EnContact: predicting enhancer-enhancer contacts using sequence-based deep learning model

Mingxin Gan[1,*], Wenran Li[2,3,4,*] and Rui Jiang[2]

[1] Donlinks School of Economics and Management, University of Science and Technology Beijing, Beijing, China
[2] MOE Key Laboratory of Bioinformatics; Bioinformatics Division and Center for Synthetic and Systems Biology, BNRist; Department of Automation, Tsinghua University, Beijing, China
[3] Department of Statistics, Stanford University, Stanford, CA, USA
[4] Department of Biomedical Data Science, Stanford University, Stanford, CA, USA
* These authors contributed equally to this work.

ABSTRACT

Chromatin contacts between regulatory elements are of crucial importance for the interpretation of transcriptional regulation and the understanding of disease mechanisms. However, existing computational methods mainly focus on the prediction of interactions between enhancers and promoters, leaving enhancer-enhancer (E-E) interactions not well explored. In this work, we develop a novel deep learning approach, named Enhancer-enhancer contacts prediction (EnContact), to predict E-E contacts using genomic sequences as input. We statistically demonstrated the predicting ability of EnContact using training sets and testing sets derived from HiChIP data of seven cell lines. We also show that our model significantly outperforms other baseline methods. Besides, our model identifies finer-mapping E-E interactions from region-based chromatin contacts, where each region contains several enhancers. In addition, we identify a class of hub enhancers using the predicted E-E interactions and find that hub enhancers tend to be active across cell lines. We summarize that our EnContact model is capable of predicting E-E interactions using features automatically learned from genomic sequences.

Submitted 14 May 2019
Accepted 10 August 2019
Published 13 September 2019

Corresponding author
Rui Jiang, ruijiang@tsinghua.edu.cn

Academic editor
Yong Wang

Additional Information and Declarations can be found on page 15

DOI 10.7717/peerj.7657

Subjects Bioinformatics, Computational Biology
Keywords Deep learning, HiChIP data, Attention-based RNN, Hub enhancers, Enhancer-enhancer contacts

INTRODUCTION

Chromatin contacts between regulatory elements are widely studied to interpret the regulation relationship of transcriptome and to understand the regulatory mechanism of complex diseases. Chromosome conformation capture (3C)-based methods, including 4C and 5C, have been developed to detect physical contacts on a local scale (*Dekker et al., 2002*; *Simonis et al., 2006*; *Dostie et al., 2006*). Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) captures chromatin interactions related to a protein of interest (*Fullwood et al., 2009*). Recently, Hi-C, Capture Hi-C, and HiChIP techniques allow genome-wide detection of interactions between all possible pairs of regions (*Rao et al., 2014*; *Mifsud et al., 2015*; *Mumbach et al., 2016*), which provides the

comprehensive landscape of three-dimension chromatin structure. However, all of these techniques require an extremely deep sequencing depth to achieve high resolution, which can hardly be applied to a large number of cell lines. Therefore, computational approaches are needed to help with the identification of finer-mapping interactions.

In the past 5 years, a series of methods have been developed to predict promoter-related interactions (*Roy et al., 2015*; *Whalen, Truty & Pollard, 2016*; *Li, Wong & Jiang, 2019*; *Cao et al., 2017*; *Schreiber et al., 2018*; *Singh et al., 2016*). *Roy et al. (2015)* implemented regulatory interaction prediction for promoters and long-range enhancers which integrates published 3C data sets with a minimal set of regulatory genomic data sets to predict enhancer-promoter interactions in a cell line-specific manner. *Whalen, Truty & Pollard (2016)* proposed a computational method, TargetFinder, to predict promoter-enhancer interactions from diverse features along the genome. With the understanding that there exist enhancer-like promoters which can regulate distal genes (*Diao et al., 2017*; *Gasperini et al., 2017*), *Li, Wong & Jiang (2019)* developed a deep learning model, DeepTACT, which pays equal attention to the prediction of promoter-enhancer interactions and promoter-promoter interactions. However, all of these methods only focused on the prediction of promoter-enhancer interactions and promoter-promoter interactions.

Over the last decade, studies have demonstrated the importance of enhancer-enhancer (E-E) interactions in the interpretation of gene regulation (*Ghavi-Helm et al., 2014*; *Ing-Simmons et al., 2015*; *Kumasaka, Knights & Gaffney, 2019*; *Mumbach et al., 2017*). *Ghavi-Helm et al. (2014)* generated a high-resolution map of enhancer 3D contacts during Drosophila embryogenesis and found each enhancer contacts multiple enhancers and promoters with similar expression, suggesting a role of co-regulation of enhancers. *Ing-Simmons et al. (2015)* reported that ~50% of deregulated genes in mouse thymocytes reside in the vicinity of enhancer elements, suggesting that gene expression was regulated through E-E interactions. *Kumasaka, Knights & Gaffney (2019)* detected frequent long-range interactions between enhancers, which were further used to interpret gene regulation and disease causality. Although studies have shown that E-E loops are closely related to gene expression by co-regulation, to the best of our knowledge, hardly any computational approach is specifically designed to provide a powerful prediction of E-E contacts.

In computational biology filed, genomic sequence information is widely used to extract motif-like patterns to predict regulatory elements (e.g., enhancers) using various statistical models (*Le et al., 2019*; *Kleftogiannis, Ashoor & Bajic, 2018*). Recently, deep learning has shown impressive performance in pattern recognition and imaging field (*Shen et al., 2017*; *Wang, Sun & Wang, 2017*). By formatting genomic sequences and other epigenomic information into numeric vectors or matrix, deep learning models have achieved great success in many biological problems such as the identification of transcription factor (TF) binding sites, the recognition of regulatory regions, and the prediction of interacting pairs (*Li, Wong & Jiang, 2019*; *Kelley, Snoek & Rinn, 2016*; *Alipanahi et al., 2015*; *Park & Kellis, 2015*; *Le, Ho & Ou, 2017*, *2018*; *Le & Nguyen 2019*). These applications demonstrate the ability of deep learning models in capturing useful genomic features and accurately predicting biological signals. This inspires us to design a deep learning model to predict E-E interactions by learning motif-like feature patterns from genomic sequences using deep neuron networks.

Table 1 Number of enhancer-enhancer interactions collected from HiChIP.

Cell type	Acronym	1v1	mvm
GM12878	GM	143,810	144,380
K562	K562	158,058	142,131
Human coronary artery smooth muscle	HCASMC	110,078	114,758
CD4+ T cell leukemia	MyLa	124,858	123,163
Naïve CD4+ T cells	Naive	128,450	146,308
T helper 17 cells	Th17	145,706	178,600
T regulatory cells	Treg	112,392	125,625
Total	–	923,352	974,965

In this paper, we develop a deep learning approach, named Enhancer-enhancer contacts prediction (EnContact), to predict E-E contacts only using genomic sequences. Through a series of comprehensive experiments in seven cell lines, we demonstrate that EnContact is able to learn context-specific features from genomic sequences and to predict E-E interactions accurately. When applying the context-specific models to HiChIP data, EnContact identifies finer-mapping E-E interactions from region-based chromatin contacts, where each region contains several enhancers. In addition, EnContact identifies a class of hub enhancers which are active across different cell lines. In summary, EnContact achieves much better performance compared with traditional machine learning methods and is capable of identifying E-E interactions from HiChIP data.

MATERIALS AND METHODS

Collection and processing of datasets

We collected chromatin contact matrix of seven cell lines (i.e., GM, K562, HCASMC, MyLa, Naïve, Th17, and Treg) from HiChIP data of *Mumbach et al. (2017)*. Permissive enhancers were downloaded from FANTOM5 (*Andersson et al., 2014*) and then uniformly extended to two kb based on their middle sites for more information in their surrounding genomic sequence. In each cell line, we converted the contact matrix into chromatin interactions and annotated each bait-level interaction with permissive enhancers, resulting in a list of enhancer-related interactions. Then, those enhancer-related interactions were divided into two subsets: a "1v1" set which contains interaction pairs with only one enhancer in each end; a "mvm" set which consists of interaction pairs with more than one enhancer in either end or both ends (Table 1).

