PREDICTION OF *CIS*-REGULATORY MODULES IN GENOMES

by

Pengyu Ni

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics

Charlotte

2020

Approved by:

_____
Dr. Zhengchang Su

_____
Dr. Juntao Guo

_____
Dr. Xinghua Mindy Shi

_____
Dr. Baohua Song

ABSTRACT

PENGYU NI. Prediction of *cis*-regulatory modules in genomes. (Under the direction of DR. ZHENGCHANG SU)

Annotating all *cis*-regulatory modules (CRMs) and constituent transcription factor (TF) binding sites (TFBSs) in genomes is essential to understand genome functions, however, the task remains highly challenging. In this dissertation, we first developed a new algorithm dePCRM2 for predicting CRMs and TFBSs by integrating numerous TF ChIP-seq datasets based on an ultra-fast motif-finding algorithm. dePCRM2 partitions genome regions covered by extended binding peaks in the datasets into a CRM candidates (CRMCs) set and a non-CRMCs set, and evaluates each CRMC using a novel score that captures the essential features of CRMs. Applying dePCRM2 to 6,092 datasets covering 77.47% of the human genome, we predicted 201 unique TF binding motif families and 1,404,973 CRMCs. And dePCRM2 largely outperforms the existing methods. Based our predictions, we estimated that about 55% and 22% of the genome code for CRMs and TFBSs, respectively. Thus, the regulatory genome is more prevalent than originally thought. Moreover, based on the highly similar evolutionary behaviors of TFBSs and inter-TFBSs spacer sequences, we provided genome-wide evidence for the continuum model of TF binding in CRMs. Additionally, as epigenomic marks determine the functional states of CRMs, thereby playing crucial roles in cell fate determination and type maintenance during cell differentiation, epigenomic marks can help to predict the functional states of CRMs. Although genomic sequences play a crucial role in establishing the unique epigenome in each cell type during cell differentiation, little is known about the sequence determinants that lead to the unique epigenomes of the cells. We developed two types of highly accurate deep convolutional neural networks (CNNs) for cell types and for histone marks. The results showed that they are powerful ways to uncover the sequence determinants of the various histone modification patterns in different cell types. We found that sequence motifs learned by the CNN models are highly like known binding motifs of TFs known to play important roles in cell

differentiation. Using these models, we can predict the importance of the learned motifs and their interactions in determining specific histone mark patterns in the cell types. Thus, the CNNs provide a way to pinpoint the influences of the motifs in epigenome marks. Finally, although several databases have been developed for predicted or experimentally determined enhancers/CRMs, they only cover a small portion of CRMs encoded in the genomes, lack constituent TFBSs, have high false positives, and are often dedicated to a single organism. To aid the use of the predicted CRMs and TFBSs by the research community, we developed a database dePCRMS (de novo predicted CRMs) (https://pcrms.uncc.edu). Currently, dePCRMS contains 1,155,151, 777,409 and 19,515 CRMs, and 89,948,206, 103,718,473, and 3,758,557 TFBSs, in *Homo sapiens*, *Mus musculus* and *Caenorhabditis elegans*, respectively. The users can use the web interface to quickly browse and visualize the CRMs and their constituent TFBSs at different significant levels in selected chromosomes in an organism. Moreover, the web interface provides three types of functional analysis modules for the user 1) to search the closest CRM to a gene, 2) to search CRMs in a given genome range around a gene, and 3) to search TFBSs in CRMs for a given TF. The dePCRMS database can be an informative tool for the users to characterize functions of regulatory genomes in important organisms.

DEDICATION

To my parents, To my advisor, To my wife and daughter

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor Dr. Zhengchang Su for his patience, encouragement, and support in my PhD career. He is an extraordinary advisor who guided and encouraged me to be professional. This dissertation would not have been realized without his guidance and persistent help. I would like to thank my committee members, Dr. Juntao Guo, Dr. Xinghua Mindy Shi, Dr. Baohua Song for their insightful suggestions and comments. I would like to thank the members in our group for their stimulating discussions. I would like to dedicate my gratitude to my family for their support. Finally, I would like to show my gratitude to the Graduate Assistant Support Plan (GASP) and NSF and NIH grants for their financial support. My thanks and appreciation also go to all staff members in Department of Bioinformatics and Genomics and the international student and scholar office (ISSO) for their supportive help.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CRM	*Cis*-regulatory module
TF	Transcription factor
CDS	coding sequence
NCS	non-coding sequence
SNP	single nucleotide polymorphism
LD	linkage disequilibrium
TFBS	TF binding site
CRMC	CRM candidate
UM	unique motif
CP	co-occurring motifs pair
NES	non-exonic sequence
ES	exonic sequence
TPR	true positive rate
FPR	false positive rate
FNR	false negative rate
FDR	false discovery rate
FOR	false omission rate
FP	FANTOM promoter
FE	FANTOM enhancer
DHS	dnase hypersensitivity site
TAS	transposase-accessible site
PPI	protein protein interaction
TSS	transcription starting site
CNN	convolutional neural networks
ReLU	rectified linear unit
ROC	receiver operating characteristic
AUC	under the curve
Tn	the native T cells
Tcm	the central memory T cells

Tem	the T effector memory cells
Temra	the terminally differentiated CD45RA+ memory cells
H1	H1 human embryonic stem cells
TBL	trophoblast-like cells
ME	mesendoderm cells
MSC	mesenchymal cells
NPC	neural progenitor cells
PWM	position weight matrices
M-Motif	Merged Motif
ChIP-seq	chromatin immunoprecipitation followed by sequencing
DNase-seq	DNase I hypersensitive sites sequencing
ATAC-seq	assay for transposase-accessible chromatin using sequencing
FAIRE-seq	formaldehyde-assisted isolation of regulatory elements sequencing
MNase-seq	micrococcal nuclease digestion with deep sequencing

CHAPTER 1: INTRODUCTION

Cis-regulatory modules (CRMs) consisting of clusters of transcription factor (TF) binding sites (TFBSs) play crucial roles in regulating the transcriptional patterns of their target genes, which eventually shape the complex phenotypes of the organisms. A comprehensive map of CRMs and their constituent TFBSs in important organisms could help the research community to elucidate the relationship between genotypes and phenotypes, thereby laying the foundation for precision medicine to prevent and treat the common complex diseases. However, the current understanding of the relationship between the causal CRM variants and the complex diseases/traits is limited due to the lack of an accurate, high resolution map of CRMs and constituent TFBSs in the human genome. With the development of next generation sequencing (NGS) technologies, enormous heterogeneous ChIP-seq data have been produced, which could provide an opportunity to predict a comprehensive map of CRMs and elucidate the underlying mechanisms of gene regulation from various perspectives. To address the challenges, this dissertation project has fulfilled three aims. Firstly, we developed a new pipeline called dePCRM2 to predict a comprehensive map of CRMs in human genome. Secondly, we developed an interpretation framework for exploring the effects of TFBSs on the epigenetic marks in cell differentiation using convolutional neural network (CNN) models. Finally, we built a database to hold the predicted comprehensive maps of CRMs and constituent TFBSs for three important organisms, and an interface to facilitate the research community to conduct the functional analysis, such as identifying the target genes, searching the comprehensive map of TFBSs for a specific TF, etc. This dissertation is organized as follows.

In chapter 1, we introduce the aims and organization of the dissertation.

In chapter 2, we predict a map of CRMs in human genome using the dePCRM2 pipeline. Firstly, we describe the current progress for predicting CRMs and the limitation of the previous works in

the background section. Then, we illustrate the dePCRM2 pipeline in details. Next, we demonstrate our hypothesis of the algorithm and the dataset features that support this hypothesis. We also show the comprehensive map and evolutionary behaviors of CRMs in human genome, and the performance comparison with the state of art algorithms. Finally, we interpret our results in the discussion section in this chapter.

In chapter 3, we propose that our predicted CRMs support the continuum model of TF binding in human genome. Firstly, we describe the historical hypothesis about the TF binding modes in CRMs in the background section. Then we classify the predicted CRMs according to their lengths and locations and provide the evidence that supports the continuum model of TF binding in full-length CRMs. Finally, we explain our results in more details in the discussion section.

In chapter 4, we propose two types of CNNs to predict the influences of TFBSs on the epigenetic marks in cell differentiation. In this chapter, we describe the background of cell differentiation, CNN and the previous works which attempt to apply CNN to explain the epigenetic marks in cell differentiation and their limitations. Then we demonstrate the data analysis pipeline, including the data preprocessing, data representation, construction of CNN, model training, validation, testing, and model interpretation, motif conservation analysis, and interaction prediction of the cognate TF of the learned motifs. Next, we evaluate the performance of the cell type CNN models which were applied in two cell differential models. We use four types of human CD4+ T cells for the sequential analysis and use the H1 human embryonic stem cells and another 4 derived cells for the generality and robustness evaluation. And then we predict the cell types in six histone mark models, including H3K27me3, H3K27ac H3K36me3, H3K4me1, H3K4me3 and H3K9me3. In the model interpretation section, we observe that a unique set of motifs could largely determine the combinational patterns of different histone marks and cell types, and the motifs learned in both models could reflect the lineage of the cell or functional relationships of the marks. Then, we explore the inference power of the learned motifs in both the cell type and the histone mark models.

Next, we also analyze the relationship between the evolutionary behavior of the motifs and their inference power in both the cell type models and the histone mark models. Finally, we predict the interactions between the cooperative TF pairs to determine the histone marks in the same cell type and distinguish cells in the same histone mark.

In chapter 5, we build a webserver called dePCRMS. It holds the comprehensive maps of CRMs and constituent TFBSs for three organisms, i.e., *Homo sapiens*, *Mus musculus* and *Caenorhabditis elegans*. Firstly, we introduce the existing databases that hold the annotated CRMs. Then we briefly describe the predicting pipeline and the technical implement details of the database in the methods section. Finally, we introduce the interface and the functional modules of the webserver.

Chapter 6 gives the conclusion.

CHAPTER 2: A MAP OF *CIS*-REGULATORY MODULES IN HUMAN GENOME

2.1. Background

Cis-regulatory sequences, also known as cis-regulatory modules (CRMs) (i.e., promoters, enhancers, silencers and insulators), are made of clusters of short DNA sequences recognized and bound by specific transcription factors (TFs), and their functional states are responsible for specific transcriptomes in various cell types in multi-cellular eukaryotes. A growing body of evidence indicates that CRMs are as important as coding sequences (CDSs) account for inter-species divergence and intra-species diversity in complex traits [1-6]. Recent genome-wide association studies (GWAS) have found that most complex trait-associated single nucleotide polymorphisms (SNPs) do not reside in CDSs, but rather lie in non-coding sequences (NCSs) [7, 8], and often overlap with or are in linkage disequilibrium (LD) with TF binding sites (TFBSs) in CRMs [9]. It has also been shown that complex trait-associated variants systematically disrupt TFBSs of TFs related to the traits [10], and that variation in TFBSs affects DNA binding, chromatin modification, transcription [11-15], and susceptibility to complex diseases [16-21] including cancer [22-29]. In principle, variation in a CRM may result in changes in the affinity and interactions between TFs and their binding sites, leading to alterations in histone modifications and target gene expressions in relevant cells. These alterations in molecular phenotypes can lead to changes in cellular and organ-related phenotypes among individuals of a species [30, 31]. However, it has been difficult to link non-coding variants to phenotypes [32-36], largely because of our lack of a good understanding of all CRMs and their constituent TFBSs in genomes.

Fortunately, the recent development of ChIP-seq techniques for locating TFBSs of a TF in the genomes of specific cell/tissue types [37-39] has led to the generation of enormous amounts of data by large consortia such as ENCODE [40-42] Roadmap Epigenomics [43, 44] and Genotype-Tissue

Expression (GTEx) [45, 46], as well as individual labs worldwide [47]. These increasing amounts of ChIP-seq data for various TFs in a wide spectrum of cell/tissue types provide an unprecedented opportunity to predict a map of CRMs and constituent TFBSs in the human genome. Many computational methods have been developed to explore these data at various levels [34, 48]. At the lowest level, as the large number of binding-peaks in a ChIP-seq dataset dwarf earlier motif-finding tools (e.g., MEME[49], WEEDER [50, 51], Seeder [52] and BioProspector [53], new motif-finders (e.g., Trawler [54], ChIPMunk [55], HMS [56], CMF [57], STEME [58], DREME [59], MEME-ChIP [60], MICSA [61], DECOD [62], RSAT [63], POSMO [64], XXmotif [65], EXTREME [66], FastMotif [67] and Homer [68]) have been designed. However, some of these tools (e.g. MEME-ChIP, MICSA and CMF) were designed to find motifs in very short sequences (~200bp) around the binding-peak summits in a limited number of selected binding peaks due to their slow speed. Some faster tools (e.g., Seeder, ChIPMunk, DECOD, RSAT, POSMO, Homer, DREME, and XXmotif) are based on the discriminative motif-finding schema[69] by finding overrepresented k-mers in a ChIP-seq dataset, but they often fail to identify TFBSs with subtle degeneracy. As TFBSs form CRMs for combinatory regulation [70-76] tools (such as SpaMo [73], CPModule [76], COPS [77], INSECT [78], CCAT [79] and others [80-82]) have been developed to identify multiple closely located motifs as CRMs in a single ChIP-seq dataset. However, these tools cannot predict CRMs containing novel TFBSs, because they all depend on a library of known motifs (e.g., TRANSFAC [83] or JASPAR [84]) to scan for cooperative TFBSs in binding peaks.

Methods for predicting CRMs based on multiple epigenetic marks have been developed using hidden Markov models [85-89], dynamic Bayesian networks[90, 91], time-delay neural networks random forest [92, 93], and support vector machines (SVMs) [48, 94]. Sequence features have also been used to predict tissue-specific enhancers using SVM [48, 95]. Several enhancer databases have also been compiled either by combining results of multiple methods [96-98] or by identifying overlapping regions of chromatin accessibility (CA) and multiple histone mark tracks in a

cell/tissue type [99-104]. In particular, the ENCODE 3 consortium recently identified 0.9 million candidate *cis*-regulatory elements (cCREs) by identifying overlapping regions of between 2.2 million DNase I hypersensitivity sites (DHSs) and active promoter histone marks H3K4me3, active enhancer mark H3K27ac and isolator mark CTCT peaks in various cell types [105]. Although CRMs predicted by these methods are often cell/tissue type-specific, their applications are limited to cell/tissue types for which the required datasets are available. Further, the results of these methods are quite inconsistent [98, 106-109], e.g., even the best-performing tools (DEEP and CSI-ANN) have only 49.8% and 45.2%, respectively, of their predicted CRMs overlap with the DHSs in Hela cells [48], and although 1.3 million enhancers have been predicted using epigenetic marks [110], few disease-associated non-coding SNPs map to them [111]. Although some predictions provide TFBSs information by finding matches to known motifs in predicted CRMs [97, 98, 103], these methods were unable to identify TFBSs of novel motifs in CRMs.

Surprisingly, while TF ChIP-seq data provide more accurate information for TF-binding locations and their combinatory patterns in CRMs than epigenetic marks [48, 107, 109, 112, 113], few efforts have been made to fully explore the increasing volume of datasets due to technical difficulty to integrate them [112-115]. With this recognition, we have previously proposed a different strategy to first predict a catalog or a static map of CRMs and constituent TFBBs in the genome by integrating all available TF ChIP-seq datasets for different TFs in various cell/tissue types [112, 113] as has been done for identifying all genes encoded in the genome [116]. Once a map of CRMs and constituent TFBSs is available, the specificity of CRMs in any cell/tissue type can be determined using two epigenetic mark datasets collected in the cell/tissue type, reflecting their functional states as demonstrated recently [105]. Although very promising results have been obtained using even insufficient datasets available then [112, 113], there are three limitations in our earlier algorithm dePCRM to integrate much larger datasets available now and in the future. First, although existing motif-finders such as DREME used in dePCRM work well for relatively small

ChIP-seq datasets from organisms with smaller genomes such as the fly [113], they are too slow for very large datasets from human cells/tissues, so we had to split large datasets into smaller ones for the motif finding [112], which may compromise the accuracy of motif finding and complicate subsequent data integration. Second, the distance and interactions of TFBSs in a CRM were not explicitly considered [112], potentially limiting the accuracy. Third, the original "branch-and-bound" approach to integrate motifs is not efficient enough to handle much larger number of motifs found in ever increasing number of large ChIP-seq datasets from human cells/tissues. To overcome these drawbacks, we developed a new CRM predictor dePCRM2 that combines an ultrafast, accurate motif-finder ProSampler [117] with a novel effective combinatory motif pattern discovery method. Applying dePCRM2 to available 6,092 ChIP-seq datasets covering 77.47% of the human genome, we predicted 201 unique TF binding motif families and 1,404,973 CRM candidates (CRMCs). Both evolutionary and independent experimental data indicate that dePCRM2 achieves very high sensitivity and specificity in predicting CRMs and constituent TFBSs.

2.2. Methods and materials

2.2.1. Datasets

We downloaded 6,092 TF ChIP-seq datasets (SUPPLEMENTARY TABLE S1) from the Cistrome database [47]. The binding peaks in each dataset were called using a pipeline for uniform processing [47]. We filtered out binding peaks with a read depth score less than 20. For each binding peak in each dataset, we extracted a 1,000 bp genome sequence centering on the middle of the summit of the binding peak. We downloaded 976 experimentally verified enhancers from the VISTA Enhancer database [118], 32,689 enhancers and184,424 promoters from the FANTOM project website[119, 120], 424,622 ClinVar SNPs from the ClinVar database [121], 91,369 GWAS SNPs from GWAS Catalog [122, 123], and 122,468,173 DHSs in 1353 datasets (SUPPLEMENTARY TABLE S2), 29,520,736 transposase-accessible sites(TASs) in 1,059 datasets (SUPPLEMENTARY TABLE S3), 99,974,447 H3K27ac peaks in 2,539 datasets (SUPPLEMENTARY TABLE S4), 77,500,232 H3K4me1 peaks in 1,210 datasets (SUPPLEMENTARY TABLE S5), and 70,591,888 H3K4me3 peaks in 2,317 datasets (SUPPLEMENTARY TABLE S6) from the Cistrome database [47].

2.2.2. Measurement of the overlap between two different datasets

To evaluate the extent to which the binding peaks in two datasets overlap with each other, we calculate an overlap score $S_0(d_i, d_j)$ between each pair of datasets d_i and d_j , which is defined as,

$$S_0(d_i, d_j) = \frac{1}{2} \times \left(\frac{o(d_i + d_j)}{|d_i|} + \frac{o(d_i + d_j)}{|d_j|} \right). \quad 2\text{-}1$$

2.2.3. Parameters for accuracy evaluation

We use the following definition of the parameters for evaluating the accuracy of datasets of predictions. Sensitivity = recall rate = TPR (true positive rate) = $\frac{TP}{TP+FN}$, Specificity = $\frac{TN}{FP+TN}$, FNR (false negative rate) = $\frac{FN}{TP+FN}$, FPR(false positive rate) = $\frac{FP}{FP+TN}$, FDR (false discorery rate) = $\frac{FP}{TP+FP}$, FOR (false ommision rate) = $\frac{FN}{FN+TN}$, where TP is true positives; FN is false negatives; FP is false positives; and TN is true negatives.

2.2.4. The dePCRM2 algorithm

Step 1. Find motifs in each dataset using ProSampler [117] (FIGURE 2-1A and B).

Step 2. Compute pairwise motif co-occurring scores and find co-occurring motif pairs: As True motifs are more likely to co-occur in the same sequence than the spurious ones, to filter out false positive motifs, we find overrepresented co-occurring motif pairs (CPs) in each dataset (FIGURE 2-1C). Specifically, for each pair of motifs $M_d(i)$ and $M_d(j)$ in each data set d, we compute their co-occurring scores S_c defined as,

$$S_c\left(M_i(i), M_j(j)\right) = \frac{o\left(M_d(i), M_d(j)\right)}{max\{|M_d(i)|, |M_d(i)|\}}, \qquad 2\text{-}2$$

where $|M_d(i)|$ and $|M_d(j)|$ are the number of binding peaks containing TFBSs of motifs $M_d(i)$ and $M_d(j)$, respectively; and $o\left(M_d(i), M_d(j)\right)$ is the number of binding peaks containing TFBSs of both the motifs in d. We identify CPs with an $S_c \geq 0.7$ (by default) (FIGURE 2-1C).

Step 3. Construct a motif similarity graph and find unique motifs: We combine highly similar motifs in the CPs from different datasets to form a unique motif (UM) presumably recognized by a TF or highly similar TFs of the same family/superfamily [124]. Specifically, for each pair of motifs $M_a(i)$ and $M_b(i)$ from different datasets a and b, respectively, we compute their similarity score S_s using our SPIC [125] metric. We then build a motif similarity graph using motifs in the

CPs as nodes and connecting two motifs with their S_s being the weight on the edge, if and only if (iff) $S_s > \beta$ (by default, $\beta = 0.8$, FIGURE 2-1D). We apply the Markov cluster (MCL) algorithm [126] to the graph to identify dense subgraphs as clusters. For each cluster, we merge overlapping sequences, extend each sequence to a length of 30bp by padding the same number of nucleotides from the genome to the two ends, and then realign the sequences to form a UM using ProSampler (FIGURE 2-1D).

Step 4. Construct the integration networks of UMs: TFs tend to repetitively cooperate with each other to regulate genes in different contexts by binding to their cognate TFBSs in CRMs. The relative distances between TFBSs in a CRM often do not matter (billboard model) [127-129] but sometimes they are constrained by the interactions between cognate TFs. To model both scenarios, we compute an interaction score between each pair of UMs,

$$S_{\text{INTER}}(U_i, U_j) \frac{1}{|D(U_i,U_j|} \sum_{d \in D(U_i,U_j)} \left(\frac{1}{|U_i(d)|} \sum_{s \in S(U_i(d),U_j(d))} \frac{150}{r(s)} + \frac{1}{|U_j(d)|} \sum_{s \in S(U_i(d),U_j(d))} \frac{150}{r(s)} \right), \quad 2\text{-}3$$

where $D(U_i, U_j)$ is the datasets in which TFBSs of motifs U_i and U_j are found, $U_k(d)$ is the subset of dataset d, containing at least one TFBS of U_k, $S(U_i(d), U_j(d))$ is the subset of d containing TFBSs of both U_i and U_j, and $r(s)$ is the shortest distance between a TFBS of U_i and a TFBS of U_j in a sequence s. We construct UM/TF interaction networks using the UMs as nodes and connecting two nodes with their S_{INTER} being the weight on the edge (FIGURE 2-1E). Therefore, S_{INTER} allows flexible adjacency and orientation of TFBSs in a CRM, and at the same time, it rewards motifs with binding sites co-occurring frequently in a shorter distance in a CRM [127-129].

