# Characterizing the sequence and prior chromatin determinants of induced TF binding with bimodal neural networks

Divyanshi Srivastava[1], Begüm Aydin[2], Esteban O. Mazzoni[2], Shaun Mahony[1*]

[1]Center for Eukaryotic Gene Regulation, Department of Biochemistry & Molecular Biology, Pennsylvania State University, University Park, PA

[2]Department of Biology, New York University, New York, NY

* Corresponding author, mahony@psu.edu

## Abstract

Transcription factor (TF) binding specificity is determined via a complex interplay between the TF's DNA binding preference and cell-type specific chromatin environments. The chromatin features that correlate with TF binding in a given cell-type have been well characterized. For instance, the binding sites for a majority of TFs display concurrent chromatin accessibility. However, concurrent chromatin features reflect the binding activities of the TF itself, and thus provide limited insight into how genome-wide TF binding patterns became established in the first place. To understand the determinants of TF binding specificity, we therefore need to examine how newly activated TFs interact with sequence and preexisting chromatin landscapes to determine their binding sites.

Here, we investigate the sequence and preexisting chromatin determinants of TF binding by examining genome-wide binding of TFs that have been induced in well-characterized chromatin environments. We develop a bimodal neural network that jointly models sequence and prior chromatin data to interpret the genome-wide binding patterns of induced TFs. We find that the preexisting chromatin landscape is a differential global determinant of TF binding; incorporating prior chromatin features substantially improves our ability to explain the binding specificity of some TFs, but not others. Furthermore, by analyzing the per-site predictors of TF binding, we show that TF binding in previously inaccessible chromatin tends to correspond to the presence of more favorable cognate DNA sequences. Our model thus provides a framework for modeling, interpreting, and visualizing the joint sequence and chromatin landscapes that determine *in vivo* TF binding dynamics.

## Introduction

Sequence-specific transcription factors (TFs) interact with the genome by binding their cognate DNA sequence motifs, using both direct base interactions and DNA structural feature recognition[1–3]. However, the presence of cognate motif instances alone is a poor predictor of TF binding[4,5]. Most TFs bind to only a small fraction of their potential target motif instances in a given cell type, and the cohort of sites which are bound can vary greatly across cell types[6–8]. These observations suggest that TF binding specificity is constrained by cell-specific chromatin landscapes[7,9,10]. For example, cell-specific nucleosome organization and stability can enable or prevent some TFs' access to DNA[5,11,12], whereas certain so-called "pioneer" TFs may be able to override such constraints[12,13]. Cooperative and antagonistic interactions with other regulatory proteins may also constrain cell-specific TF binding[14,15]. However, it remains unclear how DNA sequence, chromatin structure, and interactions with other regulators act in concert to determine cell-type specific binding across a range of TFs.

Computational models of genome-wide TF occupancy are often developed with the goal of gaining insight into cell-type specific TF binding mechanisms. Several methods integrate DNA sequence with information about the chromatin landscape in which the TF is binding (i.e. "concurrent" chromatin information) to characterize genome-wide TF binding specificity[16–18]. However, TFs and their recruited regulatory complexes often alter local chromatin landscapes upon binding to DNA[19,20]. Therefore, concurrent chromatin landscapes are not determinants of TF binding but rather parallel measurements of TF binding itself. Models that integrate DNA sequence and concurrent chromatin information can thus only provide limited insights into how a TF's DNA-binding occupancy became established in the first place.

In order to understand the chromatin determinants of *in vivo* TF binding specificity, we must examine chromatin landscapes that exist in a given cell type prior to TF expression, and then characterize which sites become bound by the TF upon induction. Here, we develop a principled framework to jointly model TF binding as a function of DNA sequence and the preexisting chromatin environment. Specifically, we model genome-wide TF binding through multi-modal deep neural networks that can learn separate representations for the heterogeneous sequence and preexisting chromatin data type modalities, while integrating these distinct representations with readily interpretable deeper layers[21]. Modeling TF binding as a function of both DNA sequence and prior chromatin enables us to estimate the relative contribution of the preexisting cell-type specific chromatin landscape to an induced TF's binding specificity, and allows us to ask whether these contributions differ across TFs.

First, we demonstrate our approach by examining the binding determinants of the pro-neural bHLH TF Ascl1 when it is over-expressed in mouse embryonic stem (mES) cells[22]. Specifically, we characterize the degree to which genome-wide Ascl1 binding depends on the prior mES cell chromatin landscape. We further use our network to examine the DNA sequence and prior chromatin determinants of Ascl1 binding at individual sites, demonstrating that Ascl1 binding occurs across a continuum of sequence and prior chromatin constraints. Second, we expand our analysis to examine the differential sequence and prior chromatin drivers across 12 TFs induced in cell-types in which the prior chromatin accessibility landscape has been characterized (mES cells and NIH-3T3 fibroblast cells)[22,23]. While we focus here on systems in which TF expression is induced in cell lines, our methods are broadly applicable to study any dynamic regulatory system in which chromatin accessibility landscapes can be assayed before TF binding activity occurs.

## Results

### A bimodal neural network integrates DNA sequence and prior chromatin accessibility to predict TF-DNA binding

To estimate the dependence of TF binding on the preexisting chromatin accessibility landscape, we use a stepwise forward classification approach. Specifically, we first train a neural network, $M_S$, to predict TF binding using DNA sequence features alone. We then assess whether an expanded network architecture that incorporates sequence and chromatin features, $M_{SC}$, leads to an improvement in predictive accuracy. Any such improvement points to predictive information in the preexisting chromatin landscape that is not captured by sequence alone.

To model the sequence specificity driving TF binding, we use convolutional neural networks due to their ability to outperform both PWMs and *k*-mer based string kernels at TF binding prediction tasks[24,25]. Specifically, our sequence-only network $M_S$ uses a convolutional layer followed by a long short-term memory (LSTM) layer and multiple dense layers (Fig. 1a). While convolutional filters identify short discriminative PWMs at bound sequences, LSTMs and deeper layers are capable of integrating information from convolutional filters to model higher-order sequence dependencies[18,24,26]. We note that in order to test if preexisting chromatin landscapes drive TF binding specificity, we must ensure that our sequence-only model does not learn sequence features associated with the preexisting chromatin accessibility landscape itself. We prevent such features from being learned using careful design of the training mini-batches (see Methods).

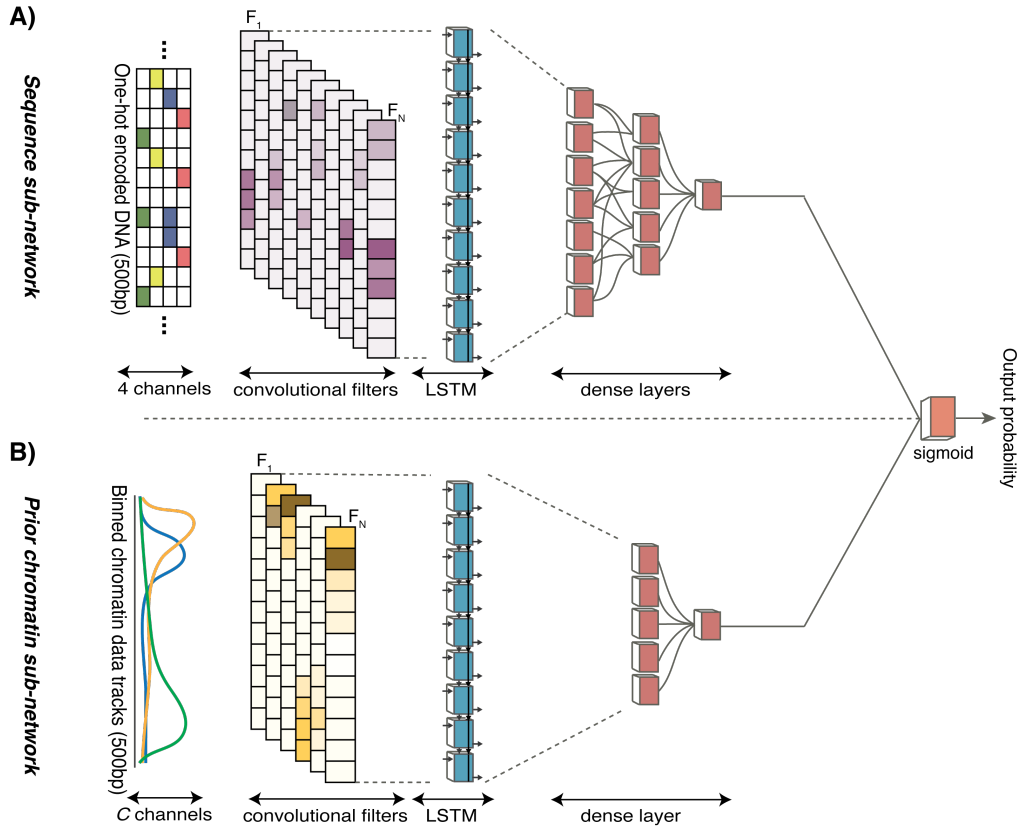



**Figure 1:** Bimodal neural network architecture containing **A)** a sub-network that trains on sequence features and **B)** a sub-network that trains on prior chromatin features. Sub-network activations are combined using a single sigmoid-activated node in order to aid interpretability of sequence and prior chromatin binding pre-determinants.