Chromatin interaction analysis by paired-end tag sequencing data of different cell lines were collected from *Tang et al. (2015)* and processed using a standard tool ChIA-PET2 (*Li et al., 2017*) with default settings, yielding 194,467 fragment-level interactions at a q value threshold 0.05. Then, we annotated these fragment-level interactions with enhancers using bedtools (*Quinlan & Hall, 2010*), resulting in a list of 37,894 E-E interactions as a validation set. DNase-seq experiments of above seven cell lines and ChIP-seq experiments of four histone marks (i.e., H3K4me3, H3K27ac, H3K4me2, H3K9ac) and 579 TFs were downloaded from ENCODE (*The ENCODE Project Consortium, 2004*). Detailed experimental information is shown in Tables S1–S3.

Structure of EnContact

The structure of EnContact can be divided into three parts: two convolution neuron networks (CNN) and a recurrent neural network (RNN). In each CNN, a convolution layer is used to learn motif-like patterns from genomic sequences, together with a rectifier operation (rectified linear unit (ReLU)) to propagate positive outputs and eliminate negative outputs. The convolution process can be devoted as

$$\mathrm{Conv}(X)_{ij} = \mathrm{ReLU}\left(\sum_{m=1}^{M}\sum_{n=1}^{N} w_{m,n}x_{i+m-1,\,j+n-1}\right), \tag{1}$$

where X is the one-hot encoding representation of sequences, $i = 1, 2, \cdots, L - M + 1$; $j = C - N + 1$. Here, L is the length of an enhancer (i.e., L = 2,000), C the number of nucleotides (i.e., C = 4). $W = (w_{mn})_{M \times N}$ is the weight matrix of a convolution kernel; $M \times N$ the size of the kernel. In our case, N is set to be 4, thus j is a constant 0. The activation function is the ReLU, defined as

$$\mathrm{ReLU}(z) = \max\{0, z\}. \tag{2}$$

Next, max-pooling layers are used to reduce dimensions and help extract higher-level features. The pooling process can be devoted as

$$\mathrm{MaxPooling}(X) = \max\left\{x_{ij}, x_{i+1,j}, \cdots, x_{i+W,j}\right\}, \tag{3}$$

where W is the size of pooling window.

Then, features learned by the above CNNs are concatenated using a merging layer, followed by a bidirectional long-short-term memory (BLSTM) layer to further learn the context features from pooled sequence patterns. As a typical representation of RNNs, BLSTM is widely used for its ability in capturing dependencies of sequences by accessing long-range context (*Graves, Jaitly & Mohamed, 2013*). To help BLSTM to pay more attention to specific sequence patterns, an attention layer is adopted in the integration module, following the BLSTM layer. The simplification of attention mechanism can be formulated as

$$\alpha_t = \frac{\exp\{f(h_t)\}}{\sum_{k=1}^{T}\exp\{f(h_k)\}} \tag{4}$$

where $f(\cdot) = \tanh(\cdot)$ can be considered as a learnable function depending on hidden layer h_t at time step t, which measures scalar importance for h_t. α_t is the weight computed at each time step t for each state $h_t, t = (1, 2, \cdots, T)$; T the number of time steps determined by the BLSTM layer. Given hidden states h_t, attention layer computes an adaptive weighted average of hidden states, θ, devoted as

$$\theta = \sum_{t=1}^{T}\alpha_t h_t. \tag{5}$$

The final layer of EnContact is a dense layer which is actually an array of hidden units with the ReLU activations feeding into a logistic regression (LR) unit that predicts the probability of interacting. In addition, we adopt batch normalization layers to accelerate the training process and dropout layers to avoid overfitting. Details for the parameters used in the deep learning model are described in Table S4. We implemented the EnContact model using Keras 1.2.0 (*Bastien et al., 2012*) on a Linux server. All experiments were carried out with 4 Nvidia K80 GPUs which significantly accelerated the training process than CPUs.

Generating negative cases based on the distance distribution

We generate negative cases based on the distance distribution of positive cases. Following existing literature (*Whalen, Truty & Pollard, 2016*), we first divide the distances between positive interaction pairs into five bins, guaranteeing each bin has an equal number of positive samples. Then, we generate negative cases within each distance bin, making sure that the number of negative cases in each bin is the same as that of positive cases.

Baseline methods

We use four baseline methods for comparison, including three typical classification models, SVM (*Wu, Lin & Weng, 2004*), LR (*Hosmer, Lemeshow & Sturdivant, 2013*), and random forest (RF) (*Liaw & Wiener, 2002*), and a deep learning model SPEID (*Singh et al., 2016*). For the typical classification models, to convert nucleotide-based information into numeric vectors, we extract k-mer features from genomic sequences using the following strategy. First, we list all combinations of four types of nucleotides (A, T, C, G) in k sites, resulting in 4^k motif-like patterns. Then, for the sequence of a given enhancer, we count the occurrence frequency of each k-mer pattern. Thus, we derive a feature vector consisting of the frequency of 4^k motif-like fragments for each enhancer. Since our goal is to predict the interaction of two enhancers, we connect the feature vectors of these two enhancers as the input for baseline methods.

We accomplish the classification process of SVM, LR, and RF using Scikit-learn package (*Pedregosa et al., 2011*) in Python. We downloaded the source code of SPEID from https://github.com/ma-compbio/SPEID and ran the model followed its instruction. Considering SPEID was designed to predict enhancer-promoter contacts, we substituted promoter-enhancer sequences to E-E sequences as the input for SPEID. The input samples and features are totally the same for SPEID and EnContact. To ensure a fair comparison, we provide the same training sets and testing sets in seven cell lines for baseline methods and our EnContact model.

Motif analysis

To convert the weights of convolution kernels learned by EnContact into probabilistic position weight matrix (PWM), we first calculate the activation scores of kernels for a given input (i.e., the sequence of an enhancer), as

$$S_i = \sum_{m=1}^{M} \sum_{n=1}^{N} w_{m,n} x_{i+m-1,n-1}, \tag{6}$$

where S_i is the activation score of the ith nucleotide of the input sequence. Then, we define a position as activated if its activation score is larger than half of the maximum value of the whole sequence, formulated as

$$\begin{cases} S_i > 0.5 * \text{MAXS} \\ \text{MAXS} = \max\{S_i | 1 \leq i \leq L\}' \end{cases} \tag{7}$$

where L is the length of the sequence; MAXS the maximum value of activation scores.

Then, we count the nucleotide occurrences of activated positions and format them into PWMs. After that, we match the resulting PWMs to known motifs derived from JASPAR database (*Mathelier et al., 2016*), which contains 1,082 motifs of 1,072 human TFs, using the tool TOMTOM v4.12.0 (*Gupta et al., 2007*) with a threshold of false discovery rate (FDR) q value < 0.1.

Definition of co-opening degree between two enhancers

We define the co-opening degree of two enhancers as the absolute value of Pearson's correlation coefficient of their openness scores. Here, the openness score of each enhancer is defined as follow. Suppose the number of replicates in a given cell line is R, the length of the regulatory element L, then the DNase signal of each regulatory element can be represented as $O^{R \times L}$, where O_{rl} $(r = 1, 2, \ldots, R;\ l = 1, 2, \ldots, L)$ is the chromatin accessibility score at each genomic site, defined as

$$O_{rl} = \frac{N_{rl}}{M_{rl}/W}, \tag{8}$$

where N is the number of reads falling at each genomic site; M the number of reads falling into a background window of length W (say, one Mb) surrounding this site. This fold change value O_{rl} is designed to remove the influence of sequencing depth according to *Li et al. (2017)*.

