DeFine: deep convolutional neural networks accurately quantify intensities of transcription factor-DNA binding and facilitate evaluation of functional non-coding variants

Meng Wang[1], Cheng Tai[2,3], Weinan E[2,3,4,*] and Liping Wei[1,*]

[1]Center for Bioinformatics, State Key Laboratory of Protein and Plant Gene Research, School of Life Sciences, Peking University, Beijing, 100871, P.R. China, [2]Center for Data Science, Peking University, Beijing, 100871, P.R. China, [3]Beijing Institute of Big Data Research, Beijing, 100871, P.R. China and [4]Department of Mathematics and PACM, Princeton University, Princeton, NJ, 08544, USA

ABSTRACT

The complex system of gene expression is regulated by the cell type-specific binding of transcription factors (TFs) to regulatory elements. Identifying variants that disrupt TF binding and lead to human diseases remains a great challenge. To address this, we implement sequence-based deep learning models that accurately predict the TF binding intensities to given DNA sequences. In addition to accurately classifying TF-DNA binding or unbinding, our models are capable of accurately predicting real-valued TF binding intensities by leveraging large-scale TF ChIP-seq data. The changes in the TF binding intensities between the altered sequence and the reference sequence reflect the degree of functional impact for the variant. This enables us to develop the tool DeFine (Deep learning based Functional impact of non-coding variants evaluator, http://define.cbi.pku.edu.cn) with improved performance for assessing the functional impact of non-coding variants including SNPs and indels. DeFine accurately identifies the causal functional non-coding variants from disease-associated variants in GWAS. DeFine is an effective and easy-to-use tool that facilities systematic prioritization of functional non-coding variants.

INTRODUCTION

The precise control of spatio-temporal gene expression is regulated by the binding of cell type-specific transcription factors (TFs) to regulatory elements, including promotors, enhancers, silencers and insulators, across the vast non-coding part of the genome (1–4). Large-scale expression quantitative trait loci (eQTL) analysis and genome-wide association studies (GWAS) have identified abundant variants that associate with diverse gene expression levels and human diseases (5,6). Most of these variants reside in the non-coding regions of the human genome, suggesting that such non-coding variants play crucial roles in human disorders by disrupting the *cis*-regulation of gene expression (7). In particular, non-coding *cis*-regulatory variants have been shown to be functional in transcriptional alterations by affecting TF binding and leading to human diseases, including Mendelian disorders, complex diseases and cancers (8–11).

The genome-wide landscape of *cis*-regulatory sequences and TF binding profiles could be decoded by high-throughput sequencing-based methodologies such as chromatin immunoprecipitation followed by massive parallel sequencing (ChIP-seq) (12). Despite the rapid advancement of functional genomic sequencing technologies, interpretation of the functional consequences of variants on regulatory elements remains a great challenge due to the complexity of cell type-specific transcription regulation systems (8,13). In addition, a limited number of non-coding variants have been functionally validated by experiments (14). Previous efforts on interpreting genomic variants are mainly concentrated on variants in the coding regions and have achieved considerably high performance (15–17). However, as GWAS reveal that the non-coding variants also play an important role in complex diseases, identifying the pathogenic functional non-coding variants from the massive neutral ones is essential in genotype–phenotype relationship research and precision medicine.

The functional genomics high-throughput sequencing data generated by the ENCODE project (18) and the NIH Roadmap Epigenomics project (19) facilitates the systematic annotation of TF binding profiles, which makes it pos-

*To whom correspondence should be addressed. Tel: +86 10 6275 5206; Fax: +86 10 6276 8628; Email: weilp@mail.cbi.pku.edu.cn
Correspondence may also be addressed to Weinan E. Tel: +1 609 258 3683; Fax: +1 609 258 1735; Email: weinan@math.princeton.edu

sible to identify cell type-specific regulatory elements and to study the TF binding intensities to various DNA sequences. The high-throughput chromosome conformation capture sequencing data helps to uncover genome-wide chromatin organization and interactions (20), enabling the identification of target genes for distal regulatory elements. The recent rapid development and wide application of deep learning technologies (21) have shown extraordinary performance in variety of tasks including functional genomics predictions (22,23). Deep learning is distinguished by its capability to automatically discover predictive signatures and handle high-dimensional data. Deep learning-based frameworks such as DeepBind (23), DeepSEA (24), Basset (25), DanQ (26) and Basenji (BioRxiv: https://doi.org/10.1101/161851) have shown remarkable advantages over conventional machine-learning methods for predicting TF binding and chromatin accessibility from DNA sequences.

However, very few *in silico* methods have been developed to assess the functional impact of non-coding variants, including CADD (27), GWAVA (28), FunSeq2 (29) and DeepSEA (24). The prediction accuracies must be further improved, and these tools suffer from various limitations. Tools measuring evolutionary conservation such as CADD have low performance for evaluating non-coding variants because most of the regulatory variants are not subjected to evolutionary constraints (30). Tools that rely on known non-coding variants related to human disease such as GWAVA and FunSeq2 are limited by the number of available training variants (14). In addition, most of the known pathogenic non-coding variants reside in the promoter regions or conserved sites, causing ascertainment bias in the training set. Tools such as DeepSEA predict the binding or unbinding binary outcome of TFs to a given sequence, ignoring the binding affinity of TFs with different sequence preferences (2). Moreover, none of these tools could indicate the target gene(s) possibly affected by the predicted functional variants. Most regulatory elements regulate distal rather than proximal genes in the primary genome sequence (31,32). Discovering the affected gene(s) is required for interpreting the functional consequence of the non-coding variants.

In this paper, we present a method based on deep convolutional neural networks (CNNs) (21) that predict the intensities of cell type-specific DNA binding of TFs using large-scale TF ChIP-seq data (18). The change in the predicted binding signal value between the altered sequence and the reference sequence reflects the extent of the functional impact of the variant. The deep learning models depend solely on genomic sequences, requiring no prior knowledge of the variants on the sequences. We demonstrate that the deep CNN models accurately predict the ChIP-seq signal values for different sequence-specific TFs in different cell types. The deep CNN models accurately capture the binding motifs of TFs. Based on these deep CNN models, we develop the tool DeFine (Deep learning-based Functional impact of non-coding variants evaluator) to assess the functional impact of all types of non-coding variants including SNPs and indels in a cellular context with single base resolution. Performance evaluation and comparison show that the classifiers based on DeFine functional scores outperform the state-of-the-art tools in three different test sets.

When applied to prioritize candidate non-coding variants associated with disease from GWAS, DeFine identifies the causal functional non-coding variants with remarkable accuracy. Furthermore, DeFine integrates cell type-specific three-dimensional genome contact maps from *in situ* Hi-C experiments (20). Thus, DeFine is able not only to determine whether a non-coding variant has a functional impact but also to indicate the potential gene(s) affected by this variant. DeFine facilitates high-throughput prioritization of non-coding variants on a large scale.