Step 5. Evaluate CRM candidates: We project TFBSs of each UM back to the genome and link two adjacent TFBSs if their distance $d \leq 300$bp (roughly the length of DNA in two nucleosome). The resulting linked DNA segments are CRM candidates (CRMCs) (FIGURE 2-1F). We evaluate each CRMC containing n TFBSs ($b_1, b_2, \ldots, b_n$) by computing a CRM score defined as,

$$S_{CRM}(b_1, b_2 \cdots, b_n) = \frac{2}{n-1} \times \sum_{i=1}^{n} \sum_{j>i} W\left[U(b_i), U(b_j)\right] \times \left[S(b_i) + S(b_j)\right], \qquad 2\text{-}4$$

where $U(b_k)$ is the UM of TFBS b_k, $W\left[U(b_i), U\left(b_j\right)\right]$ is the weight on the edge between the motifs of $U(b_i)$ and $U\left(b_j\right)$ in the interaction networks, and $S(b_k)$ is the binding affinity score of b_k based on the position weight matrix (PWM) of $U(b_k)$. Only TFBSs with a positive score are considered.

Step 6. Predict CRMs: We create the Null interaction networks by randomly reconnecting the nodes with the edges in the interaction networks constructed in Step 4. For each CRMC, we generate a Null CRMC that has the same length and nucleotide compositions as the CRMC using a third order Markov chain model [117]. We compute a S_{CRM} score for each Null CRMC using the Null interaction networks, and the binding site positions affinity with the UMs based on their PWMs in the corresponding CRMC. Based on the distribution of the S_{CRM} scores of the Null CRMCs, we compute a p-value for each CRMC, and predict those with a p-value smaller than a preset cutoff as CRMs in the genome (FIGURE 2-1G).

Step 7. Prediction of the functional states of CRMs in a given cell type: For each predicted CRM, we predict it to be active in a cell/tissue type, if its constituent binding sites of the UMs whose cognate TFs were tested in the cell/tissue type overlap the original binding peaks of the TFs; otherwise, we predict the CRM to be inactive in the cell/tissue type. If the CRM does not overlap any binding peaks of the TFs tested in the cell/tissue type, we assign its functional state in the cell/tissue type as “to be determined” (TBD).

2.2.5. Generation of control sequences for validation

To create a set of matched control sequences for validating the predictions for each predicted CRMC, we produced a control sequence by randomly selecting a sequence segment with the same

length as the CRMC from the genome regions covered by the extended binding peaks. To calculate the S_{CRM} score of a control sequence, we assigned it the TFBS positions and their UMs according to those in the counterpart CRMC. Thus, the control set contains the same number and length of sequences as in the CRMCs, but with arbitrarily assigned TFBSs and UMs.

2.3. Results

2.3.1. The dePCRM2 pipeline

TFs in eukaryotes tend to cooperatively bind to their TFBSs in CRMs[130]. Different CRMs of the same gene are structurally similar and closely located, and functionally related genes are often regulated by the same sets of TFs in different cell types during development and in maintaining homeostasis[131]. Thus, if we extend the called binding peaks of a TF ChIP-seq dataset from the two ends and reach the typical size of a CRM (500~3000bp)[118], then the extended peaks may include TFBSs of cooperative TFs [112, 113]. For instance, if two different TFs regulate the same group of target genes cooperatively in several cell types, then at least some of the extended peaks of datasets for the two TFs from these cell types should contain the TFBSs of both TFs, or even have some overlaps if the CRMs are reused in different cell types. Therefore, with sufficient amount of various TFs from different cell types are produced, the datasets for some cooperative TFs are likely to be included, and their TFBSs of the cooperative TFs may co-occur in some extended peaks. Based on these observations, we designed the pipelines dePCRM [112, 113] and dePCRM2 to predict CRMs and constituent TFBSs by identifying overrepresented co-occurring patterns of motifs found by a motif-finder in a large number of TF ChIP-seq datasets[112, 113]. We overcome the aforementioned shortcomings of dePCRM [112, 113] as follows. First, using an ultrafast, accurate motif-finder ProSampler [117], dePCRM2 can find significant motifs in any size of available ChIP-seq datasets without the need to split large datasets into small ones (FIGURE 2-1A and B). Second, after identifying highly co-occurring motifs pairs (CPs) in the extended binding peaks in each dataset (FIGURE 2-1C), dePCRM2 clusters highly similar motifs in the CPs and finds a unique motif (UM) in each resulting cluster (FIGURE 2-1D). Third, dePCRM2 models interactions among cognate TFs of the binding sites in a CRM by constructing interaction networks of the UMs based on the distance between the binding sites and the extent to which biding sites in the UMs cooccur (FIGURE 2-1E). Fourth, dePCRM2 identifies as CRMCs closely located clusters

of binding sites of the UMs along the genome (FIGURE 2-1F), thereby partitioning genome regions covered by the extended binding peaks in the datasets into a CRMCs set and a non-CRMCs set. Fifth, dePCRM2 evaluates each CRMC using a novel score that considers the quality of binding sites as well as the strength of interactions among the corresponding UMs defined in the interaction networks (FIGURE 2-1G). Lastly, dePCRM2 computes a p-value for each S_{CRM} score, so that CRMs and constituent TFBSs can be predicted at different significant levels using different S_{CRM} score or p-value cutoffs. Clearly, as the number of UMs is a small constant number constrained by the number of TF families encoded in the genome, the downstream computation based on the set of UMs runs in a constant time, thus dePCRM2 is highly scalable.

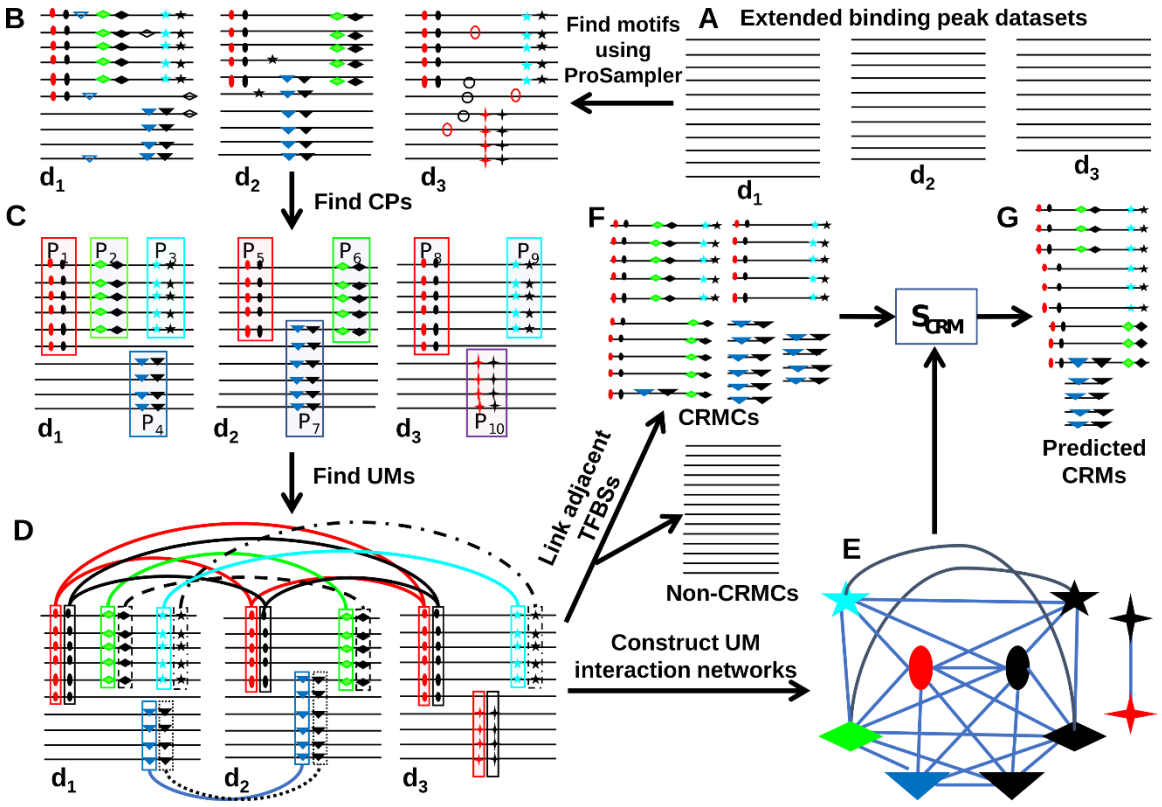

FIGURE 2-1: Schema of the dePCRM2 algorithm. A. Extend each binding peak in each dataset to its two ends to reach a preset length, e.g., 1,000bp. B. Find motifs in each dataset using ProSampler. C. Find CPs in each dataset. For clarity, only the indicated CPs are shown, while those formed between motifs in pairs P1 and P2 in d1, and so on, are omitted. D. Construct the motif similarity graph, cluster similar motifs and find UMs in the resulting motif clusters. Each node in the graph represents a motif, weights on the edges are omitted for clarity. Clusters are connected by edges of the same color and line type. E. Construct UM interaction networks. Each node in the networks represents a UM, weights on the edges are omitted for clarity. F. Project binding sites in the UMs back to the genome and identify CRMCs along the genome. G. Evaluate each CRMC by computing its S_{CRM} score and the associated p-value.

2.3.2. Extended binding peaks in different datasets have extensive overlaps

After filtering out low-quality peaks in the 6,092 ChIP-seq datasets, we ended up with 6,070 non-empty datasets for 779 TFs in 2,631 cell/tissue/organ types. The datasets are strongly biased to few cell types (FIGURE 2-2A). For example, 532, 475 and 309 datasets were collected from mammary gland epithelium, colon epithelium, and bone marrow erythroblast, respectively, while only one dataset was generated from 129 cell/tissue types, including heart embryonic fibroblast, fetal skin fibroblast, and bone marrow haematopoietic progenitor, and so on. The datasets also are strongly biased to few TFs (FIGURE 2-2B). For example, 370 and 263 datasets were collected for TFs CTCF and ESR1, respectively, while just one dataset was produced for 324 TFs, such as MSX2, RAX2, and MYNN, and so on. The number of called binding peaks in a dataset is highly varying, ranging from 2 to 100,539, with an average of 19,314 (FIGURE 2-2C). For instance, datasets for STAT1, and NR3C1 have the smallest number of 2 binding peaks in HeLa-S3, and HEK293 cells, respectively, while datasets for CEBPB, BRD4 and FOXA2 have the largest number of 115,776, 99,646, and 99,512 binding peaks in HepG2, U87, and Mesenchymal Stem Cells, respectively. The highly varying numbers of binding peaks in the datasets suggest that different TFs might bind a highly varying number of sites in the genomes of cells. However, some datasets with very few binding peaks might be resulted from technical artifacts, thus are of low quality (see below), even though they passed our first quality filter. The lengths of binding peaks in the datasets range from 75 to 10,143bp, with a mean of 300bp (FIGURE 2-2D), and 99.12% of binding peaks are shorter than 1,000pb. All the binding peaks in the 6,070 datasets cover a total of 1,265,474,520bp (40.98%) of the genome (3,088,269,832bp). For each binding peak in each dataset, we extracted a 1,000bp genome sequence centering on the middle of its binding summit, thereby extending the lengths for most binding peaks. We have shown that extension of binding peaks to 500~1,000bp could substantially increase the chance of finding TFBSs of cooperative TFs of the ChIP-ed/target TFs, while the introduced noise had a little effect on identifying the primary motifs of target TFs [117]. The extended binding peaks contain a total of 115,710,048,000bp, which is 37.5 times the size of

the genome, but cover only 2,392,488,699bp (77.47%) of the genome, leaving the remaining 22.53% of the genome uncovered, indicating that they have extensive overlaps. Nonetheless, by extending the original binding peaks, we increased the coverage of genome by 89.04% (77.47% vs 40.98%). Notably, we may not know the functional states of some predicted binding sites in extended parts of the original binding peaks from a cell/tissue type if no binding peaks for other TFs tested in the cell/tissue type overlap the extended parts. We trade this drawback for a more complete prediction of the catalogs/map of CRMs and TFBSs in the genome [112]. We expect that when more diverse, less biased data for untested TFs in untested cell/tissue types are generated in the future, a larger proportion of the functional genome can be covered. As dePCRM2 predicts CRMs and constituent TFBSs based on overlapping patterns between datasets of cooperative TFs [112, 113] (Methods and materials), we evaluated the extent to which the extended binding peaks in different datasets overlap one another. To this end, we hierarchically clustered the 6,070 datasets using an overlap score between each pair of the datasets (Methods and materials). As shown in FIGURE 2-3A, there are extensive distinct overlapping patterns among the datasets. As expected, clusters are formed by datasets for largely the same TF in different cell/tissue types, and/or by datasets for different TFs that are known or potential collaborators in transcriptional regulation. For instance, a cluster is formed by 1, 1 and 48 datasets for RAD21, SMC3 and CTCF, respectively, in various cell/tissue types (FIGURE 2-3B). It is well-known that RAD21, SMC3 and CTCF are core subunits of the cohensin complex, and are widely colocalized in mammalian genomes [132]. In another example (FIGURE 2-3C), a cluster is formed by 50 datasets for 45 TFs in various cell/tissue types. Multiple sources of evidence indicate that these 45 TFs have extensive physical interactions for DNA binding and transcriptional regulation (FIGURE 2-3D) [133, 134]. For example, it has been shown that LEF1 interacts with the TGF beta activating regulator SMAD4 [135], E2F4 helps to recruit PML to the TBX2 promoter [136], and FOXK 1 and TP53 can form a distinct protein complex on and off chromatin [137]. These overlapping patterns between the extended binding peaks in the datasets warrant us to predict CRMs and constituent TFBSs in the covered 77.47% of the genome

[112, 113](Methods and materials). In addition, 10 sets of experimentally determined CRM function-related elements (Methods and materials) highly enriched in the covered 77.47% regions compared to the uncovered 22.53% genome regions, including 785 (80.43%) VISTA enhancers [118], 402,730 (94.84%) of ClinVar single nucleotide polymorphisms (SNPs) [121, 138], 181,436 (98.38%) FANTOM promoters (FPs) [119], 32,029 (97.98%) FANTOM enhancers (FEs) [120], 82,378 (90.16%) of GWAS SNPs [123], 121,075,184 (98.86%) DNase I hypersensitive sites (DHSs) [122, 123], 29,195,778, 98,297,240, 7,5467,050, 69,282,044 transposase-accessible sites(TASs)[139](98.90%), H3K27ac peaks [140](98.32%), H3K4me1 peaks[141] (97.38%), and H3K4me3 peaks[141](98.14%). We will evaluate the sensitivity of dePCRM2 to recall these elements at different S_{CRM} scores and associated p-value cutoffs.

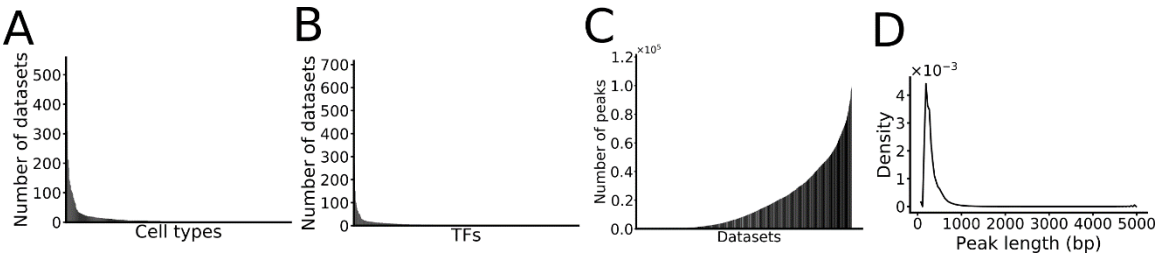

FIGURE 2-2: Properties of the datasets. A. Number of datasets collected in each cell/tissue types sorted in descending order. B. Number of datasets collected for each TF sorted in descending order. C. Number of peaks in each dataset sorted in ascending order. D. Distribution of the lengths of binding peaks in the entire datasets.

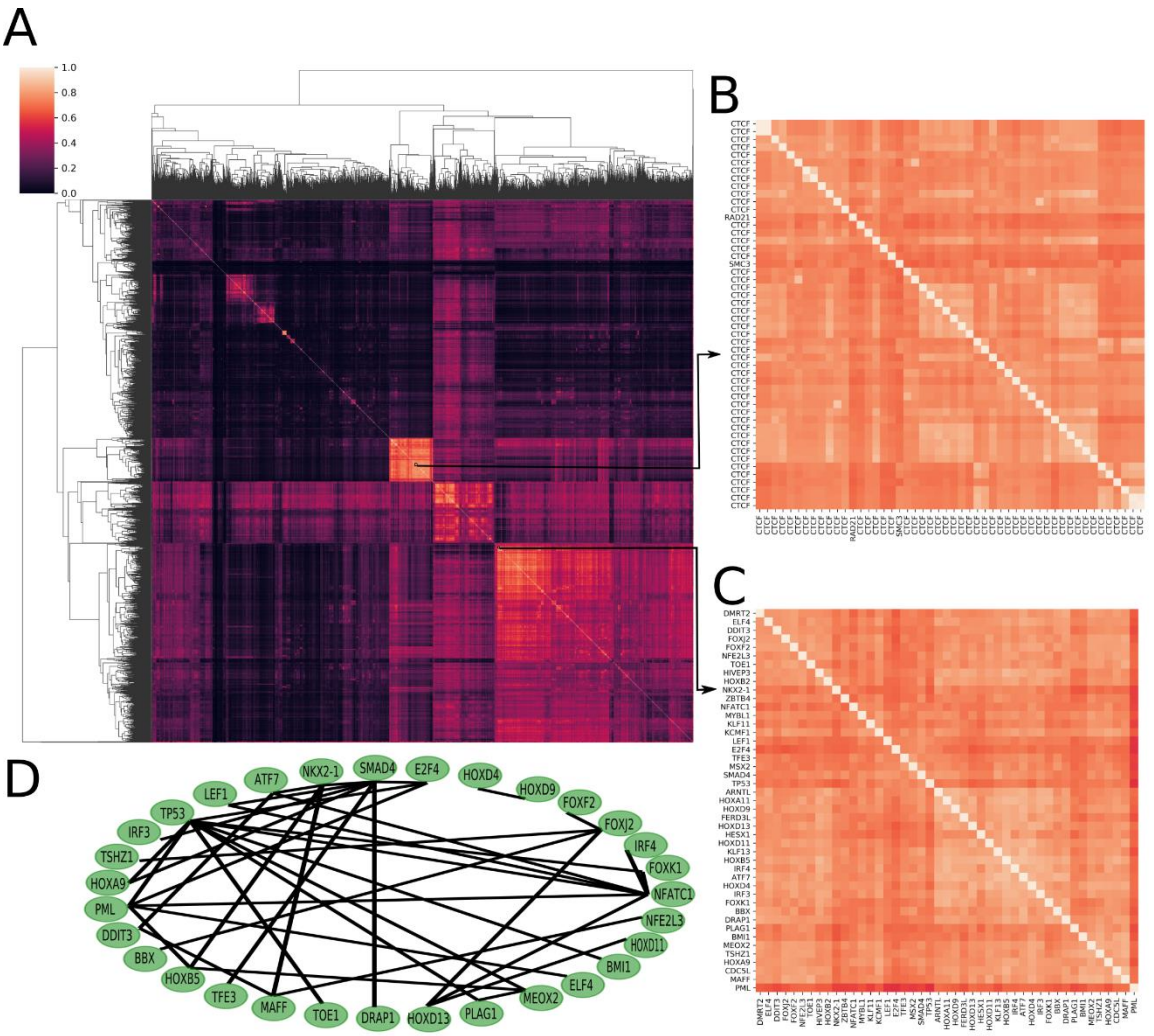

FIGURE 2-3: Overlap of extended binding peaks. A. Heatmap of overlaps of extended binding peaks between each pair of the datasets. B. A blowup view of the indicated cluster in, formed by 48 datasets for CTCF in different cell/tissue types, as well as one dataset for each of its two collaborators, RAID21 and SMC3. C. A blowup view of the indicated cluster in D, formed by 50 datasets for 45 TFs. B. Known physical interactions between the 45 TFs whose 50 datasets form the cluster in E.

2.3.3. Unique motifs recover most known TF motifs families

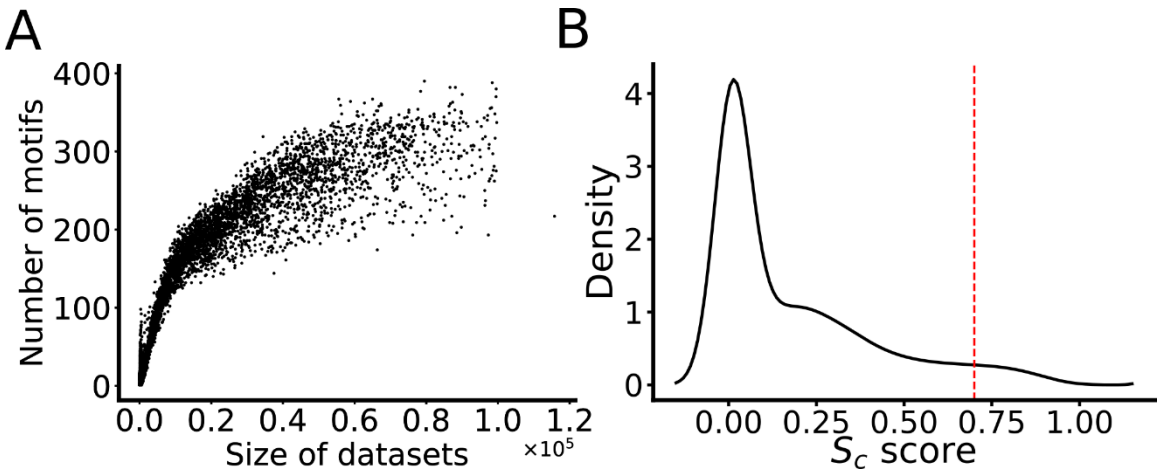

FIGURE 2-4: Identification of motifs and Sc score. A. Relationship between the number of predicted motifs in a dataset and the size of the dataset (number of binding peaks in the dataset). The datasets are sorted in ascending order of their sizes. B. Distribution of cooccurrence scores (Sc) of motif pairs found in a dataset. The dotted vertical line indicates the cutoff value of Sc for predicting cooccurring pairs (CPs).

