DARDN: Identifying transcription factor binding motifs from long DNA sequences using multi-CNNs and DeepLIFT

Hyun Jae Cho[1*], Zhenjia Wang[2†], Yidan Cong[2,3†], Stefan Bekiranov[3‡], Aidong Zhang[1*], Chongzhi Zang[2†]

1 Department of Computer Science, University of Virginia, Charlottesville, VA, USA

2 Center for Public Health, University of Virginia, Charlottesville, VA, USA

3 Department of Biochemistry and Molecular Genetics, Institution Name, University of Virginia, Charlottesville, VA, USA

* hc2kc@virginia.edu

Abstract

Motivation: Characterization of regulatory elements in DNA sequence is a key task in functional genomics. CTCF exhibits specific binding patterns in the genome of cancer cells and has a non-canonical function to facilitate oncogenic transcription program by cooperating with transcription factors bound at flanking distal regions. Identification of sequence motifs from a broad genomic region surrounding cancer-specific CTCF binding sites can help find active transcription factors in a cancer type. However, the long DNA sequences without localization information makes it difficult to perform conventional motif enrichment analysis.

Results: We present DNAResDualNet (DARDN), a computational method that utilizes convolutional neural networks (CNNs) coupled with feature discovery using DeepLIFT, for identifying DNA sequence features that can differentiate two sets of lengthy DNA sequences. Evaluation on DNA sequences associated with CTCF binding sites in T-cell acute lymphoblastic leukemia (T-ALL) and other cancer types demonstrates DARDN's ability in classifying DNA sequences surrounding cancer-specific CTCF binding from control constitutive CTCF binding and identifying sequence motifs for transcription factors potentially active in each specific cancer type. We identified motifs for potential oncogenic transcription factors in T-ALL, acute myeloid leukemia (AML), breast cancer (BRCA), colorectal cancer (CRC), lung adenocarcinoma (LUAD), and prostate cancer (PRAD). Our work demonstrates the power of advanced machine learning and feature discovery approach in finding biologically meaningful information from complex DNA sequence data.

Author summary

We present DNAResDualNet (DARDN), a deep learning model designed for identifying DNA sequence motifs in long DNA sequences, particularly for revealing oncogenic transcription factors in various cancers. Building on our previous work (Fang et al., Genome Biol. 2020), which uncovered the non-canonical role of cancer-specific CTCF binding in oncogenesis, DARDN uses a dual convolutional neural network with residual connections and the DeepLIFT method to discern complex patterns in DNA. We successfully applied DARDN to identify unique sequence motifs in T-cell acute lymphoblastic leukemia and other cancers like AML, BRCA, CRC, LUAD, and PRAD, showcasing its capability in cancer genomics research. This work underscores the potential of deep learning in functional genomics, especially in interpreting biological data and discovering new insights in gene regulation and cancer genomics.

Introduction

Identification of cis-regulatory elements in the non-coding genome is a key task in functional and regulatory genomics research. Active cis-regulatory elements usually function as transcription factor (TF) binding sites, containing specific DNA sequence recognized and bound by the TF(s) that regulate gene expression. Most TF binding sites are located in distal enhancer regions in the genome that can be far away from their regulatory target genes. This makes it difficult to identify regulatory sequence motifs from a long DNA sequence. Distal enhancer-binding TFs interact with co-factors to execute their regulatory functions. One such example is CCCTC-binding factor (CTCF), a zinc finger protein that binds to DNA and can induce DNA looping, functioning by anchoring at topologically associating domain (TAD) boundaries and blocking cross-domain interactions [1]. Disruption of individual CTCF binding in the genome causing aberrant chromatin interaction and differential gene expression has been observed in many cellular systems [2,3]. We previously showed that specific CTCF binding patterns frequently occur in many cancer types, and such aberrant CTCF binding events are induced by oncogenic TF binding at distal regions [4]. Therefore, the wide genomic regions flanking cancer-specific CTCF binding sites should contain sequence features for specific oncogenic TFs, and knowing the oncogenic factors is important for understanding the mechanisms of cancer development. We aim to find DNA sequence features enriched at genomic regions associated with cancer-specific CTCF binding sites but not at regions near constitutive CTCF binding sites that exist in most cell types.

Conventional TF motif search methods are not feasible for this problem because the relative genomic location of the target oncogenic TF binding site relative to the cancer-specific CTCF site is unknown and can be very far, and it varies across different cancer-specific CTCF sites. Conventional DNA sequence motif search methods are not feasible also because the search space is huge and without appropriate control sequences. In fact, direct DNA sequence motif search in the gained sites was unable to yield any motifs unambiguously enriched other than CTCF itself [4]. TF-binding ChIP-seq data-based methods like BART [6] are potentially feasible but are limited to pre-identified cis-regulatory element repertoire such as the union open chromatin regions. Therefore, new computational methods need to be developed for tackling this unique but important problem with significant biological meaning.

Advanced machine learning approaches such as deep convolutional neural networks (CNN) have been popular for applications in genomics and cancer research [8–11]. In addition to solving a classification problem using deep learning, we also focus on the interpretation of the CNN model to make biologically meaningful discoveries. Specifically, with a well-trained deep neural network classifier, one can use feature discovery tools such as DeepLIFT (Deep Learning Important FeaTures) [12] to identify features from the model that imply functionally important biological insights. Compared with traditional bioinformatics algorithms, deep learning models are specifically suitable for this task because of the complexity of the problem, i.e., ultra-long DNA sequences with a huge number of features and relatively rare cancer-specific CTCF binding events.

To address this problem, we introduce DNAResDualNet (DARDN), a computational method that utilizes convolutional neural networks (CNNs) coupled with feature discovery using DeepLIFT, to identify DNA sequence features enriched in a set of long DNA sequences compared with another set of DNA sequences as control. DARDN trains a pair of deep CNN models, alongside residual connections, to enhance classification accuracy for extended input DNA sequences. It is designed to rely exclusively on DNA sequences for training without integrating other data types, making it simple to train and become versatile to be applied to similar sequence data from other biological scenarios. We demonstrate the effectiveness of DARDN in finding the simulated sequence motif from synthetic sequence data and finding the sequence motif for known oncogenic TFs such as Notch1 for T-cell acute lymphoblastic leukemia (T-ALL) data. We then applied DARDN to identify sequence motifs for potential oncogenic transcription factors for acute myeloid leukemia (AML), breast cancer (BRCA), colorectal cancer (CRC), lung adenocarcinoma (LUAD), and prostate cancer (PRAD).

Materials and methods

0.1 Data and Its Representation

Genomic DNA sequences used in this study are from human hg38 genome version. Foreground cancer-specific CTCF sites and constitutive CTCF sites data are from our previous work [4], which identified cancer specific CTCF binding patterns by integrative analysis of over 771 high-quality CTCF ChIP-seq datasets across a variety of different human cell types including both normal and cancer [4]. For each of the six cancer types included in this work, tens to thousands of cancer-type-specific CTCF binding sites are identified in each cancer type, while 22,097 constitutive sites in the genome are conserved across cell types.

To alleviate the problem of data imbalance between the 72 T-ALL-specific CTCF sites and 22,097 constitutive CTCF sites, we performed data augmentation by reverse complementing and shifting the gained sites. Specifically, we shift the original sequences and their reverse complements to the left and to the right stochastically between 1 to 5 base pairs (bps) (Figure 1a).

Each DNA sequence containing a CTCF binding site is then represented as a one-hot encoding in order to be processed by the deep neural network model. The matrix consisting of one-hot encoded DNAs is passed to a deep neural network to train the model (Figure 1b). The dimension of the matrix is $4 \times L$, where L is the length of the DNA sequence. Hence, the model is flexible with DNA sequences with various length. The model produces a binary prediction of whether there is a cancer-specific CTCF binding site for each input sequence (0- or 1-labelled). We generated 10 kilo-base (kb) genomic DNA sequence centered at each T-ALL-specific CTCF site as positive signal with label 1 and those centered at constitutive CTCF sites with label 0 and trained our model to classify any 10kb DNA sequence as either 0 or 1.

