Briefings in Bioinformatics, 22(4), 2021, 1–10

https://doi.org/10.1093/bib/bbaa245
Problem Solving Protocol

GuidingNet: revealing transcriptional cofactor and predicting binding for DNA methyltransferase by network regularization

Lixin Ren, Caixia Gao, Zhana Duren and Yong Wang

Corresponding author: Yong Wang, CEMS, NCMIS, MDIS, Academy of Mathematics and Systems Science, National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Sciences, Zhongguancun East Road, Beijing 100190, China. Tel.: 86-10-82541372; Fax: 86-10-82541372. E-mail: ywang@amss.ac.cn

Abstract

The DNA methyltransferases (DNMTs) (DNMT3A, DNMT3B and DNMT3L) are primarily responsible for the establishment of genomic locus-specific DNA methylation patterns, which play an important role in gene regulation and animal development. However, this important protein family's binding mechanism, i.e. how and where the DNMTs bind to genome, is still missing in most tissues and cell lines. This motivates us to explore DNMTs and TF's cooperation and develop a network regularized logistic regression model, GuidingNet, to predict DNMTs' genome-wide binding by integrating gene expression, chromatin accessibility, sequence and protein–protein interaction data. GuidingNet accurately predicted methylation experimental data validated DNMTs' binding, outperformed single data source based and sparsity regularized methods and performed well in within and across tissue prediction for several DNMTs in human and mouse. Importantly, GuidingNet can reveal transcription cofactors assisting DNMTs for methylation establishment. This provides biological understanding in the DNMTs' binding specificity in different tissues and demonstrate the advantage of network regularization. In addition to DNMTs, GuidingNet achieves good performance for other chromatin regulators' binding. GuidingNet is freely available at https://github.com/AMSSwanglab/GuidingNet.

Key words: DNA methyltransferase; network regularization; transcription cofactor; data integration

Introduction

DNA methylation is essential for mammalian development and plays crucial roles in various biological processes, including regulation of gene expression, maintenance of genomic stability, genomic imprinting and X chromosome inactivation [1, 2]. It is catalyzed by DNA methyltransferases (DNMTs) to bind to DNA-specific region and add methyl groups to cytosine residues. For example, DNMT3A and DNMT3B alone or in a complex with DNMT3L are known to de novo establish DNA methylation, whereas DNMT1 mediates DNA methylation maintenance [3–6].

An urging question is how DNMTs recognize their binding sites in the genome in different tissues. The researchers were puzzled by the fact that DNA methylation established by DNMTs is highly locus and tissue specific but DNMTs do not have binding specificity by serving as general chromatin regulators. ChIP-seq experiments can measure DNMTs' genome-wide binding locations in specific cellular context [7]. However, ChIP-seq technique requires a large amount of sample material and high-quality antibody. Thus, DNMTs' genomic binding in most tissues are still missing. We observed that large-scale transcriptomic and epigenomic data across tissues are more easily measured

Lixin Ren is a graduate student in Inner Mongolia University. His research interests are bioinformatics, machine learning, optimization models and algorithms.
Caixia Gao is an Associate Professor in Inner Mongolia University. Her research interests are optimization models and algorithms.
Zhana Duren is an Assistant professor in Clemson University. His research interests are biostatistics and computational biology.
Yong Wang is a Professor in the Chinese Academy of Sciences. His research interest is computational systems biology.
Submitted: 8 June 2020; **Received (in revised form):** 15 August 2020

and rapidly accumulated in ENCODE and ROADMAP database. This motivates us to develop computational method to integrate the available omics data, predict DNMTs' binding and further understand their binding specificity mechanism.

Our recent study suggests that chromatin regulator (CR) is likely to be recruited to a regulatory element (RE) if RE is open and is bound by transcription factors (TFs), which have protein interaction propensity with the CR [8]. This allows us to hypothesize that DNMTs' binding is guided by TFs to a particular locus to methylate cytosines. TFs is known to express specifically in certain tissue types, recognize specific DNA motifs, bind regions with open chromatin structure and have direct and indirect protein–protein interactions with DNMTs. In addition, DNA methylation is almost exclusively found in CpG dinucleotides, and it suggests that DNMTs tend to bind to GC-rich sequences. Taken together, it is promising to integrate those evidence from available genome sequence, gene expression, chromatin accessibility and protein–protein interaction data to reveal DNMTs' binding sites and mechanism for binding specificity.

Collecting the existing DNMTs' ChIP-seq data as gold-standard positives, we model the DNMTs' binding site prediction as a binary classification problem. The purpose is to mine the matched expression and accessibility data (i.e. measured on the same sample), and context nonspecific data (sequence data and protein–protein interaction data), to recover a significant portion of the information in the missing data on DNMTs' binding location. Numerous statistical methods have been successfully applied in binary classification, and logistic regression (LR) is a powerful discriminative method. LR provides predicted probabilities of class membership and easy interpretation of the feature coefficients. In our case, we need to select TFs to provide rich interpretation of the DNMTs' binding mechanism. Regularized logistic regression (RLR) provides different choices of regularization terms such as the lasso (l_1-norm) regularization [9] and the elastic net [10], which generalizes the lasso by adding a l_2 penalty. Furthermore, the adaptive lasso regularizes different coefficients in the l_1-regularization [11] and provides a very general framework for setting feature weights [12–16].

In this paper, we develop a network regularized logistic regression framework, GuidingNet, to predict DNMTs' binding by integrating multiple dataset. Our major contribution is to reconstruct TF protein–protein interaction and coexpression network for regularization, to choose weights efficiently in adaptive lasso and to improve prediction accuracy and biological interpretability. GuidingNet predicts binding of Dnmt3b and Dnmt3l for E14 mESC, DNMT3A for human MCF-7 cells and DNMT3B for human HepG2 cells and foreskin keratinocyte. GuidingNet shows superior performance in both prediction and feature selection. In addition, GuidingNet outputs a TF cofactor network to interpret the mechanism of DNMTs' binding specificity in genomic locus and in different tissue contexts.

Methods

Overview of GuidingNet model

We propose GuidingNet to model the physical process that TF recruits DNMTs to a specific regulatory element (RE) to methylate cytosines (Supplementary Figure 1), i.e. DNMT is guided by TF to recognize the specific DNA motifs and bind in a RE. This requires that RE should be context-specifically accessible to TF binding, and TFs have context nonspecific protein interaction propensity with DNMT. Together GuidingNet takes those context specific and nonspecific genomic data as input and outputs the DNMT's binding probability on a given RE and the guiding TF network. As depicted in Figure 1, GuidingNet has three components, respectively, (i) extracting predictive genomic features and combining feature with physical meaning, (ii) generating the candidate guiding TF network and (iii) training the model and outputting DNMT binding and TF cofactors.

Figure 1 shows the example for Dnmt3b's binding prediction in mouse E14. The input data include chromatin openness, expression, sequence and protein–protein interaction. Features are extracted from the input data to construct training dataset. The training labels are from Dnmt3b's ChIP-seq data. Independently, the candidate GuidingNet are generated and weighted based on the TF protein–protein interaction network and co-expression. We used the TFs and weights of the candidate GuidingNet to train a network regularized logistic regression model. Model output are the probability that a specific region is the binding site of Dnmt3b in E14 and the underlying GuidingNet. The model components are described in Table 1 with definitions of notations.