Then, for each enhancer, we obtain a vector of openness scores by averaging $O^{R \times L}$ across replicates R. Finally, the co-opening degree of two regulatory elements is defined as the absolute value of Pearson's correlation coefficient of two vectors of openness scores. If the P-value of a Pearson's correlation coefficient of an interaction pair is significant (i.e., P-value > 0.05), we consider this E-E pair as co-opening.

Activity of hub enhancers

For each peak i in experiment j, we define an activity score, PAS_{ij}, by calculating the fold change between the number of reads falling into this peak and the number of reads falling into a background region surrounding the peak, formulated as

$$\text{PAS}_{ij} = \frac{N_{ij}/P_{ij}}{M_{ij}/W}, \tag{9}$$

where W is the length of background window, defaulted as one Mb; P_{ij} the length of peak i in experiment j; N_{ij} the number of reads falling into peak i in experiment j; M_{ij} the number of reads falling into the background region of peak i in experiment j.

Then, we define the activity score of an enhancer as the maximum activity score of peaks overlapping with this enhancer,

$$\mathrm{EAS}_{ij} = \max\{\mathrm{PAS}_{kj} | k \in S_i\}, \tag{10}$$

where S_i is the set of peaks overlapping with enhancer i. Finally, we consider an enhancer as active when its activity score is greater than 1. For each hub enhancer or non-hub enhancer, we count the number of experiments where it is active to assess its activity across cell lines.

RESULTS

Design of EnContact model and training strategy

We developed a deep learning model, named EnContact, to identify E-E interactions using features learned from genomic sequences. As shown in Fig. 1A, we adopted a one-hot encoding strategy to convert the sequence of an enhancer into a four-dimensional matrix, where each genomic site has a four-element vector with the nucleotide bit set to be 1. For a given E-E pair, EnContact learns patterns from the encoded sequences of two enhancers using two separate convolutional neuron networks. Then, an attention-based RNN was applied to extract high-level features from the concatenation of patterns learned by CNNs. Finally, EnContact predicts the interacting probability of the given two enhancers using a LR layer (Fig. 1B). Details of EnContact model are shown in Methods.

To construct context-specific training data for EnContact, we developed the following strategy. First, we collected the chromatin interactions of seven cell lines from the HiChIP data of *Mumbach et al. (2017)*. Then, we annotated those interactions with permissive enhancers derived from FANTOM5 (*Andersson et al., 2014*) to obtain E-E interactions. As shown in Fig. 1C, the E-E interactions are further divided into two subsets: one subset consists of interactions that have only one enhancer located at each end (1v1); the other subset contains interactions which have more than one enhancer in either end or both ends (mvm). Thus, we derived an average of 131,907 interacting E-E pairs for "1v1" subsets and an average of 1,023,134 interacting pairs for "mvm" subsets. The numbers of E-E interactions collected from the HiChIP data of seven cell lines are shown in Table 1. Next, we use the unambiguous E-E interactions in "1v1" subsets to construct positive cases for model training and cross validation. After a context-specific model is trained using "1v1" subset, we then apply the model to identify true E-E interactions from ambiguous E-E pairs in "mvm" subset.

EnContact accurately predicts enhancer-enhancer contacts

To evaluate the ability of EnContact in predicting E-E interactions, we designed a series of systematical experiments. For each cell line, the unambiguous E-E interactions in "1v1" subset were regarded as positive cases. Meanwhile, we considered three ways to generate negative cases: (1) sampling negative cases based on the distance distribution of positive cases (random contacts; Fig. 2A); (2) randomly sampling negative cases with one end of positive E-E pairs fixed (random enhancers; Fig. 2A); (3) randomly sampling negative

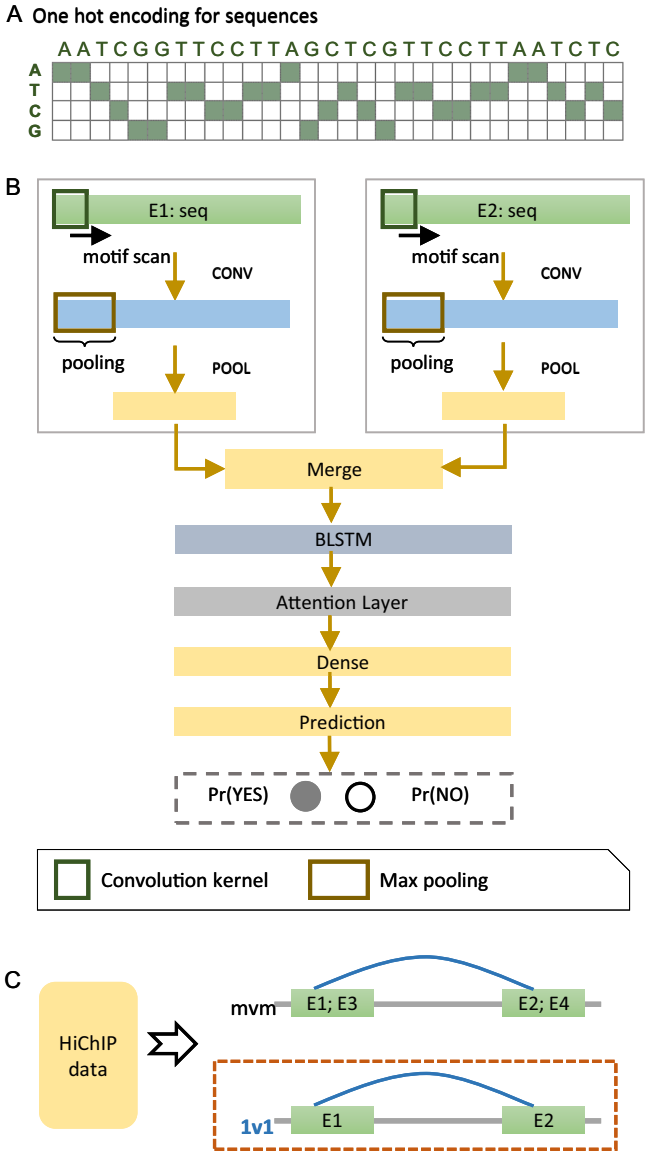

Figure 1 The EnContact method. (A) One-hot encoded sequence matrix. (B) Schematic illustration of the deep neural network architecture in EnContact. See "Methods" for details. (C) Region-based interactions in HiChIP data are divided into two sets: one set consists of interacting regions with only one enhancer in each region (1v1); the other set contains interacting regions where each region contains more than one enhancer (mvm). Full-size DOI: 10.7717/peerj.7657/fig-1

cases from all possible combinations of enhancers (random pairs; Fig. 2A). Then, we uniformly divided the union of positive cases and negative cases into 10 groups: one for testing and the others for training. All assessments were conducted on the testing sets.

