Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Method article

DeepCORE: An interpretable multi-view deep neural network model to detect co-operative regulatory elements

Pramod Bharadwaj Chandrashekar [a,b,1], Hai Chen [c,d,1], Matthew Lee [c], Navid Ahmadinejad [c,d], Li Liu [c,d,*]

[a] *Waisman Center, University of Wisconsin-Madison, Madison, WI 53705, USA*
[b] *Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53076, USA*
[c] *College of Health Solutions, Arizona State University, Phoenix, AZ, United States*
[d] *Biodesign Institute, Arizona State University, Tempe, AZ, United States*

ARTICLE INFO

Keywords:
Gene regulation
Deep learning
Epigenetics
Cooperative regulatory elements

ABSTRACT

Gene transcription is an essential process involved in all aspects of cellular functions with significant impact on biological traits and diseases. This process is tightly regulated by multiple elements that co-operate to jointly modulate the transcription levels of target genes. To decipher the complicated regulatory network, we present a novel multi-view attention-based deep neural network that models the relationship between genetic, epigenetic, and transcriptional patterns and identifies co-operative regulatory elements (COREs). We applied this new method, named DeepCORE, to predict transcriptomes in various tissues and cell lines, which outperformed the state-of-the-art algorithms. Furthermore, DeepCORE contains an interpreter that extracts the attention values embedded in the deep neural network, maps the attended regions to putative regulatory elements, and infers COREs based on correlated attentions. The identified COREs are significantly enriched with known promoters and enhancers. Novel regulatory elements discovered by DeepCORE showed epigenetic signatures consistent with the status of histone modification marks.

1. Introduction

Gene transcription displays complicated spatiotemporal patterns that vary between tissue and cell types, developmental stages, disease phenotypes, and environmental exposures [1,2]. Such variations are regulated by a set of mechanisms that induce or repress gene transcription as part of a large network [3,4]. Many factors attribute to gene transcription regulation, such as genetic alterations [5,6], epigenetic changes [7,8], and chromatin structure [9–11]. Deciphering and cataloging these regulatory codes are a grand challenge.

Computational mining of multi-omics data is a promising approach to investigate the mechanisms of gene transcriptional regulation. As early attempts, several models used genetic sequence information such as transcription factor binding sites (TFBS) to predict gene transcription levels. [12–18] However, relying on genetic sequences that cannot capture tissue-specific information is a major limitation of these models. Epigenetic features, such as histone post-translational modifications (hPTMs), are introduced to address this issue. DeepChrome [19] is one of the early deep learning method that models the relationship between epigenetic and transcriptional profiles. It retrieves the hPTM signals in the ± 5kbps region around the transcription start site (TSS) of a gene, uses a convolutional neural network (CNN) to extract local features, and passes these features to a feedforward neural network (FNN) to make binary prediction of gene transcription levels. ExPecto [20] expands the TSS-flanking region to 40kbps, includes ChIP-seq data of hundreds of TFs as input, and predicts gene expression in continuous values. Although these models reported high accuracy of predicting gene transcription levels, they do not identify regulatory elements (REs) in the genome that are essential to understanding the regulatory mechanisms.

Because deep neural networks (DNN) are often considered a black box, extracting biological meanings from these models can be challenging. Recently, several algorithms have been developed to interpret and visualize patterns learned in DNN [21–23]. DeepChrome summarized the hPTM patterns coded in the CNN model, which were consistent with known active and repressive marks. However, a high-level summary cannot identify and locate REs. Furthermore, epigenetic signals

[*] Corresponding author at: College of Health Solutions, Arizona State University, Phoenix, AZ, United States.
E-mail address: liliu@asu.edu (L. Liu).

[1] co-first authors.

https://doi.org/10.1016/j.csbj.2023.12.044
Received 29 August 2023; Received in revised form 14 December 2023; Accepted 27 December 2023
Available online 29 December 2023

highly depend on genetic features. For example, chromatin structure changes involving TFBS will have a larger impact on gene transcription than those outside TFBS. In this study, we present a novel method to address this knowledge gap and model co-operative regulatory elements (COREs).

This new method, named DeepCORE, uses a multi-view architecture to integrate genetic and epigenetic profiles in a DNN. It captures short-range and long-range interactions between REs through bidirectional Long short-term memory (BiLSTM). It leverages the attention mechanism to pinpoint the most informative regions harboring REs and enhance the interpretability of the model. The output of DeepCORE includes prediction of gene transcription level, locations of REs in the genome, and correlations of multiple REs. We applied DeepCORE on various tissues and cell lines and showed that DeepCORE has significantly higher accuracies than existing state-of-the-art methods. DeepCORE model has good generalizability across samples and identifies COREs in high resolution. These putative REs are enriched with known promoters and enhancers.

2. Methods and materials

2.1. Overall design and data sets

In the ENCODE [24] project and the RoadMap Epigenomics project (REMC) [25], we searched for samples that had transcriptomic data (RNA-seq) and epigenetic data (ChIP-seq of H3K4me, H3K4me3, H2K27ac, H3K9me3, and H3K27me3). We randomly selected two of these samples (E061 and E071) and used them for algorithm development and parameter tuning. We randomly selected additional 23 samples to systematically evaluate the performance of DeepCORE, DeepChrome, and ExPecto. To compare with Xpresso [26] predictions, we trained DeepCORE on two more samples (E116 and E123) that were tested in the Xpresso study. These samples included cancer cell lines, embryonic stem cell-derived cell lines, primary cell cultures, and primary tissues (Supplementary Table 1).

DeepCORE has two components (Fig. 1A). The first component is a deep neural network (DNN) that predicts transcription level of a gene based on its genetic and epigenetic features. The second component is an interpreter that analyzes the attention matrices in the DNN to discover COREs.