0.2 Evaluation Metrics

We use the Matthew's correlation coefficient (MCC) to evaluate DARDN's classification accuracy for predicting CTCF gained versus constitutive sites. MCC measures the correlation between the true labels and predicted labels, ranging from -1 to +1. A value of +1 indicates perfect prediction, -1 indicates total disagreement between prediction and truth, and 0 is the expected value for random guessing. MCC is calculated by dividing the covariance of the true and predicted labels by the product of their standard deviations, which is represented as:

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

Using HOMER [7], we identified enriched motifs on CTCF gained sites in T-ALL, guided by prior findings of oncogenic motifs. We evaluated DeepLIFT's performance by examining the ranking of the RBPJ motif, associated with an oncogene in T-ALL. For other cancer types, we relied on literature to identify highly ranked oncogenic motifs. Our pipeline was tested for robustness using varying input sequence lengths, sampling gained sites, and sampling constitutive sites.

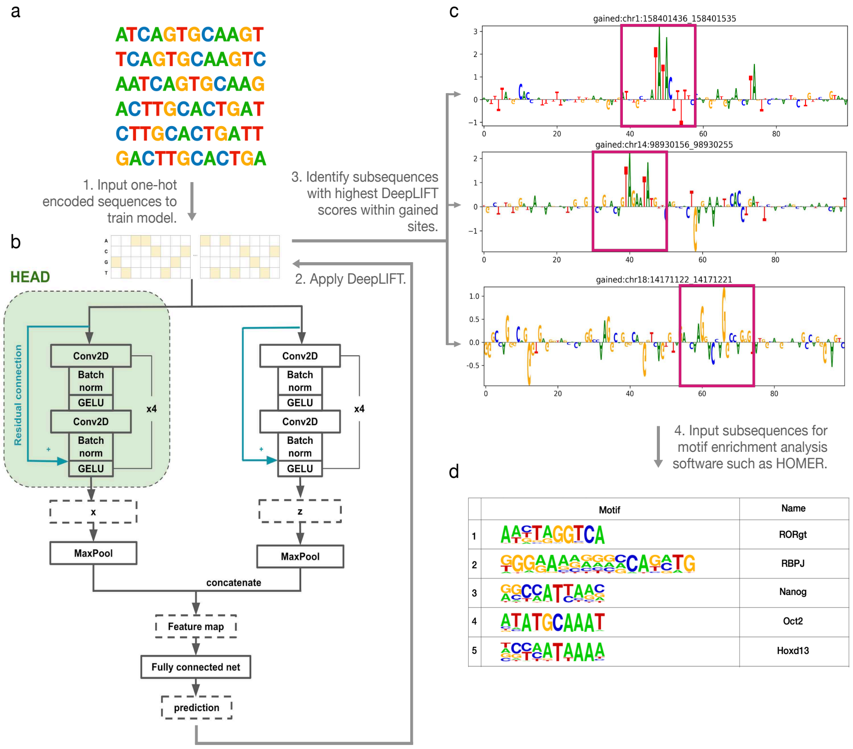
a
ATCAGTGCAAGT
TCAGTGCAAGTC
AATCAGTGCAAG
ACTTGCACTGAT
CTTGCACTGATT
GACTTGCACTGA
1. Input one-hot encoded sequences to train model.
b
HEAD
Residual connection
Conv2D
Batch norm
GELU
Conv2D
Batch norm
GELU
x4
x
z
MaxPool
concatenate
Feature map
Fully connected net
prediction
2. Apply DeepLIFT.
3. Identify subsequences with highest DeepLIFT scores within gained sites.
c
gained:chr1:158401436_158401535
gained:chr14:98930156_98930255
gained:chr18:14171122_14171221
4. Input subsequences for motif enrichment analysis software such as HOMER.
d
Motif
Name
1 RORgt
2 RBPJ
3 Nanog
4 Oct2
5 Hoxd13

Fig 1. Schematic of overall computational framework design. a. data augmentation: original sequence, its left/right shifts, the reverse complement, and its left/right shifts. b. DARDN Model: uses two deep convolutional networks to process a 2D one-hot encoded sequence for binary classification. c. DeepLIFT is applied for sequence feature selection. Subsequences with the highest moving average DeepLIFT scores are selected for motif analysis. d. HOMER motif analysis result.

0.3 Model

Thanks to its exceptional capability in hierarchical feature extraction and the characteristic of being location invariant, convolutional neural networks (CNNs) have been used as a promising approach for generating informative latent feature maps as well as for various tasks using DNA sequences [8–11, 13–17]. However, while plain CNN models are typically location-invariant and can be effective for certain types of DNA sequences, we found that they are ineffective for our purposes, as shown in Table 1.

To tackle the limitations of plain CNN models, we have developed DARDN (DNAResDualNet), a CNN-based model that is capable of learning of intricate relationships among distant DNA sequences even in the existence of deep convolutional layers. DARDN, as the name suggests, employs two CNNs with distinct initial kernel sizes for DNA sequence classification and residual connections in it to preserve complex

Model	True Positives	True Negatives	False Positives	False Negatives	MCC
CNN	47	5067	31	62	0.5
DARDN	76	5108	2	21	0.87

Table 1. Evaluation on hold-out data consisting of 78 T-ALL-specific CTCF gained sites and 5,129 constitutive CTCF sites of length 10,000 base pairs. The number of hold-out gained sites includes augmented sites. A plain CNN model is unable to make accurate classification due to class imbalance and sequence length while our model, DARDN, can achieve significantly superior performance. MCC stands for Matthew's Correlation Coefficient.

relationships between distant DNA sequences. Having two input kernels of different sizes leverages the variability in gene sequence lengths to enable the CNNs to learn important features at different levels of granularity. We use 4 and 8 base pair input kernel sizes, which are hyperparameters that may need to be optimized depending on the input DNA sequences' lengths and types. We have tried various different input kernel sizes, and 4 and 8 base pair-long initial kernel sizes together worked well.

We compared the performance of models with one, two, and three deep CNN networks, and found that all converged to similar classification performance. However, models with one or two CNN networks converged faster than the model with three networks due to having fewer parameters. Furthermore, models with two and three networks produced significantly higher logits at the output neurons than the model with a single deep convolutional network, suggesting higher confidence in their predictions. Given the faster convergence and higher confidence of models with multiple CNN networks, we chose to implement our DNA sequence classifier with two networks. The two-channel CNN model yielded superior classification performance, demonstrating the benefits of our approach for accurately classifying DNA sequences.

Furthermore, a skip (residual) connection [18] is established from the input of the first CNN layer to the second non-linear activation to maintain important signals across sequential convolutional layers. DARDN's architecture is visualized in Figure 1b.

Finally, the binary classification prediction of gained and constitutive CTCF sites is generated by merging the outputs from each deep CNN and passing them through a fully connected layer. To train the DARDN model, the binary cross entropy (BCE) loss is computed between the predicted probability of each sample being a CTCF gained site p_i and the true label y_i for each input sequence i:

$$\text{BCELoss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)]$$

where N is the number of input sequences. By minimizing the binary cross entropy loss, DARDN can learn to make accurate predictions on whether a CTCF site is gained or constitutive.

In this work, we demonstrate the effectiveness of our model DARDN (DNAResDualNet) in identifying oncogenic transcription factors (TFs) associated with T cell lymphoblastic leukemia (T-ALL). Specifically, we show that DARDN is capable of accurately identifying TFs that bind to known cancer-specific CTCF binding sites in long DNA sequences.

