Refining the cis-regulatory grammar learned by sequence-to-activity models by increasing model resolution

Nuria Alina Chandra[1], Yan Hu[2,3], Jason D. Buenrostro[2,3], Sara Mostafavi[1,4$], Alexander Sasse[1,5,6,7$]

[1] Paul G. Allen School of Computer Science and Engineering, University of Washington, WA, USA, 98195
[2] Gene Regulation Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, 02142 USA.
[3] Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, 02138 USA.
[4] Canadian Institute for Advanced Research, Toronto, ON, Canada, MG51ZB
[5] Heidelberg University, Heidelberg, Germany, 69120
[6] Center for Molecular Biology Heidelberg (ZMBH), Heidelberg, Germany, 69120
[7] Center for Synthetic Genomics, Heidelberg, Germany, 69120

[$] Co-senior authorship, to whom the correspondence should be addressed to: saramos@cs.washington.edu, a.sasse@zmbh.uni-heidelberg.de

Keywords: ATAC-seq, chromatin, gene regulation, deep learning, sequence-to-function models

Abstract

Chromatin accessibility can be measured genome-wide with ATAC-seq, enabling the discovery of regulatory regions that control gene expression and determine cell type. Deep genomic sequence-to-function (S2F) models link underlying genomic sequences to the measured chromatin state and identify motifs that regulate chromatin accessibility. Previously, we developed AI-TAC, a S2F model that predicts chromatin accessibility across 81 immune cell types and identifies sequence patterns that control their differential ATAC-seq signals. While AI-TAC provided valuable insights into the regulatory patterns that govern immune cell differentiation, later research established that ATAC-seq profiles (the distribution of Tn5 cuts) contain additional information about the exact location and strength of TF binding. To make use of this additional information, we developed bpAI-TAC, a multi-task neural network which models ATAC-seq at base-pair resolution across 90 immune cell types. We show that adding ATAC-profile information consistently improves predictions of differential chromatin accessibility. We also demonstrate that simultaneous learning of related cell types through multi-task modeling leads to better predictions than single task models. We then present a systematic framework for comparing how differences in model performance can be attributed to differences in what the model has learned. To understand what additional information bpAI-TAC gleans from ATAC-profiles, we use sequence attributions and identify motifs that have different effect sizes when trained on profiles. We conclude that modeling ATAC-seq at base-pair resolution enables the model to learn a more sensitive representation of the regulatory syntax that drives differences between immunocytes, and therefore will improve predictions of variant effects.

Introduction

The establishment of distinct cell types depends on differential gene expression, a process that is regulated by thousands of transcription factors (TFs) that bind to open chromatin regions (OCRs) and recruit transcriptional machinery [1,2]. TFs are able to recognize specific DNA sequences in the genome, and the combination of these molecular interactions produces a complex regulatory language that is engraved in cis-regulatory elements (CRE) across the entire genome [3,4]. This cis-regulatory language precisely controls when, where, and how much of each gene is transcribed in a cell type specific manner. Chromatin accessibility can be measured genome-wide with ATAC-seq [5], using the Tn5 enzyme's ability to cut DNA in accessible OCRs and insert sequencing adapters. The number of Tn5 insertions per region is a measure of *accessibility* [6]. Open chromatin regions span a couple hundred base pairs, and from the ATAC-seq signal alone it is not straightforward to determine which bases are bound by TFs that regulate accessibility [7].

To overcome this limitation, Deep genomic sequence-to-function (S2F) models have been developed to link chromatin accessibility to the cis-regulatory grammar that orchestrates accessibility [8–12]. AI-TAC is a S2F model that was developed to identify cis-regulatory sequence elements that control differential chromatin accessibility across mouse immune cells [10]. It uses a convolutional neural network that takes as input DNA sequences and predicts the cumulative number of Tn5 insertions (hence, accessibility) across different immune cell types in a multi-task manner [6]. While AI-TAC has shown promising results, it has recently been established that the distribution of Tn5 insertions contains valuable information about the exact location and strength of transcription factor (TF) binding sites, a phenomenon also called TF "footprints" [13–15]. BPnet [16] and its ATAC-seq successor ChromBPnet [17] recently demonstrated that training on base-pair (bp) resolution profiles significantly improves models' ability to discover regulatory motifs and their precise binding strength [11,12,18].

Here, we present bpAI-TAC, a multi-task model that predicts ATAC-seq counts at base-pair resolution across immune cell types (**Figure 1a**). Similar to AI-TAC, bpAI-TAC was trained using ATAC-seq data from 90 different mouse immune cell types collected by the Immunological Genome Project [6]. While ChromBPnet is trained as an ensemble of models, with a single model for every cell type to avoid "motif leakage", we trained bpAI-TAC as a multi-task model to enable cross-cell type comparisons that are not confounded by different model training runs. This approach also reduces the time and resources required for training. We show that additional information within ATAC-seq profiles consistently results in improved predictions of differential chromatin accessibility compared to AI-TAC. Moreover, we demonstrate how modeling choices, the training objective and resolution affect model performance. We find that multi-task modeling improves predictions, and that bias-factorization is not necessary for the model to make better predictions. To understand what additional information the model learns from ATAC-profiles, we use sequence attributions and determine additional motifs of TFs and unknown factors that bpAI-TAC utilizes for its predictions in specific cell types while AI-TAC ignores them in this context [19,20].

Results

BpAI-TAC uses information of ATAC-profiles to improve predictions of differential chromatin accessibility

AI-TAC is a convolutional neural network (CNN) that predicts the number of Tn5 insertions in 250bp sequences around peaks of OCRs across 81 different immunocytes [10]. Instead of predicting an aggregate value of chromatin accessibility per genomic region, bpAI-TAC predicts per base-pair number of Tn5 insertions (**Figure 1a**). In particular, following previous model architectures [16,17] the model takes as input 1Kb genomic DNA and makes predictions from two output heads: the “profile head” predicting the distribution of Tn5 insertions for 250bp in its center, and the “accessibility head” predicting the aggregate number of insertions for that region. The multiplication of the predicted profile and the accessibility per region recreates the number of insertions per base [16,18] (**Supp. Methods**).

As shown previously [17], the Tn5-enzyme has a strong sequence preference that impacts local distribution of Tn5 insertions [13,14], and so can hide TF footprints. Under this assumption, without correction, the model will have to learn the sequence motifs that are associated with the Tn5 enzyme and might miss signals that come from TF motifs [14,16,17]. Nevertheless, we note that the total number of insertions in a given region are purely measures of biological accessibility and hence the learned motifs from a “binned” output model do not reflect the base-pair resolution Tn5 bias. To address the “profile bias” for base-pair resolution modeling, we follow the strategy of (Chrom)BPnet and model the distributions Tn5 insertions with a second CNN. We train this bias model on protein-free DNA from [13] (**Supp. Methods**). This model learns the distribution of Tn5 insertions without interference of DNA binding proteins. Just as (Chrom)BPnet, bpAI-TAC adds the ‘log-likelihood’ of the bias-model to the ‘log-likelihood’ of the modeled ATAC-profiles before the sum is being transformed to the predicted ATAC profile via a softmax function. Past research has shown that adding a bias model is necessary to force the model to only learn the TF footprints during the profile prediction and ignore the strong sequence signal of the Tn5 enzyme [17,18].

While correctly modeling the bp-profiles is important for “deep footprinting” (i.e. determining the exact binding location and binding strength of a TF) [12,17], our analysis primarily focuses on how well these models can predict differences in chromatin accessibility between cell types. To quantify this, we use Pearson correlation R across cell types as a metric to compare bpAI-TAC to an equivalent version of AI-TAC (**Figure 1b,c**). This version of AI-TAC uses the exact same architecture as bpAI-TAC, simply without training on the bp-profile predictions. We observe that bpAI-TAC consistently outperforms AI-TAC (**Figure 1b, S1a**), suggesting that it learns additional grammar that distinguishes chromatin accessibility across cell types. bpAI-TAC also improves predictions across test set OCRs for all cell types (**Figure 1c, S1b**), suggesting that the profile information also helps the model to discover cell type specific factors that contribute to the variance between OCRs. Stratified on the coefficient of variation (CV), we observe larger gains for OCRs with high CVs across cell types (**Figure 1d**), suggesting that bpAI-TAC is better at learning sequence motifs that predict accessibility in cell-type specific OCRs. In summary, we observe that modeling ATAC-seq profiles improves the model’s understanding of chromatin grammar across different metrics.

Tn5 profiles improve open chromatin predictions consistently across of modeling choices

BpAI-TAC uses a hyperparameter (lambda) to combine separate objectives for profile and accessibility heads, as in (Chrom)BPNet, into a composite loss function [16,17] (**Figure 1a, Supp. Methods**). First, we compared different loss functions, composite and combined (i.e directly applied to the multiplication of cumulative counts and profiles), to determine the best suited objective function for our model. Surprisingly, although ATAC-seq counts follow a Poisson distribution, we found that training our model directly on per base count predictions results in worse predictive performance than AI-TAC (**Figure 2a*, *S2a**). On the other hand, the approximation of the Poisson loss as Mean squared error (MSE) of logged counts and the multinomial negative log-likelihood (MNLL) [16] significantly improved predictions at the best mixing fraction lambda (**Supp. Methods**). However, while the composite loss with the MNLL represents the theoretically correct approximation of the Poisson loss [16], we observe that a mixture of MSE and cross-entropy (CE) loss achieves the best performance for predicting differences between cell types. The CE can be interpreted as an unweighted MNLL, where each loss from comparisons of distributions is weighted equally, while they are weighted by the cumulative counts in the MNLL (**Supp. Methods**), therefore paying more attention to profiles with high numbers of counts.