For a given gene, DeepCORE focuses on the ± 5kbps region

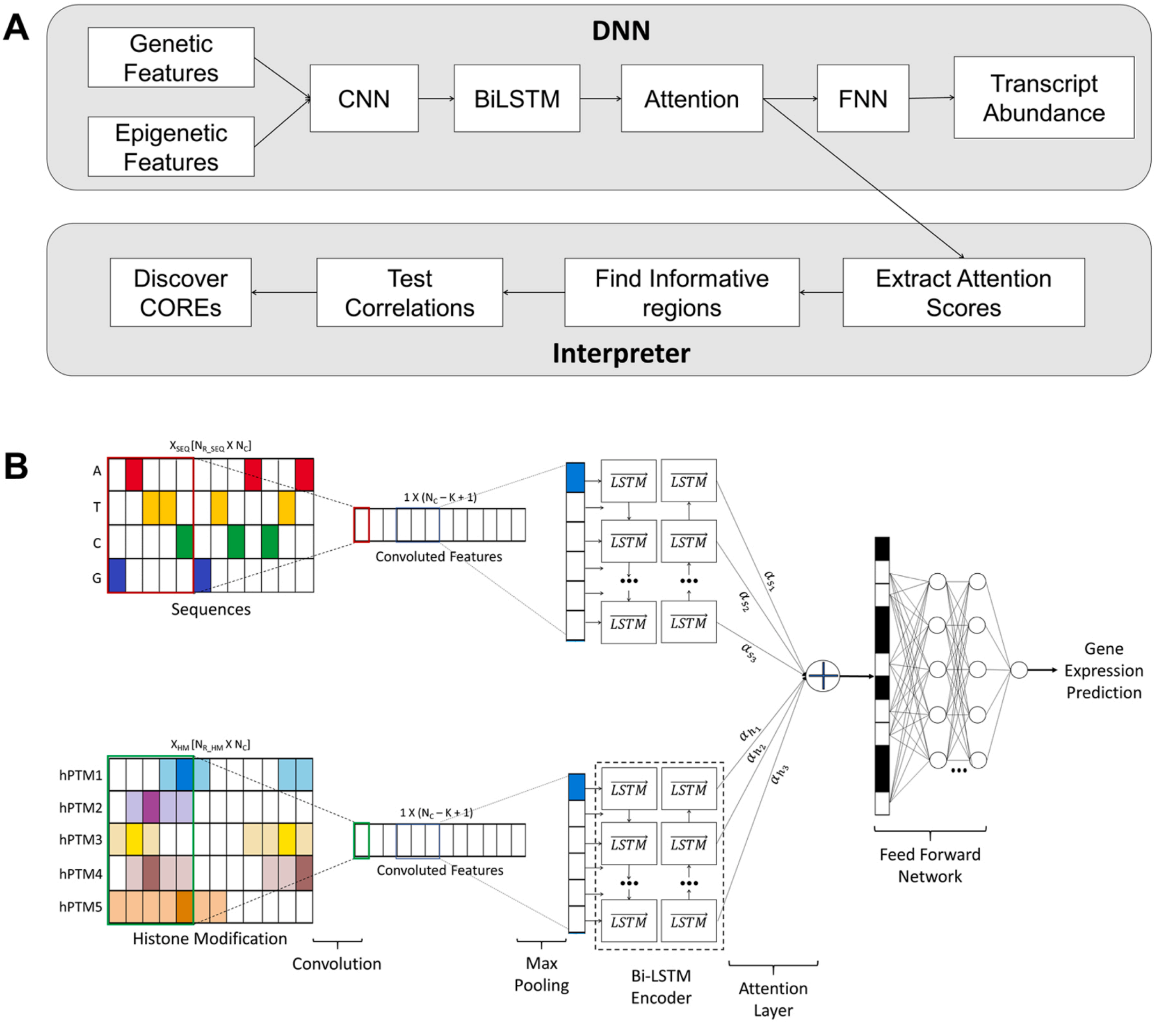


Fig. 1. The DeepCORE architecture. (A) DeepCORE consists of two components. In the DNN component, genetic and epigenetic signals go through a CNN layer, a BiLSTM layer, an Attention layer, and an FNN layer to predict transcript abundance of a gene. In the Interpreter component, attention scores extracted from the output of the Attention layer is analyzed to identify informative and correlated regions as COREs. (B) The DNN has a genetic view and an epigenetic view, each consisting of a CNN layer and a BiLSTM layer. These two views are joint before fed into an attention layer and subsequently an FNN layer to predict gene transcription level. Nc = 10,000 in both genetic and epigenetic feature matrices.

surrounding the TSS. To derive genetic features, we extracted the ± 5kbps nucleotide sequences for each gene and converted into a one hot encoding format. This gives us the genetic feature matrix with four rows corresponding to the nucleotides and 10,000 columns corresponding to genomic regions with the value in each cell corresponding to the presence or absence of a specific nucleotide. It is worth noting that the human reference genome sequence is used, and genetic alterations are not considered.

To derive epigenetic features, we obtained ChIP-seq data of 25 tissue or cell line samples from the ENCODE [24] project and the REMC [25] project. The ChIP-seq data contained normalized read counts measuring five hPTMs including known transcription activator marks (H3K4me, H3K4me3 and H2K27ac) and repressor marks (H3K9me3 and H3K27me3) [27]. Given a gene, we examined the ± 5kbps TSS-flanking region and recorded position-specific normalized read count for each histone modification mark. These data from each sample were organized into an epigenetic feature matrix with five rows corresponding to hPTMs and 10,000 columns corresponding to genomic positions.

The ENCODE and REMC projects also contained RNA-seq data. For each sample, we obtained the Reads Per Kilobase of transcript per Million mapped reads (RPKM) value for each gene. These data were organized into a single-column vector with rows corresponding to genes. For each sample, we removed genes with missing values of RNA-seq data and missing values of ChIP-seq data across all five histone marks.

2.2. Multi-view attention-based DNN

The DNN architecture consists of two separate paths representing the genetic view and the epigenetic view, respectively (Fig. 1B). Each path starts with a CNN layer consisting of N_{CNN} filters with size k and stride 1. The output of the CNN layer is passed to a ReLU function connected to max pooling over non-overlapping interval of length p. These steps produce a vector of size $(N_C - K + 1)/p$ for filter $f_i, i \in \{1, \ldots, N_{CNN}\}$ encoding sequence motifs and another vector of equal size encoding combinations of histone modification patterns. To avoid overfitting, dropout [28] with a probability of 0.5 is applied to the max-pooled vector. While CNN captures local patterns within a genomic region, it does not consider interactions between regions. Since enhancers and promoters separated by thousands of base pairs can interact to regulate gene transcription, DeepCORE uses bi-directional long short-term memory (BiLSTM) networks [29] to capture short-range and long-range dependencies.

As the input sequences to the BiLSTM get longer, it becomes more difficult for the hidden states to capture the context, leading to decreased performance [30,31]. Furthermore, epigenetic signals are abundantly distributed throughout the human genome, but not all epigenetically modified regions have regulatory functions. To pinpoint the most functionally important elements and capture their local and distal interactions, DeepCORE employs an encoder-decoder [32] with attention mechanism [33]. The encoder is the BiLSTM model, and the decoder predicts the importance score of the next genomic region based on importance scores of the regions it has already predicted. This allows the prediction to be made based on a series of important hidden states from the encoder instead of only the last state. Furthermore, DeepCORE highlights the most informative regions contributing to gene transcription regulation by replacing the default softmax function in the attention model with a sparsemax function [34] that introduces sparsity of probability distribution. The learnt attention is a vector of length equal to the number of output nodes from the CNN layer containing importance score of each genomic region. DeepCORE then gives the decoder outputs to a fully connected FNN to predict continuous gene transcription levels.

2.3. Training DNN

We trained a multi-view attention-based DNN model for each sample. Given a sample, the data were randomly split into disjoint training, validation, and test sets, each comprised of 80 %, 10 %, 10 % of all genes, respectively. The test set was kept hidden and was used only after hyperparameter tuning and parameter learning were completed to avoid information leak. Mean Squared Error (MSE) was computed as the loss function and fed back to the network through backpropagation. We used Adam (Adaptive Moment Estimation) optimizer [35] to train the model for 100 epochs. Early stopping criteria (training is stopped if the model performance on the validation set does not improve over 5 epochs) is employed to avoid overfitting. We noted that the performance of all models stabilized before reaching 50 epochs, after which the training was terminated.