ProSampler identified at least one motif in 5,991 (98.70%) datasets but failed to find any motifs in the remaining 79 (1.30%) datasets that all contain less than 310 binding peaks, indicating that they are likely of low quality. As shown in FIGURE 2-4A, the number of motifs found in a dataset generally increases with the increase in the number of binding peaks in the dataset, but enters a saturation phase and stabilizes around 250 motifs when the number of binding peaks is beyond 40,000. In total, ProSampler identified 856,793 motifs in the 5,991 datasets with at least one motif found. dePCRM2 finds co-occurring motif pairs (CPs) in each dataset (FIGURE 2-1C) by computing a cooccurring score S_c for each pair of motifs in the dataset (Formula 2-2). As shown in FIGURE 2-4B, S_c scores show a three-mode distribution. dePCRM2 selects as CPs the motif pairs that account for the mode with the highest S_c scores, and discards those that account for the other two modes with lower S_c scores (by default, S_c<0.7), because these low-scoring motif pairs are likely to co-occur by chance. In total, dePCRM2 identified 4,455,838 CPs containing 226,355 (26.4%) motifs from 5,578 (93.11%) of the 5,991 datasets. Therefore, we filtered out 413 (6.89%) of the 5,991 datasets because each had a low S_c score compared with other datasets. Clearly, more and balanced datasets are needed to rescue their use in the future for more complete predictions.

Clustering the 226,355 motifs in the CPs resulted in 245 clusters consisting of 2~72,849 motifs, most of which form a complete similarity graph or clique, indicating that member motifs in a cluster are highly similar to each other (FIGURE 2-5). dePCRM2 found a UM in 201 of the 245 clusters (FIGURE 2-6 and SUPPLEMENTARY TABLE S7-8) but failed to do so in 44 clusters due to the low similarity between some member motifs (FIGURE 2-5). Binding sites of the 201 UMs were found in 39.87~100% of the sequences in the corresponding clusters, and in only 1.49% of the clusters binding sites were not found in more than 50% of the sequences due to the low quality of member motifs (FIGURE 2-7). Thus, this step retains most of putative binding sites in most clusters. The UMs contain highly varying numbers of binding sites ranging from 64 to 13,672,868 with an average of 905,288 (FIGURE 2-8A and SUPPLEMENTARY TABLE S7-8), reminiscent of highly varying number of binding peaks in the datasets (FIGURE 2-8A). The lengths of the UMs range from 10bp to 21pb with a mean of 11pb (FIGURE 2-8B), which are in the range of the lengths of known TF binding motifs, although they are biased to 10bp due to the limitation of the motif-finder to find longer motifs. As expected, a UM is highly similar to its member motifs that are highly similar to each other (FIGURE 2-5). For example, UM44 contains 250 highly similar member motifs (FIGURE 2-9A). Of the 201 UMs, 117 (58.2%) match at least one of the 856 annotated motifs in the HOCOMOCO [142] and JASPAR [143] databases (SUPPLEMENTARY TABLE S7-8), and 92 (78.63%) match at least two, suggesting that most UMs might consist of motifs of different TFs of the same TF family/superfamily that recognize highly similar motifs, a well-known phenomenon [144, 145]. Thus, a UM might represent a motif family/superfamily for the cognate TF family/superfamily. For instance, UM44 matches known motifs of nine TFs of the “ETS” family ETV4~7, ERG, ELF3, ELF5, ETS2 and FLI1, a known motif of NFAT5 of the “NFAT-related factor” family, and a known motif of ZNF41 of the “more than 3 adjacent zinc finger factors” family (FIGURE 2-9B and SUPPLEMENTARY TABLE S7-8). The high similarity of these motifs suggest that they might form a superfamily. On the other hand, 64 (71.91%) of the 89 annotated motifs TF families match one of the 201 UMs (SUPPLEMENTARY TABLE S7-8), thus our

predicted UMs include most of the known TF motif families. To model interactions between cognate TFs of the UMs, dePCRM2 computes interaction scores between the UMs (Formula 2-3). As shown in (FIGURE 2-10A), there are extensive interactions between the UMs, which indeed reflect the interactions among their cognate TFs or TF families in transcriptional regulation. For example, in a cluster formed by 10 UMs (FIGURE 2-10B), seven of them (UM126, UM146, UM79, UM223, UM170, UM103 and UM159) match known motifs of MESP1/ZEB1, TAL1::TCF3, ZNF740, MEIS1/TGIF1/MEIS2/MEIS3, TCF4/ZEB1/CTCFL/ZIC1/ZIC4/SNAI1, GLI2/GLI3 and KLF8, respectively. At least a few of them are known collaborators in transcriptional regulation. For example, GLI2 cooperates with ZEB1 for repressing expression of CDH1 gene in human melanoma cells via direct binding two close binding sites at CDH promoter [146] , ZIC and GLI cooperatively regulate neural and skeletal development through physical interactions between their zinc finger domains [147], and ZEB1 and TCF4 could regulate the transcription of WNT target gene reciprocally[148], to name a few.

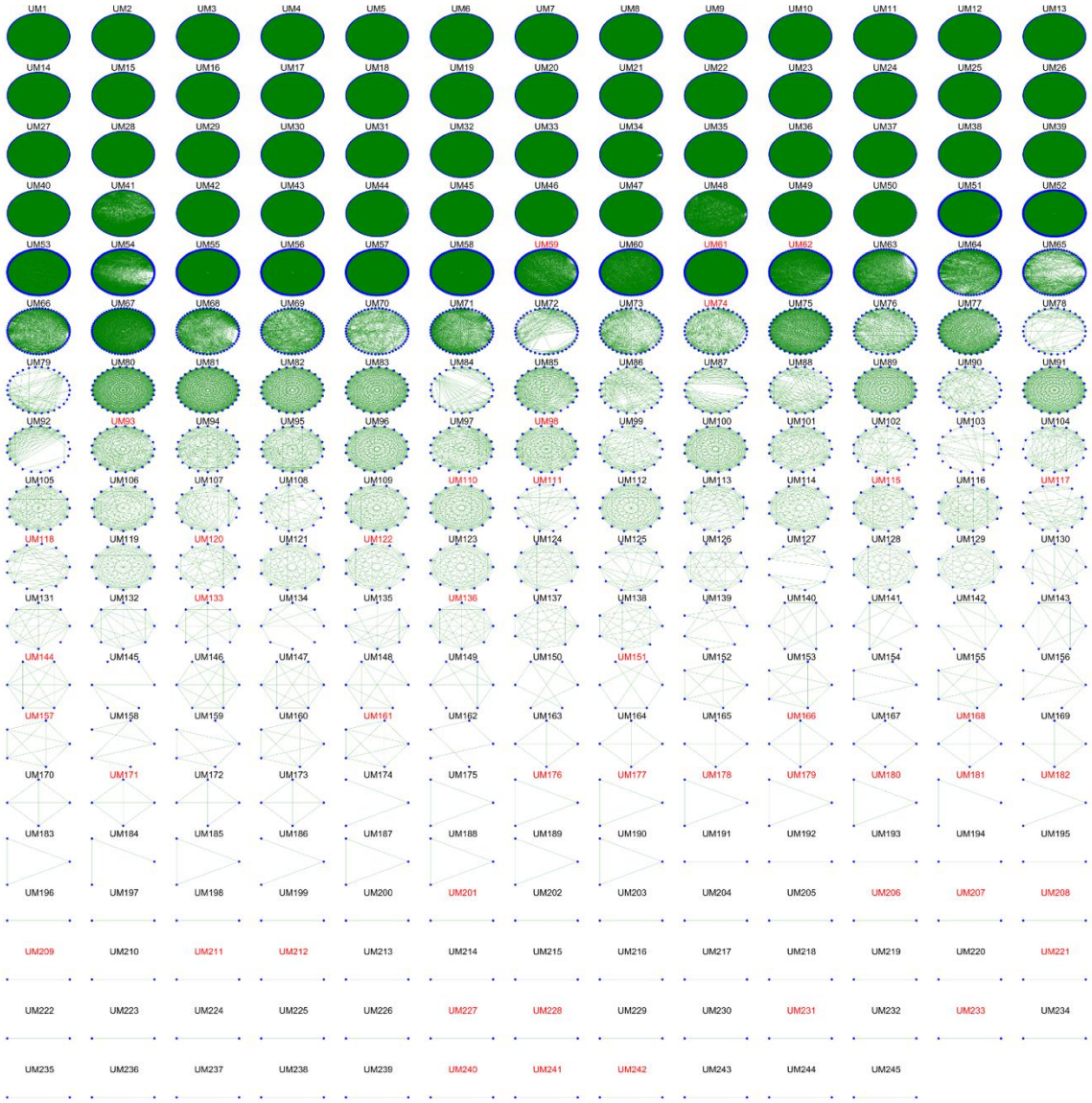

FIGURE 2-5: Graphs of member motifs of UMs. A. Similarity graphs of member motifs in the 245 motif clusters. In each graph, a node in blue represents a member motif of the cluster, and two member motifs are connected by an edge in green if their similarity is greater than 0.8 (SPIC score). Clusters with the names in RED font are those in which a UM cannot be found.

FIGURE 2-6: Logos of the UMs. A. Logos of the 201 UMs found in the corresponding clusters.

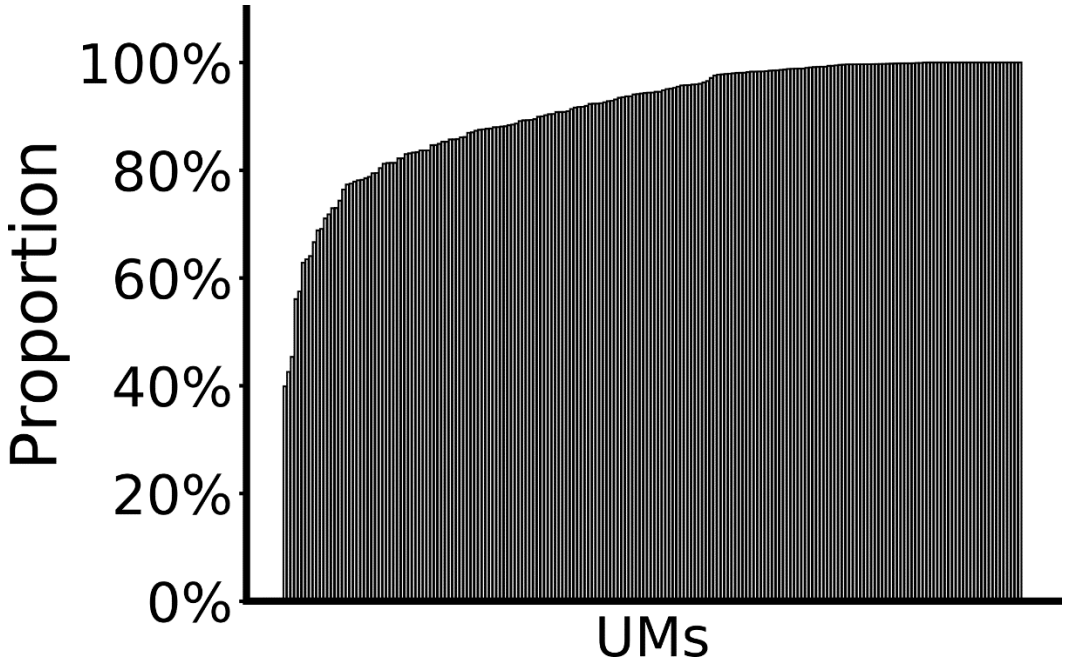

FIGURE 2-7: Recovery rate of the UMs. . Proportion of sequences of the member motifs of a UM in which binding sites were found. UMs are sorted in ascending order of the proportion.

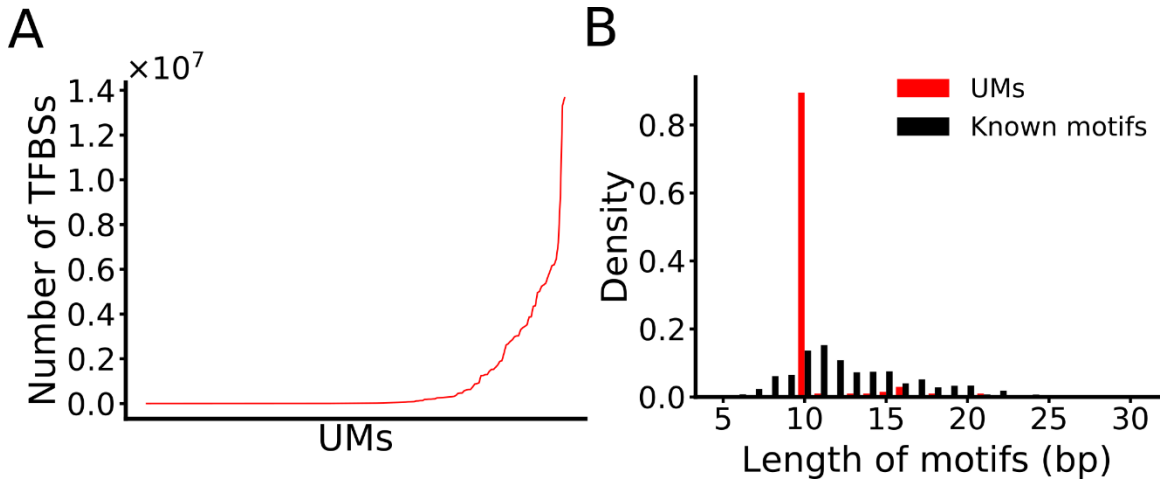

FIGURE 2-8: Properties of the UMs. . A. Number of putative binding sites in each of the UMs sorted in ascending order. B. Distribution of the lengths of the UMs and known motifs in the HOCOMOCO and JASPAR databases.

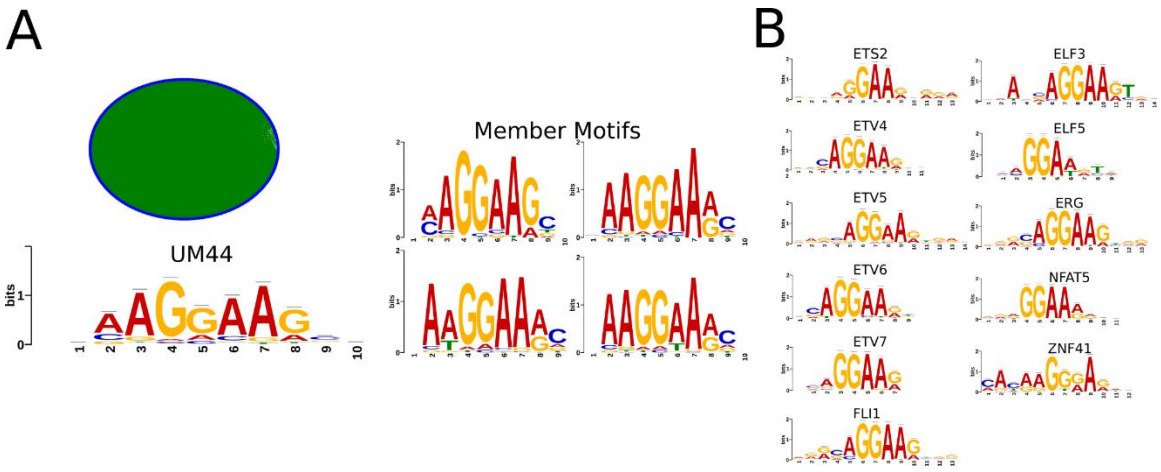

FIGURE 2-9: Example of a UM. A.The logo and similarity graph of the 250 member motifs of UM44. In the graph, each node in blue represents a member motif, and two member motifs are connected by an edge in green if their similarity is greater than 0.8 (SPIC score). Four examples of member motifs are shown in the left panel. B. UM44 matches known motifs of nine TFs of the "ETS", "NFAT-related factor", and "more than 3 adjacent zinc finger factors" families.

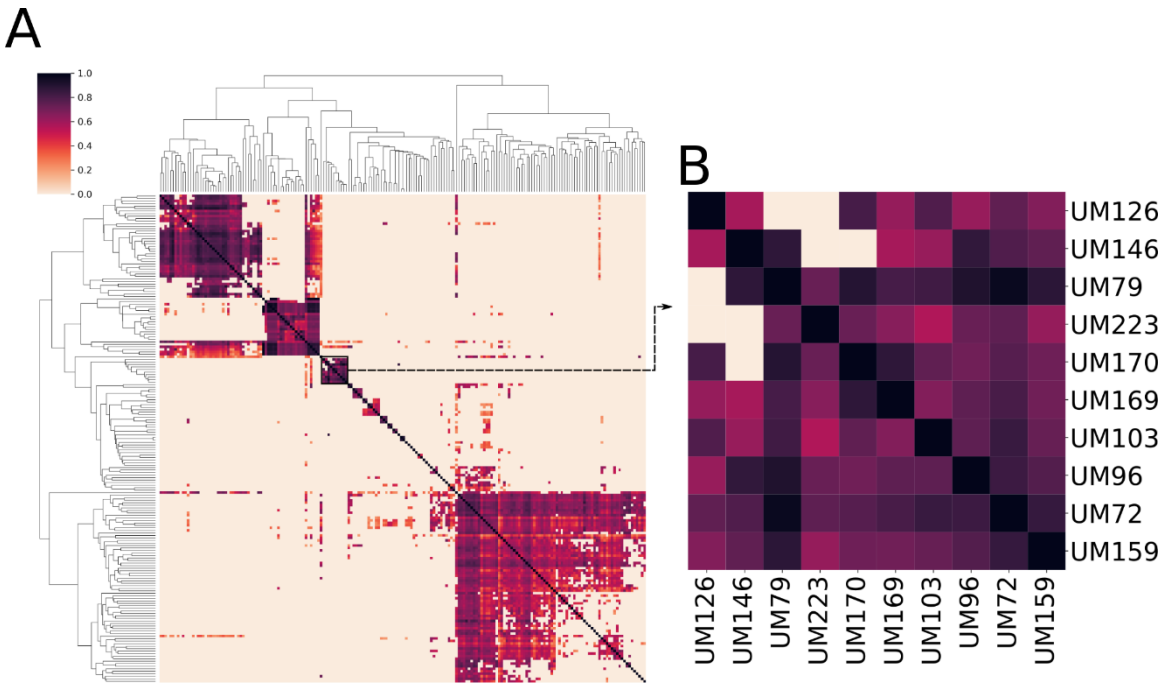

FIGURE 2-10: Interactions between the UM pairs. A.Heatmap of the interaction networks of the 201 UMs, names of most UMs are omitted for clarity

2.3.4. Extension of original binding peaks increase the power of datasets

By concatenating closely located binding sites of the UMs along the genome, dePCRM2 partitions the extended binding peak-covered genome regions (2,392,488,699bp) in two exclusive sets FIGURE 2-1, the CRMCs set containing 1,404,973 CRMCs with a total length of 1,359,824,275bp (56.84%) and the non-CRMCs set containing 1,957,936 sequence segments with a total length of 1,032,664,424bp (43.16%), covering 44.03% and 33.44% of the genome, respectively. Interestingly, 57.88% (776,999,862bp) of genome nucleotide positions of the CRMCs overlap those of the original peaks, while the remaining 42.12% (565,448,583bp) overlap those of the extended parts of the original peaks. Hence, in predicting CRMCs, dePCRM2 used only 61.40% of positions and abandoned the remaining 38.60% positions covered by the original binding peaks (1,265,512,389bp), suggesting that this portion of called binding peak position might not enrich for binding sites, which is in agreement with earlier studies [149-151]. Meanwhile, dePCRM2 predicted 565,448,583pb (42.12%) CRMC positions covered by the extended parts of original binding peaks, suggesting that that TFBSs of cooperative TFs are enriched in the extended parts as we showed earlier [117]. Thus, appropriate extension of original binding peaks could greatly increase the power of datasets. If a CRMC overlaps an original binding peak in a cell/tissue type, we predict the CRMC to be active in the cell/tissue type. Thus, we could predict functional states of about 57.88% of the CRMCs in at least one of the cell/tissue types, from which data were used in the prediction. However, we could not predict functional states of the remaining 42.12% of the CRMCs that do not overlap any original binding peaks.

2.3.5. Most CRMCs have low p-values

As shown in FIGURE 2-11A, the distribution of the S_{CRM} scores of the CRMCs is strongly right-skewed relative to that of the Null CRMCs (Methods and materials), indicating that the CRMCs generally score higher than Null CRMCs, and thus are unlikely produced by chance. Based

on the distribution of the S_{CRM} scores of Null CRMCs, dePCRM2 computes a p-value for each CRMC (FIGURE 2-11A). With the increase in the S_{CRM} cutoff α ($S_{CRM} \geq \alpha$), the associated p-value drops rapidly, while both the number of predicted CRMs and the proportion of the genome predicted to be CRMs decrease slowly, indicating that dePCRM2 might achieve high prediction specificity (FIGURE 2-11B). Specifically, with α increasing from 56 to 922, p-value drops precipitously from 0.05 to 1.00x10-6, while the number of predicted CRMs decreases from 1,155,151 to 327,396, and the proportion of the genome predicted to be CRMs decreases from 43.47% to 27.82% (FIGURE 2-11B). Predicted CRMs contain from 20,835,542 (p-value ≤ 1x10-6) to 31,811,310 (p-value ≤ 0.05) non-overlapping putative TFBSs that consist of from 11.47% (p-value≤ 1x10-6) to 16.54% (p-value ≤ 0.05) of the genome (FIGURE 2-12A). In other words, dependent on p-value cutoffs (1x10-6 ~0.05), 38.05~41.23% of predicted nucleotide positions in the predicted CRMs are made of putative TFBSs (FIGURE 2-12B). As expected, most of the predicted CRMs (93.99~95.46%) and constituent TFBSs (93.20~94.67%) are located in non-exonic sequences (NESs) (FIGURE 2-12A), comprising 26.66~42.47% and 10.94~16.03% of NESs, respectively (FIGURE 2-12B). Surprisingly, dependent on p-value cutoffs (1x10-6 ~0.05), the remaining 4.54~6.01% and 5.33~6.80% of the predicted CRMs and constituent TFBSs, respectively, are in an exon (FIGURE 2-12A), comprising 76.82%~85.50% and 35.42~38.17% of exonic sequences (ESs, including CDSs, 5'- and 3'-untranslated regions), respectively (FIGURE 2-12B).