Feature collection and combination

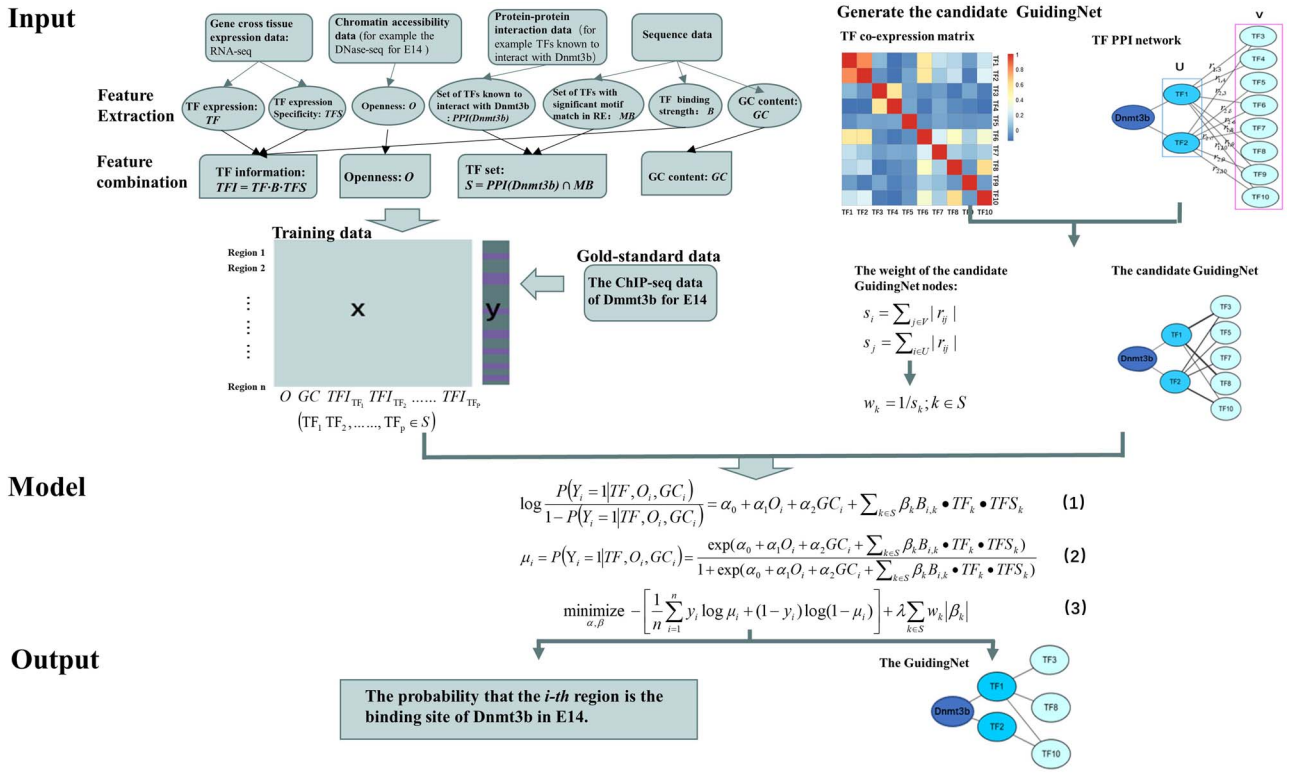
We extracted five features including openness (*O*), TF binding strength (*B*), TF expression (*TF*), TF expression specificity (*TFS*) and GC content (*GC*) from our input data (see details in Supplementary Materials). These features are not independent. Different combinations indicate different recruitment patterns of DNMT binding to the RE. We used the following feature transformation to model the three physical steps for DNMTs' binding processes that (i) RE is open, (ii) TF binds to the RE and (iii) RE has proper GC content for methylation. *B·TF·TFS* indicates that TF has significant motif match on the RE, highly expressed and specific expression across tissues (Figure 1). Then we have three features openness *O*, TF binding information *B·TF·TFS* and GC content *GC* for each RE.

Logistic regression model

We model DNMTs' binding to REs by a logistic regression, which is a statistical method for a binary classification problem. Given a DNMT, we can measure its ChIP-seq data and extract peaks as the gold-standard positive data (GSP) for binding. The gold-standard negative data (GSN) are randomly sampled from the nonbinding regions (see details in Supplementary Materials). Assume the entire training gold-standard data (includes GSP and GSN) and genomics feature data have n regions and p predictors. We denote whether a DNMT3 binding to regions i as $Y_i \in \{0,1\}$, $Y_i = 1$ means region i is bound by DNMT and $Y_i = 0$ means not. The genomic features are openness (*O*), TF binding information (*B·TF·TFS*) and GC content (GC). A linear relationship between the genomic features and the log-odds of the event that $Y_i = 1$ can be written in the following mathematical form:

$$\log \frac{P(Y_i = 1 \mid TF, O_i, GC_i)}{1 - P(Y_i = 1 \mid TF, O_i, GC_i)} = \alpha_0 + \alpha_1 O_i + \alpha_2 GC_i + \sum\nolimits_{k \in S} \beta_k B_{i,k} \cdot TF_k \cdot TFS_k \quad (1)$$

$$\mu_i = P(Y_i = 1 \mid TF, O_i, GC_i) = \frac{\exp(\alpha_0 + \alpha_1 O_i + \alpha_2 GC_i + \sum_{k \in S} \beta_k B_{i,k} \cdot TF_k \cdot TFS_k)}{1 + \exp(\alpha_0 + \alpha_1 O_i + \alpha_2 GC_i + \sum_{k \in S} \beta_k B_{i,k} \cdot TF_k \cdot TFS_k)}, \quad (2)$$



$$\log\frac{P(Y_i=1|TF,O_i,GC_i)}{1-P(Y_i=1|TF,O_i,GC_i)}=\alpha_0+\alpha_1O_i+\alpha_2GC_i+\sum_{k\in S}\beta_kB_{i,k}\bullet TF_k\bullet TFS_k \quad (1)$$

$$\mu_i=P(Y_i=1|TF,O_i,GC_i)=\frac{\exp(\alpha_0+\alpha_1O_i+\alpha_2GC_i+\sum_{k\in S}\beta_kB_{i,k}\bullet TF_k\bullet TFS_k)}{1+\exp(\alpha_0+\alpha_1O_i+\alpha_2GC_i+\sum_{k\in S}\beta_kB_{i,k}\bullet TF_k\bullet TFS_k)} \quad (2)$$

$$\underset{\alpha,\beta}{\text{minimize}}\left[-\frac{1}{n}\sum_{i=1}^{n}y_i\log\mu_i+(1-y_i)\log(1-\mu_i)\right]+\lambda\sum_{k\in S}w_k|\beta_k| \quad (3)$$

Figure 1. Overview of GuidingNet model.
GuidingNet is illustrated from three parts: input, model and output. Dnmt3b in mouse E14 is used as an example for DNMTs' binding prediction. Model input includes the construction of training data and the candidate GuidingNet. Openness, expression, sequence and protein–protein interaction data are collected and processed as the GuidingNet's input. Seven genomic features are extracted from the input data. TF expression, TF expression specificity and TF binding strength are further combined as one feature. This together gives training data X. The training labels are from the ChIP-seq data. Meanwhile the candidate GuidingNet is generated and weighted based on TFs' protein–protein interaction network and co-expression. GuidingNet is trained by an adaptive lasso-regularized logistic regression framework with the candidate GuidingNet and corresponding training data. The model components are described in Table 1 with mathematical notations. Model output are the probability that the region i is the binding site of Dnmt3b in E14 and the underlying GuidingNet.

