Virtual ChIP-seq: predicting transcription factor binding by learning from the transcriptome

Mehran Karimzadeh[1,2,3] and Michael M. Hoffman[1,2,4,5]

[1]Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada
[2]Princess Margaret Cancer Centre, Toronto, ON, Canada
[3]Vector Institute, Toronto, ON, Canada
[4]Department of Computer Science, University of Toronto, Toronto, ON, Canada
[5]Lead contact: michael.hoffman@utoronto.ca

May 10, 2018

Abstract

Motivation:

Identifying transcription factor binding sites is the first step in pinpointing non-coding mutations that disrupt the regulatory function of transcription factors and promote disease. ChIP-seq is the most common method for identifying binding sites, but performing it on patient samples is hampered by the amount of available biological material and the cost of the experiment. Existing methods for computational prediction of regulatory elements primarily predict binding in genomic regions with sequence similarity to known transcription factor sequence preferences. This has limited efficacy since most binding sites do not resemble known transcription factor sequence motifs, and many transcription factors are not even sequence-specific.

Results:

We developed Virtual ChIP-seq, which predicts binding of individual transcription factors in new cell types using an artificial neural network that integrates ChIP-seq results from other cell types and chromatin accessibility data in the new cell type. Virtual ChIP-seq also uses learned associations between gene expression and transcription factor binding at specific genomic regions. This approach outperforms methods that use transcription factor sequence preferences in the form of position weight matrices, predicting binding for 31 transcription factors (Matthews correlation coefficient $>$ 0.3).

Availability:

The datasets we used for training and validation are available at https://virchip.hoffmanlab.org. We have deposited in Zenodo the current version of our software (`http://doi.org/10.5281/zenodo.1066928`), datasets (`http://doi.org/10.5281/zenodo.823297`), predictions for 31 transcription factors on Roadmap Epigenomics cell types (`http://doi.org/10.5281/zenodo.1243913`), and predictions in Cistrome as well as ENCODE-DREAM *in vivo* TF Binding Site Prediction Challenge (`http://doi.org/10.5281/zenodo.1209308`).

1 Introduction

Transcription factor (TF) binding regulates gene expression. Each TF can harmonize expression of many genes by binding to genomic regions that regulate transcription. Cellular machinery utilizes these master regulators to regulate key cellular processes and adapt to environmental stimuli. Alteration in sequence or quantity of a given TF can impact expression of many genes. In fact, these alterations can be the primary cause of hereditary disorders, complex disease, autoimmune defects, and cancer[1].

TFs bind to accessible chromatin based on weak non-covalent interactions between amino acid residues and nucleic acids. DNA's primary structure (sequence)[2], secondary structure (shape)[3], and tertiary structure (conformation)[4] all play roles in TF binding. Many TFs form a complex with others as well as chromatin-binding proteins and therefore bind to DNA indirectly. Some TFs also have different isoforms and undergo various post-translational modifications. *In vitro* assays, such as high throughput systematic evolution of ligands by exponential enrichment (HT-SELEX)[5] and protein binding microarrays[6], have provided a compelling understanding of context-independent TF sequence and shape preference[7]. Yet, for the aforementioned reasons, performance of models trained on these *in vitro* data are poor when applied on *in vivo* experiments[8,9]. To address this challenge, we must explore how to better model DNA shape, TF-TF interactions, and context-dependent TF binding.

Chromatin immunoprecipitation and sequencing (ChIP-seq)[10] and similar methods, such as ChIP-exo[11] and ChIP-nexus[12], can map the presence of a given TF in the genome of a biological sample. To map TFs, these assays require a minimum of 1,000,000 to 100,000,000 cells, depending on properties of the TF itself and available antibodies. Such large numbers of cells are not often available from clinical samples. Therefore, it is impossible to systematically assess TF binding in most disease systems. Assessing chromatin accessibility through transposase-accessible chromatin using sequencing (ATAC-seq)[13], however, requires only hundreds or thousands of cells. One can obtain this many cells from many more clinical samples. While chromatin accessibility does not determine TF binding, several methods use this information together with knowledge of TF sequence preference, genomic conservation, and other genomic features to predict TF binding[14,15,16].

Predicting TF binding with motif discovery tools within chromatin accessible regions has helped us understand the role of several TFs in various disease. For example, He et al.[17] used motif discovery tools to identify the role of OCT1 and NKX3-1 after prolonged androgen stimulation in prostate cancer. Similarly, Bailey et al.[18] discovered that a known breast cancer risk single nucleotide polymorphism (SNP) upstream of *ESR1* disrupts GATA3 binding and enhances expression of *ESR1*. We propose that using more accurate tools to predict TF binding will allow understanding the role of TF binding in more contexts.

Previous studies have used various approaches to predict TF binding. Several methods use unsupervised approaches such as hierarchical mixture models[14] or hidden Markov models[15] to identify transcription factor footprint using chromatin accessibility data. These approaches use sequence motif scores to attribute footprints to different transcription factors. Convolutional neural network models can boost precision by learning sequence preferences from *in vivo*, rather than *in vitro* data[20,21]. Variation in sequence specificity and cooperative binding of some transcription factors prevents these methods from accurately predicting binding of all transcription factors. A more recent approach uses matrix completion to impute TF binding using a 3-mode tensor representing genomic positions, cell types, and TF binding[22]. This method doesn't rely on sequence specificity, but can only predict TF binding in well-studied cell types with many ChIP-seq datasets. This means one cannot use it to predict binding in a cell type where ChIP-seq is not possible, such as limited clinical samples.

Identifying the best approach for predicting TF binding remains a challenge, because most

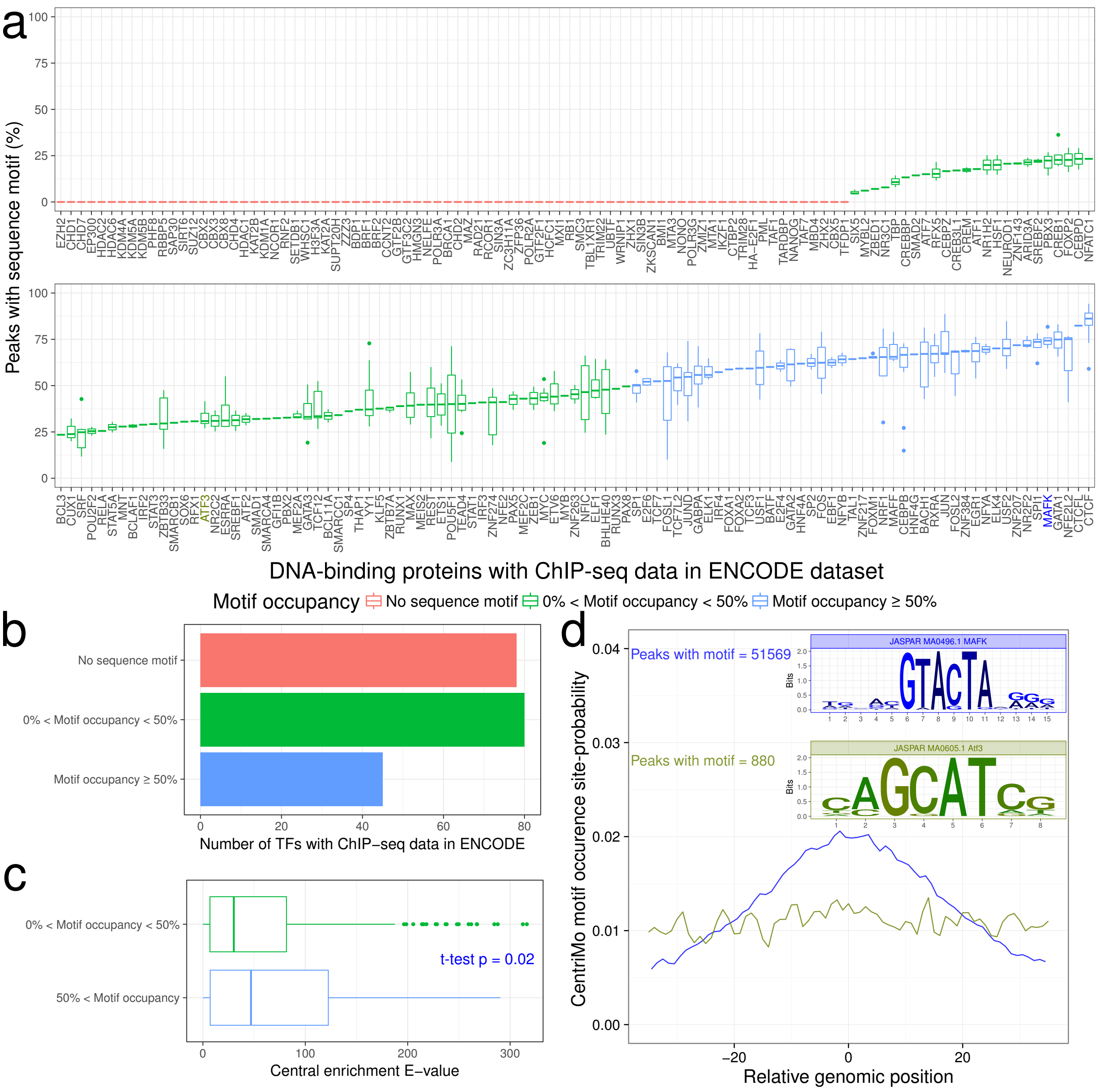

Figure 1: **Most ChIP-seq peaks lack the TF's sequence motif. (a)** Fraction of ENCODE ChIP-seq peaks for a TF with any JASPAR sequence motif from the TF's family. Boxplots show the distribution among datasets from different cell types and replicates. Horizontal line of boxplot: median. Box range: interquartile range (IQR). Whisker: most extreme value within quartile ±1.5 IQR. Individual points: outliers beyond a whisker. **(b)** Number of TFs with no sequence motif (red), TFs where less than 50% of peaks have the sequence motif (low motif occupancy, green), and TFs where 50% or more of peaks have the sequence motif (high motif occupancy, blue). **(c)** Central enrichment[19] of a TF's motif is lower for TFs with motif occupancy of less than 50% compared to TFs with motif occupancy of 50% or more. **(d)** For TFs with a small number of peaks matching sequence motif of the same TF, such as ATF3, central enrichment of the motif is also low. In contrast, most MAFK peaks both contain its sequence motif and show central enrichment.

studies use different benchmarking approaches. For example, one earlier study[14] only assesses prediction on genomic regions that match the TF's sequence motif. By excluding ChIP-seq peaks not matching the TF's sequence motif from benchmarking, it underestimates false negative peaks and overestimates prediction accuracy. Most previous studies benchmark their predictions using the area under receiver operating characteristic curve (auROC) statistic[22,23,24]. When test data is imbalanced, meaning it has very different numbers of positive and negative examples, using auROC misleads evaluators[25,26]. Unfortunately, the TF binding status of genomic regions is highly imbalanced, making auROC alone a poor metric for evaluating TF binding prediction. Evaluation is further complicated by wildly varying prediction performance across different TFs. Recently, the ENCODE-DREAM *in vivo* TF Binding Site Prediction Challenge (DREAM Challenge) introduced guidelines for assessing TF binding prediction[27]. They recommend reporting both auROC, which assesses false negative predictions and the area under precision-recall curve (auPR), which also assesses false positives.

RNA-seq allows us to obtain transcriptome data from samples with small cell counts, including patient samples. We hypothesized that we could leverage the transcriptome to better predict TF binding. Previous methods have predicted gene expression using information on active regulatory elements[28,29,30]. Others have predicted chromatin accessibility using gene expression data[31], but they haven't predicted TF binding using transcriptome data, as we do below.

Here, we introduce Virtual ChIP-seq, a novel method for more accurate prediction of TF binding. Virtual ChIP-seq predicts TF binding by learning from publicly available ChIP-seq experiments. Unlike Qin and Feng[23], it can do this in new cell types with no existing ChIP-seq data. Virtual ChIP-seq also learns from other data such as genomic conservation, and the association of gene expression with TF binding.

Virtual ChIP-seq also accurately predicts the locations of DNA-binding proteins without known sequence preference. This would be impossible for most existing methods, which rely on sequence preference. Strictly speaking, only some of these proteins are TFs, but we usually refer to all DNA-binding proteins as TFs in this paper for ease of communication and comparison with other methods.

Virtual ChIP-seq predicted binding of 31 TFs in new cell types with a minimum Matthews correlation coefficient (MCC) of 0.3. These TFs had minimum accuracy (fraction of all predictions that were correct) of 0.99 and minimum specificity (fraction of negative predictions that were correct) of 0.99. Precision (fraction of positive predictions that were correct) ranged between 0.16 and 0.78 (Table 1). We predicted binding of these 31 TFs on 34 Roadmap Epigenomics[32] cell types and provide these predictions as a track hub for community use (https://virchip.hoffmanlab.org).

2 Results

2.1 Sequence motifs are absent in most TF binding sites

2.1.1 Most ChIP-seq peaks lack the TF's relevant sequence motif

Many computational tools predict TF binding using sequence preference data[14,15]. Most tools represent TF sequence preference in position weight matrix (PWM) format. PWMs encode the likelihood for presence of each nucleotide at different positions of a sequence motif. With tools such as FIMO[33], we can efficiently search and rank genomic regions that match TF sequence motifs.