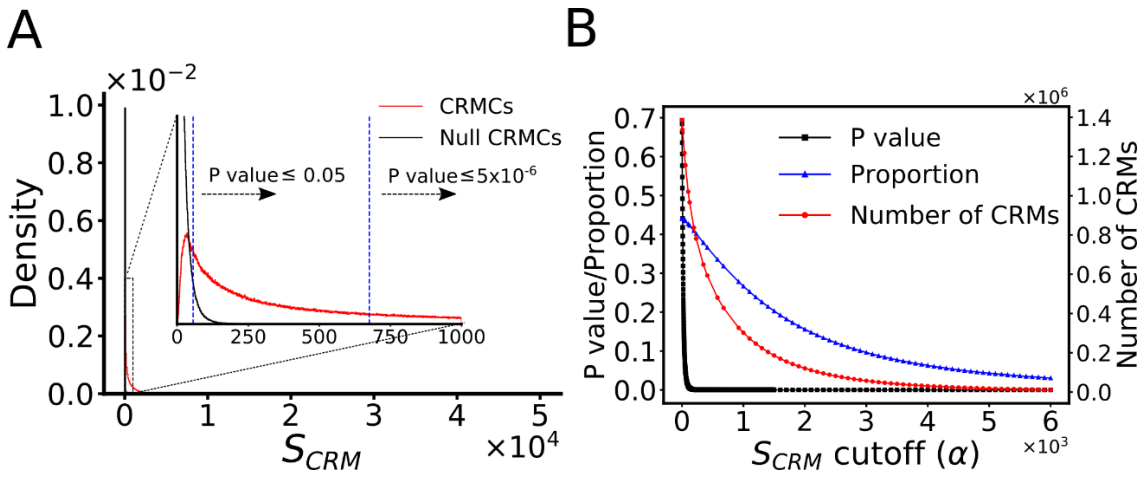

FIGURE 2-11: Prediction of CRMs using different S_{CRM} cutoffs. . A. Distribution of S_{CRM} scores of the CRMCs and Null CRMCs. B. Number of the predicted CRMs, proportion of the genome predicted to be CRMs and the associated p-value as functions of the S_{CRM} cutoff α.

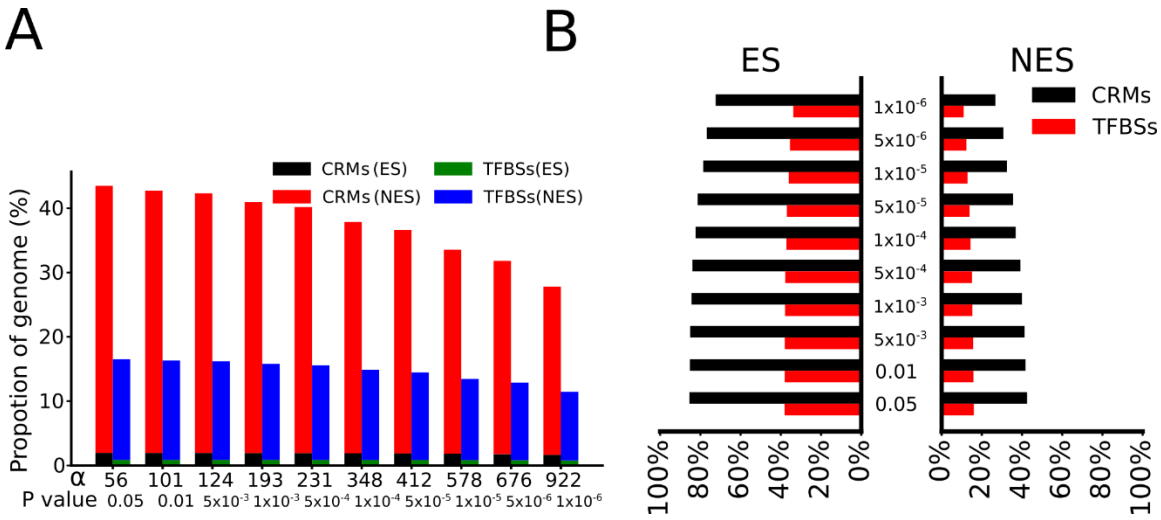

FIGURE 2-12: Coverage of the predicted CRMs at different p-value. . A.Percentage of the genome that are predicted to be CRMs and TFBSs in ESs and NESs using various S_{CRM} cutoffs and associated p-values. B. Percentage of NESs and ESs that are predicted to be CRMs and TFBSs using various S_{CRM} cutoffs and associated p-values.

2.3.6. The S_{CRM} score captures the length feature of long enhancers

We designed the S_{CRM} score to capture essential features of enhancers including their lengths. To see whether it achieves this goal, we compared the distribution of the lengths of predicted CRMCs at different S_{CRM} cutoffs α and associated p-values with those of functionally verified

VISTA enhancers. As shown in FIGURE 2-13, most of CRMCs are quite short (average length 981bp) compared to VISTA enhancers (average length 2,049bp). Specifically, almost half (621,842 or 44.26%) of the 1,404,973 CRMCs are shorter than the shortest VISTA enhancer (428bp), suggesting that most of these short CRMCs are likely short CRMs (such as promoters) or parts/components of long enhancers. However, these short CRMCs (< 428bp, 44.26%) consist of only 7.42% of the total length of the CRMCs, while the remaining 733,132 (55.74%, >428bp) CRMCs comprise 92.58% of the total length of the CRMCs. With the increase in α (decrease in p-value cutoff), the distribution of the lengths of the predicted CRMs shifts to right and gradually approaches to that of VISTA enhancers (FIGURE 2-13). Intriguingly, at α=676 (p-value ≤ $5x10^{-6}$), the distribution fits very well to that of VISTA enhancers, indicating that the corresponding 428,628 predicted CRMs have similar length distribution (average length 2,292bp) to those of VISTA enhancers (average length 2,049bp) (FIGURE 2-13), and thus they are likely full-length VISTA-like enhancers. These results demonstrate that short CRMs and CRM components tend to have smaller S_{CRM} scores than full-length enhancers and can be effectively filtered out by a higher S_{CRM} cutoff α (a smaller p-value). Therefore, S_{CRM} indeed captures the length property of full-length enhancers. Although these 428,628 putative full-length VISTA-like CRMs consist of only 30.51% of the 1,404,973 CRMCs, they comprise 72.25% (982,470,181bp) of the total length (1,359,824275bp) of the CRMCs, while the remaining 976,345 (69.49%) short CRMCs consist of only 27.75% of the total length of the CRMCs, indicating that full-length VISTA-like enhancers dominate the CRMCs in length. The failure to predict full-length CRMs of short CRM components might be due to insufficient data coverage on the relevant loci in the genome. This is reminiscent of our earlier predicted, even shorter CRMCs (average length 182bp) using a much smaller number and less diverse 670 datasets [112]. As we argued earlier [112] and confirmed here by the much longer CRMCs (average length 982bp) predicted using the much larger and more diverse datasets albeit still strongly biased to a few TFs and cell/tissue types. We anticipate that full-length CRMs

of these short CRM components can be predicted using even larger and more diverse TF ChIP-seq data when available in the future.

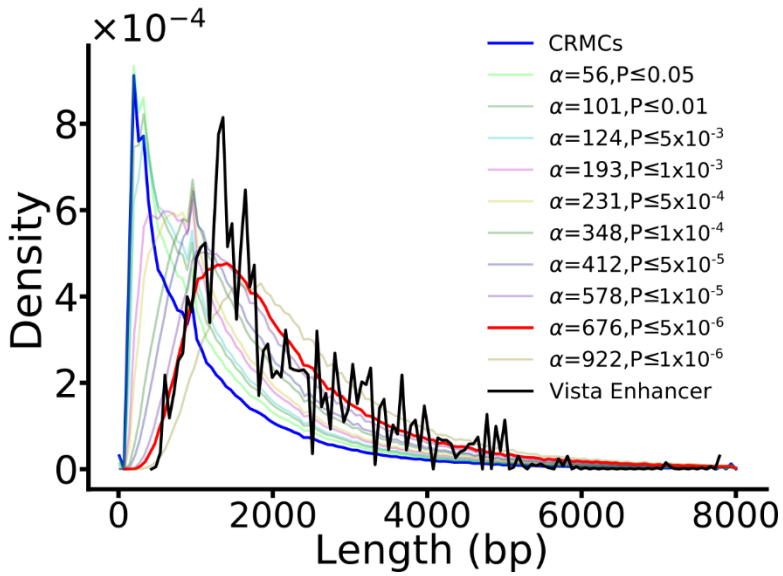

FIGURE 2-13: Distribution of the lengths of CRMs. predicted using different S_{CRM} cutoffs and associated p-values.

2.3.7. Predicted CRMs tend to be under strong evolutionary selections

To see how effective that dePCRM2 partitions the covered genome regions into the CRMCs set and the non-CRMCs set, we compared their evolutionary behaviors using the GERP [152] and phyloP [153] scores of their nucleotide positions in the human genome. Both the GERP and the phyloP scores quantify conservation levels of nucleotide positions in the genome based on nucleotide substitutions in alignments of multiple vertebrate genomes. The larger a positive GERP or phyloP score of a position, the more likely it is under negative selection; and a GERP or phyloP score around zero means that the position is selectively neutral or nearly so. For convenience of discussion, we consider a position with a GERP or phyloP score within an interval centering on 0 [$-\delta$,$+\delta$] (δ >0) to be selectively neutral or nearly so, and a position with a score greater than δ to be under negative selection. We define proportion of neutrality of a set of positions to be the size

of the area under the density curve of the distribution of the scores of the positions within the window [$-\delta$,+ δ]. Because ESs evolve quite differently from NESs, we focused on the CRMCs and constituent TFBSs in NESs and left those that at least partially overlap ESs in another analysis. The choice of δ = 0.5, 1, 2, and 3 gave similar results (data not shown), so we choose a=1 in subsequent analysis. Intriguingly, the distribution of the GERP scores of the non-CRMCs (1,034,985,426 bp) in NESs displays a sharp peak around score 0, with low right and left shoulders and a high proportion of neutrality 0.71 (FIGURE 2-14A), suggesting that the most of the non-CRMCs are selectively neutral or nearly so, and thus at least most of them are unlikely to be functional. In sharp contrast, the distribution of the GERP scores of the 1,292,356 CRMCs (1,298,719,954bp) in NESs has a blunt peak around score 0, with high right and left shoulders and a small proportion of neutrality 0.31 (FIGURE 2-14A). These results strongly suggest that CRMCs are subject to much stronger evolutionary selection than are the non-CRMCs. To see how known CRMs evolve, we plotted the distribution of conservation scores of all the 976 VISTA enhancers. Clearly, the distribution of GERP scores of the VISTA enhancers are similar to that of CRMCs, also with a blunt peak around score 0, high right and left shoulders, and a mall proportion of neutrality 0.23 (0.31 for the CRMCs) (FIGURE 2-14A). Thus, like the CRMCs, the VISTA enhancers are also under much stronger evolutionary selections that are the non-CRMCs as expected. Notably, however, the distribution of VISTA enhancers has peal around score 3, indicating that VISTA enhancers tend to be more conserved than the CRMCs (FIGURE 2-14A). This is not surprising as the VISTAT enhancers are biased [154] to ultra-conserved, development related enhancers[155, 156]. The similar evolutionary behavior between the CRMCs and VISTA enhancers strongly suggest that at least most of the CRMCs might be functional. Moreover, dramatic differences between the evolutionary behaviors of the non-CRMCs and those of the CRMCs as well as the VISTA enhancers strongly suggests that dePCRM2 largely partitions the covered genome regions into a functional CRMC set and a non-functional non-CRMC set. Similar results were obtained using the phyloP scores (FIGURE 2-15A).

To see why dePCRM2 abandoned 38.60% nucleotide positions covered by the original binding peaks in predicting the CRMCs, we plotted the distribution of conservation scores of the abandoned positions. These abandoned positions have a GERP score distribution almost identical to those in the non-CRMCs (FIGURE 2-14A), thus they are unlikely to be functional, strengthening our earlier argument that this portion (38.60%) of the original binding peaks might not contain TFBSs. Therefore, dePCRM2 is able to accurately distinguish functional and non-functional parts in both the original binding peaks and their extended parts. Interestingly, the uncovered 22.53% genome regions have a GERP score distribution and a proportion of neutrality (0.59) in between those of the covered regions (0.49) and those of the non-CRMCs (0.71) (FIGURE 2-14A). These results indicate that the uncovered regions are more evolutionarily selected than the non-CRMCs, but less evolutionary selected than the covered regions. This implies that the uncovered regions contain functional elements such as CRMs, but their density could be lower than that of the covered regions. Assuming that the density of CRMs is proportional to the size of evolutionarily selected regions, the density of CRMs in the uncovered regions could be estimated to be (1-0.59)/(1-0.49)=79.40% of the covered regions. Similar results could be obtained using the phyloP scores (FIGURE 2-15A).

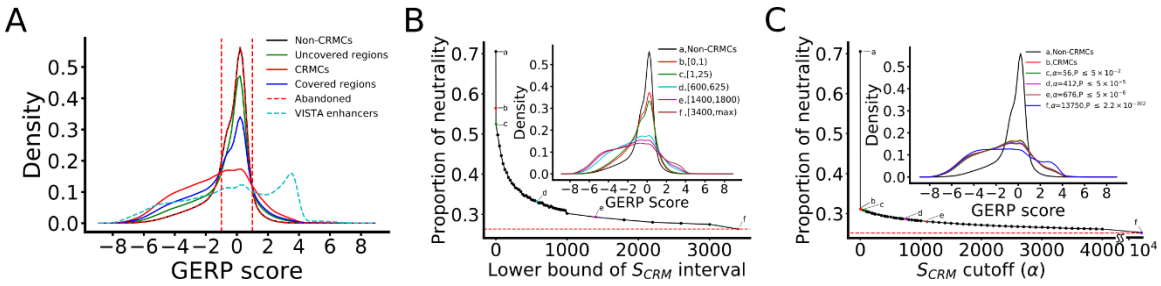

FIGURE 2-14: Distribution of GERP score on the CRMCs and non-CRMCs in NESs. A. Distributions of the GERP scores of nucleotides of the predicted CRMCs, non-CRMCs, abandoned genome regions covered by the original binding peaks, genome regions covered by the extended binding peaks and genome regions uncovered by the extended binding peaks. The area under the density curves in the score interval [-1, 1] is defined as the proportion of neutrality of the sequences. B. Proportion of neutrality of CRMCs with a S_{CRM} score in different intervals in comparison with that of the non-CRMCs (a). The inset shows the distributions of the GERP scores of the non-CRMCs and CRMCs with S_{CRM} scores in the intervals indicted by color and letters. C. Proportion of neutrality of CRMs predicted using different S_{CRM} score cutoffs and associated p-values in comparison with those of the non-CRMCs (a) and CRMCs (b). The inset shows the distributions of the GERP scores of the non-CRMCs, CRMCs and the predicted CRMs using the S_{CRM} score cutoffs and p-values indicated by color and letters.

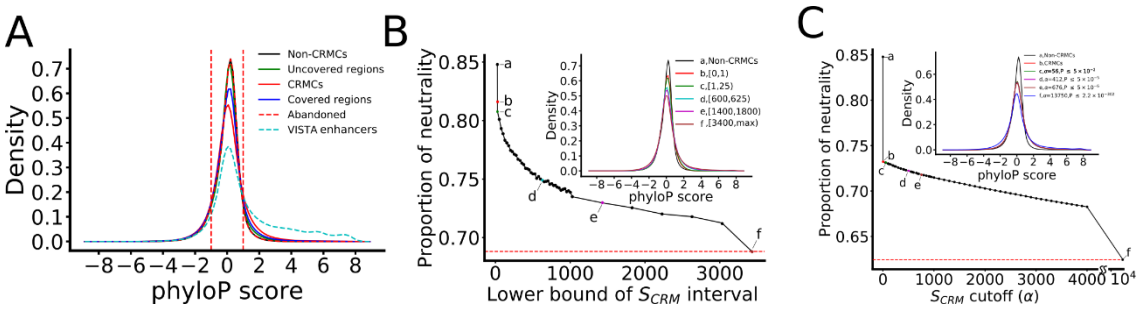

FIGURE 2-15: Distribution of phyloP score on the CRMCs and non-CRMCs in NESs. . A. Distributions of the phyloP scores of nucleotides of the predicted CRMCs, non-CRMCs, abandoned genome regions covered by the original binding peaks, genome regions covered the extended binding peaks, and genome regions uncovered by the extended binding peaks. The area under the density curves in the score interval [-1, 1] is defined as the proportion of neutrality of the sequences. B. Proportion of neutrality of CRMCs with a S_{CRM} score in different intervals in comparison with that of the non-CRMCs (a). The inset shows the distributions of the phyloP scores of the non-CRMCs and CRMCs with S_{CRM} scores in the intervals indicated by colors and letters. C. Proportion of neutrality of CRMs predicted using different S_{CRM} score cutoffs and associated p-values in comparison with that of the non-CRMCs (a) and CRMCs (b). The inset shows the distributions of the phyloP scores of the non-CRMCs, CRMCs and predicted CRMs using the S_{CRM} scores and p-values indicated by colors and letters.

2.3.8. The S_{CRM} score captures the evolutionary feature of CRMs

We then investigated the relationship between the conservation scores of the CRMCs and their S_{CRM} scores. To this end, we plotted the distributions of the conservation scores of the CRMCs with a S_{CRM} score in different non-overlapping intervals. Remarkably, the CRMCs with S_{CRM} scores in the lowest interval [0, 1) have a smaller proportion of neutrality (0.56) than the non-CRMCs (0.71) (FIGURE 2-14B), indicating that even these low-scoring CRMCs with short lengths (FIGURE 2-13) are more likely to be under strong evolutionary constraints than the non-CRMCs, and thus might be functional. With the increase in the lower bound of S_{CRM} intervals, the proportion of neutrality of the CRMCs in the intervals drops rapidly, followed by a slow linear decrease around the interval [1000,1400) (FIGURE 2-14B). Therefore, the higher the S_{CRM} score of a CRMC, the more likely it is under strong evolutionary constraint, suggesting that the S_{CRM} score captures the evolutionary behavior of a CRM as a functional element, in addition to its length feature (FIGURE 2-13). The same conclusion can be drawn from the phyloP scores (FIGURE 2-15B).

We next examined the relationship between the conservation scores of the predicted CRMs and S_{CRM} score cutoffs α (or p-value cutoffs) used for their predictions. As shown in FIGURE 2-14C, even the CRMs predicted at the low S_{CRM} cutoffs have a much smaller proportion of neutrality (e.g., 0.31 for α=0, i.e., the CRMC set) than the non-CRMCs (0.71), suggesting that most of the predicted CRMs might be authentic, while the non-CRMCs might contain few false negative CRMCs. With the increase in the S_{CRM} cutoff α (decrease in p-value), the proportion of neutrality of the predicted CRMs decreases further but linearly and slowly, entering a saturation phase (FIGURE 2-14C). Interestingly, at a high S_{CRM} cutoff such as α= 13,750, the resulting predicted CRMs evolve more similar to the VISTA enhancers with a peak in the GERP score distribution around score 3 (FIGURE 2-14A and C). Thus, the higher the S_{CRM} cutoff α (i.e., the smaller the p-value cutoff), the more likely the predicted CRMs are under strong evolutionary constraints, and thus functional, suggesting again that the S_{CRM} scores capture the essence of CRMs as functional elements. Moreover, the infinitesimal decrease in the proportion of neutrality of the CRMs predicted with the increase in S_{CRM} cutoffs (decrease in p-value cutoff) (FIGURE 2-14C) strongly suggests that the resulting predicted CRMs are under similarly strong evolutionary constraints, and thus, the specificity of predicted CRMs might approach a saturated high level that the VISTA enhancers achieve. However, without the availability of a no gold standard negative CRM set in the genome[157], we could not explicitly calculate the specificity of the predicted CRMs at different p-value cutoffs. Similar results are observed using the phyloP scores (FIGURE 2-15C).

2.3.9. Performance of dePCRM2 for recovering functional elements

To further evaluate the accuracy of the predicted CRMs, we calculated the sensitivity (recall rate or true positive rate (TPR)) of CRMs predicted at different S_{CRM} cutoffs α and associated p-values for recalling a variety of CRM function-related elements located in the covered genome regions in 10 experimentally determined datasets in various cell/tissue types (Methods and

materials), including 785 VISTA enhancers [118], 402,730 of ClinVar SNPs[121, 158], 181,436 FANTOM promoters (FPs) [159], 32,029 FANTOM enhancers (FEs) [160], 82,378 of GWAS SNPs [123], 121,075,184 DHSs [122, 123], 29,195,778 transposase-accessible sites (TASs)[139], 98,297,240 H3K27ac peaks [140], 75,467,050 H3K4me1 peaks[141], and 69,282,044 H3K4me3 peaks[141]. Here, if a predicted CRM and an element overlap each other by at least 50% of the length of the shorter one, we say that the CRM recalls the element. As shown in FIGURE 2-16A, with the increase in the p-value cutoff, the sensitivity for recalling the elements in all the 10 datasets increases rapidly and becomes saturated well before p-value increases to 0.05 ($\alpha \geq 56$). FIGURE 2-17 A~J show examples of the predicted CRMs recalling the elements in the 10 datasets by overlapping them. Particularly, at p-value cutoff $5x10^{-5}$ (α=412), the predicted 593,731 CRMs recall 100% of the VISTA enhancers[118] and 97.43% of ClinVar SNPs[121] (FIGURE 2-16A). The very rapid saturation of the sensitivity for recalling these two types of validated functional elements at very low p-value once again strongly suggest that the dePCRM2 also achieves very high specificity,although we could not explicitly compute it for the aforementioned reason. On the other hand, even at a relatively higher p-value cutoff 0.05 (α=56), the predicted 1,155,151 CRMs only achieve varying intermediate levels of sensitivity for recalling FPs (88.77%), FEs)[160] (81.90%), DHSs[161](74.68%), TASs[139](84.32%), H3K27ac[140](82.96%), H3K4me1[141] (76.77%), H3K4me3[141](86.96%) and GWAS SNPs[122](64.50%), although all are significantly higher than that of randomly selected sequences (15%) with matched lengths from genome regions covered by the data (FIGURE 2-16A).

