtained from https://www.10xgenomics.com/datasets/5k-human-pbmcs-stained-with-totalseq-B-human-TBNK-cocktail-NextGEM.

The aggregated mRNA half-life feature for human genes was obtained from a previously published meta-analysis [22] (file: 13059_2022_2811_MOESM3_ESM.xlsx).

The human PBMC CD4 T cell PRO-seq data were obtained from a previous publication by Danko et al. [23], downloaded from the Gene Expression Omnibus with accession number GSE85337.

6 Code availability

Implemented code is available on GitHub, at https://github.com/lapohosorsolya/accessible_seq2exp.

7 Competing interests

No competing interest is declared.

8 Supplementary material

Supplementary Notes, Tables and Figures in PDF format and Supplementary Data in Excel sheets are available online.

9 Author contributions statement

All authors contributed to the design of the study and the writing of the manuscript. A.E. and G.J.F. supervised the study. O.L. performed the experiments and analyzed the results. O.L., G.J.F., and A.E. read and approved the final manuscript.

10 Funding

This work was supported by grants from Natural Sciences and Engineering Research Council of Canada (NSERC) [RGPIN-2019-04460] (AE) and Canada Foundation for Innovation (CFI) JELF [project 40781] (AE). This research was enabled in part by support provided by Calcul Québec (www.calculquebec.ca) and the Digital Research Alliance of Canada (alliancecan.ca).

References

[1] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell, 4th Edition*. Garland Science, New York, 2002.

[2] Peter C. FitsGerald, Andrey Shlyakhtenko, Alain A. Mir, and Charles Vinson. Clustering of dna sequences in human promoters. *Genome Research*, 14:1562–1574, 8 2004.

[3] Gabrielle D. Smith, Wan Hern Ching, Paola Cornejo-Páramo, and Emily S. Wong. Decoding enhancer complexity with machine learning and high-throughput discovery. *Genome Biology*, 24:1–25, 5 2023.

[4] Sven Heinz, Casey E. Romanoski, Christopher Benner, and Christopher K. Glass. The selection and function of cell type-specific enhancers. *Nature Reviews Molecular Cell Biology*, 16:144–154, 2 2015.

[5] Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, and Christopher K. Glass. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Molecular Cell*, 38:576–589, 5 2010.

[6] Gregory J. Fonseca, Jenhan Tao, Emma M. Westin, Sascha H. Duttke, Nathanael J. Spann, Tobias Strid, Zeyang Shen, Joshua D. Stender, Mashito Sakai, Verena M. Link, Christopher Benner, and Christopher K. Glass. Diverse motif ensembles specify non-redundant dna binding activities of ap-1 family members in macrophages. *Nature Communications*, 10, 12 2019.

[7] Jian Zhou, Chandra L. Theesfeld, Kevin Yao, Kathleen M. Chen, Aaron K. Wong, and Olga G. Troyanskaya. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature Genetics*, 50:1171, 8 2018.

[8] David R. Kelley. Cross-species regulatory sequence activity prediction. *PLOS Computational Biology*, 16:e1008050, 7 2020.

[9] Vikram Agarwal and Jay Shendure. Predicting mrna abundance directly from genomic sequence using deep convolutional neural networks. *Cell Reports*, 31:107663, 5 2020.

[10] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, Agnieszka Grabska-Barwinska, Kyle R. Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R. Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18:1196–1203, 10 2021.

[11] Jiaqi Li, Jingjing Wang, Peijing Zhang, Renying Wang, Yuqing Mei, Zhongyi Sun, Lijiang Fei, Mengmeng Jiang, Lifeng Ma, Weigao E, Haide Chen, Xinru Wang, Yuting Fu, Hanyu Wu, Daiyuan Liu, Xueyi Wang, Jingyu Li, Qile Guo, Yuan Liao, Chengxuan Yu, Danmei Jia, Jian Wu, Shibo He, Huanju Liu, Jun Ma, Kai Lei, Jiming Chen, Xiaoping Han, and Guoji Guo. Deep learning of cross-species single-cell landscapes identifies conserved regulatory programs underlying cell types. *Nature Genetics*, 54:1711–1720, 10 2022.

[12] Jacob Hepkema, Nicholas Keone Lee, Benjamin J. Stewart, Siwat Ruangroengkulrith, Varodom Charoensawan, Menna R. Clatworthy, and Martin Hemberg. Predicting the impact of sequence motifs on gene regulation using single-cell data. *Genome Biology*, 24:1–22, 12 2023.

[13] Lieke Michielsen, Marcel J.T. Reinders, and Ahmed Mahfouz. Predicting cell population-specific gene expression from genomic sequence. *Frontiers in Bioinformatics*, 4:1347276, 3 2024.