One cannot determine a TF's binding sites based solely on its sequence preference. We can identify some additional properties, such as co-binding partners, from high-throughput experiments. For other properties, such as post-translational modifications to the TF, we lack corresponding large-scale data. Therefore, we expect existing computational prediction methods to be more accurate for

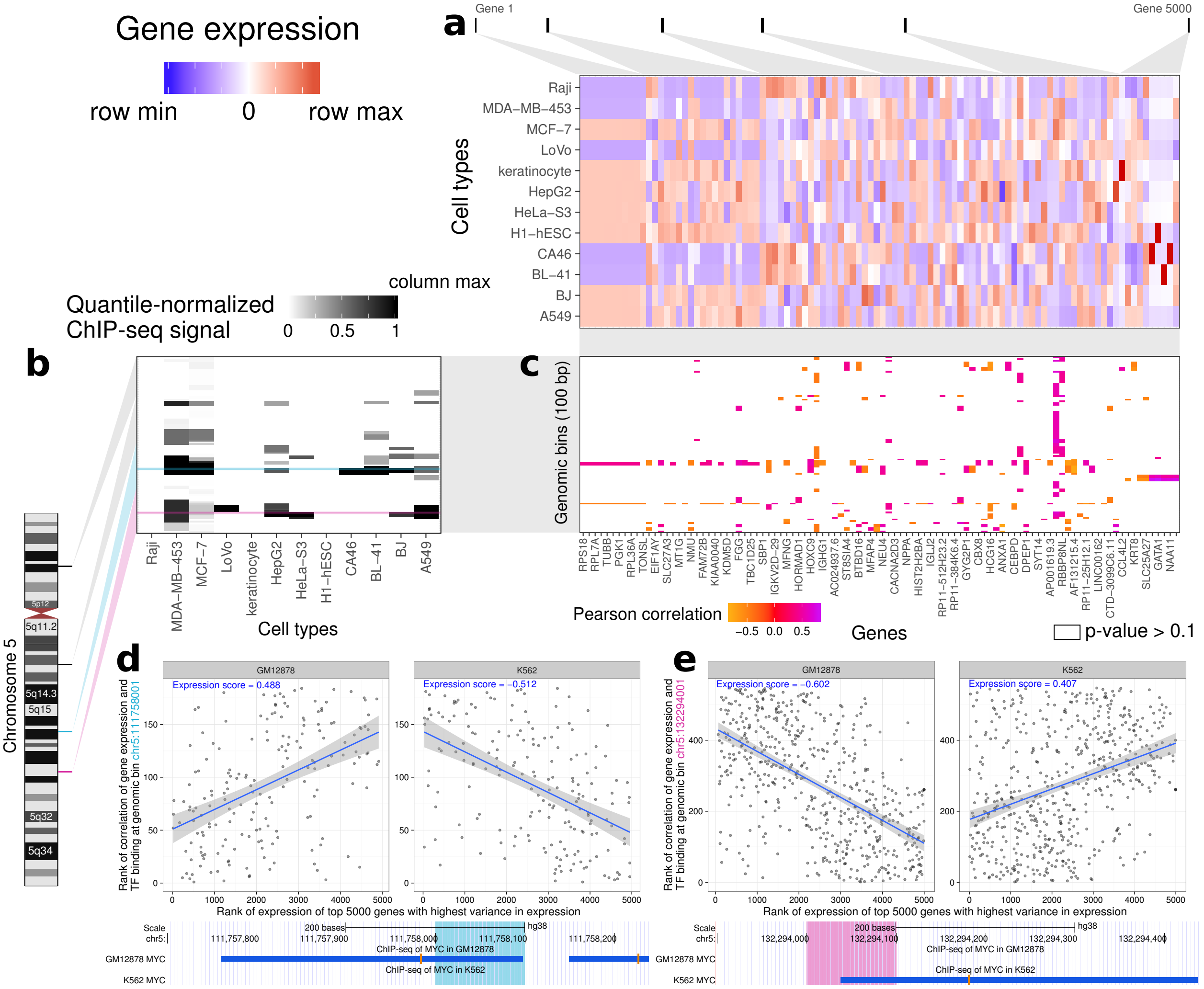

Figure 2: **Virtual ChIP-seq learns from association of gene expression and TF binding at each genomic bin.** This example shows Virtual ChIP-seq analysis for the MYC TF. **(a)** Gene expression levels for 5000 genes in 12 cell types. We ranked RNA-seq RPKM expression values within each cell type. This matrix shows a subset of 5000 high-variance genes, sorted by variance of each gene's expression between cell types. Blue: row minimum; white: median expression; red: row maximum. **(b)** ChIP-seq signal for 100 bp bins in 12 cell types, taken from four larger regions (25 bins each) on chromosome 5. We quantile-normalized ChIP signal from MACS software among cell types. This matrix shows a subset of the 54,037 bins on chromosome 5 which have TF binding in at least one training cell type. White: column minimum (0.0); black: column maximum (1.0). Cyan: a region in the *NREP* locus with *MYC* binding in GM12878; magenta: a region upstream of *SLC22A4* with *MYC* binding in K562. **(c)** Association matrix: gene expression–ChIP signal correlation between 100 genomic bins and 5000 high-variance genes. This is a subset of the larger 54,037 × 5,000 association matrix for chromosome 5. Each cell shows the Pearson correlation for 12 cell types between expression for a particular gene and ChIP signal at a particular genomic bin. Orange: negative correlation; white: p-value of Pearson correlation greater than 0.1 (NA); Purple: positive correlation. **(d)** (*Top*) Expression score plots for a 100 bp bin in the *NREP* locus. Each plot has one point for each of 184 genes with non-NA correlation values at that bin in the association matrix. Each point displays *(Continued on next page.)*

Figure 2: *(Continued from previous page.)* the rank of correlation value for that gene among one row of the association matrix against the rank of expression for that gene among 5000 high-variance genes in (*left*) GM12878 and (*right*) K562 cell types. The expression score at a bin for a cell type is Spearman's rank correlation coefficient ρ between those two ranks. Blue line: best linear fit to data; grey region: 95% confidence interval of the fit. (*Bottom*) UCSC Genome Browser display of 550 bp around that region. Blue rectangle: *MYC* ChIP-seq peak in GM12878 or K562. Here, *MYC* binds only in GM12878. **(e)** Expression score plot and Genome Browser display for a 100 bp bin upstream of *SLC22A4*. Here, *MYC* binds only in K562.

TFs where post-translational modifications and co-binding partners contribute less to TF binding. For TFs with more complex biology, however, we expect computational prediction methods to fail.

Using ChIP-seq data on 201 DNA-binding proteins in 54 different cell types, we investigated whether the majority of binding sites matched the sequence motif of the same TF. Among these 201 proteins, 76 lacked a sequence motif in JASPAR (Figure 1a, Supplementary Table 1). Some of these motif-free proteins, such as EZH2 and HDAC, are chromatin-binding proteins rather than true TFs. For simplicity in describing the prediction task, we refer to them as TFs nonetheless. Others are TFs without known sequence preference. For sequence-specific TFs, the fraction of peaks that match a sequence motif ranges from 4.55% (for SIX5) to 94.2% (for CTCF) with a mean of 49.4% (Figure 1b).

2.1.2 Many sequence motifs are not centrally enriched

Central enrichment measures how close a sequence motif occurs to a set of ChIP-seq peak summits. High central enrichment indicates direct TF binding[19]. We used CentriMo[19] to measure central enrichment. We compared central enrichment between TFs with low motif occupancy ($<$ 50% of ChIP-seq peaks contain the motif) and high motif occupancy ($\geq$ 50% of peaks contain the motif; Figure 1c). TFs with low motif occupancy had weaker central enrichment (t-test; $p = 0.02$). For example, 30.87% of ATF3 peaks overlapped with the MA0605.1 JASPAR motif. ATF3 peaks also had lower central enrichment than MAFK peaks, which had 74.29% overlap with the MA0496.1 JASPAR motif (Figure 1d).

2.2 Model, performance, and benchmarking

2.2.1 Datasets

Virtual ChIP-seq learns from the association of gene expression and TF binding in publicly available datasets. Our method requires ChIP-seq data of each TF in as many cell types as possible, with matched RNA-seq data from the same cell types. We used ChIP-seq data (from Cistrome DB[34] and ENCODE[35]) and RNA-seq data (from CCLE[36] and ENCODE[37]) to assess Virtual ChIP-seq's binding predictions for 63 DNA-binding proteins in new cell types.

In addition to benchmarking on our own held-out test cell types, we wanted to compare against the DREAM Challenge[27]. To do this, we also used their datasets, which include ChIP-seq data for 31 TFs. For most of these TFs, the DREAM Challenge held out test chromosomes instead of test cell types. The DREAM Challenge included ChIP-seq data for only 12 TFs in completely held-out cell types. Completely holding out cell types better fits the real-world scenarios that require binding site prediction. Using the datasets we generated, we had matched data in enough cell types to train and validate models for 9 of these 12 TFs (CTCF, E2F1, EGR1, FOXA1, GABPA, JUND, MAX, REST, and TAF1).

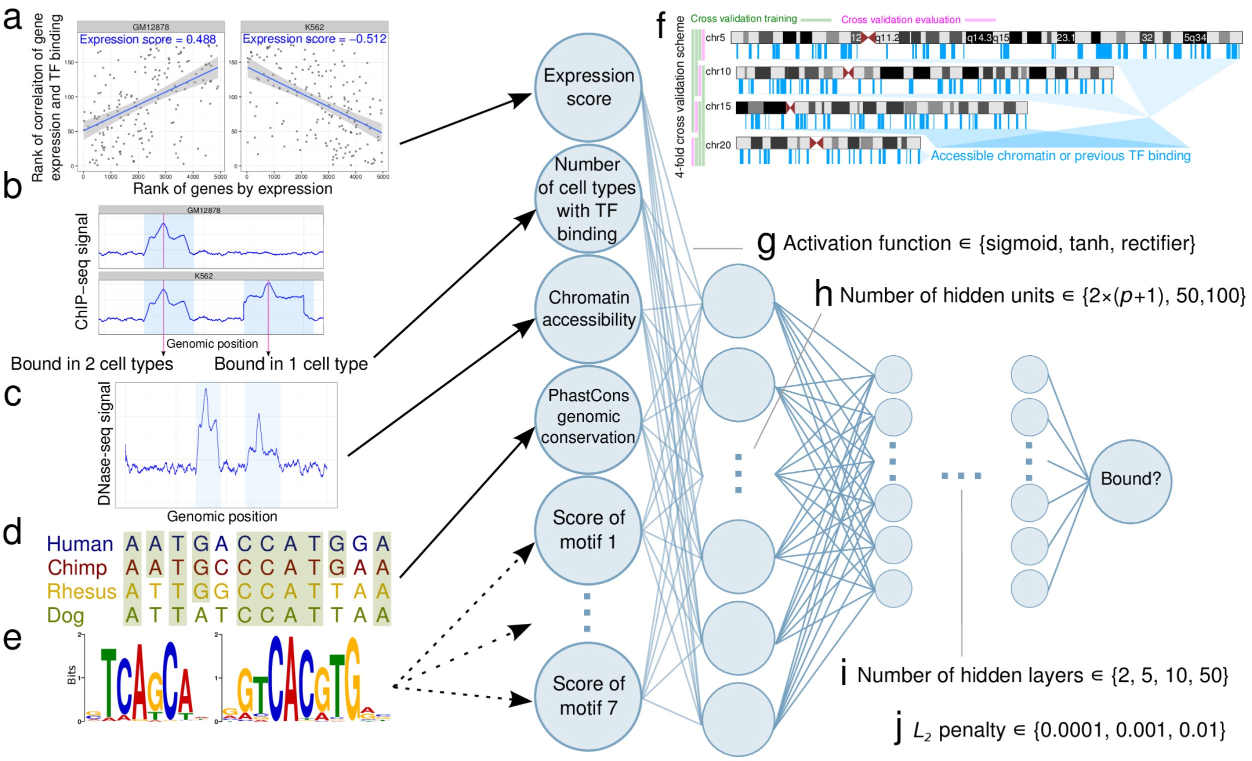
a
Rank of correlation of gene expression and TF binding
GM12878
Expression score = 0.488
K562
Expression score = −0.512
Rank of genes by expression
b
ChIP–seq signal
GM12878
K562
Genomic position
Bound in 2 cell types
Bound in 1 cell type
c
DNase-seq signal
Genomic position
d
Human A A T G A C C A T G G A
Chimp A A T G C C C A T G A A
Rhesus A T T G G C C A T T A A
Dog A T T A T C C A T T A A
e
Bits
Expression score
Number of cell types with TF binding
Chromatin accessibility
PhastCons genomic conservation
Score of motif 1
Score of motif 7
f
Cross validation training
Cross validation evaluation
4-fold cross validation scheme
chr5
chr10
chr15
chr20
Accessible chromatin or previous TF binding
g Activation function ∈ {sigmoid, tanh, rectifier}
h Number of hidden units ∈ {2×(p+1), 50,100}
Bound?
i Number of hidden layers ∈ {2, 5, 10, 50}
j L2 penalty ∈ {0.0001, 0.001, 0.01}

Figure 3: **Optimizing and training a multi-layer perceptron.** We used a number of features to predict TF binding in each bin. These include **(a)** expression score (Figure 2d–e), **(b)** the number of training cell types with binding of that TF, **(c)** chromatin accessibility, **(d)** PhastCons genomic conservation in placental mammals, and **(e)** any sequence motif corresponding to that TF in the JASPAR database. In JASPAR, some TFs have no sequence motifs, while others have up to seven different sequence motifs. This led to a number of features $p \in [4, 11]$, excluding features from HINT footprints or CREAM peaks not used in the main model. **(f)** For each TF, we trained a multi-layer perceptron using these features for selected bins in four chromosomes (5, 10, 15, and 20). Specifically, we selected bins with accessible chromatin or ChIP-seq signal in at least one training cell type (selected regions with vertical blue bars are for illustration purpose). To optimize hyperparameters, we repeated the training process with different hyperparameters using four-fold cross validation (CV), excluding one chromosome at a time. For each TF, we performed a grid search over **(g)** activation function (sigmoid, tanh, and rectifier), **(h)** number of hidden units per layer ($2(p + 1)$, 50, or 100), **(i)** number of hidden layers (2, 5, 10, or 50), and **(j)** L_2 regularization penalty (0.0001, 0.001, or 0.01). We chose the quadruple of hyperparameters which resulted in the highest mean Matthews correlation coefficient (MCC) over all four chromosomes.

2.2.2 Learning from the transcriptome

Different cell types have distinct transcriptomic and epigenomic states[38]. Changing gene expression levels can affect patterns of TF binding and chromatin structure. We hypothesized that some gene expression changes would lead to consistent and observable changes in TF binding. As an extreme example, eliminating expression of a TF would eventually eliminate binding of that TF genome-wide. Other changes in gene expression could lead to competitive, cooperative, allosteric, and other indirect effects that would affect TF binding. To exploit this model, we identified genes with significant positive or negative correlation with TF binding at any given genomic bin. We did this for genes all over the genome, irrespective of distance from the binding site.

For each TF, we created an *association matrix* measuring correlation between gene expression and binding of that TF in previously collected datasets (Figure 2a–c). In this matrix, each value corresponds to the Pearson correlation between ChIP-seq binding of that TF at one genomic bin and the expression level of one gene. We used missing values when there was no significant association between gene expression and TF binding ($p > 0.1$).