Next, we trained and evaluated EnContact model in seven cell lines (Table 2). When negative cases were sampled on account of the distance distribution of positive cases, EnContact yields AUROCs of 0.803–0.858 and AUPRCs of 0.773–0.850. Otherwise, when negative cases were sampled based on random enhancers or random pairs, EnContact yields AUROCs of 0.811–0.867 and AUPRCs of 0.800–0.875 (for random enhancers), and

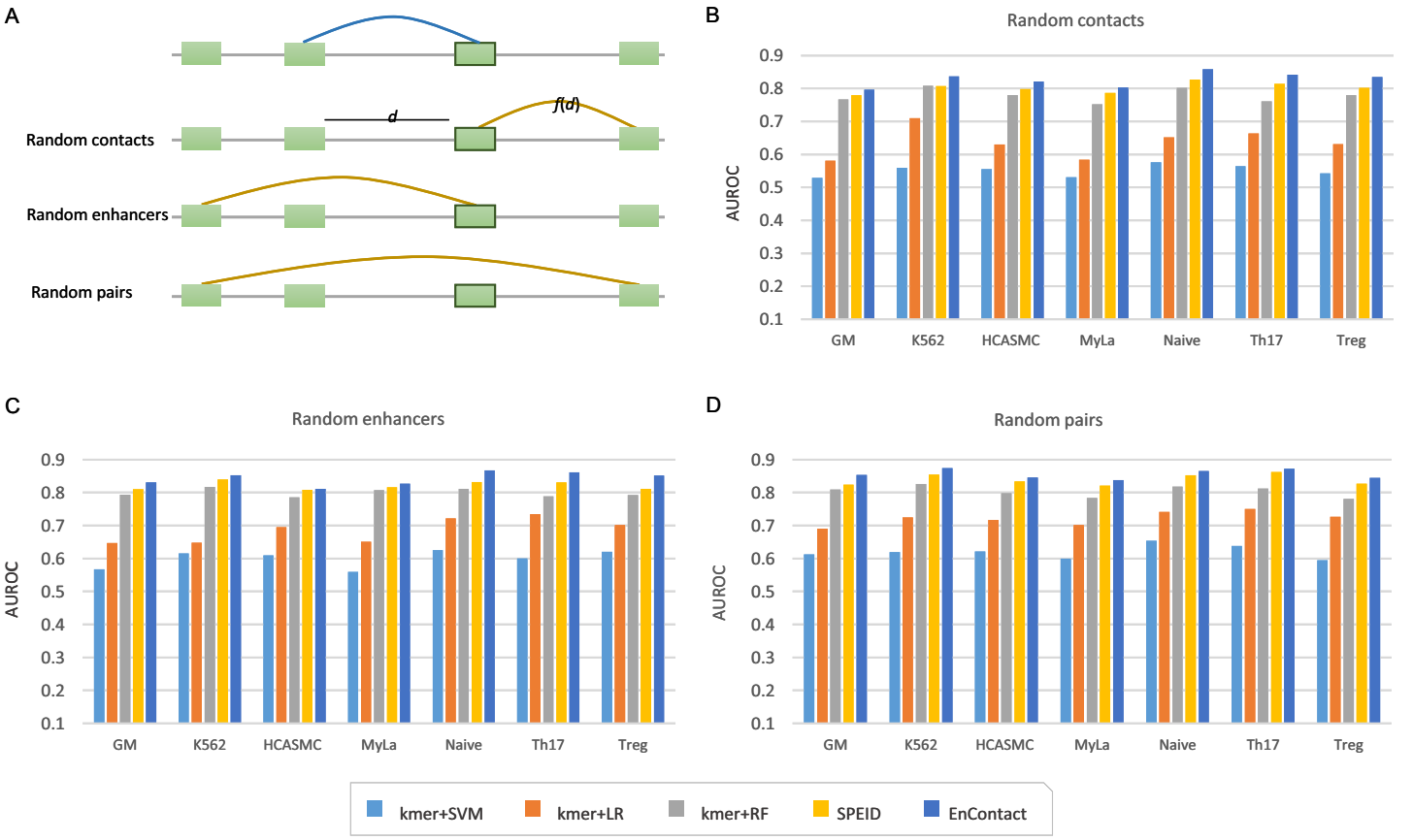

Figure 2 Comparison with baseline methods. (A) Three types of background enhancer-enhancer pairs. Random contacts are generated based on the distance distribution, $f(d)$, of positive interactions; d the distance between two enhancers. Random enhancers are sampled with one end of positive interactions fixed. Random pairs are randomly selected from all possible combinations of any two enhancers. (B–D) Performance of EnContact and other baseline methods across seven cell lines on account of random contacts (B), random enhancers (C), and random pairs (D) as background.
Full-size DOI: 10.7717/peerj.7657/fig-2

AUROCs of 0.838–0.874 and AUPRCs of 0.851–0.891 (for random pairs). This result demonstrates that our EnContact model can achieve decent performance no matter how negative cases were generated. To check whether the shared enhancers between the training set and testing set will bring bias to the model evaluation, we also designed another training and testing sets which involves different groups of enhancers. Specifically, we considered interactions derived from chromosome 1–18 as training set, and interactions derived from other chromosomes as testing set. Our model achieves comparable performance on the new training sets and testing sets (Table S5).

Finally, we compared our EnContact model with four baseline models: SVM (*Wu, Lin & Weng, 2004*), LR (*Hosmer, Lemeshow & Sturdivant, 2013*), RF (*Liaw & Wiener, 2002*), and SPEID model (*Singh et al., 2016*). In EnContact model and SPEID model, we used a one-hot encoding strategy to convert sequences into a numeric matrix and then extracted features by scanning along the matrix with a convolution layer. For other machine learning models which do not have convolution layers, we need a strategy to extract features from sequences for downstream classification. Here, to derive sequence

Table 2 Model performance of EnContact across seven cell lines.

Cell line	Random contacts		Random enhancers		Random pairs	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
GM	0.806	0.773	0.831	0.830	0.853	0.863
K562	0.837	0.850	0.852	0.826	0.874	0.887
HCASMC	0.821	0.804	0.811	0.815	0.846	0.857
MyLa	0.803	0.792	0.827	0.800	0.838	0.851
Naive	0.858	0.848	0.867	0.875	0.865	0.885
Th17	0.842	0.832	0.861	0.868	0.872	0.891
Treg	0.835	0.828	0.851	0.854	0.845	0.864
Mean	0.829	0.818	0.843	0.838	0.856	0.871

features, we counted the number of k-mer (say, $k = 6$) located in each enhancer and constructed a k-mer frequency vector for each enhancer. Then, we concatenated the frequency vector of two enhancers of an E-E pair as the input for baseline models. With the same training sets and testing sets, EnContact achieves a mean AUROC of 0.827, compared with 0.551 of SVM, 0.636 of LR, 0.779 of RF, and 0.802 of SPEID (random contacts; Fig. 2B); a mean AUROC of 0.843, compared with 0.599 of SVM, 0.686 of LR, 0.799 of RF, and 0.821 of SPEID (random enhancers; Fig. 2C); a mean AUROC of 0.856, compared with 0.620 of SVM, 0.722 of LR, 0.804 of RF, and 0.839 of SPEID model (random pairs; Fig. 2D). This shows that our model outperforms other baseline models in all seven cell lines under different backgrounds.

Taken together, the above results show that our EnContact model is capable of extracting useful features from genomic sequences to predict E-E interactions and can achieve much better performance than other traditional methods.

EnContact captures context-specific features which can be mapped to transcription factors

To explore the biological meaning of features learned by EnContact, we developed a strategy to convert the parameters of the first convolution layer of EnContact into PWM (see "Methods"). Each convolution kernel can be regarded as a pattern recognizer, which recognizes a specific sequence pattern by scanning along the input matrix. Then, we matched the resulting PWM to known motifs and TFs derived from JASPAR database (*Mathelier et al., 2016*). As a result, we found that EnContact captures different sets of TFs in different cell lines and some of the TFs have been reported to be key TFs in the corresponding cell line (Table 3). For example, IRF4 was identified as a key TF in Th17 cell line by our strategy and was previously reported to be required during the development of inflammatory Th17 cells (*Brüstle et al., 2007*; *Lohoff et al., 2002*). In Treg cell line, we matched a convolution kernel to the motif of ETS1, which was previously reported to control the development and function of natural Treg cells (*Mouly et al., 2010*). The motif analysis shows that EnContact model can capture context-specific sequence patterns and simultaneously convert input sequences into higher-level features. To sum up,

Table 3 Key TFs captured by convolution kernels of EnContact.