The optimization was performed in two stages., In the first stage, the hyperparameters in the DNN model were optimized via grid search (Supplementary Table 2). The optimal configuration was selected based on the performance on the validation set. The second stage of optimization is done on the attention mechanism to achieve sparsity. The parameters in the DNN model for both stages were jointly learned. The model was trained on Titan Xp GPU donated by the NVIDIA Corporation. The total runtime was recorded by varying the sequence length from 500 bps to 10kbps (Supplementary Fig. 1).

2.4. Interpreting attention matrices to discover COREs

The DeepCORE model trained on each sample contains an attention probability matrix with rows corresponding to genes and columns corresponding to 50 bps windows (bins). For each gene, we extracted the tissue-specific attention probabilities of the bins and computed the cumulative distribution (CDF) of the attention probabilities. We then calculated empirical p-values based on the CDF and applied correction of multiple comparison to derive the false discover rate (FDR). Bins with FDR< 0.05 indicated genomic regions with significant regulatory function.

After extracting the significant bins for each gene across all tissues, we obtained a matrix with rows representing different tissue or cell line samples and columns representing the bin probabilities. Pearson's pairwise correlation [36] was then applied to this matrix to estimate correlations between bins to infer interactions of different genomic regions. Blocks of bins that have significantly correlated attention probabilities and are at least 1kbps apart are putative COREs, i.e., regulatory elements that co-operatively modulate gene transcription.

3. Results

3.1. DeepCORE accurately predicts within- and cross- sample gene transcription levels

We first used two samples of different types: E061 (melanocyte cells) and E071(brain hippocampus middle) to assess the performance of DeepCORE models with various DNN architectures and the performance of two existing methods (ExPecto and DeepChrome). The DeepCORE DNN hyperparameters selected via grid search are K= 50, N_{CNN}= 50, p = 50, N_{LSTM}= 15, N_{ATTN}= 25, and N_{FNN}= 100. With this setting, genetic sequences and epigenetic signals in each 50 bps window are convolved separately. BiLSTM with attention layer produces 200 bins (50 bps long), each receiving attention probabilities before being fed to the FNN.

We initialized all the trainable weights by 0 to avoid randomness in the model's performance. To avoid any other randomness in the performance, we repeated the data splitting step and the model training step 20 times. Evaluated on the held-out test sets from these 20 repeats, the multi-view DNN reported the lowest error rate in both samples (mean and standard deviation of $RMSE_{E061} = 1.80 \pm 0.014$ and $RMSE_{E071} = 1.79 \pm 0.015$) compared to the two baseline single-view models (genetic-view model: $RMSE_{E061} = 2.68 \pm 0.004$ and $RMSE_{E071} = 2.31 \pm 0.002$, and epigenetic-view model: ($RMSE_{E061} =$

1.87 ± 0.015 and $RMSE_{E071} = 1.85 \pm 0.055$, Fig. 2A). It also produced lower error rates than ExPecto ($RMSE_{E061} = 2.86 \pm 0.034$ and $RMSE_{E071} = 3.02 \pm 0.019$). As DeepChrome only makes binary classifications, we converted the continuous prediction scores from DeepCORE and ExPecto using a median cutoff. Again, DeepCORE reported the highest accuracy ($F1_{E061} = 0.845 \pm 0.11$ and $F1_{E071} = 0.749 \pm 0.089$) compared to the genetic-view model ($F1_{E061} = 0.69 \pm 0.37$ and $F1_{E071} = 0.627 \pm 0.062$) epigenetic-view model ($F1_{E061} = 0.834 \pm 0.25$ and $F1_{E071} = 0.734 \pm 0.084$), ExPecto ($F1_{E061} = 0.82 \pm 0.057$ and $F1_{E071} = 0.753 \pm 0.3$), and DeepChrome ($F1_{E061} = 0.706 \pm 0.017$ and $F1_{E071} = 0.69 \pm 0.023$, Fig. 2B). Because multi-view models consistently outperformed single-view models in these two samples, we built only multi-view models in the subsequent analyses.

For each of the 25 cell-line and tissue samples, we built a DeepCORE multi-view model. As expected, samples of similar origins shared similar epigenetic profiles and those of distinct origins showed different epigenetic profiles (Supplementary Fig. 2). Using these samples, we evaluated the performance of DeepCORE, ExPecto, and DeepChrome in predicting gene expression levels. Across all sample, DeepCORE consistently reported a lower error rate (RMSE) than ExPecto on predicting continuous gene expression levels (Fig. 2C). The best