This rather unexpected result may be due to the fact that bpAI-TAC is a multi-task network, predicting accessibility for multiple cell types simultaneously. To investigate if multi-tasking helps our model to extract more information from the data than an ensemble of single-track models, we trained an ensemble of models for the average ATAC-signal of cell lineages (10 lineages in 90 cell types). We compared the predictions of the ensemble against the predictions of a model that was trained on the same data in a multi-task fashion, and the lineage-averaged predictions from our model trained on individual cell types. Both multi-task models show improved predictions across cell lineages compared to the ensemble of single-task models (**Figure 2b**). Additionally, predictions are also improved in the other direction for individual cell lineages across held-out OCRs, suggesting that the multi-task model does not only improve scaling between tasks, but also extracts more information about the regulatory code (**Figure S2b**).

The composite loss functions enable us to adjust the models' focus on the profile predictions with the hyperparameter lambda. To optimize this hyperparameter and investigate its influence on the model's predictions, we trained models with different weighting of the profile head by increasing lambda from 0 (only training on scalar head, i.e. AI-TAC) to 1 (only training on profile). We found that models which included base resolution profiles in their loss, even with non-optimal lambdas showed superior prediction for total accessibility compared to AI-TAC (**Figure 2c**). However, when the model was weighted too strongly towards profiles (lambda > 0.6) the performance for predictions of total chromatin accessibility declined again and became near random for a model that was only trained on profiles (lambda = 1.0). We observed the same trend for the correlations for individual cell types across OCRs (**Figure S2c**).

Next, we investigated which modeling choices contributed to the superior performance of bpAI-TAC besides the loss function (**Figure 2d, S3a**). First, we replaced the deep architecture that used 9 dilated residual convolutional blocks with a shallower version that only used 4 blocks (*shallow*), and observed minimal impact on performance. When we replaced the body of the model entirely with a single convolutional layer (*base*), we observed significantly decreased performance, likely due the inability of this model to learn positional effects of motifs, motif context, or motif interactions. Replacing the entire scalar head with global mean-pooling of the base pair representations over the entire OCR (*GlobPoolAccHead*), as used by (Chrom)BPnet, also had a negative impact on the model performance. Replacing the global mean-pooling with smaller max-pooling over 50 base pairs (*Max50AccHead*) led to even worse performance. On the other hand, when we replaced the scalar head with a massive fully connected layer (*FCAccHead*) instead of multiple convolutions and pooling layers, the model significantly underperformed and reached almost random performance, suggesting that purely fully connected layers are inferior to well suited convolutional layers for this task.

Then, we investigated the influence of the bias model on the model's ability to learn differential accessibility (**Figure 2e, S3b**). We trained three bias models solely on the profile loss with the bpAI-TAC architecture using: 1) protein-free DNA from [13], and 2) on accumulated profiles from closed OCRs across cell types. Closed OCRs were defined by the probability from peak calling in each of the 90 cell types (**Supp. Methods**). Individual closed regions only contain very few reads, so we decided to sum over the counts of this region across cell types in which the OCR is closed. This strategy also controls the number of data points and speeds up training. 3) We trained a shallower bias model on the protein-free data. All three models' profiles were highly concordant (**Figure S4 a,b**). On held-out chromosomes and held-out protein-free DNA, both deep models were able to predict their own and the other data's profiles with high performance (**Figure S4 c,d**). In addition, to test the importance of bias modeling, we trained a model without bias. While it has been shown that it is more difficult to interpret such a model [17], we observe that this model is performing equally well or slightly better at predicting differences across cell types than models with a bias (**Figure 2e**), and also performs similar to the other models for cell types across held-out OCRs (**Figure S3b**)

One of the main bottlenecks for modeling at bp-resolution with a multi-task model is the required memory. Modeling every base-pair increases the data requirement by a factor equal to the length of the sequence. On the other hand, TF binding sites are between 5-20 bp wide [21,22], so that bp-resolution may not be necessary to extract information about TF footprints. To investigate this hypothesis, we trained models on profiles at 5, 10, and 20bp resolution. Here, we also added the Tn5-bias to the single base pair resolution log-likelihood of our model but then summed over the respective number of base-pairs in the bin before taking the softmax (**Supp. Methods**). As before, we compared the predicted versus the measured accessibility on the held-out set. We observe that larger resolution is not able to extract as much information from the profiles, resulting in slightly worse performance in predicting total accessibility than the model that was trained on single base pair resolution (**Figure 2f**). We hypothesize that this might be due to the rough binning that can contain only partial foot prints or footprints form TFs that do not show clear valleys of multiple base pairs (see below).

BpAI-TAC learns additional motifs that drive immune cell differentiation from TF footprints.

Biologically meaningful prediction improvements would also require bpAI-TAC to learn a more comprehensive cis-regulatory language. To investigate this, we performed a careful model interpretation, comparing motifs learned by bpAI-TAC with AI-TAC. Specifically, we identified sets of sequences that bpAI-TAC consistently predicted well, and better than AI-TAC across ten model initializations (**Figure 3a**, **Supp. Methods**). We identified the motifs in each of these sequences that were driving model predictions for each cell lineage from the 90 cell types [19]. From visual inspection, we observed that bpAI-TAC consistently identified clear motifs in these sequences while AI-TAC missed motifs completely or partially (**Figure 3b, S5**). We then extracted seqlets of identifiable motifs from the attributions of these sequences from AI-TAC and bpAI-TAC, jointly clustered them, and determined motifs that are recognized by either one or both models (**Supp. Methods**). We determined 36 motif clusters that were significantly enriched in attributions from the bpAI-TAC model over attributions from AI-TAC (Fisher exact test, **Figure S6**).

The three most enriched cluster groups were associated with RELA/REL-like motifs, Ptf1a/TFEB-like motifs and a group of motifs (cluster 53) that did not resemble a known TF. Both RELA and TFEB have important roles in immune response but their motifs are not picked up by AI-TAC in the investigated regions for these cell lineages [23–25]. To investigate if AI-TAC misses these motifs globally or just in the context of these sequences, we inserted them into 1,000 random sequences and determined the predicted effect over random sequences for both models (**Figure 3c**). We observe that these motifs in fact also have a predicted effect in AI-TAC, however the effect sizes differ significantly for many cell lineages compared to bpAI-TAC (Wilcoxon rank-sum test), especially in the stroma, T.act, and abT cell lineages. These results suggest that instead of learning about completely new motifs bpAI-TAC improves the prediction of the effect sizes for signals in the sequence.

We hypothesized that the improved effect size predictions were learned from footprints in the base-pair profiles. To investigate this, we performed *in silico* footprinting with our model (**Supp. Methods**), setting the bias to zero during inference to record the residual from the models' predictions, which explains what the model has learned about TF binding profiles [17]. Indeed, we found that the three motifs produced noticeable footprints compared to AI-TAC (**Figure 3d**), suggesting that the profiles help the model to learn about the precise impact of motifs, consistent with the results of our ablation study (**Figure 2b,c**). On the other hand, AI-TAC does not return any meaningful footprints, but instead increases the noise at the position of these motifs, suggesting that it learns to associate these motifs with signals in the data but not in a meaningful way that would contribute to the predictions in the sequence's contexts.

Discussion

We showed that modeling ATAC-seq data at bp-resolution not only improves interpretability of the models' predictions [12,17,18], but also our ability to accurately model differential activity across highly related cell types. bpAI-TAC builds on modeling choices of (Chrom)BPnet [16,17], but focuses on modeling the data in comparative multi-task fashion, which facilitates training and improves comparisons of differential activity across cell types. Moreover, we observed that the choice of loss function and model architecture can have substantial impacts on model performance. On the other hand, the choice of the bias model had only a minor impact on the model's ability to extract additional information from the ATAC-seq footprints. Further, we observed that models without bias correction were still able to learn additional motif grammar from profiles, resulting in improved performance. We hypothesize that bias correction is not necessarily important in our modeling framework since we separate predictions for profiles and accessibilities. This allows us to derive attributions for accessibility separately, and therefore, by design, ignore sequence motifs that primarily influence the distribution of the data, such as the Tn5 sequence bias, when making accessibility predictions (**Figure S7**). However, we note that this observed bias-model invariance may not hold across all data sets.

When we investigated differences in the learned sequence grammar to AI-TAC, we observed that AI-TAC systematically overlooked specific motifs in certain contexts. However, when we inserted these motifs into random sequences, we were able to identify activity of these motifs for AI-TAC, but with differing effect sizes across cell types, suggesting that bpAI-TAC is learning a more precise estimate of these motifs' effects with respect to sequence context and cell types. The predicted residual profiles (Tn5 bias removed) for individual sequences were not as well aligned with the location of the motifs as in ChromBPnet [17]. Also of note, our deep footprinting was less clear compared to their results (**Figure 3d**). While the variation of the profile around the location of the motif becomes minimal, for some motifs, we cannot observe a clear footprint with a width of several bases. Instead, we observe spikes within the footprints, similar to some motifs in ChromBPnet of weakly binding TFs, or examples in which a suboptimal bias model was used. It is unclear if this is the result of our multi-task modeling approach on data with low number of reads (2.5-40 million reads, 15 million on average), the protein-free or aggregated data that we trained the bias model on, or another modeling choice that we are unaware of.

In summary, we developed and assessed a multi-task convolutional modeling framework that models ATAC-seq data at bp-resolution. In concordance with past research[16,17,26,27], we observe significant improvements in model interpretation when we introduce genomic data at a base-pair resolution. While other approaches have focused on the results of deep footprinting, we show that the dual design for total accessibility and accessibility profiles can also be used to effectively improve prediction performance by harnessing base-pair resolution information. Further, our study found that multi-tasking significantly improves model predictions, and that simple modeling choices, such as the loss function, are important for improved predictions and interpretations.