Power analysis (Methods) identified which correlations the $p > 0.1$ cutoff would exclude depending on the number of available cell types with matched ChIP-seq and RNA-seq data. For CTCF, which had the largest number of cell types available—21 cell types with matched ChIP-seq and RNA-seq—this cutoff provided 80% power to detect an absolute value of Pearson correlation $|r| \geq 0.52$. Many TFs had only 5 cell types with matched data and the cutoff provided 80% power to detect only larger correlations, $|r| \geq 0.92$.

We calculated an *expression score* for a TF in a new cell type using the association matrix and RNA-seq data for the new cell type, but no ChIP-seq data. The expression score is the Spearman correlation between the non-NA values for that genomic bin in the association matrix and the expression levels of those genes in the new cell type (Figure 2d, Figure 3a). We used the rank-based Spearman correlation to make the score robust against slight differences in analytical methodology used to estimate gene expression.

2.2.3 Learning from other predictive features

We included a number of other predictive features beyond expression score. Virtual ChIP-seq includes as input for each genomic bin the frequency of the TF's presence in existing ChIP-seq data (Figure 3b). Since most TF binding occurs within accessible chromatin[39], we also used evidence of chromatin accessibility from DNase-seq or ATAC-seq (Figure 3c).

While many intra-species genomic differences lie in the non-coding genome[40], we expect some regulatory elements to be conserved among closely related species. Previous studies highlight the association of genomic conservation and TF binding in organisms as simple as yeast[41] or as complex as human[42]. To learn from patterns of genomic conservation, we used PhastCons[43,44] scores from a 7-way primate and placental mammal comparison (`http://hgdownload.cse.ucsc.edu/goldenPath/hg38/phastCons7way`) in our model (Figure 3d).

We used sequence motif score where available (Figure 3e). Relying only on TF sequence preference, however, would prevent accurate prediction of most true TF binding sites[9] (Figure 1). For each TF, we represented sequence preference using the FIMO score of JASPAR sequence motifs of that TF or a similar TF. JASPAR has no motif for some TFs, such as EP300. Where JASPAR has more than one motif for a TF, additional motifs often represent different versions of the motif such as SREBF2 (MA0596.1) and SREBF2-var2 (MA0828.1). In some cases, the additional motif represents a preference of a cooperative TF heterodimer, such as MAX-MYC (MA0059.1). Regardless of reason, we included all of each TF's motifs as features in its model (Supplementary Table 2).

We also investigated potential improvements by adding a couple of additional integrative features available for a limited number of TFs and cell types (Supplementary Table 2). First, we used the output of Hidden Markov model-based Identification of TF footprints (HINT)[15] which identifies TF footprints within accessible chromatin. Second, we used a boolean feature indicating overlap of each genomic bin with clusters of chromatin accessibility peaks identified by CREAM[45].

2.2.4 Selecting hyperparameters and training

We created an input matrix with rows corresponding to 200 bp genomic windows and columns representing the features described above. Specifically, these features included expression score (Figure 3a), previous evidence of binding of TF of interest in publicly available ChIP-seq data (Figure 3b), chromatin accessibility (Figure 3c), genomic conservation (Figure 3d), sequence motif scores (Figure 3e), HINT footprints, and CREAM peaks. We used sliding genomic bins with 50 bp shifts, where most 200 bp bins overlap six other bins. This provided a maximum resolution of 50 bp in binding prediction. This resulted in a sparse matrix with 60,620,768 rows representing each bin in the GRCh38 genome assembly[46]. The sparse matrix used in the main model had between 4 and 11 columns, depending on the number of available sequence motifs. When we added HINT footprints and CREAM peaks, the matrix had between 6 and 13 columns instead. We trained on an imbalanced subset of genomic regions which had TF binding or chromatin accessibility (FDR $< 10^{-4}$) in any of the training cell types. To speed the process of training and evaluation, we further limited training input data to four chromosomes (chr5, chr10, chr15, and chr20). For validation, however, we used data from these same four chromosomes in completely different cell types held out from training. We evaluated the performance on all of the 9,635,407 bins in these four chromosomes (Figure 3f), not just those with prior evidence of TF binding or chromatin accessibility.

To build a generalizable classifier that performs well on new cell types with only transcriptome and chromatin accessibility data, we concatenated input matrices from 12 training cell types: A549, GM12878, HepG2, HeLa-S3, HCT-116, BJ, Jurkat, NHEK, Raji, Ishikawa, LNCaP, and T47D (Supplementary Table 3).

2.2.5 The multi-layer perceptron

The multi-layer perceptron (MLP) is a fully connected feed-forward artificial neural network[47]. Our MLP assumes binding at each genomic window is independent of upstream and downstream windows (Figure 3). For each TF, we trained the MLP with adaptive momentum stochastic gradient descent[48] and a minibatch size of 200 samples. We used 4-fold cross validation to optimize hyperparameters including activation function (Figure 3g), number of hidden units per layer (Figure 3h), number of hidden layers (Figure 3i), and L_2 regularization penalty (Figure 3j). In each cross validation fold, we iteratively trained on 3 of the 4 chromosomes (5, 10, 15, and 20) at a time, and assessed performance in the remaining chromosome. We selected the model with the highest average Matthews correlation coefficient (MCC)[49] after 4-fold cross validation. MCC incorporates all four categories of a confusion matrix and assesses performance well even on imbalanced datasets[50]. For 23 TFs the optimal model had 10 hidden layers, and for another 23 TFs the optimal model had 5 hidden layers, and for the final 17 TFs, the optimal model had only 2 hidden layers. For 57 TFs, the best-performing model had 100 hidden units in each layers. The optimal model of 6 TFs had 10–24 hidden units in their hidden layers. Different activation functions—sigmoid, hyperbolic tangent (tanh), or rectifier—proved optimal for different TFs (Supplementary Table 4).

2.2.6 Virtual ChIP-seq predicts TF binding with high accuracy

We evaluated the performance of Virtual ChIP-seq in validation cell types (K562, PANC-1, MCF-7, IMR-90, H1-hESC, and primary liver cells) which we did not use in calculating the expression score, training the MLP, or optimizing hyperparameters. Before predicting in new cell types, we chose a posterior probability cutoff for use in point metrics such as accuracy and F_1 score. When a TF had ChIP-seq data in more than one of the validation cell types, we chose the cutoff that maximizes MCC of that TF in H1-hESC cells. Then, we excluded H1-hESC when reporting threshold-requiring metrics. For these TFs, we pre-set a posterior probability cutoff of 0.4, the mode of the cutoffs for other TFs (Supplementary Table 5).

We used area under precision-recall (auPR) curves to compare performance of Virtual ChIP-seq in validation cell types with other available methods. Virtual ChIP-seq predicts binding of 31 TFs in validation cell types with MCC > 0.3, auROC > 0.9, and $0.3 <$ auPR < 0.8 (Figure 4a, Table 1,

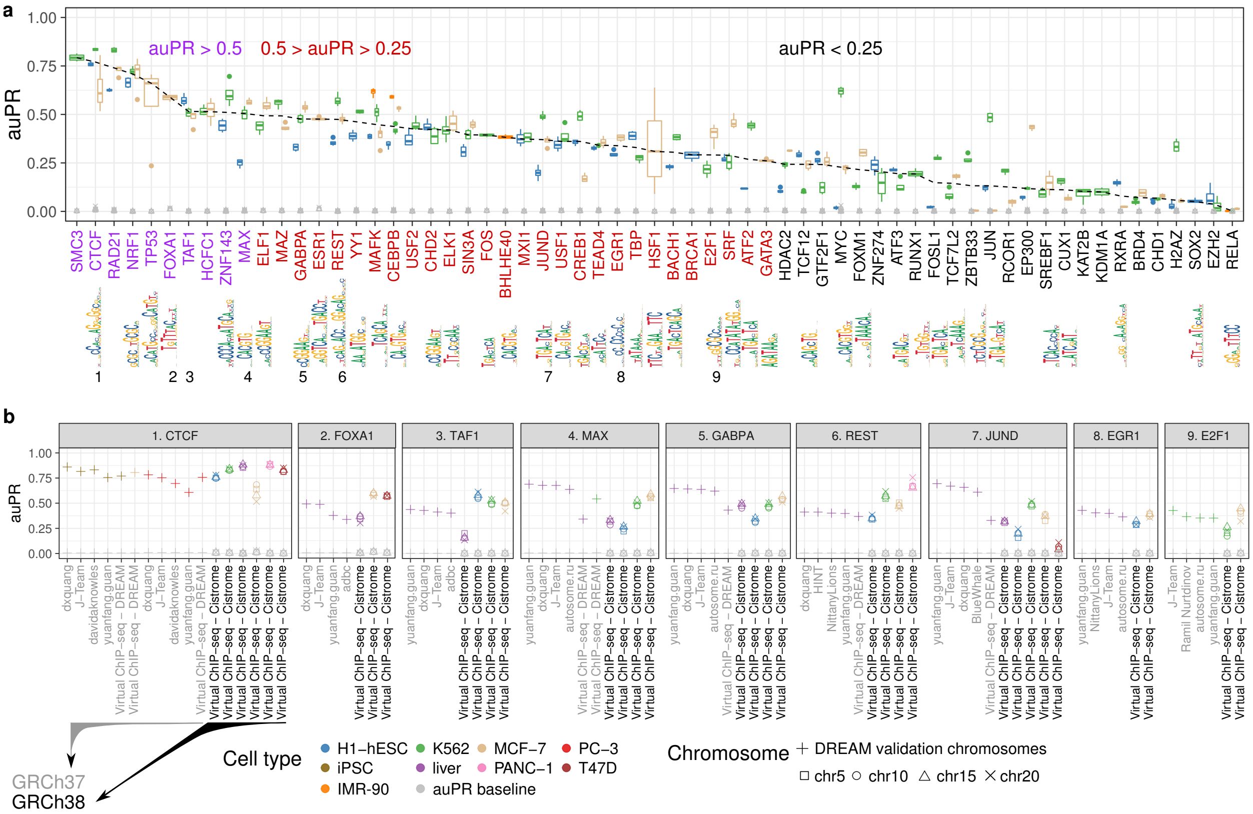

Figure 4: **Virtual ChIP-seq predicts TF binding with high accuracy.** Using ChIP-seq and RNA-seq data, we learned from the association of gene expression and TF binding for 63 TFs. **(a)** Box plots show distribution of auPR among 4 chromosomes (5, 10, 15, and 20) for 63 TFs assessed in four cell types (blue: H1-hESC; orange: IMR-90; green: K562; brown: MCF-7). Dashed line: medians. Grey shapes: prevalence of bound bins in the chromosome, the auPR baseline. Axis label colors categorize median auPR (purple: greater than 0.5, red: between 0.25 and 0.5, black: below 0.25). Sequence logos indicate one of a TF's JASPAR motifs, when available. When multiple motifs existed, we displayed the shortest motif here. Numbers 1–9: The nine TFs that the DREAM Challenge also evaluated in its final round. **(b)** We compared Virtual ChIP-seq's performance to that of the top 4 performing methods in the DREAM Challenge across-cell type final round. For CTCF, MAX, GABPA, REST, and JUND, we had enough cell types to train and validate the performance of Virtual ChIP-seq on DREAM data. For these TFs we trained on chromosomes 5, 10, 15, and 20 in training cell types and validated performance on merged data of chromosomes 1, 8, and 21 in validation cell types. For other TFs, we trained the model and validated our performance using publicly available Cistrome and ENCODE data. Axis label color: reference genome assembly (grey: GRCh37, black: GRCh38).

Supplementary Table 6).

2.2.7 Virtual ChIP-seq correctly predicts binding sites in genomic locations not found in training data

We evaluated the performance of Virtual ChIP-seq for 63 TFs with binding in validation cell types. For 59 of these TFs, Virtual ChIP-seq predicted true TF binding in regions without conservation among placental mammals. For 44 out of 63 TFs, Virtual ChIP-seq predicted true TF binding

TF	F_1	Accuracy	MCC	auROC	auPR	N
BHLHE40	0.334±0.021	0.997±0.000	0.356±0.010	0.974±0.002	0.382±0.01	1
CEBPB	0.510±0.091	0.992±0.002	0.515±0.072	0.964±0.017	0.534±0.073	3
CHD2	0.399±0.038	0.998±0.000	0.406±0.034	0.950±0.012	0.386±0.046	1
CREB1	0.362±0.131	0.997±0.002	0.371±0.121	0.868±0.135	0.335±0.174	2
CTCF	0.667±0.126	0.995±0.004	0.675±0.092	0.985±0.050	0.841±0.108	6
ELF1	0.431±0.047	0.997±0.001	0.456±0.038	0.949±0.042	0.493±0.066	2
ELK1	0.430±0.069	1.000±0.000	0.465±0.054	0.991±0.009	0.420±0.054	2
ESR1	0.372±0.103	0.993±0.006	0.430±0.049	0.883±0.033	0.461±0.019	2
FOS	0.333±0.027	0.997±0.001	0.393±0.020	0.861±0.004	0.394±0.008	1
FOSL1	0.319±0.006	0.994±0.001	0.316±0.006	0.929±0.006	0.272±0.012	1
FOXA1	0.433±0.082	0.997±0.004	0.492±0.072	0.981±0.022	0.568±0.117	3
GABPA	0.298±0.049	0.994±0.002	0.393±0.036	0.986±0.012	0.496±0.036	3
HCFC1	0.459±0.021	0.999±0.000	0.487±0.024	0.990±0.005	0.515±0.044	2
JUN	0.218±0.127	0.998±0.001	0.311±0.153	0.983±0.009	0.456±0.257	2
JUND	0.341±0.163	0.993±0.002	0.386±0.135	0.979±0.019	0.326±0.161	4
MAFK	0.354±0.041	0.997±0.001	0.423±0.028	0.989±0.005	0.513±0.103	3
MAX	0.400±0.045	0.996±0.002	0.444±0.059	0.961±0.012	0.491±0.111	3
MAZ	0.370±0.025	0.997±0.001	0.422±0.019	0.987±0.005	0.493±0.070	2
MXI1	0.394±0.018	0.999±0.000	0.402±0.017	0.993±0.004	0.381±0.025	1
NRF1	0.658±0.042	1.000±0.000	0.664±0.038	0.994±0.014	0.720±0.051	3
RAD21	0.593±0.062	0.996±0.002	0.626±0.056	0.983±0.033	0.740±0.095	3
REST	0.482±0.120	0.999±0.001	0.493±0.091	0.985±0.008	0.567±0.095	3
SIN3A	0.389±0.048	0.998±0.002	0.394±0.029	0.966±0.004	0.411±0.037	3
SMC3	0.733±0.016	0.999±0.000	0.734±0.016	0.998±0.001	0.792±0.018	1
SRF	0.353±0.060	0.998±0.001	0.364±0.070	0.982±0.008	0.365±0.115	2
TAF1	0.378±0.073	0.999±0.001	0.437±0.097	0.987±0.009	0.490±0.168	3
TEAD4	0.344±0.061	0.990±0.002	0.385±0.020	0.967±0.023	0.343±0.019	2
TP53	0.275±0.103	1.000±0.000	0.382±0.086	1.000±0.008	0.660±0.222	1
USF1	0.353±0.047	0.993±0.001	0.382±0.040	0.891±0.012	0.372±0.046	1
USF2	0.410±0.040	0.999±0.000	0.427±0.028	0.982±0.007	0.437±0.032	1
YY1	0.397±0.049	0.996±0.001	0.408±0.058	0.945±0.043	0.417±0.104	2

