BMC Genomics

RESEARCH Open Access

Identifying transcription factors with cell-type specific DNA binding signatures

Aseel Awdeh[1,2], Marcel Turcotte[1] and Theodore J. Perkins[1,2,3*]

Abstract

Background Transcription factors (TFs) bind to different parts of the genome in different types of cells, but it is usually assumed that the inherent DNA-binding preferences of a TF are invariant to cell type. Yet, there are several known examples of TFs that switch their DNA-binding preferences in different cell types, and yet more examples of other mechanisms, such as steric hindrance or cooperative binding, that may result in a "DNA signature" of differential binding.

Results To survey this phenomenon systematically, we developed a deep learning method we call SigTFB (Signatures of TF Binding) to detect and quantify cell-type specificity in a TF's known genomic binding sites. We used ENCODE ChIP-seq data to conduct a wide scale investigation of 169 distinct TFs in up to 14 distinct cell types. SigTFB detected statistically significant DNA binding signatures in approximately two-thirds of TFs, far more than might have been expected from the relatively sparse evidence in prior literature. We found that the presence or absence of a cell-type specific DNA binding signature is distinct from, and indeed largely uncorrelated to, the degree of overlap between ChIP-seq peaks in different cell types, and tended to arise by two mechanisms: using established motifs in different frequencies, and by selective inclusion of motifs for distint TFs.

Conclusions While recent results have highlighted cell state features such as chromatin accessibility and gene expression in predicting TF binding, our results emphasize that, for some TFs, the DNA sequences of the binding sites contain substantial cell-type specific motifs.

Keywords Transcription factor binding, Differential binding, Cell-type specificity, Deep learning

Introduction

Transcription factors (TFs) bind to gene promoters and enhancers to regulate gene expression, and are therefore major determinants of cell fate decisions, metabolic activity, and, when regulation goes awry, of disease [1–3]. TFs bind relatively short preferred DNA sequences, or motifs, typically 5 to 20 bases long [4, 5]. Because these motifs are so short, the human genome often harbors millions of potential matches for a given motif [6]. Yet, ChIP-seq studies of TF binding show that in any given condition, a TF typically binds only several thousands or tens of thousands of those sites [7]. Moreover, that same TF will bind some overlapping but some distinct sites when comparing different cell types or disease conditions [8]. There are many mechanisms that can drive differential binding of a TF, including: differential expression [9], chromatin accessibility [10], conformational changes or complexing with other regulatory factors [11], cooperative or competitive binding [12, 13], and alternative splicing [14].

*Correspondence:
Theodore J. Perkins
tperkins@ohri.ca
[1] School of Electrical Engineering and Compute Science, University of Ottawa, 800 King Edward Ave., Ottawa K1N 6N5, Ontario, Canada
[2] Regenerative Medicine Program, Ottawa Hospital Research Institute, 501 Smyth Rd., Ottawa K1H 8L6, Ontario, Canada
[3] Ottawa Institute of Systems Biology, Department of Biochemistry, Microbiology and Immunology, University of Ottawa, 451 Smyth Rd., Ottawa K1H 8M5, Ontario, Canada

One of the lesser-studied mechanisms of differential binding is a change in the DNA preference of the TF itself. Indeed, one often assumes the reverse—that the DNA binding preference of a TF is the same regardless of the cell type or condition in which it is expressed. Binding motif databases such as JASPAR [15] and HOCOMOCO [16], and experimental methods such as HT-SELEX [17], are predicated on this assumption. Their success shows that, to a substantial extent, the assumption is good. Nevertheless, there are several well-documented cases of DNA preference switching by TFs. For instance, estrogen receptor α binds distinct DNA patterns in different cancerous lines, such as breast cancer and endometrial cancer [18]. The strongest binding sites are bound across all cancer lines, but different lower affinity sites are bound depending on the binding co-factors available. Similarly, in human embryonic stem cells, SOX2 binds regions with distinct motifs depending on whether it is co-binding with PAX6, leading to hECS neural differentiation, or OCT4, leaning to self-renewal [19]. Wang et al. performed a systematic comparison of human TF binding preferences, including a study of other TF motifs present in the peaks of five TFs in five cell types [20]. Arvey et al. used machine learning to study cell-type specific determinants of TF binding in different cell types, with the key finding that cell-type specific sequences were a key factor in predicting binding [21]. Keilwagen et al. studied 31 transcription factors, identifying key features useful in predicting cell-type specific binding [22]. Given the greater wealth of data available today, the time is right for a re-investigation of cell-type specificity in TF-DNA binding sites.

Motivated by the possibility that certain TFs might have cell-type specific DNA signatures at or in the vicinity of their binding sites, we set out to perform a comprehensive and systematic search for the phenomenon. To perform this search, we developed SigTFB (Signatures of TF Binding), a deep learning framework to quantify the *degree* to which cell-type specific DNA signatures are present in a TF's binding sites. One of the advantages of SigTFB is that it can accommodate TF binding data from any number of cell types, without knowing which subset(s) of cell types, if any, may show different DNA binding signatures. Traditional differential motif enrichment analysis can identify known or de novo motifs that may vary between two datasets [23, 24], but it cannot identify subsets of datasets that vary relative to others. Moreover, its computational complexity and the necessary multiple comparison corrections scales quadratically with the number of datasets–problems that SigTFB avoids. Moreover, like other deep learning frameworks that have been highly successful for analyzing TF binding [25–42], SigTFB is capable in principle of learning sophisticated DNA patterns that influence TF binding. This is a good choice when many different or even unknown molecular mechanisms could be generating binding signatures. However, our problem formulation is unique compared to previous work, which has largely focused on separating binding sites from non-binding sites—while sometimes optimizing other criteria such as the spatial resolution of predictions [30–32], interpretability of results [35], or data efficiency [37]. In contrast, in our formulation, *all instances are bona fide, empirical binding sites for a given TF*, determined by replicate ChIP-seq experiments. Our task is to predict *in which cell types* those sites are bound. By comparing that prediction performance with performance when the target cell type is hidden, we can quantify the extent to which deep learning can identify cell-type specific DNA patterns. Our work also stands distinct from, and complementary to, recent studies of how multiple TFs combine to discriminate genuine/functional binding sites across the genome in a single cell type [43, 44]. Moreover, our question is distinct from the question of whether a TF has different roles or functions in different cell types. A TF may bind different sites in different cell types, and regulate different genes, and yet the DNA motif it binds in those sites can be identical. Conversely, differences in DNA binding motif in different cell types do not necessarily imply that the functional role of that TF is noticeably different, particularly if there are key binding sites that remain the same between cell types. The question of DNA binding signatures is one that most directly pertains to the mechanisms that guide a TF to bind its target sites in different cell types.

Using our SigTFB method, we investigated the binding of 169 distinct human TFs assayed by one or more antibodies (AB) (for a total of 199 distinct TF-AB pairs) across multiple cell types (ranging from 2 up to 14 for any given TF; 35 distinct human cell types in total). We found that different TFs show varying degrees of cell-type specific DNA binding signatures, with approximately two-thirds of TFs having significant cell-type specific signatures. Importantly, we found that the mere presence of differential binding is not the same as having a DNA signature of differential binding. Many TFs bind very different sites in different cell types, yet show no specific, discriminating DNA signatures at those differential sites. In such cases, TF binding differences may be due to mechanisms that do not leave strong local signals in the DNA, such as chromatic accessibility [10]. We also compared our results when analyzing data from the same TF assayed by different antibodies, and find that, with a few exceptions, there is generally good agreement on whether a TF displays cell-type specific DNA binding signatures. Finally, we show that across all TFs and cell types, differences in DNA signatures commonly emerge as differences in the frequency of presence of the same motifs, but in several cases, as radical switches to the inclusion of different motifs.

The remainder of the paper is organized as follows. First, we introduce the supervised learning problem formulation that we propose for quantifying the presence of cell-type specific DNA binding signatures. Next, we describe our deep learning-based SigTFB method for solving that problem. We then analyze two TFs in detail: ATF7, which shows substantial cell-type specificity in the DNA sequences of its peaks, and CTCF, which does not. We then provide a summary analysis of all TFs. Next, we perform some embedding and motif analyses to further investigate cell-type specific sequences in peaks and the deep learning representations of those sequences. Finally, we conclude with a discussion of our results, its strengths and limitations, and directions for future work.

Results

A supervised learning formulation for detecting cell-type specific DNA-binding signatures

In this section, we describe a novel supervised learning problem whose solution allows one to identify and quantify cell-type specific DNA-binding signatures in a collection of known binding sites for a single TF across a set of cell types. The essential idea is to take DNA sequences from known binding sites, and employ supervised learning to predict whether a TF binds that sequence in a given cell type. In addition, the supervised learning is asked to make the same prediction, but when the target cell-type information is hidden. The difference in predictive performance between the cell-type specific instances and the cell-type hidden (or general) instances quantifies the extent to which the learner is able to pick up cell-type specific DNA signatures that improve binding prediction.

For our study, we turned to the ENCODE project peak calls [7] to identify high-quality, known TF binding sites. Because different antibodies for a TF can have different specificities or biases, we chose not to mix data from different antibodies. We identified 169 TFs satisfying the following criteria: 1) The TF is assayed by the same antibody in at least two human cell types; 2) for each cell type, "experiment" level peaks are available. Such peaks are present in at least two replicate ChIP-seq experiments and pass an irreproducible discovery rate test at a 2% threshold; and 3) there are at least 1000 such peaks for each cell type. Some TFs satisfied these criteria for more than one antibody, so in total we identified 199 transcription factor-antibody (TF-AB) combinations that we could use to study cell-type specificity in DNA binding. The full list of experiment accession numbers is available in Supplementary Table 1. We downloaded the called peaks for these accession numbers from the ENCODE website.