MATERIALS AND METHODS

Overview of the DeFine deep learning model

The deep CNN in DeFine consisted of a hierarchical architecture that used raw DNA sequence as input and predicted the real-valued ChIP-seq signal value, which measured the *in vivo* TF-DNA binding intensity (Figure 1A). The deep CNN model in DeFine consisted of convolution layers, rectification layers, pooling layers and fully connected layers. Each input sequence was converted to a one-hot matrix with 4 rows and 300 columns. The four rows corresponded to the four nucleotides A, G, T and C. For positions with N, the whole corresponding column was filled with zeroes. As the DNA is a double-helix, TFs could recognize either strand of the DNA at a given position. Thus, both the forward sequence and its reverse complementary sequence were simultaneously modeled in our deep CNN. The reverse complementary sequence of each input sequence was also encoded by the one-hot matrix. Both the input matrixes for the forward and reverse-complement sequences were simultaneously fed into the convolution layers. The convolution layers for the forward sequence and its reverse-complement sequence shared the same set of filters. This strategy of reverse-complement parameter sharing is similar to the one proposed by a parallel manuscript (BioRxiv: http://dx.doi.org/10.1101/103663), which explicitly constrained the presentation of a reverse-complement filter for each learned filter, and illustrated the effectiveness of modeling both strand of DNA sequences together. In our convolution layer, the same set of filters was learned from both input sequence and its reverse-complement sequence together in the single end-to-end neural network, by sharing filters between the forward part convolution and the reverse-complement part convolution (Figure 1A). Sixteen filters were used in each convolution layer. Each filter was a 4-by-24 matrix. These filters automatically extracted predictive features from input sequences during model training. After convolution, the rectified linear units (ReLU) were used to output the filter scanning results that were above the thresholds, which were learned during model training. Both max pooling and average pooling were utilized in the pooling layer. Max pooling was applied to find the most significant activation signal in a sequence for each filter. The average pooling considered the whole sequence context by averaging the filter scanning results at each position of the sequence. All the pooling results of both forward and reverse strand sequences were combined in one vector, resulting in a vector with a size of 64. The vector was batch-normalized (33) before inputting it into the fully connected layer. Two fully connected layers were employed in our model, each with 128 nodes. A
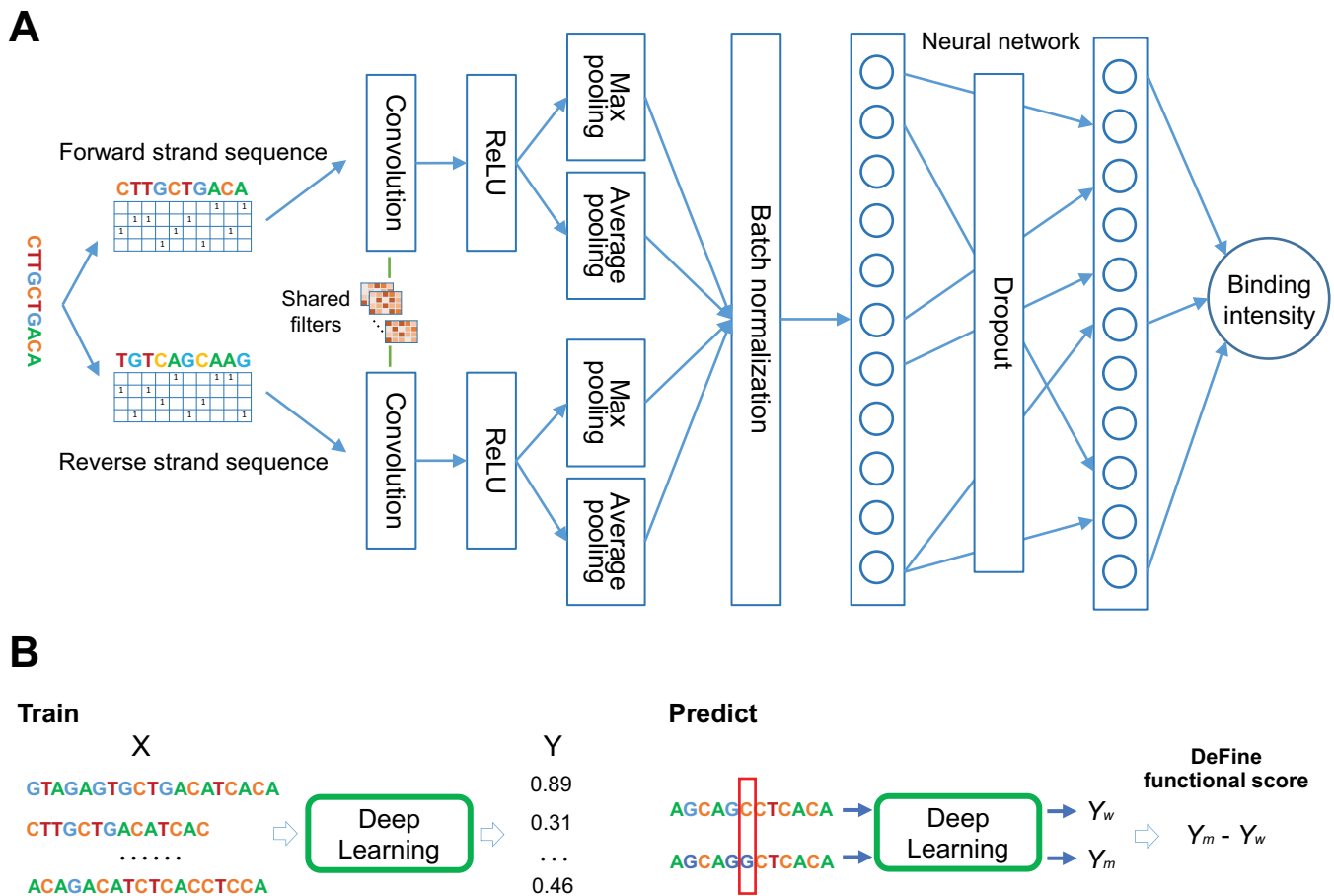
A

B

Figure 1. The deep learning method for assessing the functional impact of non-coding variants. (**A**) The architecture of the CNN in DeFine for predicting TF-DNA binding intensities from DNA sequences. (**B**) DeFine functional score for assessing the functional impact of non-coding variants. The DeFine functional score of a variant for one TF is derived as the difference in TF binding intensities predicted by the deep learning model between the reference sequence and the altered sequence centered at the variant.

dropout (34) layer with a probability of 0.5 was added between the two fully connected layers to improve the generalization capability of the model and avoid overfitting. The output layer was a regression layer, which outputted the predicted ChIP-seq signal intensity. The deep CNN was implemented with Torch7 (http://torch.ch).

ChIP-seq data sources

The ChIP-seq data for the sequence-specific TFs in the K562 and GM12878 cell lines from the ENCODE project were employed. All the data were downloaded from https://www.encodeproject.org. To ensure the quality of the data for reliable downstream analysis, datasets without biological replicates were discarded. For the ChIP-seq datasets of the same TF from different experiments, they were used as biological replicates. The raw ChIP-seq short sequencing reads of all biological replicates and the corresponding control library raw sequences for each TF were downloaded in fastq format if available. When the raw reads in fastq format were not available for a TF, the mapped reads in bam format were downloaded, and the sequences were extracted and then converted into fastq format. ChIP-seq datasets of TFs with <1000 peaks called by our analyzing pipeline were discarded. The accession IDs for all the utilized TFs ChIP-seq data of the K562 and GM12878 cell lines are provided in Supplementary Tables S1 and 2, respectively.

Unified processing pipeline for peak calling from ChIP-seq raw reads

The raw short reads of all replicates and controls were first cleaned to remove adaptor sequences and truncate low-quality reads using Trimmomatic (35). If the trimmed reads were <25 bp, then they were discarded. The reads qualities before and after cleaning were assessed. Burrows-Wheeler Aligner (BWA) (36) was used to map all the cleaned short reads to the reference genome with default parameters. As both K562 and GM12878 cell lines were derived from female individuals, the GRCh37 reference genome without the Y chromosome and random contigs was employed as the reference genome during mapping. The mitochondrial sequence was included in the reference genome. Polymerase chain reaction duplicates were then removed using Picard (http://broadinstitute.github.io/picard), and the uniquely mapped reads were retained. The processed bam files for all the replicates were merged together to create a pooled sample bam file. If there were multiple control

libraries for one TF, all the processed bam files from all the controls were merged and used as a single control bam file. The pooled sample bam file was randomly separated into two files with an equal number of sequences to create two pseudo-replicates. Next, all the bam files of the original replicates, the two pseudo-replicates and the pooled sample bam file were used as the input for calling peaks by SPP tool (37). The called peaks in each replicate were ranked according to their signal values, and only peaks that appeared in both replicates were retained. The final reproducible peaks for each TF were derived using the IDR (irreproducible discovery rate) framework (38) with a cutoff of 0.01. If multiple replicates were available, pairwise comparisons were performed. The maximum of the reproducible peaks of original replicates and the two pseudo-replicates were used as the final optimal number of peaks.