Methods that model multiple modalities in TF binding predictions tasks often use early-fusion; i.e. they integrate the modalities into a single input vector[17,25]. However, low-level correlations between heterogeneous sequence and chromatin accessibility inputs may not always be meaningful or interpretable. To incorporate preexisting chromatin features into our predictive framework, we define a bimodal network architecture that models the sequence and chromatin accessibility through independent sub-networks combined with an additive sigmoidal dense node (Fig. 1b). DNA sequences from 500bp windows are used as input to the sequence sub-network, whereas binned ATAC-seq and histone mark ChIP-seq data are used as input to the chromatin sub-network (see Methods).

To test our framework, we focused on characterizing where the bHLH TF Ascl1 binds when expressed in mES cells. Specifically, we trained the networks to predict Ascl1 ChIP-seq data, assayed after 12 hours of ectopic Ascl1 expression in mES cells. We incorporated publicly available ATAC-seq

as well as H3K27ac, H3K27me3, H3K4me1/me2/me3, H2A.Z, acH2A.Z, H3K9ac, H3K36me3 and H3K9me3 ChIP-seq data from mES cells as inputs into the chromatin sub-network (see Methods). Due to the imbalanced nature of the problem, we use the genome-wide area under the precision-recall curve (auPRC) as a performance metric for both the sequence-only and the bimodal sequence-chromatin networks.

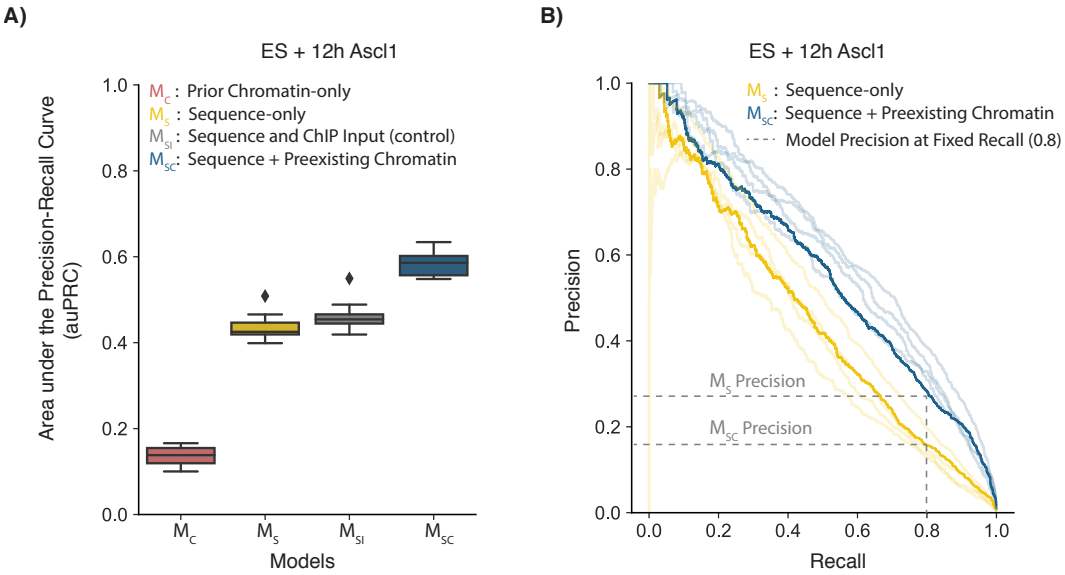



**Figure 2: A)** The distribution of neural network classification performance (auPRC) on held-out chromosomes 10-14 for the prior chromatin-only ($M_C$), sequence-only ($M_S$), sequence and ChIP input control ($M_{SI}$) and the sequence and prior ES chromatin ($M_{SC}$) models. **B)** The precision-recall curves for the sequence-only model compared to the precision-recall curves for the sequence and prior ES chromatin models. Model performance for held-out chromosome 10 is highlighted in solid yellow and blue lines for illustration; performance for chromosomes 11-14 is represented with lighter (alpha=0.2) traces.

We find that the bimodal sequence and preexisting chromatin model $M_{SC}$ outperforms the sequence-only model $M_S$ when trained on induced Ascl1 ChIP-seq data (Fig. 2a). While the median auPRC across test chromosomes for the sequence model $M_S$ is 0.42, the median auPRC for the joint sequence and preexisting chromatin model $M_{sc}$ is 0.59 (Fig. 2a). The improved performance of the $M_{sc}$ model is driven mostly by improved specificity. At a fixed false positive rate (FPR=0.01), a large majority of Ascl1-bound sites are correctly predicted by both models (Suppl. Fig. 1a). However, at a fixed recall of 80%, the number of false positives predicted by the joint sequence and chromatin network is less than half the number of false positives predicted by the sequence-only network (Fig. 2b). As a negative control, a joint sequence and chromatin model trained using a sequenced input control experiment instead of pre-induction chromatin data leads to only a marginal improvement in

performance over the sequence-only model (auPRC=0.45; Fig. 2a). We also confirmed that an additive bimodal design does not perform worse than a model with more complex interactions between sequence and prior chromatin features (Suppl. Fig. 1b). Incorporating prior chromatin features therefore leads to an increase in specificity of induced Ascl1 binding predictions, suggesting that Ascl1 binding is somewhat dependent on the preexisting chromatin landscape.

**A range of sequence and prior chromatin constraints determine induced Ascl1 binding at individual sites in mouse embryonic stem cells**

Our bimodal network design has a distinct advantage in that it integrates outputs from the sequence and chromatin sub-networks in an easily interpretable additive fashion. Specifically, let $M_{SC}$ be the model that incorporates both sequence and preexisting chromatin features. The sequence and chromatin sub-networks can be thought of as transforming the input feature vectors $x_s$ and $x_c$ into a network embedded space. If $\phi_s(x_s)$ and $\phi_c(x_c)$ represent transformed feature vectors, then the network output node $y_{sc}$ acts an additive logistic model that estimates the coefficients for the embedded sequence and chromatin feature vectors.

$$logit(y_{sc}) = \beta_0 + \beta_1\phi_s(x_s) + \beta_2\phi_c(x_c)$$

The weighted activations $S_j^i$ for each sub-network can then be interpreted as scores assigned by sub-network $j$ to genomic locus $i$:

$$S_j^i = \beta_j\phi_j(x_j^i)$$

Here, $\beta_j$ is the weight assigned to sub-network $j$, and $\phi_j(x_j^i)$ is the non-linear transformation applied by sub-network $j$ to input feature vector $x_j^i$. The bimodal network thus maps each genomic locus to a two-dimensional space defined by the weighted sequence and chromatin sub-network activations.

We make use of the two-dimensional network embedding to examine the sequence and chromatin sub-network activations at individual Ascl1 binding sites. We embed Ascl1-bound genomic regions (orange) as well as randomly sampled unbound genomic regions into the two-dimensional network transformed space (Fig. 3a). As stated in the previous section, a large majority of Ascl1 binding sites receive high activation scores from the overall $M_{SC}$ network. Interestingly, the high overall scores at Ascl1 binding sites correspond to a broad range of compensatory sequence and chromatin sub-network activations in the two-dimensional embeddings. Ascl1-bound sites with low sequence scores are on average scored highly by the chromatin sub-network. Conversely, bound sites with low chromatin scores are on average scored highly by the sequence sub-network. We quantify

this effect using the median sequence score at bound sites ($S_M$) as a threshold to show that the marginal chromatin-score distributions differ at high-scoring ($S_s^i > S_M$) versus low-scoring ($S_s^i < S_M$) Ascl1 binding sites (Fig. 3b). Thus, the network learns a model in which Ascl1 binds target sequences that exhibit a broad range of sequence and chromatin sub-network scores, and some degree of compensation between sequence and prior chromatin predicts genome-wide Ascl1 binding.

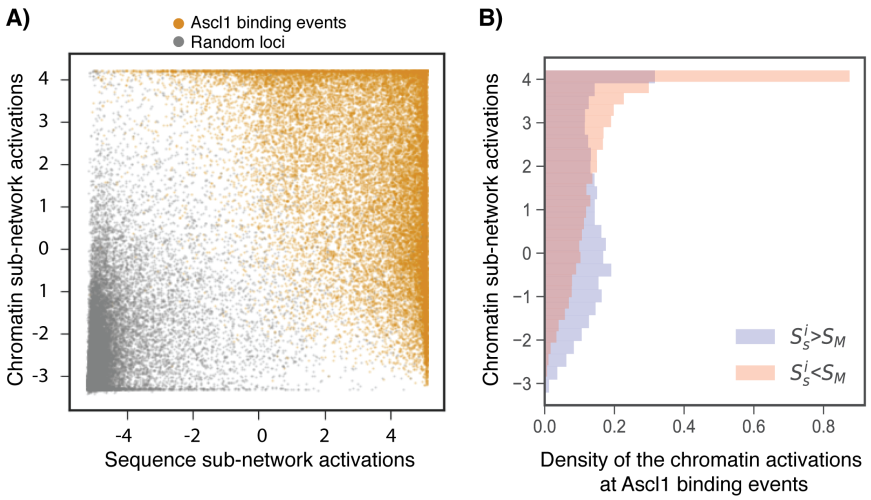



**Figure 3: A)** The contribution of sequence and preexisting chromatin to binding predictions at Ascl1 binding events (orange) and randomly sampled unbound genomic windows (grey). **B)** Distributions of chromatin sub-network scores from **A)** for Ascl1 binding events with sequence scores greater than or less than the median sequence network score.

## Motif multiplicity and motif flanks drive variation in sequence sub-network activations at individual Ascl1 binding sites

Next, we aimed to interpret the trained network to identify the sequence and prior chromatin features that drive the observed variation in sequence and chromatin sub-network activations at *in vivo* TF binding sites. First, we used integrated gradients based feature attribution[27] to confirm that the network learns features associated with Ascl1's cognate E-box binding preference (Suppl. Fig. 2a,b). Interestingly, we noticed that at many loci, multiple Ascl1 binding E-box motifs were assigned high attribution scores (Suppl. Fig. 2), suggesting that motif multiplicity is a predictive feature of Ascl1 binding.

