seq2VPlot: An Interpretable Deep Learning Model for Two-Dimensional ATAC-seq Prediction

Justin Currie, Sneha Mitra, Ritambhara Singh, Christina Leslie

Computer Science ScB. Honors Thesis, Spring 2025

Abstract

Chromatin accessibility profiles at *cis*-regulatory elements (CREs) are shaped by DNA-binding events of various scale, including individual Transcription Factor (TF) binding, TF complex formation, and nucleosome positioning. DNA binding is largely sequence driven, as TFs preferentially bind to recurring sequence motifs. This relationship suggests a learnable mapping between DNA sequence and chromatin accessibility. Here, we propose seq2VPlot, an interpretable deep learning model that accurately predicts two-dimensional ATAC-seq profiles from DNA sequences at base-pair resolution. By extracting sequence attribution scores from trained seq2VPlot models, we identify motif patterns that govern chromatin accessibility across cell-types. Two-dimensional ATAC-seq profiles offer enhanced interpretability relative to one-dimensional profiles, allowing the impact of TF binding motifs and disease-associated regulatory variants to be explored at finer resolution. seq2VPlot is part of a growing class of sequence-to-function models that may guide the design of wet-lab experiments exploring how DNA sequence and TF binding shape cell-state specific chromatin accessibility landscapes and gene regulatory programs.

Introduction

The binding of Transcription Factors (TFs) to *cis*-regulatory elements (CREs) plays a critical role in the regulation of gene expression. A gene can be regulated by several CREs, and different TF-binding combinations at each CRE precisely tunes gene expression levels. TFs preferentially bind to short sequence segments called TF-binding motifs (TFBMs), and TF-binding frequently occurs in a cooperative manner. Pioneer TFs can bind to condensed chromatin, increasing chromatin accessibility at CREs and recruiting other TFs to bind at nearby sites. The fact that TF-binding is both sequence specific and cooperative suggests the existence of a motif grammar, or a set of rules that determines which TFBM combinations will lead to certain TF binding patterns at a given CRE. TF binding also varies with cell-state. During differentiation, lineage-specific TFs bind to CREs and recruit co-factors, driving the expression of cell-type specific genes. Furthermore, TFs that bind ubiquitously across the genome may associate with other factors in a cell-state specifc manner. These conditions suggest that motif grammars are specific to certain cellular states.

DNA-binding events can impact local chromatin accessibility patterns through several mechanisms. At a basic level, the formation of a TF complex or the binding of a nucleosome occludes the DNA at the binding site itself. This alters the signal from Chromatin accessibility assays such as Assay for Transposase Accessible Chromatin with sequencing (ATAC-seq). In an ATAC-seq protocol, chromatin is exposed to a Tn5 transposase enzyme. Tn5 cuts DNA at accessible sites, creating fragments that can be sequenced and mapped back to a reference genome. Genomic loci with a high density of reads mapping to them are likely to be CREs, and many ATAC-seq analyses focus on investigating the differential accessibility of CREs across cell types. However, analyzing ATAC-seq data at base-pair resolution reveals information regarding local TF-binding patterns. Using bulk or pseudo-bulk ATAC-seq data, the frequency of Tn5 insertions along a given region can be plotted at base pair resolution, creating a

1-dimensional profile. DNA-binding proteins protect their binding sites from the Tn5 enzyme, creating small dips or "footprints" in the Tn5 insertion profile. Computational methods such as TOBIAS[1] and HINT-ATAC[2] have been developed to identify footprints in ATAC-seq signal. These footprints can be used in conjuction with motif information and genomic sequence to infer TF-binding events. However, since many TFs have very short residence times, it is estimated that ∼50% of TFs leave detectable footprints[1]. It is especially difficult to identify footprints of repressive TFs, as their binding generally decrease chromatin accessibility. While binding prediction for TFs that leave weak footprints can be improved by taking nucleosome position in to account[3], these conditions present challenges for predicting the binding of TFs genome-wide.

Beyond occluding DNA at their binding sites, TF binding may result in the recruitment of additional co-factors, non-coding RNAs (ncRNAs), and chromatin remodeling complexes. These factors may further alter nucleosome density and positioning through nucleosome eviction or chemical modification of histone tails and DNA. The existence of cell-type specific motif grammars that direct TF binding coupled with the fact that TF binding events shape local chromatin accessibility patterns suggest a learnable relationship between DNA sequence and chromatin accessibility.

The use of deep neural networks to learn the relationship between DNA sequence and chromatin accessibility has become increasingly popular[3,4,5]. One such model, ChromBPNet[4], uses DNA sequence to predict a 1-dimensional Tn5 insertion profile along with total accessibility for a given ATAC-seq peak. Once a ChromBPNet model has been trained, the sequence features that contribute the most to the model's predictions can be inferred using deep-learning interpretability frameworks such as DeepLIFT[6]. High-scoring sequence segments overwhelmingly correspond to TF-binding sites that are supported by Chromatin Immunoprecipitaion sequencing (ChIP-seq) experiments[4]. Analyzing recurring patterns in sequence attribution scores from models trained on accessible regions in a variety of cell-types allows for the investigation of cell-type specific motif grammars.

While a 1-dimensional Tn5 insertion profiles can reflect DNA-binding driven footprints, the interpretability of base-pair resolution ATAC-seq data can be improved by taking the length of mapped fragments into account. As DNA-binding proteins and complexes increase in size, increasingly long fragments are needed to span them, shifting the fragment length distribution around binding sites towards longer fragment lengths. Fragment length-driven visualizations of ATAC-seq data, such as V-Plots, can improve the interpretation of complex chromatin structure at CREs (Figure 1B,C). While not leveraging fragment length directly, PRINT[3] creates interpretable visualizations of ATAC-seq peaks by using a variable window-size to detect footprints, which allows for easier differentiation between individual TF-binding events, larger TF-complexes, and nucleosomes. The recently developed deep learning model seq2PRINT[3] can be trained to predict PRINT footprints from DNA sequence. Sequence attribution scores for seq2PRINT models can be analyzed in the same manner as those from ChromBPNet, however the impact of high-scoring regions is more interpretable due to the two-dimensional profile predictions. Generating high-quality training data for seq2PRINT models requires high-quality multiscale footprints. As ATAC-seq data becomes increasingly sparse, multiscale footprinting may become less reliable. This may lead to difficulties in training models on pseudobulked ATAC-seq data from small cell sub-populations of interest, such as transition cell-states in developmental trajectories. In contrast, ChromBPNet models do not rely on the initial footprinting step. Regions that may be too sparse for accurate footprinting still provide valuable information regarding the relationship between sequence and accessibility. Training across thousands of accessible peaks allows models to learn subtle motif grammars, even when the majority of peaks are sparse[4].

Here we propose seq2VPlot, which leverages fragment length to predict a two-dimensional ATAC-seq profile. By training on fragments directly, this approach avoids issues with footprinting on sparse data while maintaining the interpretability of a two-dimensional profile. seq2VPlot model predictions illustrate distinctions between DNA-binding events of various size and accurately learn the distribution of ATAC-seq fragment lengths across accessible regions. Using trained seq2VPlot models, we can generate sequence attribution scores that are corrected for Tn5 bias, allowing for the identification of TFBMs that impact local chromatin accessibility. Training seq2VPlot models on pseudobulked ATAC-seq data from different cell-types allows for the identification of cell-type specific motif grammars. Furthermore, assessing seq2VPlot predictions on perturbed sequences allows for the interpretation of the role that individual base pairs or sequence motifs play in shaping local chromatin accessibility. By perturbing TFBMs with high attribution scores, we can interpret how TF-binding events shape accessibility patterns. We can

further investigate the functional impact of disease-associated regulatory variants by introducing them *in silico* and analyzing changes in seq2VPlot predictions. Finally, by mutating pairs of TFBMs in tandem, we can identify potentially instances of cooperative TF binding. Investigating the cell-state specific relationships between DNA sequence and chromatin accessibility has implications for therapeutically-driven modulation of cell-state and synthetic CRE design.

Methods

Data preprocessing

ATAC-seq Peak Calling

scATAC-seq fragments files were downloaded from the NCBI Gene Expression Omnibus under accession number GSE216462. Fragments originating from early erythrocytes and CD4 cells were extracted using the provided cell-barcodes and cell-type annotations, creating a pseudo-bulk ATAC-seq datasets for both cell types. Peak calling was performed using ArchR's `getPeakSet()` function[6]. PyTorch tensors representing V-plots for each peak were created using the Pysam Python package, which allows for the efficient extraction of fragments residing within provided peak coordinates. Only fragments of length 25 - 250 were retained, as the fragment density outside of this range is sparse and could confuse seq2VPlot models.

Tn5 bias correction

Tn5 insertions are not uniformly distributed across accessible DNA. The enzyme has a sequence bias that causes a greater frequency of insertions in GC-rich regions[7]. Tn5 bias creates thin V-shapes centered on preferred insertion sites that are visible in V-Plots (Figure 2A). Sequence models that are trained to predict ATAC-seq data learn Tn5 insertion bias concurrently with TFBMs. This prevents the extraction of sequence attribution scores that purely reflect DNA-binding events. ChromBPNet performs bias correction by training a separate bias model on sparse non-peak regions. The lack of structure in these regions prevents the model from learning anything other than the Tn5 bias. The weights of the bias model are frozen, and a separate accessibility model is trained on peak-regions. The bias model output is added to the accessibility model output during training, regressing out the Tn5 bias signal and allowing the accessibility model to learn DNA-binding patterns. Feature attribution scoring is performed on the accessibility model to extract TFBMs.

It is difficult to apply this approach in the two-dimensional case. The fragment length distribution is impacted by DNA-binding events, so a Tn5 bias model should not learn to correctly distribute predicted insertion counts along the fragment length axis. Furthermore, the fragment length distribution varies across peak regions, meaning predicted insertion counts from a bias model cannot be distributed across the fragment length axis using the overall fragment length distribution. Therefore, seq2VPlot uses a different approach to correct for Tn5 bias. Prior to model training, Gaussian smoothing is applied to V-plot matrices to blur the bias signal.

Gaussian smoothing of V-Plots is performed using the `gaussian_blur()` function from `torchvision.transforms.functional`, with `kernel_size=(41,41)` and `sigma=(40,40)`. This approach retains the important features of a V-Plot (including nucleosome positioning and DNA-binding events), while removing the Tn5 bias signal. Furthermore, this approach is less expensive computationally relative to other bias correction approaches. ChromBPNet requires training a separate Tn5 bias model, as well as identifying an appropriate peak set to train the bias model[4]. seq2PRINT requires training a separate bias model on naked DNA and pre-computing two-dimensional bias-corrected footprints across all peaks[3]. Both approaches require training separate models, while Gaussian blurring of a peak set can be performed in minutes.