Table 1: **Performance of Virtual ChIP-seq for 31 TFs on validation cell types.** Each row displays median values ± standard deviation of several performance metrics for prediction of a TF across 4 chromosomes for each available validation cell type. MCC: Matthews correlation coefficient, auROC: area under receiver-operating characteristic curve, auPR: area under precision-recall, N: number of validation cell types for 31 TFs with MCC > 0.3. We reported auROC and auPR across all the validation cell types across all posterior probability cutoffs. Black TFs: we found the posterior probability cutoff which maximized MCC in H1-hESC, and then reported F_1, accuracy, and MCC of the other validation cell types.

in regions without TF binding in any of the training ChIP-seq data. From these 63 TFs, 43 are sequence-specific, and for all of these TFs, Virtual ChIP-seq predicted true binding for regions that did not match the TF's sequence motif. For 47 TFs, Virtual ChIP-seq even correctly predicted TF binding in regions that didn't overlap chromatin accessibility peaks (Supplementary Table 7).

Most of these regions were frequently bound to the TF in publicly available ChIP-seq data. These predictions showed that the MLP learned to leverage multiple kinds of information and predict TF binding accurately, even in the absence of features required by previous generations of binding site classifiers.

2.2.8 Comparison with DREAM Challenge

DREAM Challenge rules forbid using genomic conservation or ChIP-seq data as training features. This also excludes the expression score, as creating its association matrix relies on ChIP-seq data. The challenge also required training and validation on its own provided datasets. These datasets have ChIP-seq data in only a few cell types. This restricts Virtual ChIP-seq's approach which leverages all publicly available datasets. The DREAM Challenge ChIP-seq datasets use only two replicates for each experiment and requires that peaks have a irreproducibility discovery rate (IDR)[51] of less than 5%. IDR only handles experiments with exactly two replicates, but most of the public ChIP-seq experiments we used had more than two replicates (Supplementary Table 8). In these cases, we included peaks that pass a false discovery rate (FDR) threshold of 10^{-4} in at least two replicates.

The DREAM Challenge assessed participant entries by measuring performance on three validation chromosomes (chr1, chr8, and chr21), combined. To assess performance of Virtual ChIP-seq on DREAM Challenge data, we did the same. To assess performance on Cistrome data, however, we measured performance on each chromosome independently. This allowed us to examine the variance in performance among these chromosomes.

Although Virtual ChIP-seq used features not allowed in the DREAM Challenge, comparing with DREAM Challenge participants is the only sound way to show how any method including these features compares to the state of the art. Before the DREAM Challenge, TF binding prediction methods mostly reported performance measurements only in those parts of a chromosome where a method had more likelihood of success. The DREAM Challenge, like Virtual ChIP-seq, instead reports performance on the intended deployment domain of such methods: whole chromosomes. Leading DREAM Challenge methods potentially could improve their performance by including the features used by Virtual ChIP-seq. We compared Virtual ChIP-seq with DREAM Challenge results when we trained and validated on either Cistrome DB data or DREAM Challenge data.

2.2.9 Prediction accuracy varies by transcription factor

The DREAM Challenge evaluates predictions on binding of 31 TFs. The final submission round evaluates predictions for 12 TFs in held-out cell types. The datasets we used, however, allow us to predict binding of 63 TFs in new cell types. Of these TFs, 41 are unique to our dataset and do not overlap any of the DREAM Challenge TFs (Supplementary Table 9). The DREAM Challenge has data on the other 22 TFs, but the challenge evaluated only 9 of these TFs in its final round.

For CTCF, FOXA1, TAF1, and REST, Virtual ChIP-seq had a higher auPR in at least one validation cell type than any DREAM Challenge participant[52,53]. For EGR1 and E2F1, Virtual ChIP-seq performed better than at least one of the four top-performing methods of the challenge in one of the validation cell types (Figure 4b). DREAM Challenge and Cistrome ChIP-seq peak calls had different class imbalances, making auPR statistics not directly comparable (Supplementary Table 10). These imbalances were not always in the same direction. In FOXA1 peak calls in liver, for example, Cistrome called 0.12% of genomic bins bound to a TF, half the fraction of the DREAM Challenge (0.25%). Our predictions for FOXA1 binding in T47D and MCF-7 using Cistrome had a higher auPR than participants of DREAM Challenge for liver. The FOXA1 peak calls for these cell types also had a higher fraction of TF-bound genomic bins: 1.36% for MCF-7, and 0.39% for T47D.

This opposed the smaller fraction of bins bound in Cistrome data in CTCF (in PANC-1, liver, and T47D), TAF1 (in liver, H1-hESC, K562, and T47D), and REST (in H1-hESC, K562, and PANC-1). The differences in class prevalence are both minor and in diverging directions. Because of this, they do not bias the baseline auPR of evaluation on Cistrome datasets in a particular direction when compared to evaluation on DREAM Challenge datasets.

The power of Virtual ChIP-seq to learn from the transcriptome data diminishes when fewer cell types are available, as in the DREAM Challenge data. Nonetheless, when trained on DREAM Challenge data, Virtual ChIP-seq outperformed 13/14 DREAM Challenge participants when predicting CTCF binding in PC-3 cells. When predicting CTCF binding in iPSC cells, Virtual ChIP-seq had a higher auPR than 8/14 Challenge participants. The Virtual ChIP-seq auPR for binding of REST in liver was also higher than that of 9/14 DREAM Challenge participants (Supplementary Table 11).

Virtual ChIP-seq predicted binding of 31 TFs with a median MCC > 0.3. These 31 TFs had a auPR between 0.27 and 0.84 (Table 1). Some of these TFs show high levels of consistent binding among different cell types, which makes predictions easier. The fraction of bins bound to a TF in at least half of training cell types, however, varies between 0 to 15.75% across all TFs. Even for TFs with a median auPR > 0.5 (purple in Figure 4a) the fraction of bins bound in half of training cell types varied from 0.5% in FOXA1 to 10.5% in NRF1. For some DNA-binding proteins, Virtual ChIP-seq fails to predict binding accurately (auPR < 0.3). DNA-binding proteins with low auPR and low MCC include chromatin modifiers such as KAT2B, KDM1A, EZH2 and chromatin binding proteins such as CHD1 and BRD4. TFs with low prediction accuracy include ATF2, CUX1, E2F1, EP300, FOSL1, FOXM1, JUN, RCOR1, RELA, RXRA, SREBF1, TCF12, TCF7L2, and ZBTB33. For some proteins, such as ATF2, EP300, EZH2, FOXM1, KAT2B, KDM1A, TCF12, and TCF7L2, in at least one validation cell type, most ChIP-seq peaks didn't overlap with chromatin accessible regions.

2.3 The choice of input features determines prediction performance

2.3.1 The most important features

To evaluate the importance of each feature in our predictive model, we performed an ablation study on training data. First, we systematically removed features. Second, we fitted the model without these features on some of the training cell types (HeLa-S3, GM12878, HCT-116, LNCaP). Third, we evaluated performance on one held-out training cell type (HepG2; Supplementary Table 12). This ablation study did not use any of the validation cell types which we used for final evaluation of the model.

We called the effect of excluding an input feature substantive only when the average increase or decrease in auPR was at least 0.05. Excluding sequence motif, HINT, or CREAM did not substantively change performance of the model for most TFs (Figure 5). Excluding publicly available ChIP-seq data, the expression score, or both decreased performance in most TFs. Excluding expression score substantively decreased median auPR in 13/21 TFs, while excluding publicly available ChIP-seq data substantively decreased auPR in 18/21 TFs.

2.3.2 Inclusion of some features have opposite effects on prediction of different TFs

Beyond the most important features—ChIP-seq and expression score—excluding other features rarely substantively decreased prediction performance (Figure 5b–c). When we excluded sequence motifs, auPR decreased substantively for ZBTB33, JUN, JUND, FOXA1, and ELF1. Excluding HINT footprints decreased auPR substantively only for CEBPB, JUN, and JUND. Excluding

CREAM clusters of chromatin accessibility peaks decreased auPR substantively only for ZBTB33, ELF1, and FOXA1.

Removing certain input features actually improved prediction for some TFs (Figure 5b–c). Associations that differed between training cell types and validation cell types suggested that these input features generalize poorly. For example, CREAM clusters' overlap with NRF1 ChIP-seq peaks was not consistent among GM12878 (7.52%), HeLa-S3 (31.8%), and HepG2 (25.78%). This represented a significant variation among these cell types (ANOVA; $p = 1.9 \times 10^{-4}$).

While most TF footprints (95.96%) overlapped NRF1 peaks, TF footprints constituted only a small fraction of NRF1 peaks (0.73%). NRF1 peaks overlapped a small proportion of TF footprints in training cell types GM12878 (1.14%) and HeLa-S3 (0.59%), but significantly greater than the 0.45% overlap in HepG2 (Welch t-test; $p = 0.007$). In HepG2, 7.28% of YY1 peaks overlap TF footprints while in the training cell type GM12878, the overlap is only 1.22% (Welch t-test; $p = 5 \times 10^{-5}$) and in the other training cell type HCT-116 the overlap is much higher (17.92%; Welch t-test; $p = 5 \times 10^{-6}$). Overlap of ZBTB33 peaks with TF footprints is much smaller in HepG2 (0.49%) compared to training cell types GM12878 (2.32%) and HCT-116 (5.27%; Welch t-test; $p = 6 \times 10^{-4}$). Features with varying and cell-specific association with TF binding complicate convergence of the MLP and may result in overfitting. As a result, the MLP achieved a higher performance on some TFs when we ablated those features.

Association of clusters of regulatory elements and TF footprints with TF binding varies among cell types. Using a CREAM feature substantively improved performance in 3/21 TFs and using a HINT feature substantively improved performance in 3/21 TFs (Figure 5b–c). In contrast, including CREAM substantively decreased performance for 1 case and including HINT for 4 cases. When we repeat this experiment by using different training and validation cell types, clusters of regulatory elements and TF footprints result in increase or decrease in performance of different TFs, while they barely result in an increase in auPR above 0.05. Because of the limited upside and apparent downside, we didn't use these two features for our final model.

2.4 Transcription factors and their targets regulate similar biological pathways

2.4.1 Gene set enrichment analysis of TF targets

To understand biological implications of transcriptome perturbation in response to TF binding, we measured how frequently each gene's expression associated with binding of each TF. We hypothesized that if expression of a gene consistently correlates with binding of a TF, it is a potential target of that TF. Similarly, if the expression of a gene negatively correlates with binding of a TF, cellular machinery upregulated by that TF might cause net suppression of that gene's expression.