Applying DeepLIFT and Motif Analysis

DeepLIFT requires a reference value, serving as a null input. It compares the differences in output values obtained by running the actual and reference inputs. This difference is allocated to each base pair through backward propagation, assigning input contribution scores. The resulting scores reflect the extent to which each base pair is responsible for the output difference from the reference. We randomly sampled 80% of the constitutive sites and used the averaged frequency at each index as the reference value. In our approach, we processed sites to be in the shape of $4 \times L$, where 4 corresponds to the four nucleotides and L represents the sequence length. Once we allocated contribution scores to each base, we then performed gating to retrieve the specific nucleotide that exists in the sequence at the intended location.

After obtaining DeepLIFT scores for each gained site, we applied a sliding window of length w bps with a 1 base pair stride across the scores associated with each gained CTCF site. Each base pair was assigned a DeepLIFT score, and w base pair subsequences were assigned a score by averaging their individual base pair scores. The sliding window method produces a total of $L - w + 1$ subsequences. We explored using 10 and 20 window sizes to determine the optimal size for identifying enriched motifs. While the resulting motif enrichments varied slightly across the different window sizes, ultimately decided to use $w = 20$ and 20 bp subsequences to use as HOMER input, as this size yielded more superior results. This process of subsequence selection is demonstrated in Figure 1c.

After obtaining the list of subsequences and their corresponding DeepLIFT scores, we filtered them for further analysis using motif enrichment software through two approaches. The first approach is to select a fixed number of subsequences with either the highest positive mean contribution scores for each gained site. The second approach entails aggregating all subsequences from each gained site and then selecting a fixed number of subsequences with the highest positive mean DeepLIFT scores. The first approach may be more suitable when dominant oncogene occurrences around each CTCF gained site are consistent, while the second approach may be more advantageous when oncogene occurrences vary across CTCF gained sites. Although both approaches are viable, we chose to implement the second approach and selected 1000 subsequences with the highest positive scores. Those subsequences were fed into HOMER, with which we perform known motif analysis using the *findMotifsGenome.pl* module and 200 base pair search space. This resulted in the list of most highly enriched motifs, as summarized in Figure 1d.

Results

Performance Evaluation through Simulation

To evaluate the validity of our method and DARDN's classification ability in detecting crucial features in DNA sequences, we conducted a preliminary test using 25,762 real DNA sequences of length 10,000 base pairs (bps) without CTCF binding sites. We replaced any occurrences of the RBPJ consensus sequence (CCTGGGAA) with a random 8bp combination. Then, we inserted the RBPJ consensus sequence at ten random locations in each of the 33% of the sites ($25,762 \times 0.33 \times 10 = 8500$ sites). We trained DARDN on the classification of RBPJ-inserted sequences and achieved 100% accuracy on hold-out data. Subsequently, we used DeepLIFT to assign contribution scores to each base pairs (bps) in the sequence. This approach allowed us to evaluate the performance of our pipeline in accurately identifying inserted RBPJ sequences and assigning relevant scores to each bp.

We first demonstrate the assignment of DeepLIFT scores to sequences that were

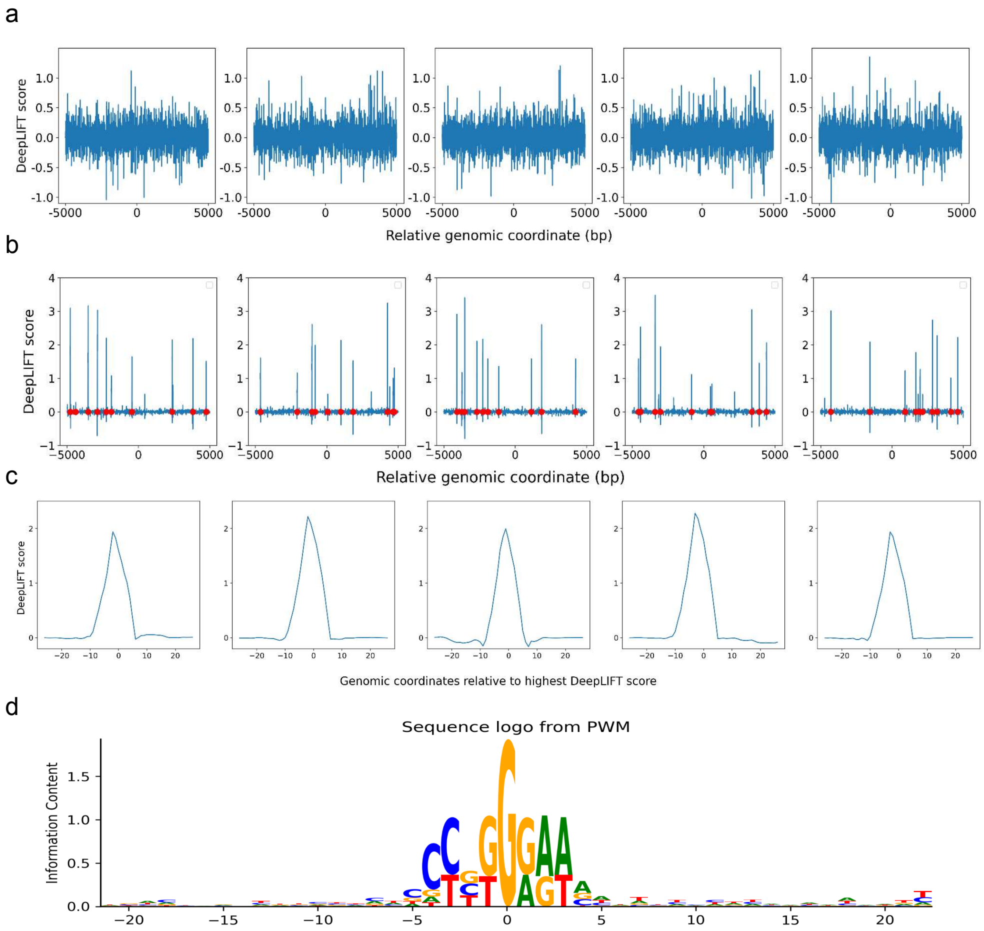

Fig 2. Evaluating the effectiveness of DARDN and motif discovery pipeline using simulation. a. DeepLIFT scores for sites without RBPJ consensus sequence insertion. b. DeepLIFT scores for sites containing 10 RBPJ consensus sequences per site, with red dots marking the locations of the insertions. c. Highest average scoring peak for each site in b. d. Sequence logo surrounding the highest peak from each RBPJ-inserted sites.

trained using DARDN without any RBPJ sites inserted. Since these sites were randomly selected from actual DNA sequences without any specific criteria, the scores do not exhibit any discernible pattern (Figure 2a). On the other hand, after DARDN was trained to classify RBPJ-inserted sequences, it is evident that the DeepLIFT scores at those particular locations with RBPJ insertions (indicated with red dots) are significantly greater than those at other locations (Figure 2b). This provides a clear evidence that DARDN and our score assignment work as expected.

Once the DeepLIFT score at each individual index is computed, we use a sliding window to compute the average scores of the subsequences in the input sequences. Specifically, the average score at index S_i is computed using the formula

$\frac{1}{w+1}\sum_{i-w/2}^{i+w/2} S_i$, where w indicates the window size. Because the RBPJ consensus sequence we inserted contains 8 bps (CCTGGGAA), for our simulation, we used $w = 8$. In our primary experiments, we tested various values of w to optimize the window size for motif enrichment identification.

Figure 2c shows the peak with the highest average DeepLIFT score for each plot in Figure 2b, after re-indexing to center at 0. Lastly, in Figure 2d, we illustrate the sequence logo generated by computing the Position Weight Matrix (PWM) for the sequences that center at the highest peak at each RBPJ-inserted site. Evidently, the inserted sequence of CCTGGGAA is displayed with the highest frequency in the center, which further validates our pipeline.