Sequence Preparation

DNA-sequences corresponding to peaks are extracted using the `getfasta` command from Bedtools[8], and sequences are converted to a 4×`seq_len` one-hot-encoded representation. Each ATAC-seq peak is 500 bp long, but an additional 500 bp is added to the end of each peak to prevent edge effects in model predictions.

Model architecture

seq2VPlot was developed using PyTorch, and largely follows the ChromBPNet architecture[4]. The model consists of a initial convolutional layer, a series of dilated convolutional layers, and a final convolutional layer. The initial convolutional layer takes in a one-hot encoded DNA sequence of shape 4×1500. The initial layer has 512 convolutional filters of shape 4×21 that learn TFBMs. Stride length is set to 1 and padding is set to "same," resulting in a 512×1500 dimensional embedding that is passed through a ReLU activation step. There are 8 dilated 1-D convolutional layers, each with 512 filters of shape 512×3. The size of the dilation and padding is set to 2^i, where i is the index of the dilated convolutional layer. This results in the identification of increasingly long-range sequence patterns. Following each dilated convolution, the embedding is passed through a ReLU activation. Residual connections made by adding the embedding produced by the previous convolutional layer to the current one. The final 1-D convolutional layer has 225 filters of shape 512×75. With a stride length of 1 and padding set to "same," this results in an output matrix of shape 225×1500 which represents a V-Plot.

Model training

During seq2VPlot training, model outputs are softmaxed by row, generating 225 multinomial distributions corresponding to predictions for each fragment length. Multinomial Negative Log Likelihood loss (MNLLL) is then calculated with respect to the counts values in the ground-truth V-plots. MNLLL is defined as follows:

$$\text{Loss} = -\log \mathbf{p}_{mult.}(k^{obs}|p^{pred}, n^{obs})$$

$$\mathbf{p}_{mult.}(k|p,n) = \frac{n!}{k_1!,...,k_d!} p_1^{k_1}...p_d^{k_d}$$

Where k^{obs} is the observed read counts from the input V-plot, p^{pred} are the predicted probabilities for each V-plot position as defined by the multinomial distribution, and n^{obs} is the total number of observed counts in the input V-plot. MNLLL ultimately reflects the negative likelihood of generating the input V-plot from the multinomial distribution parameterized by the predicted V-plot. The genomic position and fragment length marginal distributions are also calculated for each true and predicted V-plot by summing along the fragment length and genomic position axes, respectively. The cosine similarity is then computed between the ground-truth marginal distributions and the predicted marginal distributions. The final seq2VPlot loss function is defined as follows:

$$\text{Loss} = -\log \mathbf{p}_{mult.}(\mathbf{k}^{obs}|\mathbf{p}^{pred}, n^{obs}) - \alpha(\cos(x^{true}, x^{pred}) + \cos(y^{true}, y^{pred}))$$

Where x_{true} and x_{pred} correspond to the true and predicted genomic position marginal distributions and y_{true} and y_{pred} correspond to the true and predicted fragment length marginal distributions. The α parameter is set to 5 by default. To update model weights we use the Adam optimizer with learning rate 0.001. Models are trained for 50 epochs and validation loss is calculated after each epoch. The best model, defined by validation loss, is retained.

Feature attribution scoring

DeepSHAP

DeepSHAP attribution scoring was performed using implementation provided in the tangermeme[9] package. DeepLIFT[10] is a general framework for extracting feature attribution scores from a given input

to a neural network by explaining the difference in model output relative to some "reference" output in terms of the difference from the input from some "reference" input. DeepLIFT contribution scores exhibit the following property:

$$\sum_{i=1}^{n} C_{\Delta x_i \Delta_t} = \Delta_t$$

Where Δ_t is the difference from reference activation of some neuron t, and $C_{\Delta x_i \Delta_t}$ is the contribution of feature x_i to Δ_t (Δx_i refers to the difference from reference input for feature x_i). This property asserts that the sum of the attribution scores for each input feature explain the difference from reference activation. DeepLIFT attribution scores can be efficiently computed via backpropagation. This is done using values called multipliers, which are defined as follows:

$$m_{\Delta_x \Delta_t} = \frac{C_{\Delta x \Delta_t}}{\Delta_t}$$

Multipliers are similar in spirit to partial derivatives, they are just computed over a finite difference from reference input as opposed to an infinitesimally small difference. The multipliers can be chained together just like partial derivatives so the contribution of the original input features to the activation of the final output neuron can be computed efficiently via a single backpropagation step. For analyzing DNA sequence models, a set of n random sequences are used as a reference (we set $n = 60$). The DeepLIFT/SHAP algorithm implemented by tangermeme uses the DeepLIFT algorithm to estimate Shapely values[11]. DeepLIFT multipliers are defined in terms of Shapely values and recursively passed backwards through the model.

The DeepLIFT/SHAP algorithm is designed to explain a single scalar output of a model, however in our case the model output is a tensor of shape 225×1500. We use the method used by BPNet[9] reduce our output V-Plots to a scalar value:

$$c = \sum_{i,j} l_{ij} p_{ij}$$

Where c is the scalar value that we explain, l_{ij} is the output logit at postion i, j in the output V-plot, and p_{ij} is the weight of the logit given by the softmax of the output V-plot. This value is essentially a weighted sum of the logits, meaning profiles with sharp distinctions in fragment density will have higher scalars associated with them. As a result, DeepSHAP scores reflect how each base contributes to the "sharpness" of the profile, assuming that the original, unshuffled sequence is associated with the highest scalar. Importantly, this objective differs from directly assessing how each base affects the full profile shape itself. DeepSHAP approximates Shapley values, which formally represent the marginal contribution of a feature averaged over all possible combinations of other features. Since it would be computationally infeasible to compute exact Shapely values for all input bases, DeepSHAP estimates Shapley values using several "reference" sequences that are shuffled copies of the original. Consequently, the DeepSHAP score for a given base represents an estimate of its Shapley value contribution to a scalar summary of the original profile. Noise introduced by this approximation process, combined with artifacts from attributing relative to a scalar proxy rather than the full profile, likely contributes to noise in the resulting DeepSHAP scores.

Cosine distance

Obtaining smooth attribution scores is necessary for identifying "seqlets," or short regions of high attribution corresponding to TFBMs. In an effort to reduce noise and compute attribution scores that properly reflect the contribution of each base to the entire predicted profile, we use a simple cosine-distance based perturbation metric. To compute cosine-distance based attribution scores, we perturb each base in an input sequence individually. For each base, we generate three alternate sequences representing all possible point mutations at that position. We then compute the average seq2VPlot-predicted profile over the three mutant sequences, as well as the predicted profile for the wild-type sequence. Profiles are flattened into a 337500-dimensional vector (225×1500) and softmaxed. The attribution score for the base is computed as the cosine distance between the averaged mutant profile and the original profile, where cosine distance is defined as follows:

$$1 - \frac{M \cdot WT}{||M|| \cdot ||WT||} = 1 - \frac{\sum_{i=1}^{n} M_i WT_i}{\sqrt{\sum_{i=1}^{n} M_i^2} \sqrt{\sum_{i=1}^{n} WT_i^2}}$$

where M represents the average of the mutant profiles, and WT refers to the wild-type profile. This is equivalent to 1 minus the cosine similarity between the two profiles. Since cosine similarity has a range between -1 and 1, this cosine distance metric has a range between 0 and 2. The fact that the contribution of each base is computed in isolation, without considering the inclusion of other base combinations, is the obvious drawback of this approach compared to estimated Shapley values. Non-linear interactions between bases may be missed, especially when two bases redundantly contribute to the same profile feature. However, cosine-distance attribution scores are markedly smoother than DeepSHAP scores and spike at TFBMs, making them ideal for high-confidence seqlet-identification.

Motif discovery

To find recurring motifs with high-attribution, we use the implementation of the TF-MoDISco algorithm[12] provided by the tfmodiscolite package[13]. For motif discovery, we run `modisco motifs` with -n (seqlets per metacluster) set to 100,000. TF-MoDISco identifies "seqlets", or short sequence spans with high attribution score and generates a gapped k-mer representation for each seqlet. The top nearest neighbors for each seqlet are identified by computing the cosine similarity between all pairs of seqlets. Among the top nearest neighbors, a fine-grained similarity is calculated as the Jaccard index between gapped k-mer representations. This results in a sparse similarity matrix which is density adapted, allowing for the use of Leiden clustering to extract motif patterns. To identify potential TFs corresponding to each motif pattern, we run `modisco report` which internally uses TOMTOM[14] to compare discovered motif patterns against a motif database (we use JASPAR CORE 2024[15]). To identify instances of each discovered motif pattern in our attribution scores, we use the `call-hits` command from Fi-NeMo[16]

Shape clustering

To investigate the impact of a sequence motif on local chromatin accessibility patterns, we employ an *in silico* mutagenesis approach. For an ATAC-seq peak containing a motif of interest, we generate 200 copies of the underlying sequence. In each copy, we replace the motif site with a random sequence of the same length. We then generate seq2VPlot predictions for all mutated sequences and average them, resulting in a single profile that is not dependent on the motif site. We also generate the seq2VPlot prediction on the original sequence. Both the original and mutated profiles are exponentiated ($e^{\alpha x}$ for profile x, $\alpha = 0.5$), to emphasize regions of high fragment density. The difference between the original and mutated profiles generates a "shape" which illustrates the impact of removing the motif. We trim each shape to be of size 225×225 with the motif site at the center of the genomic position axis.

After generating shapes for each Fi-NeMo hit across all peaks, we normalize the values in each shape to be between -1 and 1 by dividing each shape by its maximum absolute value element. This prevents the magnitude of accessibility differences from impacting shape clustering. Each shape is flattened into a 50625-dimensional vector (225×225), and all shapes are combined to form a $n \times 50625$ matrix, where n is the number of Fi-NeMo hits. An `AnnData` object is then formed from this matrix, and we perform dimensionality reduction with the PCA implementation provided by `scanpy.tl.PCA()` with `n_comps=15` and `svd_solver="randomized"`. We then compute a nearest neighbors graph from the PCA embedding `scanpy.pp.neighbors()` with `n_neighbors=100` and `metric="cosine"`. Finally, we perform Leiden clustering from the nearest neighbors graph using `scanpy.tl.leiden()` with `resolution=0.7` and `adjacency=adata.obsp['connectivities']`.