In our formulation, each TF-AB combination is studied separately. For each TF-AB, we begin by constructing a "unified" set of peaks across all the cell types in which that TF was assayed by that AB, using the approach developed by Basset et al. [45] in their study of chromatin accessibility (Fig. 1a). Starting with the set of all peaks identified in all cell types (for that TF-AB pair), we repeatedly merge any two peaks that overlap by at least 30bp, keeping track of which cell types contribute to each merged peak. At the end of this process, we have a set of unified peaks annotated with source cell types (at least one, and as many as all cell types). The unified peaks can be of varying sizes depending on the sizes of the original peaks, their degree of overlap, and how many different peaks are combined. To "normalize" them for ease of supervised learning, the center of each unified peak is taken and extended by 50bp in each direction, such that the length of the intervals is 101bp. (This window size has been common in other, similar studies, and our own pilot study showed degradation of performance below 101bp, and no gain with bigger windows. See Discussion for more information.) Where there are C number of cell types, each unified peak is translated into $2C$ supervised learning instances. In C of those instances, the input is the DNA sequence of the unified peak center along with a one-hot encoding of one of the cell types, and the output is one or zero depending on whether that cell type had a peak or not at that location. These are called the cell-type specific instances. The other C instances associated with the unified peak are identical, except that the part of the input vector encoding the cell type is zeroed out. We call these the cell-type general instances. As mentioned above, the intuition behind this formulation is that the difference in predictive performance between the cell-type specific instances and the cell-type general instances is a measure of the extent to which knowing the cell type informs ones interpretation of the input DNA sequence. In other words, it is a measure of the presence of cell-type specific binding site sequences, or DNA binding signatures, for this TF-AB pair, across this set of cell types.

SigTFB: a two-stage deep learning model to study DNA-signatures associated to TF binding

To solve the learning problem described in the previous section, we developed a deep learning architecture and two-stage training approach called SigTFB (Fig. 1b). The two-stage approach is modeled after that of Nair et al. [46]. Stage 1 of training is meant to help initializing the first DNA sequence-interpreting layers of the network, and is described further in the Methods section. In stage

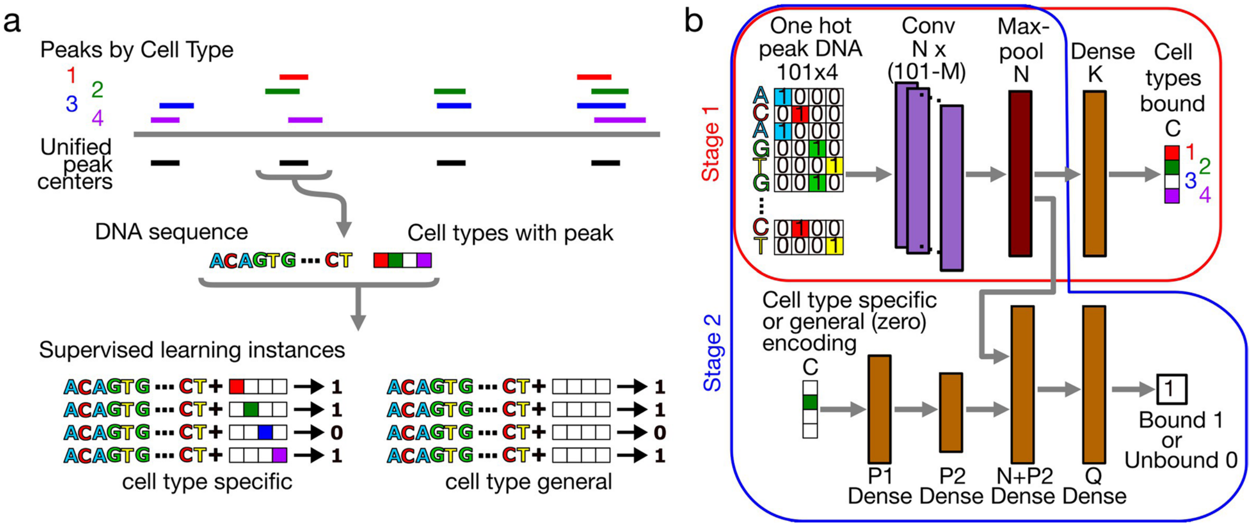

Fig. 1 Supervised learning formulation and deep learning architecture. **a** Empirical binding sites are unified across cell types. In "cell-type specific" instances, each site's DNA sequence and one-hot encoded cell type is associated to a binary bound/unbound output. In matching "cell-type general" instances, the cell-type information is hidden, but DNA input and bound/unbound output remain the same. **b** Simplified diagram of deep learning architecture, and its division into Stage 1 and Stage 2 Models. Stage 1 is shown in the red outline, and Stage 2 is shown in the blue outline. In Stage 1, the input instance is a one-hot encoded DNA sequence of size 101 × 4. This is passed through the convolutional layer (Conv) with *N* filters of size *M*, through a maxpool layer (Maxpool) of length *N*, a fully connected layer (Dense) of length *K*, then the output layer of length *C* to predict if the TF is bound or not in the different cell types. Stage 2 takes cell-type information of length *C* as input as well, as depicted in panel (**a**). This is first passed through fully connected layers of lengths *P*1 and *P*2, and then concatenated with the output of the maxpool layer from Stage 1. The concatenated output is passed through a fully connected layer of length *Q* to predict whether the sequence is bound or not in that cell type

2, the network's inputs and outputs are as described above. Our network includes a modified version of DeepBind [25] of one hidden layer convolutional layer followed by one fully connected layer (Fig. 1b). Unlike DeepBind, the number of channels in our model is set as a hyperparameter. We also investigated the use of more complex models, with more than one convolutional layer. However, one convolutional layer gave the best results in terms of validation accuracy and loss. The lower part of the network in Fig. 1b takes as input the length C binary vector encoding a specific cell type or a zero vector, processes it through dense layers and then combines that with the DNA-processing side of the network through additional dense layers until reaching a binary output node.

We use negative log likelihood loss for training. The entire training procedure is performed in 10x cross-validation, with held-out performance being recorded for each cell type, along with the macro-average across cell types. Within each fold, performance is also averaged over 10 random initial weight sets and training trajectories. During both training and testing, instances are randomly chosen in mini-batches to have the same number of positive and negative instances from each cell type, and the same number of cell-type specific and cell-type general instances, avoiding any problems with class imbalance. Finally, all of that is wrapped within Ax [47] for tuning the various network layer size hyperparameters *M*, *N*, *K*, *P*1, *P*2 and *Q* shown in Fig. 1b. The AUROC is computed for each cell type and for cell-type specific and general instances separately. The macro-averaged AUROC is computed across cell types, and the difference in macro-averaged AUROC between cell-type specific and general instances is used as our measure of cell-type specificity.

In the next two sections, we provide a detailed analysis of our results for two transcription factors. First, we examine a transcription factor with a high degree of cell-type specificity in its DNA binding signature. Then, we present an example where SigTFB found little evidence of cell-type specificity in the DNA binding signature.

ATF7 binding shows cell-type specific DNA binding signatures

To illustrate our approach, we first focus on Activating Transcription Factor 7 (ATF7). As a member of the ATF family, ATF7 binds to the cyclic AMP response element (CRE) with the consensus DNA sequence "TGACGT CA" [48, 49]. Members of the ATF family are basic leucine zipper (bZIP) factors that complex with other bZIP factors to form homodimers or heterodimers [48–51]. These ATF TFs exhibit varying functionalities in different tissues and cancerous cell types, including tumour suppressive and oncogenic functions [49]. For instance, the deletion of ATF7 results in the spread of lymphoma [49]. Conversely, the activation of ATF7 in gastric or hepatocellular carcinoma promotes the proliferation of cancer cells. As such, ATF7 may be used as a biomarker for the early detection of tumours in liver and gastric cell types. Due to the differences observed, we suspect ATF7 to bind to different places along the genome in different cell types.

Our ENCODE [7] data compendium (see "A supervised learning formulation for detecting cell-type specific DNA-binding signatures" section) includes ATF7 peaks in four cell types: GM12878, K562, HepG2 and MCF-7. The cancerous cell types HepG2, MCF-7 and K562 correspond to liver hepatocellular carcinoma, breast cancer and myelogenous leukemia respectively. GM12878 is a non cancerous lymphoblastoid cell type. Figure 2a shows a Venn diagram of the peak overlaps between the four cell types. The number of peaks per cell type are shown after the cell-type name in brackets. A mere 1.36% of the total number of peaks across all four cell types overlap, and the majority of the peaks are unique to one of the four cell types. For example, 22.36% of the K562 peaks do not overlap with peaks from other cell types. There is greater peak overlap between the pairs HepG2 and MCF-7, and GM12878 and K562.

The lack of overlap between peaks in the four cell types does not imply cell-type specificity in DNA binding preference, as sequences in those peaks may be very similar. Differences in output may be due to dissimilarities in terms of noise, bias or even the number of peaks of the ChIP-seq experiments. For instance, HepG2 has over 40,000 peaks while MCF-7 has fewer than 30,000. Therefore, no more than 75% of HepG2 peaks could possibly overlap with MCF-7 peaks.