Table 1. Model components for GuidingNet

Description of data and variables	Notation	Example
Expression of TF	TF_k :expression of TF k	$TF_{Nanog} = 30.95$ in E14 mESC
Accessibility of a genomic region	O_i :degree of openness of region i	$O_{chr7:12,922,178\text{-}12,922,505} = 12.83$in E14 mESC
Binding status of Dnmt3b in a genomic region	Y_i : indicator for whether the Dnmt3 is binding to region i.	Dnmt3b binds genome at chr7:12,922,178-12,922,505 in E14 mESC
Interacting TFs for Dnmt3b	*PPI*(*Dnmt3b*) :set of TFs known to interact with Dnmt3b	*PPI*(*Dnmt3b*) contains Nanog
TFs with motif match in a genomic region	*MB* : set of TFs with significant motif match in region i	Nanog has motif match at chr7:12,922,178-12,922,505 in E14 mESC
Motif matching strength of TF in a genomic region	$B_{i,k}$: binding strength of TF k in region i	$B_{chr7:12,922,178-12,922,505,Nanog} = 5.39$
GC fraction of a genomic region	GC_i : GC content of region i	$GC_{chr7:12,922,178\text{-}12,922,505}$ =0.61

where $S = PPI(DNMT3) \cap MB$. The log-likelihood function of equation (1) is defined as

$$l(\alpha,\beta)=\frac{1}{n}\sum_{i=1}^{n}y_i\log\mu_i+(1-y_i)\log(1-\mu_i), \quad (3)$$

where $\alpha=(\alpha_0,\alpha_1,\alpha_2)$ is a vector of TF-independent coefficients, $\beta=(\beta_1,\beta_2,\ldots\ldots)$ is a vector of TF-dependent coefficients and $y_i \in \{0,1\}$ is the training label of region i. We can estimate the parameters by minimizing the negative log-likelihood function.

Network regularization

We further introduce network regularization term in the above logistic regression model (3) to constrain the parameters and enhance model interpretation. The regularization is achieved based on our hypothesis that TFs physically interact and guide DMNT to DNA binding. These TFs do not work independently and tends to form a complex. Our network regularization aims to dig the biological relationships between them. First, we check the physical protein–protein interactions (PPI) among TFs and extracted the TF PPI data. Second, we check the TF co-expression relationships from a large amount of gene expression data in diverse context and reconstruct the TF co-expression network. In summary, we have two biological relationships among TFs,

namely protein–protein interaction and mRNA expression correlation, and the network structure can be constructed by integrating the TF PPI network and the TF co-expression network into a candidate GuidingNet. Then we use candidate GuidingNet's topology to derive the feature weights as network regularization in model (3).

Specifically, the generation of candidate GuidingNet consists of three steps (Figure 1 for Dnmt3b in mouse E14 example):

Step 1. TF set is constructed as $S = PPI(DNMT3) \cap MB$ by considering protein–protein interaction (PPI) with DNMT3b and motif occurrence. S is further divided into U and V. The set of TFs having direct PPIs (first order neighbor) with Dnmt3b is defined as U. The set of TFs having indirect PPIs (second order neighbor) with Dnmt3b is defined as V. Then we construct a bipartite graph $G_1(U, V, E_1)$ from PPI network with nodes in U and V. The edges of this graph are defined if TF i in U has PPI with TF j in V (as shown in Figure 1). For co-expression network, a complete bipartite graph $G_2(U, V, E_2)$ is constructed and we use c_{ij} to represent the weight of the edge between TF i and TF j. c_{ij} is calculated by the correlation between TF i and TF j for the expression levels across diverse tissues.

Step 2. We set r_{ij} as the weight of the edge between TF i and TF j in $G_1(U, V, E_1)$.

Step 3. Let $s_i = \sum_{j \in V} \mid r_{ij} \mid$, and $s_j = \sum_{i \in U} \mid r_{ij} \mid$. If $s_k = 0; k \in S$, we delete TF k in $G_1(U, V, E_1)$. Then we generated the candidate GuidingNet and constructed feature weights $w_k = 1/s_k; k \in S$.

With the candidate GuidingNet and derived weight, we regularize parameter β in logistic regression model (3) as follows:

$$\underset{\alpha,\beta}{\text{minimize}} - \left[\frac{1}{n}\sum_{i=1}^{n} y_i \log \mu_i + (1 - y_i)\log(1 - \mu_i)\right] + \lambda \sum_{k \in S} w_k |\beta_k|, \tag{4}$$

where w_k is the weight of the kth TF and is calculated from TF protein–protein interaction and co-expression network in an unsupervised manner. In practice, we fit the network regularized model using the vector of weight as 'penalty factors' in the glmnet R package.

Data sources

We collected the ChIP-seq of DNA methyltransferases 3 (DNMT3) family proteins including Dnmt3b and Dnmt3l for E14 mouse embryonic stem cells (mESC), DNMT3A for human MCF-7 cells, DNMT3B for human HepG2 cells and foreskin keratinocyte form the Cistrome Data Browser (http://cistrome.org/db). The corresponding paired RNA-seq and DNase-seq are downloaded from the ENCODE project. We collected RNA-seq data from other 145 mouse (Supplementary Table 1) and 832 humn (Supplementary Table 2) samples from the ENCODE and ROADMAP. We also collected the WGBS data of mouse E14 [17]. Both mouse and human protein–protein interaction data are from the BIOGRID database. The RNA-seq, DNase-seq and WGBS of mouse embryo development are downloaded from the ENCODE Web site.

Result

Openness, TF information and GC content are predictive for DNMTs' binding

We first quantitatively assess the usefulness of openness, TF information and GC content feature for DNMT3 binding prediction by area under the curve (AUC) value with univariate ordinary logistic regression. As shown in Figure 2, all three genomic features are good predictors for the DNMT3 prediction in five scenarios. All AUC values of single feature predicting exceed 0.5 and demonstrate all the features are informative than random guess. The performance of openness, TF information and GC content of binding prediction for the same DNMT3 protein in different cell types and different DNMT3 proteins were distinctive. In particular, the predictive ability of TF information is great except in human HepG2, demonstrating the importance of TF guidance in the process of DNMT3 binding to the regulatory elements. The predictive ability of openness and GC content are varied in different cell types. For example, the AUC for openness prediction in DNMT3A for MCF-7 is 0.82, but only 0.53 in Dnmt3b for mouse E14. The AUC for GC content in Dnmt3b for E14 is 0.85, but only 0.59 in Dnmt3l for mouse E14. Then we compared the GC content between Dnmt3b and Dnmt3l binding regions and without any methyltransferase binding regions in E14. Compared with Dnmt3l, the difference of GC content between the Dnmt3b binding regions and DNMT nonbinding is larger (Supplementary Figure 27). This indicates that Dnmt3b prefers to bind to regions with higher GC content. For DNMT3B, three feature performance are different in E14 mESC, human foreskin keratinocyte and HepG2 cells. This illustrates that the same DNA methyltransferase had a distinctive binding pattern in different cell types.