Cell type	Key TFs
GM	PKNOX1, PKNOX2, FOS, BHLHE40, JUND, USF1, EBF1
K562	EGR1, MGA, BHLHE40, MNT, CREB1, FOXA1, NFYB, USF1
HCASMC	ETS1, IRF2, HSF2, MZF1, HOXB3, JUND, CREB1, FOXO3, POU6F1, NFE2, ZNF263, BATF3, ATF7, NKX6-2
MyLa	ETS1, FLI1, RORA, CREB1, ELK4, NFATC1, ETV3, ZBTB33, REST, ELK3, MLX, NR3C1, SP2, ETV6
Naive	IRF4, PLAG1, IRF2, HSF2, MZF1, HOXB3, MEF2D, JUND, CREB1, EMX1, FOXO3, NFYB, MEF2A, POU6F1
Th17	IRF4, IRF2, FOXP3, FOXP1, TP53, RXRB, SMAD3, NR2C2, POU6F1, CREB1, TFE3, IRF7, PRDM1
Treg	ETS1, NFATC2, MAX, RORA, ELF4, NR2C2, RUNX1, ELF1, RELA, ZBTB33

EnContact demonstrates the ability to capture known context-specific sequence patterns and provides us an opportunity to explore novel context-specific TFs which have not been identified by experiments yet.

EnContact provides a finer mapping of enhancer-enhancer interactions

As discussed in the previous section, we divided E-E interactions derived from HiChIP data into "1v1" subset and "mvm" subset. Chromatin contacts in "mvm" subset are interactions with more than one enhancer in either end or both ends, which therefore can be regarded as a bunch of candidate E-E pairs (Fig. 3A). For a given cell line, once the context-specific model was trained using E-E interactions in "1v1" subset, we can apply this model to predict true interactions from candidate E-E pairs in "mvm" subset.

For each of the seven cell lines, we applied the well-trained EnContact model to identify true interactions from candidate E-E pairs, which can contribute to reducing the false positive rate of E-E pairs for HiChIP data. For each candidate E-E pair, EnContact provides a probability that two enhancers of this pair are interacting with each other. Setting threshold as 0.5, we considered E-E pairs with probabilities larger than the threshold as positive predictions, otherwise as negative predictions. In total, we identified 1,545,180 positive E-E interactions from candidate E-E pairs in seven cell lines.

Next, we compared the co-openness of two enhancers in positive predictions with that in negative predictions. Here, we calculated the co-opening degree of two enhancers based on the consistency of their DNase I hypersensitivity signal (Fig. 3B; see "Methods"). We assumed that if two enhancers are interacting with each other, they should have a larger probability to co-open. Indeed, we found that the percentage of co-opening pairs of positive predictions is larger than that of negative predictions in all seven cell lines (Fig. 3C). This result shows that E-E interactions predicted by EnContact are more likely to be co-opening than other candidate interactions, suggesting that the predicted E-E interactions are more reasonable than those derived from original HiChIP data.

Furthermore, we built a validation dataset using E-E interactions derived from ChIA-PET data of several cell lines to check the consistency between predicted E-E pairs and the validation dataset. We regarded positive predictions inferred by EnContact as

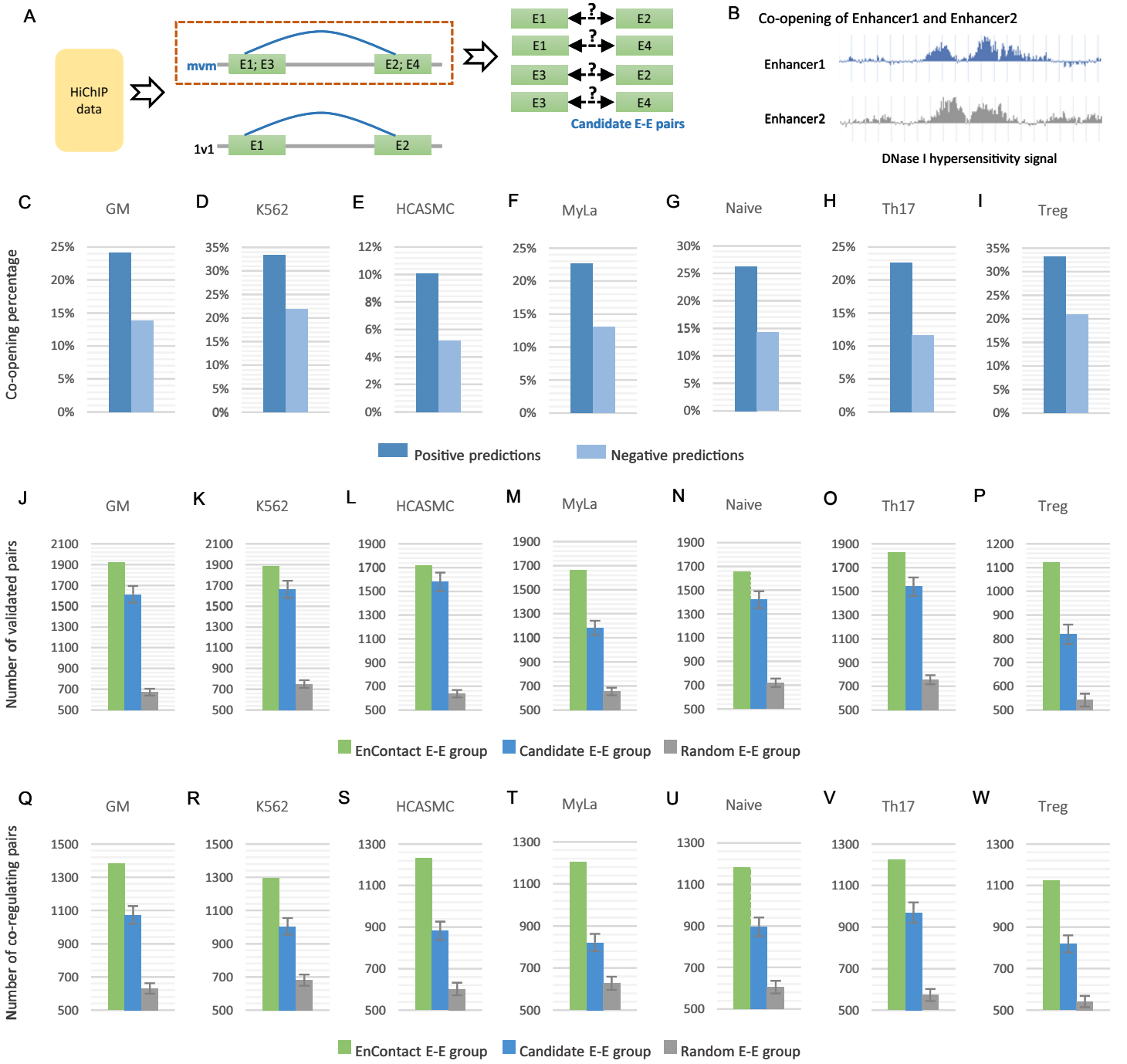

Figure 3 Fine-scale enhancer-enhancer interactions predicted by EnContact. (A) Region-based interactions in HiChIP data are divided into two sets. Interacting regions with more than one enhancer located on each end may result in several candidate enhancer-enhancer interactions. (B) Concept of enhancer co-opening based on DNase I hypersensitivity signal. (C–I) Co-opening percentages of positive predictions and negative predictions across seven cell lines. (J–P) Comparison of validated enhancer-enhancer interactions within E-E pairs predicted by EnContact, E-E pairs derived from candidate interactions, and random E-E pairs. (Q–W) Comparison of three E-E groups (i.e., EnContact E-E group, Candidate E-E group, and Random E-E group) in terms of the number of E-E pairs whose two enhancers regulate the same promoter.