To determine if there are cell-type specific DNA signatures in the ATF7 peaks, we applied our deep learning method, SigTFB, as described in the previous section. Figure 2b shows the receiver operating characteristic (ROC) curves for each cell type with and without the cell-type identity being provided, as well as averaged performance across all cell types. The plot shows high variability in site prediction across across cell types. Predictions for HepG2 (solid red curve) are significantly better than for MCF-7 and K562 (solid purple and green), which are better than for GM12878 (solid blue). In this case, predictions are more accurate when the network is informed of cell type than when it is not (e.g. solid curves versus dashed curves). This trend is also true for the macro-averaged ROC curve (gray color in Fig. 2b). Figure 2c shows the area under the ROC curve (AUC) per cell type per condition for the ATF7 TF, where the shaded and unshaded bars are cell-type general and cell-type specific cases respectively. For each cell type, as well as the macro-averaged result, there is a clear difference between the two conditions. Cell-type specific classification outperforms cell-type general classification with a macro-averaged AUC difference of 0.2 ($p \ll 0.05$; one-sample t-test on AUC difference). Thus, we can conclude that the network has detected DNA signatures discriminating peaks in different cell types. Further below, we investigate what exactly those signatures might be.

CTCF binding does not show cell-type specific DNA binding signatures

We next examine CCCTC-binding factor (CTCF), which can function as a transcriptional repressor, transcriptional activator, or as an insulator barrier between genomic domains. The CTCF binding domain is defined by 11 zinc fingers, and binding preference is believed to be invariant across cell types [52–54]. Importantly, this does not mean that CTCF always binds the same sites in different cell types, nor does it mean that it has the same function in different cell types. Indeed, by binding different sites or by binding the same sites with different binding partners, CTCF can have cell-type specific functions. For instance, CTCF has been shown to bind specific groups of regulatory elements in different brain cell types [55–57], where it specifically regulates memory-related genes, among others [58]. Conversely, these cell-type specific functions or binding sites do not imply any difference in direct CTCF-DNA binding preference or other signature. Therefore, we used SigTFB to test whether CTCF binding sites had any cell-type specific DNA signatures.

In our ENCODE data compendium, CTCF is assayed by five different antibodies. Here, we focus on the antibody that was used the most, giving us empirical binding sites for CTCF in 14 different cell types: smooth muscle cell, GM23338, bipolar neuron, neural progenitor cell, fibroblast of dermis, myotube, PC-3, astrocyte, HCT116, hepatocyte, osteoblast, OCI-LY7, MCF-7 and SK-N-SH. The percentage overlap of ChIP-seq peaks between each pair of cell types is shown in Fig. 2d, where each entry of the heatmap shows the percentage of peaks of the row's cell type overlapping peaks in the column's cell type. Additionally, the number of peaks per cell type are shown in brackets after the row cell type label. Overlap percentages range from approximately 50% to 90%, with an average of 77%. Cell types with fewer peaks tend to be better covered by cell types with more peaks, suggesting an element of peak detection power is at play. For instance, the astrocyte dataset has the fewest peaks at ≈37,000, which are more than 90% covered by the CTCF peaks in every other cell type – even distantly related cell types such as osteoblasts or fibroblasts (first row in Fig. 2d).

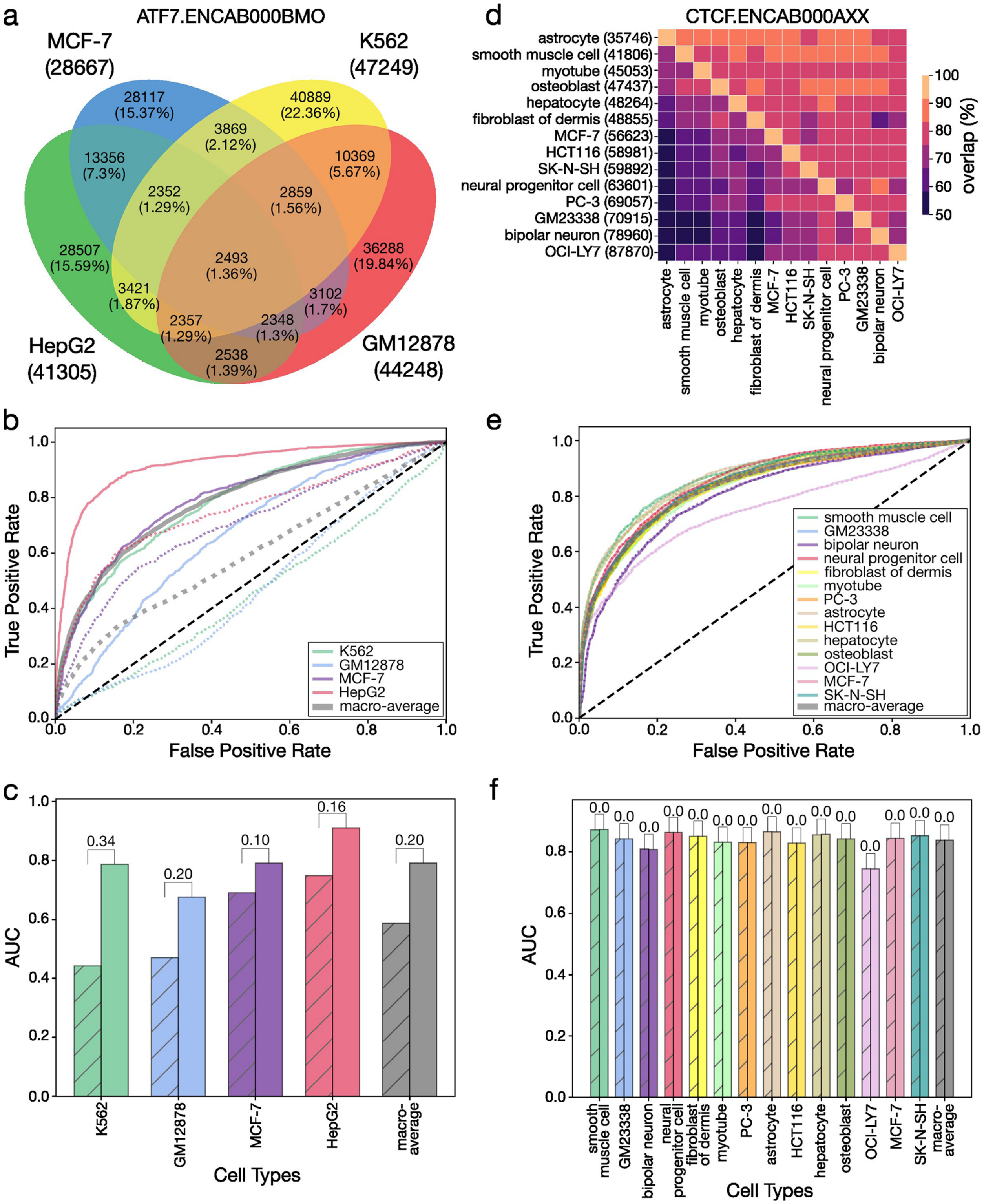

Fig. 2 Illustration of SigTFB on ATF7 and CTCF. **a** Venn diagram of percentage overlap between cell types for ATF7. **b** ROC curves per cell type per condition: cell-type general (dashed line) and cell-type specific (solid line) for ATF7. **c** AUC per cell type per condition: cell-type general (shaded) and cell-type specific (not shaded) for ATF7. Numbers at the tops of pairs of bars are the AUC difference between cell-type general and specific instances. **d** Heatmap of percentage overlap between 14 cell types in CTCF assayed by antibody ENCAB000AXX. **e** ROC curves per cell type per condition: cell-type general (dashed line) and cell-type specific (solid line) for CTCF. **f** AUC per cell type per condition: cell-type general (shaded) and cell-type specific (not shaded) for CTCF. Numbers at the tops of pairs of bars are the AUC difference between cell-type general and specific

Figure 2d gives some intuition about the datasets. However, as seen for ATF7, a simple intersection analysis is not sufficient to determine cell-type specificity. We further investigated the binding activity of CTCF by training SigTFB on CTCF and its 14 corresponding cell types. Figure 2e shows the ROC curves for each of the cell types and the macro-averaged ROC across all cell types. Compared to ATF7, there is relatively little difference in binding site predictability across cell types and nearly no difference in predictability for a given cell type, with or without cell-type identity information. Cell types OCI-LY7 (lavender line) and bipolar neuron (indigo line) have the worst prediction performance, and also have the highest number of peaks. Possibly, some fraction of these peaks are less reliable, which would explain both inflated peak numbers and prediction difficulty. Figure 2f shows there is little to no difference in the area under the ROC curves (AUC) between cell-type specific (solid line) and cell-type general (broken line) conditions for each cell type ($p > 0.5$; one-sample t-test on percentage differences). Consequently, these results illustrate the ubiquitous non-cell-type specific nature of CTCF DNA binding preferences. Importantly, they also demonstrate the specificity of SigTFB, in that it does not incorrectly report cell-type specificity where there is none to be found.