To find out the reasons for such different performance of dePCRM2 on different datasets, we plotted the distribution of GERP scores of the recalled and uncalled elements in the 10 datasets. As there is no uncalled VISTA enhancer and we have already plotted the distribution of the entire set of the 976 VISTA enhancers (FIGURE 2-14A), we instead plotted the distribution of CRMs that overlap and recall the 785 VISTA enhancers in the covered genome regions (VISTA-CRMs).

As expected, recalled elements in all the datasets evolve similarly to the predicted 1,155,151 CRMs at p-value <0.05 (FIGURE 2-16B), recalled 785 VISTA enhancers have almost identical distribution (data not shown) to the entire set of 976 VISTA enhancers (FIGURE 2-14A). Like the VISTA enhancers (FIGURE 2-14A), VISTA-CRMs as well as the recalled ClinVar SNPs and FPs, are all under stronger evolution constraints than the other recalled element types (FIGURE 2-16B). These results are not surprising, as we indicated earlier that VISTA enhancers are biased[154] to ultra-conserved, development related enhancers[155, 156], while ClinVar SNPs were identified for their large effect sizes[121, 158], and promoters are well-known be to more conserved than are enhancers[162]. In stark contrast, all unrecalled elements in the 10 datasets evolve similarly to the non-CRMCs, with the exception that the 2.57% (10,350) of unrecalled ClinVar SNPs display a bimodal distribution and there are no unrecalled VISTA enhancers (FIGURE 2-16B). However, it is notable that the proportions of neutrality of unrecalled PEs (0.59) and PFs (0.63) are smaller than that of the non-CRMCs (0.71) (FIGURE 2-16B), suggesting we might miss a small portion of authentic PEs and PFs (see below for false negative rate (FNR) estimations of our CRMs). Nevertheless, assuming that at least most of unrecalled elements in the datasets except the VISTA and ClinVar datasets are non-functional, we estimate the false discovery rate (FDR) of remaining eight datasets might be up to from 11.23% (1-0.8877) for FPs to 35.32% (1-0.6450) for GWAS SNPs, which are consistent with an earlier study showing that histone marks and CA data resulted in high false positives for predicting enhancers[107]. Interestingly, the bimodal distribution of GERP scores of the 2.57% of unrecalled ClinVar SNPs displays a peak around score 0 with a proportion of neutrality 0.40 (FIGURE 2-16B), indicating that at most 40% of the relevant SNPs are selective neutral, and thus might be non-functional. We therefore estimate the FDR of the ClinVar SNP dataset to be < 0.40*2.57%=1.03%. Hence, like the VISTA enhancers, ClinVar SNPs are a reliable set with only a small portion of false positives (<1.03%) for evaluating CRM predictions. The other peak of the unrecalled ClinVar SNPs are located around score 3 (FIGURE 2-16B), indicating that the relevant SNPs are under strong purifying selection, and thus might be

functional that were missed by our algorithms. We therefore estimate our predictions (at p-value <0.05) might have a FNR < 2.57%-1.03%=1.54%. In other words, the real sensitivity (=1-FNR) for dePCRM2 to recall authentic ClinVar SNPs might be higher than the calculated 97.54% (FIGURE 2-16A). These estimates are consistent with the zero FNR and 100% sensitivity for our predicted CRMs to recall VISTA enhancers (FIGURE 2-16A) and a simulation to be described later.

The zero, very low (<1.26%) and low (11.23%) FDRs of the VISTA enhancers, ClinVar SNPs and FPs, respectively, indicate that the experimental methods used to characterize them might be more reliable. However, their high accuracy might also be related to the their larger effect sizes and more conserved functions, which may facilitate their correct determinations as indicated by the facts that they are under stronger evolutionary selections than the elements characterized by other experimental methods (FIGURE 2-16B). In this regarding, we note that the intermediately high FDRs of FE(18.10%), DHS(25.32), TAS (15.68%), H3K4m3 (13.04%), H3K4m1 (23.23%) and H3K27ac (17.04%) datasets might be due to the facts that bidirectional transcription[163], CA[107, 109, 164] and histone marks[107, 109] are not unique to active enhancers as pointed out in an earlier study[107]. The high FDR of GWAS SNPs (35.5%) might be due to the fact that a lead SNP associated with a trait may not necessarily be located in a CRM and causal; rather, some variants in a CRM, which are in LD with the lead SNP, are the culprits[122, 123]. Example of GWAS SNPs in LD with positions in a CRM are shown in FIGURE 2-17 K and L. Interestingly, many recalled ClinVar SNPs (42.59%) and GWAS SNPs (38.18%) are located in critical positions in predicted binding sites of the UMs (e.g., FIGURE 2-17D and F).

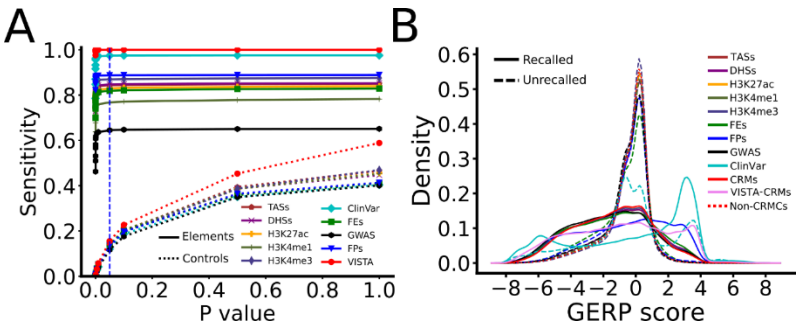

FIGURE 2-16: Validation of the predicted CRMs.. A. Sensitivity (TPR) of the predicted CRMs and control sequences as a function of p-value cutoff (FPR) for recovering the elements. The dotted vertical lines indicate the p-value $\leqslant 0.05$ cutoff. B. Distributions of the GERP scores of the recalled and unrecalled elements in comparison with those of the predicted CRMs at $p \leqslant 0.05$ and non-CRMCs.

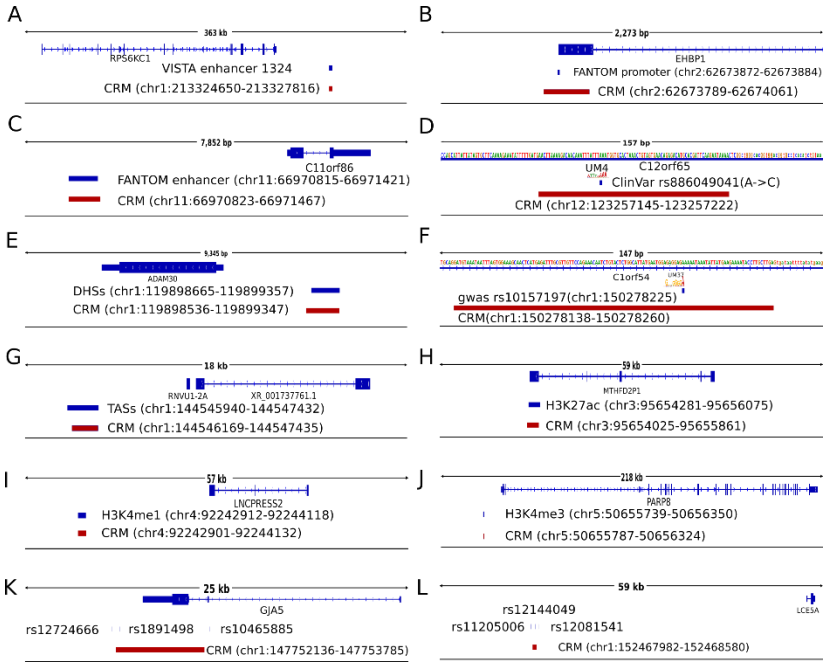

FIGURE 2-17: Examples of CRMs that recover validated functional elements. . A. A CRM (chr1:213324650-2733277816) recovers VISTA enhancer 1324 downstream of gene RPS6K1. B. A CRM (chr2:62673789-62674061) recovers a FANTOM promoter (chr2: 62673872-62673884) for gene EHBP1. C. A CRM (chr11:66970823-66971467) recovers a FANTOM enhancer (chr11: 66970815-66971421) upstream of gene C11orf86. D. A CRM (chr3:10145256-10145421) recovers three ClinVar point mutants (chr3:10145341C>T, chr3:10145379A>G, 10145381G>A) in an intron of gene VHL, related to hereditary cancer-predisposing syndrome. Each mutation disrupts a critical position of the binding- site of the indicated UMs. E. A CRM (chr1:119898536-119899347) recovers a DHS (chr1: 119898665-119899357) upstream of gene ADAM30. F. A CRM (chr1:150278138-150278260) recovers a GWAS SNP rs10157197 (chr1:150278225) in an intron of gene C1orf54. The SNP is located in a critical position of UM37. G. A CRM (chr1:114546169-144547435) recovers a TAS (chr1:114545940-144547432) is located upstream of gene RNVU1-2A. H. A CRM (chr3:95654025-95655861) recovers a H3K27ac peal located downstream of the gene MTHFD2P1. I. A CRM (chr4:92242901-92244132) recovers a H3K4me1 peak (chr4:92242912-92244118) downstream of the gene LNCPRESS2. J. A CRM (chr5:50655787-50656324) recovers a H3K4me3 peak (chr5:50655739-50656350) upstream of the gene PARP8. K. A CRM (chr1:147752136-147753785) recovers GWAS SNP rs1891498 upstream of gene GJA5, while two unrecovered GWAS SNPS rs12724666 and rs10465885 located upstream of and in an intron of the gene, respectively, are in LD with rs1891498. L. A CRM (chr1:152467982-152468580) recovers a GWAS SNP rs12144049 upstream of gene LCE5A, while two unrecovered GWAS SNPS rs11205006 and rs12081541 located upstream and downstream of rs12144049, respectively, are in LD with it.

2.3.10. dePCRM2 largely correctly predicts the lengths of CRMs

After showing that dePCRM2 is able to capture the length feature of CRMs (FIGURE 2-13), we now evaluated the accuracy of dePCRM2 for predicting the lengths of CRMs. To this end, we compared the distributions of the lengths of the 785 recalled VISTA enhancers and the 26,233 recalled FEs with those of the 836 and 22,235 recalling CRMs (p-value < 0.05), respectively. As shown in FIGURE 2-18A, the recalling CRMs have a largely similar length distribution to the recalled VISTA enhancers, although the former have longer median length (3,799bp) than the latter (1,613bp). However, it is unclear whether or not the VISTA enhancers are in full-length, as a portion of an enhancer could be still partially functional[131]. On the other hand, a few (7.92%) VISTA enhancers are recalled by multiple short CRMCs (FIGURE 2-18A), suggesting that some of our predicted CRMCs might not in full-length, but only components of long CRMCs. Nonetheless, these results suggest that dePCRM2 is able to largely correctly predict the length of VISTA enhancers, albeit not perfect. Moreover, it is well-known that development-related enhancers that most VISTA enhancers belong to tend to be longer than other types of enhancers. In agreement with this, all the VISTA enhancers are recalled at a rather low p-value cutoff $5x10^{-5}$ ($\alpha = 412$) (FIGURE 2-16A) that filters out most short CRMCs (FIGURE 2-13). At even a lower p-value cutoff $5x10^{-6}$ ($\alpha = 676$) the resulting 428,628 predicted CRMs with almost an identical length distribution as the entire sets of VISTA enhancers (FIGURE 2-13) still recall 99.10% of the VISTA enhancers. Thus, it appears that dePCRM2 is able to predict full-length long CRMs at low p-value cutoffs.

By contrast, the recalled FEs at p-value cutoff 0.05 are generally shorter (mean length 293bp) than the recalling CRMs (median length 2,371bp) (FIGURE 2-18B), and a considerable proportion (26.69%) of the recalled FEs overlap the same CRMs at different parts (FIGURE 2-19A and B). To see whether the FEs also are generally shorter than their overlapping VISTA enhancers, we compared the distribution of the lengths of 113 FEs with that of their 91 overlapping VISTA

enhancers (FIGURE 2-18C). The low overlapping rates between VISTA enhancers and FEs (11.59% for VISTA enhancers and 0.35% for FEs) indicate that they might belong to quite different types of enhancers. Nonetheless, the 113 FEs (median length 319bp) are much shorter than the 91 VISTA enhancers (median length 2,686bp), and in a few cases, multiple FEs overlap different parts of the same VISTA enhancers (FIGURE 2-19A and B). Thus, it is likely that the CRMs that match multiple FEs might be in full-length, and that the eRNA-seq based method tends to identify short CRMs or CRM components of otherwise long CRMs, instead of full-length enhancers. Interestingly, the CRMs that recall EHs tend to be shorter than those that recall VISTA enhancers (median length 2,371bp vs 3,799pb) (FIGURE 2-18A and B), strengthening our earlier argument that they might be different types of enhancers. This conclusion is consistent the fact that most VISTA enhancers are development-related while most FEs are not as they were mainly determined in adult primary tissues and cell line[160]. Furthermore, decrease in p-value cutoff from 0.05 to $5x10^{-6}$ largely decrease the sensitivity for recalling FEs from 81.90% to 74.12% (FIGURE 2-16B), indicating that short recalling CRMs are filtered out at low p-value cutoffs. Taken together, these results suggest that dePCRM2 is able to accurately predict the length of either short or long CRMs, although a small portion of the predicted CRMCs might be components of longer CRMs.

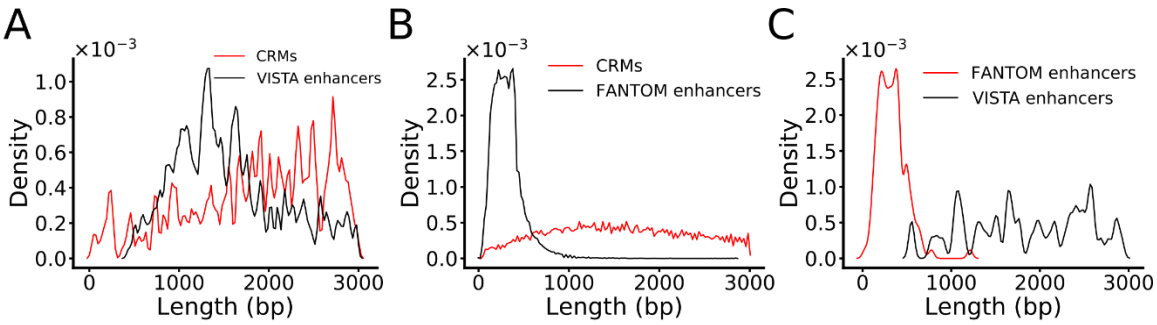

FIGURE 2-18: dePCRM2 can predict the lengths of VISTA enhancers. A. Distributions of the lengths of the recovered VISTA enhancers and the recovering CRMs. B. Distributions of the lengths of the recovered FANTOM enhancers and the recovering CRMs. C. Distributions of the lengths of FANTOM enhancers and the overlapped VISTA enhancers.

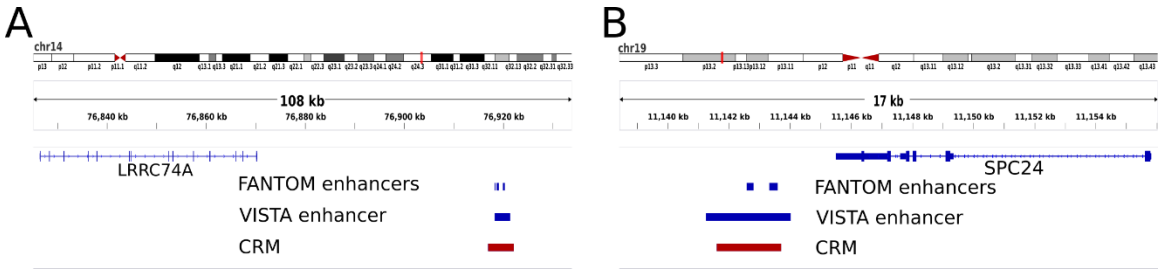

FIGURE 2-19: Examples of overlap between CRM, FANTOM and VISTA enhancers. A.Three different FANTOM enhancers (chr14:76918231-76918434, chr14:76918637-76919122, chr14:76919959-76920278) overlap VISTA enhancer 1466 (chr14:76918298-76921354) and a predicted CRM (chr14:76916942-76922129) downstream of gene LRRC74A. B. Two FANTOM enhancers (chr19:11142586-11142806, chr19:11143315-11143595) overlap VISTA enhancer 1754 (chr19:11141255-11144027) and a predicted CRM (chr19:11141600-11143708) downstream of gene SPC24.

2.3.11. dePCRM2 outperforms state-of-the-art algorithms

We compared our predicted CRMs at p-value ≤ 0.05 ($S_{CRM} < 56$) with three most comprehensive sets of predicted enhancers/promoters, i.e., GeneHancer 4.14 [98], EnhancerAtals2.0 [103] and cCREs[105]. For convivence of discussion, we call these three sets enhancers or cCREs. GeneHancer 4.14 is the most updated version containing 394,086 non-overlapping enhancers covering 18.99% (586,582,674bp) of the genome (FIGURE 2-20A). These enhancers were predicted by integrating multiple sources of both predicted and experimentally determined CRMs, including ENCODE 2 enhancer-like regions [165], ENSEMBL regulatory build [96], dbSUPER [166], EPDnew promoters [167], UCNEbase [168], CraniofacialAtlas [169], VISTA enhancers [118], FPs[159] and FEs [160]. Enhancers from ENCODE 2 and ESEMBL were predicted based on multiple tracks of epigenetic marks using the well-regarded tools ChromHMM [87] and Segway [90]. Of the GeneHancer enhancers, 388,407 (98.56%) have a at least one nucleotide located in the covered genome regions, covering 18.89% of the genome (FIGURE 2-20A). EnhancerAtlas 2.0 contains 7,433,367 overlapping cell/tissue-specific enhancers in 277 cell/tissue types, which were predicted by integrating 4,159 TF ChIP-seq, 1,580 epigenetic, 1,113 DHS-seq, and 1,153 other enhancer function related datasets, such as FEs [102]. After removing redundancy (identical enhancers in difference cell/tissues), we ended up with 3,452,739 EnhancerAtlas enhancers that

may still have overlaps, covering 58.99% (1,821,795,020bp) of the genome (FIGURE 2-20A), 3,417,629 (98.98%) of which have at least one nucleotide located in the covered genome regions, covering 58.78% (1,815,133,195bp) of the genome (FIGURE 2-20A). cCREs represents the most recent CRM prediction by the ENCODE consortium[105], containing 926,535 non-overlapping enhancers covering 8.20% (253,321,371bp) of the genome. The cCREs were predicted based on overlaps among 703 DHS, 46 TAS and 2,091 histone mark datasets in various cell/tissue types produced by ENCODE phases 2 and 3 as well as the Roadmap Epigenomics project[105]. Of these cCREs, 917,618 (99.04%) have at least one nucleotide located in the covered genome regions, covering 8.13% (251,078,466bp) of the genome (FIGURE 2-20A). Thus, both the number (1,155,151) and genome coverage (43.47%) of our predicted CRMs (at. P-vaule,0.05) are larger than those of GeneHancer enhancers (388,407 and 18.89%) and of cCREs (917,618 and 8.12%) regions, but smaller than those of EnhancerAtlas enhancers (3,417,629 and 58.78%), which at least partially overlap the covered regions.

To make the comparison relatively fair, we first computed recall rates of these enhancers that at least partially overlap the covered genome regions, for recalling VISTA enhancers, ClinVar SNPs and GWAS SNPs. We omitted FPs, FEs, DHSs, TASs and the three histone marks for the valuation as they were used in predicting CRMs by GeneHancer 4.14, EnhancerAtlas 2.0 or ENCODE phase 3 consortium. We included VISTA enhancers for the evaluation as they were not included in EnhancerAtlas enhancers and cCREs, although they were parts of GeneHancer 4.14. Remarkably, our predicted CRMs outperform EnhancerAtlas enhancers for recalling VISTA enhancers (100.00% vs 94.01%) and ClinVar SNPs (97.43% vs 7.03%) (FIGURE 2-20B), even though our CRMs cover a smaller proportion of the genome (43.47% vs 58.78%, or 35.22% more) (FIGURE 2-20A), indicating that dePCRM2 has both higher sensitivity and specificity than the method behind EnhancerAtlas 2.0 [103]. However, our predicted CRMs underperform EnhancerAtlas enhancers for recalling GWAS SNPS (64.50% vs 69.36%, or 7.54% more) (FIGURE 2-20B). As we indicated

earlier, the lower sensitivity of dePCRM2 for recalling GWAS SNPs might be due to the fact that an associated SNP may not necessarily be causal (FIGURE 2-17D). The higher recall rate of EnhancerAtlas enhancers for GWAS SNPs might be simply thanks to their 35.22% more coverage of the genome (58.78%) than that of our predicted CRMs (43.47%) (FIGURE 2-20A). Our predicted CRMs outperform cCREs for recalling VISTA enhancers (100% vs 85.99%), ClinVar SNPs (97.43% vs 18.28%) and GWAS SNPs (64.50% vs 15.74%) (FIGURE 2-20B). Our predicted CRMs also outperform GeneHancer enhancers for recalling ClinVar SNPs (97.43% vs 33.16%) and GWAS SNPs (64.50% vs 34.11%) (FIGURE 2-20B). However, no conclusion can be drawn from these results about the specificity of our predicted CRMS compared with the other three predicted enhancer sets, because our predicted CRMs cover a higher proportion of the genome (43.47%) than GeneHancer enhancers (18.89%) and cCREs (8.20%) (FIGURE 2-20A). Nevertheless, both GeneHancer enhancers (33.16%) and cCREs (18.28%) outperform EnhancerAtlas enhancers (7.03%) for recalling ClinVar SNPs (FIGURE 2-20B), even though the former two (18.89% and 8.20%, respectively) have a much smaller genome coverage than the latter one (58.78%) (FIGURE 2-20A), indicating the former two have higher specificity than the latter.