Full-size DOI: 10.7717/peerj.7657/fig-3

"EnContact E-E group." Then, we constructed a comparison group, named "Candidate E-E group," by randomly sampling from all candidate pairs with the same sample size as EnContact E-E group. Additionally, we constructed a control group which contains E-E pairs generated based on the distance distribution of positive E-E pairs. The comparison group and control group were generated 1,000 times to remove randomness. In each cell line, we calculated the overlaps between these three E-E groups and the validation dataset. The result shows that in all seven cell lines, EnContact group has the largest overlap with validation interactions, compared with candidate groups and random groups (Fig. 3D). This indicates that EnContact can provide finer mapping E-E interactions than candidate E-E interactions derived from original HiChIP-data.

Finally, considering that E-E interaction may be related with the common promoter that they may co-regulate. We analyzed the E-E interaction based on whether they regulate the same promoter in ChIA-PET data. Specifically, for each E-E group described above (i.e., EnContact E-E group, Candidate E-E group, and Random E-E group), we counted the number of E-E pairs whose two enhancers regulate the same promoter. The result shows that E-E interactions predicted by EnContact model are more likely to co-regulate the same promoters than candidate E-E interactions and random E-E interactions (Fig. 3E). This indicates that EnContact model can identify E-E interactions which are related to the co-regulation of promoters.

Collectively, the above analysis suggests that features learned by EnContact from genomic sequences can be used to predict context-specific E-E interactions.

Characterization of hub enhancers

To further discuss the characterization of predicted interactions, we constructed an enhancer-based network using E-E interactions identified by EnContact. We checked the degree distribution of the enhancer network and observed that only a small portion of enhancers have significantly high degrees (Fig. 4A). In the following analysis, we take the data of GM cell line as an example. To check the characterization of enhancers frequently interacting with others, we defined those enhancers with top 10% highest degrees as hub enhancers. For comparison, we prepared hub enhancers of a network constructed using candidate interactions. In addition, we randomly selected non-hub enhancers with the same sample size as another control group. Next, we collected plentiful epigenomic data to explore distinct features of hub enhancers and also to compare the difference of characterization of hub enhancers defined by EnContact interactions, candidate interactions, and non-hub enhancers.

We asked whether hub enhancers are more active across different cell lines than other enhancers. To answer this question, we collected 637 ChIP-seq experiments of 128 cell lines for four key histone markers: H3K4me3, H3K27ac, H3K4me2, and H3K9ac (*The ENCODE Project Consortium, 2004*). For each hub enhancer, we calculated its activity across different experiments and counted the number of experiments where this enhancer is active. Thus, we derived a distribution of experiment numbers where all hub enhancers defined by EnContact interactions (i.e., E-E pairs predicted by EnContact) are active. Next, we also calculated another two distributions of experiment numbers

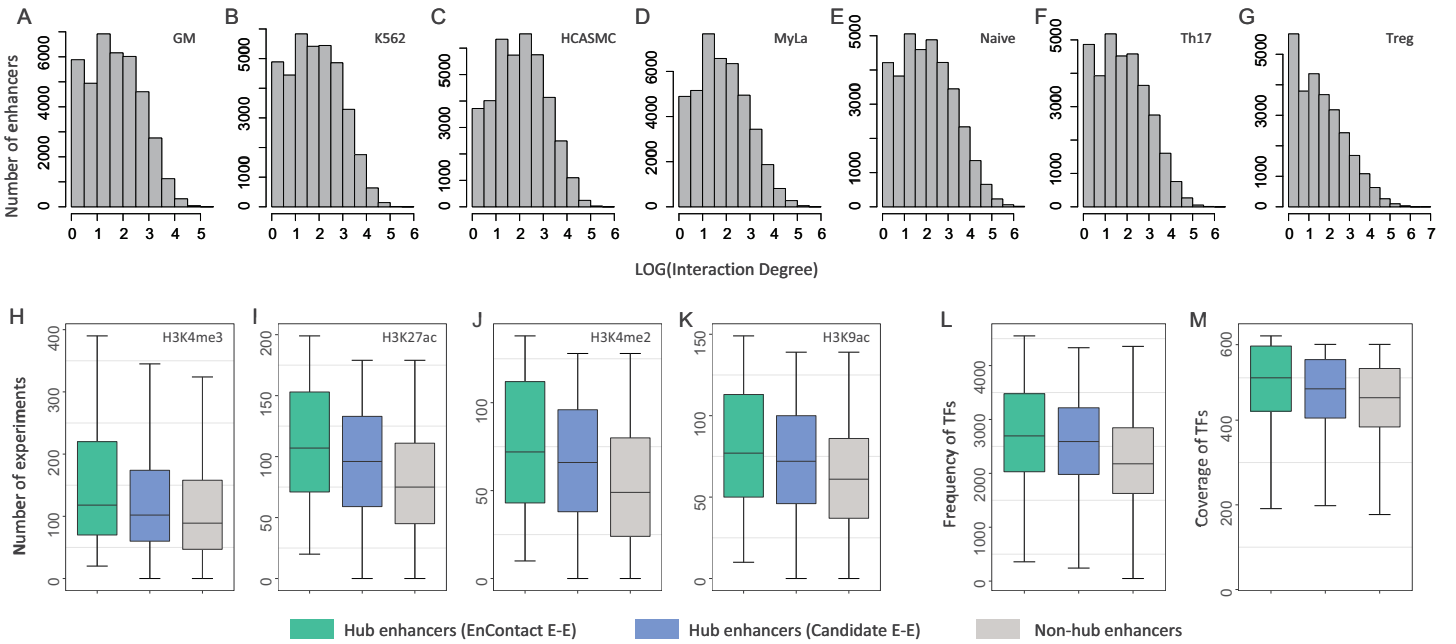

Figure 4 Characterization of hub enhancers. (A–G) Distribution of enhancer interaction degrees in different cell lines. (H–K) Comparison of hub enhancers derived from E-E pairs predicted by EnContact, hub enhancers derived from candidate E-E pairs, and non-hub enhancers in terms of four histone marker H3K4me3 (H), H3K27ac (I), H3K4me2 (J), and H3K9ac (K). (L) Comparison of the number of ChIP-seq experiments where hub enhancers are active. (M) Comparison of the number of TFs included in the experiments where hub enhancers are active. The *y*-axis represents the number of experiments where a promoter is active.
Full-size DOI: 10.7717/peerj.7657/fig-4

where hub enhancers defined by candidate interactions (i.e., E-E pairs derived from original HiChIP data) and non-hub enhancers are active. Comparing these three distributions, we observed that hub enhancers defined by EnContact interactions are generally active in significantly more experiments than hub enhancers defined by candidate E-E pairs or non-hub enhancers (Fig. 4B; *P*-values $< 2.2 \times 10^{-16}$, one-sided Wilcoxon rank-sum tests). This indicates that hub enhancers are indeed more active across different cell lines. Besides, the comparison result further supports that EnContact is able to extract true E-E interactions from original HiChIP data.

Moreover, we downloaded 4,383 ChIP-seq experiments of 144 cell lines for 579 distinct TFs (*The ENCODE Project Consortium, 2004*). Similarly, for each hub enhancer, we counted the number of experiments where this enhancer is active and the number of TFs included in these experiments. Then, we compared the number of experiments where hub enhancers define by the three E-E groups are active. As shown in Fig. 4C, hub enhancers defined by EnContact interactions are active in significantly more experiments and covered much more TFs than the other two enhancer groups (*P*-values $< 2.2 \times 10^{-16}$, one-sided Wilcoxon rank-sum tests). This again indicates that hub enhancers tend to be active across different cell lines.