Comprehensive analysis of cell-type specific DNA-signatures in 169 transcription factors

Motivated by our results for ATF7 and CTCF, we expanded our study to investigate cell-type specific DNA binding signatures in all 169 TFs (199 TF-AB pairs). Figure 3a displays a scatter plot of the mean AUC of prediction when the network is (y-axis) or is not (x-axis) told what cell type it is predicting for. Each point corresponds

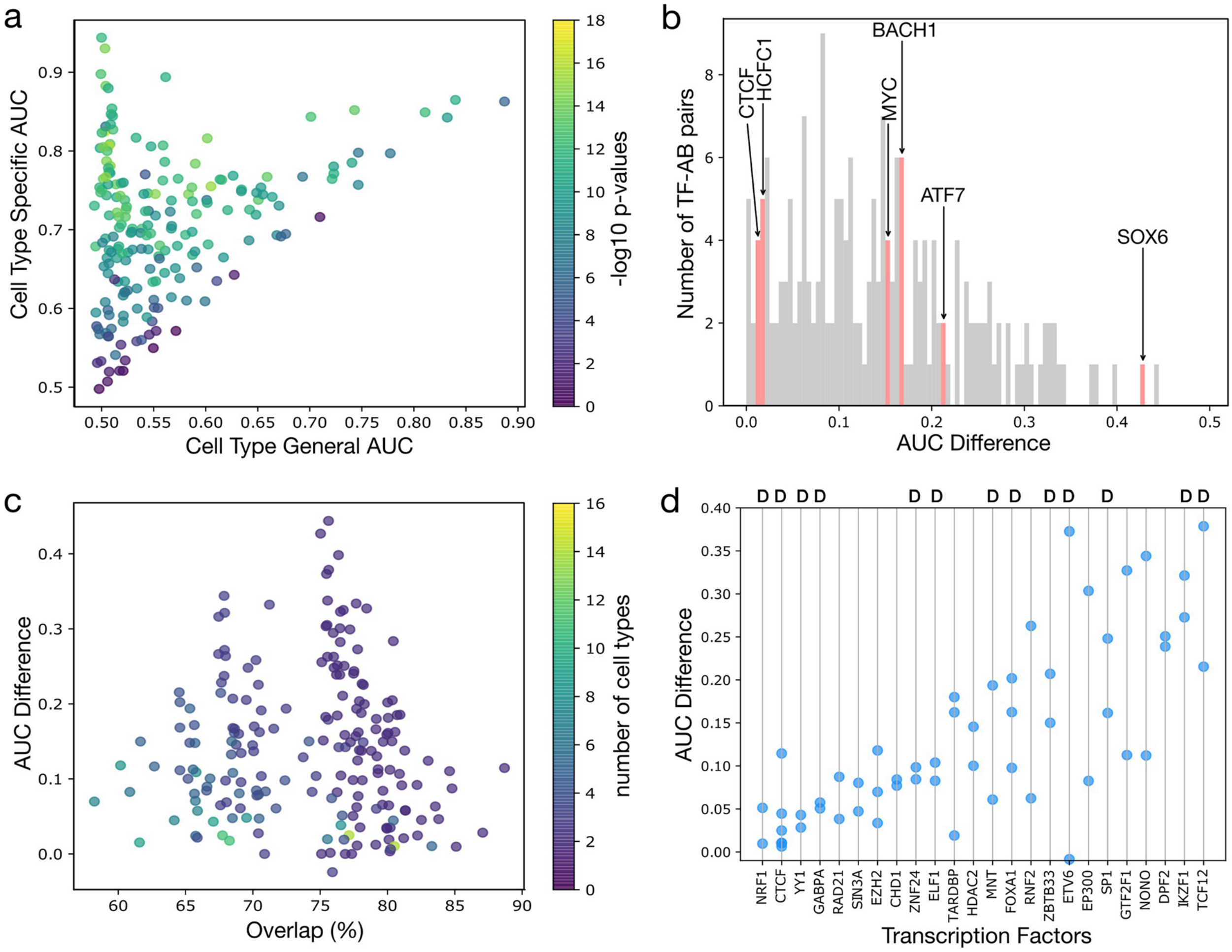

Fig. 3 Summary statistics from our comprehensive study on DNA signatures in TF binding using SigTFB. **a** Scatter plot of cel-type specific AUC versus cell-type general AUC with the color gradient depending on the -log10 p-values. **b** Bar chart of AUC differences. **c** Scatter plot of percentage overlap(%) and AUC differences, with the color gradient depending on number of cell types per TF-AB. **d** Scatter plot of AUC differences for TFs with more than two ABs. The letter "D" at the top of the panel indicates a TF with strong evidence of direct sequence-specific DNA binding, as opposed to possible indirect DNA binding through intermediaries

to a TF-AB combination. The color gradient depends on the negative log 10 p-values for statistical significance of difference between the cell-type specific and cell-type general predictions, across the 10 folds of cross validation. We observe a continuum of cell-type specificity, where TFs with the least cell-type specificity lie in the $x = y$ diagonal of the scatter plot. For these TFs, the cell-type information does not improve prediction. The position of a point along the diagonal may depend on the extent to which there are common motifs for the TF across cell types, the extent to which the peaks themselves overlap across cell types, data set quality, or other factors. Conversely, points lying *above* the diagonal indicate that the network predicts binding better when informed of the cell type; these are TFs with the most significant cell-type specificity, where the network responds differently to input DNA sequences depending on the cell type for which it is predicting. Points in the the upper left corner correspond to TFs where cross-cell-type prediction is virtually impossible, but is highly accurate for specific cell types. For such TFs, each cell type is expected to have specific DNA motifs that discriminate its binding sites.

Figure 3b shows a histogram of the distribution of AUC differences between cell-type specific and general predictions for the different TF and AB combinations, with select TFs highlighted. Out of 199 TF-AB combinations, 127 TF-ABs, or 116 distinct TFs, have a statistically significant AUC difference of at least 0.1, suggesting that a majority of TFs have some degree of cell-type specific DNA signatures in their binding sites. TFs that play a pivotal role in cancer either as oncogenes or suppressors, such as MYC [59], BACH1 [60], ATF7 [49], and SOX6 [61], show a relatively higher cell-type specificity than other TFs, such as CTCF [53] and HCFC1 [62], that are involved in chromatin regulation or other cellular processes. Supplementary Table 1 lists the AUC differences for all TF-AB pairs.

As explained above, the lack of overlap between binding sites in different cell types is not evidence per se of any differential DNA signature. We next examined whether there is any association between the two. Figure 3c plots the mean pairwise percentage peak overlap versus the mean AUC difference for each TF-AB. No clear relationship between the two variables is seen (Spearman correlation r = -0.1). With either a high percentage overlap between 75% and 80% or a lower percentage overlap between 65% and 70%, the AUC difference ranges between roughly zero and 0.4. This confirms that peak overlap is not in itself an indicator of cell-type specificity in DNA binding signatures.

We also considered the possibility that AUC differences might somehow be an artifact of the number of cell types in the analysis. For instance, if there were only two cell types, perhaps it is more likely that some one of the two would contain some spurious signal that allows peak discrimination. Conversely, perhaps the more cell types are assayed, the more likely there is to be an "outlier" cell type with spurious DNA signals in the peaks. We found little evidence of such a phenomenon. The color gradient in Fig. 3c indicates the number of cell types tested. The correlation of AUC difference to number of cell types is a minor (r = -0.1).

Of the 169 distinct TFs we studied, 24 were assayed with multiple ABs. Lack of consistency across ABs due to different off-target biases or binding affinities may impact the TF's DNA signatures. Moreover, different ABs may have been used on different sets of cell types. Nevertheless, we may be reassured of the generality of our results if our measure of cell-type specificity is consistent between different sets of experiments with different ABs for the same TF. Figure 3d shows a plot of the AUC differences for the 24 TFs assayed by least 2 ABs. For example, CTCF was assayed with six different ABs, all of which returned relatively low estimates of cell-type specificity (five of six being below 0.04). Conversely, several TFs show consistently high cell-type specificity across multiple ABs, including TCF12, SPI1, MNT, IKZF1 and DPF2. The least consistency is seen for the TF ETV6. Surprisingly, both datasets for ETV6 explore the same two cell types GM12878 and K562, yet produce very differing results for cell-type specificity: essentially 0 for one antibody and 0.37 for the other. This may be due to differences in the ABs used, or could be a result of differences in the total number of peaks per dataset for each TF-AB combination. Overall, however, there is strong consistency in our measure of cell-type specificity of TF binding, even when assayed by different ABs.

We also marked TFs in Fig. 3d with a "D" above them when there was strong evidence from prior research that the TF directly binds DNA in a sequence specific manner, as opposed to indirectly as part of a complex. We took as "strong evidence" the TF being annotated with the Gene Ontology molecular function "RNA polymerase II cis-regulatory region sequence-specific DNA binding", and having a DNA binding motif in one or both of the JASPAR [15] and HOCOMOCO [16] databases. Despite the possibility that direct- versus indirect-DNA binders might have different propensities towards cell-type specificity, we see no obvious trend in that regard.

TFs with cell-type specificity show differential enrichment for known TF-DNA binding motifs

As mentioned above, SigTFB attempts to identify the presence of cell-type specific DNA signatures in a TF's peaks, but doesn't explicitly tell us what those signatures are. In this section, we investigate further what those signatures might be, starting with ATF7. First, adopting a similar approach to AI-TAC [63], we used the t-SNE algorithm to represent each ChIP-seq peak per cell type by its activation values in two dimensions across the neurons of the final fully connected layer of the stage 2 model. Figure 4a and b show the ATF7 t-SNE plots for cell-type general and specific instances respectively. Each point/instance is colored depending on which cell type(s) it belongs to. The cell-type general instances (Fig. 4a) appear as a single cluster, although the peaks from some cell types do tend to be on one side or the other of the mass. The network's internal representations of the cell-type specific instances however, group perfectly by cell type (Fig. 4b). A similar analysis of the CTCF learned model (Fig. 4c, d) shows a single cluster for both general and specific instances. Although the cell-type specific instances show some increasing grouping by cell type, it apparently has little effect on predictive power, as we saw negligible AUC difference above. These results reinforce our contention that SigTFB is able to learn cell-type specific representations of DNA binding sequences, especially for ATF7 in comparison with CTCF.