Robustness Evaluation

To comprehensively evaluate the robustness of DARDN, we subjected it to four distinct test conditions and observed the enrichment of RBPJ, which we noted in our previous research as the most enriched motif for T-ALL [4]. The test conditions we considered are 1) modifying subsequence lengths for HOMER input: this scenario involves examining how changes in subsequence lengths influence motif rankings. This is equivalent to the window size with which we compute the running average DeepLIFT scores; 2) altering input sequence lengths: we explore how motif enrichment changes with input sequences of various lengths, specifically 5,000, 10,000, and 20,000 base pairs (bps); 3) sampling background control sequences from constitutive CTCF sites: this entails studying the effect of sampling constitutive sites on motif rankings; 4) sampling foreground sequences from cancer-specific CTCF sites: we investigate the impact on motif rankings when gained sites are sampled. In our experiments, we carefully selected 150 most statistically significant T-ALL-specific CTCF gained sites and 22,097 constitutive CTCF sites, which were subsequently centered within the sequences.

In our investigation, we explored subsequence lengths ranging from 10 to 20 base pairs (bps) and discovered that adopting a subsequence length of 20 bps consistently yielded superior rankings for RBPJ, irrespective of the input sequence length (5kbps, 10kbps, or 20kbps). In Figure 4a, we present the percentile rank of RBPJ across various combinations of input sequence length and subsequence length. The x-axis represents

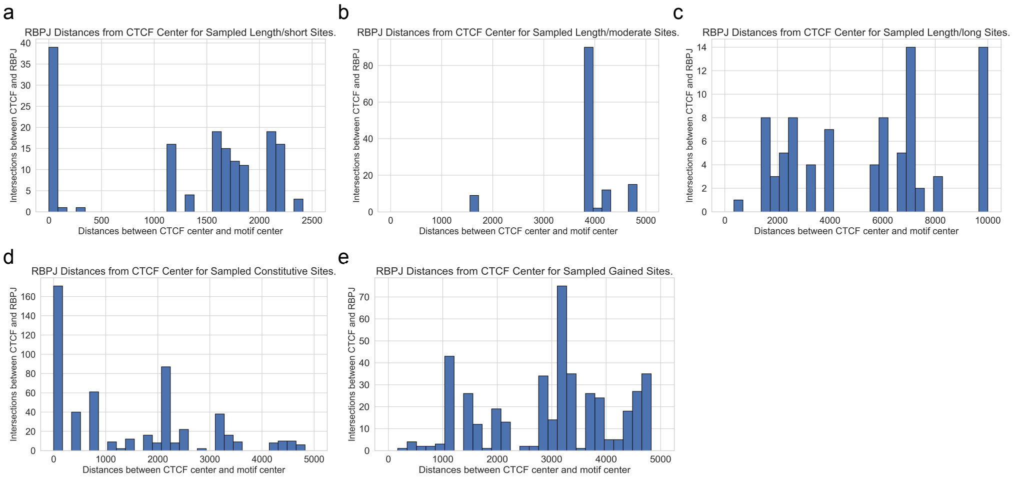

Fig 3. The distribution of center-to-center distances between T-ALL-specific CTCF sites and identified RBPJ sites was examined under various robustness tests. a-c. The distributions when the input sequence lengths are 5k, 10k, 20k bps respectively. d. The distribution obtained by independently sampling constitutive sites 5 times. This is the aggregate distribution of the 5 sampling experiments. e. The distribution obtained by independently sampling gained sites 5 times. This is the aggregate distribution of the 5 sampling experiments.

the sequence lengths of 5kbps, 10kbps, and 20kbps, denoted as “short,” “moderate,” and “long,” respectively.

Using 10bp subsequences, RBPJ achieved the 91st percentile (3 out of 32 enriched motifs) for the short input sequence length, the 97.3th percentile (7 out of 264 enriched motifs) for the moderate input sequence length, and the 96.4th percentile (9 out of 264 enriched motifs) for the long input sequence length. On the other hand, when utilizing 20bp subsequences, RBPJ achieved the 99.7th percentile (1st out of 264 enriched motifs) for the short input sequence length, 99.2th percentile (2nd rank among 264 enriched motifs) for the moderate input sequence length, and 98.5th percentile (4th rank among 264 enriched motifs) for long input sequence length. Regarding the classification accuracy, DARDN demonstrated a Matthews correlation coefficient (MCC) of 0.91 for the short input sequence length, as well as 0.87 for both the moderate and long input sequence lengths.

To further evaluate the robustness of DARDN, we conducted five samplings of the background constitutive sites and five separate samplings of the foreground specific sites. In each trial, we randomly selected 15,000 out of 22,097 (approximately 68%) background constitutive sites and 72 out of 150 foreground specific sites. The performance of DARDN was individually evaluated on each set of sampled sites, and the respective results are presented in Figure 4b.

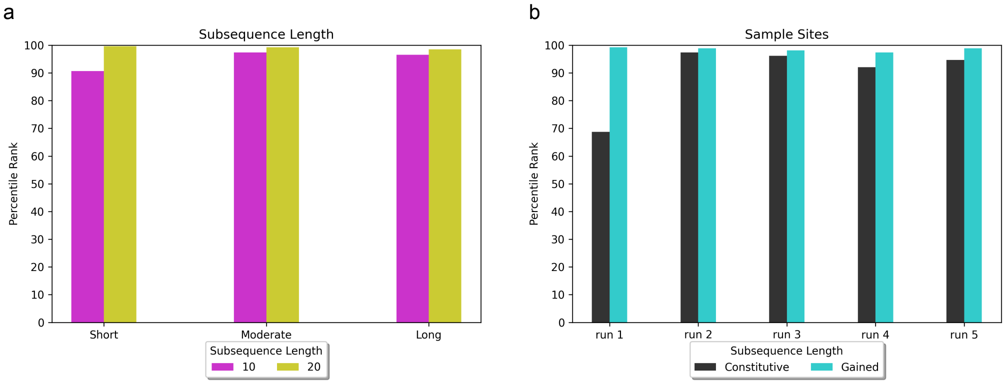

Fig 4. Evaluating the robustness of DARDN and our motif discovery pipeline under varying conditions, such as subsequence length, input sequence length, sampling constitutive CTCF, and sampling specific CTCF sites. Short, moderate, and long indicate input sequence length of 5k, 10k, and 20k bps. a. Observing the impact of subsequence length, ranging between 10bps and 20bps, on RBPJ rank. b. Observing the impact of sampling constitutive and gained sites.

During one of the tests using sampled background constitutive sites (run 1 in Figure 4b), we observed a decline in the average rank of RBPJ compared to our previous trials that involved the complete set of 22,097 constitutive sites. RBPJ achieved the following percentiles and rankings in the five trials: 69 percentile (10th out of 32 enriched motifs), 97.3 percentile (7th out of 264 enriched motifs), 96.2 percentile (10th out of 264 enriched motifs), 92.1 percentile (21st out of 264 enriched motifs), and 94.7 percentile (14th out of 264 enriched motifs). The corresponding MCC values for classification were 0.88, 0.92, 0.81, 0.80, and 0.84, respectively. This outcome was expected as reducing the number of background constitutive sites not only diminishes the pool of negative samples and also can weaken the robustness of DeepLIFT reference values.