Since gradient-based feature attribution can be susceptible to network parameterizations, we also used an orthogonal strategy to confirm that motif multiplicity defined the high-scoring Ascl1 binding sites. We first calculated the number of Ascl1 cognate E-box motif instances in each 500bp window bound by Ascl1 (Fig. 4a). We divided bound loci into categories based on their motif

multiplicity, and measured the sequence sub-network activations in each category. Only 5% of all Ascl1-bound loci lack exact matches to the core Ascl1 E-box motif CAGSTG, and these loci were assigned the lowest median scores by the sequence sub-network (Fig. 4b). However, a large fraction of Ascl1-bound windows contains more than one exact match to the Ascl1 binding E-box CAGSTG, and the number of motif occurrences is strongly positively correlated with sequence sub-network scores (Fig. 4b). To systematically examine the relationship between motif multiplicity and sequence sub-network scores, we inserted between one and four randomly spaced motif instances in a set of randomly simulated sequences. We found that the sequence sub-network scores increase with increasing motif multiplicity, indicating that the network indeed uses multiplicity as a feature in predicting Ascl1 binding (Fig. 4c).

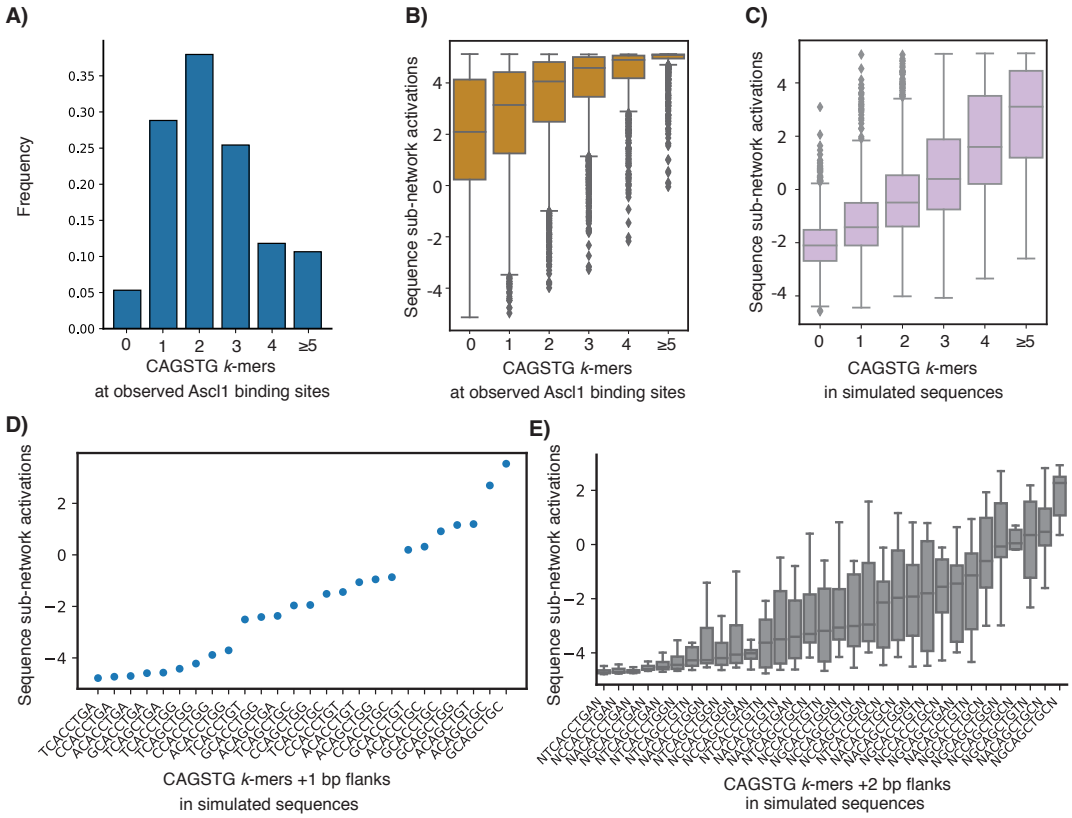



**Figure 4: A)** Frequencies of CAGSTG *k*-mers at all Ascl1 binding sites. **B)** Sequence sub-network scores increase with motif multiplicity at Ascl1 binding windows. **C)** Embedding CAGSTG motifs in simulated sequences confirms that the network uses the number of motif occurrences as predictors of sequence scores. **D) & E)** Embedding the CAGSTG motif with defined 1bp or 2bp flanks shows that the network sequence scores for Ascl1 are driven by motif flanks.

In addition to the primary motif, motif flanking nucleotides have been shown drive genome-wide Ascl1 binding specificity in mES cells[22]. To test whether the nucleotides flanking the Ascl1 motif CAGSTG affect sequence sub-network activations, we used two complementary simulation strategies. First, we constructed a single 500 bp reference sequence in which each position is encoded as a $[0.25,0.25,0.25,0.25]^T$ vector; i.e. each base $[A, T, G, C]^T$ occurs with equal probability at each position along this sequence. We embedded all combinations of the Ascl1 CAGSTG motif flanked by a single nucleotide on either end into this reference sequence, and extracted the sequence sub-network activation at each such simulated sequence. We find large variation in the sequence sub-network output based on the flanking nucleotides, suggesting that the network is learning motif flanking information as a predictor of induced Ascl1 binding (Fig. 4d). While certain 8-mers such as GCAGCTGC are scored highly by the sequence sub-network, others such as ACAGCTGCA lead to low sequence sub-network scores. Since an artificially constructed reference baseline may introduce biases into our estimation of sequence activations, we also embed the Ascl1 motif + 1bp flanks into 10,000 randomly generated 500bp sequences, resulting in scores consistent with *k*-mers embedded into a reference sequence (Suppl. Fig. 3). Embedding Ascl1 motifs + 2bp flanks results in a further large variation in scores for each 8-mer (motif +1bp flanks), suggesting that the network learns information beyond 1bp flanking nucleotides (Fig. 4e). The sequence sub-network thus uses both motif multiplicity and flanking nucleotide information to assign variable sequence scores to *in vivo* Ascl1 binding sites.

**Higher-order chromatin information drives variation in chromatin sub-network activations at individual Ascl1 binding sites**

In order to identify the drivers of prior chromatin sub-network activations, we first calculated the distributions of chromatin sub-network activation scores at enrichment domains for each input histone modification ChIP-seq dataset. The median scores for preexisting domains of chromatin accessibility, active histone marks H3K4me1/2/3, H3K27ac, H3K9ac, and active histone variants H2A.Z and acH2A.Z were positive, suggesting that some degree of induced Ascl1 binding is associated with regions that already displayed signs of regulatory activity in the preexisting pluripotent cell state. Conversely, median scores for preexisting H3K9me3 and H4K20me3 domains were strongly negative (Fig. 5a), suggesting that the chromatin sub-network uses preexisting repressive/heterochromatic histone modifications as negative predictors of induced Ascl1 binding.

To further examine the determinants of variation in chromatin activations at Ascl1-bound sites,

we divided Ascl1 binding sites into quartiles based on their chromatin sub-network activations and calculated their composite tag enrichment profiles for each chromatin track (Fig. 5b). We found that the genomic windows associated with the highest chromatin sub-network activations were enriched for chromatin accessibility, H3K4me2, H3K4me3 and H3K27ac in the prior cell state. Genomic windows in the second and third quartiles were enriched for H3K4me2 along with H3K4me1 and H2A.Z, but show lower chromatin accessibility, H3K9ac, and H3K27ac. The fourth data quartile, associated with the lowest chromatin sub-network activations, lacked significant enrichment for any histone modifications (Fig. 5b).