Training data generation

For each TF, the peak regions and corresponding signal values were obtained from the peak calling results. By manually examining the ChIP-seq peaks, we found that most of the peaks with extremely high signal values located at genomic regions with low complexity. Thus, to remove the outliers with extremely high signal values, the peak regions with the top 1% signal values for each TF were discarded. The signal values were log-transformed and normalized by min-max scaling between 0 and 1. The genomic sequences of each peak were extracted from the reference genome according to the peak regions. Each sequence was then refined using a reference-guided local re-assembly based on the whole genomes sequences of K562 and GM12878, utilizing Pilon (39). This step corrected the cell type-specific variants in the genome sequences of each cell line. The K562 whole-genome sequencing reads were downloaded from https://www.ncbi.nlm.nih.gov/sra/SRX118400. The GM12878 whole-genome sequencing reads were obtained at ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/NIST_NA12878_HG001_HiSeq_300x/RMNISTHS_30xdownsample.bam. The peak sequences were diverse in length, and most sequences were 300 bp. We fixed the sequence length at 300 bp. For sequences that were longer than 300 bp, they were cropped at the 5′ and 3′ ends, centering at the summit of each peak. For sequences that were shorter than 300 bp, both ends were padded with equal numbers of Ns to form 300-bp sequences. For each TF, the corrected peak sequences were used as the model inputs, and the corresponding transformed ChIP-seq peak signal values were employed as model output. Furthermore, to augment the training set, we randomly selected sequences from genomic regions that showed no binding of any known TFs, and they were applied as training sequences with zero signal values. This is a simple method to augment the data but not optimal (40). More conservative method could be employed to generate samples with zero signal values. The number of randomly selected sequences with zero signal values were the same as the number of ChIP-seq peaks for each TF.

The training sequences and corresponding ChIP-seq signal intensities of each TF were partitioned such that 70% of the data were used for training, 15% were used for valida-tion and 15% were used for model testing. The validation dataset was used in the grid search process during model training to determine the optimal hyper-parameters in the model.

Training of deep CNN

The mean square error with L2 regularization (weight decay) was employed as the loss function. The parameters in the deep CNN model were randomly initialized using Gaussian distribution with mean value 0 and standard deviation 1. We trained the deep CNN with a mini-batch stochastic gradient decent algorithm. The mini-batch size was 128. The gradients were calculated using backpropagation. All the parameters in the model were updated based on the gradients during each mini-batch training. After each epoch of training, the loss in the validation set was assessed and monitored. When the loss in the validation set did not decrease in 50 successive epochs of training, the model training process was stopped (early stopping). The model with the smallest loss in the validation set was saved. The optimal hyper parameters, including the learning rate and lambda for weight decay, were determined by the grid search. All the deep CNN models were trained on GPU.

Generating motifs learned by deep CNN

The learned motif of the deep CNN model for each TF were revealed by extracting all the test sequences (15% total data for each TF ChIP-seq) that were active in the rectification layer (return a non-zero value) after filter scanning in the convolution layer. For each sequence in the test dataset of one TF, we sought the position that had the maximum convolution value among all the filters in both the forward and reverse strands. The 24-bp (filter width) subsequence starting at this position of the test sequence was extracted. All of the 24-bp subsequences with the maximum convolution value for each sequence in the test set were pooled together and aligned. The frequencies of the four nucleotides at each position were then calculated, and the position weight matrix representing the TF motif was derived.

Compiling test variants

Three different types of test datasets were compiled with functional non-coding variants and neutral variants: HGMD (Human Gene Mutation Database)-based, GWAS (Genome Wide Association Study)-based and eQTL-based variants dataset. For the functional variants in the HGMD-based dataset, regulatory non-coding variants from the HGMD professional database (release 2016.1) were employed. For the functional variants in the GWAS-based dataset, intergenic variants that were significantly associated with disease in the GWAS catalog (6) (https://www.ebi.ac.uk/gwas/docs/downloads, downloaded at 20160817) were utilized. For functional variants in the eQTL-based dataset, the best associated eQTL in the lymphoblastoid cell lines (LCL) of the EUR population (5) were used (downloaded from http://www.ebi.ac.uk/arrayexpress/experiments/E-GEUV-3/files/analysis_results/). For the neutral variants in each

of the three datasets, non-coding variants from the 1000 Genomes Project were randomly selected, with a matched variant number and matched allele frequency distribution. For each dataset, five neutral variant sets were composed, and the performance of the five sets was averaged in downstream analyses based on these datasets. Both SNPs and indels were included in these datasets, and their ratio was matched in each positive and negative set. The 1000 Genomes variants were downloaded from ftp://ftp-trace. ncbi.nih.gov/1000genomes/ftp/release/20130502/ALL. wgs.phase3_shapeit2_mvncall_integrated_v5b.20130502. sites.vcf.gz. For the region-restricted analysis, the neutral variants were randomly selected from the 1000 Genomes variants within 5 kb of the positions around the functional variants in each dataset.

Training of the gradient boosting decision tree classifier

To investigate the capability of DeFine functional scores in discriminating functional non-coding variants from neutral ones, we trained the gradient boosting decision tree (GBDT) classifier based on the functional scores of all the TFs in one cell line for each of the three test datasets compiled above. In addition to the DeFine functional scores, we also leveraged four widely employed conservation scores, namely, GERP_NR (41), GERP_RS (41), PhyloP (42) and PhastCons (43), as input features for the GBDT classifier. We built three classifiers for each dataset. The classifier named DeFine-regression utilized DeFine-predicted binding intensities as features. The classifier named DeFine-classification was based on the modified deep CNN model that predicted the binary outcome of TF-DNA binding/unbinding. The classifier named DeFine-combine employed both the regression version and the classification version of the DeFine scores as features. The GBDT was implemented using xgboost (https://github.com/dmlc/xgboost). The maximum training round was 1000. Early stopping was performed when training the classifiers. All the hyper parameters in the gradient boosting tree, including the learning rate, maximum depth, gramma, lambda, subsample rate and column sample rate, were determined using the grid search. The GERP scores for the whole genome sites were downloaded from http://mendel.stanford.edu/SidowLab/downloads/gerp/. The PhyloP scores could be accessed at http://hgdownload. cse.ucsc.edu/goldenpath/hg19/phyloP100way/hg19. 100way.phyloP100way.bw. The PhastCons scores were available at http://hgdownload.cse.ucsc.edu/goldenpath/ hg19/phastCons100way/hg19.100way.phastCons.bw.

Performance evaluation and comparison

The performance of the DeFine functional scores-based classifiers was measured using 10-fold cross validation for each of the three test datasets. Since all of the test datasets were balanced, the receiver operating characteristic (ROC) curves were generated using the three datasets to measure the performance of the GBDT classifier (44). The area under the curve (AUC) was compared with the other tools, including CADD, DeepSEA and GWAVA. The CADD scores were predicted at http://cadd.gs.washington.edu/score, and

the Phred scores were used. The DeepSEA scores for each test variants were predicted at http://deepsea.princeton. edu/job/analysis/create/, and the functional significance scores were used to calculate its AUC on each test set. GWAVA was installed and run locally. The source code of GWAVA was downloaded from http://www.sanger.ac.uk/ science/tools/gwava.

Prioritizing candidate disease-related non-coding variants

The candidate non-coding variants associated with Hirschsprung disease were from a family-based association study (45). The 12 screened variants located in the 5′ *RET* gene and the first intron of *RET* were included (Table 1 of Emison *et al.* (45)). For the colorectal cancer risk-associated variant list (46), all 13 common variants (Supplementary Table S2 of Lubbe *et al.* (46)) in the 14q22.2 genomic region screened in a large cohort from a targeted association study were employed. The genomic coordinates of the variants were converted to GRCh37-based positions. These variants were ranked by scores given by DeFine, CADD and DeepSEA. GWAVA was not used to prioritize these variants because of the causal variants in the two lists presented in the training variants of GWAVA.

Hi-C genome contact map integration

To find the target genes regulated by the functional non-coding variants, we integrated cell type-specific three-dimensional genome contact maps revealed by *in situ* Hi-C experiments. The contact maps for the K562 and GM12878 cell lines were downloaded from GEO (https://www.ncbi. nlm.nih.gov/geo/) under accession number GSE63525. For the K562 cell line, maps with a 5-kb resolution derived from reads with a mapping quality of at least 30 were used. For GM12878, maps with 1-kb resolution derived from reads with a mapping quality of at least 30 were integrated. To reduce noise from the maps, raw contact metrics with values of <10 were discarded from the contact maps. The KR-normalized metrics were then employed to build the final contact maps.

