Incorporating Gene Expression in Genome-wide Prediction of Chromatin Accessibility via Deep Learning

Qiao Liu[1,2], Wing Hung Wong[2,*] and Rui Jiang[1,*]

[1] Department of Automation, Tsinghua University, Beijing 100084, China;
[2] Department of Statistics, Stanford University, Stanford, CA 94305, USA
whwong@stanford.edu, ruijiang@tsinghua.edu.cn

Abstract. Regulatory elements (REs) in human genome are major sites of non-coding transcription which lack adequate interpretation. Although computational approaches have been complementing high-throughput biological experiments towards the annotation of the human genome, it remains a big challenge to systematically and accurately characterize REs in the context of a specific cell type. To address this problem, we proposed DeepCAGE, an deep learning framework that incorporates transcriptome profile of human transcription factors (TFs) for accurately predicting the activities of cell type-specific REs. Our approach automatically learns the regulatory code of input DNA sequence incorporated with cell type-specific TFs expression. In a series of systematic comparison with existing methods, we show the superior performance of our model in not only the classification of accessible regions, but also the regression of DNase-seq signals. A typical scenario of usage for our method is to predict the activities of REs in novel cell types, especially where the chromatin accessibility data is not available. To sum up, our study provides a fascinating insight into disclosing complex regulatory mechanism by integrating transcriptome profile of human TFs.

Keyword: Gene expression, Chromatin accessibility, Deep learning

1.Introduction

Identifying functional sites in human genome can help us understand regulatory mechanisms underlying genetic signals that are statistically associated with diseases [1]. However, a majority of known genetic variants lie in noncoding regions which lack adequate interpretation, making it imperative to character genome-wide regulatory elements (REs) and accurately decipher their implications in a comprehensive manner [2, 3]. It has been argued that a genetic variant may result in the disruption of its hosting regulatory element (RE) and hence predispose to the occurrence of a disease [4]. Currently, our inability to precisely and efficiently identify the regulatory elements in human genome impedes progress towards precision medicine to a large extend.

Regulatory elements (REs) in human genome tent to be enriched in DNA accessible regions which often involve transcription factors (TFs), RNA polymerases and other cellular machines to regulate gene expression [5]. Recently, the exploration of regulatory landscape has been accelerated by the development of high-throughput technologies, such as DNase-seq [6], ChIP-seq [7] and ATAC-seq [8], making it possible for assaying accessible chromatin. So far, these technologies have only been applied to a small subset of all biological cell or tissue types. Towards this concern, researchers have proposed computational models for predicting various chromatin signals, such as DNA accessibility, transcription factor binding sites (TFBSs), histone markers, DNA methylation and chromatin interaction [9-14].

Here we investigate the problem of predicting chromatin accessibility by incorporating TFs gene expression data and deep learning technology. Deep learning models, especially deep neural networks, have achieved breakthrough in many fields, including speech recognition, visual object detection and natural language processing [15]. It inspires us to build DeepCAGE, a **deep** densely connected convolutional network for predicting **c**hromatin **a**ccessibility by incorporating **g**ene **e**xpression (Figure 1). DeepCAGE overcomes two major limitations of previous approaches. First, sequence-based methods, such as DeepSEA [16], DanQ [17] and Basset [18], are unable to make prediction of new cell types especially where the chromatin accessibility data is not available. Second, expression-based methods, such like BIRD [19], fail to utilize the comprehensive DNA sequence information, hence cannot make prediction of new loci that are not contained in the training dataset. DeepCAGE takes both the DNA sequence information and TFs gene expression data into consideration and adopts the architecture of densely connected convolutional neural network which has been experimentally proved to alleviate vanishing-gradient problem [20]. In a series of systematic evaluation, DeepCAGE not only achieves state-of-the-arts performance in the binary chromatin accessible status classification experiments, but also recovers continuous DNase-seq signal in regression settings. To make DeepCAGE more understandable, we proposed a strategy for visualizing the weights in the first convolutional layer. Interestingly, many known motifs were successfully recovered by DeepCAGE. We finally summarize DeepCAGE, as an effective and precise predictive model for dissecting regulatory code, which could shed light on the understanding the comprehensive regulatory mechanism.

MATERIAL AND METHODS

Overview of DeepCAGE model

DeepCAGE consists of a hybrid architecture which takes DNA sequence and gene expression data as inputs and predicts chromatin accessibility in a binary value (classification) or continuous signal (regression) (Figure 1). The input 1000 base pair (bp) DNA sequence is first extracted from the hg19 reference genome around the midpoint of a 200 bp bin and then converted to a one-hot matrix. The four columns of the matrix correspond to the four nucleotides A, C, G and T and each row represents a nucleotide in DNA sequence. For any position in DNA sequence with 'N', the corresponding row is filled with all zeros. After one-hot encoding, the DNA sequence then was fed to a densely connected deep convolutional neural network which contains three dense blocks. In each dense block, there are five layers and each layer connects to every