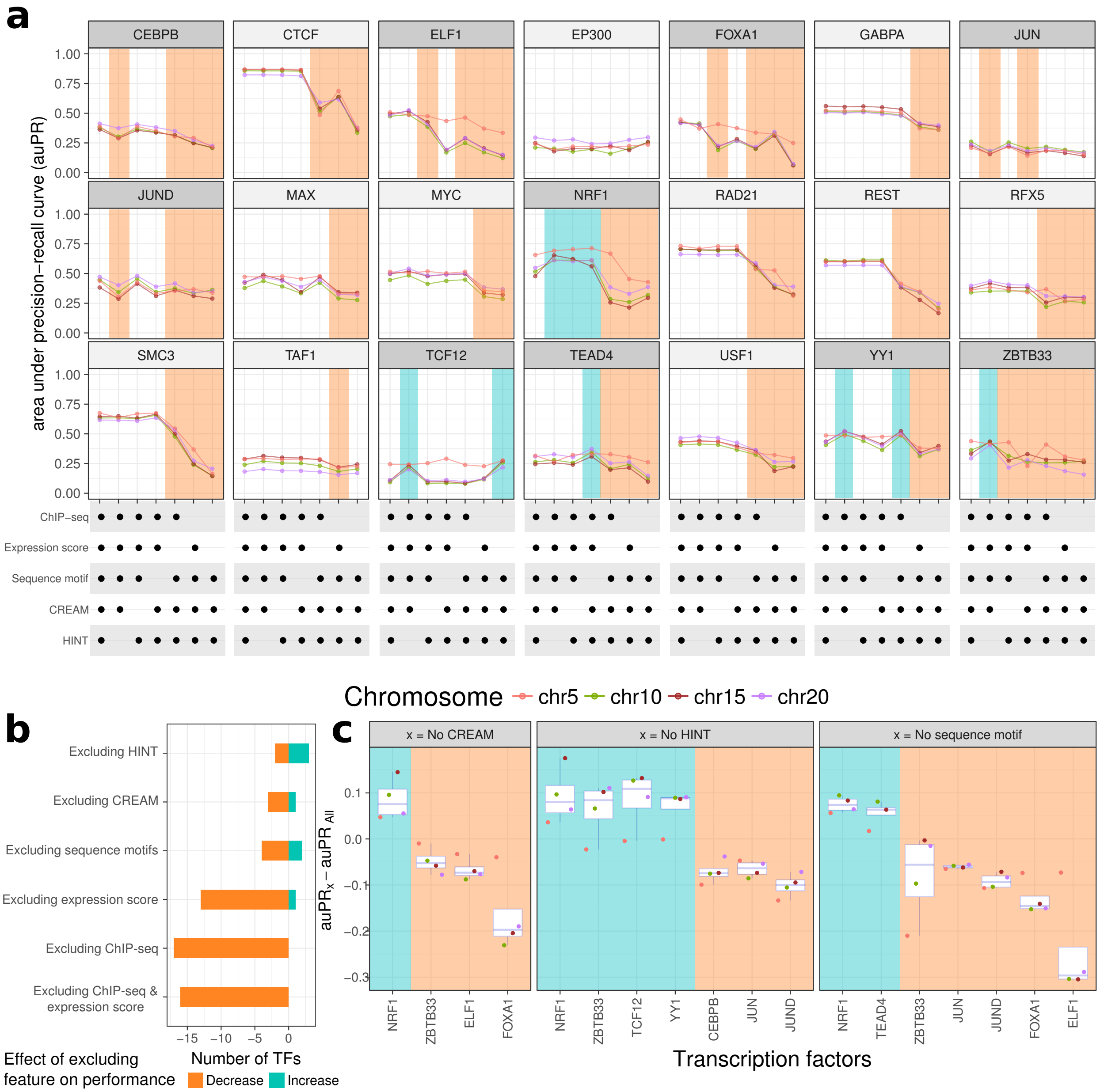

Figure 5: **Virtual ChIP-seq's most important features consist in ChIP-seq data and expression score. (a)** Area under precision-recall curve (auPR) for predicting a TF's binding sites after training on only a subset of input features. We trained on five cell types (HeLa-S3, GM12878, HCT-116, and LNCaP) and predicted on either HepG2. Ablating a feature caused either substantive decrease (orange), substantive increase (turquoise), or no substantive change in auPR. An UpSet[54]-like matrix shows the subset of features used for each column. Dark grey strip above facet: when ablating HINT, CREAM, or sequence motifs substantively changed auPR. **(b)** Double-ended bar plot of the number of TFs with average auPR increase or decrease of at least 0.05 when ablating each feature. Bars show the number of TFs where ablation caused the average auPR to decrease (orange, left) or increase (turquoise, right). **(c)** Change in auPR for those TFs with an average auPR increase or decrease of at least 0.05 when we excluded clusters of regulatory elements (CREAM), footprints (HINT), or sequence motifs. Backgrounds indicate auPR decrease (orange) or increase (turquoise).

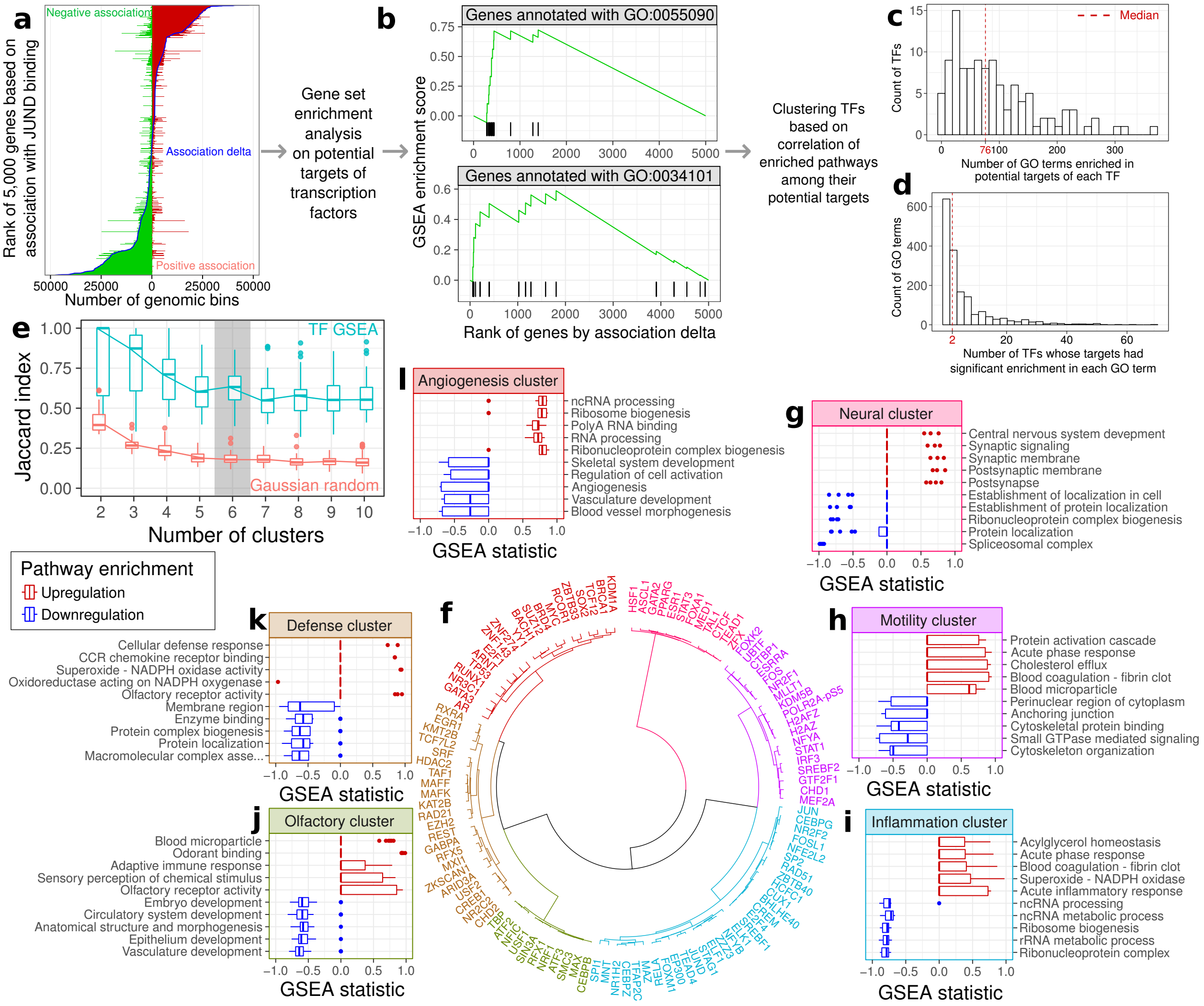

Figure 6: **Top biological pathways regulated by potential targets of TF clusters.** Each gene may have both positive and negative correlation with TF binding at some genomic bins. For each TF, we ranked 5,000 genes by an association delta that summarizes how many genomic bins associated with binding. Specifically, the association delta takes the number of bins that positively associated with a gene's expression and subtracts the number of bins that negatively associated. **(a)** Example of the association ranking process for JUND binding. Double-ended bar plot for each of the 5,000 genes, with bars for positive association (red) and negative association (green). Superimposed blue curve: association delta for each gene. **(b)** Gene Set Enrichment Analysis (GSEA)[55] identified pathways with significant enrichment in potential targets of each TF. Vertical black bars: rank of association delta for genes annotated with each GO term. Green line: GSEA enrichment score. **(c)** Histogram showing how many of 1,681 GO terms are enriched in potential targets of each TF. **(d)** Histogram showing how many of 63 TFs have potential targets with enrichment in each GO term. **(e)** Boxplot of cluster stability, as measured by Jaccard index, between clusters found in both the subsampled correlation matrix of TFs by GSEA (turquoise) and a subsampled random Gaussian matrix of the same dimensions (red). Grey background: the smallest number of clusters where cluster stability of the GSEA matrix increased but the random Gaussian matrix did not. **(f)** Dendrogram of 6 clusters identified in correlation matrix of TFs. We defined 6 clusters based on correlation of enrichment in 1,681 GO terms. **(g–l)** Boxplots of GSEA statistic for the top 5 pathways enriched in genes positively correlated with TF binding (red) and the top 5 pathways enriched in genes negatively correlated with TF binding (blue) for each cluster.

To identify such genes, for each TF, we ranked genes by subtracting the number of genomic bins they are positively correlated with from the number of genomic bins they are negatively correlated. We call this difference the *association delta*. For each TF, we identified the 5,000 genes with the highest variance in expression among cells with matched RNA-seq and ChIP-seq data (Figure 2a). We measured correlation of expression of each of the 5,000 genes with TF binding at every 100 bp genomic window in 4 chromosomes (chr5, chr10, chr15, and chr20). This approach identified genes that have consistent positive or negative association with TF binding (Figure 6a). We considered these genes as potential targets of each TF, and used the Gene Set Enrichment Analysis (GSEA) tool[55] to identify pathways with significant enrichment in either direction (Figure 6a.) Only the rank of association delta affects these results, and we presumed that there would be little difference in using all chromosomes instead of just 4. The 4-chromosome analysis for JUND had no significant rank difference from an analysis of chromosome 10 alone (Wilcoxon rank sum test $p = 0.3$). We only investigated Gene Ontology (GO) terms annotated to a minimum of 10 and a maximum of 500 out of a total of 17,106 GO-annotated genes.

We identified 1,681 GO terms with significant enrichment (GSEA $p < 0.001$) among potential targets of at least one of the 113 TFs we investigated (Figure 6b). Only 63 of these 113 TFs had matched ChIP-seq and RNA-seq in at least 5 of the training cell types and one of the validation cell types we used for learning from the transcriptome. Each TF had potential targets with significant enrichment in a mean of 92 terms (median 76; Figure 6c). Each of the 1,681 terms had significant enrichment in potential targets of a mean of 6 TFs (median 2; Figure 6d). Furthermore, 300 of these GO terms had significant enrichment in potential targets of at least 10 TFs.

To identify TFs involved in similar biological processes, we searched for enrichment of any of the 1,681 GO terms in 113 TFs. This analysis relied on the GSEA enrichment score as a normalized test statistic. We examined the pairwise correlation between the vector of enrichment scores for each pair of TFs. These pairwise correlations constitute a symmetric correlation matrix. We hypothesized that TFs with high correlation are involved in similar biological processes.

To identify groups of TFs involved in similar biological processes, we performed hierarchical clustering on the correlation matrix. We sought to identify clusters of TFs, and the best number of clusters between 2 and 10, inclusive. As a control, we generated a correlation matrix of same dimensions from a matrix of random Gaussian values (Methods). For each matrix we repeatedly generated random subsamples and clustered them. For each subsample, we found the set of pairs of TFs with the same cluster membership. For couples of these subsamples, we identified the Jaccard index between these sets as a measure of cluster stability[104] (Methods). We then compared the increase or decrease in Jaccard indices from each number of clusters to the number of clusters one larger.

The smallest number of clusters with an increase in Jaccard index only for the correlation matrix was 6 (Figure 6e–f). We assigned names to these clusters based on their enriched biological pathways. We then examined the TFs included in those clusters. The Neural cluster (Figure 6g) includes ASCL1[56], HSF1[61], GATA2[60], and PPARγ[62]. These TFs play important roles in the development of the nervous system and are implicated in neurological disorders[56,60,61,62]. The top 5 GO terms enriched in the potential targets of these TFs are all related to nervous system development and function (Figure 6g). The downregulated pathways of the Motility cluster (Figure 6h) relate to cytoskeletal organization. The included TFs, CTBP1[66], KDM5B[67], MEF2A[68], and STAT1[69], all play a role in the epithelial-to-mesenchymal transition, which involves re-organization of the cytoskeleton. Similarly, we found that for other clusters, specific upregulated or downregulated pathways of cluster's targets are also regulated by many of the cluster's TFs (Figure 6i–l, Table 2).

TF cluster	**Upregulated pathways**	**Downregulated pathways**	**TFs in cluster with relevant biology**
Neural	**Neural activity and development**	Protein biosynthesis	ASCL1[56], CTCF[57], ESR1[58], FOXA1[59], GATA2[60], HSF1[61], PPARγ[62], STAT3[63], TAL1[64], TEAD1[65]
Motility	Inflammation	**Cytoskeletal organization**	CTBP1[66], KDM5B[67], MEF2A[68], STAT1[69]
Inflammation	**Inflammation**	RNA biosynthesis	BHLHE40[70], CEBPG[71], CUX1[72], ELK1[73], FOXM1[74], JUN[75], JUND[76], RELA[77]
Olfactory	**Olfactory perception**	Vasculature, blood, and structural development	NFIC[78], ATF2[79], ATF3[80], SIN3A[81], CEBPB[82], RFX1[83]
Defense	**Cell defense and chemokine signaling**	Protein biogenesis and localization	ARID3A[84], CREB1[85], EGR1[86], KAT2B[87], KMT2B[88], MAFF[89], RFX5[90], RXRA[91], SRF[92]
Angiogenesis	RNA biosynthesis	**Angiogenesis and vasculature**	AR[93], ARNT[94], BACH1[95], BRCA1[96], BRD4[97], E2F1[98], GATA3[99], KDM1A[100], MYC[101], RUNX1[102], TP53[103]

Table 2: **Many of TFs within each biological function cluster are involved in the same pathways as their potential target genes.** We summarized each cluster of TFs according to top over-represented GO terms in the first 3 columns. TFs in the 4th column are involved in the same biological mechanism as the bold pathways mentioned in 2nd or 3rd column.

2.5 A compendium of TF binding predictions for 34 tissues and cell types

2.5.1 Predicting TF binding in Roadmap datasets

The Roadmap Epigenomics Project[32] performed DNase-seq on 55 and RNA-seq on 39 human tissues and cell types, but not ChIP-seq of any TF. For 34 of these tissues, they produced matched DNase-seq and RNA-seq data. This makes the Roadmap data an ideal application for Virtual ChIP-seq.