Next, we explored the filters from the convolutional layer by converting the filters into PWMs and using Tomtom [64] to search for the PWMs in the JASPAR database [15]. In the ATF7 network, we found that most filters had a match or partial match to a small number of known TFs. For example, ≈40% of the filters matched bested to a JUND motif. ATF7 and JUND both have basic leucine zipper domains, with similar consensus binding sequences, and are known to physically interact [65]. Another set of filters matched the motif for SP2, which has a very different zinc finger binding domain that prefers a gapped sequences of G's or C's.

To assess more systematically which TF motifs might be present in ATF7's binding sites, we constructed 35 base pair windows around the positions that in silico mutagenesis found to have the most influence on network output, and then ran FIMO to identify significant motifs hits for a library of 400 high-confidence human TF-DNA binding motifs from JASPAR [66]. (The 35 base pair window size is larger than any motifs in the library, and provides a more focused analysis less likely to include the many short TF motifs by random chance.) Figure 4e shows the fraction of ATF7 peaks in each cell type that included motif hits for the 33 motifs with most significant results. ATF7 peaks in HepG2 and MCF-7 cell types have very similar enrichment patterns, and include expected high enrichment of ATF family motifs, as well as other similar motifs. The peaks in K562 are surprisingly low in such motifs, although the enrichment levels are statistically significant. Instead, the K562 peaks are enriched in KLF- and SP-family motifs. The GM12878 peaks have some enrichment for all of these families, along with several other unique results such as ELF and ETV motifs. The presence of these different cell-type specific motif hits, along with the convolutional filter analysis, suggests that the ATF7 model may be looking at motif frequencies and/or the presence of other TF's motifs, to help discriminate ATF7 peaks in different cell types. A parallel analysis of CTCF's peaks found nearly identical motif enrichment across all cell types (Fig. 4f), further confirming the lack of cell-type specific DNA signatures.

We extended the motif analysis to all 199 TF-AB pairs. Space limitations prevent a comprehensive presentation of the results, but Fig. 5 contains a heatmap of motif presence frequency for a broad selection of TFs from major families [67, 68] and for the motifs with highest average scores. Many insights and hypotheses can be obtained from detailed examination of the matrix, but several major observations can be made immediately. A common trend of many TFs is that the canonical motif and similar motifs are enriched to different degrees in different cell types. For instance, cluster A shows that MAX and USF family factors' binding sites are, unsurprisingly, enriched for MAX and USF motifs. Similarly for clusters B, C, D1, D2 and E, which can been seen closer up in Supplementary Figure 1. However, there are interesting exceptions to these trends. For instance, in the A cluster, totally different experiments with two separate antibodies agree that MNT peaks in MCF-7 are lacking many motifs that are seen more abundantly in MNT peaks in HepG2 and K562 cell types, as well as in the peaks of many other TFs in the same group.

Another interesting phenomenon is the presence of alternative motifs, including primary motif variants, specifically in some cell types. For instance, in cluster D2, the peaks for JUND in the five cell types at the right of Supplementary Figure 1D2 (HepG2, K562, MCF-7, SK-N-SH,

(See figure on next page.)

Fig. 4 Neural network interpretation and motif enrichment in peaks. **a**-**d** tSNE embedding of unit activations in final dense layer. Points represent individual peaks, colored by originating cell type(s). **a** ATF7 peak embedding when cell-type information is withheld. **b** ATF7 peak embedding when cell-type information is provided. **c** CTCF peak embedding when cell-type information is withheld. **d** CTCF peak embedding when cell-type information is provided. **e**-**f** Enrichment of top JASPAR motifs in ATF7 (**e**) and CTCF (**f**) peaks, as the fraction: number of motif-containing peaks using default FIMO parameters, divided by total number of peaks

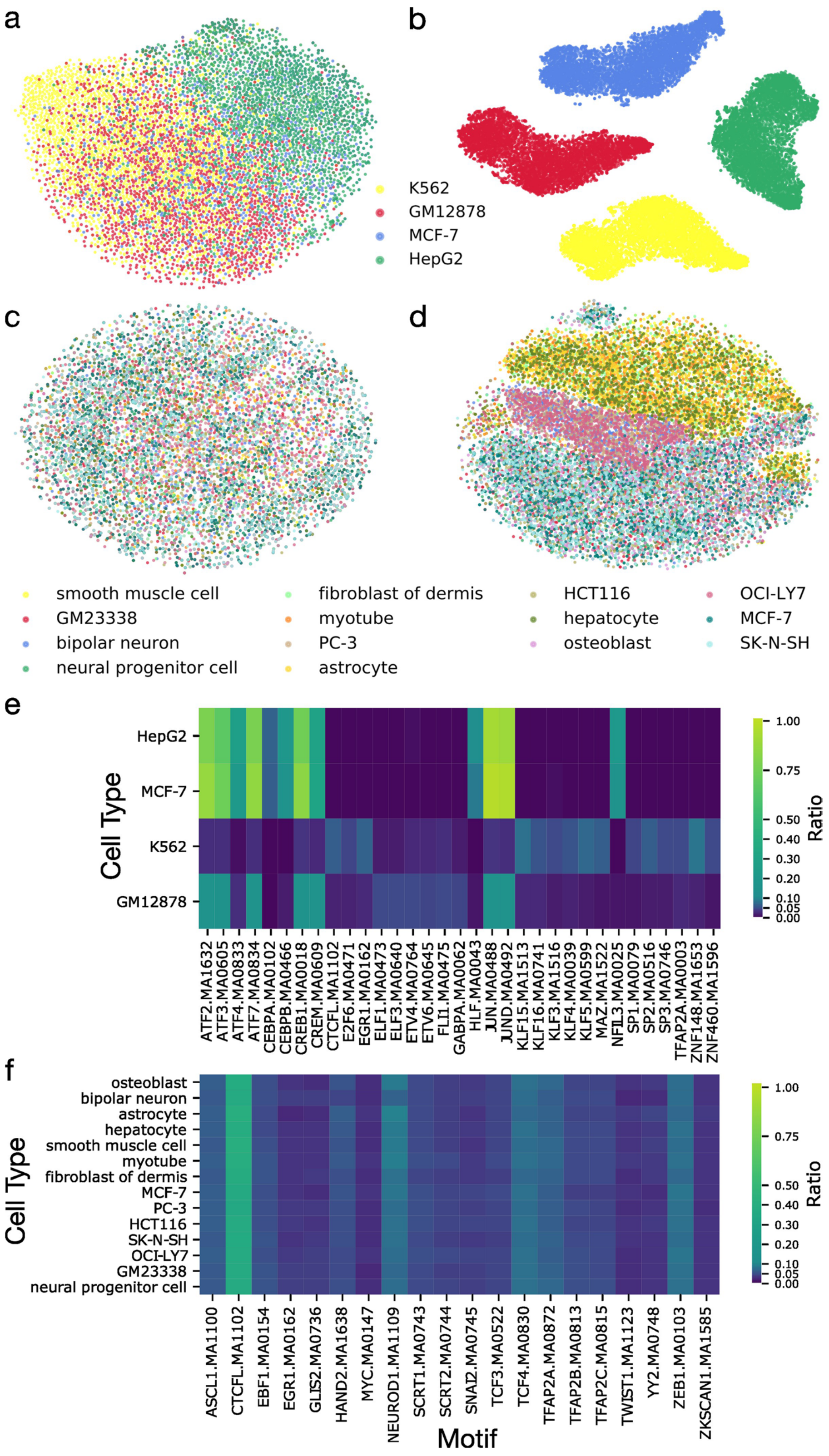

Fig. 4 (See legend on previous page.)

