BMC Genomics

RESEARCH

Open Access

Neural network modeling of differential binding between wild-type and mutant CTCF reveals putative binding preferences for zinc fingers 1–2

Irene M. Kaplow[1,2*], Abhimanyu Banerjee[3] and Chuan Sheng Foo[1,4*]

Abstract

Background: Many transcription factors (TFs), such as multi zinc-finger (ZF) TFs, have multiple DNA binding domains (DBDs), and deciphering the DNA binding motifs of individual DBDs is a major challenge. One example of such a TF is *CCCTC-binding factor* (CTCF), a TF with eleven ZFs that plays a variety of roles in transcriptional regulation, most notably anchoring DNA loops. Previous studies found that CTCF ZFs 3–7 bind CTCF's core motif and ZFs 9–11 bind a specific upstream motif, but the motifs of ZFs 1–2 have yet to be identified.

Results: We developed a new approach to identifying the binding motifs of individual DBDs of a TF through analyzing chromatin immunoprecipitation sequencing (ChIP-seq) experiments in which a single DBD is mutated: we train a deep convolutional neural network to predict whether wild-type TF binding sites are preserved in the mutant TF dataset and interpret the model. We applied this approach to mouse CTCF ChIP-seq data and identified the known binding preferences of CTCF ZFs 3–11 as well as a putative GAG binding motif for ZF 1. We analyzed other CTCF datasets to provide additional evidence that ZF 1 is associated with binding at the motif we identified, and we found that the presence of the motif for ZF 1 is associated with CTCF ChIP-seq peak strength.

Conclusions: Our approach can be applied to any TF for which in vivo binding data from both the wild-type and mutated versions of the TF are available, and our findings provide new potential insights binding preferences of CTCF's DBDs.

Keywords: Mutated transcription factor, CTCF, Zinc finger, Motif, Deep neural network, Binding strength

Background

Mutations of individual DNA binding domains (DBDs) within transcription factors (TFs) have been associated with developmental [1, 2] and bleeding [3] disorders, and differences between species in individual DBDs within TFs have been associated with species-specific gene expression [4] and speciation [5]. Although DNA binding motifs of thousands of metazoan TFs have been characterized, many TFs have multiple DNA binding domains (DBDs) whose specific binding preferences are unknown. In fact, the most common TF family in humans, Cys2His2 (C2H2) zinc finger (ZF) TFs [6, 7], consists of TFs with multiple ZF DBDs, and many of these ZFs' individual binding preferences have not been investigated.

A previous study investigated the binding preferences of ZFs within C2H2 ZF TFs by doing in vitro Bacterial 1-Hybrid (B1H) assays of over 160,000 ZFs [8] to determine the individual 3 bp [9] binding preferences of each

*Correspondence: ikaplow@cs.stanford.edu; csfoo@cs.stanford.edu
[1] Departments of Computer Science, Stanford University, 240 Pasteur Drive, Stanford, California 94305, USA
Full list of author information is available at the end of the article

ZF. The study then presented a machine learning model trained on this data to predict the position weight matrices (PWMs) of C2H2 ZF TFs. Unfortunately, for many TFs, less than two thirds of PWM columns were predicted correctly, demonstrating the limitations of using in vitro assays of individual DBDs to determine binding preferences of DBDs within a full TF. Another study described how DBDs can influence each other's binding within the context of a TF [10], further illustrating the limitations of studying binding preferences of individual DBDs out of context.

To identify the binding preferences of DBDs within a TF within the context of the other DBDs, previous studies have introduced loss-of-function mutations within specific DBDs, assayed the sequences to which the mutants bind, and used the results of the assay to determine the specific components of TFs' motifs that interact with a DBD [1, 11, 12]. In particular, one of these studies induced loss-of-function histidine-to-arginine mutations separately each of the 11 ZFs of mouse *CCCTC-binding factor* (CTCF), a C2H2 ZF TF that has been implicated in diverse roles in transcriptional regulation [13, 14] due to its ability to anchor DNA loops [15, 16] likely through interactions with cohesin [17–19], and did ChIP-seq on each mutant [11]. The study found that ZFs 3 through 7 interact with part of CTCF's known core motif, a 19 base-pair sequence that has been shown to bind CTCF in many studies [20]. The study also found that ZFs 8 through 11 interact with an upstream motif that had been identified by a few earlier studies [21–23] (Supplemental Figure 1), demonstrating the viability of assaying binding of mutated TFs to understand individual DBD binding preferences. These findings were supported by additional studies; one study used CTCF deletions to show that only ZFs 4 through 7 interact with base pairs 4 through 15 of its core motif [24], and another used electrophoretic mobility shift assays (EMSA) of CTCF with parts of its motif mutated to suggest that ZF 7 or 8 binds to base pairs four through six of the core motif [25]. In addition, recent studies showed that mutations in ZFs 1 and 10 disrupt DNA loops [26, 27]; another recent study showed that CTCF-s, a CTCF isoform that does not have ZFs 1–3, is unable to interact with cohesin [28]; and an additional recent study showed that mutations in CTCF's ZFs have been found in cancer and that some, including a mutation in ZF 2, lead to a loss of binding [29]. These studies demonstrate the potential value in understanding the ways that CTCF's ZFs that do not bind to the core motif interact with DNA.

To better leverage in vivo experiments of mutated TFs to decipher the binding preferences of individual DBDs, we developed a novel approach to analyzing the data from mutant TF ChIP-seq experiments [11]. In contrast to the earlier study, which did de novo motif discovery on the sequences within the peaks from wild-type CTCF and then scanned the peaks from the mutated CTCF for the discovered motifs [11], we directly leverage the differences between the wild-type and mutant datasets. We do this by setting up a differential peak prediction task, in which we train a deep convolutional neural network [30, 31] to use DNA sequence to predict whether a peak from wild-type TF ChIP-seq is preserved in the mutant dataset or is significantly stronger in the wild-type dataset. Our intuition is that, if a model can predict whether a peak is significantly stronger in the wild-type dataset than in the mutant dataset, then the model should have learned sequence patterns related to the binding preferences of the mutated DBD, and interpreting the model should reveal these binding preferences.

We applied our approach to the CTCF mutant ChIP-seq datasets [11] and interpreted what each model learned to identify motifs associated with each ZF. We trained a separate model for every ZF because we identified over ten thousand significantly differential peaks between the wild-type and each mutant, suggesting that every ZF plays some role in CTCF binding. Our model interpretations recapitulated earlier findings about which ZFs interact with the core and upstream motifs, illustrating the success of our approach. The interpretations also identified a novel downstream motif, GAGCCA, that may be bound by ZF 1. We found that the core motif followed by our discovered downstream motif occurs in CTCF HT-SELEX reads from the final cycle, that the core motif followed by the discovered downstream motif occurs more frequently in CTCF ChIP-seq peaks that do not overlap CTCF-s ChIP-seq peaks than in those that do overlap CTCF-s ChIP-seq peaks, and that the discovered downstream motif matches in vitro data based computational predictions of the ZF 1 motif and has been shown to bind CTCF in a previous EMSA study. We also found that the presence of the discovered downstream motif is correlated with CTCF peak strength. Our approach can be applied to any TF with multiple DBDs for which wild-type and mutated DBD in vivo binding data are available, and our results from applying our approach to CTCF provide the first insights into the in vivo binding sequence associated CTCF's most downstream ZF.

Results

Putative motifs of CTCF's zinc fingers identified by interpreting wild-type versus mutant differential peak prediction models

To identify motifs related to the binding of each ZF in CTCF, we trained and interpreted a neural network for predicting whether a peak would be significantly weaker according to DESeq2 [32] in the mutant dataset than in

the wild-type dataset (Methods, Supporting Website). Interestingly, although multiple ZFs have been implicated in interacting with RNA [19, 27] and ZF 1 is thought to have more interactions with RNA than DNA [18], for every ZF including ZF 1, we found over ten thousand peaks that are significantly stronger in the wild-type than they are in the mutant (13,307 for ZF 1, Supplemental Table 1), suggesting that every ZF may play some role in CTCF's interaction with DNA. We therefore trained a separate model for each ZF mutant ChIP-seq dataset from [11] (Supplemental Figure 1). Upon finding that our models had good performance, we used DeepLIFT with the Rescale rule [33] followed by TF-MoDISco [34] to identify motifs (called "TF-MoDISco motifs") that the model had learned (Fig. 1, Methods, Supporting Website). We identified TF-MoDISco motifs for all the ZFs in CTCF.

Neural network outperforms models with original motif hit scores as features

To evaluate our neural network and the TF-MoDISco motifs, we compared three approaches for predicting whether a CTCF peak would be substantially weaker in a mutant CTCF dataset: our neural networks, logistic regressions with the motif hit score of the best TF-MoDISco motif hit as the feature, and logistic regressions with motif hit scores of motif hits from [11] as features. We found that all models performed well for ZFs 3–7, but our neural networks and the logistic regressions with the TF-MoDISco motif hit score alone had substantially better performance than the logistic regressions with the original motif hit scores for the other ZFs (Fig. 2a). For ZFs 8–11, we also compared the performances of our neural networks and the logistic regressions with the TF-MoDISco motif hit score to the performances of the score from single motif consisting of the original upstream motif followed by five base pairs (the most common spacing found in [11]) followed by the original core motif. These logistic regressions' performances were comparable to those of the logistic regressions with the TF-MoDISco motif hit score for ZFs 9–11 and worse than other methods for ZF 8 (Supplemental Figure 2).