Identifying nonlinear sequence interactions

To identify nonlinear interactions between base pairs, we utilize a cosine-distance based interaction score, defined as follows:

$$\text{Interaction}(A, B) = \frac{D_{cos}(\neg A, \neg AB))}{D_{cos}(\neg B, WT)}$$

$$\text{Interaction}(B, A) = \frac{D_{cos}(\neg B, \neg AB))}{D_{cos}(\neg A, WT)}$$

Where:

- WT - Model prediction on the unperturbed (Wild-Type) sequence.
- $\neg A$ - model prediction with base A mutated.
- $\neg B$ - model prediction with base B mutated.
- $\neg AB$ - model prediction with both A and B mutated.
- D_{cos} - cosine distance (see section on "Cosine distance" above)

If A and B are working cooperatively, we would expect $D_{cos}(\neg A, \neg AB)$ and $D_{cos}(\neg B, \neg AB)$ to be small, since perturbing either A or B would have roughly the same effect on the model prediction as perturbing both A and B together. Computing interaction scores in both directions (both Interaction(A, B) and Interaction(B, A) allows for the capture of directional relationships between sequence elements. For instance, binding of TF 2 to a motif containing B may require that TF 1 is bound to motif containing A. However, TF 1 may be able to bind to motif A on its own. In this scenario, we would expect Interaction(A, B) to be significantly smaller than Interaction(B, A), since there would be a large difference between just perturbing B and perturbing both A and B. If there is a redundancy between A and B, we expect both $D_{cos}(\neg A, \neg AB)$ and $D_{cos}(\neg B, \neg AB)$ to be much larger than the mean. The denominator normalizes the interaction score relative to the importance of A (or B) to the model prediction. If A occurs within a TFBM while B is relatively unimportant, then we would expect $D_{cos}(\neg B, \neg AB)$ to be quite large. However this is driven by the perturbation of base A and does not reflect a meaningful interaction between A and B. This effect is removed by normalizing the score with $D_{cos}(\neg A, WT)$. When performing a base-pair level sequence interaction analysis, we compute model outputs for all possible sequence perturbations and average them to obtain the perturbed output. For instance, to compute $\neg A$ where A is a single nucleotide, we compute the model output with all 3 possible mutations of A and average them together. To compute $\neg AB$ where AB is a pair of nucleotides, we compute all 9 possible ways of mutating A and B together and average to obtain the final perturbed V-plot.

Results

seq2VPlot accurately predicts a two-dimensional ATAC-seq signal

seq2VPlot predicts two-dimensional ATAC-seq profiles in the form of a V-Plot. The V-Plot of an accessible genomic region shows the midpoint of each ATAC-seq fragment that maps to it along the axis of genomic position (x-axis) and fragment length (y-axis). DNA-binding events, as well as Tn5 sequence preference, give the V-Plot its characteristic V-structure. Since a TFBM is intermittently occluded by a bound TF, the frequency of Tn5 insertions decreases in the motif. This creates a V-shaped fragment depletion, as the midpoints of increasingly long fragments with ends within the motif are plotted further from the motif center (Figure 1D). Larger DNA-binding proteins or complexes require longer fragments to "span" them, meaning that fragments that are shorter than the spanning length will be depleted at the binding site. This allows for the distinction between protein complexes (Figure 1D) and nucleosomes (Figure 1E), which occupy 147 base pairs when bound[17]. Distinguishing binding events of different size is more straightforward when they are separated by fragment length compared to when only a 1-dimensional profile is available (Figure 1F,G).

Convolutional Neural Networks (CNNs) robustly learn TFBMs[18], making them ideal for sequence-to-function modeling. seq2VPlot utilizes the BPNet[19] model architecture (Figure 1A), taking in a one-hot encoded DNA sequence and leveraging a series of 1-dimensional convolution layers to learn TFBMs and the relationships among them. A series of dilated convolutions with exponentially increasing dilation size allows for the detection of progressively longer range sequence patterns (see Methods for details). The final layer of the model uses 225 1-dimensional convolutional filters to construct a V-plot of shape $225 \times n$, where n is the length of the input DNA sequence. Each filter in the final layer essentially corresponds to a fragment length between 25bp and 250bp. Since the vast majority of ATAC-seq fragments lie within the range of 25bp to 250bp, fragments outside of this range are filtered out. During model training, V-Plot predictions are softmaxed and treated as multinomial distributions. Multinomial Negative Log-Likelihood loss, along with cosine similarity between the genomic position (x) and fragment length (y) marginal distributions are used as a loss function.

For model training, we used a 874,480-cell SHARE-seq of the human bone marrow[3]. Using the ATAC-seq portion of the dataset, we used ArchR[6] for peak calling and used provided cell-type annotations to call cell-type specific peaks. Here, we report on two models: one trained using 47,286 from 37,830 cells annotated as early erythrocytes, and another trained using 28,009 peaks from 139,195 annotated as CD4 cells. For each model, we randomly assigned 80% of the peaks to a training set and 20% to validation. seq2VPlot learns a robust relationship between DNA sequence and chromatin accessibility, predicting profiles that accurately reflect chromatin structure observed in ground truth V-Plots. seq2VPlot model predictions accurately depict binding complexes of various size (Figure 1B), as well as nucleosome positions (Figure 1C). Since the model is trained across thousands of accessible regions, it can predict detailed structure within CREs that have low read depth. For individual regions, the model accurately learns the distribution of Tn5 insertions along the genomic position axis (Figure 1H), as well as the distribution of fragment lengths(Figure 1I). The fragment length distribution over all regions is impacted by DNA-binding events of various size, as well as the periodicity of the DNA helix, which causes Tn5 insertions to be made at minimum 10 bp apart. seq2VPlot accurately predicts the fragment length distribution across all validation regions (Figure 1J), including oscillations caused by DNA periodicity. Unlike ChromBPNet, seq2VPlot does not have a head that predicts the total Tn5 insertions in a given region and instead only predicts the shape of the two-dimensional ATAC-seq profile. However, an estimate of the total counts prediction obtained by summing the logits of seq2VPlot predictions reveals that the model learns a moderate relationship between DNA sequence and total accessibility (Figure 1K).

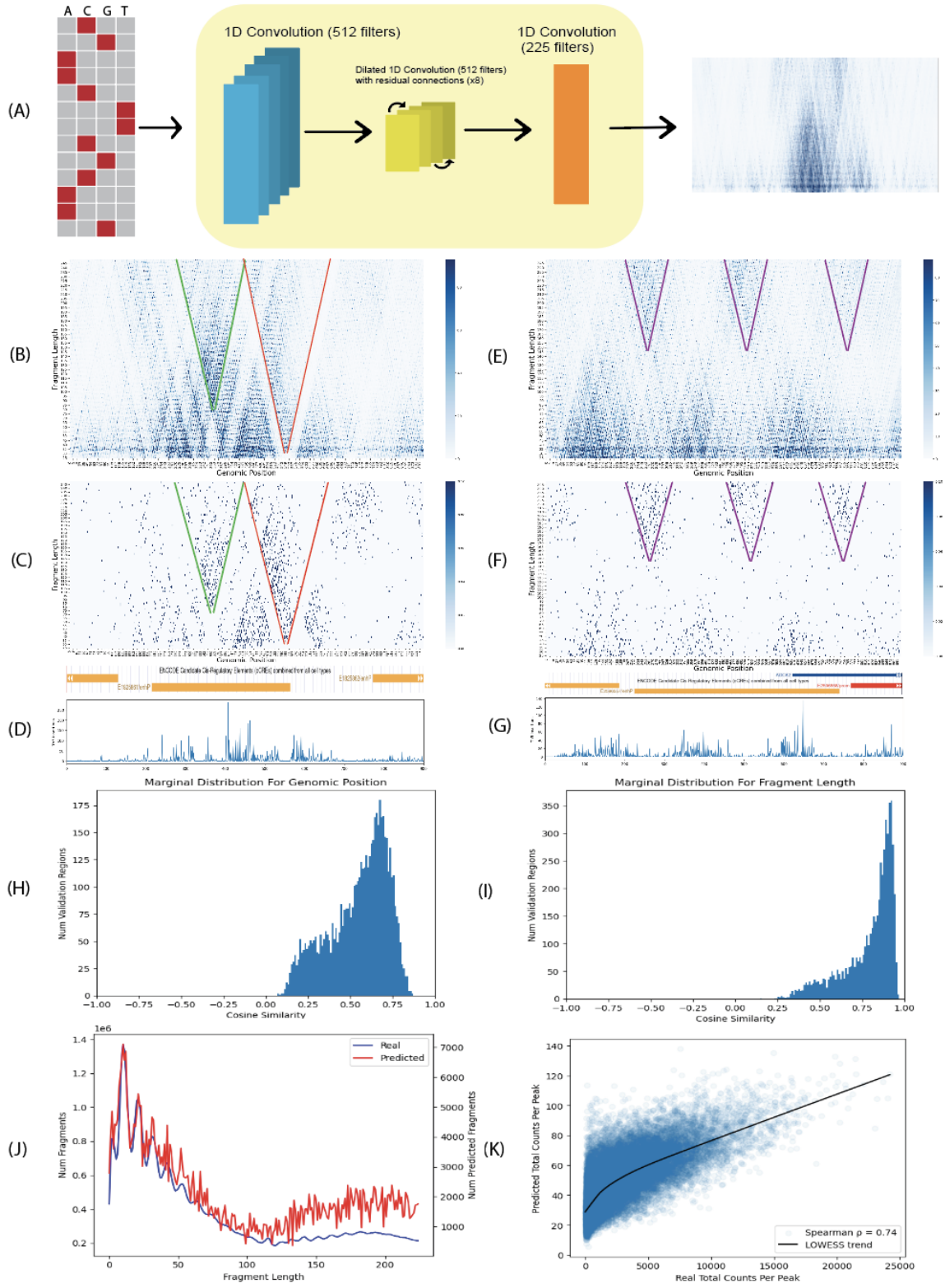

Figure 1: seq2VPlot accurately predicts ATAC-seq profiles in two-dimensions. A: Schematic of the seq2VPlot model architecture. B,C: Predicted (B) vs. actual (C) V-Plots of held out validation region chr16:71564612-71565512. An individual TF-binding site or small complex is outlined in red, and a larger protein complex is outlined in green. D: One-dimensional profile for the same region. E,F: Predicted (E) vs. actual (F) V-Plots of held out validation region chr7:140672528-140673128. Nucleosomes are highlighted in purple. G: One-dimensinoal profile for the same region. H,I: Distribution of cosine similarities between true and predicted marginal distrbiutions for genomic position (H) and fragment length (I) over all validation regions. J: True fragment length distribution (blue) and predicted (red) over all validation regions. K: True vs predicted total counts per validation region.

Gaussian smoothing prevents seq2VPlot from learning Tn5 insertion bias

Gaussian smoothing of V-Plots retains the key V-Plot features, including nucleosome positioning and protection patterns from binding events, while eliminating the V-patterns resulting from Tn5 bias (Figure 2A,B). seq2VPlot models trained on non-blurred V-Plots generate predictions that have many V-shaped patterns, clearing reflecting Tn5 bias (Figure 2C), while models trained on Gaussian blurred V-Plots predict only the important features of the V-Plot, ignoring the bias signal (Figure 2D).