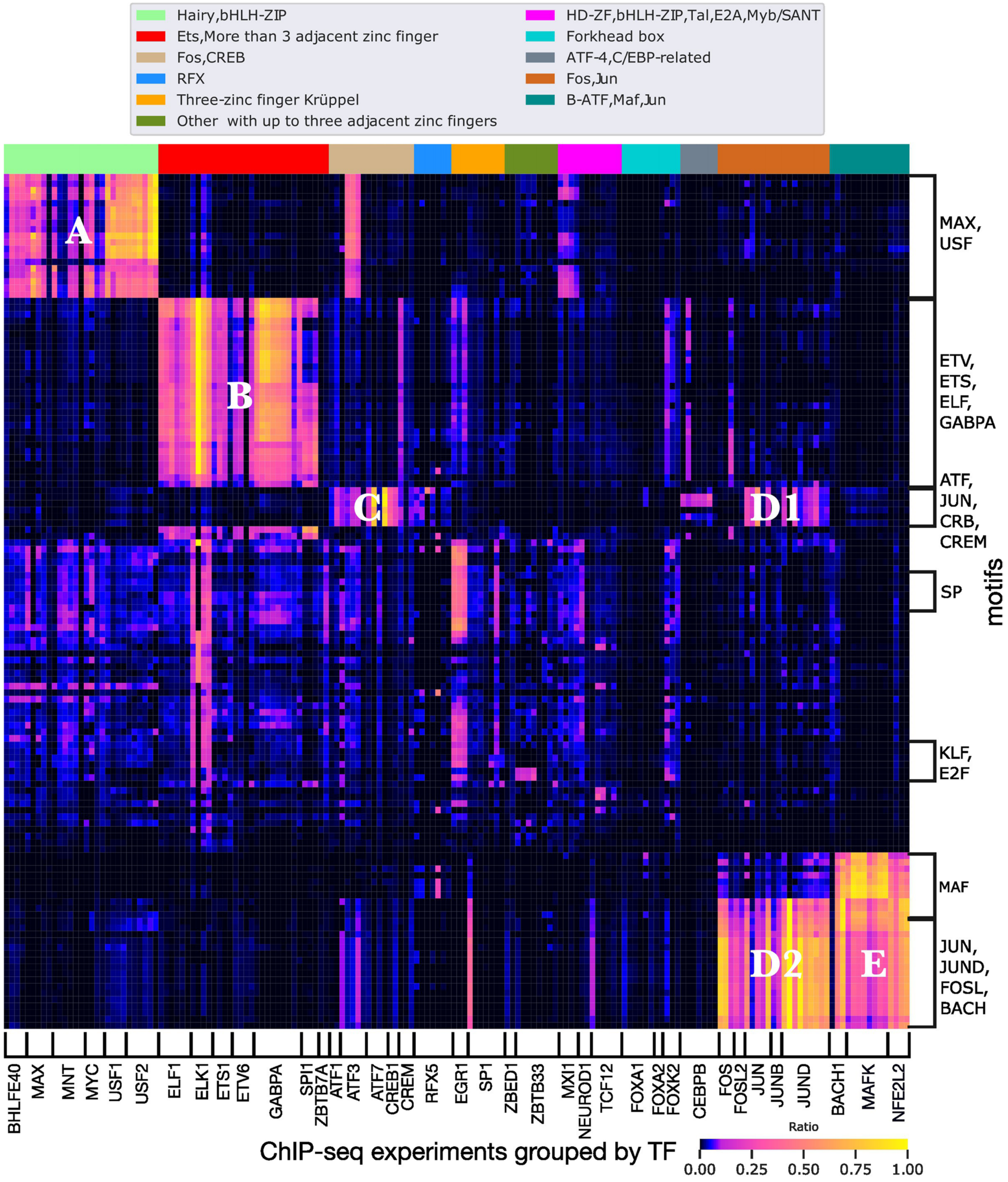

Fig. 5 Enrichment fractions of a broad range of JASPER motifs (rows) in peaks of TFs in 11 major families across numerous tissue types (columns). Close up views of regions of the heatmap A, B, C, D1, D2 and E can be found in Supplementary Figure 1

liver) show some signifiant enrichment for MAFK motif variants (at top of plot), whereas peaks in GM12878 or HCT116, which are generally higher on enrichment for JUN and FOS motifs, have little or no enrichment for the MAF variants at the top of the plot. Similar results can be seen in cluster A, for example where USF2 peaks in SK-N-SH are absent MYCN motifs, although MYCN appears commonly in USF2 peaks in other cell types. The USF2 peaks in SK-N-SH are also unusually low on the MAX motif variant MAX.MA0058.2 (concensus CAC ATG) but high on MAX.MA0051 (concensus CACGTG)–quite possibly because CACGTG is also a concensus motif for USF2 itself, but suggesting that USF2 might also bind the CACATG variant significantly (which is not among any of the USF2 concensus sequences in JASPAR).

Returning to Fig. 5g, one can see many other interesting exceptions to family-wise binding. For instance, near the top middle we see that in three of five cell types, peaks for ATF3 are substantially enriched in MAX/USF family motifs, whereas the other two cell types are virtual without these motifs. Numerous other such examples are found throughout the matrix, and are suggestive of many potential hypotheses regarding co-expression of TFs or lack thereof, co-binding, competitive binding, etc.

Discussion

The complex structural and biochemical nature of protein-DNA interactions has made it difficult to fully understand how various factors influence transcriptional regulation and differential binding. We conducted a wide-scale investigation of 169 TFs across various cell types to identify and quantify differential binding preferences of TFs. We found that different TFs display varying degrees of cell-type specificity in their binding preferences, with approximately two-thirds of those we tested having statistically significant DNA signatures of differential binding. We observed that TFs that play a pivotal role in cancer either as oncogenes or suppressors, such as MYC, BACH1, ATF7 and SOX6, show a relatively higher cell-type specificity than other TFs, such as CTCF and HCFC1, that are involved in chromatin regulation or other cellular processes. Our work constitutes a broad survey of the possibility and prevalence of such DNA signatures. However, the signatures found by SigTFB could reflect many different factors that influence the preferences of a TF, such as its intrinsic binding preference, chromatin accessibility or co-operative or competitive binding of other factors. Further experimental validation is needed if we are to determine mechanisms underlying these signatures. For instance, for a number of TFs we observed the increased presence of binding motifs for other TFs. This could be tested experimentally by first verifying that those other TFs are expressed in the cell types of interest, and then performing ChIP-seq experiments on those TFs to confirm binding at the same sites. If co-operativity/complexing is suspected, reciprocal IP experiments could be performed to identify physical interactions between the different TFs, or knockdown of one TFs could be performed followed by ChIP-PCR or ChIP-seq to determine if binding of the other TF is affected. As another example, if the direct DNA binding preference of a TF is suspected to have changed in different cell types, in vivo affinity assays using enhancers constructs could be performed to verify this change. Therefore, much more experimental work, and potentially computational work, is needed to test our findings.

Other deep learning approaches, such as MTTFSite [69] and Phuycharoen et al. [70], have also explored differential binding of TFs across cell types. While MTTFSite and Phuycharoen et al. adopt a similar learning framework to SigTFB in stage 1 training, in terms of using a multi-task model, their problem formulation and objective fundamentally differ. In MTTFSite, for example, prior to training, shared non-unique cell-type instances are defined as bound regions across cell types that overlap by at least 100bp, while the remaining bound instances that do not overlap are cell-type specific. In SigTFB, however, the model is given all instances as input and learns to differentiate non-specific versus cell-type specific instances. The negative instances for a specific cell type in SigTFB are bound regions in other cell types, while in MTTFSite and Phuycharoen et al. negative instances are unbound regions in all cell types. SigTFB essentially learns to differentiate between shared and unique motifs in cell types from only bound regions. Additionally, the scale of the study differs. MTTFSite and Phuycharoen et al. investigate TFs in a total of five and three cell types respectively, while SigTFB explores the hundreds of TFs in ENCODE with at least more than two cell types available, or a total of 35 distinct cell types across all TFs.

Similar to Novakovsky et al. [71] and ChromDragoNN [46], SigTFB displays the effectiveness of transfer learning in a multi-task deep learning framework for the prediction of binding profiles genome wide. Unlike these approaches, however, which mainly focus on cross cell-type prediction, where models are trained on some cell types and tested on other cell types with limited data, we use transfer learning to acquire exclusive features per cell type. The multi-task setting in the first stage of learning allows the model to learn generalizable shared and unique features across cell types. In stage 2, the model is constrained to learn cell-type specific features, allowing the learning of a set of motifs that are associated to cell-type specificity. In addition to the type of learning used, the data representation, the criteria chosen for model evaluation, and the hyperparameters selected are important factors we account for during the learning phase to achieve a more accurate prediction of binding profiles at cell-type resolution.

Most deep learning approaches, such as DeepBind [25], MTTFSite [69] and DanQ [27], were not used to investigate the differences in ABs for the same TF when analyzing ChIP-seq experiments. We hypothesized that ABs could greatly influence the quality of ChIP-seq experiments. The polyclonal nature of ABs in ENCODE, for example, may result in ABs targeting the same protein to have different specificities, affinities and off-target binding. As a result, due to the lack of study on the consistency of functionality and performance of ABs across TFs ENCODE wide, we separated experiments from different ABs for a particular TF, and investigated the consistency in binding preference across different ABs for the same TF. Overall, we found consistency across ABs for most TFs, although a few results were inconsistent. Lack of consistency may be due to factors such as the quality of the ChIP-seq datasets, the controls selected for peak calling, or the cell types available.

Although we have uncovered substantial evidence for DNA signatures associated to cell-type specific binding, we acknowledge some limitations to our study. First, the fact that a TF does not show cell-type specificity in the cell types available from ENCODE does not imply that it will not show cell-type specificity in other cell types. The human genome contains almost 1400 TFs [72], and despite the enormous effort of the ENCODE consortium, we found only 169 distinct TFs assayed in more than one cell type and meeting our other data set criteria. (This number has increased somewhat since we began our study, but remains far smaller than 1400.) It is thus impossible to detect cell-type specific binding for the vast majority of TFs, and it is uncertain whether other TFs may show specificity in other cell types. This underlines the importance of continued empirical study of TF binding in a wide range of cell types. A second limitation is that, despite best efforts, deep learning can at times fail to solve a prediction problem, even when a solution is possible in principle. There may be TFs for which we failed to detect a cell-type specific signal, even when one is present. On the other hand, our careful checks against overfitting suggest that when a cell-type specific signal is present, it is likely genuine, especially when it is backed up by additional motif enrichment analyses. Thus, our results are best viewed as providing evidence for cell-type specific DNA signatures in many TFs, while providing evidence against the same, without ruling it out, for other TFs. Thirdly, assumptions made regarding the network architecture, such as the 101 bp input sequence or fixed filter widths, may limit the learning capabilities of SigTFB. For instance, its inability to detect widely spaced motifs or motif pairs with fixed spacing, suggests that some DNA signatures relevant for cell-type specificity may be possibly missing. Furthermore, in this work, we used ChIP-seq data due to its high availability and accessibility for multiple TFs and cell types. While ENCODE has many standards in place to ensure high data quality, other experimental approaches, such as ChIP-exo [73] or CUT&TAG [74], may provide less noisy, higher resolution and/or more precise estimates of TF-DNA binding, and thus may ultimately improve the search for DNA signatures. Finally, it is important to note that our confirmatory motif enrichment analysis is limited by the current state of knowledge. Like ENCODE, JASPAR includes data on a relatively small fraction of all human TFs. Not all motifs are in JASPAR, and not having matches may result from a key TF not being in the database. While one could repeat the analysis with motif collections from other databases, such as HOCOMOCO [16] or TRANSFAC [75], a fundamental limitation remains that the majority of known or predicted TFs have not been assayed even once in any cell type. Motif analysis could be extended in other directions, however. For instance, although we opted for a deep learning approach to reduce a priori bias in looking for certain types of motifs, and to avoid a large number of individual or differential motif analyses, one could nevertheless carry out de novo motif finding on all of the datasets [76], and carry out differential motif finding [77], particularly between cell types or groups thereof that our SigTFB analysis suggests are substantially different.