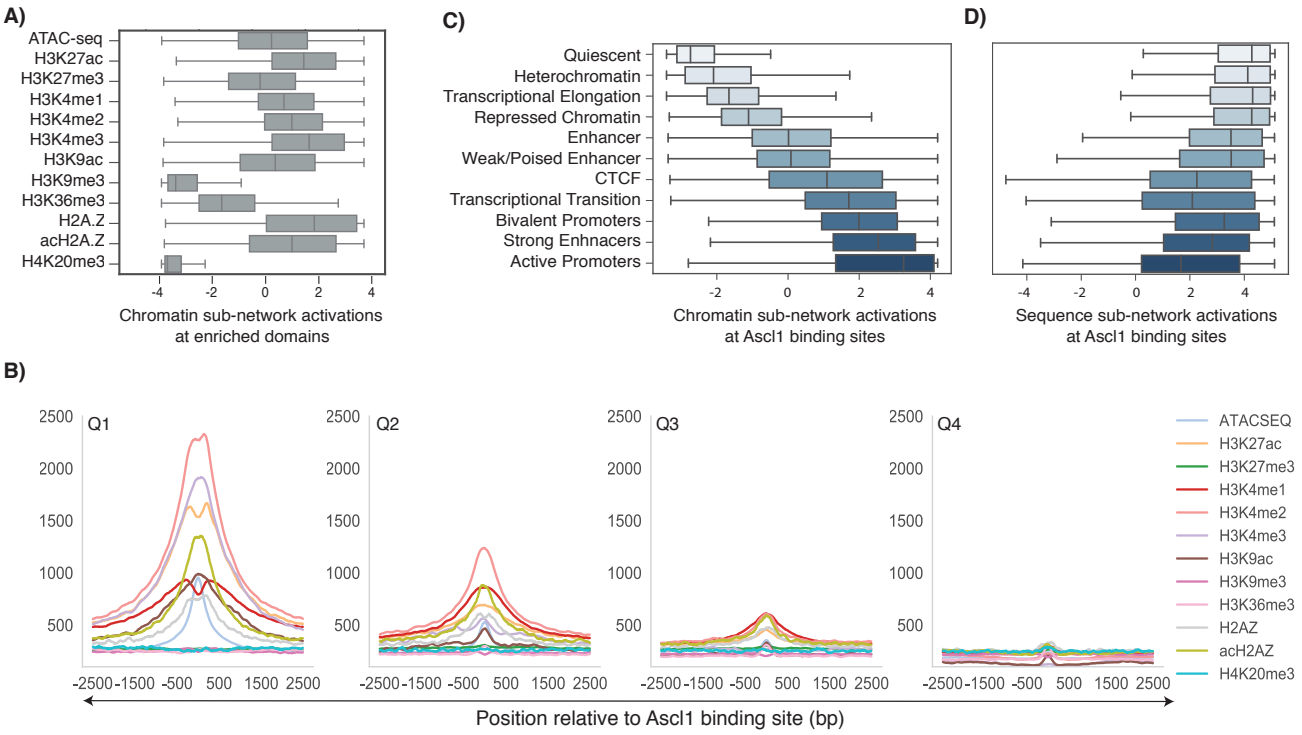



**Figure 5: A)** The distribution of chromatin sub-network activation scores at enrichment domains for each of 12 mES chromatin datasets. **B)** Composite plots of mES chromatin tag enrichment at induced Ascl1 binding sites divided into quartiles based on their associated chromatin sub-network scores. **C)** Chromatin and **D)** sequence sub-network scores vary based on the preexisting chromatin states (annotated with ChromHMM).

To explicitly test for the association of chromatin sub-network activations with particular prior chromatin states, we segmented the genome into twelve states using ChromHMM, and calculated the chromatin sub-network activations at each of these states (Fig. 5c). We found that preexisting active promoters and strong enhancers were assigned the highest median chromatin scores, followed by weak enhancers and bivalent promoters. Consistent with our previous results, Polycomb-repressed chromatin (marked by H3K27me3), heterochromatin (marked by H3K9me3) and quiescent genomic

regions were assigned the lowest median chromatin scores. Further, consistent with the compensatory behavior observed in the network embeddings, sites where Ascl1 becomes bound in regions associated with mES quiescent and repressed states contain on average higher sequence scores than sites in mES active states (Fig. 5d).

## The preexisting chromatin accessibility is a differential determinant of *in vivo* TF binding specificity

Having examined in depth the ability of the network to characterize the sequence and prior chromatin determinants of TF binding for Ascl1, we next applied our method to compare the differential sequence and prior chromatin drivers across a broader range of TFs. We collected publicly available data for TFs that were induced and profiled via ChIP-seq in cell-types with pre-assayed chromatin accessibility landscapes (see Methods). Further, to maintain consistency across TFs, we considered only those TFs that were profiled 12 hours post induction, resulting in a dataset comprising of 12 TFs, induced in either mES cells or NIH-3T3 fibroblasts (Suppl. Table 3).

We first asked whether incorporation of prior chromatin data improves the ability of a sequence-only model to predict genome-wide TF binding. We find that the sequence and prior chromatin bimodal network $M_{SC}$ outperforms the sequence-only network $M_S$ for all 12 TFs analyzed (Fig. 6a). However, we note that the auPRC is susceptible to the imbalance in the data, and thus at a fixed misclassification rate for bound and unbound data, the auPRC will be lower for a TF with fewer binding sites[28]. Thus, as an additional measure for comparison across TFs, we measured the model recall for each TF at a fixed false positive rate (FPR) of 0.01. In addition to reporting the recall, we derived the posterior distribution of the recall for each TF to quantify our degree of belief in our reported recall estimates (see Methods, Fig. 6b)[29]. Binding models for the TFs Dlx6, Bhlhb8, Six6 and Hlf show greater than 2-fold increases in recall at a fixed FPR, suggesting that the binding of these TFs are highly constrained by the prior chromatin accessibility environment. On the other hand, binding models for Ascl1, Neurog2, Cdx2 and Duxbl show a smaller gain in recall at a fixed FPR on incorporation of prior chromatin accessibility data (Fig. 6b). We note that the contribution of prior accessibility to the binding of these TFs is not immediately evident in direct comparisons with prior cell type ATAC-seq data, as these TFs bind their target motifs in both pre-accessible and pre-inaccessible chromatin. As a negative control, we show that using a sequenced input control experiment as input to the chromatin sub-network instead of prior ATAC-seq does not lead to similar gains in model performance for any of the 12 TFs tested. (Suppl. Fig. 4). The differential gain in model

predictive ability on the incorporation of preexisting chromatin accessibility data suggests that the prior chromatin accessibility landscape differentially constrains *in vivo* TF binding, and that our framework can identify this differential contribution.

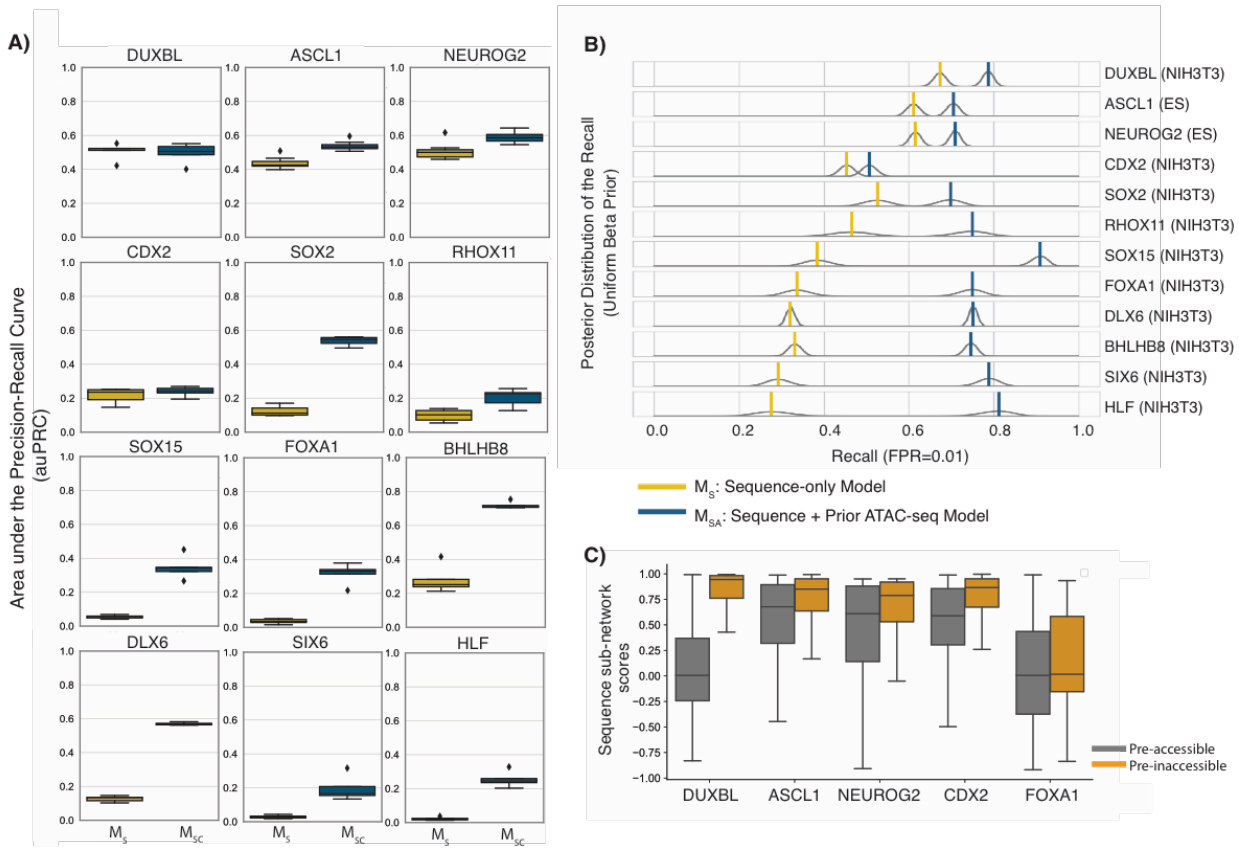



**Figure 6**: **A)** Neural network classification performance evaluated for 12 TFs induced in ES or NIH3T3 cells for sequence-only model $M_S$ (yellow) and bimodal sequence-chromatin $M_{SC}$ (blue) models. Boxplots present the auPRC on sequentially held-out chromosomes chr10-chr15. **B)** Differential gain in the model recall at a fixed False Positive Rate (FPR) for the 12 TFs. **C)** Pre-inaccessible TF-bound windows are scored higher than pre-accessible bound windows by the sequence sub-network for the TFs DUXBL, CDX2, ASCL1, NEUROG2, FOXA1 and BHLHB8. For SOX2, SOX15, DLX6, SIX6 and RHOX11, pre-inaccessible bound windows are assigned lower median scores than pre-accessible bound windows.

Finally, we focused on TFs that can bind target motifs in pre-inaccessible chromatin (TFs for which at least 10% of their binding occurred in pre-inaccessible chromatin), and asked whether the network learns compensatory sequence-chromatin models for these TFs. Specifically, we asked whether the distribution of sequence scores for these TFs was different between pre-accessible versus pre-inaccessible chromatin (Fig. 6c). We found that the network-assigned sequence scores at pre-inaccessible binding regions were consistently higher than sequence scores at pre-accessible

binding regions, compatible with a model that allows certain sequence signatures to override unfavorable chromatin features. In summary, our results show that different degrees of compensation between sequence and prior chromatin landscapes define observed *in vivo* TF binding specificity, with TFs that bind extensively in pre-inaccessible chromatin showing stronger signatures of compensation.