To affirm that our Gaussian smoothing approach robustly prevents seq2VPlot from learning Tn5 insertion bias, we used Tangermeme[9] to compare DeepSHAP[11] sequence attribution scores extracted from our CD4-cell seq2VPlot model to attribution scores extracted from a ChromBPNet model trained on the same data. While Tn5 generally prefers GC-rich regions, the ends of ATAC-seq fragments have been used to construct a motif reflecting Tn5 insertion preference[7] (Figure 2E). We used Find Individual Motif Occurrences (FIMO) tool from the MEME Suite[14] to identify Tn5 motif sites in CD4-cell ATAC-seq peaks. We then calculated the average DeepSHAP attribution score within each Tn5 motif, for four separate models: (1) a seq2VPlot model trained on V-Plots without Gaussian blur, (2) a full ChromBPNet model (bias model plus accessiblity model), (3) a seq2VPlot model trained on V-Plots with Gaussian blur, and (4) just the accessibility component of the ChromBPNet model. Attribution scores for seq2VPlot models trained on unblurred V-Plots appear similar to those for full ChromBPNet models, as both model types reflect Tn5 bias and DNA-binding information concurrently. Similarly, both seq2VPlot models trained on Gaussian blurred V-Plots and ChromBPNet accessibility models are bias-corrected, reflecting only DNA-binding motifs. Therefore, the distribution of average attribution scores within Tn5 motif sites for model types 1 and 2 is skewed right and has a much longer tail relative to the same distribution for model types 2 and 4 (Figure 2F,G). The visibly similar change in the distribution of average attribution scores within Tn5 hits for seq2VPlot and ChromBPNet after bias correction indicates that Gaussian smoothing is an effective approach.

Despite bias correction, DeepSHAP sequence attribution scores retain considerable noise (Figure 2I). For a given input feature to a neural network, DeepSHAP calculates the contribution of that feature to the output of any neuron in the network. However, DeepSHAP scores can only be computed for scalar outputs, meaning that profile predictions must be reduced to scalar values in order to compute attribution scores. ChromBPNet computes profile attribution scores by computing the dot product between the model logits and the softmax of the logits, which results in a scalar output that is compatible with DeepSHAP attribution scoring. However, reducing profiles to a scalar value likely contributes to noise in the resulting attribution scores. Obtaining smooth attribution scores is necessary for identifying "seqlets," or short regions of high attribution corresponding to TFBMs. In an effort to reduce noise and compute attribution scores that properly reflect the contribution of each base to the entire predicted profile, we use a simple cosine-distance based perturbation metric. Cosine-distance attribution scores are markedly smoother than DeepSHAP scores (Figure 2H), and clearly spike at TFBMs (Figure 2I), making them ideal for high-confidence seqlet identification.

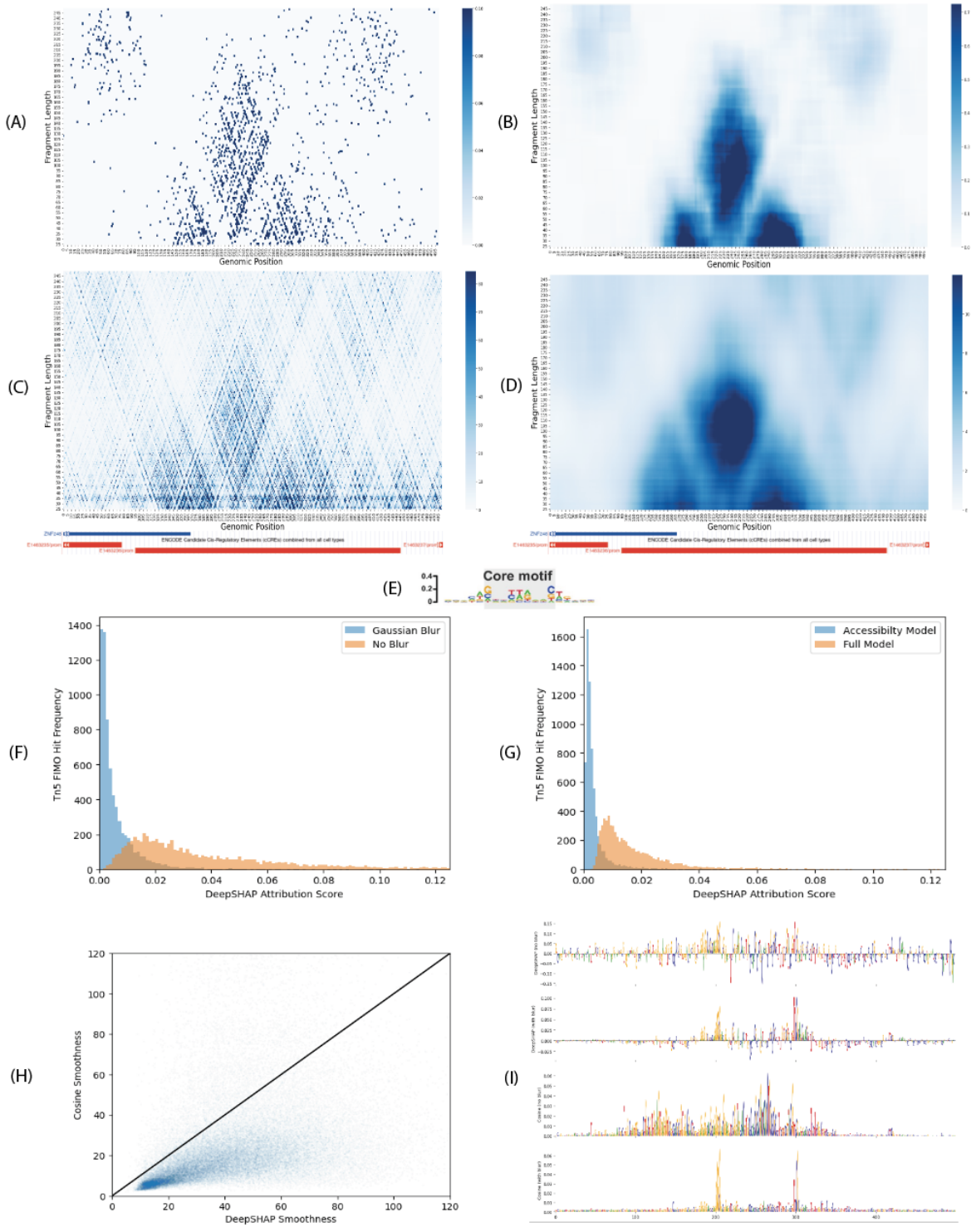

Figure 2: Gaussian smoothing prevents seq2VPlot from learning Tn5 insertion bias. A,B: Original (A) vs Gaussian smoothed (B) V-Plot for chr10:37857443-37857943, the ZNF248 promoter region. Guassian blurring maintains important features of the V-Plot while removing Tn5 bias. C,D: seq2VPlot predictions for the same region after being trained on non-blurred (C) vs. blurred (D) V-Plots. E: The Tn5 insertion motif as determined by Zhang *et al.*, 2021. F: distribution of average DeepSHAP attribution score within Tn5 FIMO hits for seq2VPlot models trained without Gaussian blurring (orange) vs with blurring (blue). G: distribution of average DeepSHAP attribution score within Tn5 FIMO hits for a full ChromBPNet model (orange) vs. only the accessibility component of the ChromBPNet model (blue). H: Smoothness of DeepSHAP attribution scores vs. cosine-distance scores across all peaks in CD4 cells. Smoothness is defined as the sum of the normalized difference in attribution between adjacent bases. I: Comparision of seq2VPlot attribution scores over chr11:74988627-74989127. top-panel: DeepSHAP scores from a model trained on unblurred V-Plots, top-middle: DeepSHAP scores from a model trained on blurred V-plots, bottom-middle: cosine distance scores from a model trained on unblurred V-Plots, bottom: cosine distance scores from a model trained on blurred V-Plots.

Cell-type specific seq2VPlot models learn distinct motif grammars

Chromatin accessibility patterns are dynamic across cell types, even at the same genomic locus. The accessibility of the MXD1 promoter varies between early erythrocytes and CD4 cells. An additional ∼150bp of DNA at the edge of the promoter is accessible in early erythrocytes but not in CD4 cells, where there is a well-positioned nucleosome instead (Figure 3A,B). Two separate seq2VPlot models trained on early erythrocytes and CD4 cells correctly predict this difference. The early erythrocyte model predicts an enrichment of short fragments in the ∼150bp region outside the promoter, while the CD4 model predicts a nucleosomal signal (Figure 3C,D). Additinally, there appears to be an enrichment of fragments on the opposite side of the promoter (closer to the TSS), that appears in the CD4 model prediction but not in the prediction from the early erythrocyte model (Figure 3C,D).

Analyzing the cosine attribution scores for the MXD1 promoter sequence for both models reveals that different sequence elements drove each model prediction, suggesting that the model has learned a cell-type specific motif grammar. It is evident that the CCAAT motif in the ∼150bp region at the end of the promoter drove the early erythrocyte model to predict increased fragment density in that region (Figure 3E). Conversely, the same CCAAT motif received very little attribution in the CD4 model, which instead predicts a nucleosome in the region (Figure 3F). The CCAAT motif is recognized by components of the NF-Y complex, which has pioneering capabilities and generally increases accessibility surrounding its binding site[20]. This does not mean that the NF-Y complex is not active CD4 cells. In fact, the CD4 model predicts high attribution associated with the CCAAT motif on the opposite end of the promoter (Figure 3F), which is not apparent in the early erythrocyte model attribution scores (Figure 3E). A possible explanation for this distinction is that the binding of the NF-Y complex is sequence-context dependent. Directly adjacent to the high-scoring CCAAT motif in CD4 cells is a GC-box that also has high attribution (Figure 3F). The GC-box is frequently bound by members of the KLF and SP TF families, which cooperate with the NF-Y complex in a variety of cellular contexts[21,22]. In this case, the KLF/SP and NF-Y complex interaction appears to be specific to CD4 cells, as the same region is assigned low attribution in early erythrocytes (Figure 3D). However, the KLF/SP TFs and the NF-Y complex may interact at other loci in early erythrocytes, meaning that this interaction is likely locus-specific. It is possible that the sequence context surrounding the GC-box and the CCAAT motif plays a role in specifying cell-type specific TF interactions.

In order to recover enriched motif patterns across all ATACs-seq peaks, we use TF-MoDISco (Transcription-Factor Motif Discovery from Importance Scores)[12], a motif discovery algorithm designed for analyzing sequence attribution scores from machine learning models. TF-MoDISco identifies high-scoring sequence segments, or "seqlets", corresponding to TFBMs and clusters them, allowing for the identification of recurring motifs with high attribution. To identify instances of each motif pattern across all peaks, we use Fi-NeMo (Finding Neural network Motifs)[16], which is designed to identify instances of high-scoring TF-MoDISco patterns from a set of attribution scores.