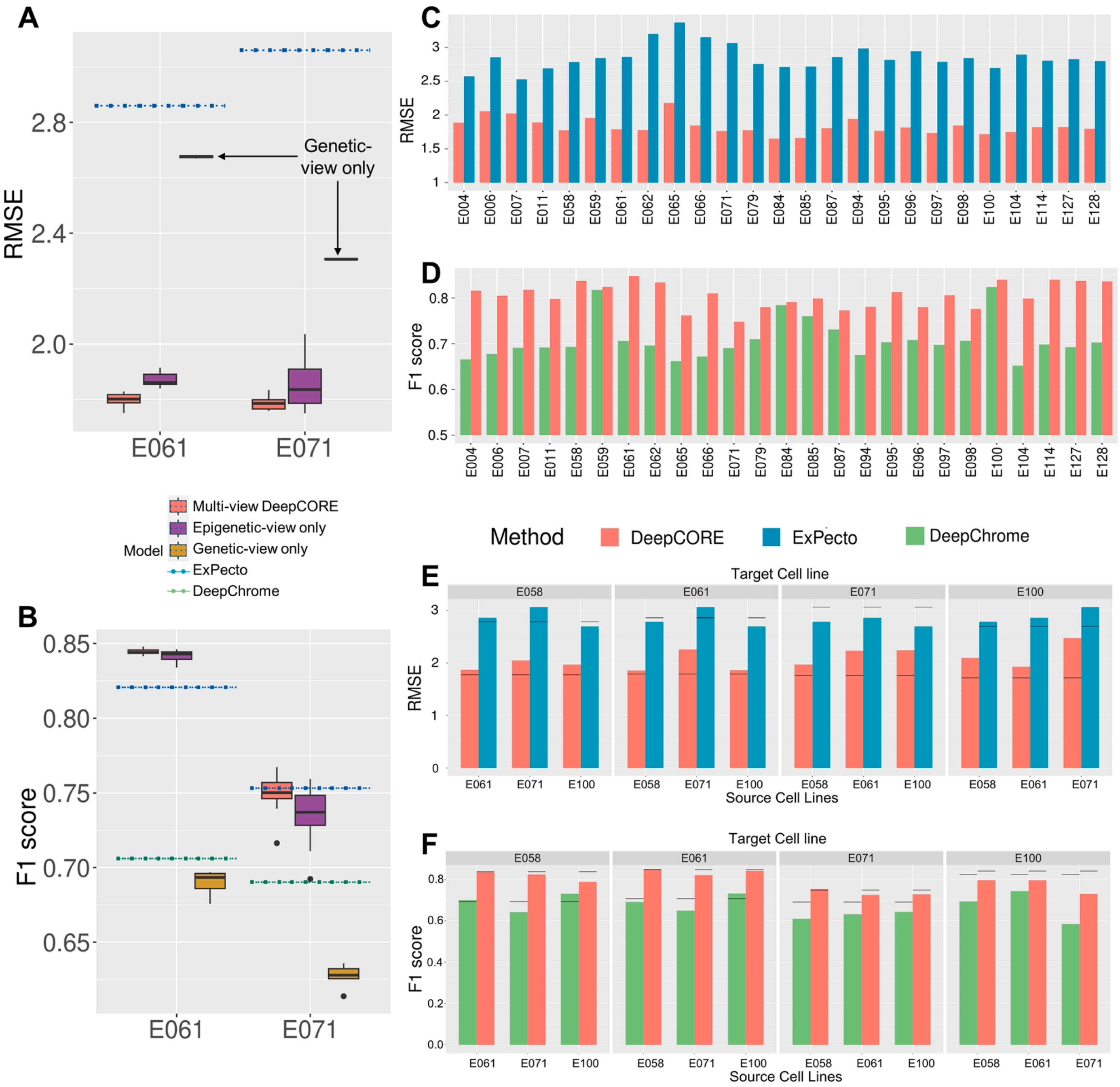


Fig. 2. Performance of DeepCORE and other methods. (A, B) Evaluated on two samples, E061 and E071, the boxplots of RMSE (A) and F1-score (B) show DeepCORE has the lowest error rate and the highest accuracy in predicting gene transcription levels, as compared to single-view DNN, ExPecto, and DeepChrome models. (C, D) Evaluated on 25 samples, DeepCORE has consistently the lowest error rate of predicting continuous gene transcription levels (C) and consistently the highest accuracy of predicting binary gene transcription classes (D). (E, F) Evaluated on cross-sample predictions in which a model trained the source sample is applied to predict gene transcription in different target samples, DeepCORE shows consistently lower error rate than ExPecto (E) and higher accuracy than DeepChrome (F). (b) Gray lines denote performance in source samples.

performance of DeepCORE was observed in the E084 sample with an RMSE of 1.65, and the lowest performance was observed in the E006 sample with an RMSE of 2.06. Similarly, DeepCORE consistently reported a higher accuracy than DeepChrome on binary classification (Fig. 2D). On average DeepCORE outperformed ExPecto and DeepChrome with an improvement of over 10 % in most samples.

Recently, another method, Xpresso [26], reported good performance of predicting gene expression levels based solely on genetic sequence. It provided sample-specific predictions for two ENCODE samples, namely E116 and E123. We trained DeepCORE on each of the two samples. Given that the gene expression values used in Xpresso and DeepCORE were in different scales, their RMSEs were not comparable. We therefore evaluated the performance based on r [2] value, Pearson correlation coefficient (PCC), and Spearman correlation coefficient (SCC) that measure the correlation between true and predicted gene expression. For both samples, DeepCORE outperformed Xpresso ($r^2 = 0.67$ vs. 0.46, PCC=0.82 vs. 0.68, SCC=0.79 vs. 0.68 for E116 sample, and $r^2 = 0.71$ vs. 0.52, PCC=0.84 vs. 0.72, SCC = 0.81 vs. 0.72 for E123 sample, Supplementary Fig. 3).

Following the success of DeepCORE on predicting within-sample gene transcription, we tested if a DNN model trained on one sample performed well on other samples, which indicate the generalizability of the model. We chose four tissues representing very different cell types (E058: keratinocyte, E061: melanocyte, E071: brain hippocampus middle, and E100: psoas muscle). In this analysis, we trained a DNN model using data from one sample and tested it using data from the remaining three samples. We first compared DeepCORE and ExPecto on predicting continuous values of gene expression levels. The results from our analysis indicated that in general, cross-sample predictions had lower performance than within-sample predictions for all three methods (Fig. 2E). The only exceptions were ExPecto models trained for E071 sample, where cross-sample prediction is better than within-sample prediction. Nevertheless, the RMSE error rate of DeepCORE on average was 27.5 % lower than that of ExPecto in cross-sample predictions (mean RMSE=2.06 vs. 2.85). We then compared DeepCORE and DeepChrome on binary classification (Fig. 2F). The F1-score was 18 % higher in DeepCORE than in DeepChrome (mean f1-score =0.79 vs. 0.671). Overall, the performance of DeepCORE decreased only slightly by 6 % for cross-sample predictions, while the ExPecto and DeepChrome showed a huge performance reduction by more than 15 % and 10 % respectively. These results implied that the patterns captured by DeepCORE likely represented general relationships between genetic, epigenetic, and transcriptional changes.

Using the E061 cell line, we experimented with training a model without BiLSTM. This reduced model reported a lower predictive accuracy than the full model that contained BiLSTM (RMSE = 3.4 vs. 1.8). The attended bins in the reduced model were closer to TSS compared to the full model (mean distance = 1835 bps vs. 3063 bps). These results demonstrated that BiLSTM detected long-range dependencies which in turn helped improve the prediction accuracy.

We also tested pooling data from multiple samples, which gave rise to models with higher errors. For example, we trained a model using pooled data from the E096 lung sample and the E071 brain sample. The RMSE of this model was 2.09 based on cross-validation. Conversely, the models trained on each of these two samples separately reported RMSE of 1.81 and 1.76, respectively.

Large language models based on the transformer architecture have reported unprecedent successes [37–39]. We also experimented with building a transformer model to predict gene expression in the E061 sample. Unfortunately, this model exhibited suboptimal performance compared to DeepCORE (RMSE=2.4 vs. 1.8, F1-score=0.74 vs. 0.85, Supplementary Fig. 4). A potential reason might be insufficient training data that fell several orders of magnitude below the scale used in large language models.