## Discussion

TFs bind subsets of their cognate motif instances in a cell-type specific fashion. Such specificity in TF binding results from an interplay between the TF's inherent sequence preferences and cell-type specific chromatin landscapes[6,30]. The question naturally arises as to which local chromatin features might enable or inhibit a given TF's binding. However, if we can measure a TF's binding occupancy using ChIP-seq, it has by definition already had its own impact on chromatin in that cell type (e.g. by making its binding sites accessible or by recruiting histone modification enzymes). Concurrent chromatin landscapes therefore predict *in vivo* TF binding in the same cell type[31,32], but cannot be used to model the causal determinants of that binding.

Here, we propose an interpretable neural network architecture, which can be used to assess the relative contribution of DNA sequence and preexisting chromatin features in specifying an induced TF's genome-wide binding sites. We demonstrate our approach on ChIP-seq data for 12 TFs that have been ectopically expressed either in mouse ES cells or NIH-3T3 fibroblasts. Our model suggests that TFs are constrained differentially by the preexisting chromatin accessibility landscape. Predictive models for the TFs such as Dlx6 and Bhlhb8 benefit significantly from the addition of prior chromatin data, suggesting that the binding of these TFs is strongly constrained by cell-type specific chromatin landscapes. On the other hand, predictive models for TFs that bind similar numbers of pre-accessible and pre-inaccessible loci *in vivo* (Ascl1, Neurog2, Duxbl and Cdx2) gain more limited information upon the addition of prior accessibility information. We note that our estimates of the dependence of TF binding on prior chromatin may still be cell-type specific as opposed to an innate feature of a given TF. For example, while our analyses suggest that binding sites for the pioneer factor Foxa1 is dependent on prior chromatin, this may be specific to the measured context of NIH-3T3 cells. It is possible, for instance, that TFs that cooperate with or otherwise predict Foxa1 binding are already present in NIH-3T3 cells, and Foxa1 may be less dependent on prior chromatin in other cell types.

Related to our work, previous studies have assessed the effects of prior chromatin landscapes on the binding of specific TFs[30,33–36]. For example, John *et al.* showed that Glucocorticoid Receptor (GR)

preferentially binds pre-accessible chromatin upon hormone induction[33]. Donaghey *et al.* characterized the relationships between the preexisting chromatin landscape and induced FOXA2, GATA4, and OCT4[13], each of which showed different propensities for binding pre-accessible regions. Our work aims to provide a unified predictive framework for quantifying and formalizing the relative contributions of sequence and chromatin pre-determinants to TF binding across a range of TFs.

Interpretation of our binding models at individual binding sites suggests that sequence and preexisting chromatin landscapes are not independent predictors of TF binding. Rather, sequence and preexisting chromatin are mutually compensatory features that define a continuum of sites that may be bound by the induced TF. While genomic loci with weaker sequence signatures may be bound by TFs given a favorable local chromatin environment, the same signatures might not be sufficient to drive TF binding at inaccessible or unfavorable chromatin. For example, Ascl1 is more likely to bind pre-inaccessible loci in the presence of certain sequence features such as high motif multiplicity and favorable motif flanks, suggesting that indirect co-operative binding may be a potential mechanism through which Ascl1 binds nucleosomal chromatin[20,37].

On the other hand, some highly accessible active promoters and enhancers are bound even with weaker sequence signatures, as defined by low activation scores from the sequence sub-network in our model. We note that some TF-bound regions with high prior chromatin sub-network activations and low sequence sub-network activations might represent artifactual ChIP-seq enrichment[38]. Alternatively, these regions may represent direct binding to weaker motifs, or indirect binding mediated by interactions with mES or NIH-3T3 cell regulators[39]. While previous studies have proposed sequence-conditional binding to inaccessible chromatin for a few TFs[30,40,41], our work suggests that this compensatory mechanism may exist across a broader range of TFs.

Finally, different TFs are expected to interact differentially with preexisting chromatin landscapes[42–44], and the same TF may be more or less dependent on prior chromatin in different cell types. It will therefore be of interest to examine how the relative contributions of sequence and prior chromatin vary in determining the binding of a wider range of TFs, and across a wider array of cell types. Identifying such sequence and chromatin predeterminants of TF binding will be crucial for understanding gene regulation in various dynamic systems such as development and cellular programming.

## Methods

**ChIP-seq & ATAC-seq data (ES cells):** Generation of the inducible iAscl1 and iNeurog2 mouse ES cell lines and corresponding ChIP-seq data is more completely described in Aydin, *et al.*[22]. Briefly, inducible cell lines were generated using the inducible cassette exchange (ICE) method as previously described[45]. TF gene constructs are inserted in single copy into the expression-competent HPRT locus. The resulting iAscl1 and iNeurog2 ES cells are differentiated on untreated plates for 2 days to form embryoid bodies, and then expression of the transgene is induced via Doxycycline. Ascl1 and Neurog2 binding was assayed by ChIP-seq 12 hours after Dox induction using the anti-Ascl1 (Abcam, ab74065) and anti-Neurog2 (Santa Cruz, SC-19233) antibodies. We assayed histone modifications as well as chromatin accessibility in EBs with ChIP-seq and ATAC-seq, respectively (Suppl. Data Table 1). We collected additional publicly available histone modification and histone variant ChIP-seq datasets from mouse ES cells (Suppl. Data Table 2). Together, our dataset defining the chromatin environment of mouse pluripotent cells consists of the following 12 data types: ATAC-seq, H2A.Z, acH2A.Z, H3K27ac, H3K27me3, H3K9me3, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H4K20me3 and H3K36me3.

**ChIP-seq & ATAC-seq data (NIH3T3 cells):** ChIP-seq data for TF inductions in mouse NIH-3T3 fibroblasts was retrieved from Raccaud *et al.* (GSE119784)[23]. We filtered for TFs that were not expressed as defined by RNA-seq in the NIH3T3 cell line[23]. We used NCIS to estimate the sequenced control-based normalization factors for each TF ChIP-seq experimnent[46]. Further, we filtered out induced TFs that had a multiGPS-reported signal fraction < 0.01 and were single-replicate ChIP-seq experiments (Suppl. Data Table 3). We used five ATAC-seq experiments (Suppl. Data Table 4) as replicates to construct the network ATAC-seq input[23].

**ChIP-seq & ATAC-seq data processing:** Fastq files were aligned to the mouse genome (version mm10) using Bowtie (1.0.1)[47] with options "-q –best –strata -m 1 –chunkmbs 1024". Only uniquely mapped reads were considered for further analysis. MultiGPS (version 0.74) was used to define transcription factor DNA binding events[48]. A q-value cutoff of 0.01 (assessed using binomial tests and Benjamini-Hochberg multiple hypothesis test correction) was used to call statistically significant binding events with respect to sequenced input material collected from the same cell line. Peak-finding statistics are reported in Suppl. Table 3. Paired-end ATAC-seq reads were aligned using

Bowtie2 (2.2.2) using the "-q –very-sensitive" options[49]. ChromHMM[50] was run using default parameters.

**Training and test set construction:** For testing, we divided the genome into 500bp non-overlapping windows. For training, we use 500bp overlapping windows, each of which are sequentially offset by 50bp. Genomic windows overlapping peak calls with a *p*-value ≤ $5x10^{-5}$ are labeled as bound. Windows overlapping non-significant peaks from MultiGPS are labeled ambiguous. All other genomic windows (~99%) are labeled as unbound. The sequence sub-network $M_{SC}$ takes as input 500bp sequences. Each nucleotide is encoded as a one-hot vector, such that only the index corresponding to the input nucleotides is set to one, and all other indices are set to zero. For each chromatin input data track, we extract the per-base read counts at each genomic locus. These raw coverage counts are binned into ten 50bp discontinuous bins (covering 500bp windows). The binned read counts are total tag normalized for each replicate, and we use the replicate average at each bin as input to our network. The chromatin datasets are stacked, resulting in a 10 x *k* chromatin input, where *k* is the number of assayed histone modifications/chromatin accessibility. For analyses (mES and NIH3T3 induced TFs) using prior ATAC-seq: *k*=1. For analysis of the mES TFs using prior ATAC-seq and other histone modification data: *k*=12.

**Neural network architectures:** In the sequence sub-network, the 500bp, one-hot encoded sequence input is first subjected to a 1-dimensional convolution layer, with each index in the one-hot encoding acting as a channel into this convolution. The convolutional layer consists of 240 x 20bp long filters. The convolutional filters within 15bp intervals are max-pooled, and the pooled convolutional output is used as input into a long short-term memory (LSTM) layer. The LSTM outputs a 32-vector, which then passes through two dense layers, both subjected to ReLU activation and dropout. The activations from the final dense layer are input into a single tanh activated dense node. The chromatin sub-network uses convolutional filters that span two input bins. The filters are followed by an LSTM to model any observable tag densities discriminative of TF binding. The LSTM activations are input into a single dense layer followed by a single tanh activated dense node. The activations of both the sequence and the preexisting chromatin sub-networks are weighted by a final sigmoid activated node, used to output binding probability. The network is trained to predict ChIP-seq by minimizing the binary cross-entropy loss *J*:

$$J = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)$$

Area under the precision-recall curve is used as a metric to measure network performance. Chromosome 17 is held out as a validation chromosome. Chromosomes 10-15 are sequentially held out as test chromosomes in a *k*-fold training procedure.