For the first sampling of the foreground specific sites, we specifically sampled 72 most significant T-ALL-specific CTCF sites, measured by the specificity and the enrichment of the occurrences. Samplings 2 through 5 involved random samplings of 72 sites from the top 150 gained sites. The MCC scores for classification and the rankings of RBPJ for these trials were notably higher than those observed in the classification involving sampled constitutive sites: 99.2nd percentile (2nd out of 264 enriched motifs), 98.9th percentile (3rd out of 264 enriched motifs), 98.1st percentile (5th out of 264 enriched motifs), 96.4th percentile (7th out of 264 enriched motifs), and 98.1st

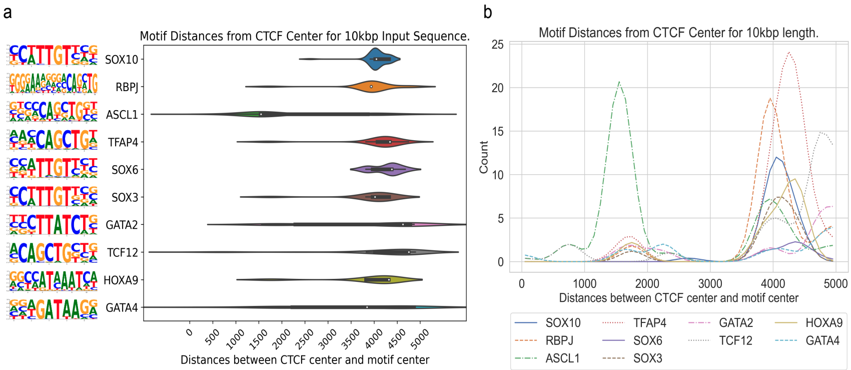

Fig 5. Distribution of distances for the most enriched motifs for T-ALL. a. Violin plots showing the distribution of individual motif's distance to the CTCF center. b. Distance distribution for the motifs in a after applying Gaussian smoothing.

percentile (3rd out of 264 enriched motifs). Both the classification accuracy as well as the ranking of RBPJ reached the most significant values among the five sampling experiments of the gained sites.

In Figure 3, we present the distributions of distances between each specific CTCF site and enriched RBPJ site under the five test criteria. Figure 3a-c showcase the distance variations for different input sequence lengths of 5k, 10k, and 20k bps, respectively. Figure 3 d and e visualize the distances obtained by independently sampling constitutive and gained sites five times. For any input sequence length, the identified RBPJ sites may occur at any distance from the foreground specific CTCF sites, suggesting that long-range interactions exist between cooperating transcription factors and specific CTCF.

In Figure 5, we show the CTCF center to motif center distance distributions for the most enriched motifs for T-ALL, including RBPJ. As shown in Figure 5a, individual motif's distances vary widely, while the median remains around from 3500 to 5000 bps away from CTCF center. In Figure 5b, the distance distributions for the motifs in Figure 5a are plotted using 1-D Gaussian smoothing. We do not observe a trend of close genomic distance between the specific CTCF binding and identified motif sites for transcription factor binding, indicating that the long-range interactions can occur at a long distance through DNA looping.

Application of DARDN to Diverse Cancer Types

To evaluate the adaptability of DARDN sequence feature identification method, we applied them on five other cancer types where cancer-specific CTCF sites were previously identified [4]: acute myeloid leukemia (AML), breast invasive carcinoma (BRCA), colorectal cancer (CRC), lung adenocarcinoma (LUAD), and prostate adenocarcinoma (PRAD), using the moderate input sequence length of 10kbps.

In all five cancer types, the motif for CTCF or CTCFL (a. k. a. BORIS, a paralog of CTCF) are highly enriched near the CTCF center. As both the foreground and background sequences are centered at specific or constitutive CTCF binding sites, respectively, the enrichment of CTCF motif indicates additional CTCF occupancy near these specific sites. This is consistent with the fact that CTCF binding exhibits a clustered pattern in the genome to maintain the higher-order chromatin structure [5].

Meanwhile, the relatively uniform distribution of the remaining motifs across the sequence length shown in the Gaussian-smoothed line plots in Figures [6-10] in the Appendix indicates potential long-range interactions between CTCF and other transcription factors through looping structures. The full list of cancer-specific enriched motifs are presented in Tables [2-6] in the Appendix.

Overall, this pattern of enrichment and distribution of different sequence motifs surrounding cancer-specific CTCF sites suggests that the regulatory mechanisms governing gene expression are specific to each cancer types and potentially involve in the specific CTCF binding events to facilitate enhancer-promoter interactions for oncogenic transcription factors to regulate their target genes.

Conclusion

This work presents DARDN, a novel deep learning computational method using dual CNNs and DeepLIFT for identifying enriched motifs in long DNA sequences. DARDN accurately classifies sequences surrounding cancer-specific vs constitutive CTCF sites. DeepLIFT selects important subsequences for motif analysis. DARDN identified simulated and known cancer motifs like RBPJ in T-ALL. Application to AML, BRCA, CRC, LUAD, and PRAD revealed distinct motifs, implying cancer-specific regulation. DARDN provides an effective framework combining deep learning and attribution for discovering functional sequence features from long genomic data without localization, addressing a key challenge in distal regulation. Our versatile approach is broadly applicable for mining insights from diverse biological sequences. DARDN represents a powerful methodology leveraging machine learning and feature discovery for extracting biological insights from complex genomic data.

References

1. Mitchell J. Rowley and Victor G. Corces. Organizational principles of 3D genome architecture. *Nature Reviews Genetics*, 19(12):789-800, 2018.

2. William A. Flavahan, Yotam Drier, Benjamin B. Liau, et al. Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature*, 529(7584):110-114, 2016.

3. William A. Flavahan, Yotam Drier, Sarah E. Johnstone, et al. Altered chromosomal topology drives oncogenic programs in SDH-deficient GISTs. *Nature*, 575(7781):229-233, 2019.

4. Chang Fang, Zhonghui Wang, Congfei Han, et al. Cancer-specific CTCF binding facilitates oncogenic transcriptional dysregulation. *Genome Biology*, 21(1):247, 2020.

5. Kentepozidou E, Aitken SJ, Feig C, et al. Clustered CTCF binding is an evolutionary mechanism to maintain topologically associating domains. *Genome Biology*, 21:5, 2020.

6. Zhonghui Wang, Mete Civelek, Cynthia L. Miller, et al. BART: a transcription factor prediction tool with query gene sets or epigenomic profiles. *Bioinformatics*, 34(16): 2867-2869, 2018.

7. Heinz Sven, Benner Christopher, Spann Nathanael, Bertolino Eric et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol Cell*, 38(4):576-589, 2010.

8. Gökcen Eraslan, Žiga Avsec, Julien Gagneur, Fabian J. Theis. Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics*, 20(7): 389-403, 2019.

9. Jian Zhou, Olga G. Troyanskaya. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, 12(10): 931-934, 2015.

10. Babak Alipanahi, Andrew Delong, Matthew T. Weirauch, Brendan J. Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8): 831-838, 2015.

11. Khoi A. Tran, Olga Kondrashova, Alexander Bradley, et al. Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Medicine*, 13(1):152, 2021.

12. Avanti Shrikumar, Peyton Greenside, Anshul Kundaje. Learning Important Features Through Propagating Activation Differences. *arXiv preprint arXiv:1704.02685*, 2017.

13. David R. Kelley, Jasper Snoek, John L. Rinn. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, 26(7): 990-999, 2016.

14. Žiga Avsec, Max Weilert, Avanti Shrikumar, et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*, 53(3): 354-366, 2021.

15. Geoffrey Fudenberg, David R. Kelley, Katherine S. Pollard. Predicting 3D genome folding from DNA sequence with Akita. *Nature Methods*, 17(11): 1111-1117, 2020.

16. Jian Zhou. Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale. *Nature Genetics*, 54(5): 725-734, 2022.

17. Ron Schwessinger, Maya Gosden, Daniel Downes, et al. DeepC: predicting 3D genome folding using megabase-scale transfer learning. *Nature Methods*, 17(11): 1118-1124, 2020.

18. Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770-778, 2016.

19. Bruno U. Mueller, Thomas Pabst, Mitsuhiro Osato, et al. Heterozygous PU.1 mutations are associated with acute myeloid leukemia. *Blood*, 100(3):998-1007, 2002.

20. Tom Verbiest, Susan Bouffler, Simon Nutt, Christophe Badie. PU.1 downregulation in murine radiation-induced acute myeloid leukaemia (AML): from molecular mechanism to human AML. *Carcinogenesis*, 36(4):413-419, 2015.

21. Hiroshi Takei, Shizuka S. Kobayashi. Targeting transcription factors in acute myeloid leukemia. *International Journal of Hematology*, 109(1):28-34, 2019.