GuidingNet improves binding prediction

Our GuidingNet holds the promise to accurately predict DNMTs' binding and provide biological mechanism in different cellular contexts. We systematically evaluated the performance in prediction and feature selection for the above five scenarios. As indicated in Figure 2, ROC curves of the GuidingNet prediction on these five scenarios show better performance, i.e. averagely 84% area under the curve. Importantly, GuidingNet is better than the predictions based on unitary feature. This demonstrated the importance of integrative multi-omics data for predicting DNA methyltransferase binding.

We collected independent data to further validate our predictions from the following three aspects. First, we used the ChIP-seq data of Guiding TF to validate the Dnmt3b's binding prediction in E14. The rationale is that DNMT's binding sites are expected to largely overlap with cofactor's ChIP-seq region. We collected ChIP-seq of 3 Guiding TFs, Nanog, Pou5f1 and Sp1 in E14 from Cistrome database. In Dnmt3b for E14 dataset, there are 259, 2723 and 8513 genomic regions with Nanog, Pou51 and Sp1 binding, respectively. For Nanog binding regions in Dnmt3b for E14 dataset, there are 249 and 10 genomic regions, respectively, predicted by our model as the Dnmt3b's binding and nonbinding region. This gives the ratio of Nanog 24.90 (Pou5f1 and Sp1 are 4.12 and 10.11, respectively, in Supplementary Table 7). Those ratios are far larger than 1 and demonstrate that the GuidingNet significantly predict the correct binding sites. Second, we used WGBS data to classify genome regions into methylated regions and unmethylated regions. The rationale is that methylation is the result after the DNMT's binding. When we used WGBS data as gold-standard positive data in Dnmt3b for E14 and applied our model in this dataset. The AUC value predicted by GuidingNet is 0.89 (Supplementary Figure 23). Third, we performed additional cross validation to further show the reliability of our tool (Supplementary Figure 24).

GuidingNet reveals biological meaningful TF cofactors to enhance interpretability

The GuidingNet performs a two-step feature selection and shows good performance. As shown in Figure 3A–D, the

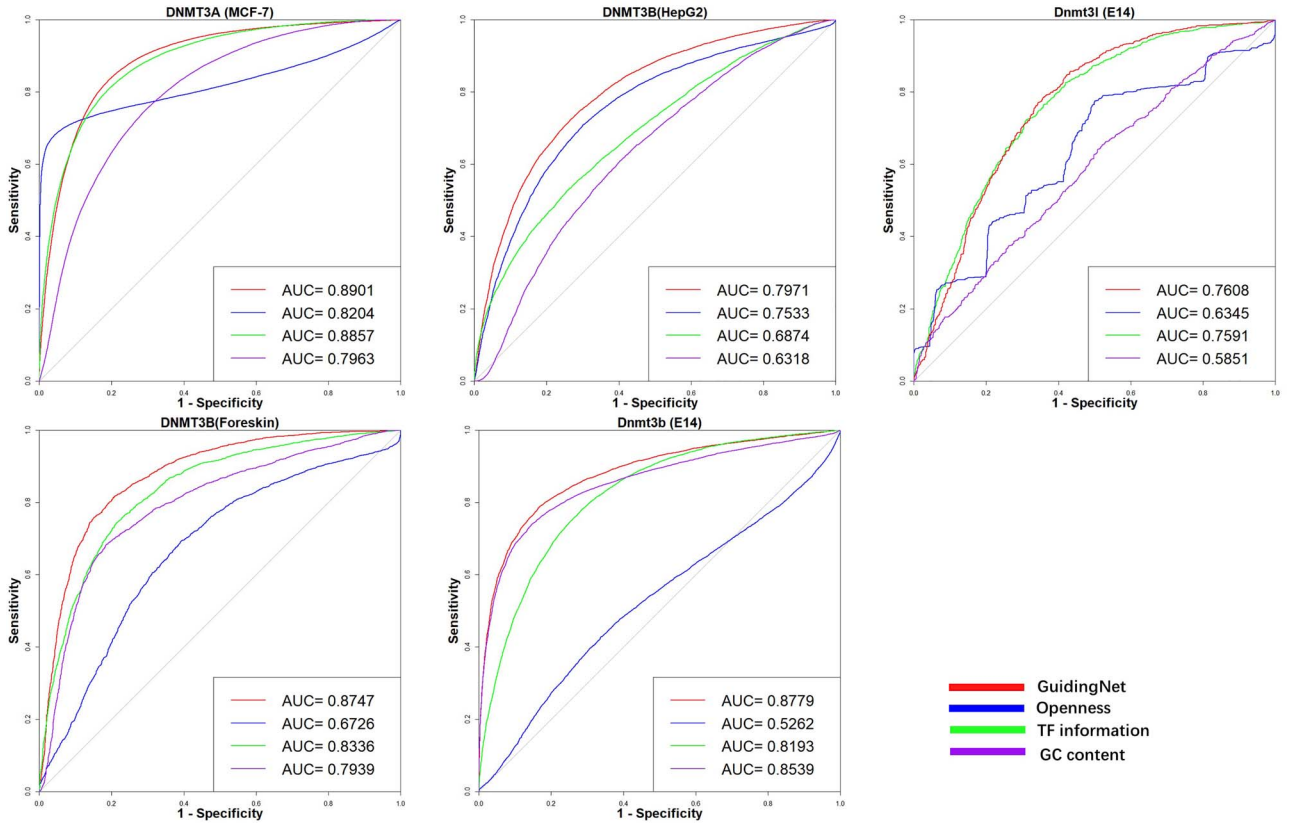


Figure 2. Every single feature is predictive and GuidingNet outperforms single features for predicting DNMTs' binding.
ROC curves of single feature and GuidingNet prediction of binding for DNMT3A in human MCF-7 cells, DNMT3B in human HepG2 cells, DNMT3B in human foreskin keratinocyte, and Dnmt3b and Dnmt3l in E14 mESC.

candidate GuidingNet includes fewer TFs than protein–protein interaction network, and the cutoff set was 0.1 for mouse and 0.7 for human cells. The cutoff value is chosen by controlling the balance between prediction accuracy and feature selection of model (see Supplementary Materials and Supplementary Figure 28). For example, the candidate GuidingNet of DNMT3A for the MCF-7 cells included approximately one-quarter of TFs in the protein–protein interaction network. Then we further filtered the candidate GuidingNet when fitting the network regularized logistic regression. Our method selected 20.8%, 15.6%, 16.8%, 53% and 16.3% TFs of protein–protein interaction network and 80%, 81%, 75%, 82% and 50% TFs of the candidate GuidingNet in DNMT3A for human MCF-7 cells, DNMT3B for human HepG2 cells and foreskin keratinocyte and Dnmt3b and Dnmt3l for E14, respectively (Figure 3A–D). Comparing with ordinary logistic regression (OLR) using openness, GC content and PPI network in Figure 3A–D (Supplementary Figure 2), GuidingNet's accuracy remains but used far fewer TFs for prediction. For example, our method used one-fifth TFs of OLR for DNMT3B binding in the human HepG2 cells.