Running this pipeline separately on cosine attribution scores from the early erythrocyte and CD4 models resulted in 16 identified sequence patterns for each cell-type (Figure 3G,H). The GC-box motif is the most frequent hit in both cell types, reflecting the widespread activity of KLF and SP family transcription factors that recognize GC-rich sequences across cellular contexts[23]. Frequent CCAAT hits across both cell types is consistent with the ubiquitous activity of NF-Y regulating nucleosomal occupancy at promoters[20], and the high frequency of CTCF motif hits reflect its integral role in 3D chromatin organization across cell types[24]. Additionally, motifs that are pivotal in the establishment of both the erythropoetic and $CD4^+$ T-cell lineages have frequent hits in their respective cell-types. For instance, the high-frequency of GATA family TFs in early erythrocytes reflect their role as master regulators of erythropoiesis[25]. RUNX1 is known to be instrumental in the maturation of $CD4^+$ T-cells[26], and the IRF family of TFs play a crucial role in the differentiation of $CD4^+$ T-cells into several subtypes[27].

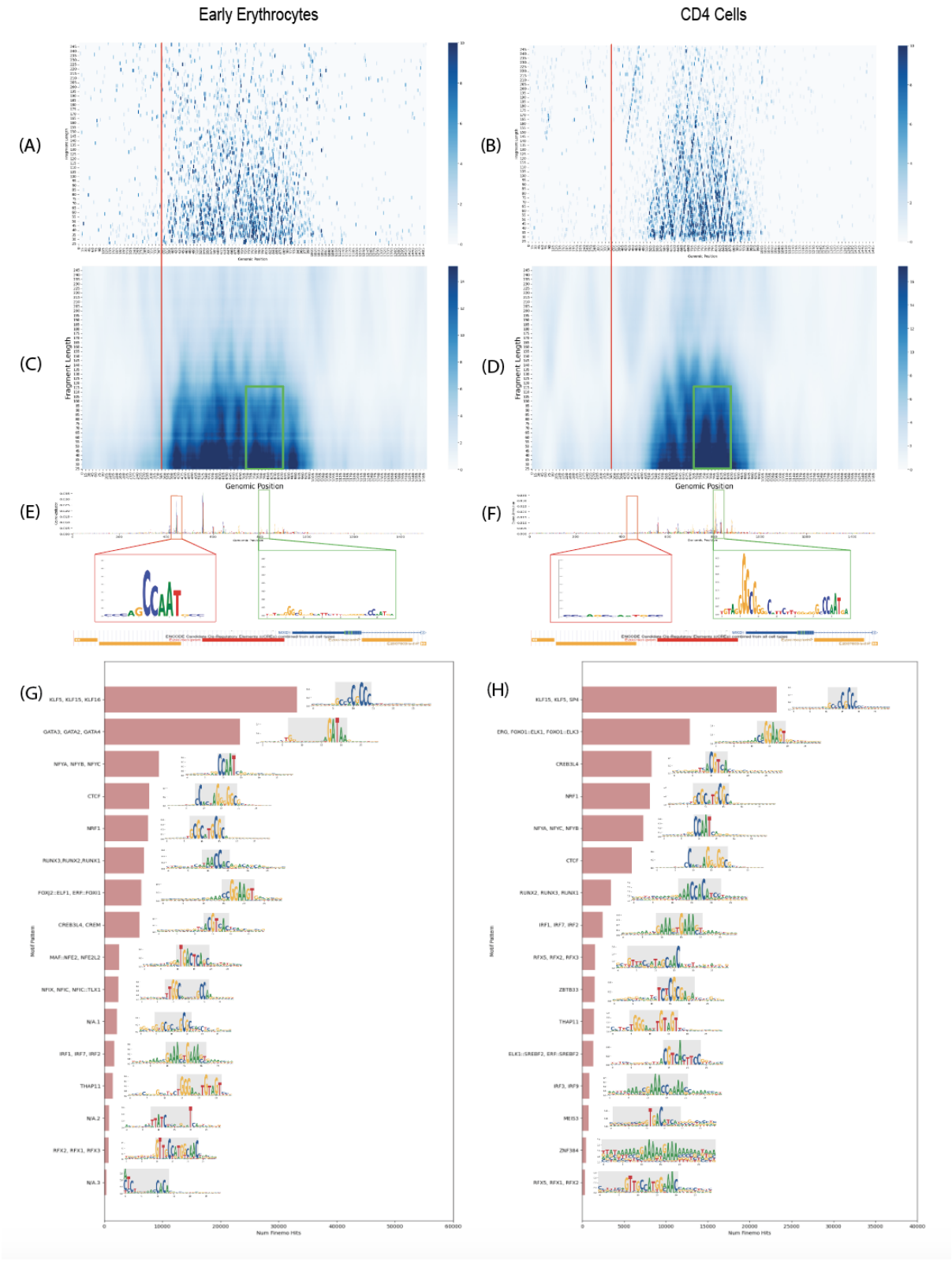

Figure 3: Cell-type specific seq2VPlot models learn distinct motif grammars. A,B: V-Plots for the MXD1 promoter (chr2:69914680-69915180) generated using ATAC-seq fragments from early erythrocytes (A) and CD4 cells (B). C,D predicted profiles for the same region generated by seq2VPlot models trained on ATAC-seq data from early erythrocytes (C) and CD4 cells (D). The red line demarcates a ∼150bp region of increased accessibility in early erythrocytes, and the green box highlights an area of increased fragment density in CD4 cells. E,F: cosine-distance attribution scores for the same region extracted from the early erythrocyte model (E) and the CD4 cell model (F). Red and green boxes highlight regions with major differences in attribution between cell types. G,H: Number of Fi-NeMo hits corresponding to sequence motifs discovered by TF-MoDISco in early erythrocytes (G) vs. CD4 cells (H).

seq2VPlot illustrates the impact of TF-binding on local chromatin accessibility patterns

Increased interpretability is the primary advantage of predicting a two-dimensional ATAC-seq profile. To explore how the binding of individual TFs may alter local chromatin accessibility patterns, we perform *in silico* mutagenesis of TFBMs. At each prospective TF-binding site, characterized by a Fi-NeMo hit in a peak region, we replace the hit sequence with 200 random sequences of the same length and average seq2VPlot model predictions across all mutated sequences. This approach effectively generates a profile that is not dependent on the sequence at the motif region. Subtracting the mutated profile from the original profile generates a "difference plot," or "shape", that reveals how the presence of the motif impacts fragment density across fragment lengths.

To investigate the impact of each motif pattern identified by TF-MoDISco on chromatin accessibilty patterns, we performed *in silico* mutagenesis and generated shapes for each Fi-NeMo hit for each motif pattern across all peaks. This resulted in 113,030 shapes from the early erythrocyte model, and 79,684 shapes for the CD4 cell model. To identify recurring patterns in the shape sets, we performed Principal Component Analysis (PCA) followed by Leiden community detection (see Methods for details), resulting in 9 shape clusters for early erythrocytes (Figure 4A,B,C,D) and 10 clusters for CD4 cells (Figure 4X,X). The shapes reveal diverse impacts for TFBMs on local chromatin accessibility patterns, including increasing low-fragment length density (early erythrocyte shapes 1-3), maintaining adjacent nucleosomes around a short region of accessibility (early erythrocyte shape 5, 6), maintaining a protein complex (early erythrocyte shape 4) and decreasing fragment density across fragment lengths (early erythrocyte shape 7).

To identify the shapes associated with each motif pattern, we created motif pattern/shape pairs for each Fi-NeMo hit. For each motif pattern, we then calculated the fraction of Fi-NeMo hits corresponding to each shape. To further identify over-represented motif pattern/shape tuples, we generated 1000 random shuffels of the labels for each motif pattern and shape across all Fi-NeMo hits, resulting in a background distribution of shape frequency for each motif pattern. Comparing the true shape frequency distribution relative to the background distribution revealed enriched shapes for each motif pattern (Figure 4E,F).

Several of the enriched motif pattern/shape pairs correspond to known biological functions of the associated TFs. For instance, the role of the pioneering role GATA-family TFs in erythropoiesis[25] is supported by the strong association of GATA motifs with a gain of short short fragments. This pattern is consistent with the displacement of nucleosomes around the GATA binding site. The CTCF motif pattern is strongly associated with a small region of increased accessibility around the binding site with two well-positioned nucleosomes on either side, which reflects the high degree of nucleosome organization commonly observed in the vicinity of CTCF binding sites[28]. The NRF1 motif demonstrates a strong association with a V-shaped pattern indicative of protein complex formation. This is consistent with the tendency of NRF1 to form homodimers[29]. NRF1 is also known to form complexes with several co-factors, including PGC (PPARγ co-activator)-1α, PGC-1β, PRC (PGC-1-related co-activator), and PARP-1 (poly(ADP-ribose) polymerase 1), which results in the recruitment of the PARP-1·DNA-PK·Ku80·Ku70·topoisomerase IIβ-containing protein complex to promoters[23–33]. In early erythroyctes, the RUNX1 motif appears strongly associated with a motif pattern associated with a decrease in fragment density around its binding site. This perhaps reflects its role as a repressor of the erythroid gene expression program[34], and its capacity to directly recruit Polycomb Repressive Complex 1 (PRC1)[35]. Notably, RUNX1 appears to be highly active in CD4 cells, but it is instead associated with a shape suggesting an increase in accessibility (Figure 4D,F). This is consistent with its role of RUNX1 as a major player in the maturation of CD4+ T cells[26].