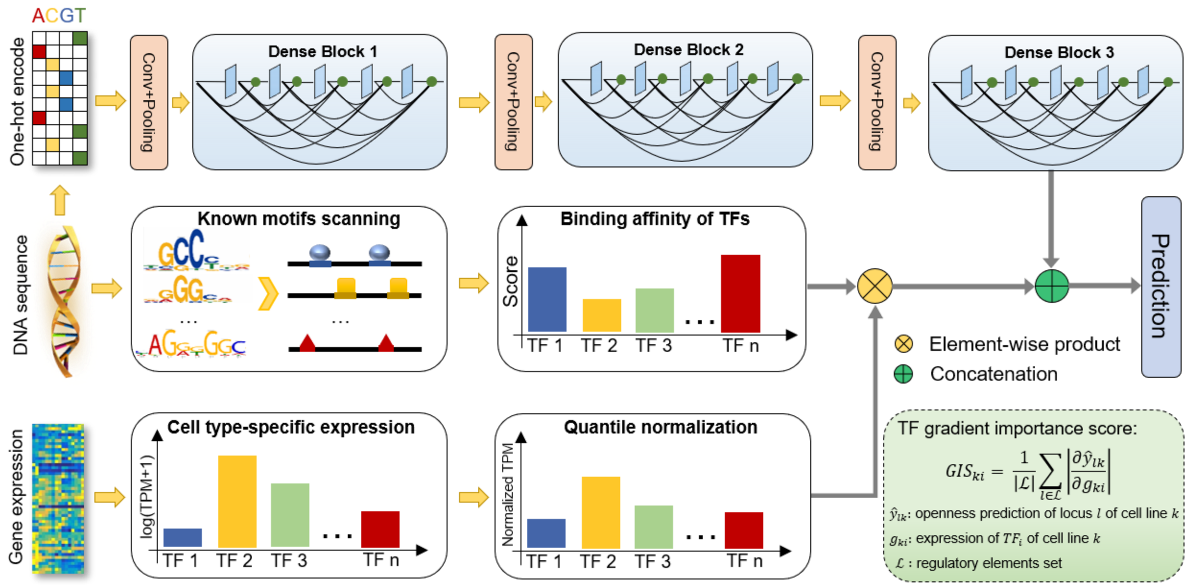

ACGT
One-hot encode
Conv+Pooling
Dense Block 1
Conv+Pooling
Dense Block 2
Conv+Pooling
Dense Block 3
DNA sequence
Known motifs scanning
...
...
Binding affinity of TFs
Score
TF 1
TF 2
TF 3
...
TF n
Prediction
Element-wise product
Concatenation
Gene expression
Cell type-specific expression
log(TPM+1)
TF 1
TF 2
TF 3
...
TF n
Quantile normalization
Normalized TPM
TF 1
TF 2
TF 3
...
TF n
TF gradient importance score:
$GIS_{ki} = \frac{1}{|\mathcal{L}|}\sum_{l\in\mathcal{L}}\left|\frac{\partial \hat{y}_{lk}}{\partial g_{ki}}\right|$
$\hat{y}_{lk}$: openness prediction of locus l of cell line k
g_{ki}: expression of TF_i of cell line k
$\mathcal{L}$: regulatory elements set

Figure 1. The schematic of DeepCAGE model. On the one hand, the input DNA sequence is first converted to a one-hot matrix and goes through a densely connected deep convolutional neural network (DenseNet) which contains three dense blocks. On the other hand, the input DNA sequence is scanned with 402 non-redundant motifs and the motif scores are extracted. The expressions of the corresponding 402 human transcription factors (TFs) are extracted and quantile normalized. We combine normalized expression and motif scores in an element-wise product manner and then concatenated by the output of the DenseNet. The prediction is made based on the hybrid features.

other layer in a feed-forward fashion. If traditional convolutional networks with have L connections then densely connected deep convolutional neural network will have L(L+1)/2 connections. Each layer in a dense block is constructed by two convolution operations with small filters size 1×1 and 3×1 of which 1×1 filters aim at reducing feature maps. Note that the first convolution operation is applied for reducing the concatenated channels to a fixed number. The dense block architecture has been proven to alleviate the vanishing-gradient problem and strengthen the feature propagation [20]. Between input and the first dense block, there is a transition module which contains a convolutional layer and max-pooling layer. The convolutional layer has 160 filters with size 4×15 for extracting low-level features and detecting DNA binding motifs while the max-pooling layer is applied for finding the most significant activation signal in a given sliding window of each filter. Between blocks, the similar transition module is utilized for extracting high-level features and dimension reduction. Note that the rectified linear units (ReLU) are used after each convolution operation for keeping positive activations and setting negative activation values to zeros. Batch normalization (https://arxiv.org/abs/1502.03167) and dropout strategy [21] are also applied after each ReLU function for reducing internal covariate shift and avoiding overfitting, respectively. Besides the neural network, the input DNA sequence is also scanned for searching potential transcription factors (TFs) binding sites with non-redundant motifs of 402 human core TFs from HOCOMOCO database [22] with HOMER motif finder tool from [23]. We then take the maximal score of all motif binding sites for each TF given a DNA sequence. For the gene expression input, only the gene expression levels (TPM values) of the above 402 human core TFs from a specific cell type are extracted. Next, the gene expression levels are log transformed after adding a pseudocount of 1 and then quantile normalized. Then we combine TFs gene expression data with motif score data in an element-wise product manner. Finally, we concatenate this feature to the output of the densely connected deep convolutional neural network, thus forming a hybrid feature. This hybrid feature will go through a fully connected layer with 512 hidden nodes and the final prediction is made by a sigmoid layer (see detailed hyperparameters in Supplementary Table 1). For DeepCAGE regression model, there are two major differences. First, the output layer directly applies a linear transformation as $\mathrm{Y} = W^T X$ instead of a sigmoid layer. Second, mean square error (MSE) is used as the loss function.

DNase-seq data processing

DNase-seq data across 55 different human cell types were downloaded from the ENCODE project [24] in narrowPeak and bam format for measuring chromatin accessibility (accession IDs are listed in Supplementary Data 1-2). The human hg19 reference genome was first divided into 200 base pair (bp) non-overlapping bins. Considering the fact that one cell type may contain multiple replicates, we denote each bin (locus) as positive if it overlaps with the peak regions from at least half of the replicates, and denote as negative otherwise (Supplementary Figure 1). Binary label y_{lk} for locus l in cell type k represents the accessible state which was used for training and testing in classification experiments. For regression experiments, multiple replicates bam files from the same cell type are first pooled together. The number of pooled reads falling into each bin was counted for each cell type. To further eliminate the effect of different sequencing depths, bin read counts for each cell type were first divided by the total count N_k, and then multiplied by a constant N ($N = min\{N_k\}$), which is the minimal read count across all cell types. After such a procedure, the raw pooled read count n_{lk} for bin l in cell type k was converted into a normalized read count $\tilde{n}_{lk} = Nn_{lk}/N_k$. The normalized read counts were further log transformed after adding a pseudocount of 1. The transformed data represent the level of DNase I hypersensitivity and were used for training and testing in the regression models.

RNA-seq data processing

RNA-seq data across the same 55 different human cell types were downloaded from the ENCODE project [24] in tsv format for quantifying gene expression level (accession IDs are listed in Supplementary Data 3). The expression levels of the 402 core human transcription factors (TF) were considered and extracted. After further log transformation and quantile normalization based on TPM values, the normalized expression within each cell type was averaged across multiple replicates and the mean expression profile of each cell type was finally used for training and testing the predictive models.