3.2. DeepCORE identifies regions with biologically meaningful histone markers

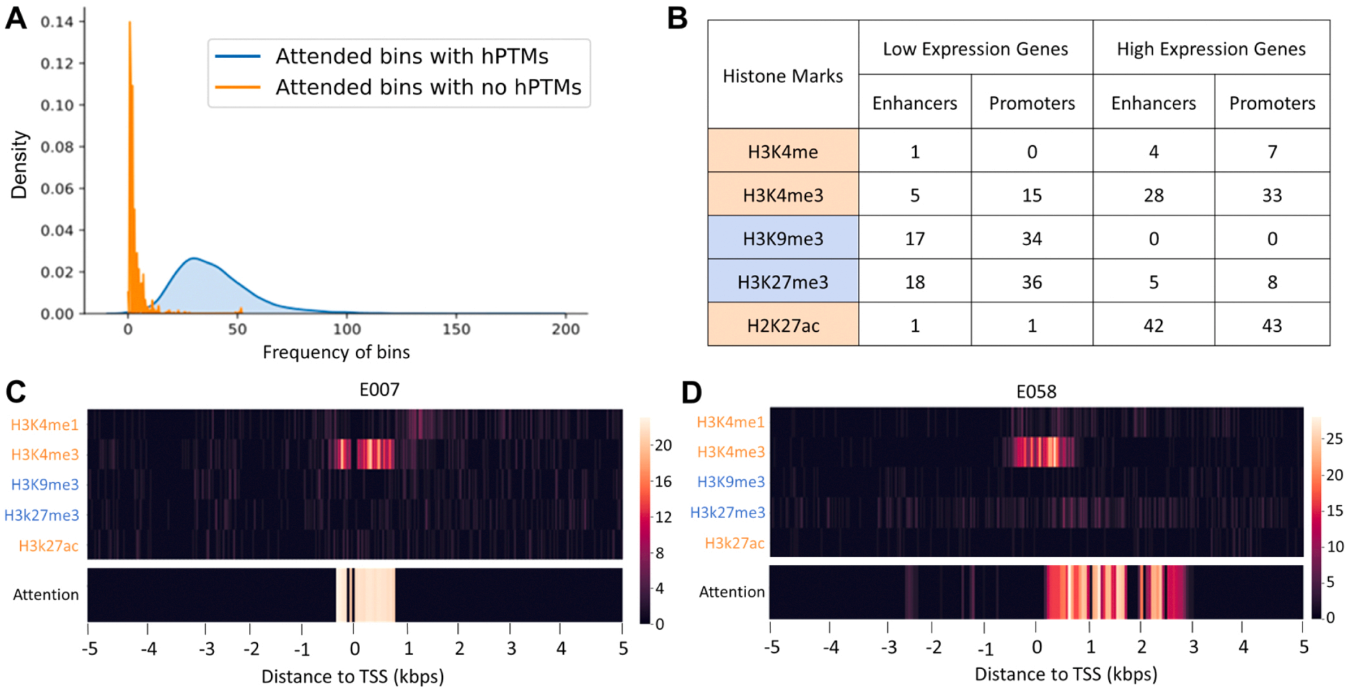
Attended bins receiving non-zero attention probabilities corresponded to genomic regions that contributed to the prediction of gene expression values. We found that hPTMs were present in most attended bins. Using genes from the held-out test set, we found that each gene on average had 33 attended bins containing hPTMs, but only 0.42 attended bins containing no hPTMs (Mann-Whitney test p-value = 9.1×10^{-25}, Fig. 3A). We then randomly selected 25 test genes that were transcribed above the median cutoff value and 25 genes transcribed below the median cutoff in the E071 sample. We extracted bins with highly significant attentions and counted the presence of hPTMs in the corresponding regions (Fig. 3B). Among genes with high transcription level, the attended genomic regions were enriched with H3K4me3 and H3K27ac that are known marks of active promoters and enhancers to enhance transcription [40,41]. Conversely, the enrichment of H3K9me3 and H3K27me3 in the attended regions near low-transcription genes were consistent with their known roles in formation of heterochromatins to repress transcription [42].

Further analysis of the attended regions of the *CYFIP2* gene in two samples revealed interesting patterns. In the E007 sample where *CYFIP2* gene was highly expressed, DeepCORE paid attention to regions that were close to the TSS and were occupied with the active histone mark H3K4me3 (Fig. 3C). In contrast, in the E058 sample where this gene was lowly expressed, DeepCORE paid attention to regions that were downstream of the TSS and were occupied with the repressor histone mark H3K27me3 marker and avoided regions with the activator histone mark H3K4me3 around the TSS (Fig. 3D). These results provided evidence that DeepCORE selects regions that are biologically relevant and reflect the underlying mechanisms of transcription regulation.

As no model can learn and explain all the features, we examined false positive attentions in the DeepCORE model trained on the E065 sample. Out of a total of 597,094 bins that contained no epigenetic signals, only 3291 received attention, indicating a very low false positive rate of 0.6 %. Our examination of false positive attentions revealed two distinct types of occurrences. In the first type, epigenetic signals were abundant in the ± 5kbps TSS-flanking regions, and bins receiving false positive attentions were next to bins receiving true positive attentions (Supplementary Fig. 5A). This phenomenon can be attributed to the CNN layer, which convolves epigenetic signals across positions, causing the spread of signals between neighboring bins. Addressing this issue might involve reducing the filter size and increasing the stride size of the CNN layer and increasing the interval of max-pooling. However, such changes will impact the bin length and subsequently the resolution of predicted. In the second type, epigenetic signals were scarce in the ± 5kbps TSS-flanking region, and bins at the leftmost or rightmost borders received false positive attentions (Supplementary Fig. 5B). This occurrence can be attributed to the BiLSTM layer, which carries epigenetic signals over an extended distance, leading to accumulation at the two ends. Potential solutions to this issue may include increasing the forget bias or the dropout rate in the BiLSTM layer. Nevertheless, considering the already very low false positive rate and the possibility that adjustments to these parameters may compromise performance, we have chosen to retain the original configuration of the DeepCORE models.

3.3. DeepCORE can identify and fine map regulatory elements

The Eukaryotic Promoter Database (EPD) [43] contains a comprehensive list of 29,598 experimentally validated human promoters. The GeneHancer [44] database annotated 250,512 candidate enhancers in the human genome. We then scanned our attended regions to identify the presence of these known promoters or enhancers. To match the attended regions with the promoters, we restricted the attended regions to be within 1kbps around the TSS. No such restrictions were applied for matching enhancers.



Histone Marks	Low Expression Genes		High Expression Genes	
	Enhancers	Promoters	Enhancers	Promoters
H3K4me	1	0	4	7
H3K4me3	5	15	28	33
H3K9me3	17	34	0	0
H3K27me3	18	36	5	8
H2K27ac	1	1	42	43

Fig. 3. Distribution of hPTMs in attended regions. (A) Density plots show distribution of attended bins with hPTMs vs. distribution of attended bins with no hPTMs. (B) Counts of bins with specific hPTMs in attended regions. Data were from randomly selected 25 highly expressed genes and 25 low expression genes. Transcription activating hPTMs are in orange background. Transcription repressing hPTMs are in blue background. (C, D) Heatmaps show the raw hPTM read counts and DeepCORE attention probabilities for the *CYFIP2* gene. Transcription of this gene was low in the E007 sample (C) and high in the E058 sample (D). The ± 5kbps TSS-flanking region is encoded into 200 bins each with an attention probability.