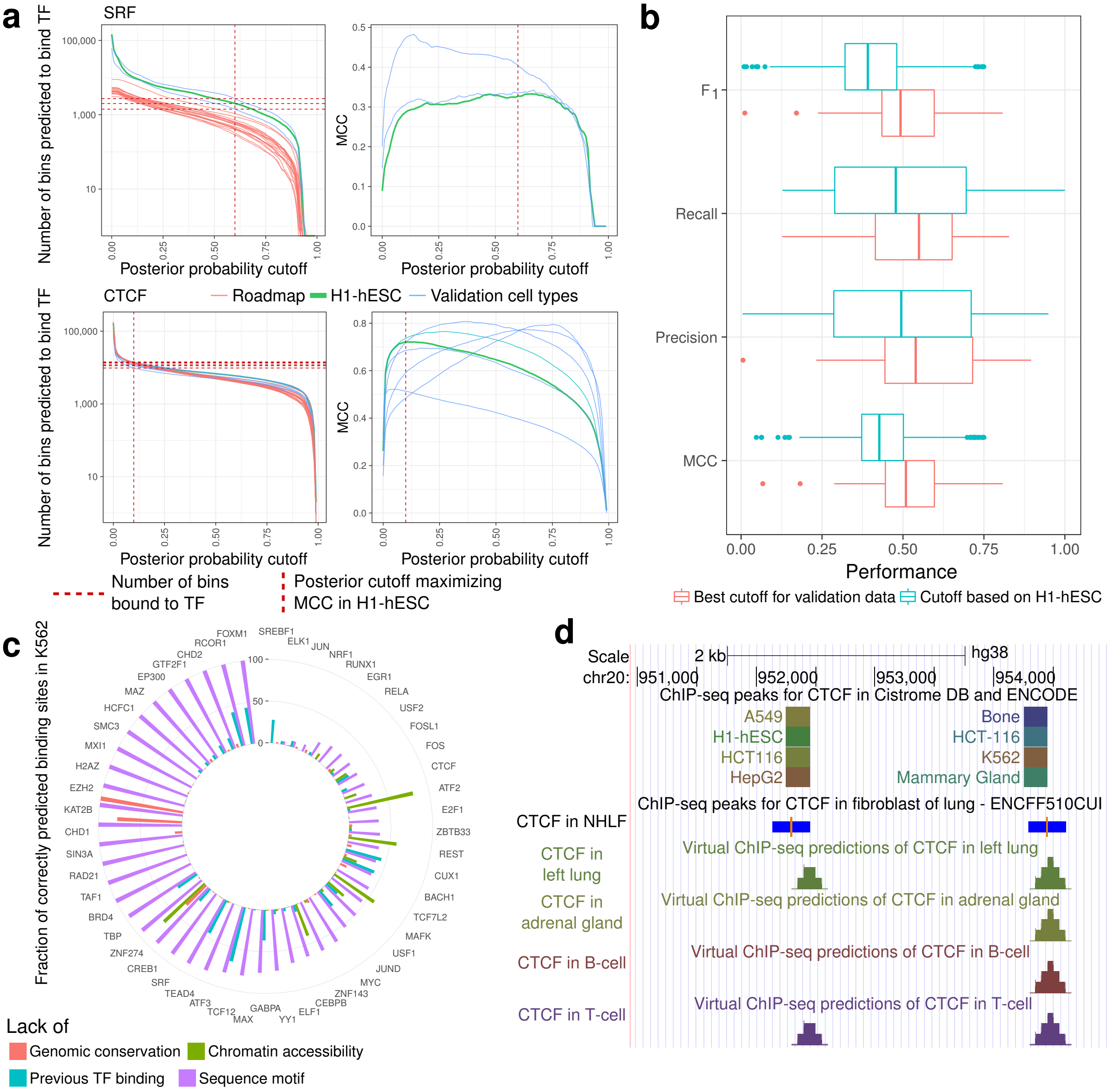

Figure 7: **TF binding predictions in validation cell types and Roadmap datasets.** **(a)** Number of genomic bins that TF is predicted to bind (left) and MCC (right) as a function of posterior probability cutoff for SRF (top) and CTCF (bottom). This relationship is shown for H1-hESC (turquoise), 2 validation cell types for SRF (blue), and 6 validation cell types for CTCF (blue). Each curve represents predictions for one of the 4 chromosomes (chr5, chr10, chr15, and chr20). Left panels also show how many genomic bins are predicted to bind the TF in 18 Roadmap datasets (red). Vertical red dashed line: posterior probability cutoff which maximized MCC of the TF in H1-hESC. Horizontal red dashed lines: number of genomic bins with TF binding in validation cell types. **(b)** Boxplot of various performance measures when using the best cutoff for each dataset (red) and the optimal cutoff in H1-hESC (turquoise). **(c)** Bar plot of the fraction of binding sites for 52 TFs correctly predicted on K562 chromosome 5 which lacked particular predictive features. These features include genomic conservation (red), chromatin accessibility (green), sequence motif (turquoise), and evidence of TF binding in another cell type (purple). For TFs with no sequence motif, we deemed every binding site to lack a sequence motif. **(d)** UCSC Genome Browser display of a 4000 bp region on *(Continued on next page.)*

Figure 7: *(Continued from previous page.)* chromosome 20 using the Virtual ChIP-seq track hub (`https://virchip.hoffmanlab.org`). The track hub has a supertrack for each TF. Each supertrack contains 35 tracks: one track specifying genomic bins bound by that TF in Cistrome and ENCODE, and one track for each of the 34 Roadmap cell types with predictions for that TF. This example shows parts of the track hub related to CTCF, including a track with experimental results in Cistrome DB and ENCODE with 7 out of 144 cell types enabled, and Virtual ChIP-seq predictions in left lung, adrenal gland, B-cell, and T-cell. The height of predictions indicates the number of overlapping genomic bins predicted to bind the TF, ranging between 0–4. Between are MACS2 narrow peak calls for CTCF in normal human lung fibroblasts (NHLF) from ENCODE (ENCFF510CUI). Blue: peaks; orange: peak summits.

We generated an annotation similar to peak calls by converting the MLP's posterior probabilities to a presence or absence call. We made this call based on a different cutoff for each TF. We defined this cutoff as the posterior probability which maximized MCC in H1-hESC. For TFs without ChIP-seq data in H1-hESC, we used the mode of cutoffs from the other different TFs (0.4). We excluded H1-hESC when reporting all performance metrics that depend on this threshold. The number of binding sites we predicted in other validation cell types and Roadmap data is similar to ChIP-seq peaks in other validation cell types (Figure 7a).

Using the cutoff which maximized MCC in H1-hESC only slightly decreased performance measurements from what one could achieve with the optimal cutoff for each cell type (Figure 7b). For example, the MCC score showed a median decrease of 0.06 and F_1 score showed a median decrease of 0.1.

Narrowing predictions to only those that pass the cutoff, we found that many correctly predicted binding sites in K562 lack important predictive features of TF binding (Figure 7c). For example, many of the correctly predicted binding sites of EZH2 and KAT2B are not conserved among placental mammals. Many correctly predicted binding sites for MAFK, REST, FOSL1, and CTCF don't overlap chromatin accessibility peaks. We correctly predicted many binding sites for TCF12, RCOR1, TEAD4, CHD1, FOXM1, GABPA, and CUX1 in regions that have no binding in other cell types. In these cases, MLP learned from other available predictive features. For example, in RCOR1, all novel correctly predicted binding sites of chromosome 5 overlapped chromatin accessibility peaks. These correct predictions also had an average genomic conservation of 0.19 which was significantly higher than other genomic bins (Welch t-test $p = 0.006$).

As a community resource, we created a public track hub (`https://virchip.hoffmanlab.org`) with predictions for 34 Roadmap cell types (Figure 7d). This track hub contains predictions for 31 TFs which had a median MCC > 0.3 in validation cell types (Table 1).

3 Methods

3.1 Data used for prediction

3.1.1 Overlapping genomic bins

To generate the input matrix for training and validation, we used 200 bp genomic bins with sliding 50 bp windows. We excluded any genomic bin which overlaps with ENCODE blacklist regions (`https://www.encodeproject.org/files/ENCFF419RSJ/@@download/ENCFF419RSJ.bed.gz`). Except where otherwise specified, we used the Genome Reference Consortium GRCh38/hg38 assembly[46].

3.1.2 Chromatin accessibility

We used Cistrome DB ATAC-seq and DNase-seq narrowPeak files for assessing chromatin accessibility (Supplementary Table 8). We mapped the signal value of peak summits to all the bins overlapping that summit. In rare cases where a genomic bin overlaps more than one summit, we used the signal value of the summit closest to the p terminus of the chromosome When data were available from multiple experiments, we averaged signal values. Because Cistrome DB does not include raw data that one can use for DNase footprinting, we limited the analysis of HINT TF footprinting and CREAM regulatory element clustering to ENCODE DNase-seq experiments on GM12878, HCT-116, HeLa-S3, LNCaP, and HepG2.

3.1.3 Genomic conservation

We used GRCh38 primate and placental mammal 7-way PhastCons genomic conservation[43,44] scores from the UCSC Genome Browser[105] (http://hgdownload.cse.ucsc.edu/goldenPath/hg38/phastCons7way). We assigned each bin the mean PhastCons score of the nucleotides within.

3.1.4 Sequence motif score

We used FIMO[33] (version 4.11.2) to search for motifs from JASPAR 2016[106] to identify binding sites of each TF that have the sequence motif of that TF. To get a liberal set of motif matches, we used a liberal p-value threshold of 0.001 and didn't adjust for multiple testing. If the motif for the TF didn't exist in JASPAR, we used other motifs with same initial 3 letters and counted any TF binding site which had overlap with any of those motifs (Supplementary Table 1).

We also used FIMO and JASPAR 2016 to identify the sequence specificity of chromatin accessible regions. For this analysis, we used a false discovery rate threshold of 0.01%. We used any sequence motif matching the initial 3 letters of a TF as a predictive feature of binding for that TF. For many TFs, more than one motif matched this criteria, and we used all as independent features in the model (Supplementary Table 2).

3.1.5 ChIP-seq data

We used Cistrome DB and ENCODE ChIP-seq narrowPeak files. We only used peaks with FDR $< 10^{-4}$. When multiple replicates of the same experiment existed, we only considered peaks that passed the FDR threshold in at least two replicates. We considered bound only those genomic bins overlapping peak summits. We calculated prevalence of bound bins in each chromosome as

$$\text{prevalence} = \frac{\text{bound}}{\text{bound} + \text{unbound}}$$

and used it as an auPR baseline[25].

3.1.6 RNA-seq data

We downloaded an ENCODE expression matrix (https://public-docs.crg.es/rguigo/encode/expressionMatrices/H.sapiens/hg19/2014_10/gencodev19_genes_with_RPKM_and_npIDR_oct2014.txt.gz)[37] with RNA-seq data for each gene, measured in reads per kilobase per million mapped reads (RPKM). We retrieved similar Cancer Cell Line Encyclopedia (CCLE) RNA-seq data using PharmacoGx[107]. Since these data are processed differently, we limited our analysis to Ensembl gene IDs shared between the two datasets, and ranked gene expression values by cell type. The two datasets have 4 shared cell types: A549, HepG2, K562, and MCF-7. Within each of these cell types, we examined the concordance of RNA-seq data between ENCODE and CCLE after possible transformations. The concordance correlation coefficient[108] of rank of RPKM (0.827) was higher compared to untransformed RPKM (0.007) or quantile-normalized RPKM (0.006; Welch t-test $p = 10^{-6}$). The DREAM Challenge, however, had processed RNA-seq of all cell types uniformly, allowing us to directly use transcripts per million reads (TPM) in analysis of DREAM Challenge datasets.

3.1.7 Expression score

We created an expression matrix for each TF with matched ChIP-seq and RNA-seq data in $N \geq 5$ training cell types with the following procedure:

1. We divided the genome into M 100 bp non-overlapping genomic bins.

2. We created a non-negative ChIP-seq matrix $\boldsymbol{C} \in \mathbb{R}_{\geq 0}^{M \times N}$ (Figure 2a). We used signal mean among replicate narrowPeak files generated by MACS2[109] for each of M bins and N cell types and quantile-normalized this matrix.

3. We row-normalized $\boldsymbol{C}$ to $\boldsymbol{C'}$, scaling the values of each row between 0 and 1.

4. We identified the $G = 5000$ genes with the highest variance among the N cell types.

5. We created an expression matrix $\mathbf{E} \in \mathbb{R}_{\in[0,1]}^{N \times G}$ containing the row-normalized rank of expression each of the $G = 5000$ genes in N cell types (Figure 2b).

6. For each bin $i \in [1, M]$ and each gene $g \in [1, G]$, we calculated the Pearson correlation coefficient $A_{i,g}$ between the ChIP-seq data for that bin $\boldsymbol{C'}_{i,:}$ and the expression ranks for that gene $\boldsymbol{E}_{:,j}$ over all cell types. If the Pearson correlation was not significant ($p > 0.1$), we set $A_{i,g}$ to NA. These coefficients constitute an association matrix $\boldsymbol{A} \in (\mathbb{R}_{\in[-1,1]} \cup \{\text{NA}\})^{M \times G}$ (Figure 2c).

We performed power analysis of the Pearson correlation test using the R pwr package[110].

To predict ChIP-seq binding for a new cell type (Figure 2d), we calculated an expression score for each genomic bin in that cell type. The expression score is Spearman's ρ for expression of the same $G = 5000$ genes in the new cell type with every row of the association matrix $\boldsymbol{A}$. Each of these rows represents a single genomic bin. An expression score close to 1 indicates that genes with high expression have high values in the association matrix, and genes with low expression genes have low values. An expression score close to -1 indicates that genes with high or low expression have opposite values in the association matrix (Figure 2d).

3.2 Training, optimization, and benchmarking

3.2.1 Training and optimization

For the purpose of training and validating the model on Cistrome datasets, we only used chromosomes 5, 10, 15, and 20. These 4 chromosomes constitute 481.78 Mbp (15.6% of the genome). For training only, we excluded any genomic region without chromatin accessibility signal and previous evidence of TF binding. For validation and reporting performance, we included these regions, using the totality of the 4 chromosomes. We concatenated data from training cell types (A549, GM12878, HepG2, HeLa-S3, HCT-116, BJ, Jurkat, NHEK, Raji, Ishikawa, LNCaP, and T47D; Supplementary Table 3) into the training matrix.

We used Python 2.7.13, Scikit-learn 0.18.1[111], NumPy 1.11.0, and Pandas 0.19.2 for processing data and training classifiers.

We optimized hyperparameters of the multi-layer perceptron (MLP)[47] using grid search and 4-fold cross validation. We used minibatch training with 200 genomic bins in each minibatch. We searched for several options to optimize the activation function (Figure 3g), number of hidden units per hidden layer (Figure 3h), number of hidden layers (Figure 3i), and L_2 regularization penalty (Figure 3j). In each round of 4-fold cross-validation, we trained on data of 3 chromosomes, and assessed best MCC on the remaining chromosome. We selected the set of hyperparameters yielding highest average MCC after 4-fold cross validation.