Important features learned by neural networks include known motifs for zinc fingers 3–11 and novel motifs for zinc fingers 1–2

Interpreting neural networks revealed known CTCF motifs

We compared the TF-MoDISco motifs to known CTCF motifs. We found that our neural network learned motifs similar to the known core motif as being indicative of a stronger peak in the wild-type for ZFs 3–7 and motifs similar to the known upstream motif for ZFs 8-11 as being indicative of a stronger peak in the wild-type,

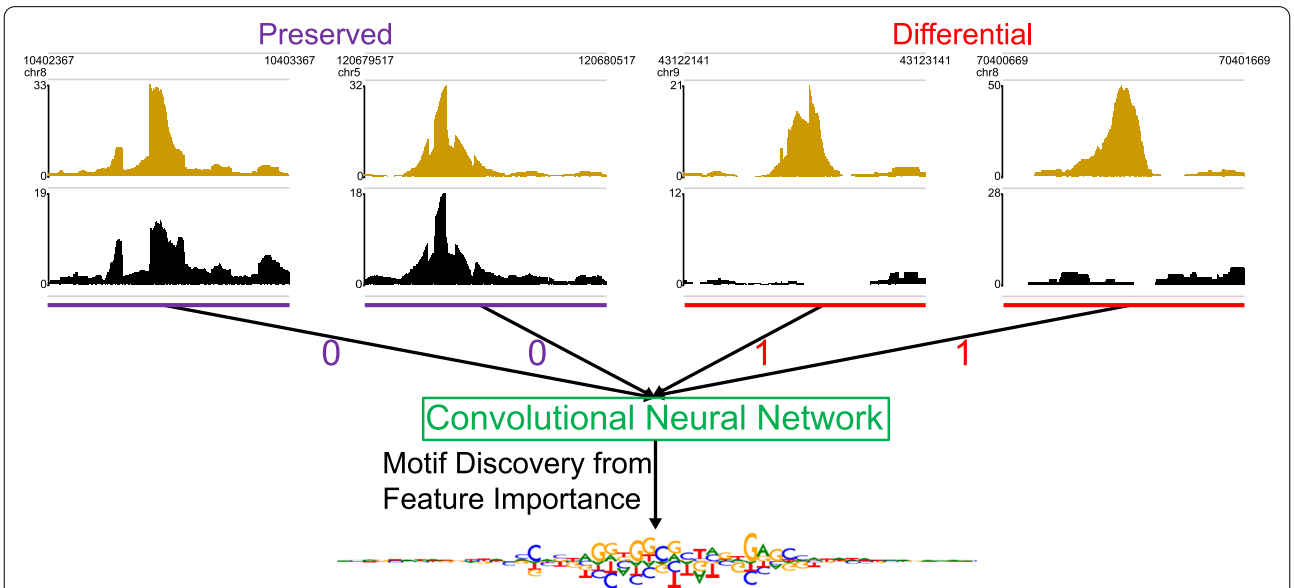

Fig. 1 Using Differential Peak Prediction to Identify Motifs of Different DNA Binding Domains. To identify the motif of a DBD, we train a deep convolutional neural network to predict whether a TF ChIP-seq peak is preserved or significantly weaker in a dataset from that TF with a mutated DBD relative to a dataset from the wild-type TF. We then use DeepLIFT followed by TF-MoDISco to identify the motifs that the neural network learned. The browser tracks in this figure are pooled replicate fold-change bigWigs from wild-type CTCF and CTCF with ZF1 mutated from [11] and were visualized using the WashU Epigenome Browser [35] with assembly mm10 [36]. The motif logo in this figure is the motif we discovered when interpreting the model for the ZF 1 mutant

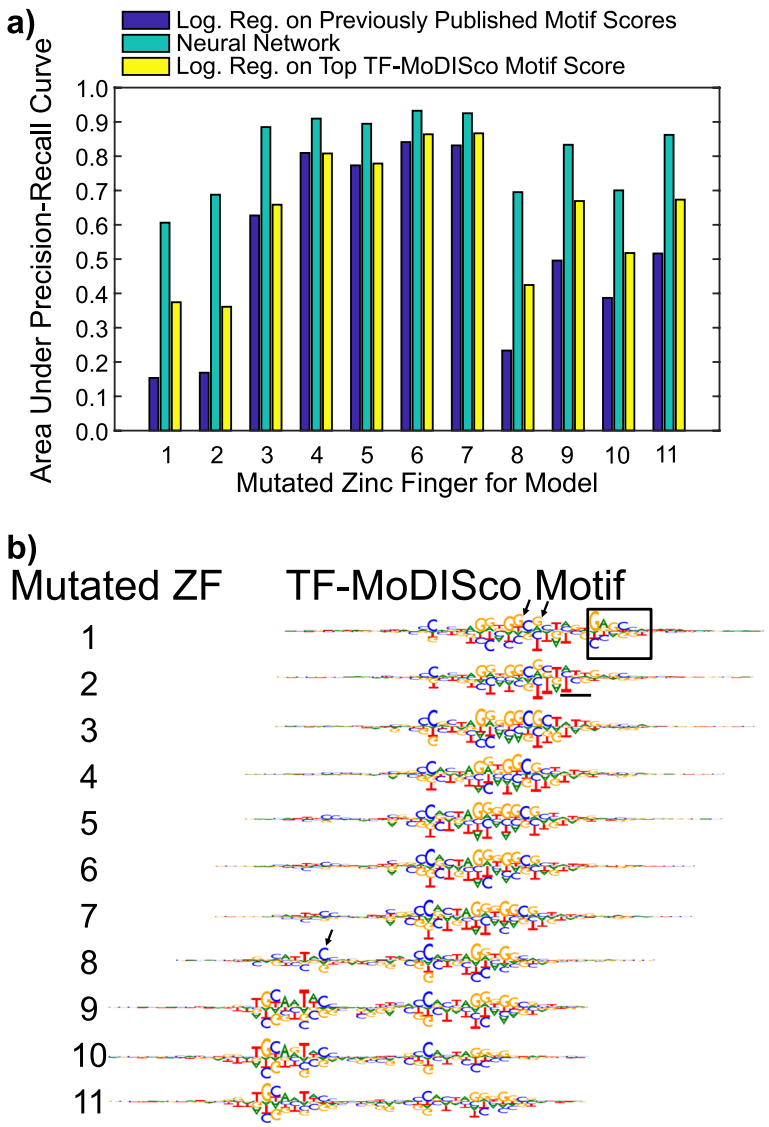

Fig. 2 Performance of Neural Networks. **a** We compared the performance of our neural networks to those of logistic regressions in which the features were the motif hit scores of the motifs from [11]. We also compared both sets of models to logistic regressions with the top TF-MoDISco motif hit scores as their only features. Performance was measured by the area under the precision-recall curve (AUPRC). **b** We aggregated the hypothetical scores of the seqlets corresponding to the motifs from DeepLIFT followed by TF-MoDISco to visualize the TF-MoDISco motifs. The box indicates the discovered downstream motif, and the underlined part indicates the weak putative motif for ZF 2. The TF-MoDISco motif for ZF 1 has a G or a T at a position where the other TF-MoDISco motifs have a G (indicated by first arrow) and a G or an A at a position where the other TF-MoDISco motifs have a G (indicated by second arrow). The TF-MoDISco motif for ZF 8 emphasizes a downstream nucleotide in the upstream motif (indicated by arrow)

which is consistent with the findings of [11] (Fig. 2b). Previous studies identified the five base pair spacing in our top TF-MoDISco motif as the most common spacing between the core and upstream motifs but also found that a six base pair spacing occurred frequently [11, 21–23, 37]. We therefore investigated all the TF-MoDISco motifs for each neural network (Supporting Website) and, for ZFs 9–11, found that the second highest-ranked TF-MoDISco motif (the TF-MoDISco motif with the second highest number of supporting seqlets) was the upstream motif, followed by six base pairs, followed by the core motif (Supplemental Figure 3).

Interpreting neural networks revealed novel motifs for ZFs 1–2 confirmed by CTCF HT-SELEX data

When identifying the important sequences for the neural networks for the mutants of ZFs 1–2, we discovered a novel downstream GAGCCA motif occurring 2 bp downstream of the core motif and a weaker ATT motif connecting the core and discovered downstream motif as being indicative of a stronger peak in the wild-type (Fig. 2b, Supplemental File 1). To investigate if CTCF can bind these motifs, we re-analyzed published HT-SELEX data for CTCF [38] to determine if there is an enrichment of reads containing the core followed by the discovered downstream motif in cycle 4 (final round) relative to cycle 0 (control) (Methods). First, to evaluate the reliability of this approach, we did this for the core motif only and found a significant enrichment ($p = 0.0$) (Supplemental Figure 4). We then found an enrichment for the core motif followed by the discovered downstream motif ($p = 1.17 \times 10^{-245}$) (Fig. 3b). In fact, the HT-SELEX reads with the best matches to the core motif followed by the downstream motif (FIMO q-value < 0.001) have ATT connecting the two motifs (Supplemental Figure 5), which is the putative motif that we found for ZF 2.

Discovered downstream motif is associated with lack of CTCF-s binding

We also compared the *p*-values of the motif hits for the core followed by the discovered downstream motif in HeLa cell ChIP-seq peaks for CTCF and CTCF-s – the alternative isoform of CTCF that is missing ZF 1, ZF 2, and part of ZF 3 – from [28]. We found that these p-values were significantly lower (negative log base ten of the p-values was significantly higher) for the CTCF peaks that do not overlap CTCF-s peaks than they were for the CTCF peaks that do overlap CTCF-s peaks ($p = 4.00 \times 10^{-6}$), suggesting that the lack of ZFs 1–3 is associated with a lack of binding to the downstream motif. We then investigated whether this result could be explained by the core motif followed by the downstream motif occurring more frequently in CTCF binding sites that are less reproducible across experiments. We did this by downloading HeLa cell CTCF ChIP-seq peaks from ENCODE [39] and comparing the core followed by the discovered downstream motif hit *p*-values for the CTCF ChIP-seq peaks from [28] that overlap the ENCODE CTCF ChIP-seq peaks to those that do not. For this comparison, we found a significant trend in the opposite direction ($p = 2.77 \times 10^{-322}$) (Fig. 3c). Since our putative motifs for ZFs 1, 2, and 3 are all part of the core followed by the downstream motif, we cannot be certain of the relative contributions of the binding of each of these 3 ZFs to these results, but they do suggest that the lack of the core followed by the discovered downstream motif is associated with the lack of binding of CTCF ZFs 1–3.