We hypothesize that the attended regions identified by DeepCORE were enriched with known REs. To test this hypothesis, we considered promoters and enhancers annotated in the EPD and GeneHancer databases as known REs, although many of these annotations were not experimentally validated. We then calculated the frequency of the attended regions containing known REs across all samples and the

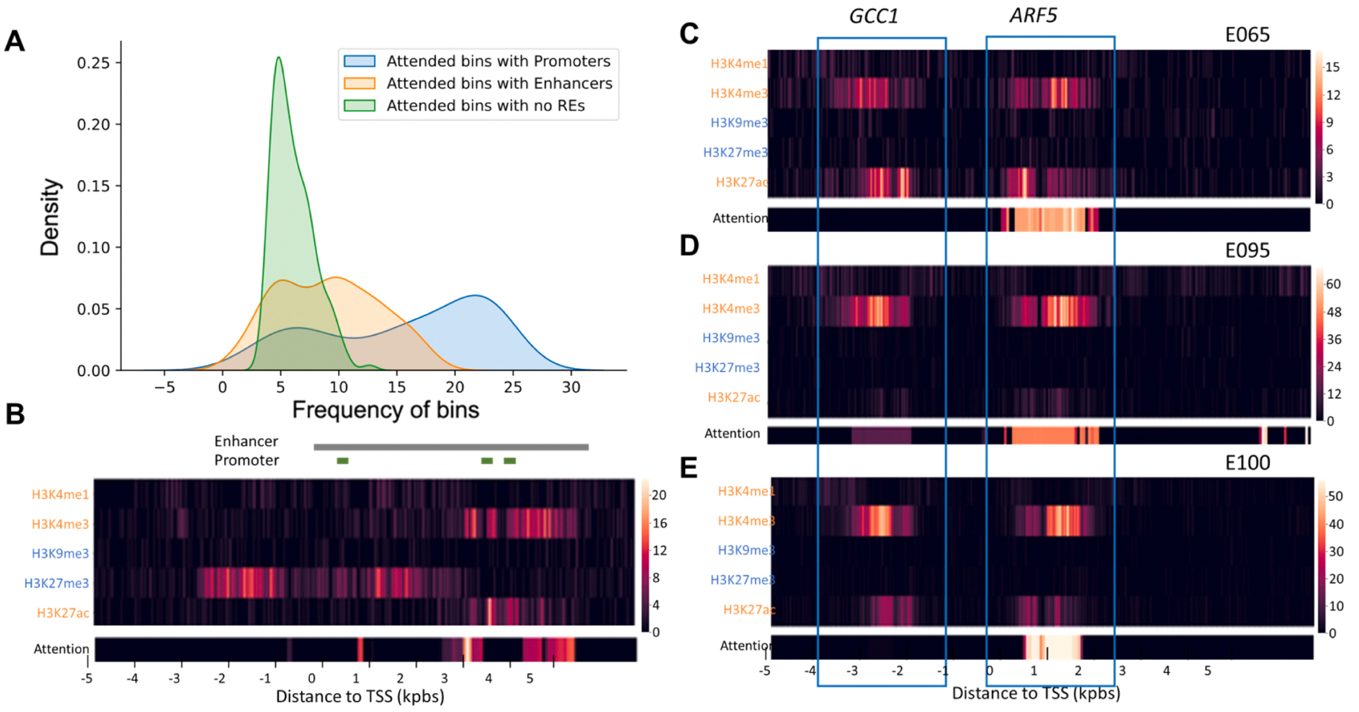


Fig. 4. Attention analysis for regulatory elements (A) Density plot of the frequency of attended bins with known promoters or enhancers across 25 samples in comparison to random bins with high attention scores. (B) In the *TMEM88* gene, attended bins matched to known enhancers and promoters. Signals from repressing hPTMs did not receive attention. (C-E) In the *ARF5* gene, hPTM signals form two clusters (indicated with blue boxes). The right cluster mapped to the promoter of the *ARF5* received attentions. The left cluster mapped to the promoter of another gene *GCC1* did not receive attention.

frequency of the remaining regions. On average, each gene had 23 attended bins containing known promoters and 10 attended bins containing known enhancers, but only 5 attended bins containing no known REs (Mann-Whitney test p-value = 4.0 $\times 10^{-25}$ and 6.1 $\times 10^{-18}$ respectively, Fig. 4A).

The *TMEM88* gene is a representative example in which the attended bins matched known promoters and enhancers. *TMEM88* was highly expressed in the E004 sample. The ± 5kbps TSS-flanking region was occupied with various active and repressive hPTMs. The EPD database annotated three promoters for this gene, one immediately upstream of the TSS and the other two towards the downstream. The enhancer annotated in the GeneHancer database spans a wide range starting at 1200 bps upstream of the TSS to 4000 bps downstream of the TSS. These REs all matched to the attended bins identified by DeepCORE (Fig. 4B). Furthermore, despite the repressive hPTM (H3K27me3) had high read counts, the DeepCORE model did not pay attention to it. Instead, activating hPTMs received attention, which was consistent with the high transcription level of *TMEM88* in this sample. Interestingly, inside the annotated enhancer region that spans more than 5kbps, only 30 bins covering 1500 bps received attention. Because only attended bins were used to predict gene transcription level, they likely were more relevant to transcription regulation than the unattended bins.

Another interesting example is the *ARF5* gene. Signals from hPTMs in three samples (E095, E065, and E100) consistently highlighted two regions (Fig. 4C–E). The right region corresponded to the promoter of this gene and received DeepCORE attention. The left region was 2500 bps upstream of the TSS and corresponded to the promoter of another gene *GCC1*. DeepCORE correctly identified the histone signals corresponding to *ARF5* gene and did not pay attention to the left peak. These results demonstrate that DeepCORE can identify and fine-map REs at a resolution of 50 bps that corresponds to the bin size of the model.

3.4. Concordant attentions identify putative COREs

The interpreter of DeepCORE includes correlation analysis of attention probabilities across samples to discover COREs. As an example, we examined the *PSMD8* gene that was consistently and highly expressed across all samples. We retrieved the attention vectors of this gene from 25 samples and calculated their pair-wise correlations (Fig. 5A). At FDR rate < 0.05, we found two blocks for which DeepCORE attentions were highly correlated (Fig. 5B). The first block is centered around the TSS and the second block is 3kbps downstream of the TSS. These two blocks received concordant attention across samples, implying that they jointly regulate transcription of the *PSMD8* gene. Indeed, these two blocks corresponded to the promoter and the enhancer of this gene.

We validated these COREs using established annotations and experimental data. As an illustrative example, we investigated the *TMUB1* gene and analyzed its attention vectors across all samples. While the EPD and GeneHancer annotations hinted at the presence of promoters and other cis-regulatory elements in this region, they offered limited resolutions (Fig. 6). DeepCORE attentions revealed three distinct blocks that finely mapped the REs in this region. The first block was located approximately 2.5kbps upstream of the TSS, the second block encompassed the TSS, and the third block was located approximately 2.5kbps downstream of the TSS. The correlated attentions observed among these three blocks strongly suggest their coordinated regulation

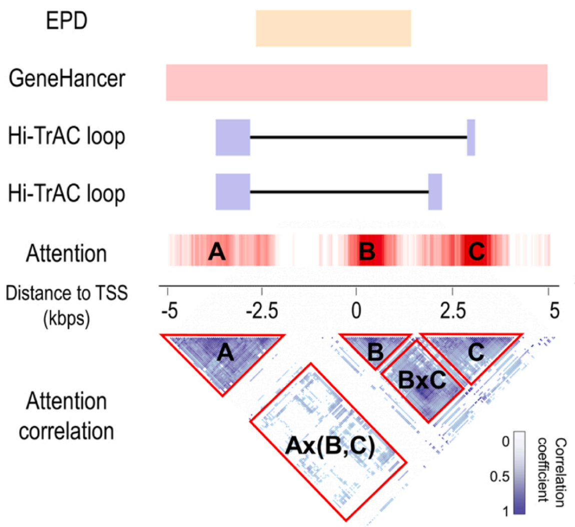


Fig. 6. CORE in the *TMUB1* Gene. The ± 5 kb TSS-flanking region is displayed, which has cis-regulatory roles as annotated in the EPD and GeneHancer databases. DeepCORE identified three blocks (A, B, and C) as putative REs, which received attentions in multiple samples. The correlation matrix of attentions revealed local interaction between B and C, and distal interaction between A and the other two elements. The distal interaction is confirmed in an Hi-TrAC study showing these REs are inside chromatin-chromatin loops.