3.2.2 Benchmarking

We used the R precrec package[112] to calculate auPR and auROC. Precision-recall curves better assess a binary classifier's performance on imbalanced test data than ROC[25,50].

3.2.3 DREAM Challenge comparison

For comparison to DREAM results, we also trained and validated the Virtual ChIP-seq model on GRCh37 DREAM Challenge data. For training the model on DREAM Challenge datasets, we used the data of chr5, chr10, chr15, and chr20 of training cell types. We evaluated performance against the union of the DREAM validation chromosomes (chr1, chr8, and chr21) in validation cell types. For CTCF, we trained on all cell types except MCF-7, PC-3, and iPSC which we used for validation. For MAX, we used all cell types except liver and K562 for training. For GABPA, REST, and JUND, we used all cell types except liver for training. We compared these metrics to those of DREAM Challenge participants in the final round of cross–cell-type competition.

3.3 Clustering TFs based on enrichment of their potential targets in GO terms

To identify groups of TFs involved in similar biological processes, we performed hierarchical clustering on the correlation matrix. We sought to identify clusters of TFs, and the best number of clusters between 2 and 10, inclusive. For use in this process, we created a Gaussian random matrix of 1,681 rows and 113 columns as a control, and calculated its correlation matrix. Then, we compared cluster stability between the original correlation matrix and the control for each potential number of clusters. To do this, we subsampled 75% of each correlation matrix rows twice without replacement. Then, we clustered TFs in each matrix into the specified number clusters. For both of these clusterings, we constructed the set of every pair of TFs present in the same cluster. We then calculated the Jaccard index between the first clustering's constructed set and that of the second[104]. We repeated this subsampling and clustering process 50 times for each number of clusters. We picked the smallest number of clusters which had an increase in Jaccard index compared to the number of clusters one smaller only in the TF correlation matrix.

3.4 TF prediction on Roadmap data

We downloaded Roadmap DNase-seq and RNA-seq data aligned to GRCh38 from the ENCODE DCC[32]. For each DNase-seq narrowPeak file with matched RNA-seq, we predicted binding of 31 TFs with MCC > 0.3 in validation cell types (Table 1, Supplementary Table 6, https://virchip.hoffmanlab.org).

4 Discussion

Performing functional genomics assays to assess binding of all TFs may never be possible in patient tissues. Nevertheless, computational prediction of TF binding based on sequence specificity of TFs has identified the role of many TFs in various diseases[1]. Scanning the genome for occurrences of each sequence motif, results in a range of 200–2000 predictions/Mbp. In some cases, this is 1,000 times more frequent than experimental data from ChIP-seq peaks. Similar observations led to a *futility conjecture* that almost all TF binding sites predicted in this way will have no functional role[113].

Nevertheless, there is more to TF binding than sequence preference. Most TFs don't have any sequence preference[9] (Figure 1), and indirect TF binding through complexes of chromatin-binding proteins complicates predictions based solely on sequence specificity. In addition to the high number of false positive motif occurrences, many ChIP-seq peaks lack the TF's sequence motif. Therefore, relying on sequence specificity alone not only generates too many false positives, but also many false negatives. We call this latter observation the *dual futility conjecture*, although it differs in degree from the original. Adding additional data about cellular state allows us to move beyond both conjectures.

We can assess TF binding through ChIP-seq or its more precise variations ChIP-nexus[12] or ChIP-exo[11]. These experiments may still not properly reflect *in vivo* TF binding due to technical difficulties such as non-specific or low affinity antibodies. Using publicly available ChIP-seq data produced with different protocols and reagents, complicates prediction of TFs more sensitive to experimental conditions[52]. Variations among training and validation cell types in our datasets, overfitted the MLP to certain input features of some TFs. More robust approaches in assessment of TF binding—such as CRISPR epitope tagging ChIP-seq (CETCh-seq)[114], which doesn't rely on specific antibodies—may provide less noisy reference data for learning and prediction of TF binding.

Virtual ChIP-seq predicted binding of 31 TFs in new cell types, using from the new cell types only chromatin accessibility and transcriptome data. By learning from direct evidence of TF binding and the association of the transcriptome with TF binding at each genomic region, most use of sequence motif scores becomes redundant. As more ChIP-seq data in diverse cell types and tissues becomes available, our approach allows predicting binding of more TFs with high accuracy. This is true even in the case of factors that are not sequence-specific. Although Virtual ChIP-seq uses direct evidence of TF binding at each genomic region as one of the input features, it is able to correctly predict new peaks which don't exist in training cell types. For 39 of 41 sequence specific TFs, Virtual ChIP-seq correctly predicted TF binding in regions without any match to sequence motifs.

The DREAM Challenge datasets provide data for training and validating machine learning models for predicting binding of 31 TFs. Our datasets, using a combination of Cistrome DB and ENCODE, allow training and validating models for predicting binding in a more extensive 63 TFs. Our provided predictions of binding of 31 high-confidence TFs in 34 different Roadmap tissue types will allow the research community to better investigate epigenomics of disease affecting those tissues (`https://virchip.hoffmanlab.org/`). In addition to providing our predictions as a resource for use by biologists, we also provide the processed datasets we use as a resource for machine learning researchers. This should accelerate the development of future methods by many groups.

Acknowledgments

We thank Shirley X. Liu for providing us with the Cistrome DB narrowPeak files. We thank the Roadmap Epigenomics Mapping Consortium and the ENCODE Project Consortium for generating the datasets which enabled this work. We thank Sage Bionetworks-DREAM and the ENCODE-DREAM Challenge organizers for providing data and results before publication. We thank Carl Virtanen and Zhibin Lu (University Health Network High Performance Computing Centre and Bioinformatics Core) for technical assistance. This work was supported by the Canadian Cancer Society (703827 to M.M.H.), the Ontario Ministry of Training, Colleges and Universities (Ontario Graduate Scholarship to M.K.), and the University of Toronto Faculty of Medicine Frank Fletcher Memorial Fund (M.K.).

Competing interests

The authors declare that they have no competing interests.

References

[1] Tong Ihn Lee and Richard A. Young. Transcriptional regulation and its misregulation in disease. *Cell*, 152(6):1237–1251, 2013.

[2] Pamela J. Mitchell and Robert Tjian. Transcriptional regulation in mammalian cells. *Science*, 245:371–378, 1989.

[3] Remo Rohs, Sean M. West, Alona Sosinsky, Peng Liu, Richard S. Mann, and Barry Honig. The role of DNA shape in protein-DNA recognition. *Nature*, 461(7268):1248, 2009.

[4] Susan Jones, Paul van Heyningen, Helen M. Berman, and Janet M. Thornton. Protein-DNA interactions: a structural analysis. *Journal of Molecular Biology*, 287(5):877–896, 1999.

[5] Nobuo Ogawa and Mark D. Biggin. High-throughput SELEX determination of DNA sequences bound by transcription factors in vitro. *Gene Regulatory Networks: Methods and Protocols*, pages 51–63, 2012.

[6] Martha L. Bulyk. Protein binding microarrays for the characterization of DNA–protein interactions. In *Analytics of Protein–DNA Interactions*, pages 65–85. Springer, 2006.

[7] Sachi Inukai, Kian Hong Kock, and Martha L. Bulyk. Transcription factor–DNA binding: beyond binding site motifs. *Current Opinion in Genetics & Development*, 43:110–119, 2017.

[8] Matthew T. Weirauch, Atina Cote, Raquel Norel, Matti Annala, et al. Evaluation of methods for modeling transcription factor sequence specificity. *Nature Biotechnology*, 31(2):126–134, 2013.

[9] Md. Abul Hassan Samee, Benoit Bruneau, and Katherine Pollard. Transcription factors recognize DNA shape without nucleotide recognition. *bioRxiv*, 2017. doi: 10.1101/143677.

[10] David S. Johnson, Ali Mortazavi, Richard M. Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830):1497–1502, 2007.

[11] Ho Sung Rhee and B. Franklin Pugh. ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. *Current Protocols in Molecular Biology*, pages 21–24, 2012.

[12] Qiye He, Jeff Johnston, and Julia Zeitlinger. ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nature Biotechnology*, 33(4):395–401, 2015.

[13] Jason D. Buenrostro, Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10(12):1213–1218, 2013.

[14] Roger Pique-Regi, Jacob F. Degner, Athma A. Pai, Daniel J. Gaffney, et al. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research*, 21(3):447–455, 2011.

[15] Eduardo G. Gusmao et al. Analysis of computational footprinting methods for DNase sequencing experiments. *Nature Methods*, 2016.

[16] Xi Chen, Bowen Yu, Nicholas Carriero, Claudio Silva, and Richard Bonneau. Mocap: Large-scale inference of transcription factor binding sites from chromatin accessibility. *Nucleic Acids Research*, 45(8):4315, 2017.

[17] Housheng Hansen He, Clifford A. Meyer, Hyunjin Shin, Shannon T. Bailey, et al. Nucleosome dynamics define transcriptional enhancers. *Nature Genetics*, 42(4):343–347, 2010.

[18] Swneke D. Bailey, Kinjal Desai, Ken J. Kron, Parisa Mazrooei, et al. Noncoding somatic and inherited single-nucleotide variants converge to promote *ESR1* expression in breast cancer. *Nature Genetics*, 48(10):1260–1266, 2016.

[19] Timothy L. Bailey and Philip Machanick. Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Research*, 40(17): e128–e128, 2012.

[20] Jian Zhou and Olga G. Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10):931, 2015.

[21] Daniel Quang and Xiaohui Xie. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Research*, 44(11):e107–e107, 2016.

[22] Wei-Li Guo and De-Shuang Huang. An efficient method to transcription factor binding sites imputation via simultaneous completion of multiple matrices with positional consistency. *Molecular BioSystems*, 13:1827–1837, 2017.

[23] Qian Qin and Jianxing Feng. Imputation for transcription factor binding predictions based on deep learning. *PLOS Computational Biology*, 13(2):e1005403, 2017.

[24] Richard I. Sherwood, Tatsunori Hashimoto, Charles W. O'Donnell, Sophia Lewis, et al. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nature Biotechnology*, 32(2):171–178, 2014.

[25] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS One*, 10(3):e0118432, 2015.

[26] Travers Ching, Daniel S. Himmelstein, Brett K. Beaulieu-Jones, Alexandr A. Kalinin, et al. Opportunities and obstacles for deep learning in biology and medicine. *bioRxiv*, 2017. doi: 10.1101/142760.

[27] ENCODE-DREAM *in vivo* Transcription Factor Binding Site Prediction Challenge. https://synapse.org/encode, 2017. Accessed: 2018-01-31.

[28] Michael A. Beer and Saeed Tavazoie. Predicting gene expression from sequence. *Cell*, 117(2):185–198, 2004.

[29] Zhengqing Ouyang, Qing Zhou, and Wing Hung Wong. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proceedings of the National Academy of Sciences*, 106(51):21521–21526, 2009.

[30] David R. Kelley and Yakir A. Reshef. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *bioRxiv*, 2017. doi: 10.1101/161851.

[31] Weiqiang Zhou, Ben Sherwood, Zhicheng Ji, Yingchao Xue, et al. Genome-wide prediction of DNase I hypersensitivity using gene expression. *Nature Communications*, 8(1):1038, 2017.

[32] Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015.

[33] Charles E. Grant et al. FIMO: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, 2011.

[34] Shenglin Mei, Qian Qin, Qiu Wu, Hanfei Sun, et al. Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Research*, 45(D1):D658–D662, 2017.

[35] ENCODE Project Consortium. An integrated Encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[36] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, et al. The Cancer Cell Line Encyclopedia enables predictive modeling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, 2012.

[37] Sarah Djebali, Carrie A Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, et al. Landscape of transcription in human cells. *Nature*, 489(7414):101, 2012.

[38] Nathan C. Sheffield et al. Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome Research*, 23(5):777–788, 2013.

[39] Robert E. Thurman, Eric Rynes, Richard Humbert, Jeff Vierstra, et al. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, 2012.

[40] Jeffrey Rogers and Richard A. Gibbs. Comparative primate genomics: emerging patterns of genome content and dynamics. *Nature Reviews Genetics*, 15(5):347–359, 2014.

[41] Moshe Pritsker, Yir-Chung Liu, Michael A. Beer, and Saeed Tavazoie. Whole-genome discovery of transcription factor binding sites by network-level conservation. *Genome Research*, 14(1):99–108, 2004.

[42] Eugene Berezikov, Victor Guryev, and Edwin Cuppen. Exploring conservation of transcription factor binding sites with CONREAL. *Methods in Molecular Biology*, 395:437–448, 2007.

[43] Adam Siepel, Gill Bejerano, Jakob S. Pedersen, Angie S. Hinrichs, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8):1034–1050, 2005.

[44] Katherine S. Pollard, Melissa J. Hubisz, Kate R. Rosenbloom, and Adam Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20(1):110–121, 2010.

[45] Seyed Ali Madani Tonekaboni, Parisa Mazrooei, Victor Kofia, Benjamin Haibe-Kains, and Mathieu Lupien. CREAM: Clustering of genomic REgions Analysis Method. *bioRxiv*, 2017. doi: 10.1101/222562.

[46] Valerie A. Schneider, Tina Graves-Lindsay, Kerstin Howe, Nathan Bouk, et al. Evaluation of GRCh38 and *de novo* haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, 27(5):849–864, 2017.

[47] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010.

[48] Diederik P. Kingma and Jimmy Ba. Adam: a method for stochastic optimization. *arXiv*, abs/1412.6980, 2014. arxiv.org/abs/1412.6980.

[49] Brian W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.

[50] Davide Chicco. Ten quick tips for machine learning in computational biology. *BioData Mining*, 10(1):35, 2017.

[51] Qunhua Li, James B. Brown, Haiyan Huang, Peter J. Bickel, et al. Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics*, 5(3):1752–1779, 2011.

[52] Daniel Quang and Xiaohui Xie. FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. *bioRxiv*, 2017. doi: 10.1101/151274.

[53] Jens Keilwagen, Stefan Posch, and Jan Grau. Learning from mistakes: accurate prediction of cell type-specific transcription factor binding. *bioRxiv*, 2017. doi: 10.1101/230011.

[54] Alexander Lex, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, and Hanspeter Pfister. UpSet: visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1983–1992, 2014.

[55] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43): 15545–15550, 2005.