Discovered downstream motif has supporting evidence from previous CTCF studies

We obtained additional evidence that our discovered downstream motif interacts with CTCF. The most non-degenerate part of this motif (GAG) is almost identical to the computationally predicted motif for ZF 1 according to multiple models that were trained on in vitro ZF binding data from B1H assays (Fig. 3a, Supplemental Figure 6), suggesting that ZF 1 interacts with this downstream motif [40–42]. In addition, a recent study showed that the upstream four nucleotides of this downstream motif are found at CTCF sites in the mouse IgH locus; this study did EMSA on multiple variants of the CTCF motif including two variants containing these upstream four nucleotides and found that CTCF was able to bind both variants [43]. Furthermore, the downstream 3 bp of this motif (CCA) is similar to the upstream 3 bp of the 4 bp downstream motif identified in CTCF-cohesin co-binding sites in [44]. Despite this evidence suggesting the existence of our downstream motif, this motif has not been previously shown to directly interact with CTCF ZF 1 in vivo.

Neural networks' nucleotide-level relative importance scores reveal putative combinatorial binding preferences that were supported by in vitro TF binding assays

Our neural networks' differences between relative importance scores of nucleotides in motifs for different ZFs provided potential insights into additional differences between CTCF's binding preferences when different ZFs interact with DNA. For example, the TF-MoDISco motif for the mutant of ZFs 1–2 had a degenerate position in the core motif that could be a G or a T and another that could be a G or an A. In contrast, the TF-MoDISco motifs for the mutants of ZFs 3–7 placed a substantially stronger importance on the G than the T in the first position, and the TF-MoDISco motifs for the mutants of ZFs 8–11 placed no importance on the T in that position.

(See figure on next page.)

Fig. 3 Comparisons of Discovered Downstream Motif to Other CTCF Data. **a** We compared our TF-MoDISco motifs from the mutants of ZFs 1 and 2 to aggregated reads from CTCF HT-SELEX cycle 4 and to computationally predicted motifs of CTCF's DBDs from the RCADE2 model, which was trained on in vitro B1H ZF binding data. **b** We compared motif hits of the core followed by discovered downstream motif in reads from CTCF HT-SELEX data in cycle 0 to cycle 4. **c** We compared the strength of the core followed by discovered downstream motif in HeLa cell *CTCF* peaks from [28] to HeLa peaks from CTCF's alternative isoform from the same study and HeLa CTCF peaks from ENCODE [39]

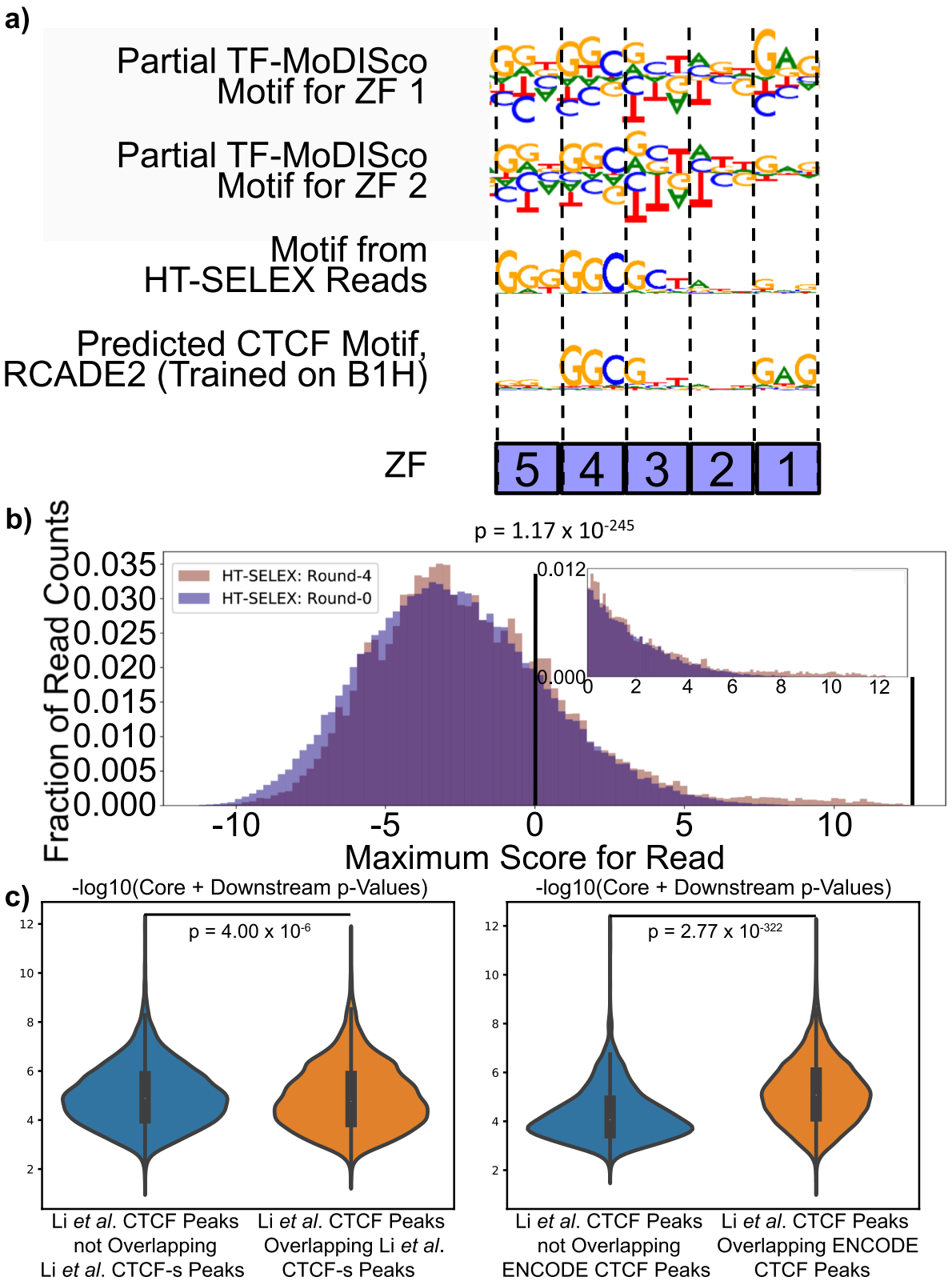

Fig. 3 (See legend on previous page.)

Interestingly, the EMSA done on the core followed by the discovered downstream motif had a T in that position [43], showing that CTCF can bind to the core motif when there is a T in that position and the discovered downstream motif is present. Likewise, the TF-MoDISco motifs for the mutants of ZFs 3–11 placed no importance on the G in the second position, providing another example of degeneracy being tolerated only in the presence of the downstream motif (Fig. 2b).

The neural networks for ZF 8 also placed higher importance on the downstream nucleotides of the upstream motif than did the neural networks for other ZF mutants (Fig. 2b). Additionally, the neural networks for ZFs that are thought to interact with upstream parts of the core motif placed a weak importance on the upstream motif, while the neural networks for ZFs that are thought to interact with downstream parts of the core motif placed a weak importance on the downstream motif. Thus, in addition to identifying putative in vivo motifs for ZFs 1–2, and our neural networks provided insights into the relative importance of various parts of motifs for interactions between DNA and different ZFs.

Presence of discovered downstream motif is associated with CTCF peak strength

Since the discovered downstream motif is present in only a strict subset of wild-type CTCF peaks, we investigated whether the presence of this motif is associated with other differences between CTCF peaks. We identified occurrences of the upstream, core, and discovered downstream motifs and combinations of these motifs in the wild-type CTCF peaks using FIMO [45] (Supporting Website). We then examined the relationship between motif presence and peak strength (Methods). We found that peaks with the core motif and the discovered downstream motif tend to be stronger than peaks with the core motif that do not have the discovered downstream motif ($p = 3.24 \times 10^{-106}$). Since the presence of a motif in a peak is dependent on the motif hit cutoff, we also applied a stricter threshold for the motif hit cutoff and obtained a similar result ($p = 7.77 \times 10^{-89}$). In addition, we evaluated the relationship between the motif hit q-values of the different motif combinations and the peak signal and found that the correlation is significantly stronger for the core motif followed by the discovered downstream motif than it is for the core motif alone ($p = 2.17 \times 10^{-13}$, Supplemental Figure 7). We repeated this for mouse liver and heart TF ChIP-seq peaks [46]. We obtained similar results for the default FIMO motif hit cutoff (liver $p = 2.81 \times 10^{-78}$, heart $p = 7.65 \times 10^{-65}$), a stricter FIMO motif hit cutoff (liver $p = 3.03 \times 10^{-25}$, heart $p = 5.22 \times 10^{-65}$), and the relationship between motif hit q-values and peak signals (liver $p = 1.10 \times 10^{-3}$, heart $p = 2.90 \times 10^{-4}$), suggesting that the relationship between the presence of the downstream motif and CTCF binding strength is consistent across tissues.