The cofactors in the output GuidingNet can help us interpret the mechanism of how cofactor assist DNA methyltransferase binding to regulatory elements. For example, Nanog is a member of Dnmt3b's guiding network (GN) in E14 and Dnmt3b and Nanog bind to the genome region chr7: 118,740,617–118,741,016 in E14. WGBS data of E14 shows that this region is methylated. The possible mechanisms for Dnmt3b's binding to genome region chr7: 118,740,617–118,741,016 is to be guided by Nanog. In other words, Nanog binds to this region by recognizing specific DNA motifs and recruits Dnmt3b by protein interaction propensity. Then Dnmt3b methylates cytosines in this region. To validate those predicted guiding TFs, we preformed motif enrichment in the ChIP-seq of DNMT3 proteins by HOMER software. As shown in Figure 3A–D, indeed the identified guiding TFs enriched in the ChIP-seq of corresponding DNMT3, for example, Nanog, Sp1, Brca1, Sin3a and Zic3 in Dnmt3b for E14. In addition, literature search shows that a number of studies have linked our predicted guiding cofactors to DNA methyltransferase. For example, TRIM28-mediated recruitment of de novo DNMT3A that leads to cytosine methylation at CpG dinucleotides can control human endogenous retroviruses [18]. In total, 24 guiding cofactors were validated to be associated with DNA methyltransferase binding by literature (Supplementary Table 3). We note that the guiding cofactors may play an important role in development and many diseases. Uncovering the cofactor's mechanisms to regulate DNA methyltransferase binding may be useful in the production of therapeutic targets.

We performed upstream and downstream analysis of the obtained network from our model. For 14 TFs of Dnmt3b's guiding network (GN) in E14, we searched for regulatory elements in the range of 100 kb around each TF gene. And we found a total of 67 regulatory elements. We did TF motif enrichment in these regulatory elements by HOMER and identified 20 enriched motifs with significant p-value (Supplementary Table 4). More importantly, enriched TFs including CTCF, Sox2 and Pou5f1 have high expression levels in E14. We identified 249 genomic regions bound with Nanog and Dnmt3b in E14 by ChIP-seq data. We searched for effector

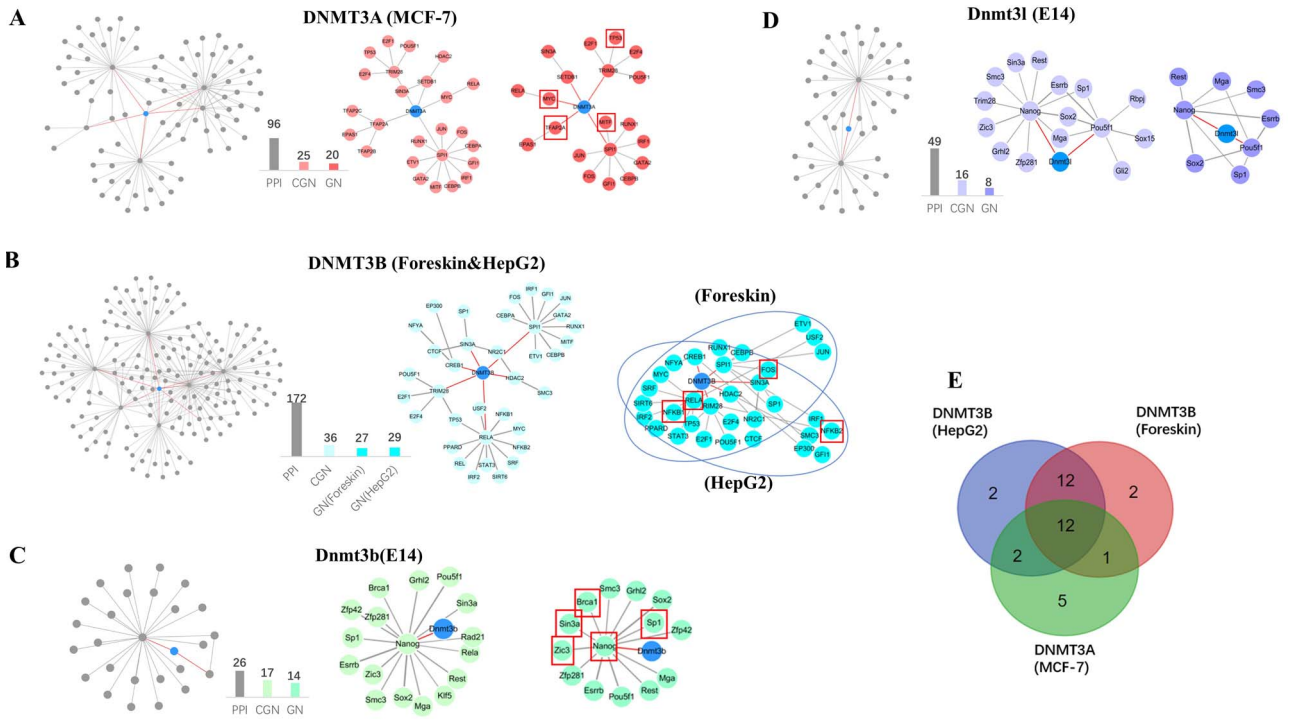


Figure 3. GuidingNet selects less TF co-factors and enhances interpretability.
(A–D) Comparison of protein–protein interaction networks (included first-order and second-order TFs interacting with the DNMT3 family protein), the candidate GuidingNet, and the GuidingNet of DNMT3A in human MCF-7 cells, DNMT3B in human HepG2 cells, DNMT3B in human foreskin keratinocyte, Dnmt3b in E14 mESC and Dnmt3l in E14 mESC. The blue nodes in protein–protein interaction networks indicate the DNMT. The red edges indicate the direct interaction between the DNMT with TFs. The width of the edge in the candidate GuidingNet indicates its weight. The wider the width of the edge, the greater its weight. Bar plot indicates the number of TFs in the PPI network (PPI), the candidate GuidingNet (CGN) and the GuidingNet (GN). The red box indicates TFs having motif enriched in the DNMT's ChIP-Sep peaks. (E) Venn diagram shows the number of shared TFs among the GuidingNet of DNMT3A in MCF-7, DNMT3B in HepG2 and DNMT3B in foreskin keratinocyte.

genes within 5 kb downstream of each region. We found a total of 324 effector genes, then we preformed Gene Ontology (GO) term enrichment analysis (Supplementary Figure 26). The most significant term is 'chromatin organization,' which is consistent with the CTCF and two pluripotent master regulators' biological function.

Then we compared the GuidingNets of these DNMT3 family proteins across different cellular contexts. It is interesting that the GuidingNets of DNMT3A and DNMT3B in human cells shared most of the guiding TFs (Figure 3E). Previous studies have shown that DNMT3A and DNMT3B are structurally similar and appear to have redundant functions overall [19]. One possible reason is that DNMT3A and DNMT3B are recruited by the same TFs. In summary, GuidingNet shows superior performance in predicting DNMTs' binding in both mouse and human and revealing their cofactors.

GuidingNet's genome-wide prediction is validated by WGBS data

We used the whole genome bisulfite sequencing (WGBS) data in mESC E14 to validate GuidingNet's binding prediction. Given a certain region, we can quantify the methylation level for this region by a simple fold change score, which is calculated as the number of methylated cytosine in this region by comparing with the total number of cytosine in this region. As indicated in Figure 4A, the methylation level of Dnmt3b binding regions is higher than non-binding regions in E14 mESC. Figure 4B shows that the higher the methylation level of this region, the higher the probability that this region is a Dnmt3b binding site. The Pearson correlation coefficient between our binding probability of Dnmt3b to the region and the methylation level of this region is 0.816, indicating that our prediction results are consistent with the independent measured WGBS data.