For each input variant, all the genomic regions that interacted with the variant position were searched through the contact maps of the selected cell line. Genes with a promoter located in these regions were then derived. The region from 1 kb upstream of the gene transcription start site (TSS) to 1 kb downstream of the TSS was considered the promoter region for that gene. All the possible target genes for each variant were sorted by their normalized contact metrics.

RESULTS

The TF binding intensities from ChIP-seq were accurately predicted from DNA sequences by deep learning models

We built deep CNNs predicting the TF binding intensity to a given sequence (Figure 1A), which was measured by the signal value of the ChIP-seq peaks (47), in a cell type-specific manner (12). We took advantage of the large-scale ChIP-seq data generated by the ENCODE project (18), leveraging the massive sequences of the ChIP-seq peaks and the corresponding signal values as training data. We built

deep CNN models for sequence-specific TFs that had adequate ChIP-seq peaks in the K562 and GM12878 cell lines, as abundant TF ChIP-seq data were available for these two cell lines in the ENCODE project. ChIP-seq datasets of 79 sequence-specific TFs in K562 (Supplementary Table S1) and 69 sequence-specific TFs in GM12878 (Supplementary Table S2) were employed, and the CNN models were trained for each of these TFs. We compiled the training data using these ChIP-seq datasets, which were re-analyzed from raw short sequencing reads, and reliable peaks were derived with the IDR framework (48) using a unified pipeline developed by our group (see 'Materials and Methods' section). The peak sequences were refined based on the whole-genome sequence of each cell line, regarding the cell type-specific variants. The number of ChIP-seq peaks ranged from 1098 to 65 232 for TFs in K562 and from 1003 to 59 037 for TFs in GM12878. To augment the training set, genomic sequences that did not show binding of any TFs in each cell line were randomly selected and incorporated into the training data with zero signal values (see 'Materials and Methods' section).

The model performance evaluation showed that the *in vivo* ChIP-seq signal intensities were accurately predicted by the *in silico* deep learning models. The correlations between the ChIP-seq experimental signal values and the predicted signal values were assessed in the testing set for each TF in each cell line. The experimental signal values and the predicted signal values were highly correlated. Examples for TFs in GM12878 were shown in Figure 2A, and examples of TFs in K562 were shown in Figure 2B. The ChIP-seq peak signals are prone to noise, especially for positions with weak binding of TFs. As shown in Figure 2A and B, the weak TF binding signals were hard to predict accurately by the model and the predicted signals got flat for these weak binding signals. For higher TF binding intensities, the predicted signals correlated with the real signals very well. The Pearson and Spearman correlation coefficients for all the TFs in the two cell lines were summarized in Figure 2C. For models of TFs in GM12878, the median Pearson correlation coefficient was 0.754, and the median Spearman correlation coefficient was 0.793. For models of TFs in K562, the median Pearson correlation coefficient and Spearman correlation coefficient were 0.793 and 0.805, respectively.

We investigated factors that may affect the model performance. First, we evaluated the effect of the input sequence length on the prediction accuracy. By default, all the models were trained and evaluated with input sequences with a length of 300 bp. We trimmed the input sequences to 75 bp centered at the summit of each ChIP-seq peak. Next, we re-trained and evaluated all the deep learning models with 75-bp input sequences. The results showed that the models with 300-bp input sequences outperformed those with 75-bp input sequences in each cell line (Figure 2D), as measured by both Pearson correlation and Spearman correlation. This result suggested that the sequence context contributed to the high performance for predicating TF binding intensities from DNA sequences.

The impact of the ChIP-seq experiment quality on the model prediction performance was assessed. The quality of the ChIP-seq experiments was measured by strand cross-correlation (37). The quality tag assigned to each ChIP-seq dataset was based on the relative strand correlation, which assessed the signal-to-noise ratio in the ChIP-seq experiments (48). As expected, results showed that models trained using ChIP-seq data with a higher experimental quality tended to have higher prediction performances in both cell lines (Supplementary Figure S1).

Unlike existing methods that solely predict the binary outcome of TF-DNA binding or unbinding, our models were able to accurately predict real-valued TF-DNA binding intensities from ChIP-seq experiments. Nevertheless, it was worthwhile to evaluate the capability of our deep CNN framework to classify TF-DNA binding or unbinding. To achieve this goal, we modified the output layer of our deep CNN to use the sigmoid function, kept the other structure of the model unchanged and switched to employ the binary cross entropy function as the loss function. The classification model for each TF in each cell type was trained using training sequences from ChIP-seq peaks as binding sequences and the same number of randomly selected genomic sequences without any TF binding as unbinding sequences. The performance of the classification model for each TF was assessed using ROC analysis with independent test sequences for each TF. The results revealed that our deep CNN framework was able to classify genomic sequences as bound or unbound by given TFs with very high accuracies (Figure 3). The median AUC for classification models of TFs in GM12878 was 0.979, and the median AUC for classifiers of TFs in K562 was 0.992. These results demonstrated the high performance of our deep CNN framework both for quantifying TF-DNA binding intensities and for classifying TF-DNA binding or unbinding.

The features learned by the deep learning models captured the binding motifs of TFs

One of the distinctive advantages of the deep learning models is the ability to automatically extract predictive features from inputs during the model training (21). We explored features that were learned in our deep CNNs by investigating the test sequences that activated the filters in the convolution layer for each TF. These activation sequences were aligned together to obtain the learned motif represented by the position weight matrix for each TF. The results showed that for 25 TFs with a motif recorded in the JASPAR database (49,50), the deep learning models revealed almost identical motifs (Supplementary Table S3). The similarities between motifs were compared by Tomtom (51) using the Pearson correlation coefficient. The median Tomtom *P*-value was 2.60e-6. Figure 4A illustrates several motifs that were learned by the deep CNN models and the corresponding motifs recorded in the JASPAR database. All the filters in the convolution layer were randomly initialized. Therefore, all the motifs were automatically learned *de novo* during model training. We tried to re-initialize all the model parameters randomly and re-trained all the models. Almost identical motifs were obtained as above for each TF, in terms of motif similarity. These results demonstrated that our models were robust and could accurately capture the key signatures of each TF.