[56] Alexandre A. Raposo, Francisca F. Vasconcelos, Daniela Drechsel, Corentine Marie, et al. Ascl1 coordinately regulates gene expression and the chromatin landscape during neurogenesis. *Cell Reports*, 10(9):1544–1556, 2015.

[57] L. Ashley Watson, Xu Wang, Adrienne Elbert, Kristin D. Kernohan, et al. Dual effect of CTCF loss on neuroprogenitor differentiation and survival. *Journal of Neuroscience*, 34(8):2860–2870, 2014.

[58] Elise Lamar and Chris Kintner. The Notch targets *Esr1* and *Esr10* are differentially regulated in *Xenopus* neural precursors. *Development*, 132(16):3619–3630, 2005.

[59] Anna L. M. Ferri, Wei Lin, Yannis E. Mavromatakis, Julie C. Wang, et al. Foxa1 and Foxa2 regulate multiple phases of midbrain dopaminergic neuron development in a dosage-dependent manner. *Development*, 134(15):2761–2769, 2007.

[60] Ryan T. Willett and Lloyd A. Greene. Gata2 is required for migration and differentiation of retinorecipient neurons in the superior colliculus. *Journal of Neuroscience*, 31(12):4444–4455, 2011.

[61] Seiji Ishii and Kazue Hashimoto-Torii. HSF modulates neural development under normal and stress conditions. In *Heat Shock Factor*, pages 115–129. Springer, 2016.

[62] Rodrigo A. Quintanilla, Elias Utreras, and Fabián A. Cabezas-Opazo. Role of PPARγ in the differentiation and function of neurons. *PPAR Research*, 2014, 2014.

[63] Seunghee Lee, Rongkun Shen, Hyong-Ho Cho, Ryuk-Jun Kwon, et al. STAT3 promotes motor neuron differentiation by collaborating with motor neuron-specific LIM complex. *Proceedings of the National Academy of Sciences*, 110(28):11445–11450, 2013.

[64] Kaia Achim, Paula Peltopuro, Laura Lahti, Hui-Hsin Tsai, et al. The role of *Tal2* and *Tal1* in the differentiation of midbrain GABAergic neuron precursors. *Biology Open*, 2(10):990–997, 2013.

[65] Xinwei Cao, Samuel L. Pfaff, and Fred H. Gage. YAP regulates neural progenitor cell number via the TEA domain transcription factor. *Genes & Development*, 22(23):3320–3334, 2008.

[66] Xiao-Ling Zhang, Cheng-Xin Huang, Jie Zhang, Akira Inoue, et al. CtBP1 is involved in epithelial-mesenchymal transition and is a potential therapeutic target for hepatocellular carcinoma. *Oncology Reports*, 30(2):809–814, 2013.

[67] Zanabazar Enkhbaatar, Minoru Terashima, Dulamsuren Oktyabri, Shoichiro Tange, et al. KDM5B histone demethylase controls epithelial-mesenchymal transition of cancer cells by regulating the expression of the microRNA-200 family. *Cell Cycle*, 12(13): 2100–2112, 2013.

[68] Wei Yu, Changshan Huang, Qian Wang, Tao Huang, et al. MEF2 transcription factors promotes EMT and invasiveness of hepatocellular carcinoma through TGF-β1 autoregulation circuitry. *Tumor Biology*, 35(11):10943–10951, 2014.

[69] Puja Kachroo, Mi-Heon Lee, Ling Zhang, Felicita Baratelli, et al. IL-27 inhibits epithelial-mesenchymal transition and angiogenic factor production in a STAT1-dominant pathway in human non-small cell lung cancer. *Journal of Experimental & Clinical Cancer Research*, 32(1):97, 2013.

[70] Chih-Chung Lin, Tara R. Bradstreet, Elizabeth A. Schwarzkopf, Julia Sim, et al. Bhlhe40 controls cytokine production by T cells and is essential for pathogenicity in autoimmune neuroinflammation. *Nature Communications*, 5:3551, 2014.

[71] Christopher J. Huggins, Radek Malik, Sook Lee, Jacqueline Salotti, et al. C/EBPγ suppresses senescence and inflammatory gene expression by heterodimerizing with C/EBPβ. *Molecular and Cellular Biology*, 33(16):3242–3258, 2013.

[72] Mathieu Darsigny, Stéphanie St-Jean, and François Boudreau. Cux1 transcription factor is induced in inflammatory bowel disease and protects against experimental colitis. *Inflammatory Bowel Diseases*, 16(10):1739–1750, 2010.

[73] Aneta Kasza, Paulina Wyrzykowska, Irena Horwacik, Piotr Tymoszuk, et al. Transcription factors Elk-1 and SRF are engaged in IL1-dependent regulation of ZC3H12A expression. *BMC Molecular Biology*, 11(1):14, 2010.

[74] David Balli, Xiaomeng Ren, Fu-Sheng Chou, Emily Cross, et al. Foxm1 transcription factor is required for macrophage migration during lung inflammation and tumor formation. *Oncogene*, 31(34):3875–3888, 2012.

[75] Bozena Kaminska. Molecular characterization of inflammation-induced JNK/c-Jun signaling pathway in connection with tumorigenesis. *Methods in Moleular Biology*, 512:249–264, 2009.

[76] H. Terence Cook, Ruth Tarzi, Zelpha D'Souza, Gaelle Laurent, et al. AP-1 transcription factor JunD confers protection from accelerated nephrotoxic nephritis and control podocyte-specific Vegfa expression. *The American Journal of Pathology*, 179 (1):134–140, 2011.

[77] Samaneh Yazdani, Mohammad Hasan Karimfar, Abbas Ali Imani Fooladi, Leila Mirbagheri, et al. Nuclear factor κB1/RelA mediates the inflammation and/or survival of human airway exposed to sulfur mustard. *Journal of Receptors and Signal Transduction*, 31(5):367–373, 2011.

[78] Hany E. S. Marei and Abd-Elmaksoud Ahmed. Transcription factors expressed in embryonic and adult olfactory bulb neural stem cells reveal distinct proliferation, differentiation and epigenetic control. *Genomics*, 101(1):12–19, 2013.

[79] Mercedes Lachn-Montes, Andrea Gonzlez-Morales, Maria Victoria Zelaya, Estela Prez-Valderrama, et al. Olfactory bulb neuroproteomics reveals a chronological perturbation of survival routes and a disruption of prohibitin complex during Alzheimer's disease progression. *Scientific Reports*, 7:9115, 2017.

[80] Shreelatha Bhat and Walton D. Jones. An accelerated miRNA-based screen implicates Atf-3 in *Drosophila* odorant receptor expression. *Scientific Reports*, 6:20109, 2016.

[81] Josefine S. Witteveen, Marjolein H. Willemsen, Thaís C. D. Dombroski, Nick H. M. Van Bakel, et al. Haploinsufficiency of MeCP2-interacting transcriptional co-repressor SIN3A causes mild intellectual disability by affecting the development of cortical integrity. *Nature Genetics*, 48(8):877–887, 2016.

[82] Adele J. Vincent, Jennifer M. Taylor, Derek L. Choi-Lundberg, Adrian K. West, and Meng Inn Chuah. Genetic expression profile of olfactory ensheathing cells is distinct from that of Schwann cells and astrocytes. *Glia*, 51(2):132–147, 2005.

[83] Chenzhuo Feng, Jiejie Li, and Zhiyi Zuo. Expression of the transcription factor regulatory factor X1 in the mouse brain. *Folia Histochemica et Cytobiologica*, 49(2):344–351, 2011.

[84] Julie M. Ward, Kira Rose, Courtney Montgomery, Indra Adrianto, et al. Disease activity in systemic lupus erythematosus correlates with expression of the transcription factor AT-rich–interactive domain 3A. *Arthritis & Rheumatology*, 66(12):3404–3412, 2014.

[85] Andy Y. Wen, Kathleen M. Sakamoto, and Lloyd S. Miller. The role of the transcription factor CREB in immune function. *The Journal of Immunology*, 185(11):6413–6419, 2010.

[86] Steven B. McMahon and John G. Monroe. The role of early growth response gene 1 (EGR-1) in regulation of the immune response. *Journal of Leukocyte Biology*, 60(2):159–166, 1996.

[87] Atsuko Masumi, I-Ming Wang, Bruno Lefebvre, Xing-Jiao Yang, et al. The histone acetylase PCAF is a phorbol-ester-inducible coactivator of the IRF family that confers enhanced interferon responsiveness. *Molecular and Cellular Biology*, 19(3):1810–1820, 1999.

[88] Chia-Hsin Su, I-Hsuan Lin, Tsai-Yu Tzeng, Wen-Ting Hsieh, and Ming-Ta Hsu. Regulation of IL-20 expression by estradiol through KMT2B-mediated epigenetic modification. *PLOS One*, 11(11):e0166090, 2016.

[89] Wael Massrieh, Anna Derjuga, Florence Doualla-Bell, Chun-Ying Ku, et al. Regulation of the MAFF transcription factor by proinflammatory cytokines in myometrial cells. *Biology of Reproduction*, 74(4):699–705, 2006.

[90] Jean Villard, Marie Peretti, Krzysztof Masternak, Emmanuèle Barras, et al. A functionally essential domain of RFX5 mediates activation of major histocompatibility complex class II promoters by promoting cooperative binding between RFX and NF-Y. *Molecular and Cellular Biology*, 20(10):3364–3376, 2000.

[91] Feng Ma, Su-Yang Liu, Bahram Razani, Neda Arora, et al. Retinoid X receptor α attenuates host antiviral response by suppressing type I interferon. *Nature Communications*, 5:5494, 2014.

[92] Lan Xie. MKL1/2 and ELK4 co-regulate distinct serum response factor (SRF) transcription programs in macrophages. *BMC Genomics*, 15(1):301, 2014.

[93] Sumiko Yoshida, Ken-ichi Aihara, Yasumasa Ikeda, Yuka Sumitomo-Ueda, et al. Androgen receptor promotes gender-independent angiogenesis in response to ischemia and is required for activation of VEGF receptor signaling. *Circulation*, 128(1):60–71, 2013.

[94] Bryan L. Krock, Nicolas Skuli, and M. Celeste Simon. Hypoxia-induced angiogenesis: good and evil. *Genes & Cancer*, 2(12): 1117–1133, 2011.

[95] Li Jiang, Meng Yin, Xiangxiang Wei, Junxu Liu, et al. Bach1 represses wnt/β-catenin signaling and angiogenesis. *Circulation Research*, 117(4):364–375, 2015.

[96] Hideki Kawai, Huchun Li, Philip Chun, Shalom Avraham, and Hava Karsenty Avraham. Direct interaction between BRCA1 and the estrogen receptor regulates vascular endothelial growth factor (VEGF) transcription and secretion in breast cancer cells. *Oncogene*, 21(50):7730, 2002.

[97] Mingcheng Huang, Qian Qiu, Youjun Xiao, Shan Zeng, Mingying Zhan, et al. BET bromodomain suppression inhibits VEGF-induced angiogenesis and vascular permeability by blocking VEGFR2-mediated activation of PAK1 and eNOS. *Scientific Reports*, 6:23770, 2016.

[98] David Engelmann, Deborah Mayoli-Nüssle, Christian Mayrhofer, Katharina Fürst, et al. E2F1 promotes angiogenesis through the VEGF-C/VEGFR-3 axis in a feedback loop for cooperative induction of PDGF-B. *Journal of Molecular Cell Biology*, 5 (6):391–403, 2013.

[99] Haihua Song, Jun-ichi Suehiro, Yasuharu Kanki, Yoshiko Kawai, et al. Critical role for GATA3 in mediating Tie2 expression and function in large vessel endothelial cells. *Journal of Biological Chemistry*, 284(42):29109–29124, 2009.

[100] Vasundhra Kashyap, Shafqat Ahmad, Emeli M. Nilsson, Leszek Helczynski, et al. The lysine specific demethylase-1 (LSD1/KDM1A) regulates VEGF-A expression in prostate cancer. *Molecular Oncology*, 7(3):555–566, 2013.

[101] Troy A Baudino, Catriona McKay, Helene Pendeville-Samain, Jonas A Nilsson, et al. c-Myc is essential for vasculogenesis and angiogenesis during development and tumor progression. *Genes & Development*, 16(19):2530–2543, 2002.

[102] Ken Iwatsuki, Kiyoko Tanaka, Tsuyoshi Kaneko, Ritsuko Kazama, et al. Runx1 promotes angiogenesis by downregulation of insulin-like growth factor-binding protein-3. *Oncogene*, 24(7):1129–1137, 2005.

[103] Farhang M. Ghahremani, Steven Goossens, David Nittner, Xavier Bisteau, et al. p53 promotes VEGF expression and angiogenesis in the absence of an intact p21-Rb pathway. *Cell Death & Differentiation*, 20(7):888–897, 2013.

[104] Gilbert Saporta and Genane Youness. Comparing two partitions: some proposals and experiments. In *Compstat*, pages 243–248. Springer, 2002.

[105] James W. Kent, Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, et al. The human genome browser at UCSC. *Genome Research*, 12(6):996–1006, 2002.

[106] Anthony Mathelier, Oriol Fornes, David J. Arenillas, Chih-yu Chen, Grgoire Denay, et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 44(D1):D110–115, 2016.

[107] Petr Smirnov, Zhaleh Safikhani, Nehme El-Hachem, Dong Wang, et al. PharmacoGx: an R package for analysis of large pharmacogenomic datasets. *Bioinformatics*, 32(8):1244–1246, 2015.

[108] Laurence I-Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1):255–268, 1989.

[109] Yong Zhang, Tao Liu, Clifford A. Meyer, Jrme Eeckhoute, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):R137, 2008.

[110] Stephane Champely. *pwr: basic functions for power analysis*, 2017. URL https://CRAN.R-project.org/package=pwr. R package version 1.2-1.

[111] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, et al. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[112] Takaya Saito and Marc Rehmsmeier. Precrec: fast and accurate precision–recall and ROC curve calculations in R. *Bioinformatics*, 33(1):145–147, 2017.

[113] Wyeth W. Wasserman and Albin Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, 5(4):276, 2004.

[114] Daniel Savic, Christopher E. Partridge, Kimberly M. Newberry, Sophia B. Smith, et al. CETCh-seq: CRISPR epitope tagging ChIP-seq of DNA-binding proteins. *Genome Research*, 25(10):1581–1589, 2015.