**Neural network training strategies**: To prevent the sequence-only network from learning accessibility-related sequences, we customize the sampling used to construct mini-batches for gradient descent-based training. We construct training batches such that in each batch, the bound and unbound training instances contain equal proportions of accessible sites, as defined by ATAC-seq data from the prior cell state. This sampling strategy reduces model false positives at preexisting accessible regions, and leads to an improvement in sequence-only model performance measured on held-out chromosomes measured via the area under the precision recall curve (auPRC) (Supp. Fig. 5a, b).

The bimodal network aims to learn both sequence and prior chromatin signatures that characterize genome-wide TF binding. We can thus no longer control for accessibility distributions across bound and unbound training sets to prevent spurious learning of prior chromatin-related sequence signatures when training the bimodal network. To address this problem, we transfer weights from the previously trained sequence-only network $M_S$ to the sequence sub-network in the bimodal network $M_{SC}$. While the lower-level layer sequence sub-network weights are kept fixed during $M_{SC}$ training, the weights for the final dense layer in the sequence sub-network are trained to fit the new data. Since higher-level layers are more likely to learn problem-specific features, re-training the final dense layer allows the network to optimize the genome-wide binding problem without learning sequence motifs related to prior accessibility at the lower-layers[51]. The joint bimodal network $M_{SC}$ can then be trained using imbalanced batches constructed by random sampling unbound data across the genome.

**Feature attribution with integrated gradients:** We use integrated gradients to estimate the relative importance of each nucleotide $(x_i)_{i=1}^{L}$ for each input sequence $\boldsymbol{x}$ of length $L$ bp. Integrated gradients consider how predictions at input feature vectors differ from reference feature vectors. More specifically, integrated gradients calculate the gradients at all points along a straight-line path from the reference feature vector $\boldsymbol{x^b}$ to the input feature. In our case, we define a reference feature

as a sequence vector such that at each position, each nucleotide is equally likely. In other words, our reference sequence is a 4 * 500 matrix, with each column defined as [0.25, 0.25, 0.25, 0.2s5]. We implemented integrated gradients as defined in Sundarajan *et al.*[27].

**The posterior distribution of the model recall:** We used the model recall at a fixed false positive rate (FPR) to compare model performance across TFs. TPs are true positives in the held-out test set, whereas FNs are false negatives in the test set.

$$Recall = \frac{TPs}{TPs + FNs}$$

However, we note that ChIP-seq signal fractions and the number of peaks called vary widely across TFs. Models trained to predict binding for TF ChIP-seq experiments that contain smaller numbers of peaks (and correlated lower signal fractions) suffer from having access to limited training data. In order to quantify our confidence in the model recall, we use a probabilistic framework that models the recall for each TF given the underlying ChIP-seq data. Specifically, analogous to Brodersen *et al.*[29], we consider the observed model recall (measured on a single held-out test chromosome) to be an actualization of an underlying true recall value $r$ given $N$ independent Bernoulli trials, where $N$ is the number of binding sites in the held-out test chromosome. Each binding site can be either labeled a true positive (success) or a false negative (failure) by the network.

$$Recall \sim Binomial(N, r)$$

We derive the posterior distribution of the recall $r$ assuming a beta (parameters $a$=1, $b$=1) prior (for details, see Brodersen *et al.*[29]). The mode of this posterior distribution is the observed model recall. If a TF ChIP-seq experiment contains a small number of peaks, the distribution of $r$ has high variance (e.g. FoxA1, Rhox11, Fig. 6b). On the other hand, a low variance in the distribution of $r$ reflects a high degree of confidence in our estimate of the recall (e.g. Ascl1, Bhlhb8, Fig. 6b).

**Availability:** Open source code (MIT license) is available from https://github.com/seqcode/iTF. ChIP-seq data have been uploaded to GEO under accession GSE114176.

**Acknowledgements:** This manuscript is based upon work supported by the National Science Foundation ABI Innovation Grant No. DBI1564466 (to SM). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. This work was also supported by the Academic Computing Fellowship (to DS),

NIGMS R01GM121613 (to SM), NICHD R01HD079682 (to EOM), and an NVIDIA GPU equipment grant.

## References

1. Bulyk, M. L. Computational prediction of transcription-factor binding site locations. *Genome Biol.* **5,** 201 (2003).
2. Gordân, R. *et al.* Genomic Regions Flanking E-Box Binding Sites Influence DNA Binding Specificity of bHLH Transcription Factors through DNA Shape. *Cell Rep.* **3,** 1093–1104 (2013).
3. Rohs, R. *et al.* The role of DNA shape in protein-DNA recognition. *Nature* **461,** 1248–53 (2009).
4. Wasserman, W. W. & Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* **5,** 276–287 (2004).
5. Slattery, M. *et al.* Absence of a simple code: how transcription factors read the genome. *Trends Biochem. Sci.* **39,** 381–399 (2014).
6. Arvey, A., Agius, P., Noble, W. S. & Leslie, C. Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Res.* **22,** 1723–34 (2012).
7. Wang, J. *et al.* Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* **22,** 1798–812 (2012).
8. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13,** 613–626 (2012).
9. Bai, L. & Morozov, A. V. Gene regulation by nucleosome positioning. *Trends Genet.* **26,** 476–483 (2010).
10. Kaplan, T. *et al.* Quantitative Models of the Mechanisms That Control Genome-Wide Patterns of Transcription Factor Binding during Early Drosophila Development. *PLoS Genet.* **7,** e1001290 (2011).
11. Xin, B. & Rohs, R. Relationship between histone modifications and transcription factor binding is protein family specific. *Genome Res.* **28,** 321–333 (2018).
12. Iwafuchi-Doi, M. *et al.* The Pioneer Transcription Factor FoxA Maintains an Accessible Nucleosome Configuration at Enhancers for Tissue-Specific Gene Activation. *Mol. Cell* **62,** 79–91 (2016).
13. Donaghey, J. *et al.* Genetic determinants and epigenetic effects of pioneer-factor occupancy. *Nat. Genet.* **50,** 250–258 (2018).

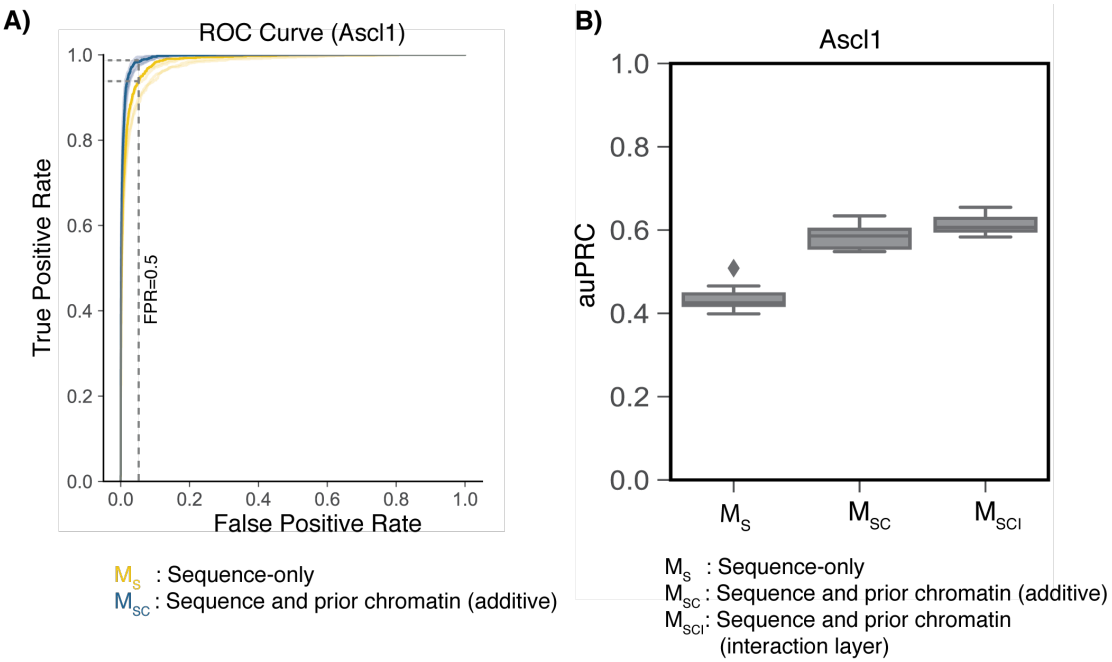
14. Slattery, M. *et al.* Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell* **147,** 1270–82 (2011).
15. Wu, W. *et al.* Dynamic shifts in occupancy by TAL1 are guided by GATA factors and drive large-scale reprogramming of gene expression during hematopoiesis. *Genome Res.* **24,** 1945–62 (2014).
16. Chen, X., Yu, B., Carriero, N., Silva, C. & Bonneau, R. Mocap: large-scale inference of transcription factor binding sites from chromatin accessibility. *Nucleic Acids Res.* **45,** 4315–4329 (2017).
17. Karimzadeh, M. & Hoffman, M. M. Virtual ChIP-seq: Predicting transcription factor binding by learning from the transcriptome. *bioRxiv* 168419 (2018). doi:10.1101/168419
18. Quang, D. & Xie, X. FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. *bioRxiv* 151274 (2017). doi:10.1101/151274
19. Li, B., Carey, M. & Workman, J. L. The role of chromatin during transcription. *Cell* **128,** 707–19 (2007).
20. Adams, C. C. & Workman, J. L. Binding of disparate transcriptional activators to nucleosomal DNA is inherently cooperative. *Mol. Cell. Biol.* **15,** 1405–21 (1995).
21. Ngiam, J. *et al.* Multimodal deep learning. in *Proceedings of the 28th international conference on machine learning (ICML-11)* 689–696 (2011).
22. Aydin, B. *et al.* Proneural factors Ascl1 and Neurog2 contribute to neuronal subtype identities by establishing distinct chromatin landscapes. *Nat. Neurosci.* **22,** 897–908 (2019).
23. Raccaud, M. *et al.* Mitotic chromosome binding predicts transcription factor properties in interphase. *Nat. Commun.* **10,** 487 (2019).
24. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33,** 831–838 (2015).
25. Quang, D. & Xie, X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.* **44,** e107–e107 (2016).
26. Greenside, P., Shimko, T., Fordyce, P. & Kundaje, A. Discovering epistatic feature interactions from neural network models of regulatory DNA sequences. *Bioinformatics* **34,** i629–i637 (2018).
27. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. in *Proceedings of the 34th International Conference on Machine Learning-Volume 70* 3319–3328 (2017).