Furthermore, the deep learning models revealed the binding motifs for some TFs that are not annotated in the JAS-
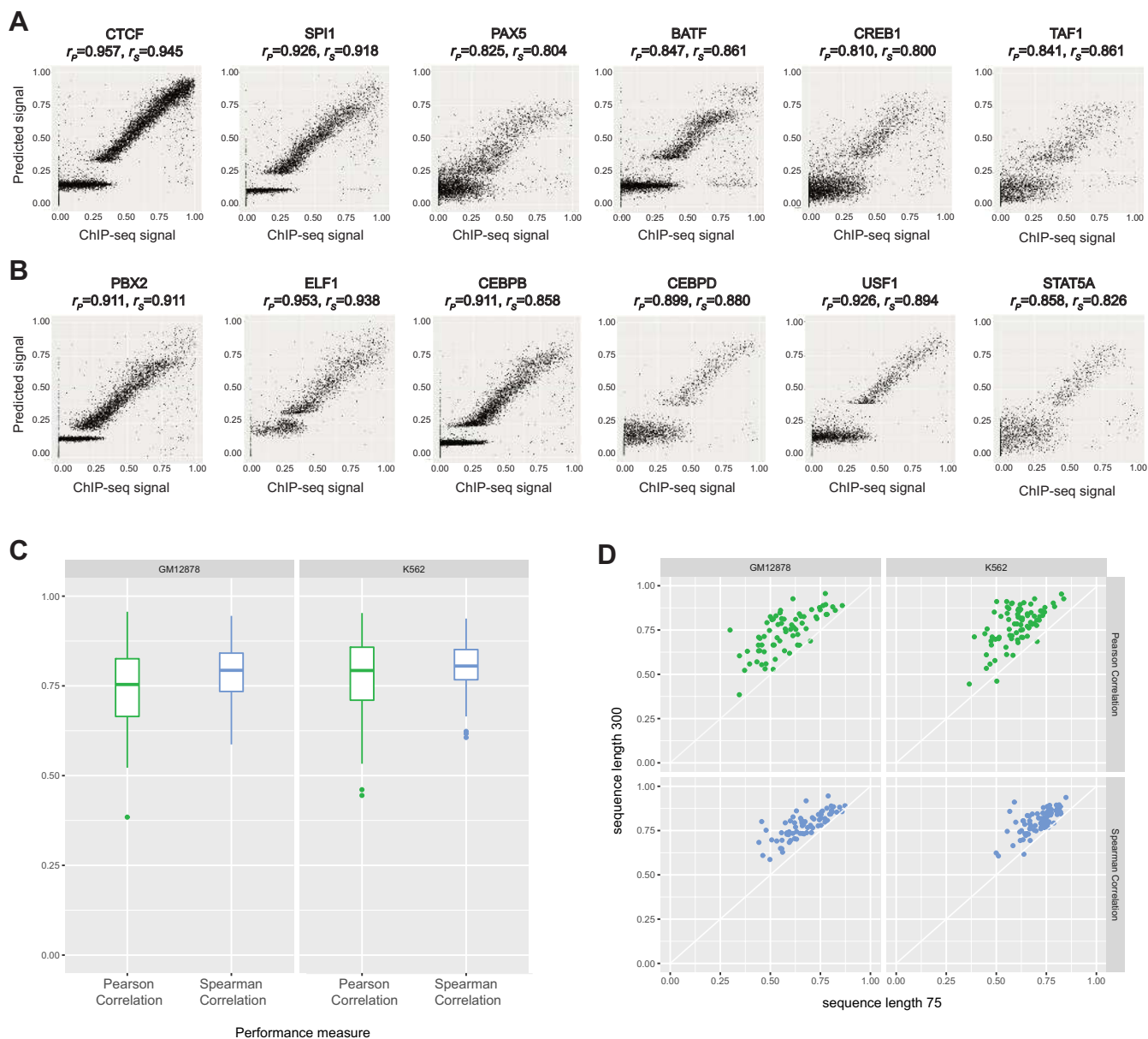
Figure 2. The *in silico* deep CNN models of DeFine accurately predicted the *in vivo* ChIP-seq signal intensities for the binding sequences of each TF in the GM12878 and K562 cell lines. (**A**) Examples of the correlation between the normalized ChIP-seq experimental signal values (*x*-axis) and the predicted signal values (*y*-axis) for TFs in the GM12878 cell line. r_P—Pearson correlation coefficient, r_S—Spearman correlation coefficient. (**B**) Examples of the correlation between the normalized ChIP-seq experimental signal values (*x*-axis) and the predicted signal values (*y*-axis) for TFs in the K562 cell line. (**C**) Summary of the performance of the deep CNN models for all collected sequence-specific TFs in the GM12878 and K562 cell lines. The performance was measured by both Pearson correlation (green) and Spearman correlation (blue) on test sets for each TF. (**D**) Model performance comparison of different input sequence lengths. The deep CNN models using 300-bp input sequences (*y*-axis) had higher performance than models using 75-bp input sequences (*x*-axis), suggesting that the sequence context contributed to accurate prediction of the TF binding intensities. The performance was measured by both Pearson correlation (green) and Spearman correlation (blue) on test sets for each TF in the GM12878 and K562 cell lines.

PAR database (Figure 4B). The *de novo* discovered motifs of NR2F2, RCOR1, STAT5A and TAL1 were all matched to the canonical GATA factor binding motif (Tomtom *P*-value <1e-4), suggesting a cooperative binding or interfering binding of these factors and GATA factors. The motif of SMAD1 matched the motif of the NFIC::TLX1 complex (Tomtom *P*-value 3.59e-5), indicating cooperative or interfering binding of these factors. The motif of CHD2 matched the motif of ZBTB33 (Tomtom *P*-value 9.50e-

7). Since CHD2 does not have a DNA binding domain, this result suggested that CHD2 might be a cofactor of ZBTB33 (52). The deep learning model incorporated all the sequences from the ChIP-seq data to learn the binding motif rather than simply utilizing the top hundreds of sequences as traditional motif discovery methods which have very high computational complexity. This capability enabled improved detection of TF binding motifs and aided the discovery of novel motifs.
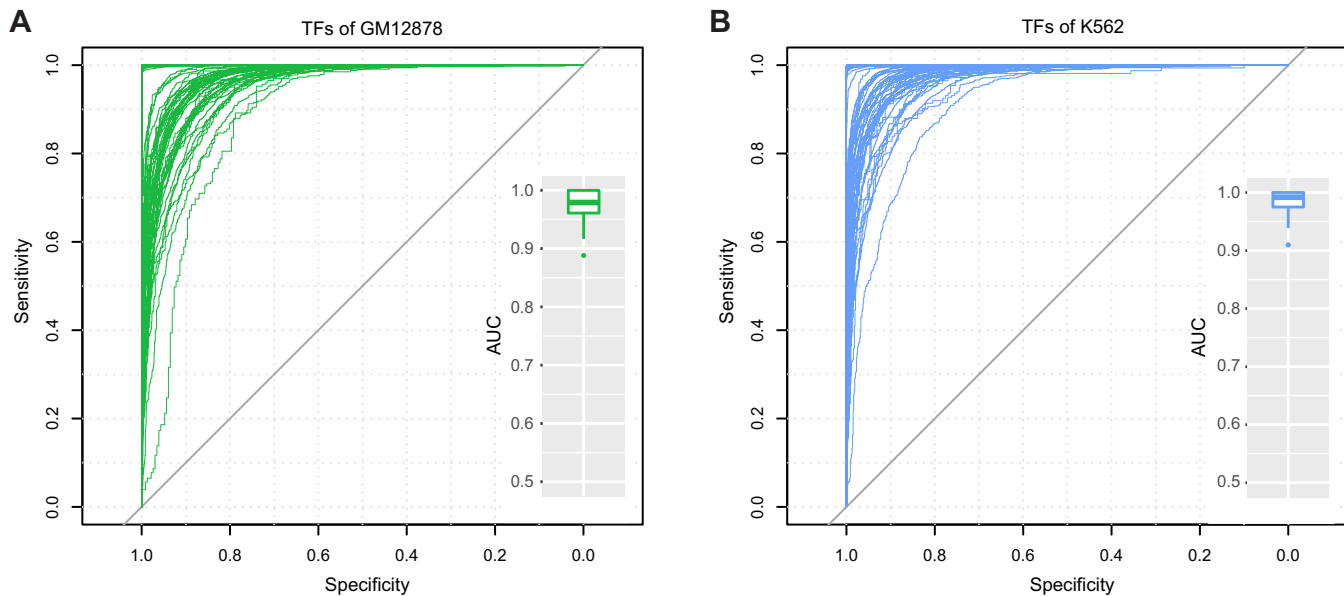
Figure 3. Performance evaluation of the deep CNN classification models to classify TF-DNA binding or unbinding. (**A**) AUCs of deep CNN classifiers for each of the 69 TFs in the GM12878 cell line. (**B**) AUCs of deep CNN classifiers for each of the 79 TFs in the K562 cell line.

DeFine functional scores helped classify disease-related and neutral non-coding variants

Based on the deep CNN model for each TF, the DeFine functional score for a variant was derived as the change between the predicted binding intensities of the altered sequence centered at the variant and that of the reference sequence (Figure 1B). We investigated the utility of DeFine functional scores for discriminating disease-related non-coding variants from neutral ones. To evaluate the performance of the DeFine functional score-based method and compare it with other tools developed for assessing non-coding variants, three different types of datasets with balanced positive and negative samples were compiled: HGMD-based, GWAS-based and eQTL-based (see 'Materials and Methods' section). The HGMD-based dataset was composed of 3360 pathogenic regulatory variants in the HGMD professional database (14) as a positive set and an equal number of randomly selected non-coding variants from the 1000 Genomes Project (53,54) as a negative set. The GWAS-based dataset consisted of 7602 GWAS catalog (6) significant variants located in the intergenic region as positive non-coding variants and an equal number of randomly selected intergenic variants from the 1000 Genomes project with an allele frequency distribution that was equivalent to the neutral variants. The functional eQTL-based dataset was compiled from the 13196 best-associated eQTL in LCLs from the EUR population (5) and a corresponding neutral set containing an equal number of randomly selected variants from the 1000 Genomes project with a matched allele frequency distribution. Both SNPs and indels were included in the three datasets, and the number of variants of the two variant types was matched in the positive set and negative set.