Methods

Data processing

Models were trained using ATAC-seq data from 90 different mouse immune cell types collected by the Immunological Genome Project [6]. As input, we used one-hot encoded sequences of length 1000bp centered around ATAC-seq peaks, from the mm10 mouse genome. Open chromatin regions were selected by [6]. Roughly, 2 to 181 samples were grouped based on hierarchical clustering with various cut-offs to estimate the peak summits for all cell lineages, which resulted in 518,845 ATAC-seq OCRs. We divided chromatin regions into training, validation, and test sets, leaving out chromosomes 12 and 15 for validation and 11 and 16 for evaluating final model performance. The training dataset included 267,237 chromatin regions, validation set included 28,329 regions, and the test set included 32,361 regions. Cumulative ATAC-seq counts were calculated by summing the number of raw Tn5 insertions in the 250bp region surrounding ATAC-seq peaks and then quantile normalizing the sums to account for different sequencing depths of 2.5-40 million. The ATAC profiles were unaffected by this normalization.

BpAI-TAC model

BpAI-TAC is composed of a body of convolutional layers and two output heads, one that predicts the sum of Tn5 insertions (accessibility head), and one that predicts the ATAC-seq profile (profile head). The body of the model starts with a convolution of width 25, followed by 9 dilated convolutions of width 3 with residual skip connections wherein dilations start at a size of 2 and double each layer. All convolutions in the body of the model have 300 filters and are followed by ReLU activation functions. The model then branches into two separate heads. The profile head starts with a convolution of width 25 and 90 filters (one for each cell type). We then add the predicted Tn5 preference for the associated input sequence, and apply a softmax function to produce a probability distribution of a Tn5 insertion at each base. The scalar head consists of three repetitions of max pooling with width 5 followed by a convolution of width 3 and 300 filters and a ReLU activation function. To produce a prediction of the total number of Tn5 insertions in the input chromatin region in each of the 90 cell types, the scalar head ends with a fully connected layer with an output of 90 dimensions. To focus model learning directly on the OCR, only the center 250bp of the predicted and actual profiles and regional counts are used to compute the model loss.

Bias-model training

We trained three bias models with the bpAI-TAC architecture, using only the profile head loss to learn the base-pair resolution Tn5 bias distribution. The models were trained on: 1) protein-free DNA from [13], and 2) on cumulative profiles from closed OCRs. The third bias model used the *shallow* bpAI-TAC architecture and trained on protein-free DNA profiles to test if model architecture influences bias learning. Protein free DNA was generated for 25 selected chromatin regions based on overlap with a manually selected set of key transcription factors and differentiation related genes [13]. The DNA of these regions was tagmented, processed and sequenced similar to the SHARE-seq ATAC-seq experiment with five replicates and pooled for sequencing.

Since there is no protein-free DNA in ATAC-seq experiments to learn about the Tn5 sequence bias, we used relatively closed OCRs as a proxy. These regions need to be accessible to the Tn5 enzyme but not occupied by any proteins. We defined closed OCRs in the 90 cell types by the probability from peak calling with MACS2 in each of the 90 cell types ($p>0.25$). Individual closed regions only contain very few reads due to the low read depth of the ATAC-seq data (2.5 - 40 million reads, 15 million on average). To generate more realistic profiles with larger sampling sizes, and to reduce the number of data points for training from this data, we summed the counts of all regions across cell types in which the OCR is closed. In this setting, the number of cuts has lost any of its meaning because it not only depends on the chromatin environment but also on the number of cell types in which the OCR was “closed”. All models trained on the protein-free and the cumulative data were only trained on predicting the profiles correctly, without the scalar head, using JSD as a loss.

Model ablations

Binning ablation

To investigate whether lower resolution accessibility profiles could also improve predictions of chromatin accessibility, we binned the ATAC profiles in windows of 5, 10, and 20 base pairs. In the profile head, we sum across the sequence within bins prior to taking the softmax. Thus, the predicted Tn5 bias is also included in this binning, and can be accounted for with a bias model. This ensures fair comparison with the no-binning condition.

Architecture ablations

To investigate how modeling choices influence the performance of the dual-headed multi-task model, we investigated changes in the base pair embedding body of the model, and the two prediction heads. First, we reduced the number of dilated convolutional layers in the model body from 9 to 4 (shallow bpAI-TAC) and in another modification removed the dilated convolutions entirely (base bpAI-TAC). For the accessibility head, we replaced the repeated convolutional layers andpoolings with a single fully connected layer (FCAccHead bpAI-TAC), a single global average pooling strategy (GlobPoolAccHead bpAI-TAC) as was used in (Chrom)BPnet, and a large window max pooling of 50bp (Max50AccHead bpAI-TAC). The other modeling choices stay the same. Please note that GlobPoolAccHead is very similar to BPnet since we modeled bpAI-TAC after BPnet’s modeling choices but added additional convolutional layers to improve accessibility head predictions.

Loss ablations

The first two loss functions that we tested measured error in the per-base pair Tn5 count prediction directly. These predictions were explicitly computed by multiplying the predicted profile with the predicted total counts (the accessibility) from the two bpAI-TAC output heads and then compared to the measured count data using the loss function. We used Poisson loss and means-squared error for this direct evaluation.

We also tested composite losses that contain a mixture parameter lambda which represents the fractional weight of the profile loss as compared to the total accessibility objective:

$$(1 - \lambda) \times MSE(log(\widehat{y}) - log(y)) + \lambda \times Loss_{Profile}(\widehat{e}, e)$$

When lambda=1 the model is trained only on the profiles, while when lambda=0 the model is solely trained on the total accessibility counts, i.e. AI-TAC.

The third loss we evaluated was a composite loss of the MSE of logged counts and the multinomial negative log likelihood of the profiles. Avsec et al. 2021 [16] showed that this composite loss approximates the Poisson loss:

$$- log(Poisson(k^{obs}, k^{pred})) \approx - \lambda \times log(p_{mult}(k^{obs}|p^{pred}, n^{obs})) + (log(n^{obs}) - log(n^{pred}))^2$$

$$\text{With } p_{mult}(k^{obs}|p^{pred}, n^{obs}) = \frac{n^{obs}!}{k_1!...k_d!} p_1^{\ k_1} ... \, p_d^{\ k_d}$$

The final loss that we evaluated was a composite loss of the MSE of logged counts and the cross entropy loss of the profiles. We did this because if we use the above relationship and substitute in p_mult, we can also see that this composite loss represents a weighting of the cross entropy loss (CRE) of each profile with the total counts of that region:

$$log(p_{mult}(k^{obs}|p^{pred}, n^{obs})) = log(\frac{n^{obs}!}{k_1^{\ obs}!...k_d^{\ obs}!}) + \sum_{i=1}^{d} k_i^{\ obs} \cdot log(p_i^{\ pred})$$

$$= const.^{obs} + \sum_{i=1}^{d} n^{obs} \frac{k_i^{\ obs}}{n^{obs}} \cdot log(p_i^{\ pred}) = const.^{obs} + n^{obs} \sum_{i=1}^{d} p_i^{\ obs} \cdot log(p_i^{\ pred})$$

$$= const.^{obs} + n^{obs} CRE(p_i^{\ obs}, p_i^{\ pred})$$

Sequence attribution and motif analysis

We used the following criteria to determine sequences that were consistently well and better predicted in bpAI-TAC over AI-TAC across ten initializations:

1. Strong signal: $\max_{\text{Cell type}}$(accessibility) > 150
2. High variability across cell types: coefficient of variation > 1
3. Good prediction with bpAI-TAC: average Pearson correlation R > 0.5
4. Consistently better than AI-TAC: all ten initializations $R_{\text{bpAI-TAC}}$ > all ten $R_{\text{AI-TAC}}$

We also looked at this for R_bpAITAC > R_AITAC and found a single region that met this criteria. We used DeepSHAP (https://github.com/jmschrei/tangermeme) to compute the attributions to the 146 sequences for the 10 cell lineages by summing the predictions for cell types in each lineage before applying DeepSHAP. We extracted seqlets with motifs if the attribution of the base that is present in the sequence was above 1.96 (equivalent to p-value < 0.05, two-tailed T-test) of the standard deviation of attributions across all sequences, and the motif consisted of at least four significant bases with single base-pair gaps between significant positions. Seqlets were extracted from the attributions of bpAI-TAC and AI-TAC across all 10 lineages and then clustered with agglomerative clustering with complete linkage, using the p-values of the strongest correlation of aligned motifs as a distance metric; this avoids different meaning of the correlation coefficient for motif pairs of different lengths. Clusters were assigned to motifs that shared at least a 0.05 p-value for their correlation with each other.

After assigning clusters to motifs, we determined motif clusters that were present significantly more often in bpAI-TAC than AI-TAC using Fisher's exact test, counting motifs in all lineages. We used Tomtom [28] with the Jaspar non-redundant vertebrate database [29] to assign TF names to motif clusters with a q-value < 0.05.

Deep footprints and motif marginalizations

We used the combined motifs from the alignment of seqlets for marginalization analysis. We introduced the extracted and clustered motifs into 1000 randomly dinucleotide shuffled OCR sequences from the test dataset and compared the predicted profiles and predictions from the bpAI-TAC model with predictions from AI-TAC. Specifically, we created deep footprints with the model's profile predictions after setting the bias of the model to zero [17]. This strategy returns the part of the profile that is unrelated to the Tn5 bias. We plotted the predicted profiles' median and the 25% and 75% percentile and compared the profiles against the median and percentiles from randomly shuffled sequences without the motif (**Figure 3d**).