28. Jeni, L. A., Cohn, J. F. & De La Torre, F. Facing Imbalanced Data Recommendations for the Use of Performance Metrics. *Int. Conf. Affect. Comput. Intell. Interact. Work. [proceedings]. ACII* **2013,** 245–251 (2013).
29. Brodersen, K. H., Ong, C. S., Stephan, K. E. & Buhmann, J. M. The Balanced Accuracy and Its Posterior Distribution. in *2010 20th International Conference on Pattern Recognition* 3121–3124 (IEEE, 2010). doi:10.1109/ICPR.2010.764
30. Gertz, J. *et al.* Distinct Properties of Cell-Type-Specific and Shared Transcription Factor Binding Sites. *Mol. Cell* **52,** 25–36 (2013).
31. Rohs, R. *et al.* Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.* **79,** 233–69 (2010).
32. Liu, L., Jin, G. & Zhou, X. Modeling the relationship of epigenetic modifications to transcription factor binding. *Nucleic Acids Res.* **43,** 3873–85 (2015).
33. John, S. *et al.* Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet.* **43,** 264–268 (2011).
34. Donaghey, J. *et al.* Genetic determinants and epigenetic effects of pioneer-factor occupancy. *Nat. Genet.* **50,** 250–258 (2018).
35. Casey, B. H., Kollipara, R. K., Pozo, K. & Johnson, J. E. Intrinsic DNA binding properties demonstrated for lineage-specifying basic helix-loop-helix transcription factors. *Genome Res.* **28,** 484–496 (2018).
36. Velasco, S. *et al.* A Multi-step Transcriptional and Chromatin State Cascade Underlies Motor Neuron Programming from Embryonic Stem Cells. *Cell Stem Cell* **20,** 205–217.e8 (2017).
37. Yan, C., Chen, H. & Bai, L. Systematic Study of Nucleosome-Displacing Factors in Budding Yeast. *Mol. Cell* **71,** 294–305.e4 (2018).
38. Wreczycka, K., Franke, V., Uyar, B., Wurmus, R. & Akalin, A. HOT or not: Examining the basis of high-occupancy target regions. *bioRxiv* 107680 (2017). doi:10.1101/107680
39. Yamada, N., Lai, W. K. M., Farrell, N., Pugh, B. F. & Mahony, S. Characterizing protein–DNA binding event subtypes in ChIP-exo data. *Bioinformatics* **35,** 903–913 (2019).
40. Zhang, Y. *et al.* Primary sequence and epigenetic determinants of in vivo occupancy of genomic DNA by GATA1. *Nucleic Acids Res.* **37,** 7024–7038 (2009).
41. Soufi, A. *et al.* Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. *Cell* **161,** 555–568 (2015).
42. Wapinski, O. L. *et al.* Hierarchical Mechanisms for Direct Reprogramming of Fibroblasts to

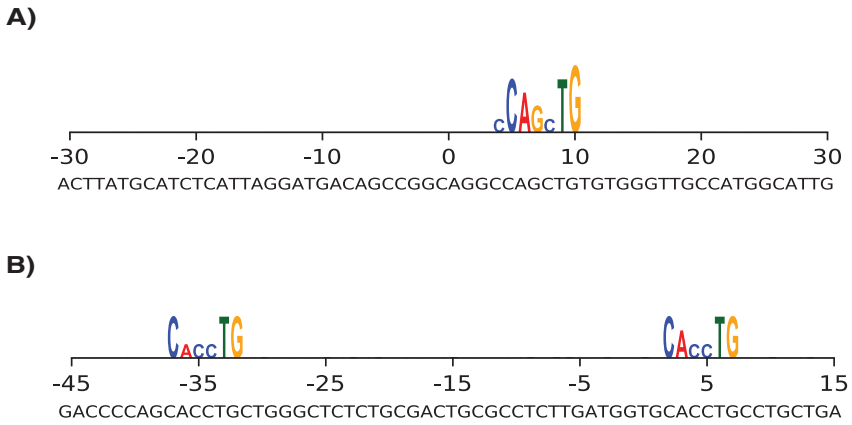
Neurons. *Cell* **155,** 621–635 (2013).

43. Cirillo, L. A. *et al.* Opening of Compacted Chromatin by Early Developmental Transcription Factors HNF3 (FoxA) and GATA-4. *Mol. Cell* **9,** 279–289 (2002).
44. Iwafuchi-Doi, M. & Zaret, K. S. Pioneer transcription factors in cell reprogramming. *Genes Dev.* **28,** 2679–92 (2014).
45. Mazzoni, E. O. *et al.* Embryonic stem cell–based mapping of developmental transcriptional programs. *Nat. Methods* **8,** 1056–1058 (2011).
46. Liang, K. & Keleş, S. Normalization of ChIP-seq data with control. *BMC Bioinformatics* **13,** 199 (2012).
47. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
48. Mahony, S. *et al.* An Integrated Model of Multiple-Condition ChIP-Seq Data Reveals Predeterminants of Cdx2 Binding. *PLoS Comput. Biol.* **10,** e1003501 (2014).
49. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).
50. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9,** 215–216 (2012).
51. Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. How transferable are features in deep neural networks? in *Advances in neural information processing systems* 3320–3328 (2014).