Taking the DeFine functional scores for all the TFs in each cell type as features, we trained GBDT-based classi-

fiers to predict each non-coding variant as functional or neutral in these three datasets. The GERP_NR, GERP_RS (41), PhyloP (42) and PhastCons (43) conservation scores were also incorporated as predictive features in our GBDT classifiers. The GBDT classifier based on the DeFine-predicted binding intensities was named DeFine-regression in the following analysis. The GBDT classifier based on the modified deep CNN model that predicted the binary outcome of binding/unbinding was named DeFine-classification. We also built a GBDT classifier that utilized both the regression version and the classification version of the DeFine scores, which was termed DeFine-combine in the following analysis. The performance was evaluated using 10-fold cross validation and compared to CADD (27), DeepSEA (24) and GWAVA (28) (see 'Materials and Methods section'). For performance comparison in HGMD, GWAVA was not included because it was trained with HGMD regulatory variants.

ROC analysis showed that the classifiers based on DeFine scores (DeFine-combine, DeFine-regression and DeFine-classification) outperformed other tools in the three different datasets in terms of the AUC (Figure 5 and Supplementary Figure S2). DeFine-combine, DeFine-regression and DeFine-classification showed almost identical performance for all evaluations. The pathogenic regulatory variants in the HGMD database were genetically or experimentally validated functional variants (14). For the HGMD-based dataset, DeFine-combine achieved the highest AUC of 0.847 using DeFine scores from the GM12878 cell line (Figure 5A), and the AUC resulting from DeFine scores from the K562 cell line was 0.851 (Supplementary Figure S2A). Precision-Recall curves also showed that classifiers based on DeFine scores outperformed CADD and DeepSEA on the HGMD-based dataset (Supplementary Figure S3). For GWAS-based and eQTL-based datasets, DeFine-based classifiers also had higher AUCs than the
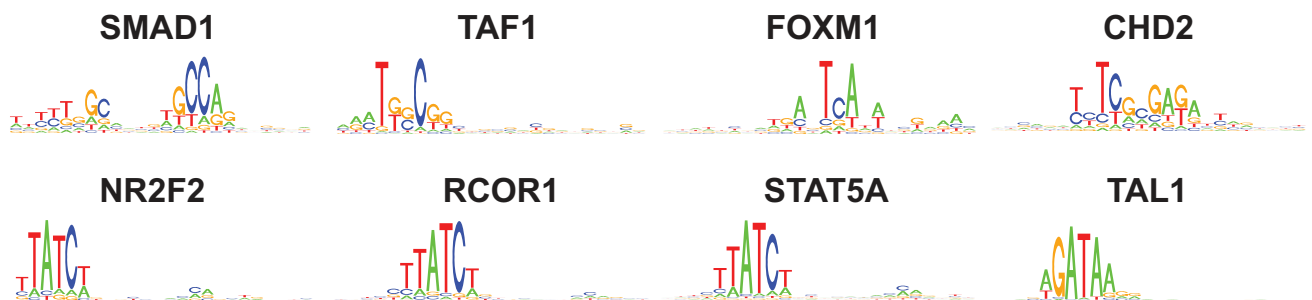
A

B

Figure 4. TF binding motifs revealed by the deep CNN models. (**A**) Examples of TF motifs learned by the deep CNN models and compared with the motifs recorded in JASPAR. The sequence signatures automatically learned by the deep CNN models of DeFine captured the known TF binding motifs. (**B**) The motifs discovered *de novo* by the deep CNN models for TFs that were not annotated in the JASPAR database.

other tools. It is noteworthy that the positive variants in GWAS and eQTL were disease and trait-associated and thus probably not the causal variants (55,56). The real functional and causal variants may be in linkage disequilibrium with the associated variants. Thus, in these two datasets, most of the positive variants were likely to be neutral variants. As expected, the AUCs of all the tools were relatively low in the GWAS-based and eQTL-based datasets compared with those in the HGMD-based dataset. This phenomenon was further illustrated when employing the region-restricted evaluation, which restricted the randomly selected neutral variants in each dataset to be within 5 kb of the positive variants (Supplementary Figure S4). Never-

theless, the classifier based on DeFine functional scores also outperformed the other tools in all region-restricted evaluations (Supplementary Figure S4).

To demonstrate that DeFine functional scores could help improve the performance of classifying disease-related non-coding variants from neutral ones, we trained GBDT-based classifiers with only the four conservation scores (GERP_NR, GERP_RS, PhyloP and PhastCons) on each of the three datasets. The cross validation results on the HGMD-based dataset revealed that the classifier with only conservation scores achieved the AUC of 0.779, which was much lower than the AUC of the classifier using DeFine functional scores (AUC: 0.847) on this dataset. The AUCs
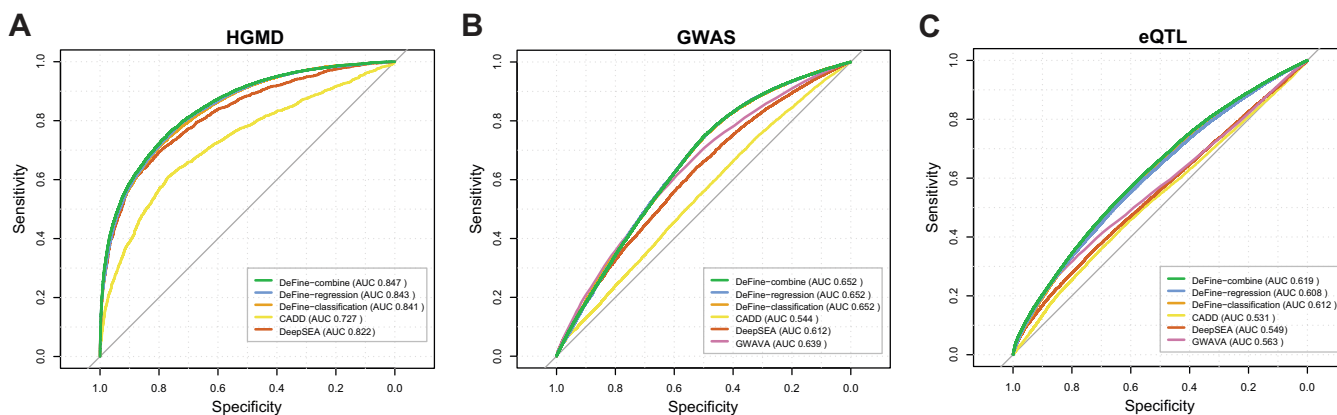
Figure 5. Performance evaluation of GBDT classifiers based on DeFine scores for the GM12878 cell line and comparison with CADD, DeepSEA and GWAVA. The DeFine-regression represented the GBDT classifier based on DeFine-predicted binding intensities. The DeFine-classification was the GBDT classifier based on the modified deep CNN model that predicted the binary outcome of binding/unbinding. DeFine-combine was the GBDT classifier based on both the regression version and the classification version of the DeFine scores. ROC analysis was performed on three different test sets: HGMD-based (**A**), GWAS-based (**B**) and eQTL-based (**C**). For the evaluation on HGMD-based test set, GWAVA was not included because it was trained on HGMD.

evaluated on GWAS-based and eQTL-based datasets were 0.645 and 0.524 for the classifiers using only conservation scores. These results showed that the classifiers integrating DeFine functional scores outperformed the classifiers using only conservation information.

Furthermore, we also evaluated the performance of the GBDT classifiers trained using HGMD-based variants to classify variants in the GWAS-based dataset. Overlapped variants were removed from the GWAS-based dataset. Comparison of the results showed that the classifier based on DeFine scores and trained on HGMD variants also outperformed the other tools in terms of AUCs in the GWAS-based dataset (Supplementary Figure S5). This result demonstrated the generalizability of the classifiers. All the above results revealed that DeFine functional scores were able to help discriminate functional non-coding variants from neutral variants.