Across-tissue predicting Dnmt3b's binding

To further evaluate our model performance, we performed across-tissue prediction of Dnmt3b's binding by GuidingNet. As shown in Figure 5A, we trained the model in E14 mESC then predicted which open regions are Dnmt3b's binding sites at 7 time points (E11.5, E12.5, E13.5, E14.5, E15.5, E16.5 and P0) of the liver during embryo development. We used the WGBS data to validate the Dnmt3b's binding prediction. As indicated in Figure 5B, the methylation level of Dnmt3b's predicted binding regions is higher than non-binding regions. This is consistent with the results of the WGBS data validation in E14 mESC. These results suggest that our method can apply in across-tissue prediction of Dnmt3b's binding. Interestingly, the methylation level of the predicted binding region in E14.5 is higher than other time points. We ask the question that whether the better prediction of E14.5 is related to the right guiding TFs. We compared predicted binding regions between E14.5 and E13.5, then extracted the 11,494 unique binding regions of E14.5 (Supplementary Figure 16A). We did TF motif enrichment in these regions by HOMER and identified 8 motifs as enriched guiding TFs (Nanog, Sp1, Smc3, Zic3, Zfp281, Esrrb,

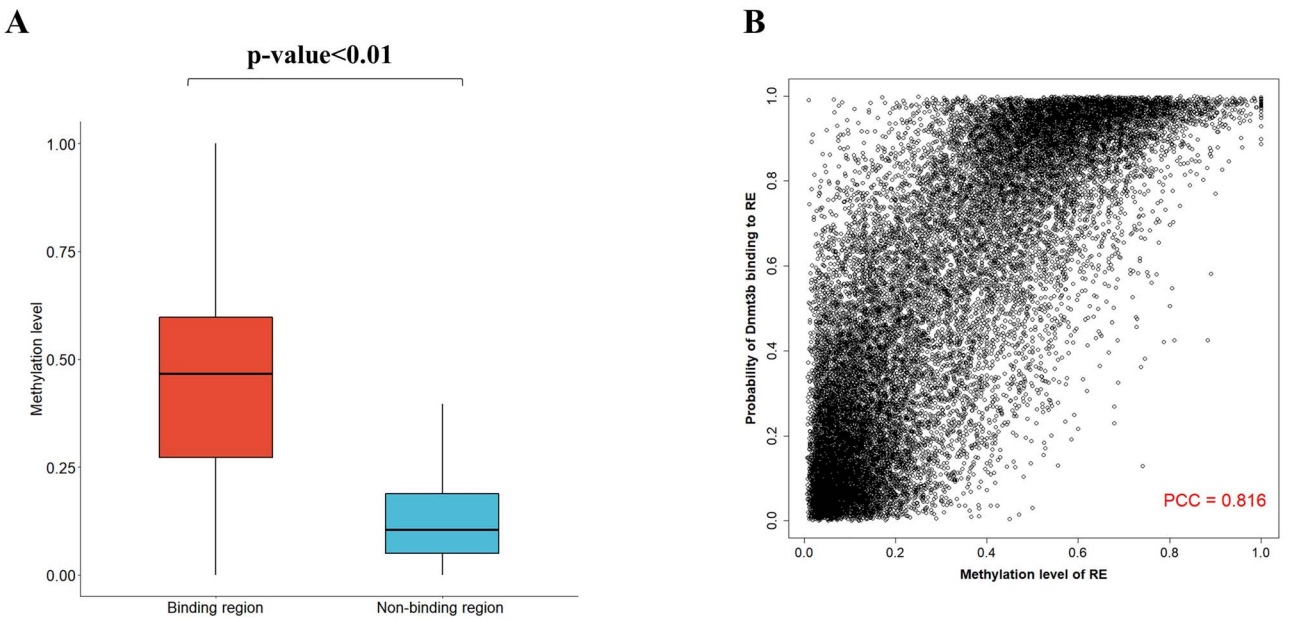


Figure 4. DNA methylation data validate Dnmt3b's binding prediction in mESC.
(A) Methylation level in Dnmt3b's predicted binding regions is significantly higher than non-binding regions (defined as the open region not covered Dnmt3b's binding region) in E14 mESC. (B) The scatter plot of the probability of GuidingNet output for Dnmt3b's binding versus the methylation level for a given region.

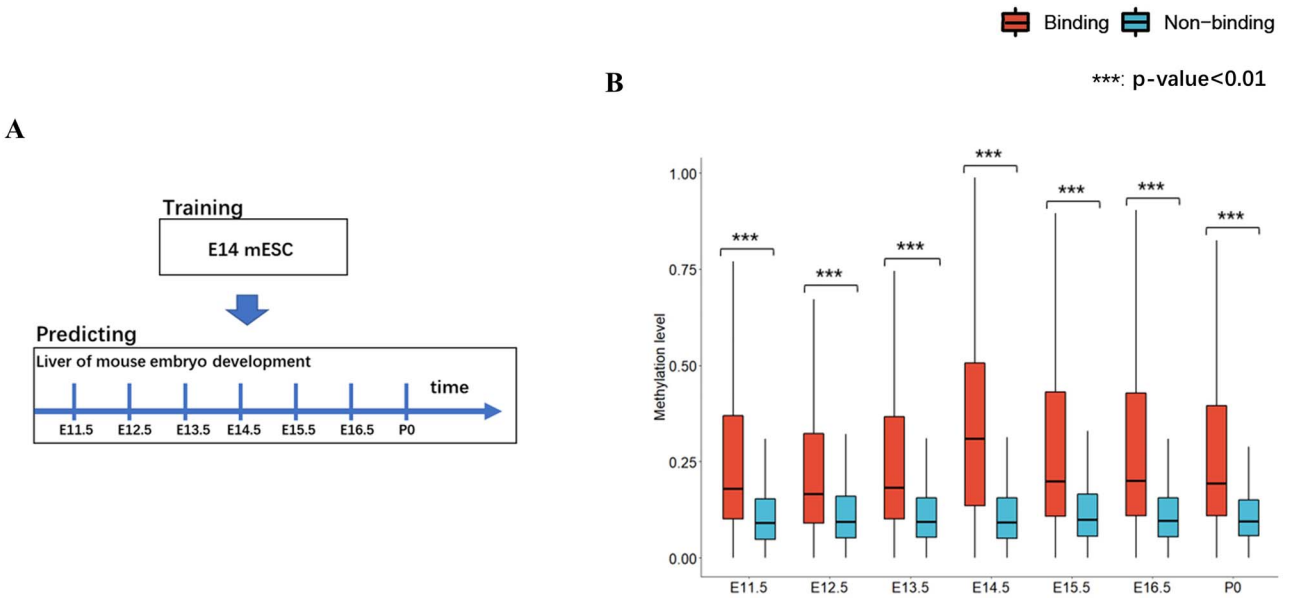


Figure 5. GuidingNet predicts Dnmt3b's across-tissue binding.
(A) Schematic of training the model and across-tissue predicting. (B) Comparison of methylation level between Dnmt3b's predicted binding regions and non-binding regions.

Pou5f1, Rest). Importantly, Smc3, Sp1, Zfp281 and Rest have high expression levels in E14.5 (Supplementary Figure 16B). In addition, the distance between predicted binding regions and the transcription start site (TSS) at E14.5 is shorter than other time points (Supplementary Figure 17). This may indicate that our method trained in mouse E14 is more accurate in the promoter regions. We expect to train our model in more tissue and cell types, and consequently, we could predict the DNMT binding for a new tissue and cell type that has not been studied yet.