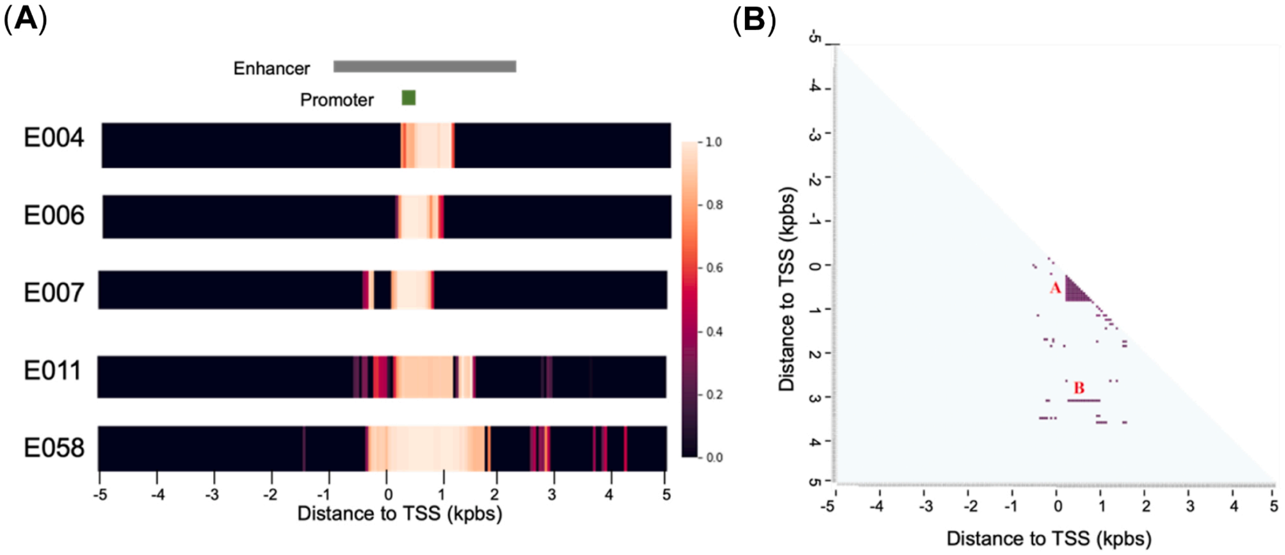


Fig. 5. COREs in the *PSMD8* gene: (**A**) Heatmaps show attentions in 5 cell lines. (**B**) Correlation plot shows two blocks (A and B) with significant correlated attentions.

of the gene's transcription and function as COREs. Furthermore, it is plausible that the distal interactions between the first block and the other two blocks occur via chromatin-chromatin loops, a phenomenon confirmed in a Hi-TrAC study [45].

4. Discussion

Multiple REs interact to regulate gene transcription. We designed the DeepCORE architecture to consider such inter-dependences in multiple aspects. In the input step, it uses two views to capture genetic and epigenetic features. In the DNN modeling step, it uses BiLSTM to allow short-range and long-range interactions, In the interpretation step, it detects correlated attentions between genomic regions. By training the DNN to predict gene transcription levels based on genetic and epigenetic features within the ± 5kbps TSS-flanking region, DeepCORE learns the most informative regions that are relevant to gene transcription regulation.

We evaluated the performance of DeepCORE and other methods on predicting gene transcription in diverse tissues and cell lines, although the assessment did not include all existing models such as Enformer [46] and Borzoi [39] for practical reasons. In these comparisons, DeepCORE was consistently the top performer. The high accuracies support that the DNN model captures informative features relevant to transcriptional regulation. It builds the foundation for subsequent analysis to further interpret the results, specifically attentions paid to each genomic region, to help mapping promoters, enhancers, and other REs. We further introduce COREs that are REs receiving concordant attentions across multiple samples.

DeepCORE uses only five hPTMs as epigenetic features. However, many other types of epigenetic signals, such as DNA methylation and transcription factor binding, provide complementary information to hPTM. Including these additional features may further increase the prediction accuracy and enhance the RE identification. Currently, DeepCORE examines ± 5kbps TSS-flanking region where promoters and proximal enhancers reside. Expanding the range to 2000kbps will allow us to detect distal REs. Furthermore, as enhancers are often clustered and selective activation of different enhancers in the same cluster is tissue-specific [47–49], concurrent modeling of multiple tissues is promising to capture the boundary between these enhancers and subsequently increase the resolution. This will also identify tissue-specific gene-promoter and gene-enhancer interactions, which is valuable knowledge that has not been annotated in existing databases.

In silico mutagenesis can be combined with gene expression prediction models to identify functional elements. For example, ExPecto [20] supports in silico mutagenesis by introducing DNA alterations into a genomic position, predicting the expression level of the target gene with and without the DNA alterations, and comparing the difference. Assuming mutations disrupting a regulatory element will be predicted with significant impact on gene expression, this approach can help identify regulatory elements. However, in silico mutagenesis usually tests simple alterations, such as single nucleotide variants, most of which have neutral or nearly neutral functional impact. Identification of regulatory elements based on these predictions may lead to many false negatives. Furthermore, DeepCORE is designed based on the rationale that epigenetic alterations can lead to gene expression changes with or without genetic alterations. However, in silico mutagenesis does not perform epigenetic alterations. The epigenetic features in our model include five quantitatively measured histone modification marks, which vary in intensity and cover different lengths of genomic regions. Epigenetic changes may involve alterations in one or several histone modification marks, entail varying extent of intensity changes, and affect different lengths of genomic regions. Given such high variability of potential epigenetic changes, it is challenging to simulate them computationally. The DeepCORE algorithm can identify regulatory elements without performing in silico genetic or epigenetic alterations, which is complementary to existing methods. In the future, we will explore if DeepCORE models combined with in silico mutagenesis can improve fine mapping of regulatory elements.

In summary, DeepCORE is a novel method to catalog cis-acting REs and COREs that influence gene transcription in tissue and cell line specific context. This knowledge can be used to discover novel REs and prioritize existing REs, which will help improve our understanding of transcription regulatory mechanisms. To facilitate evaluation and further analysis, we created an interactive web server at https://liliulab.shinyapps.io/deepcore/, which allows users to query, view, and download of DeepCORE predictions and attended bins in 27 human tissue and cell line sample. The DeepCORE source code is available at the GitHub site https://github.com/liliulab/DeepCORE.

Author statement

All authors approved the manuscript contents and affirmed that this work is not under consideration or published elsewhere. We have no other submission that contains overlapping content with this manuscript. We claim no conflict of interest.

CRediT authorship contribution statement

L.L. conceived the study. L.L. and P.B.C. designed the methodology and experiments. P.B.C., H.C curated data required for the analysis and implemented the software. P.B.C., H.C., N.A., and M.L. performed analysis and visualization. P.B.C. and H.C. implemented the software. P.B.C and L.L wrote and edited the manuscript. All authors read and approved the final manuscript.