We have proposed a supervised learning problem formulation that allows one to quantify the degree of cell-type specificity in the DNA sequences of a TF's binding sites. We solved that supervised learning problem with a deep learning approach, SigTFB. However, any number of other approaches could be explored, such as position weight matrices, logistic regression, decision trees or forests, support vector machines, or other deep learning formulations. Furthermore, whereas SigTFB's approach makes relatively little a priori assumptions about what may constitute a discriminative DNA signature, one could test alternative representations of peak sequences, for instance using known motifs or k-mers. All the supervised learning data we used is available at https://doi.org/10.20383/103.0605, so that anyone can try alternate approaches. In some scenarios, it may also make sense to alter the supervised learning formulation. For instance, we currently treat each different cell type as a monolithic, distinct entity. But in various senses, some cell types are more naturally similar to others. Perhaps some TFs behave one way in certain cancer types and a different way in healthy cells. Or perhaps a TF behaves one way in brain cells and a different way in the skin. In general, cell types might be represented by some metadata features, and we could learn if any of those metadata features associate with differential DNA signatures.

Conclusion

Many TFs are known to bind to different genomic sites in different cell types. Here, we demonstrated that for many of these TFs, different binding sites are associated with different DNA signatures. We developed a deep learning prediction framework, SigTFB, that is capable of detecting such DNA signatures, and used explanation techniques (tSNE representation embedding, in silico mutagenesis and motif enrichment analysis) to elucidate signatures from the trained networks. Our results have implications for ongoing efforts to predict TF binding in un-assayed cell types: the existence of cell-type specific signatures of binding implies some limitations to the success of such approaches that may not have been previously appreciated. Our findings also have implications for the representation of DNA binding preferences of TFs, suggesting that monolithic, cell-type independent representations, such as PWMs, may not be a satisfactory approximation in the long run for some TFs. Finally, our results set the stage for deeper investigation into mechanisms of differential TF binding, suggesting certain TFs where investigation is more relevant and more likely to succeed.

Methods

SigTFB's two-stage training process

The training procedure we describe below proceeds in two stages, which use slightly different supervised learning formulations, which we call SL1 and SL2. To make clear the difference, we first introduce some notation. Let there be U unified peaks, C total cell types, and let A_{ij} be a binary indicator of whether unified peak $i \in \{1, \ldots, U\}$ includes a peak originally found in cell type $j \in \{1, \ldots, C\}$. Let D_i be the length-404 one-hot encoded DNA sequence of unified peak i.

In SL1, each unified peak contributes one instance to the dataset. The input vector is D_i, the one-hot encoded DNA sequence of the 101bp window centered on the peak, and the output vector is $A_{i\cdot}$, the vector telling which cell types contributed to this peak. This multi-task formulation is used to pre-train part of the network, but is not ultimately the formulation that we want solved. We train using negative log likelihood loss function, and starting from random initial weights.

In SL2, as described above in "A supervised learning formulation for detecting cell-type specific DNA-binding signatures" section, each unified peak contributes $2C$ instances to the dataset, or equivalently, there are *two* instances for each unified peak and each cell type. For unified peak i and cell type j, one of the instances has as input vector D_i concatenated with a length-C binary vector having a 1 in position j. This instance has a single binary output value (or label) which is equal to A_{ij}. Intuitively, this instance can be interpreted like, "The DNA sequence D_i in cell type j was ($A_{ij} = 1$) or wasn't ($A_{ij} = 0$) bound by the TF". This is called a cell-type specific instance, because the target cell type we are querying about is given. The second instance associated to each unified peak and each cell type is just like the first one, except that the length-C binary vector part of the input is set to all zeros. Such an instance, which we call cell-type general, can be interpreted like, "The DNA sequence D_i was ($A_{ij} = 1$) or wasn't ($A_{ij} = 0$) bound, in a cell type who's identity is being kept hidden". We train on this data using the negative log likelihood loss function. The initial weights for the "upper" part of the network are taken from the SL1 training, but they are not frozen and so may change during stage 2 training. Other weights are initialized randomly.

To allow us to optimize hyperparameters and avoid over-fitting the data, we use a nested cross-validation scheme which divides the data into training, validation, and test sets. The outer loop is a standard 10-fold cross validation, which generates a 90% train/10% test split for each fold. Within each training set, we further divide as 80% train/20% validate, where the validation set is used for hyperparameter optimization. Along with network layer size parameters described above, we also optimize learning rate, weight decay, initial weight scales for convolutional and dense layers, and number of training epochs. We train using PyTorch 1.5.0 (GPU) with the Adam optimizer.

Motif analysis

To analyze the peak DNA sequences per cell type per TF-AB model, we use FIMO 5.0.3 to search for motifs in the subsequences using known JASPAR human motifs [15] that are based on at least 1000 sites and have log p-values of at least 100. This gives us a total of 400 JASPAR motifs. For each cell type per TF-AB, and each motif, we find the ratio of the number of significant motif hits identified by FIMO to the number of total peaks for that cell type. By using this approach, we account for enrichment as well as the number of peaks per cell type per TF-AB. To construct the large enrichment heatmap in Fig. 5, we find the top 20 motifs with the highest enrichment ratio per cell type, and take the union of the these motifs across the cell types.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-024-10859-1.

Additional file 1: Supplementary Table 1. Lists the ENCODE datasets on which our study was based, along with the AUC differences achieved on those datasets. Supplementary Table 2. Lists all the motif enrichment scores for all peak datasets. Supplementary Figure S1. Provides zoom-ins of certain clusters of enrichment within Figure 5.

Authors' contributions

All authors contributed to the research design and writing of the manuscript. The work was primarily carried out by AA.

Funding

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), funding reference number RGPIN-2019-06604 to TJP and a Compute Canada (www.computecanada.ca) Resources-for-Research-Groups grant to TJP. This research was enabled by support provided by a Queen Elizabeth II Graduate Scholarship in Science and Technology (QEII-GSST) to AA. None of the agencies that funded this work had any role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

The processed data used for deep learning is available at https://doi.org/10.20383/103.0605. The source code for SigTFB including the pre-processing of ChIP-seq data, classification and downstream genomic analysis, is available at https://github.com/aawdeh/SigTFB.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

All authors have read and consented to publishing this manuscript.

Competing interests

The authors declare no competing interests.

Received: 10 July 2024 Accepted: 2 October 2024
Published online: 14 October 2024