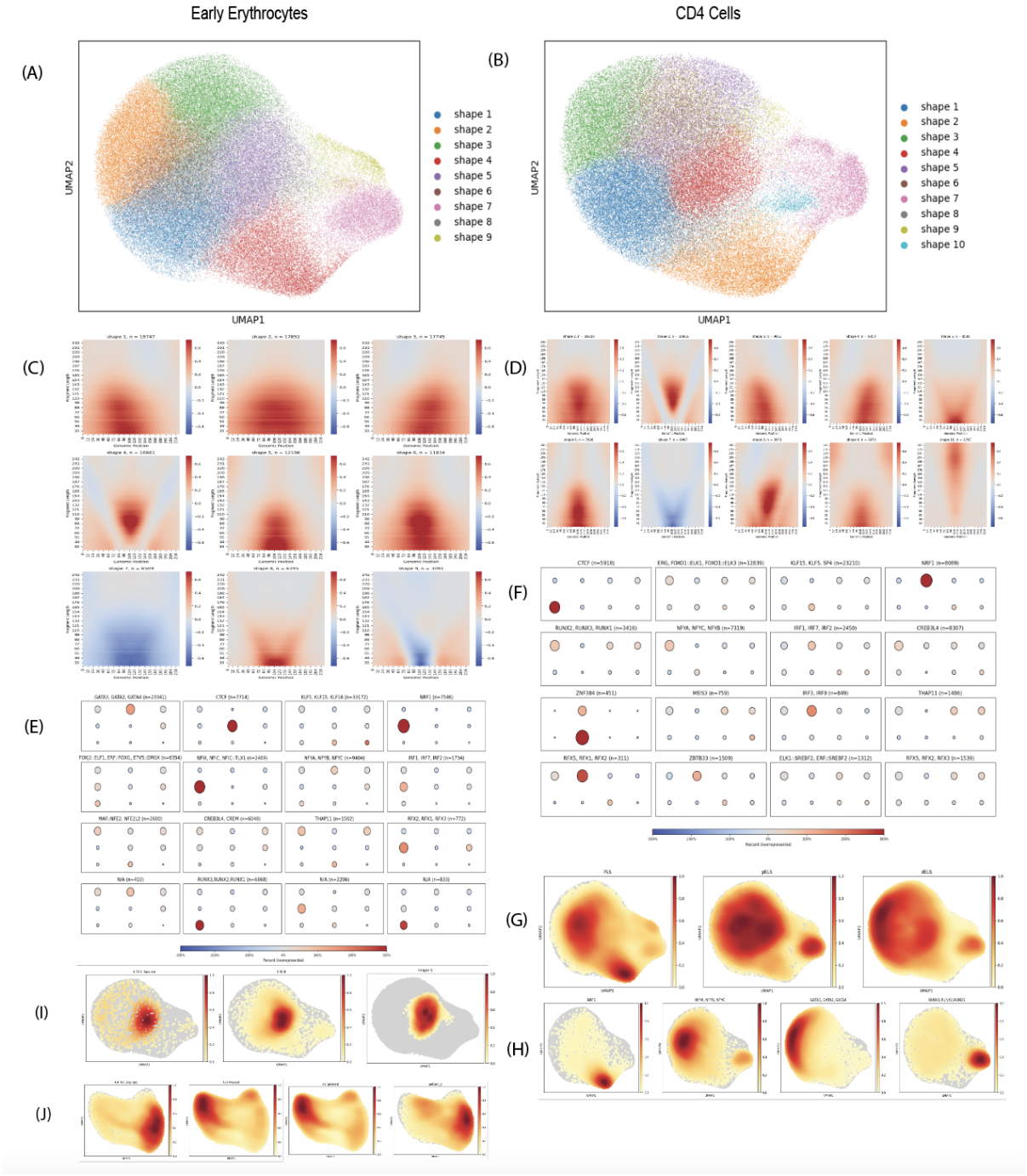

Figure 4: seq2VPlot illustrates the impact of TF-binding on local chromatin accessibility patterns. A,B: UMAPs generated from the first 15 components of PCA embeddings generated from all shapes associated with Fi-NeMo hits in early erythrocytes (A) and CD4 cells (B). Each dot represents the shape associated with a Fi-NeMo hit, colored according to its cluster assignment from Leiden community detection. C,D: average of each shape cluster discovered by Leiden community detection for early erythrocytes (C) and CD4 cells (D). The shape profiles are arranged in ascending order by assigned cluster number. E,F: Breakdown of shapes associated with each TF-MoDISco discovered motif in early erythrocytes (E) and CD4 cells (F). Each motif pattern is labeled with possible corresponding TFs identified by TOMTOM. Each dotplot "grid" corresponds to the grid of shape clusters in C,D. The size of each dot reflects the fraction of Fi-NeMo hits corresponding to each shape, while the color represents how over or under represented a given motif/shape pair is based on the overall frequency of motifs and shapes. G: Density of overlaps of Fi-NeMo hits with ENCODE annotated CRE-types (overlayed on the UMAP in A). H: Density of Fi-NeMo hits corresponding to labeled motif patterns (overlayed on the UMAP in A). I: Density of ENCODE annotated CTCF-bound sites (left), CTCF Fi-NeMo hits (middle) and the strongest CTCF-associated shape (right). J: UMAP embedding of a PCA performed on all seq2VPlot predictions of all early erythrocyte peaks. Density plots show the presence/absence of peaks overlapping with GATA1 ChIP-seq peaks, and peaks containing/missing the Fi-NeMo hits for the GATA motif.

We next asked whether our motif patterns and shapes associate with different classes of CREs. Using the ENCODE annotations of CREs as having a Promoter-Like Signature (PLS), proximal Enhancer-Like Signature (pELS), or distal Enhacer-Like Signature (dELS)[36], we used Bedtools[8] to compute the overlap between annotated CREs and Fi-NeMo hits. This analysis revealed that CRE signatures associate with distinct shapes (Figure 4G). For instance, shapes corresponding the NF-Y complex and NRF1 binding show strong association with promoters, consistent with the frequent binding of these TFs at promoter regions[20,29] (Figure 4H). Furthermore, shapes corresponding to GATA-family binding show strong association with distal enhancers, consistent with GATA1 ChIP-seq data in early erythrocytes[37] (Figure 4H). Interestingly, the repressive activity of RUNX1 appears to occur at proximal and distal enhancer elements as opposed to promoters (Figure 4H). ENCODE additionally includes annotations for CTCF-bound regions (which are not otherwise annotated as PLS, pELS, or dELS). The density of CTCF-bound regions overlaps with Fi-NeMo hits corresponding to the CTCF motif, as well as the CTCF-associated shape (Figure 4I), supporting the argument that motifs with high attribution scores represent active TF binding sites. This assertion is further supported by the overlap between GATA1 ChIP-seq peaks in erythrocytes and accessible regions with high-scoring GATA motifs from the early erythrocyte model (Figure 4J).

seq2VPlot predicts the impact of disease-associated regulatory variants on local chromatin accessibility patterns

The success of generating informative attribution scores from single-base mutations suggested that seq2VPlot may be effective predicting changes in chromatin accessibility profiles stemming from Single Nucleotide Polymorphisms (SNPs). We examined a set of fine-mapped variants for 11 heritable traits relevant to blood disorders from the UK Biobank[38]. After filtering for variants with posterior inclusion probability > 0.1 and those within called early erythrocyte ATAC-seq peaks, we had a set of 392 variants. We then introduced each variant individually *in silico* and generated model predictions on the mutated sequences. For variant, we generated an associated "shape" illustrating its impact on local chromatin accessibility patterns by subtracting the predicted profile from the mutated sequence from the predicted profile from the wild-type sequence (just as was done for each Fi-NeMo hit). To identify variants with a similar impact on accessibility, we performed PCA followed by Leiden clustering on their associated shapes (Figure 5G). We discovered 8 shape clusters that illustrate the diverse impacts of regulatory variants on accessibility patterns, including a reduction in short-fragment length density (shape 1), an increase in short-fragment length density (shape 2), and a disruption of protein-DNA complex formation (shape 8) (Figure 5H).

To illustrate the ability of seq2VPlot to predict changes in local accessibility patterns resulting from regulatory variants, we highlight two examples. The first is rs17534048, a C¿G mutation associated with Mean Corpuscular Volume (MCV) (PIP=0.99) or the average size of red blood cells. The variant occurs in a GATA motif adjacent to proximal enhancer regions within a BTG2 intron. The GATA motif in which the variant occurs also overlaps with a GATA1 ChIP-seq peak[39]. In the seq2VPlot prediction from the wild-type sequence, a high-density region of short fragments is observed above the intact GATA motif (Figure 5A), likely because GATA1 is binding and driving an accessible chromatin state. In the predicted profile from the mutated sequence, however, the region of accessibility above the motif is significantly diminished, likely because GATA1 no longer recognizes the motif (Figure 5B). It is possible that the decreased accessibility resulting from lack of GATA1 binding in this region impacts the expression of the BTG2 gene, which has been linked to MCV in mice[40].

For the second example, we use the CD4 cell model to predict the impact of rs112401631, a fine-mapped variant for asthma (PIP=0.27). The variant also occurs in a distal enhancer region predicted to regulate the CCR7 gene in CD4 cells by SCARlink[41], and it colocalizes with an eQTL linked to CCR7, chr17:40600717:G (PP.H4=0.9022; coloc[42]). Furthermore, overexpression of CCR7 in $CD4^+$ T-cells has been linked to patients with atopic asthma[43]. Relative to the prediction on the wild-type sequence (Figure 5D), the seq2VPlot prediction on the mutated sequence displays markedly increased accessibility levels along the enhancer (Figure 5E). It is possible that the increased accessibility caused by the variant increases the activity of the distal enhancer, leading to increased CCR7 expression and greater asthma risk.