Data and code availability

This paper analyzes existing, publicly available data from Yoshida et al. 2019 DOI: 10.1016/j.cell.2018.12.036. The GEO accession number for the ATACseq data reported in this paper is GSE100738. Processed ATAC-seq data and called peaks can be found at: https://sharehost.hms.harvard.edu/immgen/ImmGenATAC18_AllOCRsInfo.csv. All original code has been deposited and is publicly available at https://github.com/nuriachandra/bpAITAC as of the date of publication.

Acknowledgments

This research was in part made possible by NIH-R24 (5R24AI072073-18), the Washington Research Foundation and the Carl-Zeiss-Stiftung. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

Conceptualization: N.A.C., A.S., S.M., Y.H., J.B. methodology: N.A.C., A.S., S.M. ; data curation: A.S., Y.H., J.B. software: N.A.C., A.S.; investigation: N.A.C, A.S.; formal analysis: N.A.C., A.S.; visualization: N.A.C. A.S.; validation: N.A.C.; writing: N.A.C., A.S., S.M.; supervision: A.S. and S.M.

Declaration of interests

The authors declare no competing interest.

Figures

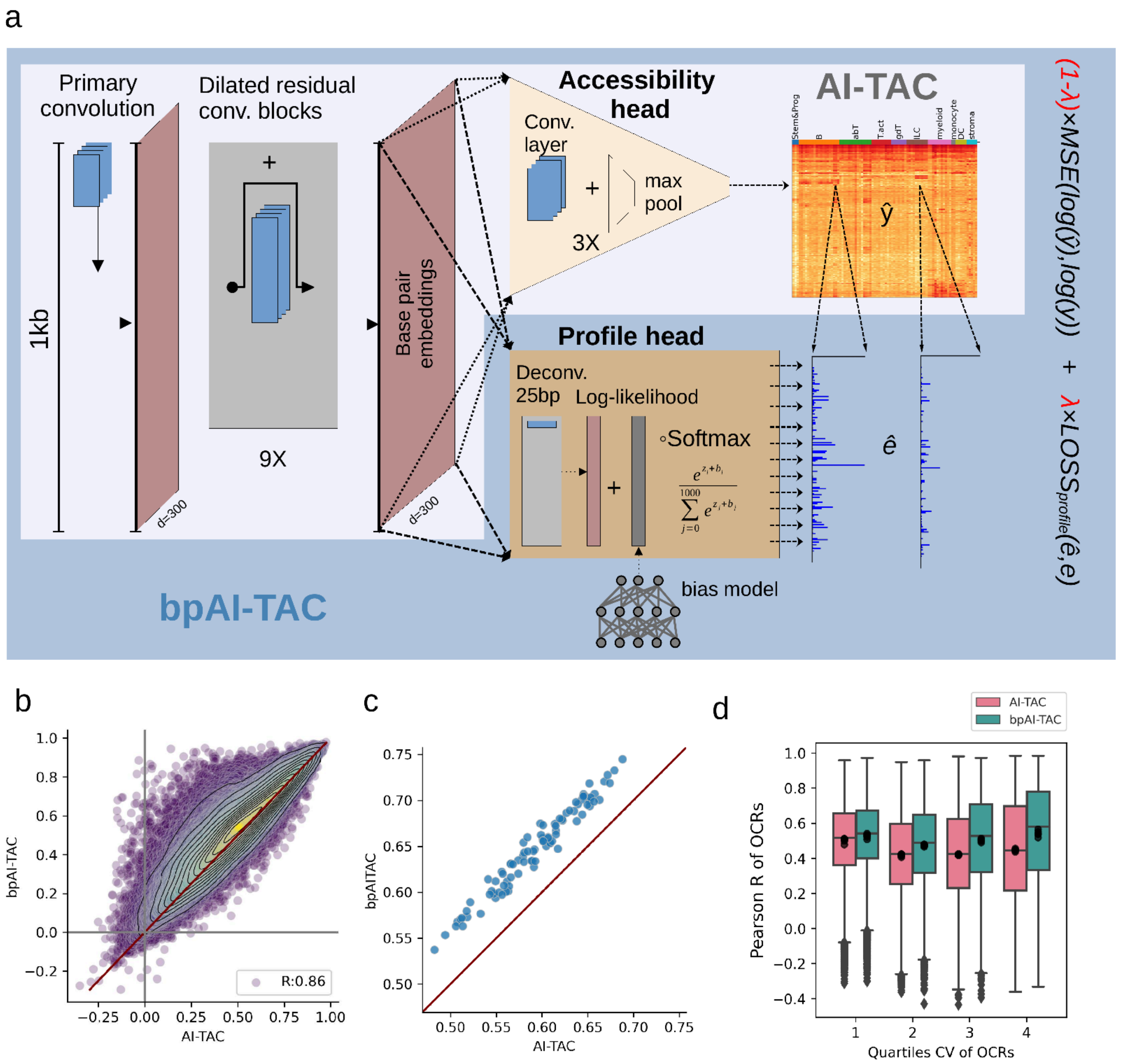

Figure 1. BpAI-TAC improves performance with dual learning of accessibility and base-pair resolution profiles.

a) BpAI-TAC is a multi-task convolutional neural network trained to predict the number of Tn5 insertions at bp-resolution across 90 cell types. It uses a dual head architecture that predicts chromatin accessibility (total number of insertions in center 250bp) and profiles of Tn5 insertion sites, including a Tn5 profile bias. The composite loss function used to train this dual objective is on the right of the figure. **b)** Pearson correlation R of held-out OCRs across cell types for bpAI-TAC and AI-TAC accessibility predictions. **c)** Pearson correlation R of cell types across held-out OCRs**. d)** Boxplots showing the distribution of Pearson correlation R of OCRs across cell types separated into quartiles defined by their coefficient of variation across cell types.

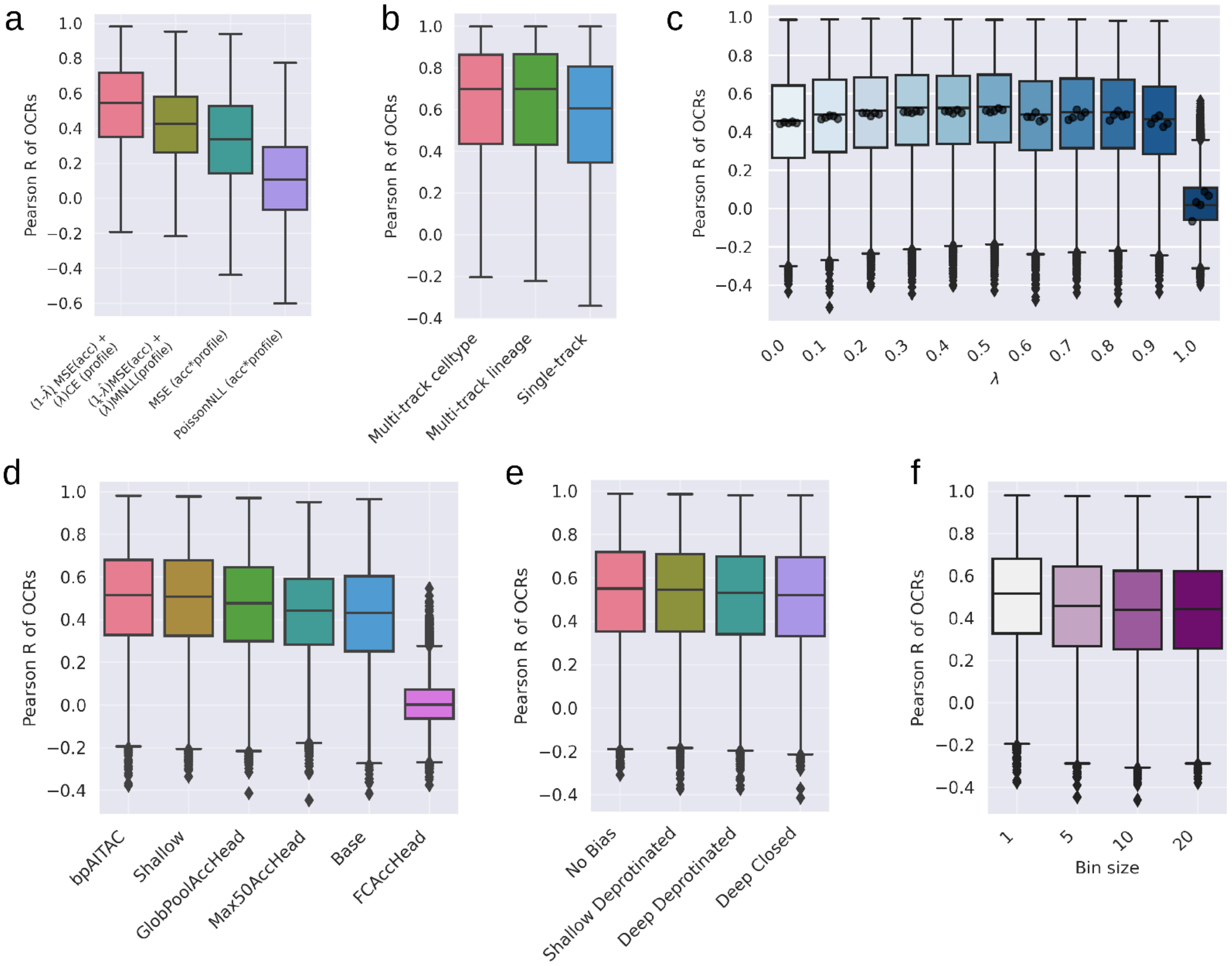

Figure 2. Tn5 profiles improve open chromatin predictions consistently across bpAI-TAC modeling choices.

a) Pearson R of OCRs across cell types for four loss functions. **b)** Pearson R of OCRs across cell lineages. The Multi-Track Celltype model is trained to predict accessibility in 90 different cell types, with predictions averaged into 10 different lineages. The Multi-Track Lineage model is trained to predict accessibility across 10 different lineages. The Single-Track model is an ensemble of the results from 10 individual models trained on individual lineages. **c)** Pearson R of OCRs across cell types across different fractional weights (lambda) on the profile loss (here Cross Entropy). **d)** BpAI-TAC architecture ablations, including reducing the size of the body (Shallow and Base), and modifying the accessibility head (GlobPoolAccHead, Max50AccHead, FCAccHead). **e)** Pearson R of OCRs across cell types for bpAI-TAC trained using four different Tn5 bias prediction approaches. From left to right these approaches are 1) no Tn5 bias prediction added, 2) a shallow CNN trained on protein-free DNA, 3) a deep CNN trained on protein-free DNA, and 4) a deep CNN trained on closed OCR regions. **f)** Pearson R of OCRs across cell types for bpAI-TAC trained on lower resolution profiles. Profiles were binned (summed over regions of sizes 5, 10, and 20) to create lower resolution Tn5 profiles.