As shown in FIGURE 2-21A, the intersections/overlaps between the four predictions are quite low. For instance, EnhancerAtlas enhancers, GeneHancer enhancers and cCREs share 926,396,395bp (50.85%), 414,806,711bp (70.72%), and 194,709,825bp (76.86%) of their nucleotide positions with our predicted CRMs, corresponding to 69.01%, 30.90% and 14.51% of positions of our CRMs (FIGURE 2-21A), respectively; and there are only 105,606,214pb shared by all the four predictions, corresponding to 5.80%, 18.00%, 41.69% and 7.87% of the number of nucleotide positions covered by EnhancerAtlas enhancers, GeneHancer enhancers, cCREs and our CRMs, respectively. To estimate the FPRs of EnhancerAtlas enhancers, GeneHancer enhancers and cCREs, we plotted the distributions of GERP scores of the positions that each of them share and do not share with our CRMs. As expected, the shared positions of the three predicted enhancer

sets all evolve similarly to our predicted CRMs, although those of GeneHancer enhancers and cCREs are under slightly higher evolutionary constraints than the our entire CRM set (FIGURE 2-21B). However, if we use a more stringent SCRM cutoff, e.g. α=3,000 ($p<2.2x10-302$, the resulting predicted CRMs are even under stronger evolutionary constraints than the shared GeneHancer enhancers and cCREs positions (FIGURE 2-16C and FIGURE 2-21B). Therefore, these shared GeneHancer enhancers and cCREs positions just evolve like subsets of our predicted CRMs. By stark contrast, the remaining 49.14%, 29.28% and 23.13% unshared positions of EnhancerAtlas enhancers, GeneHancer enhancers and cCRES, respectively, evolve similarly to the non-CRMCs, although they all have slightly smaller proportion of neutrality than that of the non-CRMCs (0.66, 0.63 and 0.61 vs. 0.71, respectively) (FIGURE 2-21B), due probablely to the small FNR (<1.54%) of our predicted CRMs. Nonetheless, these results suggest that the vast majority of the unshared positions of the three sets of predicted enhancers are selectively neutral, and thus might be nonfunctional. Therefore, it appears that predicted enhancers in the three sets that overlap our CRMs are likely to be authentic, while most of those that do not overlap with our CRMs might be false positives. Thus, we estimate the FDR of EnhancerAtlas enhancers, GeneHancer enhancers and cCRES might be slightly smaller than 49.14%, 29.28% and 23.13%, respectively. These results also strongly suggest that GeneHancer 4.14 and cCREs might largely under-predicted enhancers even with a rather high FDR up to 29.28% and 23.12%, respectively (FIGURE 2-21B), while EnhancerAtlas 2.0 might largely over-predicted enhancers with a very high FPR up to 49.14% (FIGURE 2-21B).

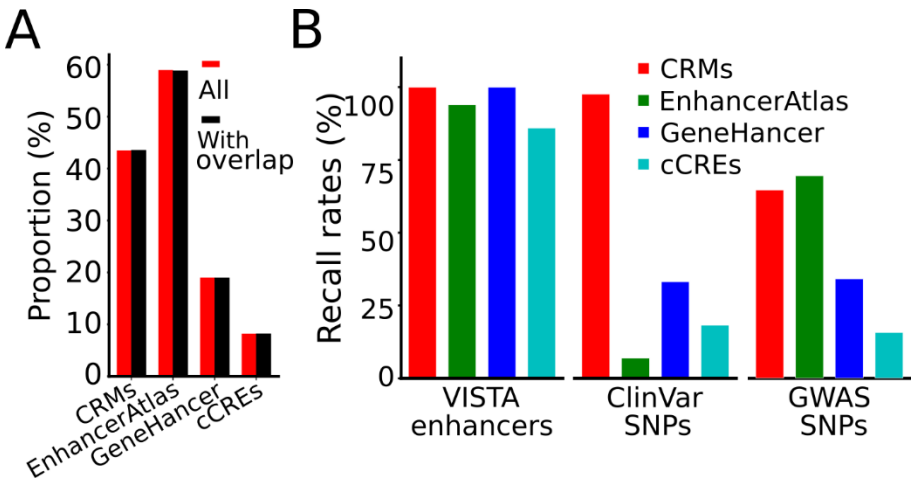

FIGURE 2-20: Comparison of dePCRM2 and three state-of-the-art methods. A. Percentage of genome regions covered by all CRMs/enhancers predicted by different methods, and percentage of genome regions covered by predicted CRMs/enhancers that at least partially overlap the covered genome regions. B, Recall rates for recovering VISTA enhancers, ClinVar SNPs and GWAS SNPs, by the predicted CRMs/enhancers that at least partially overlap the covered genome regions.

Finally, we also compared the lengths of the four sets of predicted enhancers/CRMs with those of VISTA enhancers. As shown in FIGURE 2-22, the distribution of the lengths of cCREs has a narrow high peak at 345bp with a mean length of 273bp and a maximal length of 350bp. It is highly likely that the substantial amountphy of authentic cCREs are just components of longer CRMs as even the shortest known enhancer in VISTA is 428bp long. The distribution of GeneHancer enhancers oscillates with a period of 166bp (FIGURE 2-22)., which might be an artifact of underlying algorithm. With a mean length of 1,488bp, GeneHancer enhancers are shorter than the VISTA enhancers (with mean length 2,049pb) (FIGURE 2-22). EnhancerAtlas enhancers have a similar length distribution to the VISTA enhancers (FIGURE 2-22). However, with a mean length of 680bp, they are shorter than the VISTA enhancers. Our predicted CRMs at p-value <0.05 have a mean length of 1,162pb, thus also are shorter than that of the VISTA enhancers (2,049bp) (FIGURE 2-22), suggesting that our CRMs might contain short component of longer CRMs. However, as we indicated earlier, with a higher p-value cutoff $5x10^{-6}$, the resulting 428,628 predicted CRMs have almost an identical length distribution as the VISTA enhancers (FIGURE

2-13). Taken together, these results unequivocally indicate that our predicted CRMs are much more accurate than the three state-of-the-art predicted enhancer/cCRE sets in both the positions and lengths.

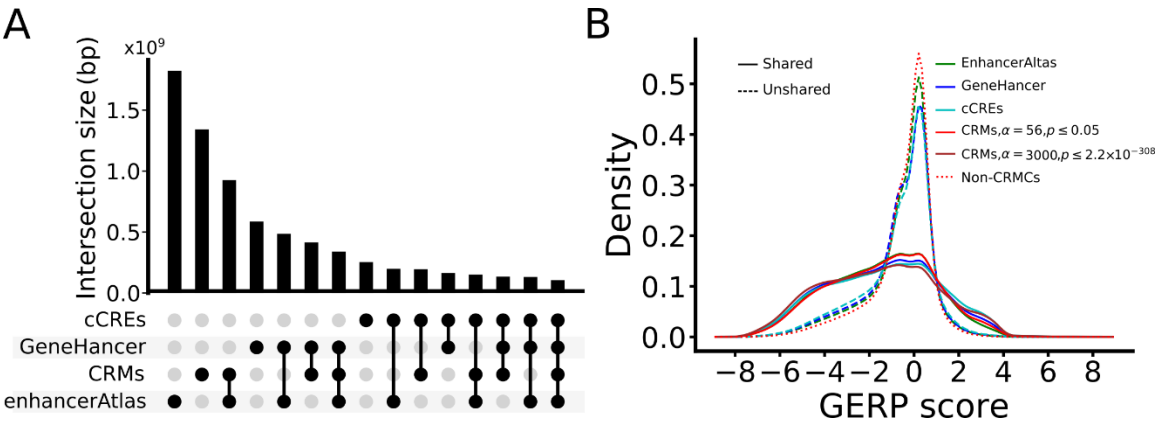

FIGURE 2-21: Overlap between each pair of CRMs, EnhancerAtlas, and GeneHancer. A.Ven diagram showing numbers of nucleotide positions shared among the predicted CRMs, GeneHancer enhancers and EnhancerAtlas enhancers. B. Distributions of GERP scores of nucleotide positions of CRMs predicted at p-value $\leqslant$ 0.05 and p-value $\leqslant$ 5X10-6, and the non-CRMCs, as well as of nucleotide positions that GeneHancer enhancers and EnhancerAtlas enhancers share or do not share with the predicted CRMs at p-value $\leqslant$ 0.05.

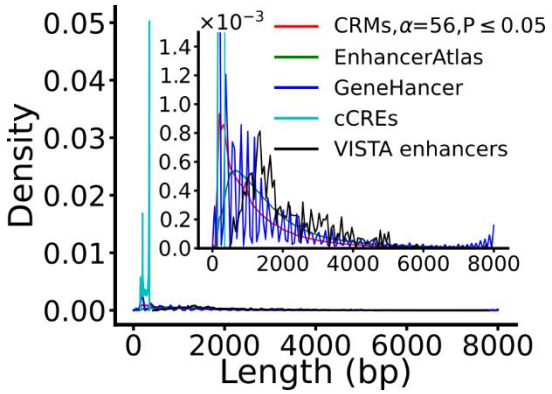

FIGURE 2-22: Distributions of lengths of the four sets of predicted enhancers/CRMs in comparison the that of the VISTA enhancers. The inset is a blow-up view of the indicated region on the Density axis.

2.3.12. At least half of the human genome might code for CRMs

What is the proportion of the human genome coding for CRMs and TFBSs? The high accuracy of our predicted CRMs and constituent TFBSs might well position us to more accurately address this interesting and important, yet unanswered question [170, 171]. To this end, we took a semi-theoretic approach. Specifically, we calculated the expected number of true positives and false positive in the CRMCs in each non-overlapping S_{CRM} score interval based on the predicted number of CRMCs and the density of S_{CRM} scores of Null CRMCs in the interval (FIGURE 2-23A), yielding 1,383,152 (98.45%) expected true positives and 21,821 (1.55%) expected false positives in the CRMCs (FIGURE 2-23B). The vast majority of the 21,821 expected false positive CRMCs have a low S_{CRM} score < 4 (inset in FIGURE 2-23A) with a mean length of 28 pb (FIGURE 2-12B), making up 0.02% (21,821x28/3,088,269,832) of the genome and 0.05% (0.02%/0.4403) of the total length of the CRMCs, i.e., a FDR of 0.05% in length (FIGURE 2-23C). On the other hand, as the CRMCs miss 1.49% of ClinVar SNPs in the covered genome regions (FIGURE 2-16A), the FNR of partitioning the genome in CRMCs and non-CRMCs would be < 2.49%(1-0.40)=1.49%, given the proportion of neutrality of 0.4 for the unrecalled ClinVar SNPs (FIGURE 2-16B). We therefore estimate false negative CRMCs make up 0.67% of the genome and 1.99% of the total length of the non-CRMCs, i.e., a false omission rate (FOR) of 1.99% (FIGURE 2-23C). Hence, the true CRM positions in the covered regions make up 44.68% (44.03%-0.02%+0.67%) of the genome (FIGURE 2-23C). In addition, as we argued earlier, the uncovered 22.53% genome regions have a 79.4% CRMC density as in the covered regions, thus, CRMCs in the uncovered regions would be about 10.32% (0.2253x0.4468x0.7940/0.7747) of the genome (FIGURE 2-23C). Taken together, we estimate that about 55.00% (44.68%+10.32%) of the genome might code for CRMs, for which we have predicted 80.02% [(44.03-0.02)/55.00]. Moreover, as we predict that about 40% of CRCs are made up of TFBSs (FIGURE 2-12A), we estimate that about 22.00% of the genome might encode TFBSs. Furthermore, assuming the mean length of CRMs is 2,049bp as VISTA enhancers and a

mean TFBS length of 10bp, we estimate that the human genome encodes about 828,965 CRMs (3,088,269,832x0.55/2049) and 67,941,963 TFBSs.

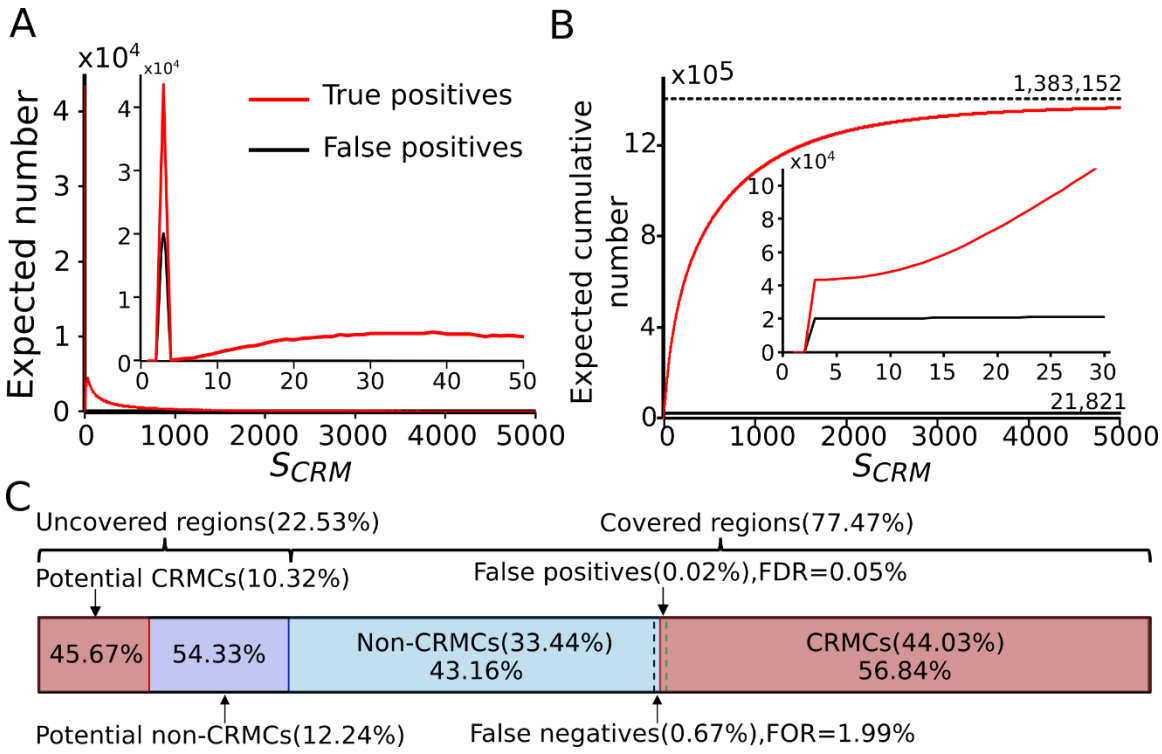

FIGURE 2-23: Estimation of the portion of the human genome encoding CRMs. A. Expected number of true positive and false positive CRMCs in the predicted CRMCs in each one-unit interval of the S_{CRM} score. The inset is a blow-up view of the axes defined region. B. Expected cumulative number of true positives and false positives with the increase in S_{CRM} score cutoffs for predicting CRMs. The inset is a blow-up view of the axes defined region. C. Proportions of the genome that are covered and uncovered by the extended binding peaks and estimated proportions of CRMCs in the regions. Numbers in the braces are the estimated proportions of the genome being the indicated sequence types, and numbers in the boxes are proportions of the indicated sequence types in the covered regions or the uncovered regions.

2.4. Discussion

Identification of all functional elements, in particular, CRMs in genomes has been the central task in the postgenomic era, and enormous CRM function-related data have been produced [157, 172]. Although great progresses have been made to predict CRMs in the genomes [96, 98, 102, 103, 105, 173] using these data, most existing methods attempt to predict cell/tissue specific CRMs using multiple tracks of epigenetic marks collected in the same cell /tissue types[87, 90, 98, 103, 105]. These methods can be only applied to few cell/tissue types for which the required epigenetic data are available[87, 90, 105], they are also limited by low resolution of predicted CRM boundaries , lack of constituent TFBS information[103, 105], and high FDRs as we show in this study (FIGURE 2-21B). To circumvent these limitations, we proposed a different strategy to first predict a static map of CRMs and constituent TFBFs in the genome [112, 113], just as the community has been practicing to find all genes encoded in the genome without necessarily knowing their functions in specific cell/tissue types [116]. It has been shown that TF binding data such ChIP-seq data are more accurate predictor of CRMs than CA and histone mark data [48, 107, 109, 112, 113], probably because it is mainly TFBSs in a CRM that define its structure and function. Therefore, we proposed to integrate as many as possible TF ChIP-seq data available for different TFs in various cell/tissue types to predict a more accurate and complete map of CRMs and constituent TFBSs. Another advantage of our approach is that we do not need to exhaust all TFs and all cell/tissue types of the organism in order to predict most, if not all, of CRMs and constituent TFBSs in the genome, because as CRMs are often repeatedly used in different cell/tissue types, developmental stages. Moreover, as binding sites of cooperative TFs of the ChIP-ed TF tend to be clustered in the adjacent regions around the summit of binding peaks of the TF [71, 73], by appropriately extending the called binding peaks in each dataset[117], we can increase the coverage of the genomes, thereby further reduce the number of datasets needed [112, 113]. In other word, we might only need a large but limited number of datasets that are less biasedly cover the entire functional genome after length extension to achieve the goal. Our earlier application of the

approach resulted in very promising results in the fly [113] and human[112] genomes even using an even relatively small number of strongly biased datasets available then. However, we were limited by the lack of a sufficiently large number of ChIP-seq datasets for more diverse TFs in various cell/tissue types [112, 113] and the inadequacy of our earlier algorithms to tackle the computational challenges imposed by the approach. In this study, we developed a new pipeline dePCRM2 to circumvents the shortcomings of the earlier version. The results we present clearly demonstrate that dePCRM2 large achieves our algorithmic goals, thereby we reveal a more prevalent cis-regulatory genome in humans than earlier thought.

More specifically, to the best of our knowledge, for the first time, dePCRM2 is able to very accurately partition the covered genome regions into two exclusive sets, i.e., the CRMCs or the regulatory genome, and the non-CRMCs, or the non-regulatory genome. Multiple pieces of evidence strongly support the highly accurate partition. First, even the CRMCs with the lowest S_{CRM} scores ((0,1]) are under stronger evolutionary constraints than non-CRMCs (FIGURE 2-14B and FIGURE 2-15B), indicating that even these small fraction (63,363 (4.51%), or 0.1% of the total length of the CRMCs) of low-scoring CRMCs are still more likely to be functional than non-CRMCs, not to mention CRMCs with higher S_{CRM} scores that are under even stronger evolutionary constraints. Second, with the increase in the S_{CRM} cutoff, the associated p-value decreases rapidly, while both the number and total length of the predicted CRMs only decrease slowly (FIGURE 2-11B), indicating the vast majority of CRMCs have small p-values, and are unlikely predicted by chance. Third, with the increase in the S_{CRM} cutoff (decrease in p-value), strength of evolutionary constraints on the predicted CRMs increase and rapidly saturate, followed by small increments, approaching the level that the VISTA enhancers are under selection (FIGURE 2-14C and FIGURE 2-15C). Fourth, all experimentally validated VISTA enhancers and almost all (97.51%) of well-documented ClinVar SNPs in the covered regions are located in the CRMCs (FIGURE 2-16), indicating dePCRM2 achieves a very low FDR in the CRMCs. Finally, our simulation study also

indicates that the partition has a very low FDR of 0.045% in the CRMCs and a low FOR of 1.99% in the non-CRMCs (FIGURE 2-23C).

We show that dePCRM2 might largely correctly predict the lengths of most CRMCs. First, at least most of the 783,132 (55.74%) of the 1,404,973 CRMCs might be in full-length as they all are longer that the shortest (428bp) known enhancer (428bp) in the VISTA database. These the remaining 621,841(44.26%) CRMCs shorter than 428bp make up only 7.42% of the total length of the CRMCs, and thus might be short CRMs or only component of longer CRMs. Second, with the increase in S_{CRM} score cutoff α (or decrease in p-value), lengths of the predicted CRMs increase (FIGURE 2-13). In particular, at α=676 (p-value ≤ $5x10^{-6}$), the resulting 428,628 predicted CRMs have an almost identical length distribution to that of the VISTA enhancers (FIGURE 2-13), while these CRMs recall 99.10% of the VISTA enhancers in the coverage regions. The failure to predict full-length CRMs of short CRM components might be due to insufficient data coverage on the relevant loci in the genome. This is reminiscent of our earlier predicted, even shorter CRMCs (average length 182bp) using a much smaller number and less diverse 670 datasets [112]. As we argued earlier [112] and confirmed here by the much longer CRMCs (average length 982bp) predicted using the much larger and more diverse datasets albeit still strongly biased to a few TFs and cell/tissue types (FIGURE 2-2 A and B). We anticipate that full-length CRMs of these short CRM components can be predicted using even larger and more diverse TF ChIP-seq data. Thus, efforts should be made in the future to increase the genome coverage and reduce data biases by including more untested TFs and untested cell types in the TF ChIP-seq data generation.