Discussion

We developed a new approach for discovering in vivo TF binding motifs for different DBDs within a TF by training and interpreting a neural network for predicting whether a wild-type TF ChIP-seq peak will be significantly stronger than the corresponding peak from mutated TF ChIP-seq data. We applied this approach to a dataset with ChIP-seq from wild-type and mutated CTCF. We identified the known motifs of CTCF ZFs 3–11, demonstrating the success of our approach. We also discovered a putative interaction between CTCF ZF 1 and a novel downstream GAGCCA motif as well as a putative weak interaction between CTCF ZF 2 and an ATT motif connecting the core and discovered downstream motifs (Fig. 4). Our discovered downstream motif was supported by in vitro studies of CTCF from HT-SELEX [38] and EMSA [43] and in vitro ZF studies from B1H assays [40, 42], the weak putative motif for ZF 2 was supported by CTCF HT-SELEX data, and the core followed by discovered downstream motif occurs more frequently in CTCF ChIP-seq peaks that are not bound by CTCF-s than by those that are.

However, we cannot be certain if our discovered motifs for ZFs 1–2 directly interact with ZFs 1–2 or tend to be present these when ZFs are indirectly stabilizing CTCF-DNA interactions. Previous studies suggested that ZF 1 interacts primarily with RNA [18], the in vitro HT-SELEX and EMSA assays supporting our findings were not limited to ZFs 1–2 [38, 43], and the stronger core followed by downstream motif matches in the CTCF peaks not bound by CTCF-s might be largely explained by ZF 3, which has been shown to directly interact with DNA [47, 48]. In addition, the motif for ZF2 was extremely weak

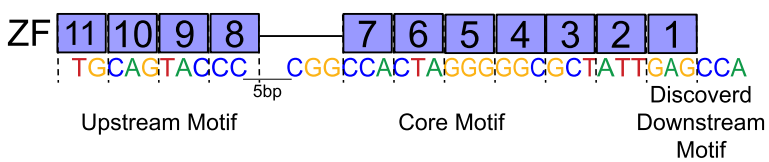

Fig. 4 Proposed Motif for CTCF Based on Findings from Interpreting Neural Networks

(Fig. 2b) and not predicted to interact with ZF 2 according to machine learning models trained on in vitro ZF binding data (Fig. 3a, Supplemental Figure 6), suggesting that this ZF may not directly interact with DNA. Likewise, a previous study crystalized CTCF interacting with DNA and neither obtained crystals for ZF 1 nor found base-specific interactions for ZF 2 [47]. While most of this study focused on an 18 bp motif that approximately matched our motifs for ZFs 3–7 and did not include our motifs for ZFs 1–2, thus not contradicting our results, the study had a short analysis on the interactions of ZFs 2–7 with a longer motif, whose downstream part was AGT (we found ATT in the ZF 2 motif and AGT in the ZF 1 motif) followed by GAG, and found that no crystals were obtained for ZF 1 and that ZF 2 did not directly interact with DNA [47, 49]. On the other hand, a recent study used molecular dynamic simulations to suggest that human CTCF mutation L309P, a mutation in ZF 2 that occurs in some cancer tumors, leads to both the loss and the creation of bonds between CTCF and DNA, even though the ZF and its mutant were found to face away from DNA, suggesting that mutating parts of ZFs that do not directly contact DNA can still lead to binding changes [29]. Other studies have shown that some ZFs interact with DNA in some contexts and RNA in others [50, 51]. While, to fully demonstrate that ZFs 1–2 bind the motifs that we found, we would need to experimentally test if ZFs 1–2 alone can interact with these sequences using an assay like EMSA, our combination of existing and novel approaches for using additional datasets to support our findings provide a foundation for following up on potential motif discoveries.

A recent study did ChIP-seq on wild-type CTCF in mouse embryonic stem cells as well as CTCF with deletions of ZFs 1, 8, 9, 10, and 11 and, in addition to finding that deleting ZF 8 led to weakening of topologically associated domains and changes in DNA methylation and gene expression, found a weak motif for ZF 1 that has some similarity but is not identical to our discovered downstream motif [52]. While this study did investigate peaks that are substantially weaker in the wild-type than in the mutant, the study had only one replicate for each mutant and for the wild-type, limiting its ability to reliably detect differential binding. In contrast, the dataset that we used had multiple replicates for the wild-type and for each mutant, allowing us to identify differential binding events that are unlikely to be explained by differences in when an experiment was done [32]. In addition, this previous study identified motifs associated with ZFs by aggregating the sequences in the differential regions, while we identified specific nucleotides that are predictive of CTCF binding being significantly stronger in the wild-type. This previous study not only failed to confidently identify our motif for ZF 1 but found that the motif ZF 8 is almost identical to the motifs for ZFs 9–11, while our motif for ZF 8 placed less importance than our motifs for ZFs 9–11 on the most upstream part of the upstream motif, matching the known biology that different ZFs interact with different parts of their TFs' motifs [9]. An additional study used protein-DNA titration to show that ZFs 1–4 could bind a different sequence from our discovered downstream motif. However, the experiment was done in vitro, and the DNA sequence did not include the core motif, so the results could either represent a motif for these ZFs that was too rare to be learned by our models or a motif that is not bound by CTCF in vivo [53].

Our results suggest that ZF 1 may help strengthen CTCF's interaction with DNA because CTCF peaks with the core and discovered downstream motifs tend to be stronger than those with only the core motif. The fact that multiple positions within the core motif in the motif associated with significantly stronger binding in the wild-type relative to ZF 1 are partially degenerate also suggests that ZF 1 may help stabilize CTCF binding in the presence of a 1 bp-mismatch to the core motif. This may explain why the previous study of this data saw little difference in CTCF ChIP-Exo signal between wild-type and the mutants of ZFs 1–2, as this previous study limited its analysis to reads overlapping the core motif [11], which may not have included 1 bp-mismatches. However, additional data is needed to fully understand the purpose of ZF 1. For example, many studies have shown that CTCF interacts with cohesin to establish DNA loops [17, 19, 54], and recent studies showed that deleting ZF 1 changes many DNA loops and suggested that these changes may occur due to interactions between ZF 1 and RNA [26, 27]. Thoroughly investigating ZF 1's role in CTCF-cohesin interactions would require mutating ZF 1 and then assaying cohesin binding. An exciting extension to this work would be to apply our method to investigate effects of mutations of different combinations of parts of CTCF, where some combinations include ZF 1, on CTCF binding and paired RAD21 binding from recent studies that used this data to illustrate roles of ZFs 9–11 and other parts of CTCF in CTCF-cohesin interactions [18, 27].

Our modeling approach enabled us to discover both known and putative novel motifs as well as the spacings between them because a neural network interpreted with DeepLIFT [33] followed by TF-MoDISco [34] does not require an explicit featurization of the sequence or assumptions about the sizes of the motifs and because the tasks for our neural networks directly contrasted the wild-type and mutant datasets. Some previous studies have used k-mer support vector machines (SVMs)

to predict TF binding [55, 56]. However, linear SVMs cannot identify relationships between nucleotides that span more than k bases, and k needs to be small (usually at most eleven) [56] so that the number of parameters does not become too large to be learned with the available data, making these models incapable of learning our longer motifs (Fig. 2b). Many additional studies have trained neural networks to predict TF binding and used interpretation methods similar to those that we used to discover known and sometimes novel motifs of TFs [57–62]. Yet the subset of these studies that predicted CTCF binding failed to identify our motif for ZFs 1–2 and many also failed to identify the known upstream motif likely because, unlike our study, their models were not designed to directly learn individual DBD binding preferences. In fact, a previous study suggested that, for TFs with multiple ZFs, some ZFs have consistent binding patterns across the majority of binding sites, while others bind at only a minority of sites and do not always have the same spacing when binding, making their motifs difficult to detect when modeling all TF binding sites together [63]. Properly evaluating differences between wild-type and mutant TF binding requires multiple high-quality replicates of in vivo binding data from each of a wild-type and mutant TF, which, unfortunately, are not always available.

Our modeling approach has several drawbacks beyond requiring in vivo binding data from a mutated TF. One limitation is that our negative set consisted of a combination of the peaks that were comparably strong in the wild-type and the mutant and peaks that were stronger in the mutant, preventing us from identifying motifs that are associated with destabilizing interactions between ZFs in CTCF and DNA. We think this is why we did not identify the downstream motif identified by the previous study of this dataset [11], a motif that is thought to destabilize the binding of ZFs 2–3 to DNA. One possible way to extend our neural network to handle this case would be to train and interpret a regression model for predicting the fold-change of the wild-type versus the mutant peak strengths.

Neural Networks also have inherent limitations, regardless of how their tasks are defined. For example, neural networks require a large number of training examples [64], so they may not always be usable for mutants that do not affect at least a few thousand peaks, and they may not be able to learn motifs that are not present in at least a few thousand peaks in the training set. This may explain why our neural network was unable to identify the slight decrease in the frequency of C relative to the T in the eighth position of the core motif identified by a previous study of the peaks that were lost in a ZF 1 mutant [26]. In addition, machine learning models may learn the minimal set of features that are necessary for achieving good predictive performance; as a result, if there are multiple highly correlated features that are associated with the model's task, the model may learn only a strict subset of them, so the motifs learned by the model may exclude some biologically relevant motifs. In addition, convolutional neural networks require a fixed-size input [65], which is why we used merged peak summits +/− 500 bp. Using a model that can handle inputs of variable sizes would enable us to incorporate additional information that has been proposed to affect TF binding, such as sequences of distal regions that loop to TF peaks. Recent advances have enabled sequences of variable sizes to be used as inputs to deep convolutional neural networks [66, 67], so such modeling may be achievable. Finally, the failure of the logistic regression with the TF-MoDISco motif hit score to reach the performance of our neural network and the lack of additional meaningful TF-MoDISco motifs for the neural networks for the mutants of ZFs 1–8 suggest that our methods for interpreting what our neural networks learned are suboptimal. Thus, improving neural network interpretation methods should enable us to use neural networks to discover additional novel biology. Our ability to discover known motifs and novel putative motifs for CTCF's ZFs despite our approach's limitations demonstrates that our approach provides a foundation for identifying motifs of TF DBDs.