Training-test data partition

We applied a five-fold cross-validation for randomly splitting training and test data. The 55 cell types were partitioned into a training dataset with 44 cell types and a test dataset with 11 cell types in each fold (see detailed five-fold partition in Supplementary Table 2-3). Due to the fact that not all genomic loci are regulatory elements, we defined 'known loci' which are denoted as positive in at least two training cell types, and we further defined 'novel loci' which are denoted as positive in at least two test cell types and do not overlap with 'known loci' in the meanwhile. Take the 5^{th} fold for example, there are 1,435,610 known loci and 54,171 novel loci respectively. The number of training examples can be as large as about 63 million $(1{,}435{,}610 \times 44)$. We provided both single-GPU and multi-GPU model for implementing DeepCAGE. In multi-GPU settings, the backpropagation process (GPU) and generation of the next batch of training data (CPU) is paralleled to ensure computing efficiency.

Model evaluation

Predictive models are evaluated based on test dataset. We defined two types of metrics in two perspectives, namely cell-type-wise evaluation metrics and locus-wise evaluation metrics respectively (see Supplementary Figure 1). Let $\boldsymbol{Y}_{L\times K}$ and $\widehat{\boldsymbol{Y}}_{L\times K}$ be the true label matrix and predicted matrix where L and K denote the number of loci and test cell types, respectively. In the case of classification experiments, y_{lk} and $\hat{y}_{lk}$ denote the true and predicted chromatin accessible state of locus l in cell type k, respectively. Due to the extremely unbalanced datasets, we applied cell-type-wise auPR (area under Precision-Recall curve) and locus-wise auPR to evaluate the performance of classification models, where cell-type-wise auPR is calculated based on $\boldsymbol{y}_{*k} = (y_{1k}, y_{2k}, \dots, y_{Lk})$ and $\widehat{\boldsymbol{y}}_{*k} = (\hat{y}_{1k}, \hat{y}_{2k}, \dots, \hat{y}_{Lk})$, locus-wise auPR is calculated based on $\boldsymbol{y}_{l*} = (y_{l1}, y_{l2}, \dots, y_{lK})$ and $\widehat{\boldsymbol{y}}_{l*} = (\hat{y}_{l1}, \hat{y}_{l2}, \dots, \hat{y}_{lK})$. In the case of regression experiments, y_{lk} and $\hat{y}_{lk}$ denote the true and predicted processed Dnase-seq signal. We applied cell-type-wise Pearson's r and locus-wise Pearson's r for evaluating the performance of regression models. Cell-type-wise Pearson's r is calculated based on $\boldsymbol{y}_{*k}$ and $\widehat{\boldsymbol{y}}_{*k}$ while locus-wise Pearson's r is calculated based on $\boldsymbol{y}_{l*}$ and $\widehat{\boldsymbol{y}}_{l*}$. We further introduced prediction squared error (PSR) which is calculated by $\mathrm{PSR} = \sum_k \sum_l (y_{lk} - \hat{y}_{lk})^2 / \sum_k \sum_l (y_{lk} - \bar{y}_{*k})^2$. Note that $\bar{y}_{*k}$ is the mean of $\boldsymbol{y}_{*k}$ $(\bar{y}_{*k} = \sum_l y_{lk} / L)$ and PSR is a statistic which considers both cell-type-wise prediction and locus-wise prediction. Besides, we introduced two statistics, namely cell range and cell variability, to describe the activity of a locus based on the true DNase-seq signal across test cell types. Cell range of locus l is calculated by $\max(\boldsymbol{y}_{l*}) - \min(\boldsymbol{y}_{l*})$ while cell variability of locus l is defined by the standard deviation of $\boldsymbol{y}_{l*}$. The various metrics in both classification and regression experiments provide a comprehensive and systematical evaluation of predictive models.

Models comparison

Basset [18], DeepSEA [16] and DanQ [17] are three sequence-based neural network models which take DNA sequences as input. They cannot make cross-cell-type prediction without incorporating cell type specific information. BIRD [19] is an expression-based model which takes gene expression data as predictor variables. It can only estimate chromatin accessibility of pre-defined regions which may ignore potential novel regulatory elements (REs) that were not included in the pre-defined regions. ChromDragoNN [25] is a recently deep learning method which takes both DNA sequence and gene expression data as input. However, it ignores the motif information of each corresponding transcription factor (TF).

Gradient importance score

We proposed a strategy for prioritizing the importance of transcription factors (TFs) given a pair of cell type and a DNA locus. Typically, we first extended the locus to -100kb to 100kb from the midpoint. Then we calculated the average absolute gradient of predicted openness within above region with respect to TFs' expression. $GIS_{ki} = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} \left| \frac{\partial \hat{y}_{lk}}{\partial g_{ki}} \right|$. Here, $\hat{y}_{lk}$ denotes the predicted openness of locus l in cell type k. g_{ki} denotes the expression of TF i in cell type k. $\mathcal{L}$ is the regulatory elements set that contains all loci within the above region. Gradient importance scores (GIS) give an intuition of which transcription factors (TFs) can play an important role in a specific cell type.

Motif analysis

We converted the weights of the filters from the first convolutional layer into position weight matrices (PWMs) by counting subsequence occurrences in a set of input sequences that activate a filter to a threshold value. All subsequences with activation value that greater than the threshold of each filter were pooled together and aligned. Then the PWMs were composed by the frequencies of the 4 nucleotides (A, C, G and T) at each position. We regard a subsequence at position i as activated if $\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} w_{m,n}^{k} x_{i+m,n}^{j} > \alpha \cdot MAV^{k}$. $M \times N$ denotes the size of the filters which is 4×15 in the first convolutional layer. MAV^{k} denotes the maximal activation value of filter k which is represented by $MAV^{k} = \max_{i,j}(\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} w_{m,n}^{k} x_{i+m,n}^{j})$. α is the control coefficient which is set to 0.7 in the experiments. We identity motifs using the TomTom v4.12.0 tool [26] with the E-value threshold 0.01 with comparison to the known motifs in the JASPAR 2018 database [27]. Besides, we calculated the information content (IC) of recovered motifs based on the information entropy. $\mathrm{IC} = \sum_{i,j}(p_{ij}\log_2 p_{ij} - b_i \log_2 b_i)$ where p_{ij} is the element in PWM matrix and i, j denote the nucleotide type and position, respectively $(\mathrm{i} = 0,1,2,3)$. b_i denotes the background frequency of nucleotide i (default: $b_i = 0.25$).