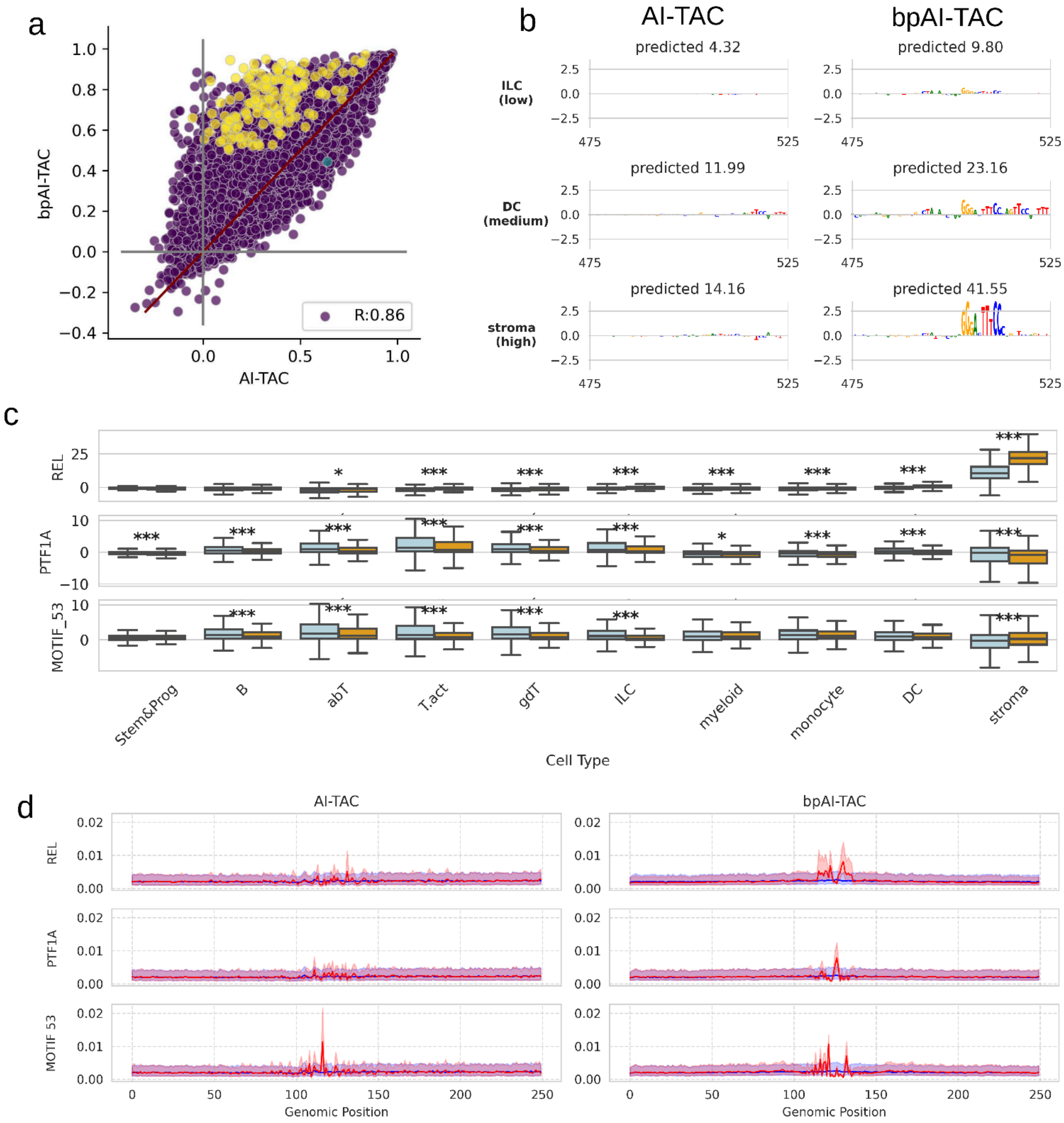

Figure 3. BpAI-TAC learns additional motifs that drive immune cell differentiation from TF footprints.

a) Comparison of Pearson correlation R of OCRs across cell types with selected regions for subsequent analysis in yellow. Regions were selected based on strong signal, superior prediction by one of the models, and high variation across cell types (**Supp. Methods**). **b)** An example of sequence attributions from AI-TAC and bpAI-TAC for one of the selected yellow regions in **a**. Attributions across three different cell lineages are shown, increasing in experimentally measured accessibility from top to bottom. **c)** The difference in accessibility predictions on 1000 dinucleotide shuffled sequences with and without motifs inserted. The difference in bpAI-TAC's predictions are displayed in orange, and the difference in AI-TAC's predictions are in blue. **d)** AI-TAC and bpAI-TAC Tn5 profile predictions on 1000 random sequences with motifs inserted (red), and controls without any motifs inserted (blue). The red and blue lines correspond to the median predicted values, and the translucent bands correspond to the interquartile range.

References

1. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The Human Transcription Factors. Cell. 2018;175:598–9.

2. Lee TI, Young RA. Transcriptional regulation and its misregulation in disease. Cell. 2013;152:1237–51.

3. Sahu B, Hartonen T, Pihlajamaa P, Wei B, Dave K, Zhu F, et al. Sequence determinants of human gene regulatory elements. Nat Genet. 2022;54:283–94.

4. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.

5. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr Protoc Mol Biol. 2015;109:21.29.1–21.29.9.

6. Yoshida H, Lareau CA, Ramirez RN, Rose SA, Maier B, Wroblewska A, et al. The cis-Regulatory Atlas of the Mouse Immune System. Cell. 2019;176:897–912.e20.

7. Yan F, Powell DR, Curtis DJ, Wong NC. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. Genome Biol. 2020;21:22.

8. Li Z, Gao E, Zhou J, Han W, Xu X, Gao X. Applications of deep learning in understanding gene regulation. Cell Rep Methods. 2023;3:100384.

9. Yuan H, Kelley DR. scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural networks. Nat Methods. 2022;19:1088–96.

10. Maslova A, Ramirez RN, Ma K, Schmutz H, Wang C, Fox C, et al. Deep learning of immune cell differentiation. Proc Natl Acad Sci U S A. 2020;117:25655–66.

11. Wang SK, Nair S, Li R, Kraft K, Pampari A, Patel A, et al. Single-cell multiome of the human retina and deep learning nominate causal variants in complex eye diseases. Cell Genom. 2022;2. Available from: http://dx.doi.org/10.1016/j.xgen.2022.100164

12. Nair S, Ameen M, Sundaram L, Pampari A, Schreiber J, Balsubramani A, et al. Transcription factor stoichiometry, motif affinity and syntax regulate single-cell chromatin dynamics during fibroblast reprogramming to pluripotency. bioRxiv. 2023; Available from: http://dx.doi.org/10.1101/2023.10.04.560808

13. Hu Y, Ma S, Kartha VK, Duarte FM, Horlbeck M, Zhang R, et al. Single-cell multi-scale footprinting reveals the modular organization of DNA regulatory elements. bioRxiv. 2023; Available from: http://dx.doi.org/10.1101/2023.03.28.533945

14. Bentsen M, Goymann P, Schultheis H, Klee K, Petrova A, Wiegandt R, et al. ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. Nat Commun. 2020;11:4267.

15. Schultheis H, Bentsen M, Heger V, Looso M. Uncovering uncharacterized binding of transcription factors from ATAC-seq footprinting data. Sci Rep. 2024;14:9275.

16. Avsec Ž, Weilert M, Shrikumar A, Krueger S, Alexandari A, Dalal K, et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. Nat Genet. 2021;53:354–66.

17. Pampari A, Shcherbina A, Kvon E, Kosicki M, Nair S, Kundu S, 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. 2024. p. 2024.12.25.630221. Available from: https://www.biorxiv.org/content/10.1101/2024.12.25.630221v1.abstract

18. Brennan KJ, Weilert M, Krueger S, Pampari A, Liu H-Y, Yang AWH, et al. Chromatin accessibility in the Drosophila embryo is determined by transcription factor pioneering and enhancer activation. Dev Cell. 2023;58:1898–916.e9.

19. Shrikumar A, Greenside P, Kundaje A. Learning Important Features Through Propagating Activation Differences. In: Precup D, Teh YW, editors. Proceedings of the 34th International Conference on Machine Learning. PMLR; 06--11 Aug 2017. p. 3145–53.

20. Sasse A, Chikina M, Mostafavi S. Quick and effective approximation of in silico saturation mutagenesis experiments with first-order Taylor expansion. bioRxiv. 2023; Available from: https://www.biorxiv.org/content/10.1101/2023.11.10.566588.abstract

21. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al. Determination and inference of eukaryotic transcription factor sequence specificity. Cell. 2014;158:1431–43.