References

1. Bintu L, et al. Transcriptional regulation by the numbers: models. Curr Opin Genet Dev. 2005;15:116–24.
2. Desvergne B, Michalik L, Wahli W. Transcriptional regulation of metabolism. Physiol Rev. 2006;86:465–514.
3. Lee TI, Young RA. Transcriptional regulation and its misregulation in disease. Cell. 2013;152:1237–51.
4. Matys V, et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 2006;34:D108–10.
5. Bryne JC, et al. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Res. 2007;36:D102–6.
6. Soleimani VD, et al. Cis-regulatory determinants of MyoD function. Nucleic Acids Res. 2018;46:7221–35.
7. Consortium EP, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57.
8. Lee B-K, et al. Cell-type specific and combinatorial usage of diverse transcriptionfactors revealed by genome-wide binding studies in multiple human cells. Genome Res. 2012;22:9–24.
9. Benedetti M, Levi A, Chao MV. Differential expression of nerve growth factor receptors leads to altered binding affinity and neurotrophin responsiveness. Proc Natl Acad Sci. 1993;90:7859–63.
10. Srivastava D, Mahony S. Sequence and chromatin determinants of transcription factor binding and the establishment of cell type-specific binding patterns. Biochim Biophys Acta (BBA) Gene Regul Mech. 2020;1863:194443.
11. Brand M, et al. Dynamic changes in transcription factor complexes during erythroid differentiation revealed by quantitative proteomics. Nat Struct Mol Biol. 2004;11:73–80.
12. Pilpel Y, Sudarsanam P, Church GM. Identifying regulatory networks by combinatorial analysis of promoter elements. Nat Genet. 2001;29(153):159.
13. Nie Y, Shu C, Sun X. Cooperative binding of transcription factors in the human genome. Genomics. 2020;112:3427–34.
14. Lowen M, Scott G, Zwollo P. Functional analyses of two alternative isoforms of the transcription factor Pax-5. J Biol Chem. 2001;276:42565–74.
15. Castro-Mondragon JA, et al. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2022;50:D165–73.
16. Kulakovskiy IV, et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-seq analysis. Nucleic Acids Res. 2018;46:D252–9.
17. Ogawa N, Biggin MD. High-throughput SELEX determination of DNA sequences bound by transcription factors in vitro. Gene Regul Netw Methods Protoc. 2012;786:51–63.
18. Gertz J, et al. Distinct properties of cell-type-specific and shared transcription factor binding sites. Mol Cell. 2013;52:25–36.
19. Zhang S, et al. OCT4 and PAX6 determine the dual function of SOX2 in human ESCs as a key pluripotent or neural factor. Stem Cell Res Ther. 2019;10:1–14.
20. Wang J, et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res. 2012;22:1798–812.
21. Arvey A, Agius P, Noble WS, Leslie C. Sequence and chromatin determinants of cell type-specific transcription factor binding. Genome Res. 2012;22:1723–34.
22. Keilwagen J, Posch S, Grau J. Accurate prediction of cell type-specific transcription factor binding. Genome Biol. 2019;20:1–17.
23. McLeay RC, Bailey TL. Motif enrichment analysis: a unified framework and an evaluation on ChIP data. BMC Bioinformatics. 2010;11:1–11.
24. Lesluyes T, Johnson J, Machanick P, Bailey TL. Differential motif enrichment analysis of paired ChIP-seq experiments. BMC Genomics. 2014;15:1–13.
25. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nat Biotechnol. 2015;33:831–8.
26. Hassanzadeh H, Wang MD. DeeperBind: Enhancing prediction of sequence specificities of DNA binding proteins. Los Alamitos: IEEE Computer Society; 2016. p. 178–83.
27. Quang D, Xie X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Res. 2016;44:e107–e107.
28. Chen C, et al. DeepGRN: prediction of transcription factor binding site across cell-types using attention-based deep neural networks. BMC Bioinformatics. 2021;22:1–18.
29. Quang D, Xie X. FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. Methods. 2019;166:40–7.
30. Li H, Guan Y. Fast decoding cell type-specific transcription factor binding landscape at single-nucleotide resolution. Genome Res. 2021;31:721–31.
31. Zhang Y, Wang Z, Zeng Y, Zhou J, Zou Q. High-resolution transcription factor binding sites prediction improved performance and interpretability by deep learning method. Brief Bioinform. 2021;22:bbab273.
32. Zhang Q, et al. Base-resolution prediction of transcription factor binding signals by a deep learning framework. PLoS Comput Biol. 2022;18:e1009941.
33. Cao L, Liu P, Chen J, Deng L. Prediction of transcription factor binding sites using a combined deep learning approach. Front Oncol. 2022;12:893520.

34. Ng JW, Ong EH, Tucker-Kellogg L, Tucker-Kellogg G. Deep learning for de-convolution of Smad2 versus Smad3 binding sites. BMC Genomics. 2022;23:525.
35. Ding P, et al. DeepSTF: predicting transcription factor binding sites by interpretable deep neural networks combining sequence and shape. Brief Bioinform. 2023;24:bbad231.
36. Zhang J, Liu B, Wu J, Wang Z, Li J. DeepCAC: a deep learning approach on DNA transcription factors classification based on multi-head self-attention and concatenate convolutional neural network. BMC Bioinformatics. 2023;24:345.
37. Wang K, et al. BERT-TFBS: a novel BERT-based model for predicting transcription factor binding sites by transfer learning. Brief Bioinform. 2024;25:bbae195.
38. Zhuang J, et al. MulTFBS: A spatial-temporal network with multichannels for predicting transcription factor binding sites. J Chem Inf Model. 2024;64(10):1549–9596.
39. Andrews G. Deep learning as a tool to better understand transcription factor binding across cell types and species. Ph.D. thesis, UMass Chan Medical School; 2024.
40. Eraslan G, Avsec Ž, Gagneur J, Theis FJ. Deep learning: new computational modelling techniques for genomics. Nat Rev Genet. 2019;20:389–403.
41. Zhang S, et al. Assessing deep learning methods in cis-regulatory motif finding based on genomic sequencing data. Brief Bioinform. 2022;23:bbab374.
42. Novakovsky G, Dexter N, Libbrecht MW, Wasserman WW, Mostafavi S. Obtaining genetics insights from deep learning via explainable artificial intelligence. Nat Rev Genet. 2023;24:125–37.
43. Singh G, et al. A exible repertoire of transcription factor binding sites and a diversity threshold determines enhancer activity in embryonic stem cells.Genome Res. 2021;31:564–575.
44. Zheng A, et al. Deep neural networks identify sequence context features predictive of transcription factor binding. Nat Mach Intel. 2021;3:172–80.
45. Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. 2016;26:990–9.
46. Nair S, Kim DS, Perricone J, Kundaje A. Integrating regulatory DNA sequence and gene expression to predict genome-wide chromatin accessibility across cellular contexts. Bioinformatics. 2019;35:i108–16.
47. Balandat M, et al. BoTorch: programmable bayesian optimization in PyTorch. 2019. arxiv e-prints arXiv–1910.
48. Maekawa T, et al. Social isolation stress induces ATF-7 phosphorylation and impairs silencing of the 5-HT 5B receptor gene. EMBO J. 2010;29:196–208.
49. Chen M, et al. Emerging roles of activating transcription factor (ATF) family members in tumourigenesis and immunity: Implications in cancer immunotherapy. Genes Dis. 2021;9(4):981–99.
50. Gozdecka M, Breitwieser W. The roles of ATF2 (activating transcription factor 2) in tumorigenesis. Biochem Soc Trans. 2012;40:230–4.
51. Meijer BJ, et al. ATF2 and ATF7 are critical mediators of intestinal epithelial repair. Cell Mol Gastroenterol Hepatol. 2020;10:23–42.
52. Kim S, Yu N-K, Kaang B-K. CTCF as a multifunctional protein in genome regulation and gene expression. Exp Mol Med. 2015;47:e166–e166.
53. Chen H, Tian Y, Shu W, Bo X, Wang S. Comprehensive identification and annotation of cell type-specific and ubiquitous CTCF-binding sites in the human genome. PLoS ONE. 2012;7:e41374.
54. Holwerda SJB, de Laat W. CTCF: the protein, the binding partners, the binding sites and their chromatin loops. Philos Trans R Soc B Biol Sci. 2013;368:20120369.
55. Li YE, et al. An atlas of gene regulatory elements in adult mouse cerebrum. Nature. 2021;598:129–36.
56. BRAIN Initiative Cell Census Network (BICCN). A multimodal cell census and atlas of the mammalian primary motor cortex. Nature. 2021;598:86–102.
57. Zu S, et al. Single-cell analysis of chromatin accessibility in the adult mouse brain. Nature. 2023;624:378–89.
58. Sams DS, et al. Neuronal CTCF is necessary for basal and experience-dependent gene regulation, memory formation, and genomic structure of BDNF and Arc. Cell Rep. 2016;17:2418–30.
59. Dang CV. MYC on the path to cancer. Cell. 2012;149:22–35.
60. Davudian S, Mansoori B, Shajari N, Mohammadi A, Baradaran B. BACH1, the master regulator gene: a novel candidate target for cancer therapy. Gene. 2016;588:30–7.
61. Guo X, Yang M, Gu H, Zhao J, Zou L. Decreased expression of SOX6 confers a poor prognosis in hepatocellular carcinoma. Cancer Epidemiol. 2013;37:732–6.
62. Wysocka J, Reilly PT, Herr W. Loss of HCF-1-chromatin association precedes temperature-induced growth arrest of tsBN67 cells. Mol Cell Biol. 2001;21:3820–9.
63. Maslova A, et al. Deep learning of immune cell differentiation. Proc Natl Acad Sci. 2020;117(25655):25666.
64. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. Genome Biol. 2007;8:1–9.
65. De Graeve F, et al. Role of the ATFa/JNK2 complex in jun activation. Oncogene. 1999;18:3491–500.
66. Fornes O, et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2020;48:D87–92.
67. Ambrosini G, et al. Insights gained from a comprehensive all-against-all transcription factor binding motif benchmarking study. Genome Biol. 2020;21:1–18.
68. Castro-Mondragon JA, Jaeger S, Thieffry D, Thomas-Chollier M, Van Helden J. RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. Nucleic Acids Res. 2017;45:e119–e119.
69. Zhou J, et al. MTTFsite: cross-cell type TF binding site prediction by using multi-task learning. Bioinformatics. 2019;35:5067–77.
70. Phuycharoen M, et al. Uncovering tissue-specific binding features from differential deep learning. Nucleic Acids Res. 2020;48:e27–e27.
71. Novakovsky G, Saraswat M, Fornes O, Mostafavi S, Wasserman WW. Biologically relevant transfer learning improves transcription factor binding prediction. Genome Biol. 2021;22:1–25.
72. Pechenick DA, Payne JL, Moore JH. Phenotypic robustness and the assortativity signature of human transcription factor networks. PLoS Comput Biol. 2014;10:e1003780.
73. Rhee HS, Pugh BF. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell. 2011;147:1408–19.
74. Kaya-Okur HS, et al. Cut &tag for efficient epigenomic profiling of small samples and single cells. Nat Commun. 2019;10:1930.
75. Wingender E, et al. TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res. 2000;28(316):319.
76. Kheradpour P, Kellis M. Systematic discovery and characterization of regulatory motifs in encode TF binding experiments. Nucleic Acids Res. 2014;42:2976–87.
77. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME suite. Nucleic Acids Res. 2015;43:W39–49.

Publisher's Note