RESULTS

DeepCAGE accurately predicts binary chromatin accessibiity status

We designed a series of experiments to systematically evaluate the performance of DeepCAGE in capturing genome accessibility code from the viewpoint of binary classification. DNase-seq and RNA-seq data across 55 cell types were downloaded from the ENCODE project [24]. The 55 cell types were partitioned into a training dataset with 44 cell types and a test dataset with 11 cell types using five-fold cross-validation. We then defined ‘known loci’ and ‘novel loci’ as potential regulatory elements which are determined by training data and test data respectively (see Methods). Predictive models were trained under ‘known loci’ of training

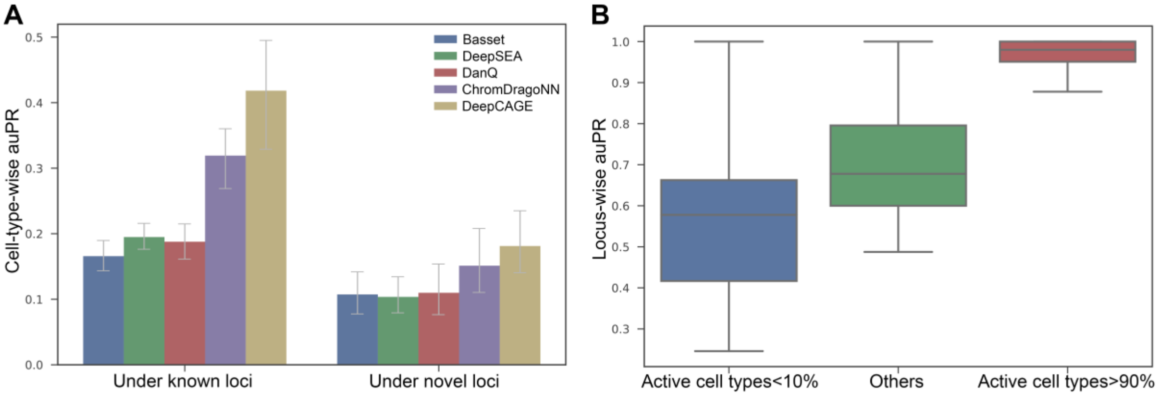


Figure 2. The performance of DeepCAGE classification model. (**A**) DeepCAGE achieves highest cell-type-wise auPR under both ‘known loci’ and ‘novel loci’ across test cell types compared to three baseline methods (Basset, DeepSEA, DanQ and ChromDragoNN). (**B**) The performance of DeepCAGE under loci with different activities across test cell types.

cell types and tested under both 'known loci' and 'novel loci' of test cell types. We compared DeepCAGE to four methods, including Basset [18], DeepSEA [16], DanQ [17] and ChromDragoNN [25]. First, we evaluated the performance of the predictive models under 'known loci' of test cell types. Due to the extremely unbalanced data with a proportion of positive loci from 2.6% to 29% across 55 cell types, these models were assessed in term of cell-type-wise auPR (area under Precision-Recall curve) (see Methods). DeepCAGE achieves the highest performance among all the baseline models with the mean cell-type-wise auPR of 0.418, compared to 0.166 of Basset, 0.195 of DeepSEA, 0.188 of DanQ and 0.319 of ChromDragoNN (Figure 2A). DeepCAGE outperforms baseline models, especially sequence-based models by a large margin as our model can discern the difference of openness landscape among test cell types while sequence-based models fail to capture the cell-type specific information. Second, we evaluated the predictive models under 'novel loci' of test cell types. It is a more difficult task as 'novel loci' are not included in the training process which can reflect the ability of cross-locus prediction. DeepCAGE also achieves the highest performance among all the baseline models with the mean cell-type-wise auPR of 0.181, compared to 0.107 of Basset, 0.104 of DeepSEA, 0.110 of DanQ and 0.151 of ChromDragoNN (Figure 2A). We then assessed the performance of DeepCAGE in term of locus-wise auPR (see Methods). Note that locus-wise auPR does not exist in sequence-based baseline methods as they cannot make any different prediction of a locus in test cell types. We divided the 'known loci' into three groups based on the percentage of test cell types in which a locus is denoted as positive. For active test cell types less than 10%, DeepCAGE achieves a mean locus-wise auPR of 0.578. The mean locus-wise auPR continues to increase as the regarding loci are active in more test cell types (Figure 2B). To sum up, DeepCAGE substantially achieves both high cell-type-wise and locus-wise prediction accuracy after incorporating with gene expression data of human core transcription factors (TFs).

DeepCAGE recovers continuous degree of chromatin accessibility

In the above classification experiments, we only consider the binary status of a locus in a specific cell type. However, binary label cannot reflect the difference of accessibility among sequences that share the same label. To address this problem, we further proposed DeepCAGE regression model (see Methods) which can predict the continuous degree of chromatin accessibility of a locus in a specific cell type. We define the degree of accessibility of a DNA sequence as the normalized average raw reads count that fall into the corresponding region (see Methods). We evaluate whether DeepCAGE can help recover continuous degree of chromatin accessibility using the same datasets. First, we compared DeepCAGE to two baseline methods, BIRD [19] and ChromDragoNN [25]. Regression models were assessed in term of cell-type-wise Pearson's r and prediction squared error (PSE) (see Methods). DeepCAGE achieves a mean cell-type-wise Pearson's r of 0.785, compared to BIRD of 0.637 and ChromDragoNN of 0.735 (Figure 3A). DeepCAGE achieves cell-type-wise Pearson's r larger than 0.85 in 18.2% of the test cell types and it even achieves cell-type-wise Pearson's r larger than 0.9 in two cell types (see examples in Figure 3B). DeepCAGE also achieves the minimal prediction square error (0.42) comparing to baseline methods (0.77 and 0.57) (Figure 3C). DeepCAGE again outperforms baseline models by a quite large margin.