22. Jolma A, Yin Y, Nitta KR, Dave K, Popov A, Taipale M, et al. DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature. 2016;534:S15–6.

23. Ghosh S, May MJ, Kopp EB. NF-kappa B and Rel proteins: evolutionarily conserved mediators of immune responses. Annu Rev Immunol. 1998;16:225–60.

24. Liu T, Zhang L, Joo D, Sun S-C. NF-κB signaling in inflammation. Signal Transduct Target Ther. 2017;2. Available from: https://doi.org/10.1038/sigtrans.2017.23

25. Brady OA, Martina JA, Puertollano R. Emerging roles for TFEB in the immune response and inflammation. Autophagy. 2018;14:181–9.

26. Dudnyk K, Cai D, Shi C, Xu J, Zhou J. Sequence basis of transcription initiation in the human genome. Science. 2024;384:eadj0116.

27. Horlacher M, Wagner N, Moyon L, Kuret K, Goedert N, Salvatore M, et al. Towards in silico CLIP-seq: predicting protein-RNA interaction via sequence-to-signal learning. Genome Biol. 2023;24:180.

28. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. Genome Biol. 2007;8:R24.

29. Rauluseviciute I, Riudavets-Puig R, Blanc-Mathieu R, Castro-Mondragon JA, Ferenc K, Kumar V, et al. JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2024;52:D174–82.

Supplementary Figures

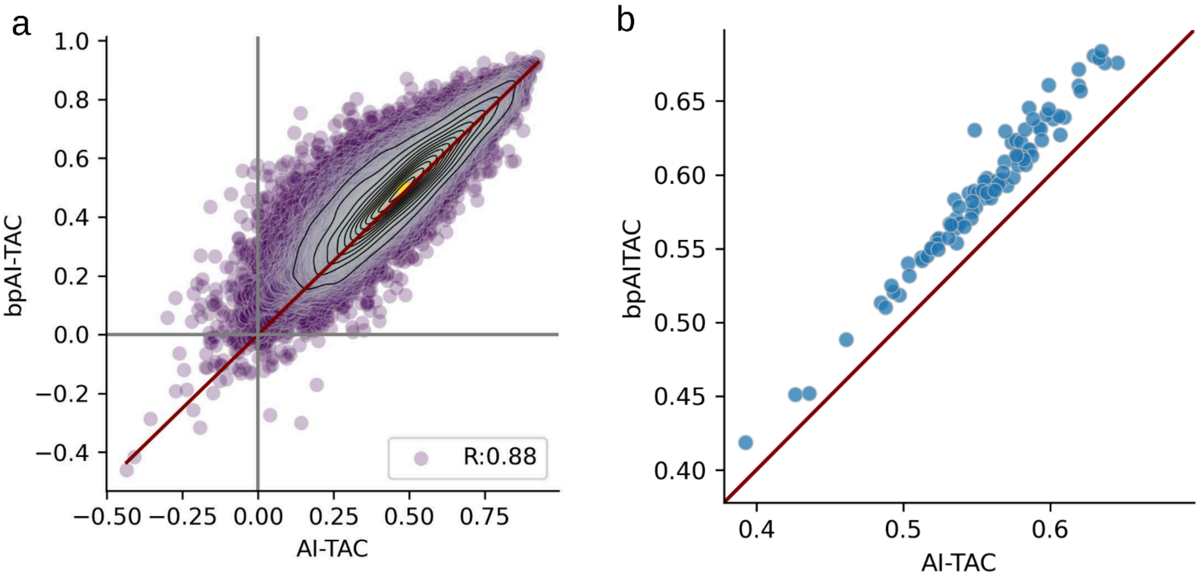

Figure S1.
a) Spearman correlation of OCRs across cell types for bpAI-TAC and AI-TAC accessibility prediction.
c) Spearman correlation of cell types across OCRs.

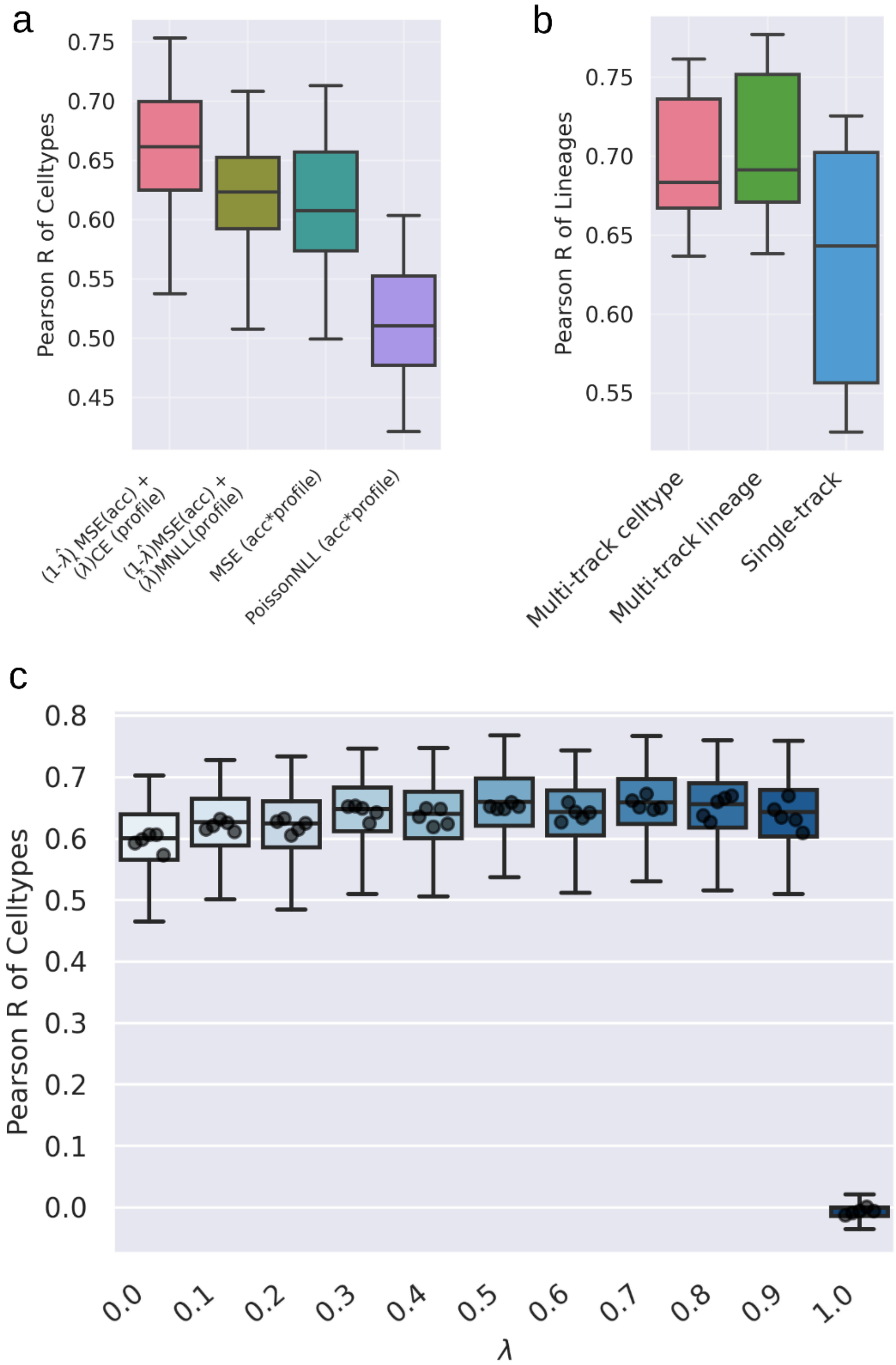

Figure S2.

a) Pearson R of cell types across OCRs for four loss functions. **b)** Pearson R of cell lineages across OCRs. The Multi-Track Celltype model is trained to predict accessibility in 90 different cell types, with predictions averaged into 10 different lineages. The Multi-Track Lineage model is trained to predict accessibility across 10 different lineages. The Single-Track model is an ensemble of the results from 10 individual models trained on individual lineages. **c)** Pearson R of cell types across OCRs for different fractional weights (lambda) on the profile loss (here Cross Entropy).