In conclusion, EnContact predicts true E-E interactions from original HiChIP data and guides us to identify a class of hub enhancers which tend to be active across different cell lines. Furthermore, the analysis of the characterization of hub enhancers again supports the predicting ability of our EnContact model.

DISCUSSION

There are several directions worth exploring in future work. First, our sequence-based model can be used to evaluate the influence of nucleotide variants on chromatin interactions. Briefly, for a given SNP located in an enhancer, we can calculate the difference of E-E interacting probability when given reference nucleotide and given variant nucleotide. This difference can be regarded as a quantitative measurement of the biological influence of the given variant. Second, in this work, we only use genomic sequences as input to predict E-E interactions. Since epigenomic features are closely related to three-dimension structure and are complementary to sequence information, we can integrate epigenomic features, which can be derived from ChIP-seq data or ATAC-seq data, with genomic sequences to achieve a better prediction of chromatin contacts among enhancers. Finally, our current model focuses on the prediction of E-E interaction in a context-specific way. With the cooperation of epigenomic signals and sequence information, it is hopeful that an integrative model to predict E-E interactions across cell lines can be developed.

CONCLUSIONS

Chromatin contacts between regulatory elements are of crucial importance for the interpretation of transcriptional regulation and the understanding of disease mechanisms. In the last decade, many computational studies have been developed to improve the resolution of three-dimension genomic data (*Whalen, Truty & Pollard, 2016*; *Zhu et al., 2016*; *Zhang et al., 2018*). However, these methods mainly focused on the interactions between enhancers and promoters, leaving E-E interactions not well explored. In this paper, we designed a deep learning model, EnContact, to predict interactions among enhancers. First, we statistically demonstrated the predicting ability of EnContact using training sets and testing sets derived from HiChIP data of seven cell lines. Then, we compared EnContact with two other machine learning models using *k*-mer features as input. Results show that our model significantly outperforms baseline models. Next, we trained a context-specific model for each cell line and applied the model to predict E-E interactions from original HiChIP data. Finally, we identified hub enhancers from the predicted E-E interactions and observed that hub enhancers tend to be active across cell lines. We summarize that EnContact is capable of predicting E-E interactions using features automatically learned from genomic sequences.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This research was supported by the National Natural Science Foundation of China (Nos. 71871019, 71471016, 61873141 and 61573207), the National Key Research and Development Program of China (No. 2018YFC0910404), and the Tsinghua-Fuzhou Institute for Data Technology. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
National Natural Science Foundation of China: 71871019, 71471016, 61873141 and 61573207.
National Key Research and Development Program of China: 2018YFC0910404.
Tsinghua-Fuzhou Institute for Data Technology.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Mingxin Gan performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables, authored or reviewed drafts of the paper.
- Wenran Li conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables, authored or reviewed drafts of the paper.
- Rui Jiang conceived and designed the experiments, approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The EnContact model source code and data is available at https://github.com/liwenran/EnContact.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.7657#supplemental-information.

REFERENCES

Alipanahi B, Delong A, Weirauch MT, Frey BJ. 2015. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nature Biotechnology* **33(8)**:831–838 DOI 10.1038/nbt.3300.

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, Ntini E, Arner E, Valen E, Li K, Schwarzfischer L, Glatz D, Raithel J, Lilje B, Rapin N, Bagger FO, Jørgensen M, Andersen PR, Bertin N, Rackham O, Burroughs AM, Baillie JK, Ishizu Y, Shimizu Y, Furuhata E, Maeda S, Negishi Y, Mungall CJ, Meehan TF, Lassmann T, Itoh M, Kawaji H, Kondo N, Kawai J, Lennartsson A, Daub CO, Heutink P, Hume DA, Jensen TH, Suzuki H, Hayashizaki Y, Müller F, Forrest ARR, Carninci P, Rehli M, Sandelin A. 2014. An atlas of active enhancers across human cell types and tissues. *Nature* **507(7493)**:455–461 DOI 10.1038/nature12787.

Bastien F, Lamblin P, Pascanu R, Bergstra J, Goodfellow I, Bergeron A, Bouchard N, Warde-Farley D, Bengio Y. 2012. Theano: new features and speed improvements. *arXiv preprint Available at http://arxiv.org/abs/1211.5590.*

Brüstle A, Heink S, Huber M, Rosenplänter C, Stadelmann C, Yu P, Arpaia E, Mak TW, Kamradt T, Lohoff M. 2007. The development of inflammatory TH-17 cells requires interferon-regulatory factor 4. *Nature Immunology* **8(9)**:958–966 DOI 10.1038/ni1500.

Cao Q, Anyansi C, Hu X, Xu L, Xiong L, Tang W, Mok MTS, Cheng C, Fan X, Gerstein M, Cheng ASL, Yip KY. 2017. Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. *Nature Genetics* **49(10)**:1428–1436 DOI 10.1038/ng.3950.

Dekker J, Rippe K, Dekker M, Kleckner N. 2002. Capturing chromosome conformation. *Science* **295(5558)**:1306–1311 DOI 10.1126/science.1067799.

Diao Y, Fang R, Li B, Meng Z, Yu J, Qiu Y, Lin KC, Huang H, Liu T, Marina RJ, Jung I, Shen Y, Guan K-L, Ren B. 2017. A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nature Methods* **14(6)**:629–635 DOI 10.1038/nmeth.4264.

Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, Rubio ED, Krumm A, Lamb J, Nusbaum C, Green RD, Dekker J. 2006. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Research* **16(10)**:1299–1309 DOI 10.1101/gr.5571506.

Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH, Chew EGY, Huang PYH, Welboren W-J, Han Y, Ooi HS, Ariyaratne PN, Vega VB, Luo Y, Tan PY, Choy PY, Wansa KDSA, Zhao B, Lim KS, Leow SC, Yow JS, Joseph R, Li H, Desai KV, Thomsen JS, Lee YK, Karuturi RKM, Herve T, Bourque G, Stunnenberg HG, Ruan X, Cacheux-Rataboul V, Sung W-K, Liu ET, Wei C-L, Cheung E, Ruan Y. 2009. An oestrogen-receptor-α-bound human chromatin interactome. *Nature* **462(7269)**:58–64 DOI 10.1038/nature08497.

Gasperini M, Findlay GM, McKenna A, Milbank JH, Lee C, Zhang MD, Cusanovich DA, Shendure J. 2017. CRISPR/Cas9-mediated scanning for regulatory elements required for HPRT1 expression via thousands of large, programmed genomic deletions. *American Journal of Human Genetics* **101(2)**:192–205 DOI 10.1016/j.ajhg.2017.06.010.

Ghavi-Helm Y, Klein FA, Pakozdi T, Ciglar L, Noordermeer D, Huber W, Furlong EE. 2014. Enhancer loops appear stable during development and are associated with paused polymerase. *Nature* **512(7512)**:96–100 DOI 10.1038/nature13417.

Graves A, Jaitly N, Mohamed A-R. 2013. Hybrid speech recognition with deep bidirectional LSTM. In: *IEEE workshop on automatic speech recognition and understanding (ASRU).* Piscataway: IEEE, 273–278.

Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. 2007. Quantifying similarity between motifs. *Genome Biology* **8(2)**:R24 DOI 10.1186/gb-2007-8-2-r24.

Hosmer DW Jr, Lemeshow S, Sturdivant RX. 2013. *Applied logistic regression.* Vol. 398. Hoboken: John Wiley & Sons.

Ing-Simmons E, Seitan VC, Faure AJ, Flicek P, Carroll T, Dekker J, Fisher AG, Lenhard B, Merkenschlager M. 2015. Spatial enhancer clustering and regulation of enhancer-proximal genes by cohesin. *Genome Research* **25(4)**:504–513 DOI 10.1101/gr.184986.114.