To further explore the performance of DeepCAGE under loci with difference properties, we introduced two statistics, cell range and cell variability (see Methods) for describing the activity of a considering locus based on the true DNase-seq signal cross test cell types. 'Known loci' and 'novel loci' were first divided into three groups based on the two statistics. 'Low' denotes the loci of which statistics were lower than the 1st quartile and 'high' represents the loci of which statistics were higher than the 3rd quartile. 'Median' means that statistics of loci were between the 1st quartile and the 3rd quartile. Loci of 'median' cell range and cell variability are relatively easier to be predicted by DeepCAGE (Figure 3D), which is consistent to BIRD model [19]. For example, DeepCAGE achieves a median locus-wise Pearson's of 0.512 under 'known loci' with median cell range across test cell types, compared to 0.435 of 'known loci' with low cell range and 0.399 of 'known loci' with high cell range. Comparing 'known loci' to 'novel loci', DeepCAGE achieves a slight lower performance under the loci within the same statistics range. Take cell variability for example, DeepCAGE achieves a median locus-wise Pearson's r of 0.384, 0.514 and 0.448 under 'known

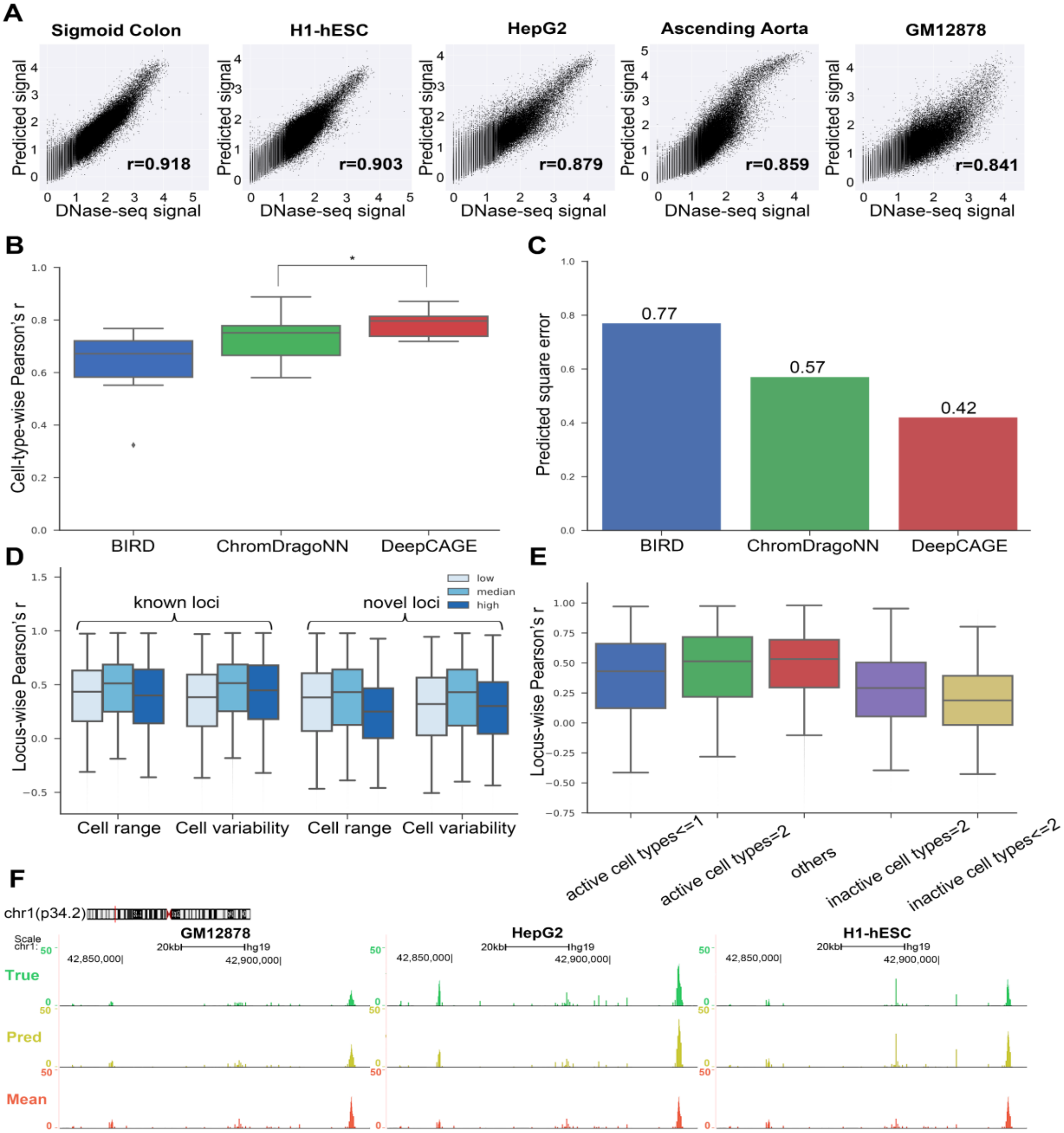
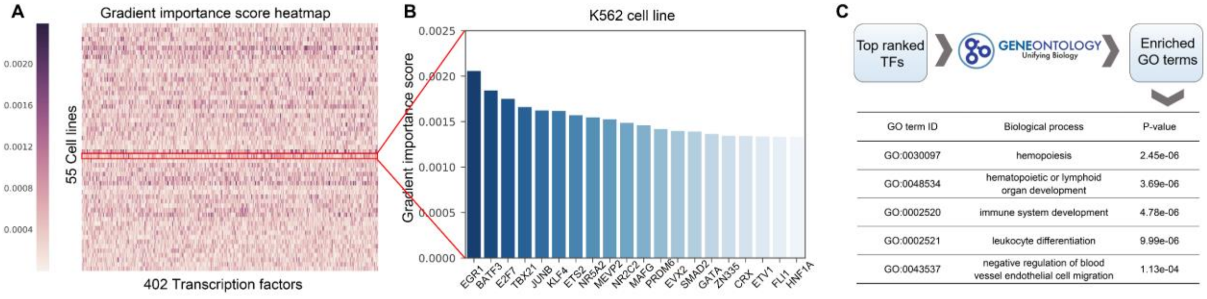


Figure 3. The performance of DeepCAGE regression model. (**A**) DeepCAGE predicts DNase-seq signals in five test cell types. (**B**) Cell-type-wise Pearson's r for three different methods across test cell types. *Two-sided Wilcoxon signed-rank test P-values=3.37×10^{-5} for comparing two methods. (**C**) The prediction square error for three difference methods. (**D**) Locus-wise Pearson's r achieved by DeepCAGE considering two locus statistics under both 'known loci' and 'novel loci'. (**E**) Locus-wise Pearson's r achieved by DeepCAGE considering cell-type specificity under 'known loci'. (**F**) An example of true (green) and predicted (yellow) DNase-seq signal of three test cell types under a same genomic region (chr1:42.83-42.93M). 'Mean' signal (red) denotes the average DNase-seq signal across training cell types.

loci' with three cell variability ranges (low, median, high) respectively. The performance decreases slightly when it comes to 'novel loci' with three cell variability ranges (0.320, 0.431 and 0.302). Based on the above