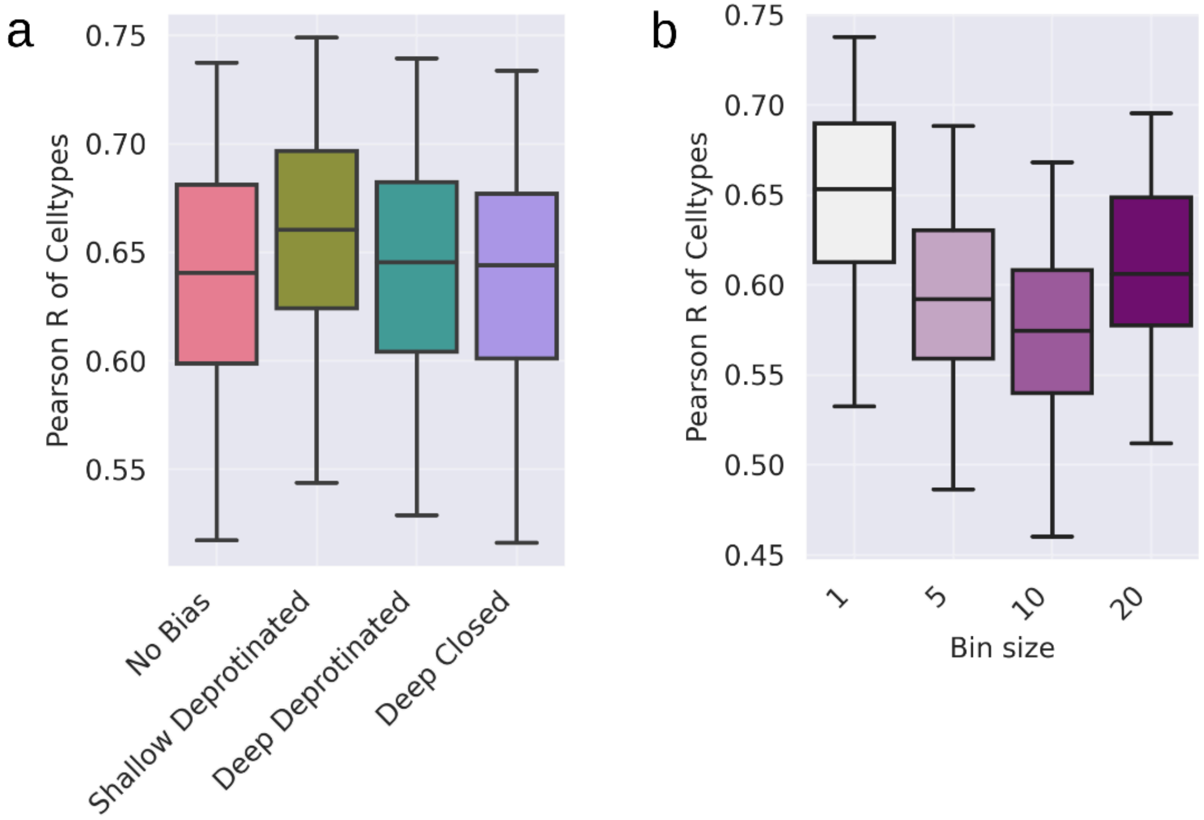

Figure S3.
a) Pearson R of cell types across OCRs for bpAI-TAC trained using four different Tn5 bias prediction approaches. From left to right these approaches are 1) no Tn5 bias prediction added, 2) a shallow CNN trained on protein-free DNA, 3) a deep CNN trained on protein-free DNA, and 4) a deep CNN trained on closed OCR regions. **b)** Pearson R of cell types across OCRs for bpAI-TAC trained on lower resolution profiles. Profiles were binned (summed over regions of sizes 5, 10, and 20) to create lower resolution Tn5 profiles.

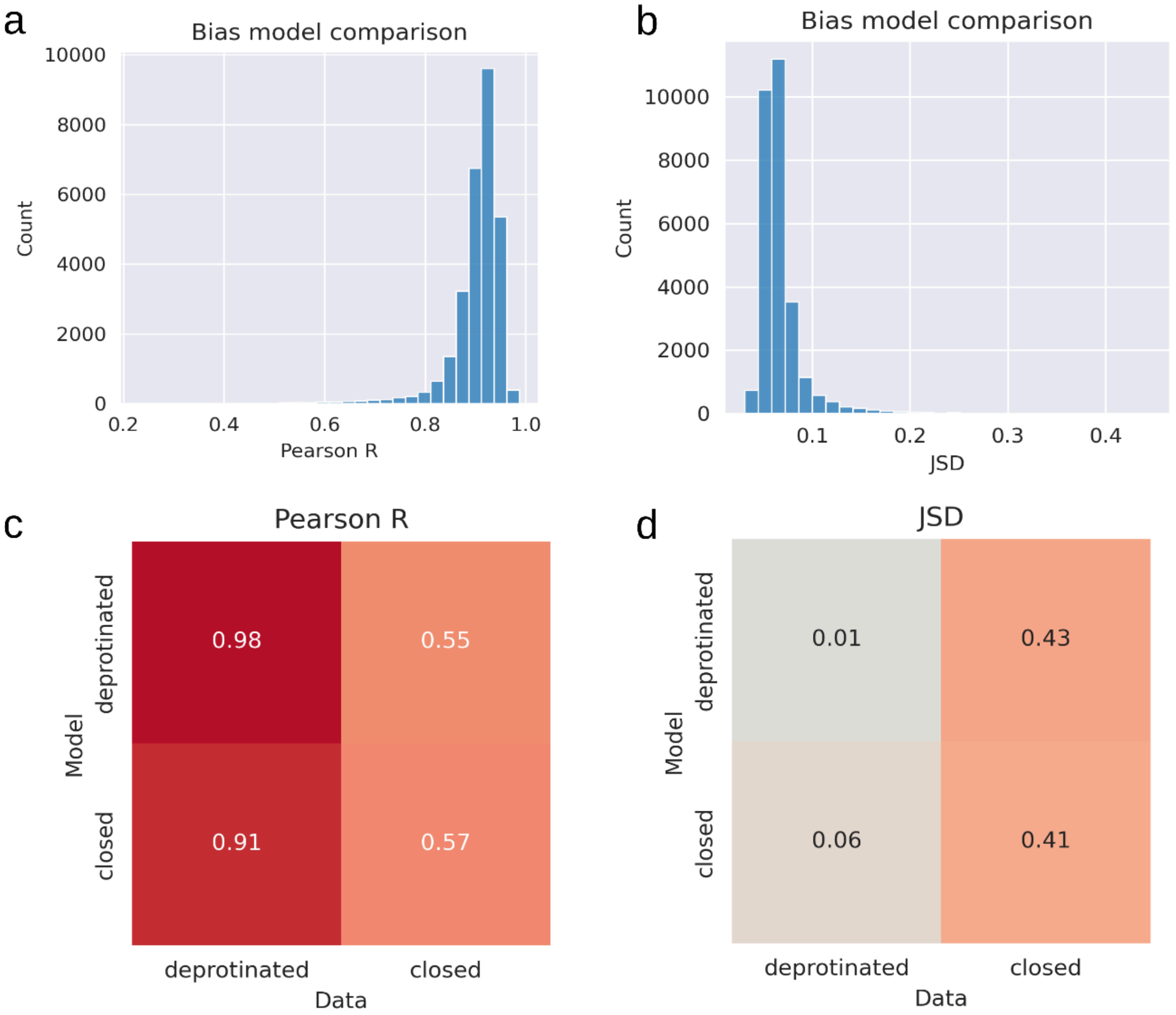

Figure S4. Comparison of Tn5 bias prediction models trained on aggregated closed DNA regions and protein-free DNA.

All analyses are performed on regions of DNA from held-out chromosomes. **a)** Mean Pearson R across profile predictions from two models trained on protein-free and aggregated closed regions on held-out test regions of closed regions of DNA. **b)** Jensen-Shannon divergence (JSD) between protein-free and closed bias model predictions on held-out test regions of closed regions of DNA. **c)** Pearson R of observed and predicted closed and protein-free DNA for both types of bias model. **d)** JSD of observed and predicted closed and protein-free DNA for both types of bias model.

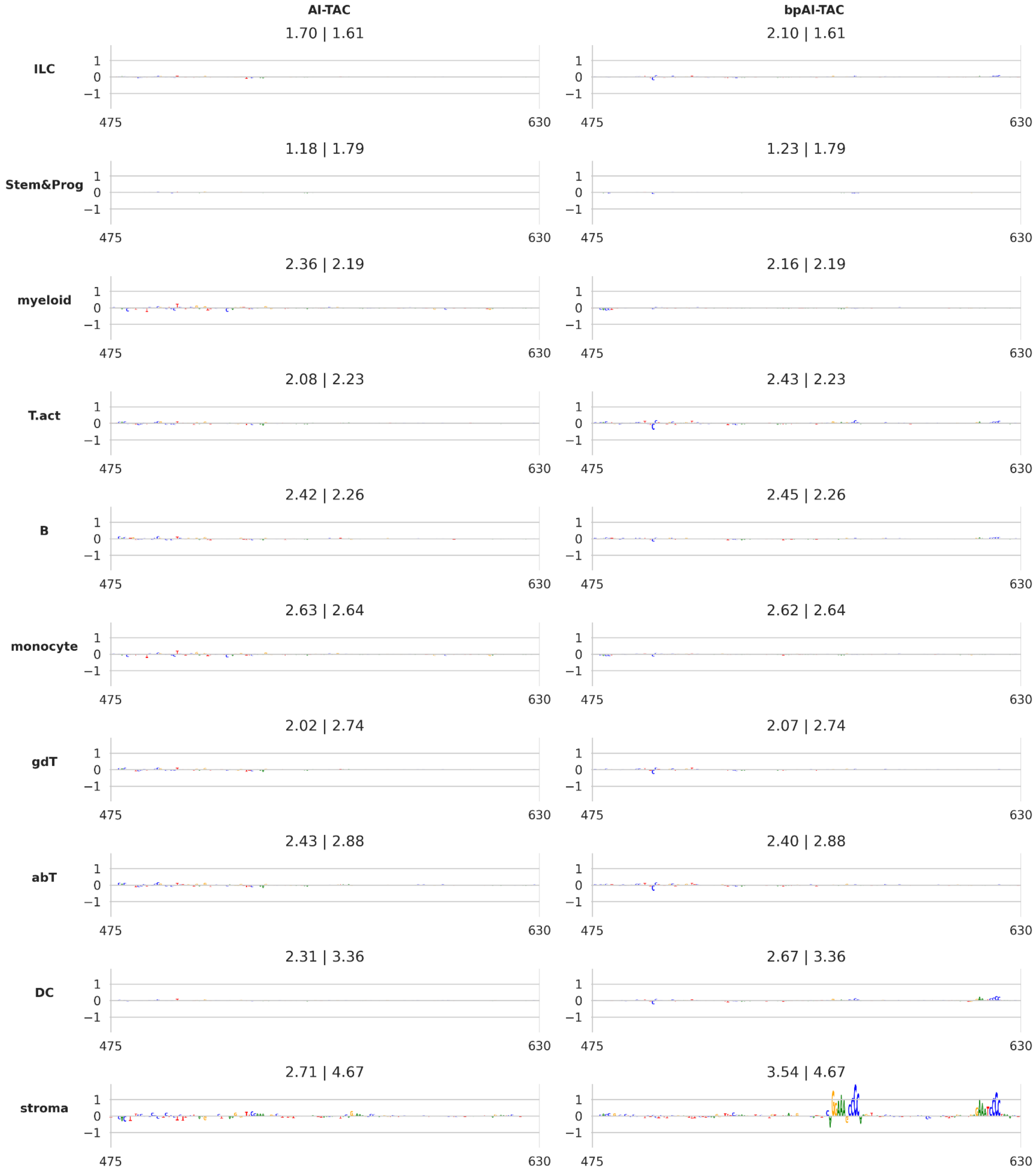

Figure S5. Example model sequence attributions across all cell lineages. The titles of each logo graph indicate predicted | observed accessibility.