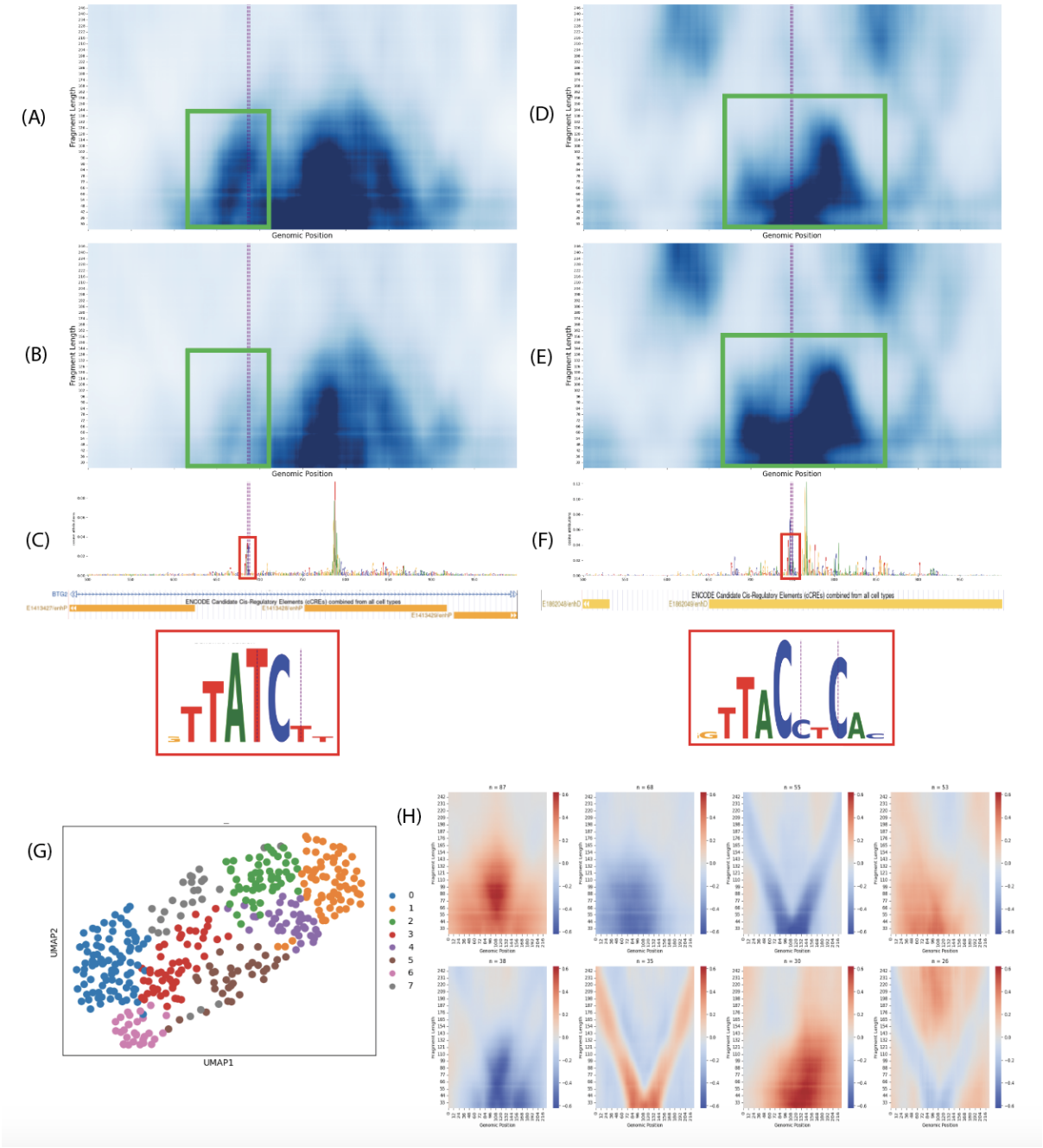

Figure 5: seq2VPlot predicts the impact of disease-associated regulatory variants on local chromatin accessibility patterns. A,B: early erythrocyte seq2VPlot model predictions of chr1:203306091-203306591, which contains rs17534048 (chr1:203306280:C:G, outlined in purple) for the sequence with the major allele (A) and the minor allele (B). Green boxes highlight changes in fragment density after the introduction of the mutation. C: cosine attribution scores for the same region, red box highlights the variant position within a GATA motif. D,E: CD4 seq2VPlot model predicitons of chr17:40607522-40609022, which contains rs112401631 (chr17:40608272:T:A, outlined in purple) for the sequence with the major allele (D) and the minor allele (E). Green boxes highlight changes in fragment density after the introduction of the mutation. F: cosine attribution scores for the same region, red box highlights the variant position. G: UMAP of a PCA embedding of all shapes generated from *in silico* introduction of blood-trait associated regulatory variants overlapping early erythrocyte ATAC-seq peaks. Colors indicate cluster assignment from Leiden community detection. H: Average of each shape cluster identified by Leiden community detection.

seq2VPlot identifies nonlinear interactions between sequence elements

seq2VPlot has great potential to investigate cooperative interactions between TFs. IF two TFs bind cooperatively to adjacent motifs, we would expect a mutation in just one of the motifs to significantly decrease the binding ability of both TFs. Consequently, the additive effect of mutating both motifs separately would far exceed the impact of mutating both together, suggesting a synergistic (nonlinear) interaction between the motifs. To identify nonlinear interactions between individual base-pairs at accessible peaks, we employ a cosine distance-based framework (see Methods for details) to compute an interaction score. Briefly, we generate mutated sequences with all possible individual and pairwise mutations for a given ATAC-seq peak and generate seq2VPlot predictions for each sequence. Using cosine distance, we compute the difference-from-additive effect of mutating all pairs of bases in tandem versus mutating them separately. This results in an interaction map for a given region that highlights nonlinear interactions between base pairs.

To demonstrate the effectiveness of this approach, we consider an accessible region in early erythrocytes with high attribution scores at the NRF1 motif, which frequently binds as a homodimer at promoters[29]. Mutating either "half" of the motif is likely to disrupt homodimer formation. Consequently, we observe a spike in interaction score between bases in either half of the motif (Figure 6A). This interaction is evident when evaluating the seq2VPlot model predictions on sequences with either neither, one, or both "halves" of the NRF1 motif intact. The model prediction on the wild-type sequence demonstrates a strong V-shaped pattern of Tn5 insertion protection surrounding the NRF1 motif (Figure 6B), consistent with homodimer formation. However, when just one half of the motif is mutated, this pattern completely disappears (Figure 6C,D), revealing that both components are necessary for homodimer formation. The protection pattern is also not observed when both motifs are mutated (Figure 6E). Exploring interactions between motifs across peaks is difficult due to the computational demands of computing pairwise interaction maps. However, this issue may be mitigated by comparing interaction scores between high-attribution regions with a background distribution.

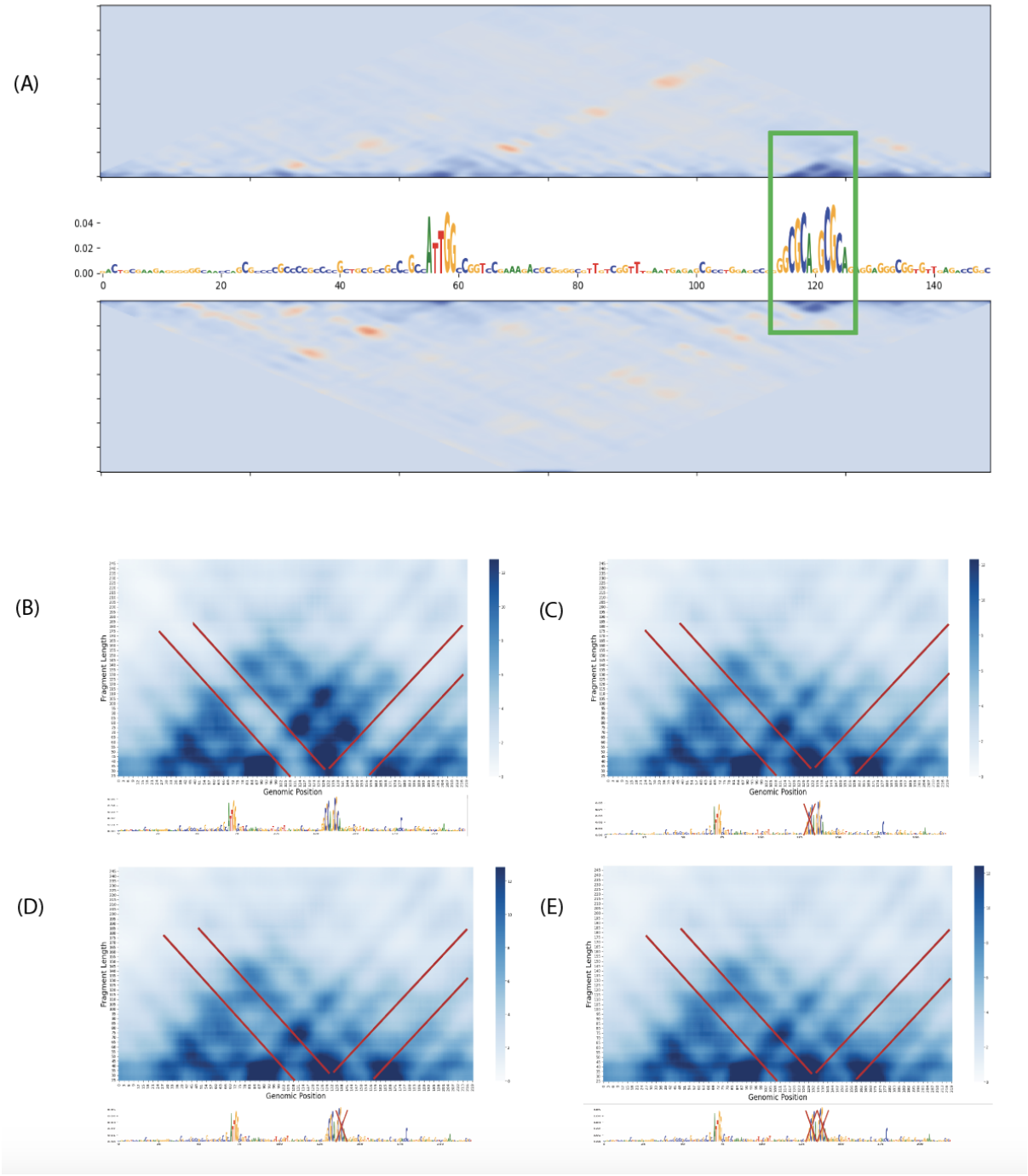

Figure 6: seq2VPlot identifies nonlinear interactions between sequence elements. A: Interaction map for chr1:108876861-1088767011 (GPSM2 Promoter). The upper triangle of the interaction map shows the Z-score of Interaction(A, B) while the lower triangle shows the Z-score of Interaction(B, A), where the index of nucleotide A is less than that of B. See Methods for details. B,C,D,E seq2VPlot predictions on chr1:108876861-1088767011 when both halves of the NRF1 motif are present (B), just one half is present (C,D), and when both are absent (E). Red lines highlight presence of V-shaped Tn5 protection caused by binding of the NRF1 homodimer in (B) and corresponding absence in C,D, and E.

Discussion

The ability of seq2VPlot models to accurately predict ATAC-seq profiles in two-dimensions suggests a robust relationship between DNA sequence and chromatin accessibility patterns. The fact that this relationship is learnable suggests that the DNA sequence within CREs consists of functional patterns that are repeated across regulatory elements. It is probable that the strength of the sequence-to-accessibility relationship is primarily driven by the sequence specificity of TFs, which bind to preferred motifs and initiate epigenetic cascades that result in changes in local chromatin accessibility patterns. Furthermore, the cell-state specificity of the sequence-to-accessibility relationship is likely driven by differential expression of TFs. Even among TFs that are ubiquitously expressed across cell-types, variability of available co-factors likely influences the ability of TFs to bind to their motifs in different cellular contexts. Consequently, distinct motif-grammars emerge that define the relationship between sequence and accessibility across cell states.

The two-dimensional profiles generated by seq2VPlot reveal the diversity of impacts on chromatin accessibility patterns caused by TF binding. The same TF may exhibit differing effects at separate binding sites both within and across cell-types, increasing local accessibility in some contexts and decreasing it in others. Many factors may contribute to differing TF behavior across binding sites, including differential isoform expression, post-translational modifications to TFs, and association with a range of co-factors. While difficult to validate, the fact that a sequence model predicts a diversity of impacts across binding sites suggests that sequence context plays a pivotal role in determining the behavior of bound TFs. In addition to dictating the localization of co-factors, sequence context around TFBMs can influence variables such as DNA shape and nucleosome positioning, both of which could play a role in TF binding and function. In future work, the relationship between sequence context and the impact of TF binding will be systematically explored. This will contribute to a comprehensive characterization of cell-state specific motif grammars.

The ability of seq2VPlot to predict the impact of disease-associated regulatory variants on chromatin accessibility patterns could dramatically increase the pace at which precision medicine is deployed to treat genetic diseases. Model predictions could aid in the identification of causative variants, allowing clinicians to predict genetic predisposition to certain disease states more accurately. Furthermore, predicting the impact of regulatory variants on chromatin accessibility patterns could prompt the use of guide-RNA localized epigenetic editing complexes that counteract the effects of the variant, potentially reducing the severity of symptoms.