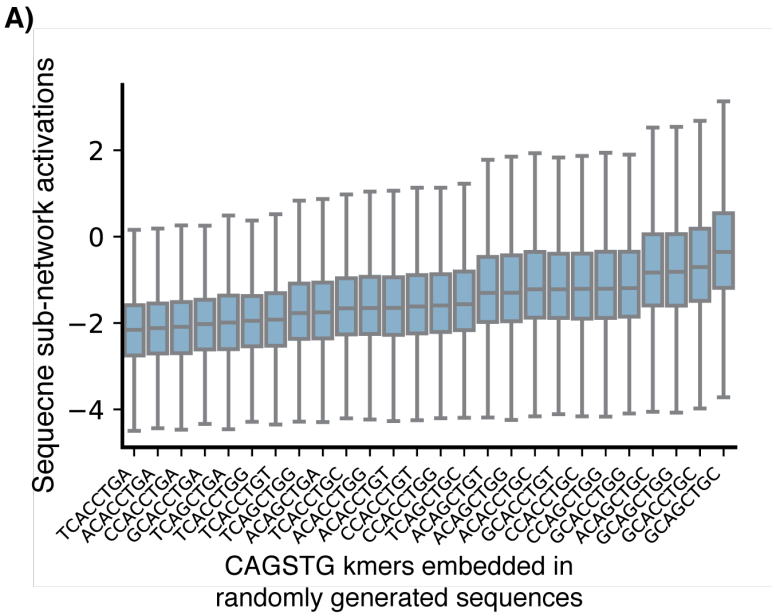
## Supplementary Material



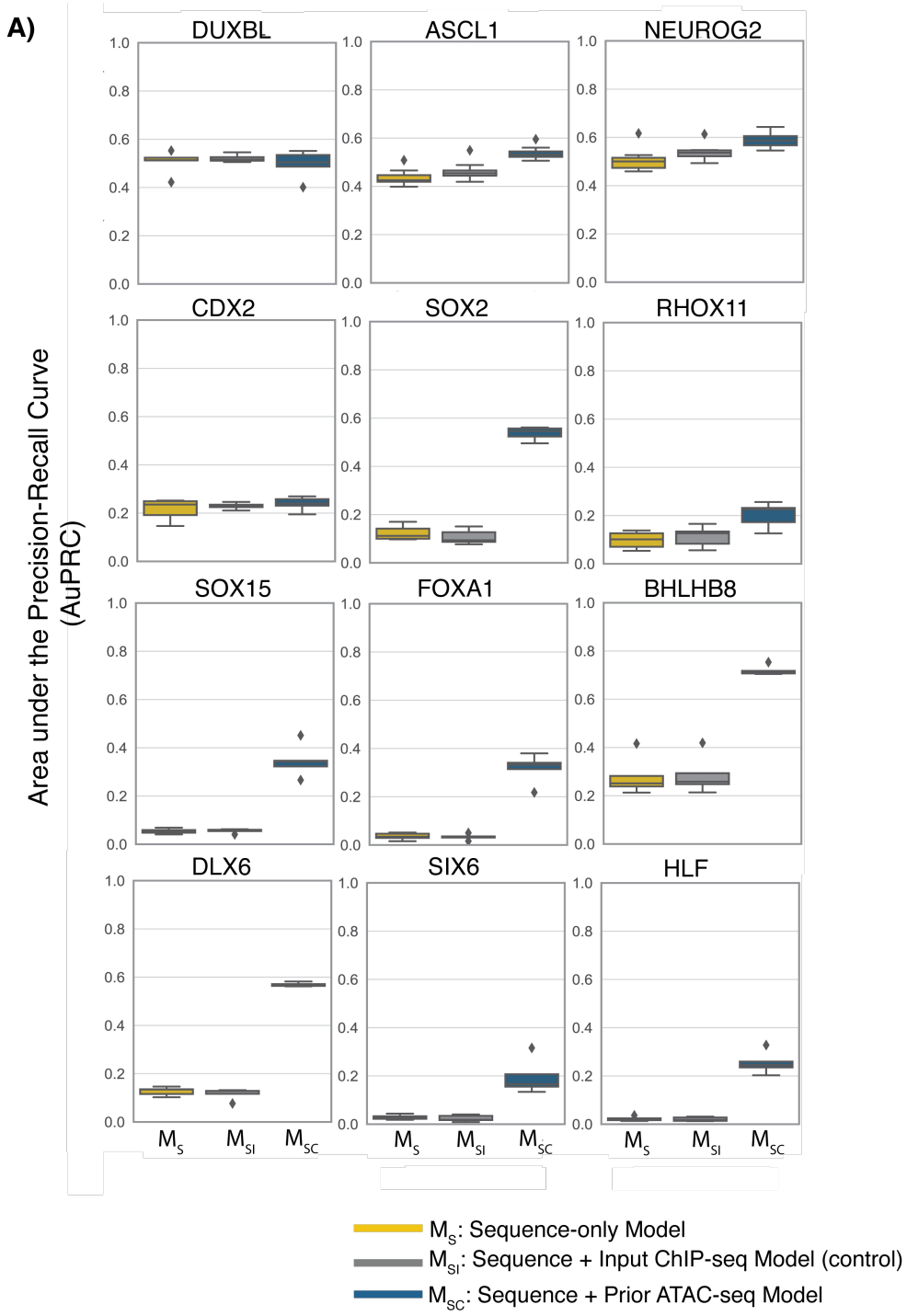

**Supplementary Figure 1: A)** The receiver operating characteristic (ROC) curve for the sequence-only and the joint sequence and prior chromatin models for a classifier trained to predict Ascl1 binding. At a fixed false positive rate (FPR) of 0.05, both models have high true positive rates (TPR > 0.9). **B)** Model auPRC distributions for held-out chromosomes 10-14. The performance of the additive versus interactive sequence and prior chromatin networks are comparable, confirming that we are not losing predictive ability with a simpler bimodal architecture.



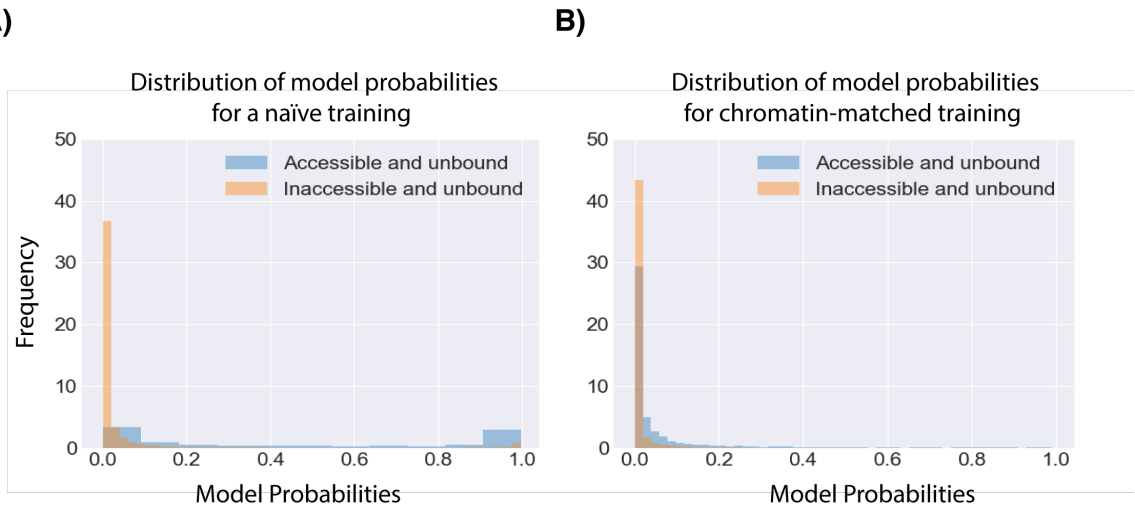

**Supplementary Figure 2: a) & b)** Feature attribution with integrated gradients at two example Ascl1 binding sites (chr10:4710120-4710170 and chr10:28136730-28136800).




**Supplementary Figure 3:** The sequence sub-network scores for CAGSTG *k*-mers +1bp flanks embedded in 10,000 randomly generated sequence windows (background nucleotide frequency-A/T:0.5 and G/C:0.5).




**Supplementary Figure 4: A)** Genome-wide auPRC distributions on held-out chromosomes 10-14. The addition of an input ChIP-seq experiment as a control does not lead to improvement in model performance when compared to the addition of pre-existing chromatin accessibility data.




**Supplementary Figure 5:** Model probabilities for Ascl1 at unbound genomic windows divided by prior accessibility. **A)** In the naïve training, a large number of pre-accessible unbound regions (blue) are incorrectly assigned a model probability close to 1. In chromatin-matched training **B)**, this bias is lost, with the model behaving more uniformly across pre-accessible and pre-inaccessible unbound windows.

**Supplementary Table 1:** Prior chromatin datasets for mouse embryoid bodies, generated in the same cell line as that used to over-express Ascl1 and Neurog2 in the current study.

| Data | Replicates | Cell Type | Cell Line |
|---|---|---|---|
| ATAC-seq[1] | 2 | EB | Ainv15 |
| H3K27ac[1] | 2 | EB | Ainv15 |
| H3K27me3[1] | 2 | EB | Ainv15 |
| H3K4me1[1] | 2 | EB | Ainv15 |
| H3K4me2[1] | 2 | EB | Ainv15 |
| H3K4me3[1] | 2 | EB | Ainv15 |

**Supplementary Table 2:** Prior chromatin datasets from mouse embryonic stem cells, sourced from referenced publications.

| Data | Replicates | Cell Type | Cell Line |
|---|---|---|---|
| H3K9ac[2] | 2 | ES | E14 |
| H3K9me3[2] | 2 | ES | E14 |
| H3K36me3[2] | 2 | ES | E14 |
| H2A.Z[3] | 1 | ES | V6.5 |
| acH2A.Z[3] | 1 | ES | V6.5 |
| H3K4me20[4] | 2 | ES | V6.5 |

**Supplementary Table 3:** ChIP-seq data for TFs induced in NIH3T3 fibroblasts, downloaded from Raccaud *et al*[5] and mES cells, downloaded from Aydin *et al*[6]. Processed with bowtie (1.0.1) and multiGPS (version 0.74).

| Data | Cell Type | Number of Peaks | Replicates | mutliGPS Signal Fraction |
|---|---|---|---|---|
| bHLHb8 | NIH-3T3 | 47,106 | 2 | 0.083,0.117 |
| Cdx2 | NIH-3T3 | 15,650 | 1 | 0.029 |
| Dlx1 | NIH-3T3 | 15,163 | 1 | 0.027 |
| Duxbl | NIH-3T3 | 21,668 | 1 | 0.04 |
| FoxA1 | NIH-3T3 | 4,736 | 2 | 0.014,0.010 |
| Hlf | NIH-3T3 | 3,131 | 1 | 0.009 |
| Rhox11 | NIH-3T3 | 4,653 | 1 | 0.01 |
| Six6 | NIH-3T3 | 5,419 | 1 | 0.013 |
| Sox15 | NIH-3T3 | 5,114 | 1 | 0.013 |
| Sox2 | NIH-3T3 | 4,421 | 1 | 0.019 |
| Neurog2 | mES | 26,643 | 3 | 0.24, 0.11, 0.18 |
| Ascl1 | mES | 21,176 | 3 | 0.08, 0.11, 0.12 |

**Supplementary Table 4:** Replicates of ATAC-seq experiments used as prior chromatin data in mouse NIH3T3 fibroblasts[5].

| Data | Cell Type | Cell Line |
|---|---|---|
| ATAC-seq | Mouse Fibroblasts | NIH3T3 |
| ATAC-seq | Mouse Fibroblasts | NIH3T3 + rtTA3G control |
| ATAC-seq | Mouse Fibroblasts | NIH3T3 + rtTA3G control |
| ATAC-seq | Mouse Fibroblasts | NIH3T3 + rtTA3G control |
| ATAC-seq | Mouse Fibroblasts | NIH3T3 + rtTA3G control |

## References

1. Velasco, S. *et al.* A Multi-step Transcriptional and Chromatin State Cascade Underlies Motor Neuron Programming from Embryonic Stem Cells. *Cell Stem Cell* **20,** 205–217.e8 (2017).
2. Yue, F. *et al.* A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515,** 355–364 (2014).
3. Ku, M. *et al.* H2A.Z landscapes and dual modifications in pluripotent and multipotent stem cells underlie complex genome regulatory functions. *Genome Biol.* **13,** R85 (2012).
4. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448,** 553–560 (2007).
5. Raccaud, M. *et al.* Mitotic chromosome binding predicts transcription factor properties in interphase. *Nat. Commun.* **10,** 487 (2019).
6. Aydin, B. *et al.* Proneural factors Ascl1 and Neurog2 contribute to neuronal subtype identities by establishing distinct chromatin landscapes. *Nat. Neurosci.* **22,** 897–908 (2019).