GO term ID	Biological process	P-value
GO:0030097	hemopoiesis	2.45e-06
GO:0048534	hematopoietic or lymphoid organ development	3.69e-06
GO:0002520	immune system development	4.78e-06
GO:0002521	leukocyte differentiation	9.99e-06
GO:0043537	negative regulation of blood vessel endothelial cell migration	1.13e-04

Figure 4. The gradient importance score (GIS) helps identify important transcription factors (TFs). (**A**) The GIS heatmap of 402 human core transcription factors (TFs) across 55 cell types. (**B**) The gradient importance score (GIS) of 20 top ranked TFs in K562 cell type. (**C**) Enriched GO terms by top ranked TFs in K562 cell type.

observation. we further divided 'known loci' into five groups based on the number of active test cell types when considering a locus. If a locus within a test cell type is denoted as active if the true normalized DNase-seq signal is greater than 1, and denoted as inactive otherwise, we found that the performance of DeepCAGE varies a lot under the loci with different number of active test cell types (Figure 3E). DeepCAGE tents to have a worse performance under loci of which the number of active/inactive test cell types is less than 2. The reason might be that the training dataset contains most of the 'median' type locus, thus DeepCAGE has a more powerful generalization ability when making a prediction under 'median' type locus. We also provided a vivid example of which we visualized both the true (green) and predicted (yellow) DNase-seq signal of a same genomic region across three test cell types (GM12878, HepG2 and H1-hESC) in the UCSC genome browser [28] (Figure 3F). Note that we also provided the 'mean signal' (red) as a reference which was calculated by taking the average DNase-seq signals across all training cell types. Obviously, DeepCAGE can well discriminate the difference of DNase-seq signals among test cell types while 'mean signal' fails.

Gradient importance score helps prioritize cell-type related TFs

We proposed a simple strategy for prioritizing cell-type related transcription factor (TF) according to the absolute gradient with regard to the expression of each TF (Figure 4A, see Methods). Take the K562 cell line for example, we calculated the average gradient importance score (GIS) of all TFs from all test loci within up-streaming 100k bp to down-streaming 100k bp of p53 gene, which was proved to have a key role in myeloid blast transformation [29]. 402 human core TFs were then prioritized by the average GIS score in K562 cell line (Figure 4B). Interestingly, many top ranked TFs in K562 cell type are related to functions in leukemia cell validated by previous literatures (Figure 4B). E.g. EGR-1 (rank^{1st}) is involved in regulating PMA-induced megakaryocytic differentiation of K562 cell line [30], the inhibition of E2F7 (rank^{3rd}) may lead to a reduction of miRNAs involved in leukemic cell lines [31], JunB (rank^{5th}) gene expression is inactivated by methylation in chronic myeloid leukemia [32]. The GO terms enriched by the top 5% prioritized TFs coding genes contain biological processes of leukocyte differentiation and hematopoietic development (Figure 4C). The gradient importance score gives us an intuitive explanation of which TF can play an important role in predicting the chromatin accessibility given a specific cell type and a region.

Model ablation analysis of DeepCAGE

In order to evaluate contributions of DNA sequence, TFs' gene expression and TF motif scores to the performance of DeepCAGE, we performed a model ablation analysis by removing the TFs' gene expression or TF motif scores. Take the DeepCAGE regression model for example, the mean cell-type-wise Pearson's correlation decreases by 11.2% and 1.4% when removing gene expression and motif scores of 402 core human TFs, respectively (Supplementary Figure 2). Gene expression data can significantly help improves the performance of DeepCAGE in cross cell type prediction while TFs motif scores achieve slight

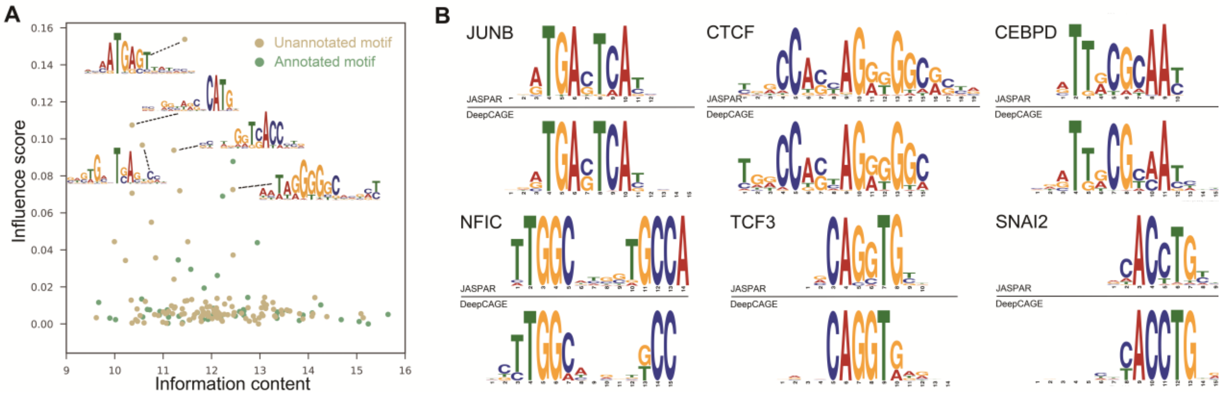


Figure 5. The convolutional layer in DeepCAGE recovers both known and novel motifs. (**A**) Green dots and yellow dots represent known motifs and novel motifs recovered by DeepCAGE, respectively. The x-axis describes the information content (see Methods) and the y-axis denotes the influence score, which is calculated by setting the filter weights to zeros and taking average changes of openness over test cell types. (**B**) we display matched motifs with an E-value threshold 0.01 in the format of sequence logos (above: known motif from the JASPAR database, below: motif learned by DeepCAGE).

incremental performance. The potential reason is that a large proportion of DNA sequence motifs have already been learned in the convolution layers of neural network. The motif scores can only achieve complementary contribution when contained in the prediction task.

DeepCAGE automatically learns the binding motifs of TFs

We explored the features that were automatically learned by DeepCAGE by investigating the weights of 160 filters from the first convolutional layer. The weights were converted into 160 position weight matrices (PWMs) through a proposed strategy (see Methods). We then compared to PWMs learned by DeepCAGE to the known motifs from JASPAR 2018 database [27]. 30% of the learned motifs can be matched to known motifs in JASPAR database with an E-value threshold of 0.01. Among the 48 matched filters, 52% have at least one matched core human TF used in DeepCAGE model. We first calculated the information content based on information entropy (see Methods) and set each of the filter weights to zeros and denoted the decreased cell-type-wise Pearson's correlation as the influence score for each motif. We showed several learned unmatched motifs that have high influence score (Figure 5A) and illustrates a few examples of learned motifs that could be matched to known motifs in JASPAR database (Figure 5B). The results demonstrated that DeepCAGE can not only help us find potential binding motifs, but also has the potential to guide us to find novel motifs which are not discovered by experiments yet.