22. Christopher P. Mill, Weiguo Fiskus, Daniel A. DiNardo, et al. Effective therapy for AML with RUNX1 mutation by cotreatment with inhibitors of protein translation and BCL2. *Blood*, 139(6):907-921, 2022.

23. Lars Bullinger, Konstanze Döhner, Hartmut Döhner. Genomics of Acute Myeloid Leukemia Diagnosis and Pathways. *Journal of Clinical Oncology*, 35(9):934-946, 2017.

24. Floriane Gonzales, Agathe Barthélémy, Pauline Peyrouze, et al. Targeting RUNX1 in acute myeloid leukemia: preclinical innovations and therapeutic implications. *Expert Opinion on Therapeutic Targets*, 25(4):299-309, 2021.

25. C. Barletta, T. Druck, S. LaForgia, et al. Chromosome locations of the MYB related genes, AMYB and BMYB. *Cancer Research*, 51(14):3821-3824, 1991.

26. Giorgio Zauli, Rachele Voltan, Maria Grazia di Iasio, et al. miR-34a induces the downregulation of both E2F1 and B-Myb oncogenes in leukemic cells. *Clinical Cancer Research*, 17(9):2712-2724, 2011.

27. Conchillo A, Pauwels D, et al. MYB Overexpression Is Directly Involved in Acute Myeloid Leukemia Pathogenesis and Could Constitute a New Therapeutic Target for Patients with Aberrant Expression of This Gene. *Blood*, 114(22):2402-2402, 2009.

28. Vidya Sasidharan Nair, Salma M. Toor, Bashir Alobeid, Eman Elkord. Dual inhibition of STAT1 and STAT3 activation downregulates expression of PD-L1 in human breast cancer cells. *Expert Opinion on Therapeutic Targets*, 22(6):547-557, 2018.

29. Jie Zhu, Yu Shen, Li Wang, Jie Qiao, Yun Zhao, Qiong Wang. A novel 12-gene prognostic signature in breast cancer based on the tumor microenvironment. *Annals of Translational Medicine*, 10(3):143, 2022.

30. Jie Zhu, Yu Shen, Li Wang, Jie Qiao, Yun Zhao, Qiong Wang. A novel 12-gene prognostic signature in breast cancer based on the tumor microenvironment. *Annals of Translational Medicine*, 10(3):143, 2022.

31. Marcin Cieślik, Shubhada A. Hoang, Nataliya Baranova, et al. Epigenetic coordination of signaling pathways during the epithelial-mesenchymal transition. *Epigenetics & Chromatin*, 6(1):28, 2013.

32. Haiyan Dong, Li Ding, Tingru Zhou, Teng Yan, Jianying Li, Chunhong Liang. FOXA1 in prostate cancer. *Asian Journal of Andrology*, 2023.

33. Brian Song, Sungyong Youl Park, Juan C. Zhao, et al. Targeting FOXA1-mediated repression of TGF-β signaling suppresses castration-resistant prostate cancer progression. *Journal of Clinical Investigation*, 129(2):569-582, 2019.

34. Minghui Teng, Shanshan Zhou, Chao Cai, Mathieu Lupien, Housheng He. Pioneer of prostate cancer: past, present and the future of FOXA1. *Protein & Cell*, 12(1):29-38, 2021.

35. Maria Del Giudice, James G. Foster, Simona Peirone, et al. FOXA1 regulates alternative splicing in prostate cancer. *Cell Reports*, 40(13):111404, 2022.

36. Haojie Cai, Sara Nørgaard Sjøgren, Anders Sjøgren, et al. In Vivo Application of CRISPR/Cas9 Revealed Implication of Foxa1 and Foxp1 in Prostate Cancer Proliferation and Epithelial Plasticity. *Cancers*, 14(18):4381, 2022.

Appendix

The DARDN pipeline was tested on five additional cancer types (AML, BRCA, CRC, LUAD, and PRAD), and the most significantly enriched motifs are listed in Tables [2, 3, 4, 5, 6]. Unlike T-ALL, the most prominent oncogenes for some of these cancers are less studied. However, as listed in detail in the Results section of this paper, we found existing literature support for some of the identified motifs, such as PU.1 (SPI1) [19–22], RUNX-related genes [21–24] and MYB gene family [22, 25–27] for AML, STAT1 [28], STAT5 [29], ASCL1 [30], for BRCA, for CRC, AP1 [31] for LUAD, and FOXA1 [32–35] and FOXP1 [36] for PRAD.

Figures [6-10] demonstrates the distribution of distances between the CTCF-center and the motif site-center for the most significantly enriched motifs associated with AML, BRCA, CRC, LUAD, and PRAD, complemented by the representation of each motif's sequence logos.

Rank	Motif	p-value	q-value (BH)
1	CTCF(Zf)/CD4+-CTCF(Barski et al.)	1E-111	< 1E-4
2	SpiB(ETS)/OCILY3-SPIB(GSE56857)	1E-67	< 1E-4
3	ELF5(ETS)/T47D-ELF5(GSE30407)	1E-55	< 1E-4
4	PU.1(ETS)/ThioMac-PU.1(GSE21512)	1E-54	< 1E-4
5	ETS1(ETS)/Jurkat-ETS1(GSE17954)	1E-46	< 1E-4
6	Fli1(ETS)/CD8-FLI(GSE20898)	1E-37	< 1E-4
7	BORIS(Zf)/K562-CTCFL(GSE32465)	1E-33	< 1E-4
8	ERG(ETS)/VCaP-ERG(GSE14097)	1E-30	< 1E-4
9	ETV1(ETS)/GIST48-ETV1(GSE22441)	1E-28	< 1E-4
10	PU.1-IRF(ETS:IRF)/Bcell-PU.1(GSE21512)	1E-28	< 1E-4
11	AMYB(HTH)/Testes-AMYB(GSE44588)	1E-28	< 1E-4
12	RUNX(Runt)/HPC7-Runx1(GSE22178)	1E-22	< 1E-4
13	RUNX1(Runt)/Jurkat-RUNX1(GSE29180)	1E-21	< 1E-4
14	RUNX2(Runt)/PCa-RUNX2(GSE33889)	1E-21	< 1E-4
15	EHF(ETS)/LoVo-EHF(GSE49402)	1E-19	< 1E-4
16	Elk4(ETS)/Hela-Elk4(GSE31477)	1E-18	< 1E-4
17	JunD(bZIP)/K562-JunD	1E-16	< 1E-4
18	EWS:ERG-fusion(ETS)/CADO_ES1-EWS:ERG(SRA014231)	1E-16	< 1E-4
19	ETS:E-box(ETS,bHLH)/HPC7-Scl(GSE22178)	1E-16	< 1E-4
20	MYB(HTH)/ERMYB-Myb-ChIPSeq(GSE22095)	1E-15	< 1E-4

Table 2. 20 most significantly enriched motifs in AML. BH q-values indicate the Benjamini-Hochberg q-values, which are multiple comparison corrected. 1301 gained sites.