[14] Jason D. Buenrostro, Beijing Wu, Howard Y. Chang, and William J. Greenleaf. Atac-seq: A method for assaying chromatin accessibility genome-wide. *Current Protocols in Molecular Biology*, 109:21.29.1, 2015.

[15] Junyue Cao, Darren A. Cusanovich, Vijay Ramani, Delasa Aghamirzaie, Hannah A. Pliner, Andrew J. Hill, Riza M. Daza, Jose L. McFaline-Figueroa, Jonathan S. Packer, Lena Christiansen, Frank J. Steemers, Andrew C. Adey, Cole Trapnell, and Jay Shendure. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, 361:1380, 9 2018.

[16] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[17] Adam Frankish, Sílvia Carbonell-Sala, Mark Diekhans, Irwin Jungreis, Jane E Loveland, Jonathan M Mudge, Cristina Sisu, James C Wright, Carme Arnan, If Barnes, Abhimanyu Banerjee, Ruth Bennett, Andrew Berry, Alexandra Bignell, Carles Boix, Ferriol Calvet, Daniel Cerdán-Vélez, Fiona Cunningham, Claire Davidson, Sarah Donaldson, Cagatay Dursun, Reham Fatima, Stefano Giorgetti, Carlos Garcıa Giron, Jose Manuel Gonzalez, Matthew Hardy, Peter W Harrison, Thibaut Hourlier, Zoe Hollis, Toby Hunt, Benjamin James, Yunzhe Jiang, Rory Johnson, Mike Kay, Julien Lagarde, Fergal J Martin, Laura Martínez Gómez, Surag Nair, Pengyu Ni, Fernando Pozo, Vivek Ramalingam, Magali Ruffier, Bianca M Schmitt, Jacob M Schreiber, Emily Steed, Marie-Marthe Suner, Dulika Sumathipala, Irina Sycheva, Barbara Uszczynska-Ratajczak, Elizabeth Wass, Yucheng T Yang, Andrew Yates, Zahoor Zafrulla, Jyoti S Choudhary, Mark Gerstein, Roderic Guigo, Tim

J P Hubbard, Manolis Kellis, Anshul Kundaje, Benedict Paten, Michael L Tress, and Paul Flicek. Gencode: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Research*, 51(D1):D942–D949, 11 2022.

[18] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019.

[20] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[21] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences, 2019.

[22] Vikram Agarwal and David R. Kelley. The genetic and biochemical determinants of mrna degradation rates in mammals. *Genome Biology*, 23, 12 2022.

[23] Charles G. Danko, Lauren A. Choate, Brooke A. Marks, Edward J. Rice, Zhong Wang, Tinyi Chu, Andre L. Martins, Noah Dukler, Scott A. Coonrod, Elia D. Tait Wojno, John T. Lis, W. Lee Kraus, and Adam Siepel. Dynamic evolution of regulatory element ensembles in primate cd4+ t cells. *Nature Ecology & Evolution*, 2:537–548, 1 2018. CD4 T cell GRO-seq from human PBMCs: resource I used to check paused transcription.

[24] Nayoung Kim, Huiram Kang, Areum Jo, Seung Ah Yoo, and Hae Ock Lee. Perspectives on single-nucleus rna sequencing in different cell types and tissues. *Journal of Pathology and Translational Medicine*, 57:52–59, 1 2023.

[25] Beata W. Solnestam, Henrik Stranneheim, Jimmie Hällman, Max Käller, Emma Lundberg, Joakim Lundeberg, and Pelin Akan. Comparison of total and cytoplasmic mrna reveals global regulation by nuclear retention and mirnas. *BMC Genomics*, 13, 10 2012.

[26] Asllhan Karabacak Calviello, Antje Hirsekorn, Ricardo Wurmus, Dilmurat Yusuf, and Uwe Ohler. Reproducible inference of transcription factor footprints in atac-seq and dnase-seq datasets using protocol-specific bias modeling. *Genome Biology*, 20:1–13, 2 2019.

[27] Alexander Karollus, Thomas Mauermeier, and Julien Gagneur. Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome Biology*, 24:1–29, 12 2023.

[28] Drew T. Bergman, Thouis R. Jones, Vincent Liu, Judhajeet Ray, Evelyn Jagoda, Layla Siraj, Helen Y. Kang, Joseph Nasser, Michael Kane, Antonio Rios, Tung H. Nguyen, Sharon R. Grossman, Charles P. Fulco, Eric S. Lander, and Jesse M. Engreitz. Compatibility rules of human enhancer and promoter sequences. *Nature*, 607:176–184, 5 2022.