DISCUSSION

Accurately predicting regulatory elements (REs) has become a fundamental problem in computation biology. In this paper, we demonstrated that the combination in software (CNNs), hardware (GPUs) and abundant genomic data can enable drastic performance boost on such problems. We introduced DeepCAGE, a deep learning framework that integrates a densely connected convolutional neural network to automatically extract DNA sequence signatures, capture TF binding motifs and implicate driving activity of transcription factors (TFs). DeepCAGE showed superior performance in both classification and regression settings. More importantly, it overcomes the limitations of previous sequence-based methods and expression-based methods. DeepCAGE could not only make cross cell type prediction, but also enables genome-wide prediction of chromatin accessibility, including novel loci that are not contained in the training set. A typical scenario to use DeepCAGE is to predict the activities of genome-wide regulatory elements in cell types where the chromatin accessibility data is not available.

To make DeepCAGE more understandable, we proposed two perspectives for model interpretation. First, the gradient importance score can give us an intuitional measurement of each transcription factor (TF) that contributes to the prediction take of each test cell type with regard to a specific locus. Second, the visualization of kernel weights in the first convolutional layer proves to contain abundant known and novel TF binding motifs. Model interpretation can help DeepCAGE well dissect regulatory landscape under various cell conditions.

Certainly, our model can be further improved from many aspects. First, DeepCAGE only considers the gene expression of 402 transcription factors (TFs) of which non-redundant binding motif could be found in HOCOMOCO database. Surely we can collect motif information of more TFs as prior knowledge by integrating multiple motif databases such as JASPAR [27], TRANSFAC [44] and UniPROBE [45] together. Second, we ignored the expression of genes which direct the synthesis of non-TF proteins. However, some proteins such as chromatin regulators (CRs), a class of enzymes with specialized function domains, can shape and maintain the epigenetic state in a cell context-dependent fashion [46], thus could also provide information for inferring chromatin accessible state. Third, DeepCAGE could not further determine the functions of a predicted region under a specific cell type currently. With the annotation of *cis*-regulatory elements in genome such as promoters, enhancers and silencers. DeepCAGE may further uncover the relationship between different kind of genomic regulatory elements and the genome-wide transcriptome profile.

To sum up, with DeepCAGE, a research can infer the chromatin accessibility code of the cell types with interests given corresponding gene expression data (RNA-seq, DNA microarray, etc.). We expect to see wide downstream applications of DeepCAGE with either public or in-house gene expression samples. We also hope DeepCAGE could help unveil the complex regulatory mechanism underlying various genetic signals.

DATA AVAILABILITY

DeepCAGE is an open-source software which can be downloaded from the GitHub repository (https://github.com/kimmo1019/DeepCAGE). Detailed instructions of DeepCAGE is provided for users.

SUPPLEMENTARY DATA

Supplementary Data are available online.

ACKNOWLEDGEMENT

We thank Mengmeng Wu, Zhana Duren and Fengling Chen for their helpful discussions.

FUNDING

National Key Research and Development Program of China No. 2018YFC0910404, the National Natural Science Foundation of China Nos. 61873141, 61721003, 61573207, and the Tsinghua-Fuzhou Institute for Data; National Institutes of Health grants P50HG007735, R01HG010359. R.J. is also supported by a RONG professorship at the Institute for Data Science of Tsinghua University.

CONFLICT OF INTEREST

None declared.

REFERENCES

1. Lowe, W.L., Reddy, T.E.: Genomic approaches for understanding the genetics of complex disease. Genome research 25, 1432-1441 (2015)
2. Haraksingh, R.R., Snyder, M.P.: Impacts of variation in the human genome on gene regulation. Journal of molecular biology 425, 3970-3977 (2013)

3. Zhang, F., Lupski, J.R.: Non-coding genetic variants in human disease. Human molecular genetics 24, R102-R110 (2015)

4. Alexander, R.P., Fang, G., Rozowsky, J., Snyder, M., Gerstein, M.B.: Annotating non-coding regions of the genome. Nat Rev Genet 11, 559-571 (2010)

5. Kellis, M., Wold, B., Snyder, M.P., Bernstein, B.E., Kundaje, A., Marinov, G.K., Ward, L.D., Birney, E., Crawford, G.E., Dekker, J.: Defining functional DNA elements in the human genome. Proceedings of the National Academy of Sciences 111, 6131-6138 (2014)

6. Crawford, G.E., Holt, I.E., Whittle, J., Webb, B.D., Tai, D., Davis, S., Margulies, E.H., Chen, Y., Bernat, J.A., Ginsburg, D.: Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome research 16, 123-131 (2006)

7. Johnson, D.S., Mortazavi, A., Myers, R.M., Wold, B.: Genome-wide mapping of in vivo protein-DNA interactions. Science 316, 1497-1502 (2007)

8. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., Greenleaf, W.J.: Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods 10, 1213 (2013)

9. Liu, Q., Xia, F., Yin, Q., Jiang, R.: Chromatin accessibility prediction via a hybrid deep convolutional neural network. Bioinformatics (2017)

10. Liu, Q., Gan, M., Jiang, R.: A sequence-based method to predict the impact of regulatory variants using random forest. BMC Syst Biol 11, 7 (2017)

11. Alipanahi, B., Delong, A., Weirauch, M.T., Frey, B.J.: Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol 33, 831-838 (2015)

12. Singh, R., Lanchantin, J., Robins, G., Qi, Y.: Deepchrome: deep-learning for predicting gene expression from histone modifications. Bioinformatics 32, i639-i648 (2016)

13. Angermueller, C., Lee, H.J., Reik, W., Stegle, O.: DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. Genome biology 18, 67 (2017)

14. Liu, Q., Lv, H., Jiang, R.: hicGAN infers super resolution Hi-C data with generative adversarial networks. Bioinformatics 35, i99-i107 (2019)

15. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature 521, 436 (2015)