Declaration of Competing Interest

All authors claim no conflict of interest.

Data and code availability

The Histone Modification signals can be downloaded from https://egg2.wustl.edu/roadmap/data/byFileType/peaks/consolidated/narrowPeak/ and the gene expressions can be downloaded from https://egg2.wustl.edu/roadmap/data/byDataType/rna/expression/. All code was implemented in Python (version 2.7.16) using TensorFlow (version 1.13.1), and the source code with demo is publicly available at https://github.com/liliulab/DeepCORE.

Acknowledgments

The authors wish to thank Dr. Jingmin Shu in Li Liu's lab for insightful discussions on this topic.

Financial disclosure statement

This work was supported by the National Institutes of Health, United States [grant number R01LM013438 and T32DK137525].

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.12.044.

References

[1] Blackwood EM, Kadonaga JT. Going the distance: a current view of enhancer action. Science 1998;281:60–3.

[2] Krivega I, Dean A. Enhancer and promoter interactions—long distance calls. Curr Opin Genet Dev 2012;22:79–85.

[3] Appella E, Anderson CW. Post-translational modifications and activation of p53 by genotoxic stresses. Eur J Biochem 2001;268:2764–72.

[4] Carey M, Peterson CL, Smale ST. Transcriptional regulation in eukaryotes. Cold Spring Harbor Laboratory Press,; 2009.

[5] Gilad Y, Rifkin SA, Pritchard JK. Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet 2008;24:408–15.

[6] Li G, Jima D, Wright FA, Nobel AB. HT-eQTL: integrative expression quantitative trait loci analysis in a large number of human tissues. BMC Bioinforma 2018;19:95.

[7] Hobert O. Gene regulation by transcription factors and microRNAs. Science 2008;319:1785–6.

[8] Wittkopp PJ, Kalay G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat Rev Genet 2012;13:59–69.

[9] Felsenfeld G, Boyes J, Chung J, Clark D, Studitsky V. Chromatin structure and gene expression. Proc Natl Acad Sci USA 1996;93:9384–8.

[10] Stavreva DA, Hager GL. Chromatin structure and gene regulation: a dynamic view of enhancer function. Nucleus 2016;6:442–8.

[11] Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. Nat Rev Genet 2019;20:207–20.

[12] Roth FP, Hughes JD, Estep PW, Church GM, Finding DNA. regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat Biotechnol 1998;16:939–45.

[13] Bräzma A, Jonassen I, Vilo J, Ukkonen E. Predicting gene regulatory elements in silico on a genomic scale. Genome Res 1998;8:1202–15.

[14] Liu X, Brutlag DL, Liu JS. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. In: Biocomputing 2001. World Scientific; 2000. p. 127–38.

[15] Sinha S, Tompa M. A statistical method for finding transcription factor binding sites. ISMB 2000;8:344–54.

[16] Beer MA, Tavazoie S. Predicting gene expression from sequence. Cell 2004;117:185–98.

[17] Yuan Y, Guo L, Shen L, Liu JS. Predicting gene expression from sequence: a reexamination. PLoS Comput Biol 2007;3:e243.

[18] Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U. Predicting expression patterns from regulatory sequence in Drosophila segmentation. Nature 2008;451:535–40.

[19] Singh R, Lanchantin J, Robins G, Qi Y. DeepChrome: deep-learning for predicting gene expression from histone modifications. Bioinformatics 2016;32:i639–48.

[20] Zhou J, et al. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nat Genet 2018;50:1171–9.

[21] Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv Prepr arXiv 2013;1312:*6034*.

[22] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: European conference on computer vision. Springer; 2014. p. 818–33.

[23] Zintgraf LM, Cohen TS, Adel T, Welling M. Visualizing deep neural network decisions: Prediction difference analysis. arXiv Prepr arXiv 2017;1702:*04595*.

[24] Consortium EP. The ENCODE (ENCyclopedia of DNA elements) project. Science 2004;306:636–40.

[25] Bernstein BE, et al. The NIH roadmap epigenomics mapping consortium. Nat Biotechnol 2010;28:1045–8.

[26] Agarwal V, Shendure J. Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks. Cell Rep 2020;(31).

[27] Kundaje A, et al. Integrative analysis of 111 reference human epigenomes. Nature 2015;518:317–30.

[28] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15:1929–58.

[29] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9:1735–80.

[30] Cho K, Van Merriënboer B, Bahdanau D, Bengio Y. On the properties of neural machine translation: Encoder-decoder approaches. arXiv Prepr arXiv 2014;1409:*1259*.

[31] Koehn P, Knowles R. Six challenges for neural machine translation. In: Proceedings of the first workshop on neural machine translation. Association for Computational Linguistics; 2017. p. 28–39. https://doi.org/10.18653/v1/W17-3204.

[32] Cho K, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv Prepr arXiv:1406 1078 2014.

[33] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv Prepr arXiv 2014;1409:*0473*.

[34] Martins A, Astudillo R. From softmax to sparsemax: A sparse model of attention and multi-label classification. Int Conf Mach Learn 2016:1614–23.

[35] Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv Prepr arXiv 2014;1412:*6980*.

[36] Benesty J, Chen J, Huang Y, Cohen I. Pearson correlation coefficient. In: Noise reduction in speech processing. Springer; 2009. p. 1–4.

[37] Vaswani A, et al. Attention is all you need. In: Advances in neural information processing systems. Curran Associates, Inc; 2017.

[38] Brown T, et al. Language models are few-shot learners. In: Advances in neural information processing systems. Curran Associates, Inc; 2020. p. 1877–901.

[39] Linder, J., Srivastava, D., Yuan, H., Agarwal, V. & Kelley, D.R. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. 2023.08.30.555582 Preprint at https://doi.org/10.1101/2023.08.30.555582 (2023).

[40] Creyghton MP, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci 2010;107:21931–6.

[41] Beacon TH, et al. The dynamic broad epigenetic (H3K4me3, H3K27ac) domain as a mark of essential genes. Clin Epigenetics 2021;13:138.

[42] Kouzarides T. Chromatin modifications and their function. Cell 2007;128:693–705.

[43] Périer RC, Praz V, Junier T, Bonnard C, Bucher P. The eukaryotic promoter database (EPD). Nucleic Acids Res 2000;28:302–3.

[44] Fishilevich S, et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database 2017;2017.

[45] Hi-TrAC reveals division of labor of transcription factors in organizing chromatin loops | Nature Communications. 〈https://www.nature.com/articles/s41467–022-34276–8〉.

[46] Avsec Ž, et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods 2021;18:1196–203.

[47] Pennacchio LA, Loots GG, Nobrega MA, Ovcharenko I. Predicting tissue-specific enhancers in the human genome. Genome Res 2007;17:201–11.

[48] Ong C-T, Corces VG. Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat Rev Genet 2011;12:283–93.

[49] Xiong L, et al. Genome-wide identification and characterization of enhancers across 10 human tissues. Int J Biol Sci 2018;14:1321.