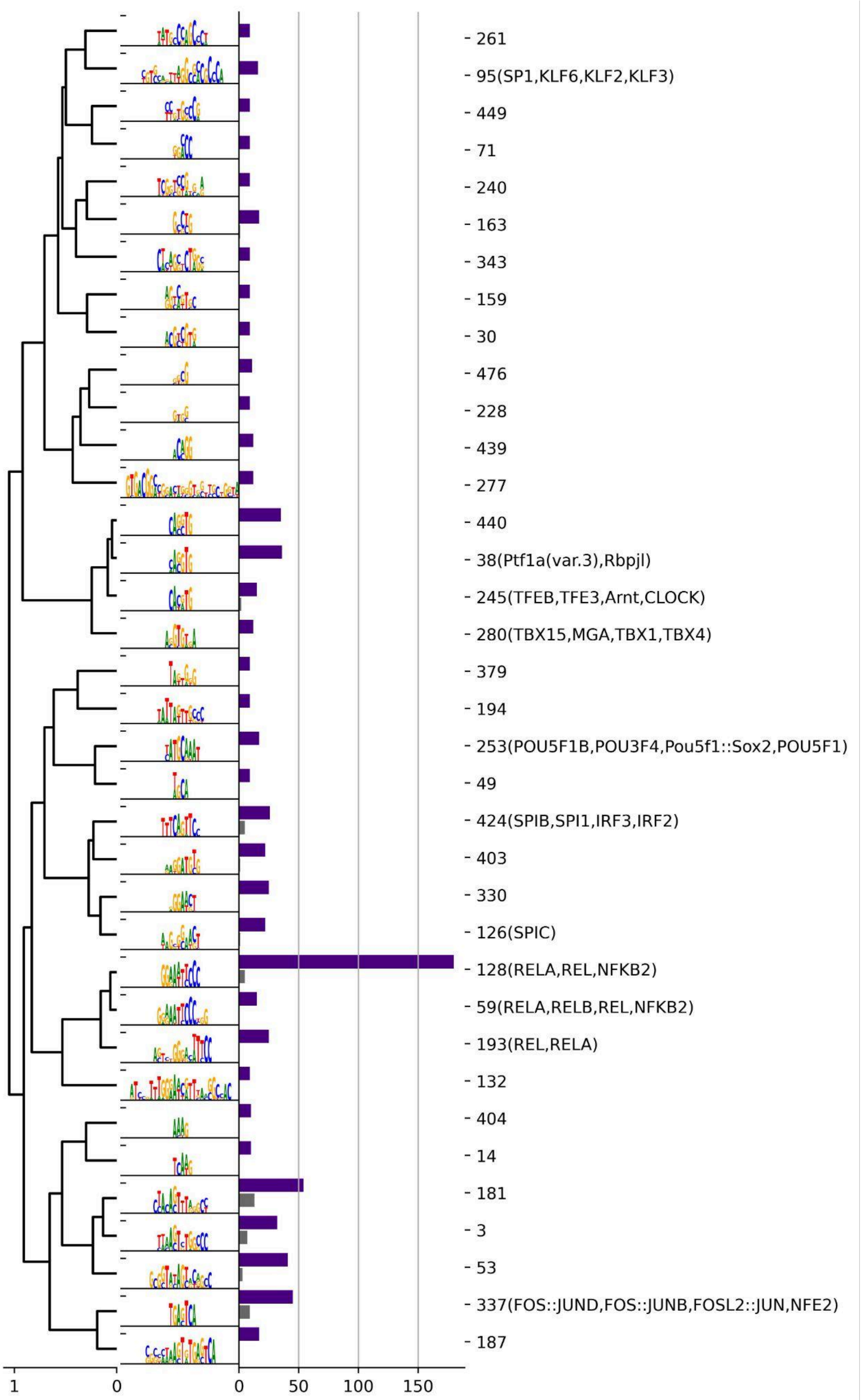

Figure S6. Combined motifs from seqlet clusters sorted on a tree with agglomerative clustering and average linkage on their correlation distance. Seqlets were extracted from sequence attributions across cell lineages and clustered (Methods). Bar plots behind motifs show the number of times the

motif cluster was found in attributions of lineages from AI-TAC (grey bars), and attributions from bpAI-TAC (purple). Cluster numbers are shown on the y-axis with associated transcription factor names in brackets (TomTom, q-value < 0.05, Jaspar non-redundant vertebrate database).

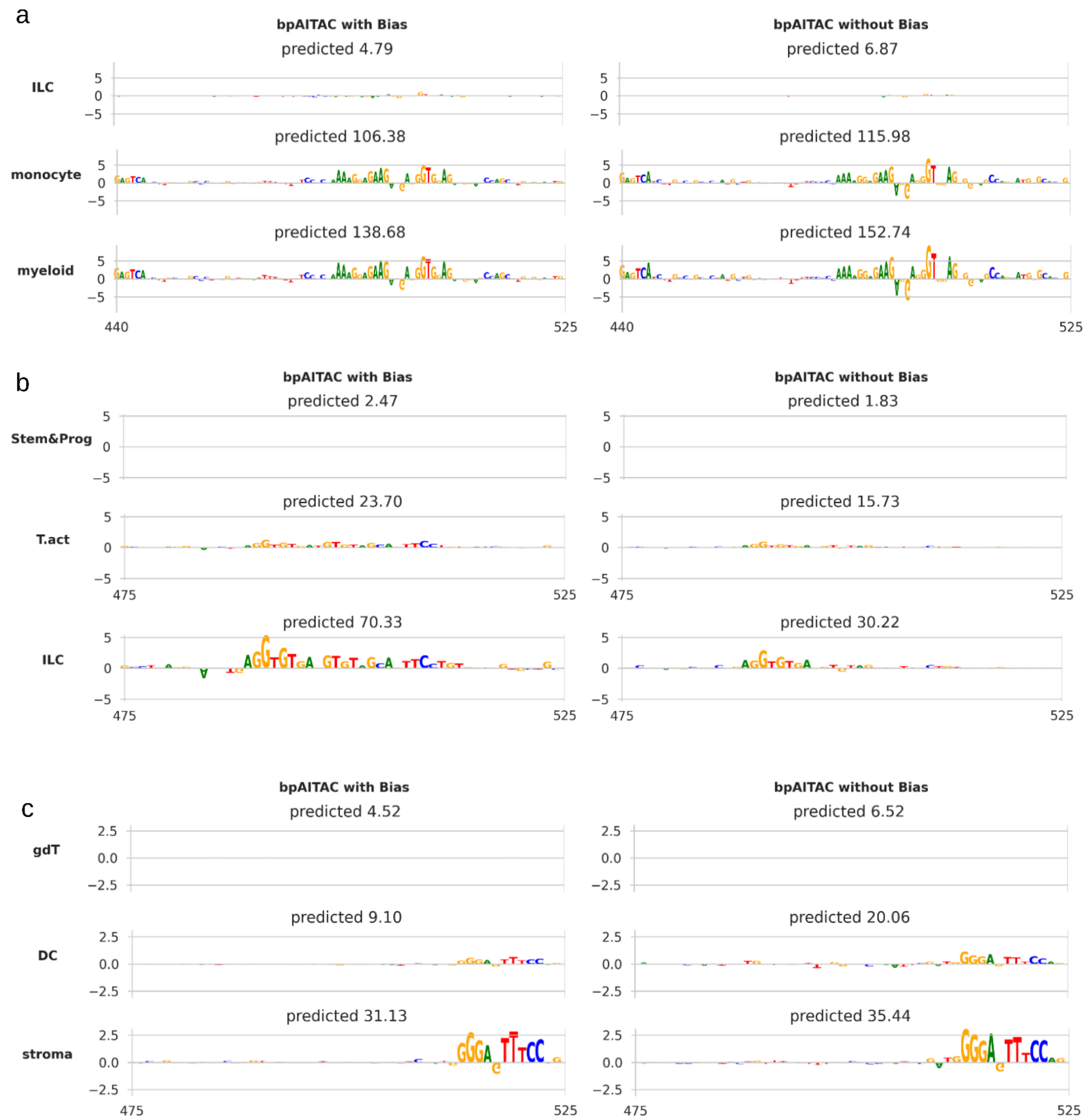

Figure S7. Tn5 bias correction shows no major effect on bpAI-TAC's attribution maps for aggregated chromatin accessibility.

Examples of bpAI-TACs attributions for chromatin accessibility of three regions, trained with and without the bias model. Attributions for the accessibility only show what contributes to the total number of Tn5 insertions, not the sequence elements that contribute to the shape of the profile.