16. Zhou, J., Troyanskaya, O.G.: Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods 12, 931-934 (2015)

17. Quang, D., Xie, X.: DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic acids research gkw226 (2016)

18. Kelley, D.R., Snoek, J., Rinn, J.L.: Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res 26, 990-999 (2016)

19. Zhou, W., Sherwood, B., Ji, Z., Xue, Y., Du, F., Bai, J., Ying, M., Ji, H.: Genome-wide prediction of DNase I hypersensitivity using gene expression. Nature communications 8, 1038 (2017)

20. Huang, G., Liu, Z., Weinberger, K.Q., van der Maaten, L.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3. (Year)

21. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research 15, 1929-1958 (2014)

22. Kulakovskiy, I.V., Vorontsov, I.E., Yevshin, I.S., Soboleva, A.V., Kasianov, A.S., Ashoor, H., Ba-Alawi, W., Bajic, V.B., Medvedeva, Y.A., Kolpakov, F.A.: HOCOMOCO: expansion and enhancement of the collection of transcription factor binding sites models. Nucleic acids research 44, D116-D125 (2015)

23. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., Glass, C.K.: Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular cell 38, 576-589 (2010)

24. Consortium, E.P.: An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57-74 (2012)

25. Nair, S., Kim, D.S., Perricone, J., Kundaje, A.: Integrating regulatory DNA sequence and gene expression to predict genome-wide chromatin accessibility across cellular contexts. Bioinformatics 35, i108-i116 (2019)

26. Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., Noble, W.S.: Quantifying similarity between motifs. Genome biology 8, R24 (2007)

27. Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J.A., van der Lee, R., Bessy, A., Chèneby, J., Kulkarni, S.R., Tan, G.: JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic acids research 46, D260-D266 (2017)

28. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., Haussler, D.: The human genome browser at UCSC. Genome research 12, 996-1006 (2002)

29. Law, J.C., Ritke, M.K., Yalowich, J.C., Leder, G.H., Ferrell, R.E.: Mutational inactivation of the p53 gene in the human erythroid leukemic K562 cell line. Leukemia research 17, 1045-1050 (1993)

30. Cheng, T., Wang, Y., Dai, W.: Transcription factor egr-1 is involved in phorbol 12-myristate 13-acetate-induced megakaryocytic differentiation of K562 cells. Journal of Biological Chemistry 269, 30848-30853 (1994)

31. Gabra, M.M., Salmena, L.: microRNAs and Acute Myeloid Leukemia chemoresistance: a mechanistic overview. Frontiers in oncology 7, (2017)

32. Yang, M.-Y., Liu, T.-C., Chang, J.-G., Lin, P.-M., Lin, S.-F.: JunB gene expression is inactivated by methylation in chronic myeloid leukemia. Blood 101, 3205-3211 (2003)

33. Morris, V.A., Cummings, C.L., Korb, B., Boaglio, S., Oehler, V.G.: Deregulated KLF4 expression in myeloid leukemias alters cell proliferation and differentiation through microRNA and gene targets. Molecular and cellular biology 36, 559-573 (2016)

34. Ge, Y., LaFiura, K.M., Dombkowski, A.A., Chen, Q., Payton, S.G., Buck, S.A., Salagrama, S., Diakiw, A.E., Matherly, L.H., Taub, J.W.: The role of the proto-oncogene ETS2 in acute megakaryocytic leukemia biology and therapy. Leukemia 22, 521 (2008)

35. Singh, J., Saxena, A., Christodoulou, J., Ravine, D.: MECP2 genomic structure and function: insights from ENCODE. Nucleic acids research 36, 6035-6047 (2008)

36. Zhang, X., Gamble, M.J., Stadler, S., Cherrington, B.D., Causey, C.P., Thompson, P.R., Roberson, M.S., Kraus, W.L., Coonrod, S.A.: Genome-wide analysis reveals PADI4 cooperates with Elk-1 to activate c-Fos expression in breast cancer cells. PLoS genetics 7, e1002112 (2011)

37. Appaiah, H., Bhat-Nakshatri, P., Mehta, R., Thorat, M., Badve, S., Nakshatri, H.: ITF2 is a target of CXCR4 in MDA-MB-231 breast cancer cells and is associated with reduced survival in estrogen receptor-negative breast cancer. Cancer biology & therapy 10, 600-614 (2010)

38. Khaled, W.T., Lee, S.C., Stingl, J., Chen, X., Ali, H.R., Rueda, O.M., Hadi, F., Wang, J., Yu, Y., Chin, S.-F.: BCL11A is a triple-negative breast cancer gene with critical functions in stem and progenitor cells. Nature communications 6, 5987 (2015)

39. Nemescu, D., Ursu, R.G., Nemescu, E.R., Negura, L.: Heterogeneous distribution of fetal microchimerism in local breast cancer environment. PloS one 11, e0147675 (2016)

40. Neupane, M., Clark, A.P., Landini, S., Birkbak, N.J., Eklund, A.C., Lim, E., Culhane, A.C., Barry, W.T., Schumacher, S.E., Beroukhim, R.: MECP2 is a frequently amplified oncogene with a novel epigenetic mechanism that mimics the role of activated RAS in malignancy. Cancer discovery (2015)

41. Yuan, Z.-Y., Dai, T., Wang, S.-S., Peng, R.-J., Li, X.-H., Qin, T., Song, L.-B., Wang, X.: Overexpression of ETV4 protein in triple-negative breast cancer is associated with a higher risk of distant metastasis. OncoTargets and therapy 7, 1733 (2014)

42. Fabre-Guillevin, E., Malo, M., Cartier-Michaud, A., Peinado, H., Moreno-Bueno, G., Vallée, B., Lawrence, D.A., Palacios, J., Cano, A., Barlovatz-Meimon, G.: PAI-1 and functional blockade of SNAI1 in breast cancer cell migration. Breast Cancer Research 10, R100 (2008)
43. Emmert-Streib, F., de Matos Simoes, R., Mullan, P., Haibe-Kains, B., Dehmer, M.: The gene regulatory network for breast cancer: integrated regulatory landscape of cancer hallmarks. Frontiers in genetics 5, 15 (2014)
44. Wingender, E., Dietze, P., Karas, H., Knüppel, R.: TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic acids research 24, 238-241 (1996)
45. Newburger, D.E., Bulyk, M.L.: UniPROBE: an online database of protein binding microarray data on protein–DNA interactions. Nucleic acids research 37, D77-D82 (2008)
46. Chen, T., Dent, S.Y.: Chromatin modifiers and remodellers: regulators of cellular differentiation. Nature Reviews Genetics 15, 93 (2014)