Rank	Motif	p-value	q-value (BH)
1	CTCF(Zf)/CD4+-CTCF(Barski et al.)	1E-319	< 1E-4
2	BORIS(Zf)/K562-CTCFL(GSE32465)	1E-318	< 1E-4
3	Tcf12(bHLH)/GM12878-Tcf12(GSE32465)	1E-39	< 1E-4
4	NeuroD1(bHLH)/Islet-NeuroD1(GSE30298)	1E-24	< 1E-4
5	Olig2(bHLH)/Neuron-Olig2(GSE30882)	1E-20	< 1E-4
6	MyoD(bHLH)/Myotube-MyoD(GSE21614)	1E-20	< 1E-4
7	Myf5(bHLH)/GM-Myf5(GSE24852)	1E-19	< 1E-4
8	SCL(bHLH)/HPC7-Scl(GSE13511)	1E-19	< 1E-4
9	EBF1(EBF)/Near-E2A(GSE21512)	1E-14	< 1E-4
10	Bcl6(Zf)/Liver-Bcl6(GSE31578)	1E-14	< 1E-4
11	Ap4(bHLH)/AML-Tfap4(GSE45738)	1E-14	< 1E-4
12	Atoh1(bHLH)/Cerebellum-Atoh1(GSE22111)	1E-14	< 1E-4
13	STAT1(Stat)/HelaS3-STAT1(GSE12782)	1E-14	< 1E-4
14	Tlx(NR)/NPC-H3K4me1(GSE16256)	1E-13	< 1E-4
15	Ptf1a(bHLH)/Panc1-Ptf1a(GSE47459)	1E-13	< 1E-4
16	STAT5(Stat)/mCD4+-Stat5(GSE12346)	1E-12	< 1E-4
17	Ascl1(bHLH)/NeuralTubes-Ascl1(GSE55840)	1E-12	< 1E-4
18	E2A(bHLH),near_PU.1/Bcell-PU.1(GSE21512)	1E-12	< 1E-4
19	MyoG(bHLH)/C2C12-MyoG(GSE36024)	1E-11	< 1E-4
20	SPDEF(ETS)/VCaP-SPDEF(SRA014231)	1E-11	< 1E-4

Table 3. 20 most significantly enriched motifs in BRCA. BH q-values indicate the Benjamini-Hochberg q-values, which are multiple comparison corrected. 1616 gained sites.

Rank	Motif	p-value	q-value (BH)
1	Tcf3(HMG)/mES-Tcf3(GSE11724)	1E-19	< 1E-4
2	CTCF(Zf)/CD4+-CTCF(Barski et al.)	1E-17	< 1E-4
3	EWS:ERG-fusion(ETS)/CADO_ES1-EWS:ERG(SRA014231)	1E-14	< 1E-4
4	Stat3+il21(Stat)/CD4-Stat3(GSE19198)	1E-10	< 1E-4
5	BORIS(Zf)/K562-CTCFL(GSE32465)	1E-10	< 1E-4
6	EWS:FLI1-fusion(ETS)/SK_N_MC-EWS:FLI1(SRA014231)	1E-10	< 1E-4
7	Elk4(ETS)/Hela-Elk4(GSE31477)	1E-09	< 1E-4
8	Elk1(ETS)/Hela-Elk1(GSE31477)	1E-09	< 1E-4
9	Fli1(ETS)/CD8-FLI(GSE20898)	1E-09	< 1E-4
10	ETV1(ETS)/GIST48-ETV1(GSE22441)	1E-09	< 1E-4
11	NF1:FOXA1(CTF,Forkhead)/LNCAP-FOXA1(GSE27824)	1E-08	< 1E-4
12	Pit1+1bp(Homeobox)/GCrat-Pit1(GSE58009)	1E-08	< 1E-4
13	AP-2gamma(AP2)/MCF7-TFAP2C(GSE21234)	1E-08	< 1E-4
14	GABPA(ETS)/Jurkat-GABPa(GSE17954)	1E-08	< 1E-4
15	ETS(ETS)/Promoter	1E-08	< 1E-4
16	Tbet(T-box)/CD8-Tbet(GSE33802)	1E-07	< 1E-4
17	Ets1-distal(ETS)/CD4+-PolII(Barski et al.)	1E-07	< 1E-4
18	Sp1(Zf)/Promoter	1E-07	< 1E-4
19	PRDM1(Zf)/Hela-PRDM1(GSE31477)	1E-07	< 1E-4
20	ELF1(ETS)/Jurkat-ELF1(SRA014231)	1E-07	<1E-4

Table 4. 20 most significantly enriched motifs in CRC. BH q-values indicate the Benjamini-Hochberg q-values, which are multiple comparison corrected. 377 gained sites.

Rank	Motif	p-value	q-value (BH)
1	BORIS(Zf)/K562-CTCFL(GSE32465)	1E-56	$<$ 1E-4
2	Jun-AP1(bZIP)/K562-cJun(GSE31477)	1E-38	$<$ 1E-4
3	Fosl2(bZIP)/3T3L1-Fosl2(GSE56872)	1E-30	$<$ 1E-4
4	Reverb(NR),DR2/RAW-Reverba.biotin(GSE45914)	1E-29	$<$ 1E-4
5	AP-1(bZIP)/ThioMac-PU.1(GSE21512)	1E-27	$<$ 1E-4
6	BATF(bZIP)/Th17-BATF(GSE39756)	1E-24	$<$ 1E-4
7	Fra1(bZIP)/BT549-Fra1(GSE46166)	1E-23	$<$ 1E-4
8	Usf2(bHLH)/C2C12-Usf2(GSE36030)	1E-17	$<$ 1E-4
9	Atf3(bZIP)/GBM-ATF3(GSE33912)	1E-16	$<$ 1E-4
10	MITF(bHLH)/MastCells-MITF(GSE48085)	1E-14	$<$ 1E-4
11	Pit1(Homeobox)/GCrat-Pit1(GSE58009)	1E-13	$<$ 1E-4
12	MafA(bZIP)/Islet-MafA(GSE30298)	1E-13	$<$ 1E-4
13	PRDM9(Zf)/Testis-DMC1(GSE35498)	1E-13	$<$ 1E-4
14	ERE(NR),IR3/MCF7-ERa(Unpublished)	1E-11	$<$ 1E-4
15	Gata4(Zf)/Heart-Gata4(GSE35151)	1E-11	$<$ 1E-4
16	Bach2(bZIP)/OCILy7-Bach2(GSE44420)	1E-10	$<$ 1E-4
17	Gata1(Zf)/K562-GATA1(GSE18829)	1E-10	$<$ 1E-4
18	Brachyury(T-box)/Mesoendoderm-Brachyury-ChIP-exo(GSE54963)	1E-09	$<$ 1E-4
19	Gata2(Zf)/K562-GATA2(GSE18829)	1E-09	$<$ 1E-4
20	RUNX1(Runt)/Jurkat-RUNX1(GSE29180)	1E-08	$<$ 1E-4

Table 5. 20 most significantly enriched motifs in LUAD. BH q-values indicate the Benjamini-Hochberg q-values, which are multiple comparison corrected. 357 gained sites.

Rank	Motif	p-value	q-value (BH)
1	CTCF(Zf)/CD4+-CTCF(Barski et al.)	1E-135	$<$ 1E-4
2	BORIS(Zf)/K562-CTCFL(GSE32465)	1E-72	$<$ 1E-4
3	NF1:FOXA1(CTF,Forkhead)/LNCAP-FOXA1(GSE27824)	1E-17	$<$ 1E-4
4	STAT5(Stat)/mCD4+-Stat5(GSE12346)	1E-15	$<$ 1E-4
5	Pit1(Homeobox)/GCrat-Pit1(GSE58009)	1E-14	$<$ 1E-4
6	Pdx1(Homeobox)/Islet-Pdx1(SRA008281)	1E-14	$<$ 1E-4
7	FOXP1(Forkhead)/H9-FOXP1(GSE31006)	1E-13	$<$ 1E-4
8	AP-2gamma(AP2)/MCF7-TFAP2C(GSE21234)	1E-12	$<$ 1E-4
9	EKLF(Zf)/Erythrocyte-Klf1(GSE20478)	1E-10	$<$ 1E-4
10	Pit1+1bp(Homeobox)/GCrat-Pit1(GSE58009)	1E-09	$<$ 1E-4
11	Maz(Zf)/HepG2-Maz(GSE31477)	1E-08	$<$ 1E-4
12	RORgt(NR)/EL4-RORgt.Flag(GSE56019)	1E-08	$<$ 1E-4
13	FXR(NR),IR1/Liver-FXR(Chong et al.)	1E-07	$<$ 1E-4
14	STAT4(Stat)/CD4-Stat4(GSE22104)	1E-07	$<$ 1E-4
15	EHF(ETS)/LoVo-EHF(GSE49402)	1E-07	$<$ 1E-4
16	Pax7(Paired,Homeobox)/Myoblast-Pax7(GSE25064)	1E-07	$<$ 1E-4
17	Rbpj1/Panc1-Rbpj1(GSE47459)	1E-07	$<$ 1E-4
18	EBF1(EBF)/Near-E2A(GSE21512)	1E-06	$<$ 1E-4
19	NF-E2(bZIP)/K562-NFE2(GSE31477)	1E-06	$<$ 1E-4
20	STAT1(Stat)/HelaS3-STAT1(GSE12782)	1E-06	$<$ 1E-4

Table 6. 20 most significantly enriched motifs in PRAD. BH q-values indicate the Benjamini-Hochberg q-values, which are multiple comparison corrected. 309 gained sites.