DeFine was capable of prioritizing disease-related functional non-coding variants

We investigated whether DeFine functional scores could help identify disease-related functional non-coding variants from a list of candidates. We employed the list of non-coding mutations screened in a family-based association study of Hirschsprung disease (HSCR) (45). This list of candidate functional HSCR-associated variants consisted of 12 SNPs located in the 5′ *RET* gene and an enhancer element in the first intron of *RET*, which included rs2435357. The functional impact of rs2435357 on the expression level of the *RET* gene was experimentally validated, and it has been demonstrated to be the disease-causing mutation (45,57). We prioritized this list using the DeFine functional score together with CADD and DeepSEA. The results showed that DeFine successfully predicted rs2435357 with the highest functional score compared with the other variants in the list (Supplementary Table S4). In contrast, neither CADD nor DeepSEA showed actual functional variants in the list (Supplementary Table S4). rs2435357 was ranked sixth by CADD and fourth by DeepSEA. Moreover, DeFine also

correctly indicated an impact of this causal mutation on the *RET* gene based on the enhancer-gene interaction map integrated in DeFine.

We next leveraged DeFine, CADD and DeepSEA to rank the common variants in the 14q22.2 genomic region that were genotyped in a targeted association study of colorectal cancer risk (46). Thirteen common variants, including rs4444235, were screened in a large cohort in this study. Among all the variants that were significantly associated with colorectal cancer risk, rs4444235 was the only one that was experimentally demonstrated to affect the enhancer activity and showed *cis*-regulation of *BMP4* gene expression (46). These *in silico* tools were employed to predict the regulatory functional impact of the 13 variants in the screened list. Prioritization results of DeFine revealed that rs4444235 had the greatest positive influence on the regulatory activity among the 13 variants, whereas CADD and DeepSEA did not predict this to be the most likely functional variant (Supplementary Table S5). Again, DeFine identified the correct target gene, *BMP4* of rs4444235, which was located ∼12 Kb downstream of the *BMP4* promoter. These results illustrated the utility of DeFine for prioritizing functional non-coding variants.

DISCUSSION

In this paper, we demonstrated that real-valued *in vivo* DNA binding intensities from TFs ChIP-seq were accurately predicted from raw sequences by our deep learning models, which allowed us to precisely measure the impact of variants on TF binding intensities. The parallel work Basenji (BioRxiv: https://doi.org/10.1101/161851) also proposed using deep learning model to predict the fine resolution quantitative genomic profiles rather than binary profiles, and showed better performance for predicting gene expression from DNA sequences. The high performance of our deep learning model was based on its ability to automatically extract sequence signatures, capture TF binding motifs and integrate the sequence context. In addition, both the forward sequence and the reverse complementary se-

quence were simultaneously considered in our deep learning models with shared filters in the CNN layer. We corrected the training sequences by incorporating cell type-specific genomic variants to reflect the actual genome sequences in that cell line rather than using reference genome sequences directly for training. This step was ignored by the available sequence-based models. These features of our model enabled us to develop the new tool DeFine to assess the functional impact of non-coding variants with increased accuracy. We showed that the deep CNN models in DeFine accurately learned the known binding motifs of TFs *de novo* during model training, suggesting that the high performance of DeFine resulted from its ability to capture *de facto* sequence features affecting TF binding intensities rather than learned batch effects of the ChIP-seq experiments or other systematic differences in the ChIP-seq data.

The DeFine functional scores, which represent differences in the model-predicted TF binding intensities between the reference sequence and the altered sequence, reflect the degree of functional impact of the variants. The DeFine functional scores can help discriminate disease-related noncoding variants from neutral ones. The classifiers employing DeFine functional scores outperformed other state-of-the-art tools in different test sets. The DeFine functional scores could facilitate prioritization of disease-related functional non-coding variants among a list of candidates. DeFine is capable of evaluating all kinds of variants, including SNPs and indels, with single base resolution. Rather than regulating nearby genes, regulatory elements commonly interact with distal target gene(s) that may be kilobases or megabases away (31,32). By integrating 3D genome contact maps, DeFine is able to not only predict the functional noncoding variants but also indicate the potential target genes affected by the variants.

The complexity of the deep learning model makes it predisposed to overfit (21). To avoid this phenomenon, we augmented the training data by the addition of randomly selected sequences from regions of the genome without TF binding. In addition, we added batch normalization and dropout layer in our models to improve their generalization capability. Furthermore, weight decay (regularization) and early stopping were performed during the model training process. These techniques together helped to eliminate overfitting of our deep learning models. Moreover, we utilized randomization to control for overfitting in our evaluations by randomly composing multiple neutral variant sets in each test set and averaged the performances in each evaluation.

Training the deep learning models in DeFine for one cell type depends on the availability of high-throughput ChIP-seq profiling of dozens of TFs in that cell type. Currently, large-scale TF ChIP-seq experiments are available for a limited number of cell types. We trained deep learning models for GM12878 and K562, which are tier 1 cell lines in the ENCODE project, and large amounts of TF ChIP-seq data are publicly available. The performance of the deep learning models in DeFine depends on the quality of the ChIP-seq experiments. When the experimental signal values are prone to noise or are too weak to be measured precisely, accurately prediction for the deep learning models are difficult. Training of the deep learning models requires many sequences.

For TFs that bind to only a few positions in the genome, the deep learning models may not be reliably trained. The DeFine functional score measures the impact of non-coding variants on TF binding. The sequence context outside the TF binding motif still plays an important role in determining the functionality of a variant, by affecting nucleosome positioning, histone modification and chromosome conformation etc. The DeFine functional score is not a direct pathogenicity measurement for non-coding variants. Nevertheless, as we showed in extensive evaluations, it could help evaluate and discriminate disease-related non-coding variants from neutral ones.

We used reference guided local reassembly to create cell type personalized genome sequences to train our deep learning models. We believe this step is necessary in sequence-based models to train the models with actual sequences in each cell line, especially for models with single base resolution. The quantitative measure of the impact on the performance of functional variants prioritization when using the cell type-specific genome sequences rather than the reference genome is worth to explore with proper test data in further research.

DeFine provides a framework for constructing cell type-specific deep learning models to assess the functional impact of abundant non-coding variants across the whole human genome. The deep CNN model in DeFine could be trained on any TF ChIP-seq dataset in any cell/tissue type with available data. With advances in functional genomics technologies and the accumulation of genomic profiling data, the deep learning models in DeFine would be available for an increasing number of cell types. As a result, the performance of DeFine will continue to improve. The high performance of DeFine would contribute to solving the challenge of interpreting the large-scale non-coding variants identified in genetics studies and clinical genetics. DeFine would facilitate the high-throughput evaluation and prioritization of non-coding variants on a genome-wide scale.

DATA AVAILABILITY

DeFine is freely available at http://define.cbi.pku.edu.cn. DeFine supports prediction through the online service, or it can be installed and run locally. DeFine provides an easy-to-use web interface and command line. The entire source code can be downloaded at the DeFine website.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