Conclusions

To our knowledge, we are the first to train machine learning models to predict whether a wild-type TF will have stronger binding than a mutated TF and the first to use differential binding between a wild-type and mutated TF to decipher binding preferences of the TF's DBDs. Our approach can aid future comparisons of wild-type TF binding to binding of TFs whose DBDs have been mutated, including TFs whose motifs are not well-characterized. In addition, our approach could be extended to comparisons of other in vivo TF binding experiments, such as differential TF binding across conditions, cell types, or time points.

Methods

CTCF ChIP-seq data processing

Since the previous study mapped the ChIP-seq reads to the mm9 genome assembly, we reprocessed the ChIP-seq data from wild-type CTCF and each of the CTCF ZF mutants so that we could map it to mm10 and ensure that it met ENCODE quality control standards after applying recently recommended methods for filtering reads and identifying reproducible peaks [68, 69]. To do this, we downloaded the data from GSE33819 [11, 70, 71]. We then mapped reads to mm10 [36] and filtered reads using the AQUAS Transcription Factor ChIP-seq processing

pipeline [72], which was also used for processing the TF ChIP-seq data for ENCODE2 and ENCODE3 [73], with default parameters.

To ensure that our datasets were sufficiently high-quality for reliable downstream analysis, we used the AQUAS pipeline [72] with default parameters to perform strict quality control evaluations. We first evaluated whether a dataset had more signal than we would expect from reads randomly dispersed in the genome, which we did by computing the normalized strand coefficient (NSC), which should ideally be at least 1.05, and the relative strand correlation (RSC), which should ideally be at least 0.8 [68]. We found that all the datasets had NSC > 1.05 and RSC > 0.8. Since all of the biological replicates for each mutant met ENCODE standards [68], we did not remove any datasets for our analyses.

We ran the AQUAS pipeline [72] separately on each replicate from each experiment to obtain irreproducible discovery rate (IDR) reproducible peaks [74] self-pseudo-replicates for each replicate. We found that the numbers of these peaks varied substantially across replicates due to different read depths per replicate. For example, replicate 2 for the ZF 4 mutant had approximately 6.2 million reads, leading to 189 reproducible peaks across self-pseudo-replicates, while replicate 3 for the ZF 4 mutant had approximately 25 million reads, leading to 24,864 reproducible peaks across self-pseudo-replicates (Supplemental Table 1).

We also pooled the reads from each dataset across the two or three biological replicates and ran the AQUAS pipeline on that [72]. For the wild-type dataset, we used the tagged data from *Mus musculus* so that the species and experimental protocol would be consistent with those of the mutants; a previous study showed that the peaks from the tagged data are consistent with those from a CTCF antibody ChIP-seq experiment done in the same lab [11]. The AQUAS pipeline [72] randomly divided the reads from each dataset into two "pooled pseudo-replicates," which are groups containing half of the reads, and identified IDR reproducible peaks [74] across pooled pseudo-replicates. We obtained tens of thousands of IDR reproducible peaks for the wild-type and for each mutant (Supplemental Table 1).

Identifying differential peaks

To identify differential peaks, which we defined as peaks that are significantly stronger in the wild-type than they are in the mutant, we merged peaks from the different datasets, computed the number of reads from each dataset in each merged peak, and evaluated whether the read depth was significantly larger in the wild-type than in each mutant. We merged all IDR self-pseudo-replicate reproducible peaks from each replicate, mutant combination, including the R339W mutant for ZF 3, and the tagged *Mus musculus* wild-type by merging peaks whose summits were within 50 bp of each other and defining the merged peak summit to be the average of the summits of the combined peaks. We used these peaks because our goal was to identify peaks that had a significantly larger difference in signal between wild-type and mutant than between replicates. Next, we used pybedtools version 7.10.0 [75, 76] and to remove reads from each replicate of each experiment mapping to mitochondrial DNA, unknown chromosome, or random chromosome parts; shift reads to the right by half of their fragment lengths from cross-correlation analysis; and count the reads overlapping the five-prime end of each merged peak. Finally, we ran DESeq2 [32], a method for identifying differential signals from read count data that accounts for differences in read depth between samples, on the read counts to compare peaks in the wild-type to those in each mutant. We defined a peak to be a member of the positive set, meaning significantly stronger in the wild-type, if the q-value was less than 0.05 and the log base 2 fold-change was less than -1 and a member of the negative set if the log base 2 fold-change was greater than or equal to 0.

Training neural networks for differential peak prediction

For each mutant except for R339W, which was not thought to have a substantial effect on binding [11], we trained a separate neural network to predict whether a merged peak was a member of the positive or negative set. Merged peaks that were members of neither set were not used. For each merged peak, we created two examples: the sequence underlying the merged peak summit $+/-500$ bp and the sequence underlying the reverse complement of the merged peak summit $+/-500$ bp. Our training set was chromosomes 3–7, 10–19, and X; our validation set was chromosomes 8–9; and our test set was chromosomes 1–2. We one-hot-encoded the sequences as four-by-one thousand matrices, where each row contained a binary vector indicating whether each position in the sequence consisted of a specific nucleotide; this encoding method has been used in previous studies that applied neural networks to predict TF binding [77–79]. We encoded Ns as all zeros. Thus, our input data did not contain any prior information about what parts of the DNA sequence are involved in CTCF binding.

The architecture that we used for each neural network was three convolutional layers [30], which were each followed by a rectified linear unit, followed by a max-pooling layer. The convolutional filters in the first layer should identify motifs that reveal whether a peak is significantly stronger in the wild-type, the filters in the following

layers should identify combinations of those motifs, and the max-pooling layer encodes the assumption that a single motif combination should not occur multiple times within a short region. The first convolutional layer had 60 4×15 filters with stride 1×1, the second convolutional layer had 60 1×15 filters with stride 1×1, and the third convolutional layer had 15 1×15 filters with stride 1×1. Each layer had dropout rate 0.2. The max-pooling layer was size 1×35 with stride 1×35. The max-pooling layer was followed by a fully connected layer with a sigmoid output. We trained the neural networks using Keras version 0.3.2 [80] with the Theano version 0.8.2 backend [81] using stochastic gradient descent with Nesterov momentum 0.85 [82] and learning rate 0.01, batch size 200, and class weights set to the fraction of peaks in the other class. We selected these hyperparameters after evaluating performance of multiple sets of hyperparameters on the validation set. The early stopping criterion was three consecutive epochs with no improvement in recall at 80% precision on the validation set. We initialized weights to be those from a pre-trained neural network with the same hyper-parameters and the negative set randomly down-sampled to be the size of the positive set. We initialized the weights for the pre-training using Keras's He normal initializer [80, 83].

Identifying important features learned by neural networks for differential peak prediction

Motifs that are important for making correct positive predictions are likely to be indicative of the binding preference of the mutant ZF because they are important for determining whether a peak will be significantly stronger in the wild-type data than in the data from the TF in which that ZF was mutated. To identify these motifs, we computed the importance of every nucleotide in each true positive example in the validation set and then used these importance values to construct motifs. We scored the importance of every nucleotide in every true positive example in the validation set using DeepLIFT, which computes the contribution of each nucleotide to a sequence's prediction relative to a reference [33]. We used DeepLIFT version 0.5.5-theano with the Rescale rule, where scores were taken from the sequence layer with the target of the final convolutional layer and our reference was a sequence of Ns. We used an extension to DeepLIFT with the Rescale rule to compute the "hypothetical scores," which can be thought of the extent to which the classifier is expecting a nucleotide, for each nucleotide at each position in each sequence [34].

We input the DeepLIFT scores and hypothetical scores into the TF-MoDISco method for constructing "TF-MoDISco motifs" learned by the model [34]. TF-MoDISco first extracts sequence patterns that frequently have high DeepLIFT scores in ChIP-seq peak sequences (called "seqlets"), next computes the pairwise similarities between seqlets, and then uses the similarities to cluster the seqlets into "TF-MoDISco motifs." We ran TF-MoDISco with these settings: seqlet FDR threshold = 0.2; gapped k-mer settings for similarity computation k-mer length = 8, number of gaps = 1, and number of mismatches = 0; final motif width = 50; and minimum number of seqlets = 200. We used the aggregated hypothetical scores of the seqlets supporting each TF-MoDISco motif to construct motif images.

To make position frequency matrices from TF-MoDISco motifs, we averaged the one-hot-encoded sequences at all the seqlet coordinates associated with the motifs. We also extracted the upstream, core, and discovered downstream motifs from our TF-MoDISco motifs (Supplemental File 1). To extract the upstream motif, we removed degenerate positions from the ends of the TF-MoDISco motif for ZF 11. We did this by first identifying the upstream-most position in which at least one nucleotide had probability > 0.60 and removing all earlier positions. We then scanned the motif until reaching another position at which no nucleotides had probability > 0.60. Because the following position was non-degenerate, we continued searching for an additional position in which no nucleotides had probability > 0.60. We removed that and all further downstream positions in the TF-MoDISco motif. To extract the core motif, we repeated the same process that we used for the upstream motif on the TF-MoDISco motif from ZF 6, except that we used a probability cutoff of 0.40 and required two consecutive bases with nucleotides passing the probability cutoff to begin extracting the motif. To extract the discovered downstream motif, we repeated the process that we used for the upstream motif on the TF-MoDISco motif from ZF 1, except that we used a probability cutoff of 0.35 and started at the downstream end of the TF-MoDISco motif, scanning upstream towards the start; we stopped when the difference in nucleotide probability for the nucleotide with the greatest probability decreased by > 0.35 between two consecutive positions. We used these upstream, core, and downstream motifs for further analyses. Finally, we constructed six motifs, which we call "mega-motifs":

1. The core motif (Supplemental File 1)
2. The upstream motif (Supplemental File 1)
3. The discovered downstream motif (Supplemental File 1)
4. The upstream motif followed by the core motif, where the motifs were separated by seven bases with nucleotide probabilities corresponding to the G/C-

content in mouse (The upstream and core motifs we identified were separated by seven bases because the nucleotide probabilities of two most upstream bases of the known core motif were not large enough to be captured in our core motif.)