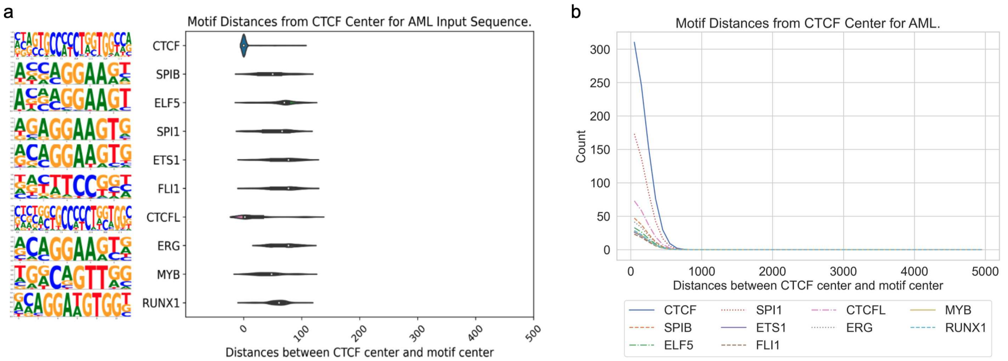

Fig 6. The distribution of center-to-center distances between cancer-specific CTCF sites and sites of enriched motifs for AML. a. Violin plot for AML. b. Gaussian-smoothed line plot for AML.

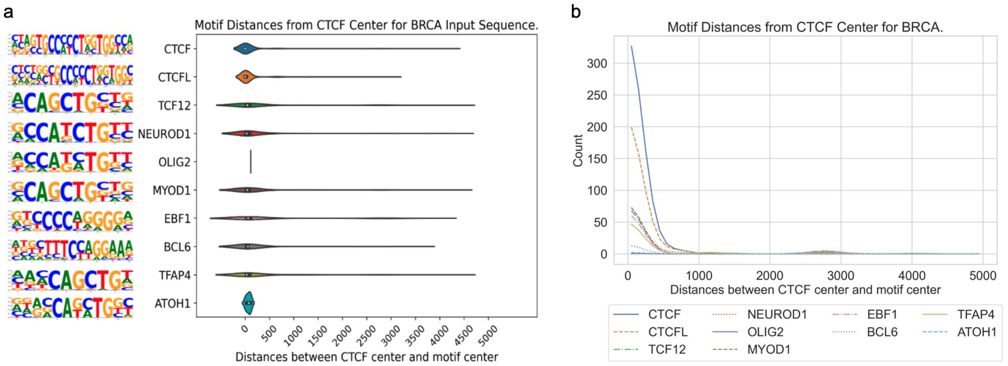

Fig 7. The distribution of center-to-center distances between cancer-specific CTCF sites and sites of enriched motifs for BRCA. a. Violin plot for BRCA. b. Gaussian-smoothed line plot for BRCA.

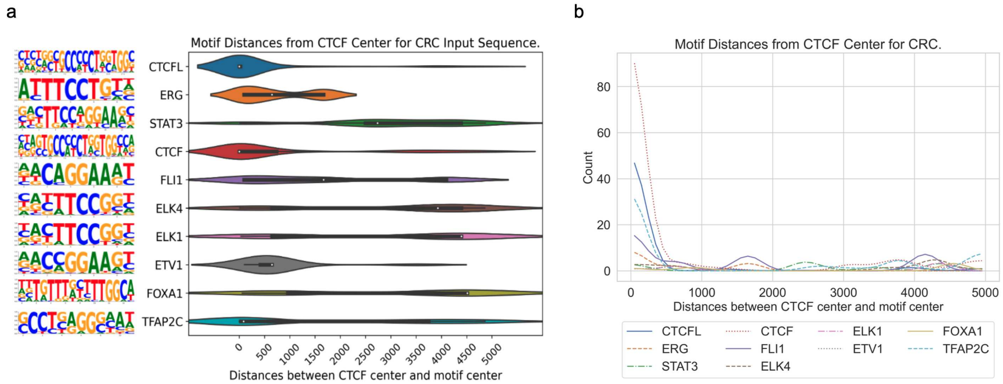

Fig 8. The distribution of center-to-center distances between cancer-specific CTCF sites and sites of enriched motifs for CRC. a. Violin plot for CRC. b. Gaussian-smoothed line plot for CRC.

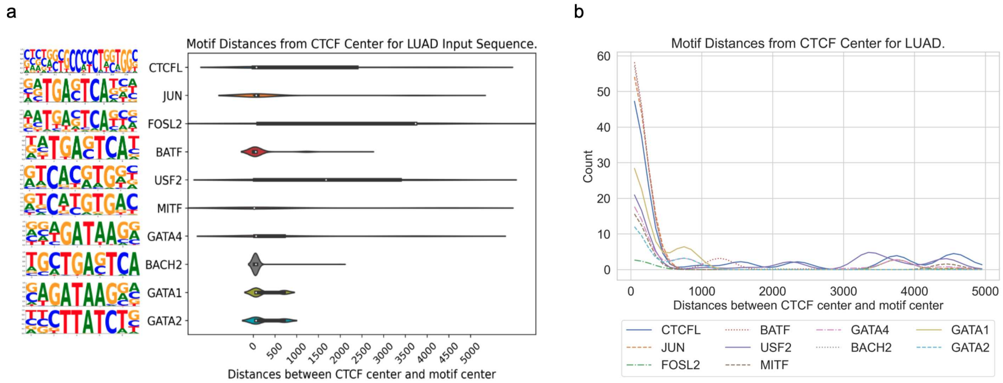

Fig 9. The distribution of center-to-center distances between cancer-specific CTCF sites and sites of enriched motifs for LUAD. a. Violin plot for LUAD. b. Gaussian-smoothed line plot for LUAD.

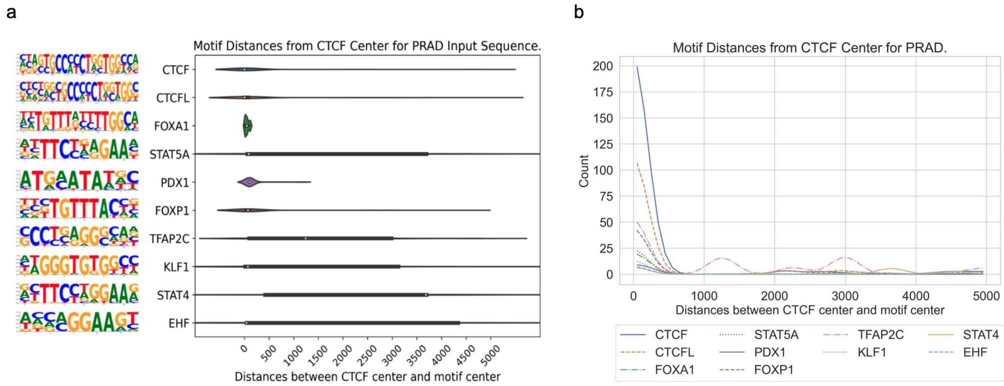

Fig 10. The distribution of center-to-center distances between cancer-specific CTCF sites and sites of enriched motifs for PRAD. a. Violin plot for PRAD. b. Gaussian-smoothed line plot for PRAD.