FUNDING

REFERENCES

1. Yanez-Cuna,J.O., Kvon,E.Z. and Stark,A. (2013) Deciphering the transcriptional cis-regulatory code. *Trends Genet.*, **29**, 11–22.
2. Levo,M. and Segal,E. (2014) In pursuit of design principles of regulatory sequences. *Nat. Rev. Genet.*, **15**, 453–468.
3. Ong,C.T. and Corces,V.G. (2011) Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat. Rev. Genet.*, **12**, 283–293.
4. Vaquerizas,J.M., Kummerfeld,S.K., Teichmann,S.A. and Luscombe,N.M. (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.*, **10**, 252–263.
5. Lappalainen,T., Sammeth,M., Friedlander,M.R., t Hoen,P.A., Monlong,J., Rivas,M.A., Gonzalez-Porta,M., Kurbatova,N., Griebel,T., Ferreira,P.G. *et al.* (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, **501**, 506–511.
6. Welter,D., MacArthur,J., Morales,J., Burdett,T., Hall,P., Junkins,H., Klemm,A., Flicek,P., Manolio,T., Hindorff,L. *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42**, D1001–D1006.
7. Maurano,M.T., Humbert,R., Rynes,E., Thurman,R.E., Haugen,E., Wang,H., Reynolds,A.P., Sandstrom,R., Qu,H., Brody,J. *et al.* (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science*, **337**, 1190–1195.
8. Albert,F.W. and Kruglyak,L. (2015) The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.*, **16**, 197–212.
9. Weinhold,N., Jacobsen,A., Schultz,N., Sander,C. and Lee,W. (2014) Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat. Genet.*, **46**, 1160–1165.
10. Khurana,E., Fu,Y., Chakravarty,D., Demichelis,F., Rubin,M.A. and Gerstein,M. (2016) Role of non-coding sequence variants in cancer. *Nat. Rev. Genet.*, **17**, 93–108.
11. Epstein,D.J. (2009) Cis-regulatory mutations in human disease. *Brief. Funct. Genomics Proteomics*, **8**, 310–316.
12. Visel,A., Blow,M.J., Li,Z., Zhang,T., Akiyama,J.A., Holt,A., Plajzer-Frick,I., Shoukry,M., Wright,C., Chen,F. *et al.* (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, **457**, 854–858.
13. Badis,G., Berger,M.F., Philippakis,A.A., Talukder,S., Gehrke,A.R., Jaeger,S.A., Chan,E.T., Metzler,G., Vedenko,A., Chen,X. *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science*, **324**, 1720–1723.
14. Stenson,P.D., Mort,M., Ball,E.V., Shaw,K., Phillips,A. and Cooper,D.N. (2014) The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.*, **133**, 1–9.
15. Ng,P.C. and Henikoff,S. (2001) Predicting deleterious amino acid substitutions. *Genome Res.*, **11**, 863–874.
16. Adzhubei,I.A., Schmidt,S., Peshkin,L., Ramensky,V.E., Gerasimova,A., Bork,P., Kondrashov,A.S. and Sunyaev,S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
17. Wang,M. and Wei,L. (2016) iFish: predicting the pathogenicity of human nonsynonymous variants using gene-specific/family-specific attributes and classifiers. *Sci. Rep.*, **6**, 31321.
18. Encode Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
19. Roadmap Epigenomics Consortium, Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
20. Rao,S.S., Huntley,M.H., Durand,N.C., Stamenova,E.K., Bochkov,I.D., Robinson,J.T., Sanborn,A.L., Machol,I., Omer,A.D., Lander,E.S. *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.
21. LeCun,J., Bengio,Y. and Hinton,G. (2015) Deep learning. *Nature*, **521**, 436–444.
22. Angermueller,C., Pärnamaa,T., Parts,L. and Stegle,O. (2016) Deep learning for computational biology. *Mol. Syst. Biol.*, **12**, 878.
23. Alipanahi,B., Delong,A., Weirauch,M.T. and Frey,B.J. (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, **33**, 831–838.
24. Zhou,J. and Troyanskaya,O.G. (2015) Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods*, **12**, 931–934.
25. Kelley,D.R., Snoek,J. and Rinn,J.L. (2016) Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.*, **26**, 990–999.
26. Quang,D. and Xie,X. (2016) DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.*, **44**, e107.
27. Kircher,M., Witten,D.M., Jain,P., O'Roak,B.J., Cooper,G.M. and Shendure,J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.*, **46**, 310–315.
28. Ritchie,G.R., Dunham,I., Zeggini,E. and Flicek,P. (2014) Functional annotation of noncoding sequence variants. *Nat. Methods*, **11**, 294–296.
29. Fu,Y., Liu,Z., Lou,S., Bedford,J., Mu,X., Yip,K.Y., Khurana,E. and Gerstein,M. (2014) FunSeq2: A framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.*, **15**, 480.
30. Schmidt,D., Wilson,M.D., Ballester,B., Schwalie,P.C., Brown,G.D., Marshall,A., Kutter,C., Watt,S., Martinez-Jimenez,C.P., Mackay,S. *et al.* (2010) Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science*, **328**, 1036–1040.
31. Sanyal,A., Lajoie,B.R., Jain,G. and Dekker,J. (2012) The long-range interaction landscape of gene promoters. *Nature*, **489**, 109–113.
32. Zhang,Y., Wong,C.H., Birnbaum,R.Y., Li,G., Favaro,R., Ngan,C.Y., Lim,J., Tai,E., Poh,H.M., Wong,E. *et al.* (2013) Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. *Nature*, **504**, 306–310.
33. Sergey Ioffe,C.S. (2015) Batch Normalization: Accelerating deep network training by reducing internal covariate shift. *Proc. Mach. Learn. Res.*, **37**, 448–456.
34. Srivastava,N., Hinton,G., Krizhevsky,A., Sutskever,I. and Salakhutdinov,R. (2014) Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, **15**, 1929–1958.
35. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
36. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
37. Kharchenko,P.V., Tolstorukov,M.Y. and Park,P.J. (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.*, **26**, 1351–1359.
38. Landt,S.G., Marinov,G.K., Kundaje,A., Kheradpour,P., Pauli,F., Batzoglou,S., Bernstein,B.E., Bickel,P., Brown,J.B., Cayting,P. *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.
39. Walker,B.J., Abeel,T., Shea,T., Priest,M., Abouelliel,A., Sakthikumar,S., Cuomo,C.A., Zeng,Q., Wortman,J., Young,S.K. *et al.* (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, **9**, e112963.
40. Worsley Hunt,R., Mathelier,A., Del Peso,L. and Wasserman,W.W. (2014) Improving analysis of transcription factor binding sites within ChIP-Seq data based on topological motif enrichment. *BMC Genomics*, **15**, 472.
41. Davydov,E.V., Goode,D.L., Sirota,M., Cooper,G.M., Sidow,A. and Batzoglou,S. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.*, **6**, e1001025.
42. Cooper,G.M., Stone,E.A., Asimenos,G., Program,N.C.S., Green,E.D., Batzoglou,S. and Sidow,A. (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.*, **15**, 901–913.

43. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.

44. Davis,J. and Goadrich,M. (2006) The relationship between Precision-Recall and ROC curves. In: *Proceedings of the 23rd international conference on Machine learning*. ACM, Pittsburgh, pp. 233–240.

45. Emison,E.S., McCallion,A.S., Kashuk,C.S., Bush,R.T., Grice,E., Lin,S., Portnoy,M.E., Cutler,D.J., Green,E.D. and Chakravarti,A. (2005) A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. *Nature*, **434**, 857–863.

46. Lubbe,S.J., Pittman,A.M., Olver,B., Lloyd,A., Vijayakrishnan,J., Naranjo,S., Dobbins,S., Broderick,P., Gómez-Skarmeta,J.L. and Houlston,R.S. (2011) The 14q22.2 colorectal cancer variant rs4444235 shows cis-acting regulation of BMP4. *Oncogene*, **31**, 3777–3784.

47. Furey,T.S. (2012) ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat. Rev. Genet.*, **13**, 840–852.

48. Bailey,T., Krajewski,P., Ladunga,I., Lefebvre,C., Li,Q., Liu,T., Madrigal,P., Taslim,C. and Zhang,J. (2013) Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS Comput. Biol.*, **9**, e1003326.

49. Sandelin,A., Alkema,W., Engstrom,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.

50. Mathelier,A., Fornes,O., Arenillas,D.J., Chen,C.Y., Denay,G., Lee,J., Shi,W., Shyr,C., Tan,G., Worsley-Hunt,R. *et al.* (2016) JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **44**, D110–D115.

51. Gupta,S., Stamatoyannopoulos,J.A., Bailey,T.L. and Noble,W.S. (2007) Quantifying similarity between motifs. *Genome Biol.*, **8**, R24.

52. Wang,J., Zhuang,J., Iyer,S., Lin,X., Whitfield,T.W., Greven,M.C., Pierce,B.G., Dong,X., Kundaje,A., Cheng,Y. *et al.* (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.*, **22**, 1798–1812.

53. The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

54. The 1000 Genomes Project Consortium. (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

55. Lewis,C.M. (2002) Genetic association studies: design, analysis and interpretation. *Brief. Bioinformatics*, **3**, 146–153.

56. Gilad,Y., Rifkin,S.A. and Pritchard,J.K. (2008) Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet.*, **24**, 408–415.

57. Emison,E.S., Garcia-Barcelo,M., Grice,E.A., Lantieri,F., Amiel,J., Burzynski,G., Fernandez,R.M., Hao,L., Kashuk,C., West,K. *et al.* (2010) Differential contributions of rare and common, coding and noncoding Ret mutations to multifactorial Hirschsprung disease liability. *Am. J. Hum. Genet.*, **87**, 60–74.