As the capabilities of generative DNA language models such as Evo2[44] improve, sequence-to-function models will become increasingly important in predicting how AI-generated DNA sequences will behave in different cellular contexts. seq2VPlot could be used to explore the accessibility landscape at synethetic CREs across cell-types. This information could inform the construction of synthetic gene regulatory modules that are designed to regulate specialized transcriptional programs in response to specific cellular cues. This technology would greatly increase our ability to modulate cell-state for therapeutic aims, contributing to applications such as cancer immunotherapy and regenerative medicine.

The recent and rapid improvement of sequence-to-function models has outpaced our ability to comprehensively validate them with wet-lab experimentation. However, it is probable that these models will become increasingly integrated into the experimental design process. seq2VPlot models could be used to screen for TFs that are likely to shift chromatin accessibility patterns in certain cell-states, informing the selection of targets for wet-lab experiments and increasing the pace of biological discovery.

References

1. **Bentsen, M., Goymann, P., Schultheis, H. *et al.*** ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nature Communications* **11**, 4267 (2020). DOI: https://doi.org/10.1038/s41467-020-18035-1

2. **Li, Z., Schulz, M.H., Look, T. *et al.*** Identification of transcription factor binding sites using ATAC-seq. *Genome Biology*, **20**, 45 (2019). DOI: https://doi.org/10.1186/s13059-019-1642-2

3. **Hu, Y., Horlbeck, M., Zhang, R. *et al.*** Multiscale footprints reveal the organization of cis-regulatory elements. *Nature* **638**, 779-786 (2025). DOI: https://doi.org/10.1038/s41586-024-08443-4

4. **Pampari, A., Shcherbina, A., Kvon, E *et al.*** ChromBPNet: bias factorized, base-resolution deep learning models of chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor footprints and regulatory variants. *bioRxiv* (2025). DOI: https://doi.org/10.1101/2024.12.25.630221

5. **Yan, H., Kelley, D.** scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural networks. *Nature Methods* **19**, 1088-1096 (2022). DOI: https://doi.org/10.1038/s41592-022-01562-8

6. **Granja, J., Corces, M., Pierce, E.*et al.*** ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature Genetics* **53**, 403-411 (2021). DOI: https://doi.org/10.1038/s41588-021-00790-6

7. **Zhang, H., Lu, T., Liu, S. *et al.*** Comprehensive understanding of Tn5 insertion preference improves transcription regulatory element identification. *NAR Genomics and Bioinformatics* **3**, 4 (2021). DOI: https://doi.org/10.1093/nargab/lqab094

8. **Quinlan, A., Hall, I.** BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 6 (2010). DOI: https://doi.org/10.1093/bioinformatics/btq033

9. **Schreiber, J., Hansen, B.V., Murphy, A., He, A.Y.** Biological sequence analysis for the modern age (Tangermeme). *GitHub* (2024). https://github.com/jmschrei/tangermeme

10. **Shrikumar, A., Greenside, P., Kundaje, A.** Learning Important Features Through Propagating Activation Differences. *ICML'17: Proceedings of the 34th International Conference on Machine Learning* **70**, 3145 - 3153 (2017). DOI: https://doi.org/10.48550/arXiv.1704.02685

11. **Lundberg, S.M., Lee, S.** A Unified Approach to Interpreting Model Predictions. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems* (2017). DOI: https://doi.org/10.48550/arXiv.1705.07874

12. **Shrikumar, A., Tian, K., Ziga, A. *et al.*** Technical Note on Transcription Factor Motif Discovery from Importance Scores (TF-MoDISco) version 0.5.6.5. *arXiv preprint* (2018). DOI: https://doi.org/10.48550/arXiv.1811.00416

13. **Schreiber, J., Raine, I., Wang, A. *et al.*** tfmodisco-lite. *GitHub* (2025). DOI: https://github.com/jmschrei/tfmo-lite

14. **Bailey, T., Johson, J., Grant, C., Noble, W.** The MEME Suite. *Nucleic Acids Research* **43**, (2015). DOI: https://doi.org/10.1093/nar/gkv416

15. **Rauluseviciute, I., Riudavets-Puig, R., Blanc-Mathieu, R., *et al.*** JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* **52**, (2024). DOI: https://doi.org/10.1093/nar/gkad1059

16. **Wang, A., Kundu, S., Yun, C.M. & Kundaje, A. *et al.*** austintwang/finemo_gpu. *GitHub*(2025). https://github.com/austintwang/finemo_gpu

17. **Richmond, T., Davey, C.** The structure of DNA in the nucleosome core. *Nature* **423**, 145-150 (2003). DOI: https://doi.org/10.1038/nature01595

18. **Kelley, D., Snoek, J., Rinn, J.** Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research* **7**, 990-999 (2016). DOI: 10.1101/gr.200535.115

19. **Avsec, Ž., Weilert, M., Shrikumar, A. et al.** Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics* **53**, 354-366 (2021). DOI: https://doi.org/10.1038/s41588-021-00782-6

20. **Oldfield, A., Henriques, T., Kumar, D. *et al.*** NF-Y controls fidelity of transcription initiation at gene promoters through maintenance of the nucleosome-depleted region. *Nature Communications* **10**, 3072 (2019). DOI: https://doi.org/10.1038/s41467-019-10905-7

21. **Suske, G.** NF-Y and SP transcription factors — New insights in a long-standing liaison *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **5**, 590-597 (2017). DOI: https://doi.org/10.1016/j.bbagrm

22. **Dolfini, D., Imbriano, C., Mantovani, R** The role(s) of NF-Y in development and differentiation. *Cell Death & Differentiation* **32**, pages195–206 (2025). DOI: https://doi.org/10.1038/s41418-024-01388-1

23. **Kaczynski, J., Cook, T., Urrutia, R.** Sp1- and Krüppel-like transcription factors. *Genome Biology* **4**, 206 (2003). DOI: https://doi.org/10.1186/gb-2003-4-2-206

24. **Kim, S., Yu, NK., Kaan, BK.** CTCF as a multifunctional protein in genome regulation and gene expression. *Experimental Molecular Methods* **47**, e166 (2015). DOI: https://doi.org/10.1038/emm.2015.33

25. **Romano, O., Petiti, L., Felix, T. *et al.*** GATA Factor-Mediated Gene Regulation in Human Erythropoiesis. *iScience* (2020). DOI: 10.1016/j.isci.2020.101018

26. **Hsu, FC., Shapiro, M., Dash, B. *et al.*** An Essential Role for the Transcription Factor Runx1 in T Cell Maturation. *Scientific Reports* **6**, 23533 (2016). DOI: https://doi.org/10.1038/srep23533

27. **Huber, M., Lohoff., M.** IRF4 at the crossroads of effector T-cell fate decision. *European Journal of Immunology* **7**, 1886-1985 (2014). DOI: https://doi.org/10.1002/eji.201344279

28. **Clarkson, CT., Deeks, E., Samarista, R. *et al.*** CTCF-dependent chromatin boundaries formed by asymmetric nucleosome arrays with decreased linker length. *Nucleic Acids Research* **47**, 21 (2019). DOI: https://doi.org/10.1093/nar/gkz908

29. **Liu, K., Li, W., Xiao, Y. *et al.*** Molecular mechanism of specific DNA sequence recognition by NRF1. *Nucleic Acids Research* **52**, 2 (2024). DOI: https://doi.org/10.1093/nar/gkad1162

30. **Andersson, U., Scarpulla., R.** PGC-1-Related Coactivator, a Novel, Serum-Inducible Coactivator of Nuclear Respiratory Factor 1-Dependent Transcription in Mammalian Cells. *Molecular and Cellular Biology* **21**, 11 (2001). DOI: https://doi.org/10.1128/MCB.21.11.3738-3749.2001

31. **Wu, Z., Puigserver, P., Andersson, U. *et al.*** Mechanisms controlling mitochondrial biogenesis and respiration through the thermogenic coactivator PGC-1. *Cell* **98**, 115-124 (1999). DOI: 10.1016/S0092-8674(00)80611-X

32. **Lin, J., Tarr P., Yang, R. *et al.*** PGC-1beta in the regulation of hepatic glucose and energy metabolism. *J Biol Chem* **33**, 30843-8 (2003). DOI: 10.1074/jbc.M303643200

33. **Hossain, M., Ji, P., Anish., R. *et al.*** Poly(ADP-ribose) Polymerase 1 Interacts with Nuclear Respiratory Factor 1 (NRF-1) and Plays a Role in NRF-1 Transcriptional Regulation. *Journal of Biological Chemistry* **284**, 13 (2009). DOI: https://www.jbc.org/article/S0021-9258(20)32409-1/fulltext

34. **Kuvardina, O., Herglotz., J. Kolodziej, S. *et al.*** RUNX1 represses the erythroid gene expression program during megakaryocytic differentiation. *Blood* **15**, 23 (2015). DOI: https://doi.org/10.1182/blood-2014-11-610519

35. **Yu, M., Mazor, T., Huang, H. *et al.*** Direct Recruitment of Polycomb Repressive Complex 1 (PRC1) to Chromatin by Core Binding Transcription Factors. **45**, 3 (2012). DOI: 10.1016/j.molcel.2011.11.032

36. **The ENCODE Project Consortium** Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020). DOI: https://doi.org/10.1038/s41586-020-2493-4

37. **Cheng, Y., Wu, W., Kumar, SA. *et al.*** Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications, and mRNA expression. *Genome Research* **12**, 2172-84. (2009). DOI: 10.1101/gr.098921.109

38. **Bycroft, C., Freeman, C., Petkova, D. *et al.*** The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018). DOI: https://doi.org/10.1038/s41586-018-0579-z

39. **Ulirsch, J., Lareau, C., Bao, E. *et al.*** Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nature Genetics* **51**, 683–693 (2019). DOI: https://doi.org/10.1038/s41588-019-0362-6

40. **Koscielny, G., Yaikhom, G., Iyer, V. *et al.*** The International Mouse Phenotyping Consortium Web Portal, a unified point of access for knockout mice and related phenotyping data. *Nucleic Acids Research* **42**, D1 (2014). DOI: https://doi.org/10.1093/nar/gkt977

41. **Mitra, S., Malik, R., Wong, W. *et al.*** Single-cell multi-ome regression models identify functional and disease-associated enhancers and enable chromatin potential analysis. *Nature Genetics* **56**, 627–636 (2024). DOI: https://doi.org/10.1038/s41588-024-01689-8

42. **Giambartolomei, C., Vukcevic, D., Schadt, E. *et al.*** Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLOS Genetics* (2014). DOI: https://doi.org/10.1371/journal.pgen.1004383

43. **Wiest, M., Upchurch, K., Yin, W. *et al.*** Clinical implications of CD4+ T cell subsets in adult atopic asthma patients. *Allergy, Asthma & Clinical Immunology* **14**, 7 (2018). DOI: https://doi.org/10.1186/s13223-018-0231-3

44. **Brixi, G., Durrant, M., Ku, J. *et al.*** Genome modeling and design across all domains of life with Evo 2. *bioRxiv* (2025). DOI: https://doi.org/10.1101/2025.02.18.638918