GuidingNet can predict other chromatin regulators' binding in human and mouse

We ask whether our model has the predictive power in other chromatin regulators' binding. The ChIP-seq datasets for 4 CRs (Ep300, Ezh2, Setdb1 and Suz12) in mouse E14 and 6 CRs (EP300, EZH2, HDAC2, SMARCC2, SMARCE1 and SUZ12) in human HepG2 cells are collected form Cistrome database and utilized as gold-standard positive data. We evaluated our predictions of the binding of these CRs. Figure 6A and 6B shows

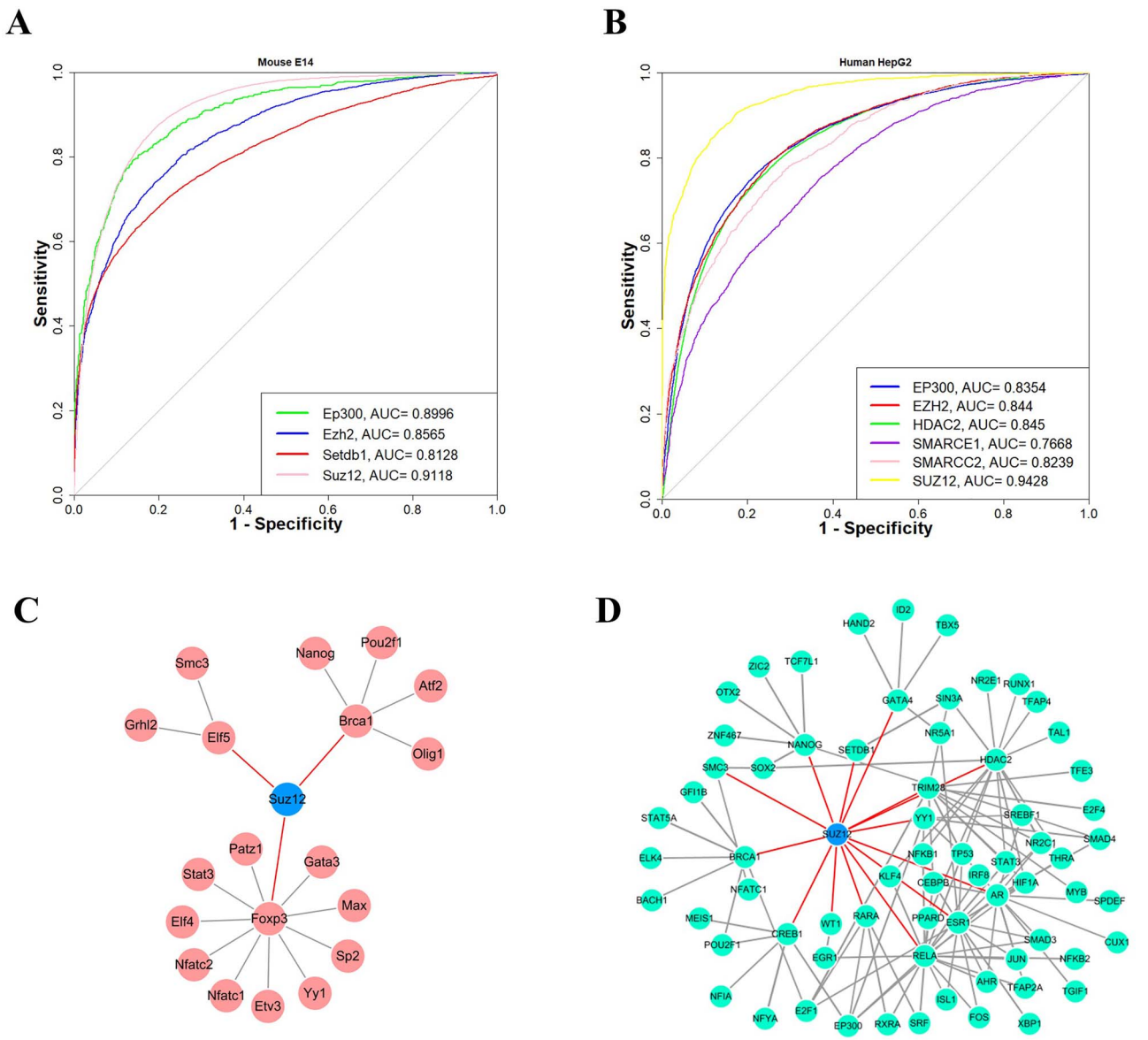


Figure 6. GuidingNet predicts other CR's binding in human and mouse.
(A) ROC curves of prediction of 4 CRs' binding in mouse E14. (B) ROC curves of 6 CRs' binding in the human HepG2. (C and D) The GuidingNet of Suz12 in mouse E14 mESC and SUZ12 in human HepG2, respectively.

that good performances (74–94% AUC) have been achieved for these CRs. We also assessed the predictive power of openness, TF information and GC content (Supplementary Figures 4 and 9). TF information showed the superior performance for all these CRs, and GC content had weak predictive power. The predictive ability of openness varies for different CRs. In addition, our method showed outstanding performance in TF-related feature selection (Supplementary Figures 5–8 and 10–15). In particular, our method had shown superior accuracy in predicting SUZ12's binding in both mouse and human. The AUCs are 0.91 and 0.94, respectively. In addition, a total of 7 TFs (BRCA1, YY1, STAT3, NFATC1, SMC3, POU2F1 and NANOG) are shared in the GuidingNet of human HepG2 (Figure 6C) and mouse E14 (Figure 6D). These TFs are conserved from mouse to human, suggesting that they play an important role in guiding SUZ12 binding to the regulatory elements. The revealed GuidingNets of these CRs in E14 mESC and HepG2 cells are biologically reasonable (Supplementary Figures 5–8 and 10–15). The strong performance both in prediction accuracy and TF selection suggests that GuidingNet is useful in other chromatin regulators' study.

Discussions and conclusions

We introduce GuidingNet, a network-regularized logistic regression framework, to integrate gene expression data, chromatin accessibility data, DNA sequence information and protein–protein interaction data for modeling DNMTs' binding. Our major contribution is to propose a network regularization to choose TF-related feature weights in the adaptive lasso base on TF protein–protein interaction and co-expression network.

Through comprehensive validation, our method shows superior performance, and the generated GuidingNet helps us to interpret DNMTs' binding mechanism in different cell types and different species. Although GuidingNet's prediction cannot substitute ChIP-seq data, it can provide a powerful reference for the actual binding site information of DNMTs on the genome, especially in those tissues or cell lines that still missing ChIP-seq data.

In addition to sensitivity analysis to show every feature is predictive, we have performed uncertainty analysis for model inputs. Three types of data including openness, TF expression and gold-standard data may have noise. To simulate different noise levels in each type of data, we randomly selected 100 times 5%, 10%, 15% and 20% of the data and added noise. Then we applied our model on the data after adding noise and recorded prediction accuracy. As shown in Supplementary Figure 18–22, different degrees of noise in openness and TF expression data have little effect on model prediction accuracy. As the degree of noise added to gold-standard data increases, the prediction accuracy of our model decreases drastically.