Interestingly, our predicted CRMs (at p-value < 0.05) achieve perfect (100.00%) and very high (97.43) sensitivity for recalling VISTA enhancers [118] and ClinVAR SNPs [158], respectively, but varying intermediate sensitivity from 64.50% (for GWAS SNPs) to 88.77% (for FPs) for recalling other eight CRM function-related elements datasets (FIGURE 2-16A). It appears that such varying sensitivity is due to varying FDRs from 0% (for VISTA enhancers) to 35.5% (for GWAS

SNPs) of the methods used to characterize the elements as CRMs or parts of CRMs (FIGURE 2-16B). Our finding that DHSs, TASs, and histone mark (H3K4m1, H3K4m3 and H3K27ac) peaks have high FDRs for predicting CRMs are consistent with an earlier study showing that histone marks or CA were less accurate predictor of enhancer activity than TF binding data[107]. In this sense, it is not surprising that our predicted CRMs substantially outperforms the three state-of-the-art sets of predicted enhancers/cCREs, i.e., GeneHancer 4.14 [98], EnhancerAtals2.0 [103] and cCREs[105], for both recalling VISTA enhancers and ClinVar SNPs (FIGURE 2-20B) and predicting full-length CRMs (FIGURE 2-22), probably because these three sets were mainly predicted based on overlaps between multiple tracks of CA and histone marks in various cell/tissue type. Although the constraint of overlaps between multiple tracks of epigenetic data might have reduced FDRs in these three sets of enhancers/cCREs[98, 103, 105], they might still suffer quite high FDRs (49.14%, 29.28% and 23.12% for EnhancerAtlas 2.0 enhancers, GeneHancer 4.14 enhancers and cCREs, respectively). It is worth pointing out that we used the smallest number of datasets to achieve the best results. More specifically, our CRMs and non-CRMCs were predicted based on 5,578 of the 6,090 TF ChIP-seq datasets, as 512 of which were filtered out due to their low quality or lack of overlaps with other datasets. In contrast, EnhancerAtlas enhancers were predicted using 8,005 datasets, including 4,159 TF ChIP-seq and 1,580 epigenetic datasets, as well as experimentally determined potential enhancers [103]. GeneHancer 4.14 is a meta-prediction based on the results of multiple algorithms[87, 90, 98] that use histone mark data. cCREs are the most updated prediction by the ENCODE phase 3 consortium[105], based on 703 DHS, 46 TAS and 2,091 histone mark datasets [105]. Although dePCRM2 can predict the functional states of CRMs in a cell/tissue type that have original binding peaks overlapping the CRMs, it cannot predict the functional states of CRMs in the extended parts of the original binding peaks in a cell/tissue if the CRMs do not overlap any available binding peaks of all TFs tested in the cell/tissue type. However, once a map of CRMs in the gnome as we predicted in this study is available, the functional state of each CRM in the map in any cell/tissue type could be studied in more focused

ways, or can be predicted based on overlap between the CRM and a single or few epigenetic mark datasets collected from the very cell/tissue type, such as CA, H3K27ac and/or H3K4m3 data. Anchored by corrected predicted CRMs, the prediction accuracy of these epigenetic marks could be high [107]. Thus, our approach might be more cost-effective for predicting both a static map of CRMs and constituent TFBSs in the genome and their functional states in various cell/tissue types.

We also show that by appropriately extending the called binding peaks in the datasets, we can substantially increase the power of available data, and therefore, substantially reduce the amount of data needed to predict to predict most, if not all, of the CRMs and constituent TFBSs encoded in the genome. For example, although originally called binding peaks in the strongly biased 6,090 TF ChIP-seq datasets used in this study cover only 40.96% of the genome, moderately extended peaks cover 77.47% of the genome, an 89.14 % increase. Remarkably, extended parts of the peaks contribute 42.12% nucleotide positions of the predicted CRMCs. On the other hand, dePCRM2 abandoned 38.60% of positions covered by the original binding peaks, which might be nonfunctional as they evolve like non-CRMCs (FIGURE 2-14A and FIGURE 2-15A). Thus, called binding peaks cannot be equivalent to CRMs or parts of CRMs as has been demonstrated earlier[149-151].

It has been estimated that the human genome encodes from 2,000 to 3,000 TFs belonging to at least 100 protein families [144, 174]. However, the exact number of TFs and TF families encoded in the genome remains unknown [144, 175]. Our prediction of the 201 UM families in the covered genome regions provides us an opportunity to estimate the number of TFs families encoded in the genome. As different TFs of the same protein family/superfamily bind indistinguishably similar motifs [145, 176], it is highly likely that a predicted UM is recognized by multiple TFs of the same family/superfamily. Indeed, 92 (78.63%) of the 117 (58.21%) UMs matching at least a known motif, match at least two. The remaining 84 (41.79%) UMs might be motifs of novel TF families that remain to be elucidated. On the other hand, the UMs recall 64 (71.91%) of the 89 known motif

families. Therefore, we estimate the lower bound of the number of TF/motif families encoded in the human genome to be around 142 (117+25) , considering that the uncovered regions of the genome might harbor novel UMs that do not appear in the covered regions.

The proportion of the human genome that is functional is a topic under hot debate [165, 177] and a wide range from 5% to 80% of the genome has been suggested to be functional based on difference sources of evidence [157, 161, 170, 177, 178]. The major disagreement is for the proportion of functional NCSs in the genome, mainly CRMs, which has been coarsely estimated to be from 8% to 40% of the genome [165, 177]. Moreover, a wide range of CRM numbers from 400,000 [165] to more than a few million [103, 157] encoded in the human genome have been suggested. However, to our knowledge, no estimate has been made on substantial evidence. Our predicted CRMCs cover a lower proportion (44.03%) of the genome than EnhancerAtlas 2.0 enhancers (58.99%)[103] that might have a FDR of 49.14%, but a higher proportion than cCREs (7.9%)[105] and GeneHancer v.4.14 enhancers (18.99%) [98], even though both sets have high FDRs (FIGURE 2-19B). The much higher accuracy of our predicted CRMs suggests that cCREs (7.9%) [105] and GeneHancer enhancer might underpredict, whereas EnhancerAtlas 2.0 might overpredict CRMs. Based on the estimated FDR and FNR of dePCRM2 in partitioning the covered genome regions into the CRMC and non-CRMC sets as well as the estimated density of CRMs in the uncovered regions relative to the covered regions (FIGURE 2-23C), we estimate that about 55.00% and 22.00% of the genome might code for CRMs and TFBSs, respectively, which encode about 828,965 CRMs and 67,941,963 TFBSs. Therefore, the number of our predicted CRMs is more than twice an earlier estimate of 400,000 [165], and they are encoded by a higher (55.00%) proportion of the genome than earlier thought 40% [165, 177]. We estimate our true positive CRMs cover 44.01% (44.03-0.02) of the genome, therefore, we might have predicted 80.02 % (44.01/55.00) CRM positions encoded in the genome. In summary, it appears that the cis-regulatory genome is more prevalent than originally thought.

2.5. Conclusion

We have developed a new highly accurate and scalable algorithm dePCRM2 for predicting CRMs and constituent TFBSs in large genomes by integrating a large number of TF ChIP-seq datasets for various TFs in a variety of cell/tissue types of the organisms. Applying dePCRM2 to all available more than 6,000 TF ChIP-seq datasets, we predicted an unprecedentedly complete, high resolution map of CRMs and constituent TFBSs in 77.47% of the human genome covered by the extended binding peaks of the datasets. Evolutionary and experimental data suggest that dePCRM2 achieves very high prediction sensitivity and specificity. With more diverse and balanced data covering the whole genomes becoming available in the future, it is possible to predict more complete maps of CRMs and constituent TFBSs in the human and other important genomes.

CHAPTER 3: CONTINUOUS MODEL OF TRANSCRIPTIONAL FACTOR BINDING

3.1. Background

Cis-regulatory modules (CRMs) are harbors of transcription factor binding sites (TFBSs), and the transcription factors regulate expressions of their target genes by interacting with the TFBSs in the CRMs in an additive manner or in a cooperative manner. To elucidate the functional mechanisms of the TF binding patterns in the CRMs, research community proposed two distinct models, the billboard model [179] and the enhanceosome model [179]. The billboard model [179, 180] hypothesizes that the TFs bind on the TFBSs in a combinatorial/additive manner, and the TFs do not form a cooperative complex via protein-protein interaction (PPI) before or concurrently binding the TFBSs, thus, this model is not strict on the spacing and orientation information of the TFBSs in the CRMs. The enhanceosome model [179, 181] hypothesizes that the cooperative TFs form a cooperative complex via PPI before or concurrently binding the underlying DNA sites to regulate the gene expression, thus, only strict spacing and orientation features of the cooperative TFBSs could facilitate this binding type. Based on the ideas of these two models, the TF collective model hypothesizes that the TFs bind on the TFBSs in a cooperative manner but the TFBSs "grammar" in the enhancers is flexible [182]. Thus, the architecture of the enhancer which can satisfy the above two assumptions might locate in a spectrum between the billboard model and the enhanceosome model. In addition, multiple evidence suggest that TF binding is highly overlapping throughout the genome [183, 184]. And the enhanceosome model is associated with more conservative regulatory regions, while the billboard model is associated with less conservative regulatory regions [185]. However, the relationship between phenomenon of the highly overlapping TFBSs and complex TFs binding mechanisms is not clear. To address this problem,

we need to know the comprehensive landscape of the TFBSs and their evolutionary behavior across the whole genome.

In this chapter, we will classify the predicted CRMCs into multiple groups based on their lengths and distances to multiple *cis*-regulatory elements, such as FAMTON enhancers, FAMTON promoters, and VISTA enhancers. Then we will systematically analyze the overlapping status of the TFBSs in all predicted CRMCs in human genome. Finally, we will analyze the evolutionary behavior of the TFBSs and the TFBS-deplete regions in the CRMs.

3.2. Methods and materials

3.2.1. Prediction of *cis*-regulatory modules (CRMs) in human genome

The CRMs was predicted using the dePCRM2 pipeline, shown in Methods and materials section in chapter 2.

3.2.2. Identification of the distance between the CRMs with nearest TSS

We download the TSSs of protein coding genes from Ensembl genome database[186], then we identify the distance between CRMs with the first occurring nearest TSSs using BEDTools[187] as follows,

bedtools closest -a CRMs_file -b TSS_file -d -t first > CRMs_with_first_occurring_nearest_TSS.

If a CRM overlaps with its nearest TSS, then we will set the distance as 0.

3.2.3. Identification of TFBSs and spacers in the CRMs

We merge all the overlapping TFBSs into TFBS islands and define the TFBS-deplete regions as spacers.

3.3. Results

3.3.1. Classification of predicted CRMs

To see if it is possible to classify predicted CRMs according to their distances from the nearest annotated transcription starting sites (TSSs), we compared the distribution of such distances of the 1,155,151 predicted CRMs at S_{CRM} cutoff α=56 (p-value ≤ 0.05) with those of entire sets of 976 VISTA enhancers [118], 184,328 FANTOM promoters [119] and 32,684 FANTOM enhancers [120]. Here, if a predicted CRM overlaps the nearest TSS, we set the distance to be 0. The distribution of the distances of VISTA enhancers has a sharp peak around 0 with a largely uniform tail to 2,053,347bp and a median of 67,885bp. Forty VISTA enhancers (4.10%) overlap the nearest TSSs, hence, they may contain a promoter in addition to an enhancer. The remaining 936 (95.90%) VISTA enhancers are largely distal (FIGURE 3-1A). The distances of FANTOM promoters display a strongly right-skewed distribution with a sharp peak around 0 and a median of 1,917bp (FIGURE 3-1A). As expected, a considerable proportion of FANTOM promoters either overlap TSSs (13.10%) which are likely core promotors or are close to the nearest TSSs (distance shorter than 500bp, 28.12%) which are likely proximal promoters. The remaining 58.78% of FANTOM promoters are more than 500bp away from the nearest TSSs (FIGURE 3-1A). The distribution of the distances of FANTOM enhancers show a broad peak around 800bp with a long tail and a median of 22,915bp (FIGURE 3-1A). Notably, FANTOM enhancers with a distance shorter than 500bp are somehow depleted (FIGURE 3-1A). More specifically, only 125 (0.38%) FANTOM enhancers overlap with TSSs, 189 (0.58%) are within 500bp from the nearest TSSs, and 32,370 (99.04%) are at least 500bp away from the nearest TSSs. These results suggest that VISTA enhancers and FANTOM enhancers are quite different in their distances to the nearest TSSs. Interestingly, the distances of predicted CRMs show a two-modal distribution ranging from 0 to 31.26Mbp (FIGURE 3-1). Mode 1 is a sharp peak at 0, containing 39,241 (3.40%) CRMs (FIGURE 3-1A) that overlap the nearest TSSs. Mode 2 consists of the remaining part of the distribution with a broad peak around

800bp and a median distance of 56.87Kbp, containing 1,115,910 (96.60%) CRMs. Similar to FANTOM enhancers, most (96.00%) of CRMs in this mode are at least 500bp away from the nearest TSSs, and CRMs with a distance shorter than 500bp (4.00%) is relatively depleted (FIGURE 3-1A).

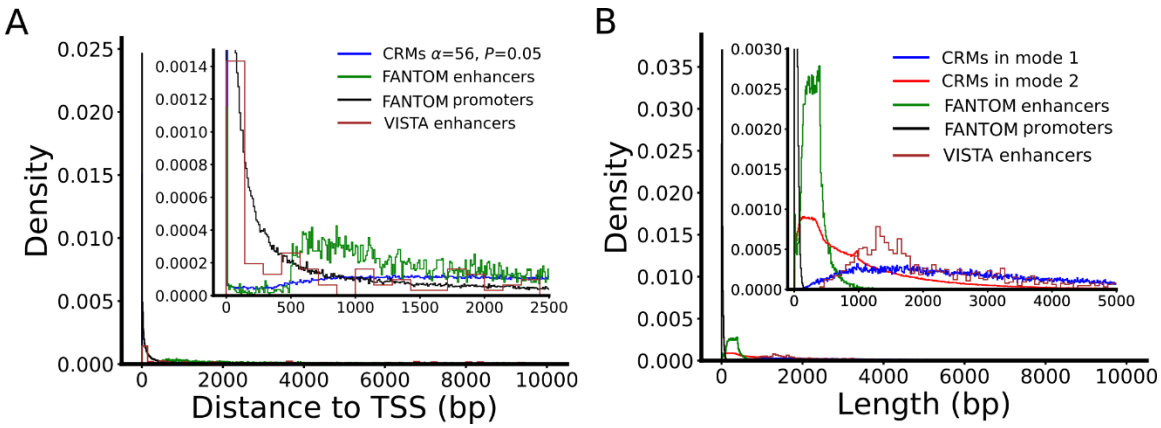

FIGURE 3-1: Classification of predicted CRMs according to their distances to the nearest TSSs. A. Bimodal distribution of the distances of predicted CRMs to the nearest TSSs in comparison with those of VISTA enhancers, FANTOM enhancers and FANTOM promoters. B. Distributions of the lengths of predicted CRMs in mode 1 (overlapping with the nearest TSS) and in mode 2 (not overlapping with the nearest TSS) in comparison with those of VISTA enhancers, FANTOM enhancers and FANTOM promoters. C. CRM (chr10:69048146-69048366) containing a core promoter plus a proximal promoter overlaps a FANTOM promoter and proximal promoter of gene ADGRE2. D. CRM (chr1:11846926-11848930) containing a core promoter and an enhancer overlaps a FANTOM promoter (chr:11847772-11847790) and VISTA enhancer 2123 of gene NPPA. E. CRM (chr10:69048146-69048366) overlaps a distal FANTOM enhancer (chr10:69048135:69048356) upstream of the target gene SRGN. F. CRM (chr11:33942659-33946495) overlaps VISTA enhancer 1858 upstream of gene LMO2.

Next, we compared the distributions of the lengths of the CRMs in the two modes with those of entire sets of VISTA enhancers, FANTOM promoters and FANTOM enhancers. As expected, FANTOM promoters are quite short with a median length of 15bp, and uniform in length with 99.41% being shorter than 100bp (FIGURE 3-1B). FANTOM enhancers display a slightly right-skewed distribution with a peak around 800bp, a median of 288bp and 99.99% being shorter than 1,000bp. Thus, like the 113 FANTOM enhancers that overlap 91 VISTA enhancers, the entire set of FANTOM enhancers are also generally shorter than the entire set of VISTA enhancers that range from 428 to 8,061 bp with a median of 1,677bp FIGURE 2-13, suggesting again that FANTOM enhancers might be short CRMs or CRM components of long enhancers due probably to the limitations of the eRNA-seq techniques for their determination. The distribution of the

lengths of CRMs in mode 1 ranges from 48 to 60,924bp with a broad peak around 1,500bp and a median of 2,638bp, covering the entire region of the lengths of VISTA enhancers (FIGURE 3-1B). Only 21 (0.54%) of the CRMs in mode 1 are shorter than 100bp, so they are likely core promoter-containing promoters. The remaining 39,220 (99.46%) CRMs in the mode are longer than 100bp, and hence might contain a core promoter plus other regulatory elements such as proximal promoters and enhancers. Of the 836 and 22,235 CRMs that recover VISTA enhancers and FANTOM enhancers, 101 (12.08%) and 3,030 (13.63%) are located in mode 1, indicating the respective recovering CRMs are similarly enriched in the mode that consists of only 3.4% of predicted CRMs. Relatively long lengths of CRMs in the mode indicate that our algorithm concatenates the core promoter with other adjacent regulatory elements due probably to the close clustering nature of these elements in the neighborhoods of TSSs. In agreement with this, 61.82% of CRMs in this mode overlap FANTOM promoters. On the other hand, the lengths of CRMs in mode 2 show a strongly right-skewed distribution with a median length of 710bp, a left narrow peak overlapping the peak of the distribution of FANTOM enhancers and a right long tail overlapping the entire range of the distribution of VISTA enhancers. As expected, 19,205 (86.37%) and 735 (87.92%) of the 22,235 and 836 predicted CRMs covering FANTOM enhancers and VISTA enhancers, respectively, are located in this mode. Therefore, CRMs in the two modes differ not only in their distances to the nearest TSSs, but also in their lengths. Based on these observations, we classify CRMs in mode 1 as core promoter-containing CRMs (3.40%) and those in mode 2 as non-core promoter-containing CRMs (96.17%). A core promoter-containing CRM contains a core promoter and a proximal promoter (FIGURE 3-2A), or a core promoter and an enhancer (FIGURE 3-2B). A non-core promoter-containing CRM is typically located at least 500bp upstream (FIGURE 3-2C) or downstream (FIGURE 3-2D) of the nearest TSSs. It appears that CRMs recovering VISTA enhancers and CRM recovering FANTOM enhancers are enriched for core promoter-containing CRMs (4.1%) and non-core promoter-containing CRMs (99.62%) compared to the expected 3.40% and 96.17% by chance, respectively.

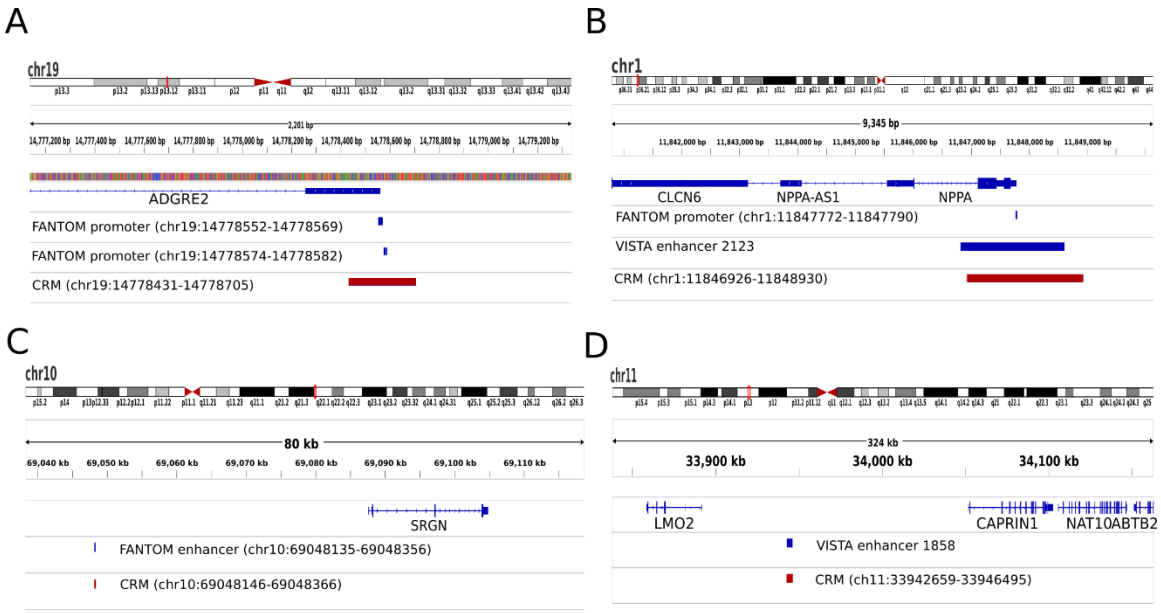

FIGURE 3-2: Examples of overlap between CRMs and FANTOM elements. A. CRM (chr10:69048146-69048366) containing a core promoter plus a proximal promoter overlaps a FANTOM promoter and proximal promoter of gene ADGRE2. B. CRM (chr1:11846926-11848930) containing a core promoter and an enhancer overlaps a FANTOM promoter (chr:11847772-11847790) and VISTA enhancer 2123 of gene NPPA. C. CRM (chr10:69048146-69048366) overlaps a distal FANTOM enhancer (chr10:69048135:69048356) upstream of the target gene SRGN. D. CRM (chr11:33942659-33946495) overlaps VISTA enhancer 1858 upstream of gene LMO2.

3.3.2. Distribution of TFBSs supports the continuum model of TF binding

Using the 428,628 putative full-length CRMs predicted with p-value $\leq 5x10^{-6}$, we further analyzed some properties of TFBSs in CRMs. TFBSs in these putative full-length CRMs have a mean length of 10.00bp (FIGURE 3-3A), in agreement of the length of UMs (10.59bp) (FIGURE 2-8B). To see how TFBSs are arranged and distributed in a CRM, we computed the distance between two immediate adjacent TFBSs in the putative CRMs. As shown in FIGURE 3-3B, adjacent TFBSs may overlap each other by 1~20bp with 10, 9 and 8bp occurring most frequently, and it is rare that two adjacent TFBSs are away from each other by more than 100bp. We merged overlapping TFBSs in a CRM to form nonoverlapping TFBS islands, which have a slightly larger mean length (12.87bp) than TFBSs (FIGURE 3-3A). The distance between adjacent TFBS islands ranges from 1 to 1,990bp with a mean of 23.24bp. A CRM contains 2.11 ~ 80.18 TFBSs/100bp with a mean of 9.66 TFBSs/100bp (FIGURE 3-4A). As 40.48% of the length of the predicted CRMs

are non-overlapping TFBSs (FIGURE 2-12A), and the remaining 59.52% are inter-TFBS spacers, each nucleotide in a TFBS island is covered by an average of 9.66x10/40.48 = 2.39 TFBSs. Thus, a nucleotide position in an island can be potentially bound by multiple TFs. In agreement with this, it has been shown that different TFs can compete for partially overlapping binding sites [188] or bind synergistically to the opposite faces of the DNA duplex [189]. TFBSs in a CRM belong to from 1 to 102 UMs with a mean of 38.55 UMs (FIGURE 3-4B), presumably reflecting interactions of cognate TF for cooperative transcriptional regulation. For example, a predicted CRM (ch2:41963100-41368696) that overlaps VISTA enhancer 2553 located upstream of gene C2orf91 is predicted to harbor 1,022 largely overlapping binding sites of 65 UMs (FIGURE 3-5A), of which 40 match known TF motif families, and 28 of them are previously known in the VISTA enhancer [190], while the remaining 12 are newly predicted in this study.