5. The core motif followed by the discovered downstream motif, where the motifs were separated by two bases with nucleotide probabilities corresponding to the G/C-content in mouse (The core and discovered downstream motifs we identified were separated by two bases.)
6. The upstream motif followed by the core motif followed by the discovered downstream motif, where the upstream and core motifs were separated by seven bases with nucleotide probabilities corresponding to the G/C-content in mouse and the core and discovered downstream motifs were separated by two bases with nucleotide probabilities corresponding to the G/C-content in mouse.

Logistic regression with motif hit scores

We compared the performance of our neural network to that of a logistic regression with the scores of motif hits of the three motifs from [11]. We received the three motifs described in [11] in MEME format [84] from the authors of [11]. We scanned the merged CTCF peaks for these motifs using FIMO version 4.12.0 [45] with default parameters, where the background was the background provided to us by the authors of [11]. We computed the smallest motif q-value in each peak for each motif and used the negative log base ten of that q-value as a feature in a logistic regression; if there were no motif hits with q-value < 0.5 for a motif in a peak, then we set the value of that feature to zero for that peak. We trained the logistic regression using Scikit-learn version 0.19.1 [85] with l2 penalty 1.0. We used the same positives and negatives that we used for our neural network. We trained the logistic regression on a combination of the training and validation sets that we used for our neural network and evaluated the logistic regression using the same test set that we used for our neural network. Note that the original motifs and spacings between them were found using all of the peaks from the wild-type, including those on the chromosomes that we held out for testing; thus, we may be underestimating the difference in performance between our neural networks and the logistic regressions with the original motif hit scores.

We also compared the performance of our neural network and of the logistic regression with the original motif hit scores to the performance of a logistic regression where the only feature was the top TF-MoDISco motif (TF-MoDISco motif with the most supporting seqlets) score and to a logistic regression in which the only feature was the score of the original upstream motif followed by five base pairs followed by the original core motif. For the latter, the nucleotide frequencies in the five base pairs between the original upstream and original core motifs were set to be the background single nucleotide frequencies provided by the authors of [11]. For both evaluations, we computed features and trained logistic regressions using the same procedures that we used for the logistic regressions with the original motif hit scores.

Area under precision-recall curve computation

We compared the performances of the logistic regressions with motif hit scores to those of our neural networks by computing the area under the precision-recall curve for each model. We computed this using PRROC [86]. We used this metric instead of AUROC because our negative set is always larger than our positive set (Supplemental Table 2).

Identifying motif combinations in reads from CTCF HT-SELEX data

We compared the core motif followed by the discovered downstream motif to reads from CTCF HT-SELEX data from [38]. We first downloaded the reads from cycle 0 (control), which were taken before the TF was introduced, and cycle 4, the final cycle, that were generated for CTCF HT-SELEX in [38]. Since the HT-SELEX reads were only 20 bp long, we constructed a partial combination of the core motif followed by the discovered downstream motif, which was the downstream 10 bp of the core motif followed by 2 bp with the G/C-content in mouse (the core and discovered downstream motif were separated by 2 bp) followed by the upstream 4 bp of the downstream motif. We then scored the motif match to each HT-SELEX. Specifically, we converted the read and its reverse complement into a one-hot-encoded sequence, computed the dot product of those sequences and the partial combination of the core motif followed by the discovered downstream motif at every possible alignment of the two matrices, and computed the maximum of the dot products. We compared the distribution of scores for reads from cycle 0 to the distribution of scores for reads from cycle 4 using a Wilcoxon rank-sum test; the histograms of these distributions are illustrated in Fig. 3b. As a control, we repeated this process with only the downstream 16 bp of the core motif, and the histograms for this comparison are in Supplemental Figure 4.

We created aggregate motifs by running FIMO [45] with the partial combination of the core motif and the discovered downstream motif on reads from CTCF HT-SELEX cycle 4 [38], one-hot-encoding the positions with motif hits, averaging the one-hot-encoded matrices, and

visualizing these averages as motif logos (Supplemental Figure 5). We defined a "motif hit" to be motif hits with FIMO q-value less than four different cutoffs – 0.05, 0.01, 0.005, and 0.001 – and created an aggregate motif for the motif hits from each of these cutoffs. We visualized the motif logos using meme2images from the MEME suite [84].

Comparison of CTCF peaks overlapping CTCF-s peaks to those that do not overlap CTCF-s peaks

To compare the CTCF peaks that overlap CTCF-s peaks to those that do not, we re-processed that biotin-tagged data from [28], identified motif hits of the core motif followed by the discovered downstream motif in the CTCF ChIP-seq peaks, and compared the *p*-values of the motif hits in different subsets of the peaks. We re-processed the data and evaluated data quality using the AQUAS pipeline [72] with the hg38 genome assembly [87] and default parameters; both the CTCF and CTCF-s data had NSC > 1.05 and RSC > 0.8. Unless otherwise indicated, we used IDR reproducible peaks across self-pseudo-replicates (Each dataset had only 1 biological replicate.) for our analyses, which gave us 15,412 IDR reproducible CTCF-s peaks and 50,967 corresponding IDR reproducible CTCF peaks. We next identified motif hits of the core motif followed by the discovered downstream motif in the CTCF ChIP-seq peaks. Specifically, we first used bedtools [76] to obtain the fasta file for the peaks, next ran the MEME suite's fasta-get-markov [84] with -m 1 on the fasta file to obtain a background file, and then ran FIMO [45] on the fasta file with the background file and the core followed by downstream mega-motif (Supplemental File 1) with settings --max-stored-scores 50,000,000 and --thresh 1. We used bedtools intersect with settings -wa and -u to obtain CTCF ChIP-seq peaks that overlap CTCF-s ChIP-seq peaks. We used bedtools subtract with setting -A to obtain CTCF ChIP-seq peaks that do not overlap any "relaxed" (includes non-reproducible across self-pseudo-replicates) [72] CTCF-s ChIP-seq peaks (299,804 "relaxed" CTCF-s peaks). We then obtained the *p*-value of the best core followed by downstream mega-motif hit in each of these subsets of CTCF ChIP-seq peaks, setting the *p*-value to 1 when no motif hit was identified. We compared the distributions of the best motif hit *p*-values for these two subsets of CTCF ChIP-seq peaks using a Wilcoxon rank-sum test.

To investigate whether our results could be explained by a relationship between core followed by discovered downstream motif occurrences and reproducibility of CTCF ChIP-seq peaks across experiments, we also compared CTCF ChIP-seq peaks from [28] to those from ENCODE [39]. Since the data in [28] came from HeLa cells, we downloaded the "optimal" IDR reproducible peaks (ENCODE entry ENCFF772LNY, 44,072 IDR reproducible CTCF peaks) and "relaxed" peaks from pooled reads across replicates (ENCODE entry ENCFF331BAX, 300,3000 "relaxed" CTCF peaks) from the deepest ENCODE HeLa cell CTCF ChIP-seq dataset [39]. We used bedtools intersect with settings -wa and -u to obtain CTCF ChIP-seq peaks from [28] that overlap ENCODE IDR reproducible CTCF ChIP-seq peaks. We used bedtools subtract with setting -A to obtain CTCF ChIP-seq peaks from [28] that do not overlap ENCODE pooled replicate CTCF ChIP-seq peaks. We then obtained the p-value of the best core followed by downstream mega-motif hit in each of these subsets of CTCF ChIP-seq peaks from [28], setting the p-value to 1 when no motif hit was identified. We compared the distributions of the *p*-values for these two subsets of CTCF ChIP-seq peaks from [28] using a Wilcoxon rank-sum test.

Computational predictions of CTCF motifs using models trained on in vitro ZF binding data

To further evaluate whether our discovered downstream motif is likely to interact with CTCF ZF 1, we compared it to predicted CTCF motifs from models trained using in vitro ZF binding data. These models were trained on in vitro data measuring the binding specificities of individual ZFs; they take ZF amino acid sequences as input and output a predicted motif. First, we used RCADE2's RC.sh to predict the motif for CTCF [42]. To explore alternative methods, we also put the sequences of CTCF's ZFs into the "Predict PWMs" function of the "Interactive PWM Predictor" [40, 41] and predicted the motif using each of the three available models: "RF Regression on B1H," "Expanded Linear SVM," and "Polynomial SVM." We additionally ran each model on ZF 1 alone to confirm that the models predicted that ZF 1 interacts with GAG. Figure 3a contains the outputs from RCADE2, and Supplemental Figure 6 contains the outputs from the other models.

Comparison of CTCF ChIP-seq peak strengths with different combinations of motifs

We compared peak strengths for different motif combinations by identifying occurrences of each mega-motif in CTCF peaks, grouping peaks based on mega-motif presences, and quantifying properties of each peak in each group. We scanned the wild-type mouse CTCF peaks for the mega-motifs using FIMO [45] with default parameters except for the threshold, which we set to 1, and the background, which we set to the output from fasta-get-markov [84] run on the sequences of the CTCF peaks with setting -m 1. We used version 4.12.0 of the MEME suite [84] for all of these analyses.