Kelley DR, Snoek J, Rinn JL. 2016. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research* **26(7)**:990–999 DOI 10.1101/gr.200535.115.

Kleftogiannis D, Ashoor H, Bajic VB. 2018. TELS: a novel computational framework for identifying motif signatures of transcribed enhancers. *Genomics, Proteomics & Bioinformatics* **16(5)**:332–341 DOI 10.1016/j.gpb.2018.05.003.

Kumasaka N, Knights AJ, Gaffney DJ. 2019. High-resolution genetic mapping of putative causal interactions between regions of open chromatin. *Nature Genetics* **51(1)**:128–137 DOI 10.1038/s41588-018-0278-6.

Le N-Q-K, Ho Q-T, Ou Y-Y. 2017. Incorporating deep learning with convolutional neural networks and position specific scoring matrices for identifying electron transport proteins. *Journal of Computational Chemistry* **38(23)**:2000–2006 DOI 10.1002/jcc.24842.

Le N-Q-K, Ho Q-T, Ou Y-Y. 2018. Classifying the molecular functions of Rab GTPases in membrane trafficking using deep convolutional neural networks. *Analytical Biochemistry* **555**:33–41 DOI 10.1016/j.ab.2018.06.011.

Le NQK, Nguyen V-N. 2019. SNARE-CNN: a 2D convolutional neural network architecture to identify SNARE proteins from high-throughput sequencing data. *PeerJ Computer Science* **5(17)**:e177 DOI 10.7717/peerj-cs.177.

Le NQK, Yapp EKY, Ho Q-T, Nagasundaram N, Ou Y-Y, Yeh H-Y. 2019. iEnhancer-5Step: identifying enhancers using hidden information of DNA sequences via Chou's 5-step rule and word embedding. *Analytical Biochemistry* **571**:53–61 DOI 10.1016/j.ab.2019.02.017.

Li G, Chen Y, Snyder MP, Zhang MQ. 2017. ChIA-PET2: a versatile and flexible pipeline for ChIA-PET data analysis. *Nucleic Acids Research* **45(1)**:e4 DOI 10.1093/nar/gkw809.

Li W, Wang M, Sun J, Wang Y, Jiang R. 2017. Gene co-opening network deciphers gene functional relationships. *Molecular BioSystems* **13(11)**:2428–2439 DOI 10.1039/C7MB00430C.

Li W, Wong WH, Jiang R. 2019. DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. *Nucleic Acids Research* **47(10)**:e60 DOI 10.1093/nar/gkz167.

Liaw A, Wiener M. 2002. Classification and regression by randomForest. *R News* **2**:18–22.

Lohoff M, Mittrücker H-W, Prechtl S, Bischof S, Sommer F, Kock S, Ferrick DA, Duncan GS, Gessner A, Mak TW. 2002. Dysregulated T helper cell differentiation in the absence of interferon regulatory factor 4. *Proceedings of the National Academy of Sciences of the United States of America* **99(18)**:11808–11812 DOI 10.1073/pnas.182425099.

Mathelier A, Fornes O, Arenillas DJ, Chen C-Y, Denay G, Lee J, Shi W, Shyr C, Tan G, Worsley-Hunt R, Zhang AW, Parcy F, Lenhard B, Sandelin A, Wasserman WW. 2016. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research* **44(D1)**:D110–D115 DOI 10.1093/nar/gkv1176.

Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, Wingett SW, Andrews S, Grey W, Ewels PA, Herman B, Happe S, Higgs A, LeProust E, Follows GA, Fraser P, Luscombe NM, Osborne CS. 2015. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics* **47(6)**:598–606 DOI 10.1038/ng.3286.

Mouly E, Chemin K, Nguyen HV, Chopin M, Mesnard L, Leite-de-Moraes M, Burlen-defranoux O, Bandeira A, Bories J-C. 2010. The Ets-1 transcription factor controls the development and function of natural regulatory T cells. *Journal of Experimental Medicine* **207(10)**:2113–2125 DOI 10.1084/jem.20092153.

Mumbach MR, Rubin AJ, Flynn RA, Dai C, Khavari PA, Greenleaf WJ, Chang HY. 2016. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nature Methods* **13(11)**:919–922 DOI 10.1038/nmeth.3999.

Mumbach MR, Satpathy AT, Boyle EA, Dai C, Gowen BG, Cho SW, Nguyen ML, Rubin AJ, Granja JM, Kazane KR, Wei Y, Nguyen T, Greenside PG, Corces MR, Tycko J, Simeonov DR, Suliman N, Li R, Xu J, Flynn RA, Kundaje A, Khavari PA, Marson A, Corn JE, Quertermous T, Greenleaf WJ, Chang HY. 2017. Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nature Genetics* **49(11)**:1602–1612 DOI 10.1038/ng.3963.

Park Y, Kellis M. 2015. Deep learning for regulatory genomics. *Nature Biotechnology* **33(8)**:825–826 DOI 10.1038/nbt.3313.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V. 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* **12**:2825–2830.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26(6)**:841–842 DOI 10.1093/bioinformatics/btq033.

Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159(7)**:1665–1680 DOI 10.1016/j.cell.2014.11.021.

Roy S, Siahpirani AF, Chasman D, Knaack S, Ay F, Stewart R, Wilson M, Sridharan R. 2015. A predictive modeling approach for cell line-specific long-range regulatory interactions. *Nucleic Acids Research* **43(18)**:8694–8712 DOI 10.1093/nar/gkv865.

Schreiber J, Libbrecht M, Bilmes J, Noble W. 2018. Nucleotide sequence and DNaseI sensitivity are predictive of 3D chromatin architecture. *bioRxiv* 103614.

Shen W, Zhou M, Yang F, Yu D, Dong D, Yang C, Zang Y, Tian J. 2017. Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. *Pattern Recognition* **61**:663–673 DOI 10.1016/j.patcog.2016.05.029.

Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, De Wit E, Van Steensel B, De Laat W. 2006. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature Genetics* **38(11)**:1348–1354 DOI 10.1038/ng1896.

Singh S, Yang Y, Poczos B, Ma J. 2016. Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. *bioRxiv* DOI 10.1101/085241.

Tang Z, Luo OJ, Li X, Zheng M, Zhu JJ, Szalaj P, Trzaskoma P, Magalska A, Wlodarczyk J, Ruszczycki B, Michalski P, Piecuch E, Wang P, Wang D, Tian SZ, Penrad-Mobayed M, Sachs LM, Ruan X, Wei C-L, Liu ET, Wilczynski GM, Plewczynski D, Li G, Ruan Y. 2015. CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell* **163(7)**:1611–1627 DOI 10.1016/j.cell.2015.11.024.

The ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia of DNA elements) project. *Science* **306(5696)**:636–640 DOI 10.1126/science.1105136.

Wang G, Sun Y, Wang J. 2017. Automatic image-based plant disease severity estimation using deep learning. *Computational Intelligence and Neuroscience* **2017**:2917536 DOI 10.1155/2017/2917536.

Whalen S, Truty RM, Pollard KS. 2016. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nature Genetics* **48(5)**:488–496 DOI 10.1038/ng.3539.

Wu T-F, Lin C-J, Weng RC. 2004. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research* **5**:975–1005.

Zhang Y, An L, Xu J, Zhang B, Zheng WJ, Hu M, Tang J, Yue F. 2018. Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus. *Nature Communications* **9(1)**:750 DOI 10.1038/s41467-018-03113-2.

Zhu Y, Chen Z, Zhang K, Wang M, Medovoy D, Whitaker JW, Ding B, Li N, Zheng L, Wang W. 2016. Constructing 3D interaction maps from 1D epigenomes. *Nature Communications* **7(1)**:10812 DOI 10.1038/ncomms10812.