We compared our model with other four classification methods: naive Bayes, SVM, CART and random forest in five scenarios. As shown in Supplementary Figure 25, the prediction performance of GuidingNet consistently outperforms naive Bayes in five scenarios. This can be expected since logistic regression doesn't require the conditional independence assumption among features. The prediction accuracy of GuidingNet is also higher than that of SVM, except for Dnmt3l (E14). GuidingNet has achieved better predictive performance for DNMT3B (HepG2), DNMT3B (Foreskin) and Dnmt3b (E14) compared to the CART. The prediction performance of nonlinear model, random forest, is slightly better than our GuidingNet model in five scenarios. The difference of AUC values between GuidingNet and random forest are within 0.06, except for Dnmt3l (E14). This indicate nonlinear model has potential to improve accuracy. However, GuidingNet selected biological meaningful features and used far fewer TFs for prediction. Our method used 20.8%, 15.6%, 16.8%, 53% and 16.3% TFs of random forest in DNMT3A (MCF-7), DNMT3B (HepG2), DNMT3B (Foreskin), Dnmt3b (E14) and Dnmt3l (E14). Taken together, logistic regression is a better choice in practice balancing the biological interpretation and prediction accuracy.

Our approach is outstanding for predicting DNMTs' binding in the following aspects. First, GuidingNet integrates TF protein–protein network and co-expression across tissue and cell types to generate the weight of features in an unsupervised manner. This is different with existing procedures to select weights by computing an initial estimate using the response variable. Second, our method has a strong ability to select features by (i) setting cutoff in the GuidingNet and (ii) fitting the regularized logistic regression. Third, we provide further biological insights into the different binding mechanism of DNMTs in different cell types. Fourth, our method is capable of making across-tissue predictions validated by independent WGBS data. Last, GuidingNet is general enough to predict other chromatin regulators' binding both in the human and mouse and shows outstanding performance.

Our model can further be improved from many aspects. First, the missing of gold-standard data limits us from training models in more tissues and cell types. Second, several studies have demonstrated that DNMT3 binding is closely related to chromatin histone modification, such as DNMT3B preferentially targets gene bodies marked with H3K36me3 [20, 21]. Many of the ChIP-seq of histone modification are available. Integrating other histone modification information in our model may benefit the prediction accuracy. Third, histone methyltransferases and demethylase also play an important role in mammal, such as KMT and KDM family are associated with embryo development and cancer occurrence and progression [22, 23]. Fourth, GuidingNet can be extended to time-series data and multiple tissue data. We can assume that DNMTs' binding at adjacent time points or in similar tissue are similar. Then continuous constraints can be applied to model parameters to borrow information. Integrating the correlation information between different time points and tissues into our model and performing joint prediction of DNMT binding in time-series or multiple tissue data may improve the prediction accuracy. Last, alterations in DNA methylation patterns have been implicated in embryo development and tumorigenesis in many studies [24–26]. We look forward to deciphering the regulatory mechanism of mediating DNA methylation status alteration and how it impacts these important biological processes.

Key Points

- We propose a network regularized logistic regression model, GuidingNet, for predicting DNMTs' genome-wide binding by integrating gene expression, chromatin accessibility, sequence and protein–protein interaction data.
- We reconstruct TF protein–protein network and co-expression network for regularization, to choose weights efficiently in adaptive lasso, and to improve prediction accuracy and biological interpretability.
- GuidingNet outputs a TF network assisting DNMTs' binding and allows us to interpret the mechanism of DNMTs' binding in different tissues and cell types.
- GuidingNet can be generalized to predict other CR binding sites in human and mouse and has achieved good performance in prediction accuracy and feature selection.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Conflict of interests

The authors declare that there is no conflict of interests.

Funding

National Science Foundation of China (Grants 11871463, 61671444 and 61621003); Shanghai Municipal Science and Technology Major Project (No. 2017SHZDZX01); CAS 'Light of West China' Program (No. xbzg-zdsys-201913); Research Program of Science and Technology, Universities of Inner Mongolia Autonomous Region (No. NJZY19005 to C.G.).

References

1. Hackett JA, Surani MA. DNA methylation dynamics during the mammalian life cycle. *Philos Trans R Soc Lond B Biol Sci* 2013;**368**:20110328.
2. Smith ZD, Meissner A. DNA methylation: roles in mammalian development. *Nat Rev Genet* **14**:204–20.

3. Li E, Okano FM, Haber DF, *et al*. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian. *Development* 1999;**99**:247–57.
4. Schuebeler D. Function and information content of DNA methylation. *Nature* **517**:321–6.
5. Jeltsch A, Jurkowska RZ. New concepts in DNA methylation. *Trends Biochem Sci* **39**:310–8.
6. Neri F, Krepelova A, Incarnato D, *et al*. Dnmt3L antagonizes DNA methylation at bivalent promoters and Favors DNA methylation at gene bodies in ESCs. *Cell* 2013;**155**:121–34.
7. Johnson DS, Mortazavi A, Myers RM, *et al*. Genome-wide mapping of in vivo protein-DNA interactions. *Science* 2007;**316**:1497–502.
8. Duren Z, Chen X, Jiang R, *et al*. Modeling gene regulation from paired expression and chromatin accessibility data, Proceedings of the National Academy of Sciences. 2007;**114**:E4914–23.
9. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc B Methodol* 1996;**58**:267–88.
10. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodology* 2005;**67**:301–20.
11. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc* 2006;**101**:1418–29.
12. Huang J, Ma S, Zhang C-H. Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica* 2008; 1603–18.
13. Breiman L. Better subset regression using the nonnegative garrote. *Dent Tech* 1995;**37**:373–84.
14. Yuan M, Lin Y. On the non-negative garrotte estimator. *J R Stat Soc Series B Stat Methodology* 2007;**69**:143–61.
15. Zou H, Zhang HH. On the adaptive elastic-net with a diverging number of parameters. *Ann Stat* 2009;**37**:1733.
16. Guan L, Fan Z, Tibshirani R. Regularization for supervised learning via the" hubNet" procedure. *arXiv preprint arXiv* 1608.05465 2016.
17. Liu Y, Siejka-Zielińska P, Velikova G, *et al*. Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. *Nat Biotechnol* 2019;**37**:424.
18. Turelli P, Castrodiaz N, Marzetta F, *et al*. Interplay of TRIM28 and DNA methylation in controlling human endogenous retroelements. *Genome Res* 2014;**24**:1260.
19. Jing L, Rahul K, Hongcang G, *et al*. Targeted disruption of DNMT1, DNMT3A and DNMT3B in human embryonic stem cells. *Nat Genet* 2015;**47**:469.
20. Rondelet G, Maso TD, Willems L, *et al*. Structural basis for recognition of histone H3K36me3 nucleosome by human de novo DNA methyltransferases 3A and 3B. *J Struct Biol* 2016;**194**:357–67.
21. Rinaldi L, Datta D, Serrat J, *et al*. Dnmt3a and Dnmt3b associate with enhancers to regulate human epidermal stem cell homeostasis. *Cell Stem Cell* 2016;**19**:491–501.
22. Amanda N, Colaiácovo MP, Yang S. Developmental roles of the histone lysine demethylases. *Development* 2009; **136**:879–89.
23. Marianne Terndrup P, Kristian H. Histone demethylases in development and disease. *Trends Cell Biol* 2010;**20**: 662–71.
24. Esteller M. Epigenetics in cancer. *N Engl J Med* 2008;**358**: 1148–59.
25. Suvà ML, Riggi N, Bernstein BE. Epigenetic reprogramming in cancer. *Science* 2013;**339**:1567–70.
26. Feinberg AP, Koldobskiy MA, Göndör A. Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nat Rev Genet* 2016;**17**:284.