We evaluated the relationship between peak strength and the presence of the downstream motif by comparing peaks with the core followed by downstream mega-motif to peaks with only the core motif. We defined motif hits as motif occurrences with FIMO p-value < 0.0001 (default from FIMO) [45]. When using the stricter motif cutoff, we defined motif hits as motif occurrences with FIMO q-value < 0.05. We then used bedtools version 2.26.0 [76] to identify peaks with the core motif, the core motif and no discovered downstream motif, and the core followed by downstream mega-motif for the different motif hit cutoffs. We defined the peak strength to be the natural log of the signal from SPP (column seven from the narrowPeak files). We compared the peak strength for peaks with the core motif and no discovered downstream motif versus peaks with the core followed by downstream mega-motif by doing a two-sided Wilcoxon rank-sum test, and we did a Bonferroni correction of the p-values by multiplying them by six (two comparisons for each of three cell types/tissues). We repeated this process for liver and heart data, which was taken from the mouse ENCODE 8-week-old mouse Ren Lab datasets [46].

Since the definition of a motif hit can be sensitive to thresholding, we also compared the peak strength of CTCF peaks to the -log base ten q-values from FIMO [45] of all of the motif occurrences from FIMO regardless of their FIMO p-value or q-value. We incorporated all motif occurrences by identifying the correlation between peak strength and the -log base ten q-values from the FIMO for the core motif, the core followed by downstream mega-motif, and the upstream followed by core mega-motif. We then compared the correlations for the core motif and each of the other two mega-motifs using a one-sided Fisher's r-to-z transformation and did a Bonferroni correction of all p-values by multiplying them by six (two pairs for each of three tissues).

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-022-08486-9.

Additional file 1.

Acknowledgements
We would like to thank A. Kundaje, A. Shrikumar, H. B. Fraser, the other members of the Fraser and Kundaje Labs, R. Casellas, and Z. Zuo for useful discussions and suggestions. We would also like to thank J. Tycko and the anonymous reviewers for feedback on the manuscript. We would additionally like to thank the members of the Casellas Lab for providing us with the original motifs from [11] and the background used for tools in the MEME suite from that study.

Authors' contributions
Study conceptualization was done by I.M.K. and C.S.F. Data curation was done by I.M.K. Resource management and methodology development were done by I.M.K. with assistance from C.S.F. Formal analysis, investigation, software implementation, and visualization were done by I.M.K. with assistance from A.B. Original draft preparation was done by I.M.K. Reviewing and editing was done by all authors. The authors read and approved the final manuscript.

Authors' information
Not applicable.

Funding
I.M.K. was funded by the Stanford Center for Computational, Evolutionary and Human Genomics Predoctoral Fellowship and the Carnegie Mellon University Computational Biology Department Lane Fellowship.

Availability of data and materials
All data used in this study was previously published and released in other studies. Mouse activated B Cell CTCF ChIP-seq data analyzed in this study was downloaded from GEO GSE33819 [11]. CTCF HT-SELEX data was downloaded from ENA PRJEB3289 [38]. Mouse liver and heart CTCF ChIP-seq data were downloaded from the ENCODE portal entries ENCFF542WEE and ENCFF-616HYA, respectively [46]. Corresponding CTCF and CTCF-s ChIP-seq data were downloaded from GSE108869. ENCODE HeLa cell CTCF ChIP-seq data was downloaded from ENCODE portal entry ENCSR000AOA [39]. The zinc finger image in Supplemental Figure 1 was taken from [88]. The core motif logo in Supplemental Figure 1 is the Hocomoco human CTCF motif downloaded from CIS-BP [20], and the upstream motif in Supplemental Figure 1 is from [11]. Code can be found at https://github.com/kundajelab/CTCFMutants. Core, upstream, and discovered downstream motifs are in Supplemental File 1. Deep neural network models, deepLIFT scores, TF-MoDISco motifs, and FIMO hits for motifs can be found on the Supporting Website.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Departments of Computer Science, Stanford University, 240 Pasteur Drive, Stanford, California 94305, USA. [2]Present address: Department of Computational Biology, Carnegie Mellon University, 5000 Forbes Avenue, Gates-Hillman Building Room 7703, Pittsburgh, PA 15213, USA. [3]Departments of Physics, Stanford University, 240 Pasteur Drive, Stanford, California 94305, USA. [4]Present address: Machine Intellection Department, Institute for Infocomm Research, 1 Fusionopolis Way, #21-01 Connexis South Tower, Singapore 138632, Singapore.

Received: 23 September 2021 Accepted: 21 March 2022
Published online: 12 April 2022

References

1. Ilsley MD, Huang S, Magor GW, Landsberg MJ, Gillinder KR, Perkins AC. Corrupted DNA-binding specificity and ectopic transcription underpin dominant neomorphic mutations in KLF/SP transcription factors. BMC Genomics. 2019;20:417.
2. Han BY, Wu S, Foo C-S, Horton RM, Jenne CN, Watson SR, et al. Zinc finger protein Zfp335 is required for the formation of the naïve T cell compartment. Elife. 2014;3:1–28.
3. Stevenson WS, Morel-Kopp MC, Chen Q, Liang HP, Bromhead CJ, Wright S, et al. GFI1B mutation causes a bleeding disorder with abnormal platelet function. J Thromb Haemost. 2013;11:2039–47.
4. Maezawa S, Alavattam KG, Tatara M, Nagai R, Barski A, Namekawa SH. A rapidly evolved domain, the SCML2 DNA-binding repeats, contributes to chromatin binding of mouse SCML2. Biol Reprod. 2018;100:409–19.
5. Schwartz JJ, Roach DJ, Thomas JH, Shendure J. Primate evolution of the recombination regulator PRDM9. Nat Commun. 2014;5:4370.

6. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human transcription factors: function, expression and evolution. Nat Rev Genet. 2009;10:252–63.
7. Fedotova AA, Bonchuk AN, Mogila VA, Georgiev PG. C2H2 zinc finger proteins: the largest but poorly explored family of higher eukaryotic transcription factors. Acta Nat. 2017;9:47–58.
8. Persikov AV, Wetzel JL, Rowland EF, Oakes BL, Xu DJ, Singh M, et al. A systematic survey of the Cys2His2 zinc finger DNA-binding landscape. Nucleic Acids Res. 2015;43:1965–84.
9. Wolfe SA, Nekludova L, Pabo CO. DNA recognition by Cys2His2 zinc finger proteins. Annu Rev Biophys Biomol Struct. 2000;29:183–212.
10. Garton M, Najafabadi HS, Schmitges FW, Radovani E, Hughes TR, Kim PM. A structural approach reveals how neighbouring C2H2 zinc fingers influence DNA binding specificity. Nucleic Acids Res. 2015;43:9147–57.
11. Nakahashi H, Kwon KRK, Resch W, Vian L, Dose M, Stavreva D, et al. A genome-wide map of CTCF Multivalency redefines the CTCF code. Cell Rep. 2013;3:1678–89.
12. Lyon MF, Jamieson RV, Perveen R, Glenister PH, Griffiths R, Boyd Y, et al. A dominant mutation within the DNA-binding domain of the bZIP transcription factor Maf causes murine cataract and results in selective alteration in DNA binding. Hum Mol Genet. 2003;12:585–94.
13. Ong C, Corces VG. CTCF : an architectural protein bridging genome topology and function. Nat Publ Gr. 2014;15:234–46.
14. Phillips JE, Corces VG. CTCF: Master weaver of the genome. Cell. 2009;137:1194–211.
15. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at Kilobase resolution reveals principles of chromatin looping. Cell. 2014;159:1665–80.
16. Tang Z, Luo OJ, Li X, Zheng M, Zhu JJ, Szalaj P, et al. CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. Cell. 2015;163:1611–27.
17. Hansen AS. CTCF as a boundary factor for cohesin-mediated loop extrusion: evidence for a multi-step mechanism. Nucleus. 2020;11:132–48.
18. Pugacheva EM, Kubo N, Loukinov D, Tajmul M, Kang S, Kovalchuk AL, et al. CTCF mediates chromatin looping via N-terminal domain-dependent cohesin retention. Proc Natl Acad Sci U S A. 2020;117:2020–31.
19. Hansen AS, Hsieh THS, Cattoglio C, Pustova I, Saldaña-Meyer R, Reinberg D, et al. Distinct classes of chromatin loops revealed by deletion of an RNA-binding region in CTCF. Mol Cell. 2019;76:395–411.
20. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al. Determination and inference of eukaryotic transcription factor sequence specificity. Cell. 2014;158:1431–43.
21. Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonalves A, Kutter C, et al. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell. 2012;148:335–48.
22. Boyle AP, Song L, Lee BK, London D, Keefe D, Birney E, et al. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. Genome Res. 2011;21:456–64.
23. Rhee HS, Pugh BF. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell. 2011;147:1408–19.
24. Renda M, Baglivo I, Burgess-Beusse B, Esposito S, Fattorusso R, Felsenfeld G, et al. Critical DNA binding interactions of the insulator protein CTCF: a small number of zinc fingers mediate strong binding, and a single finger-DNA interaction controls binding at imprinted loci. J Biol Chem. 2007;282:33336–45.
25. Li W, Shang L, Huang K, Li J, Wang Z, Yao H. Identification of critical base pairs required for CTCF binding in motif M1 and M2. Protein Cell. 2017;8:544–9.
26. Saldaña-Meyer R, Rodriguez-Hernaez J, Escobar T, Nishana M, Jácome-López K, Nora EP, et al. RNA interactions are essential for CTCF-mediated genome organization. Mol Cell. 2019;76:412–422.e5.
27. Nora EP, Caccianini L, Fudenberg G, So K, Kameswaran V, Nagle A, et al. Molecular basis of CTCF binding polarity in genome folding. Nat Commun. 2020;11:1–13.
28. Li J, Huang K, Hu G, Babarinde IA, Li Y, Dong X, et al. An alternative CTCF isoform antagonizes canonical CTCF occupancy and changes chromatin architecture to promote apoptosis. Nat Commun. 2019;10:1535.
29. Bailey CG, Gupta S, Metierre C, Amarasekera PM, O'Young P, Kyaw W, et al. Somatic mutations in CTCF zinc fingers produce cellular phenotypes explained by structure-function relationships. In: bioRxiv; 2021. https://www.biorxiv.org/content/10.1101/2021.01.08.425848v1. Accessed 6 Dec 2021.
30. LeCun Y, Jackel LD, Boser B, Denker JS, Graf HP, Guyon I, et al. Handwritten digit recognition: applications of neural network chips and automatic learning. IEEE Commun Mag. 1989;27:41–6.
31. Ciresan D, Meier U, Masci J. Flexible, high performance convolutional neural networks for image classification. Int Jt Conf Artif Intell. 2011;2:1237–42.
32. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
33. Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. Int Confer Machine Learn. 2017;70:3145–53.
34. Shrikumar A, Tian K, Shcherbina A, Avsec Ž, Banerjee A, Sharmin M, et al. TF-MoDISco v0.4.2.2-alpha: Technical Note. In: arXiv; 2018. http://arxiv.org/abs/1811.00416. Accessed 16 May 2019.
35. Zhou X, Maricque B, Xie M, Li D, Sundaram V, Martin EA, et al. The human epigenome browser at Washington University. Nat Methods. 2011;8:989–90.
36. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, et al. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002;420:520–62.
37. Xiao T, Wongtrakoongate P, Trainor C, Felsenfeld G. CTCF recruits Centromeric protein CENP-E to the Pericentromeric/Centromeric regions of chromosomes through unusual CTCF-binding sites. Cell Rep. 2015;12:1704–14.
38. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, et al. DNA-binding specificities of human transcription factors. Cell. 2013;152:327–39.
39. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.
40. Persikov AV, Singh M. De novo prediction of DNA-binding specificities for Cys2His2 zinc finger proteins. Nucleic Acids Res. 2014;42:97–108.
41. Persikov AV, Osada R, Singh M. Predicting DNA recognition by Cys2His2 zinc finger proteins. Bioinformatics. 2009;25:22–9.
42. Dogan B, Kailasam S, Corchado AH, Nikpoor N, Najafabadi HS. A DNA recognition code for probing the in vivo functions of zinc finger transcription factors at domain resolution. In: bioRxiv; 2020. https://www.biorxiv.org/content/10.1101/630756v3. Accessed 23 Apr 2020.
43. Ciccone DN, Namiki Y, Chen C, Morshead KB, Wood AL, Johnston CM, et al. The murine IgH locus contains a distinct DNA sequence motif for the chromatin regulatory factor CTCF. J Biol Chem. 2019;294:13580–92.
44. Li Y, Huang W, Niu L, Umbach DM, Covo S, Li L. Characterization of constitutive CTCF/cohesin loci: a possible role in establishing topological domains in mammalian genomes. BMC Genomics. 2013;14:553.
45. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011;27:1017–8.
46. Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, Gilbert DM, et al. An encyclopedia of mouse DNA elements (mouse ENCODE). Genome Biol. 2012;13:418.
47. Hashimoto H, Wang D, Horton JR, Zhang X, Corces VG, Cheng X. Structural basis for the versatile and methylation-dependent binding of CTCF to DNA. Mol Cell. 2017;66:711–720.e3.
48. Yin M, Wang J, Wang M, Li X, Zhang M, Wu Q, et al. Molecular mechanism of directional CTCF recognition of a diverse range of genomic sites. Cell Res. 2017;27:1365–77.
49. Sehnal D, Bittrich S, Deshpande M, Svobodová R, Berka K, Bazgier V, et al. Mol* viewer: modern web app for 3D visualization and analysis of large biomolecular structures. Nucleic Acids Res. 2021;49:W431–7.
50. Font J, MacKay JP. Beyond DNA: zinc finger domains as RNA-binding modules. Methods Mol Biol. 2010;649:479–91.
51. Wang D, Horton JR, Zheng Y, Blumenthal RM, Zhang X, Cheng X. Role for first zinc finger of WT1 in DNA sequence specificity: Denys–Drash syndrome-associated WT1 mutant in ZF1 enhances affinity for a subset of WT1 binding sites. Nucleic Acids Res. 2018;46:3864–77.
52. Soochit W, Sleutels F, Stik G, Bartkun M, Basu S, Hernandez SC, et al. CTCF chromatin residence time controls three-dimensional genome organization, gene expression and DNA methylation in pluripotent cells. Nat Cell Biol. 2021;23:881–93.
53. Xu D, Ma R, Zhang J, Liu Z, Wu B, Peng J, et al. Dynamic nature of CTCF tandem 11 zinc fingers in multivalent recognition of DNA as revealed by NMR spectroscopy. J Phys Chem Lett. 2018;9:4020–8.

54. Li Y, Haarhuis JHI, Sedeño Cacciatore Á, Oldenkamp R, van Ruiten MS, Willems L, et al. The structural basis for cohesin–CTCF-anchored loops. Nat. 2020;578:472–6.
55. 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.
56. Ghandi M, Lee D, Mohammad-Noori M, Beer MA. Enhanced regulatory sequence prediction using gapped k-mer features. PLoS Comput Biol. 2014;10:e1003711.
57. Avsec Ž, Weilert M, Shrikumar A, Krueger S, Alexandri A, Dalal K, et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. Nat Genet. 2021;53:354–66.
58. Lanchantin J, Singh R, Wang B, Qi Y. Deep motif dashboard: visualizing and understanding genomic sequences using deep neural networks. Pac Symp Biocomput. 2017;22:254–65.
59. Greenside P, Shimko T, Fordyce P, Kundaje A. Discovering epistatic feature interactions from neural network models of regulatory DNA sequences. Bioinformatics. 2018;34:i629–37.
60. Li H, Guan Y. Fast decoding cell type–specific transcription factor binding landscape at single-nucleotide resolution. Genome Res. 2021;31:721–31.
61. Liu G, Zeng H, Gifford DK. Visualizing complex feature interactions and feature sharing in genomic deep neural networks. BMC Bioinformatics. 2019;20:401.
62. Zheng A, Lamkin M, Zhao H, Wu C, Su H, Gymrek M. Deep neural networks identify sequence context features predictive of transcription factor binding. Nat Mach Intell. 2021;3:172–80.
63. Zuo Z, Billings T, Walker M, Petkov P, Fordyce P, Stormo GD. Quantitative analysis of ZFY and CTCF reveals dependent recognition of tandem zinc finger proteins. In: bioRxiv; 2021. https://www.biorxiv.org/content/10.1101/637298v2. Accessed 25 Nov 2021.
64. Angermueller C, Pärnamaa T, Parts L, Oliver S. Deep learning for computational biology. Mol Syst Biol. 2016;12:1–16.
65. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Adv Neural Inf Process Syst. 2012;25:1097–105.
66. He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell. 2015;37:1904–16.
67. Graves A, Liwicki M, Fernández S, Bertolami R, Bunke H, Schmidhuber J. A novel connectionist system for unconstrained handwriting recognition. IEEE Trans Pattern Anal Mach Intell. 2009;31:855–68.
68. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 2012;22:1813–31.
69. Marinov GK, Kundaje A, Park PJ, Wold BJ. Large-scale quality analysis of published ChIP-seq data. G3 Genes|Genomes|Genetics. 2014;4:209–23.
70. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets - update. Nucleic Acids Res. 2013;41:D991–5.
71. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30:207–10.
72. Lee JW, Boley N, Kundaje A. AQUAS TF and histone ChIP-seq pipeline. In: GitHub; 2016. https://github.com/kundajelab/chipseq_pipeline. Accessed 9 Oct 2016.
73. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, et al. The encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. 2018;46:D794–801.
74. Li Q, Brown JB, Huang H, Bickel PJ. Measuring reproducibility of high-throughput experiments. Ann Appl Stat. 2011;5:1752–79.
75. Dale RK, Pedersen BS, Quinlan AR. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. Bioinformatics. 2011;27:3423–4.
76. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
77. 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.
78. 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.
79. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods. 2015;12:931–4.
80. Chollet F. Keras. In: GitHub; 2015. https://keras.io/. Accessed 9 Feb 2016.
81. Theano Development Team, Al-Rfou R, Alain G, Almahairi A, Angermueller C, Bahdanau D, et al. Theano: A Python framework for fast computation of mathematical expressions. In: arXiv; 2016. http://arxiv.org/abs/1605.02688. Accessed 19 May 2019.
82. Nesterov Y. A method of solving a convex programming problem with convergence rate O(1/k^2). Sov Math Dokl. 1983;27:372–6.
83. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision; 2016. p. 1026–34.
84. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME suite: tools for motif discovery and searching. Nucleic Acids Res. 2009;37:W202–8.
85. Pedregosa F, Varoquaux G. Scikit-learn: machine learning in Python. JMLR. 2011;12:2825–30.
86. Grau J, Grosse I, Keilwagen J. PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. Bioinformatics. 2015;31:2595–7.
87. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921.
88. Manske M. File:zinc finger.Png. In: Wikimedia Commons; 2004. https://creativecommons.org/licenses/by-sa/3.0/legalcode. Accessed 20 Nov 2019.

Publisher's Note