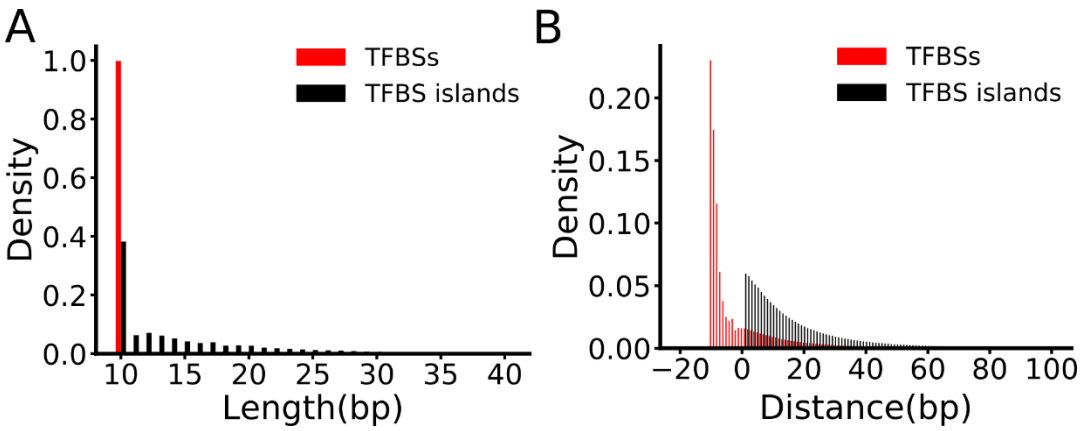

FIGURE 3-3: Features of putative TFBSs in the predicted full-length CRMs. A. Distribution of the lengths of putative TFBSs and TFBS islands in predicted full-length CRMs. B. Distribution of the distance between adjacent of putative TFBSs and TFBS islands in the predicted full-length CRMs.

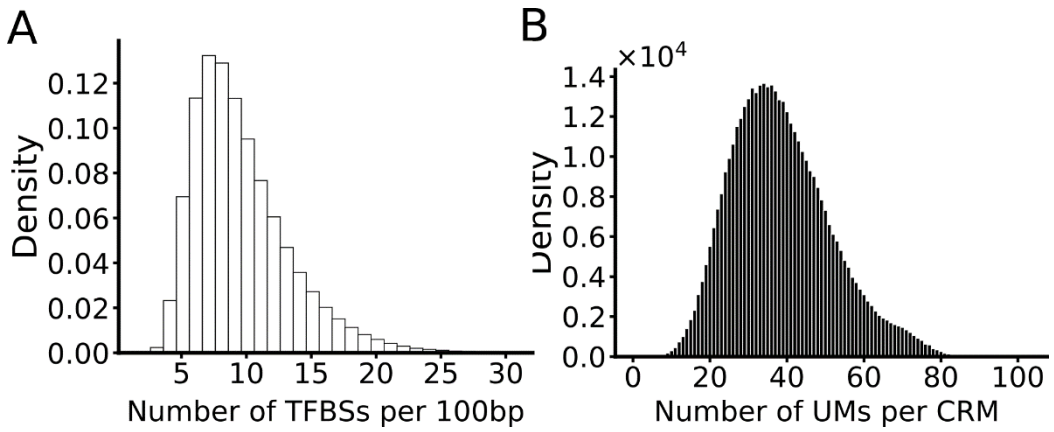

FIGURE 3-4: Binding properties of the TFBSs. A. Distribution of the predicted full-length CRMs with different number of TFBSs per 100bp. B. Distribution of the predicted full-length CRMs with biding sites for different numbers of UMs.

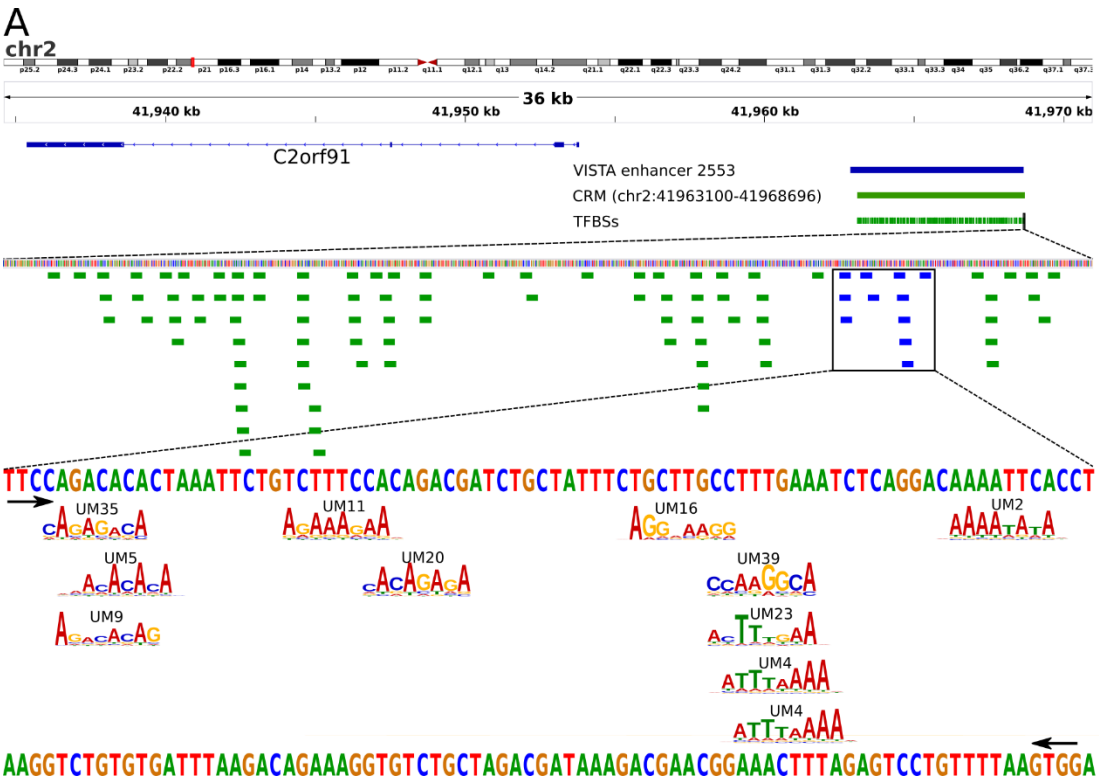

FIGURE 3-5: Example CRM of C2orf91 gene. A predicted CRM (ch2:41963100-41368696) that overlaps VISTA enhancer 2553 located upstream of gene C2FOR91 harbors numerous overlapping binding sites for different TFs. The blew-up view shows the locations of 12 putative TFBSs and the logos of the UMs that each belongs to. Note that the sites for UM11 and UM16 are on the reverse strand.

As shown in FIGURE 3-6A, TFBS islands and inter-TFBS spacers occupy an average of 41.17% and 58.83% of the length of the putative full-length CRMs, respectively. To see how TFBS islands

and inter-island spacers evolve, we compared the distributions of the GERP and phyloP scores of their nucleotide positions. Interestingly, the distribution of the GERP scores of spacers differs only slightly from that of TFBS islands, with spacers (0.3003) having a slightly higher (6.30%) proportion of neutrality than TFBS islands (0.2825), indicating that spacers are only slightly less conserved than TFBS islands. Nevertheless, spacers are under much stronger evolutionary constraints than the non-CRMCs (FIGURE 3-6B), suggesting that spacers also might play a role in CRM functions. Similar results are observed using the phyloP scores (FIGURE 3-6C). These results support the continuum model of TF binding in CRMs for transcriptional regulation, in which many TFs compete for overlapping sites with a continuum spectrum of binding affinity [144, 191].

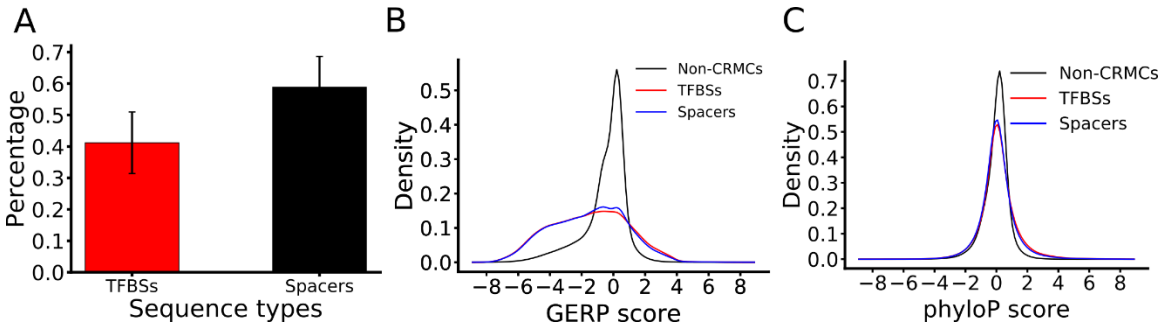

FIGURE 3-6: Evolutionary constraints on TFBS islands and spacers in CRMs. A. Proportion of the lengths of putative TFBS islands and spacers in the predicted full-length CRMs. B,C. Distribution of the GERP and phyloP scores of TFBS islands and spacers in the predicted full-length CRMs in comparison with that of the non-CRMCs.

3.4. Discussion

The landscape of the CRMs across the whole genome and TF binding modes are interesting topics. By analyzing the distances between the predicted CRMs to the CRMs in known databases, we observed that the distribution of the distances of the predicted CRMs to the nearest TSSs is bimodal (FIGURE 3-1A). Based on the length distribution modes of the predicted CRMs (FIGURE 3-1B), we classified the predicted CRMs (at p-value ≤ 0.05) into the core promoter-containing CRMs (3.40%) that overlap TSSs, and the non-core promoter-containing CRMs (96.6%) that are generally located at least 500bp away from the nearest TSSs. The core promoter-containing CRMs (median length 2,638bp) tend to be longer than the non-core promoter-containing CRMs (median length 210bp) (FIGURE 3-1B). Since an unexpectedly higher proportion (4.10% vs 3.40%) of the relatively long VISTA enhancers map to the core promoter containing CRMs, while the opposite is true for shorter FANTOM enhancers (0.38% vs 3.40%), we might have predicted most CRMs in this category in full-length. On the other hand, since the vast majority of both the VISTA enhancers (95.90%) and the FANTOM enhancers (99.62%) map to the non-core promoter-containing CRMs, we might have predicted a considerable proportion of CRMs in this category in full-length and in part-length. It appears that the currently available TF ChIP-seq datasets favor the prediction of the core promoter-containing CRMs in full-length, and the VISTA enhancers are biased to this category, while the FANTOM enhancers are biased to the non-core promoter-containing CRMs. It is unclear why the FANTOM enhancers tend to be shorter than the VISTA enhancers and our predicted CRMs (FIGURE 3-7). However, it may be related to the limitations of the eRNA-seq techniques for determining the FANTOM enhancers. As different FANTOM enhancers overlap the same VISTA enhancers and/or our predicted CRMs, it is possible that most FANTOM enhancers might be the components of the full-length CRMs.

Our results show that although only about 41% of the nucleotides in the CRMs are TFBSs, a CRM contains an average of 9.66 TFBSs/100bp, implying that each nucleotide in a TFBS island is

covered by an average of 2.39 TFBSs. Thus, TFBSs in a CRM tend to overlap with each other extensively. This result is in consistent with the earlier reports in the fly [192, 193] and mammals including human based on the extensive overlapping of the binding peaks in various TF ChIP data [118]. To our surprise, we find that the inter-spacers in a CRMs are under almost the same strength of evolutionary constraints as TFBSs (FIGURE 3-8B), suggesting that the vast majority of spacers might also be functional. These results strongly support the continuum model of TF binding in CRMs for gene transcriptional regulation [191], in which different TFs compete for overlapping sites with a continuum spectrum of binding affinity [191, 194]. The landscape of binding and the effects on gene transcription are determined by the concentration of the TFs [195] and the number of their binding sites with various affinity in the CRMs [196].

3.5. Conclusion

We classified the CRMs into two modes, the core promoter-containing CRMs (3.40%) and the non-core promoter-containing CRMs (96.17%). A core promoter-containing CRM contains a core promoter and a proximal promoter or a core promoter and an enhancer near the TSS, while a non-core promoter-containing CRM contains the distal enhancers which are located at least 500bp upstream or downstream of the nearest TSS. And then we compared the length distribution of the CRMs in the two modes with the VISTA enhancers, the FANTOM enhancers and the FANTOM promoters. We also analyzed the distribution and conservation properties of the TFBS in the full-length CRMs, and we noted that each nucleotide in a TFBS island is covered by an average of 2.39 TFBSs, thus, a nucleotide position could be bounded by different TFs with different affinities or be bounded by TFs at the opposite faces of the DNA duplex. To see the conservation properties of the TFBSs in the full-length CRMs, we compared the GERP and phyloP score distribution of the TFBSs and the inter-spacers in the full-length CRMs, then we observed that the spacers are under a slightly less evolutionary constrain than that of the TFBSs, but are under much stronger evolutionary constrains than that of the non-CRMCs. Thus, both the distribution and the evolutionary constraints of the TFBSs support the continuum model of the TF binding in CRMs for the transcriptional regulation, in which many TFs compete for overlapping sites with a continuum spectrum of binding affinity.

CHAPTER 4: DECIPHERING EPIGENOMIC CODE USING DEEP LEARNING

4.1. Background

Cell differentiation is achieved by the remodeling of the same genome that each cell inherits from the zygote. Genome remodeling involves alterations of methylation of certain cytosine residues in the genomic DNA and changes in various covalent modifications of histones in the nucleosomes, conferring a unique epigenome to each resulting cell type that expresses a unique set of gene products [197]. Increasing lines of evidence have suggested that the epigenome in a cell type is established step-wisely along the developmental lineage through the interplay of genomic sequence, chromatin remodeling systems and extracellular environmental cues [198-201]. As the latter two factors are the results of interactions of the products of genomic sequences, the epigenome of a cell type is ultimately determined by the genomic sequence that recruits the chromatin remodeling systems [36, 202-204]. For example, in a recent study, Whitaker and colleagues [36] have shown that short DNA motifs enriched in the epigenetically modified genomic regions could predict the specific histone modifications in specific cell types using a random forest-based method. However, this method could not discover sequence determinants ab initio because pre-selected motifs were needed to train the models. Therefore, new methods are needed to gain a better understanding of the sequence determinants that specify the unique epigenome of each cell type produced during cell differentiation.

Recent progress in machine-learning has demonstrated that deep convolutional neural networks (CNNs) can achieve very high accuracy in predicting transcription factor (TF) binding affinity [205] and epigenetic marks in various cell types [206-208]. Unlike traditional neural networks, the kernels in the convolutional layers in a CNN can automatically learn the features of the objects (i.e., the sequence motifs in epigenetically modified regions), and thus the learned features can provide

insights into the underlying mechanisms of the modeling systems. Although efforts have been made to explain the learned motifs in epigenetically modified regions in biological contexts types [206-208], the mixed CNN models employed in these earlier studies lack the power of comparison, limiting their ability to explain the learned motifs for their roles in determining the unique epigenetic modification patterns in different cell types. To overcome these shortcomings, we developed two types highly accurate CNN models to facilitate the explanation of the learned motifs: the cell type model to predict different histone marks in a given cell by learning motifs that specify the histone marks in the cell type, and the histone mark model to predict different cell types by learning motifs that determine different patterns of a given histone mark in different cell types. To evaluate the capability of the models to learn the histone mark-determining motifs, we applied them to a dataset of six histone marks obtained in four human CD_4^+ T cell types produced at different stages of cell differentiation [209], i.e., the native T (Tn) cells, central memory T (Tcm) cells, T effector memory (Tem) cells and CD_4^+ terminally differentiated $CD_{45}RA^+$ memory (Temra). The relatively rich knowledge about the regulators and the differentiation process of these T cell subpopulations could facilitate the validation of predictions. Indeed, we found that many sequence motifs learned in the CNN models of both the cell types and histone modifications are highly similar to known binding motifs of TFs known to play important roles in CD_4^+ T cell differentiation. Intriguingly, the shared motifs learned in different cell models support the linear model of CD_4^+ T cell development, consistent with the earlier results based on the patterns of changes in DNA methylation and DNase accessibility of the genome as well as transcriptomes in the cells [209], while the shared motifs learned in different histone mark models reflect the functional relationships of the marks. Furthermore, by computing the s of the learned motifs on the prediction of the CNNs, we were able to pinpoint specific roles and interactions of their cognate TFs in determining unique histone modification patterns in different cell types, thereby providing new insights into the underlying mechanisms of histone modifications during cell differentiation.

4.2. Methods and materials

4.2.1. Datasets

Human CD4+ T cells dataset. We downloaded from European Genome-Phenome Archive the ChIP-seq datasets for six histone marks H3K4me1, H3K4me3, H3K27me3, H3K27ac, H3K9me3 and H3K36me3 in four different human CD_4^+ T cell types native T (Tn), central memory T (Tcm), T effector memory (Tem), and CD_4^+ terminally differentiated $CD_{45}RA^+$ memory T (Temra) cells [209].

Human embryonic stem cells dataset. We downloaded from the Roadmap Epigenomics Project [210] the ChIP-seq datasets for six histone marks H3K4me1, H3K4me3, H3K27me3, H3K27ac, H3K9me3 and H3K36me3 in H1 human embryonic stem cells (H1) and in four cell types derived from H1, including trophoblast-like (TBL), mesendoderm (ME), mesenchymal (MSC) and neural progenitor (NPC) cells.

4.2.2. Peak calling, filtering and merging

To identify genome regions that are modified by different histone marks, we called tight and broad histone modification peaks [36] using MACS2 [211]. The tight peaks including H3K27ac, H3K4me1 and H3K4me3 are typically < 1 kbp. The broad peaks including H3K27me3, H3K36me3 and H3K9me3 are typically > 1 kbp. The tight peaks were called as follows:

macs2 callpeak -t bam/tagAlign file -n name -c control file –outdir output dir -g hs -q 0.05 –nomodel –extsize fragment length

The broad peaks were called as follows:

macs2 callpeak -t bam/tagAlign file -n name -c control file –outdir output dir -g hs –broad –broad-cutoff 0.1 –nomodel –extsize fragment size

The fragment sizes were estimated using phantompeakqualtools [212, 213].

We discarded peaks whose $-log10(qvalue)$ was less than 2 or whose length was greater than 10,000 bp for their low quality or too long length. We also removed the peaks that overlapped the blacklisted regions of the human genome [214], which are regions showing artificially high signal in all NGS experiments. To ensure only regions of high confidence were considered, we only used the intersection of at least two replicates when possible. We extracted and merged the peaks using BedTools [187], and used the CRCh37/hg19 genome assembly for all the analyses.

4.2.3. Data representation

To prepare the input for the deep CNN models, we segmented the human whole genome (CRCh37/hg19) into 200-bp bins[206]. For a cell model, we labeled each bin with a binary vector with each bit indicating whether it was modified by the corresponding histone mark (1) or not (0) in the cell type. For a histone mark model, we labeled each bin with a binary vector with each bit indicating whether it was modified by the mark in the corresponding cell type (1) or not (0). We say that a bin overlaps with a peak if the overlapping portion of the bin with the peak is above a threshold. To achieve the best prediction results, we tested different thresholds of 0.5, 0.6, 0.7, 0.8 and 0.9, and chose the threshold with the highest accuracy for the final analysis. We discarded the bins that had no overlap with any histone modifications. We then extended the 200-bp bin into 1,000-bp sequence centered on the middle of the 200-bp bin for context learning [206]. Each extended 1,000-bp sequence was represented by a 1,000 × 4 binary matrix as the input to the CNN models, and each row was one hot vector to represent the presence or absence of A, C, G, T at the nucleotide position. If a nucleotide position is N in the genome, we represented it as [0.25, 0.25, 0.25, 0.25][208].

4.2.4. Convolutional neural networks

CNNs are a type of feed-forward artificial neural networks, usually consisting of an input layer, multiple convolutional layers, one or more fully connected layers and an output layer. Our CNN models (FIGURE 4-1) are made of a stack of three units each consisting of a convolutional layer, a pooling layer and a batch normalization layer, followed by a fully connected layer and an output layer. We apply a rectified linear unit (ReLU) transform as the activation function after a convolution layer (FIGURE 4-1), which helps to prevent vanishing gradient problem [215, 216]:

$$Convolution(X)_{lk} = ReLU\left(\sum_{l=0}^{L-1}\sum_{d=0}^{D-1} W_{ld}^{k} X_{i+l,d}\right), \quad 4\text{-}1$$

where X is the input, L is the input length, D is the input dimension, i is the output position, and k is filters' index. ReLU is defined as,

$$ReLU(x) = \max(0, x) \quad 4\text{-}2$$

To decease internal covariate shift and accelerate training, we apply a batch normalization layer after the convolutional layer[217]. Furthermore, we apply a max pooling layer after the batch normalization layer, which extracts the maximum activation value from each receptive field in the prior layer. Three convolutional layers contain 320, 300 and 300 kernels, respectively, and the fully connected layer has 1,000 units with a sigmoid activation function feeding into the output layer (FIGURE 4-1). We use a sigmoid function as the activation function of the output layer to conduct multi-task prediction,

$$y(X) = Sigmoid(X) = \frac{1}{1 + e^{-WX}} \quad 4\text{-}3$$

where y(X) is the prediction of the output layer, X is the output of the previous layer, and W is the weight matrix of the output layer. We implemented the CNN models using Theano [218] and Lasagne [219].

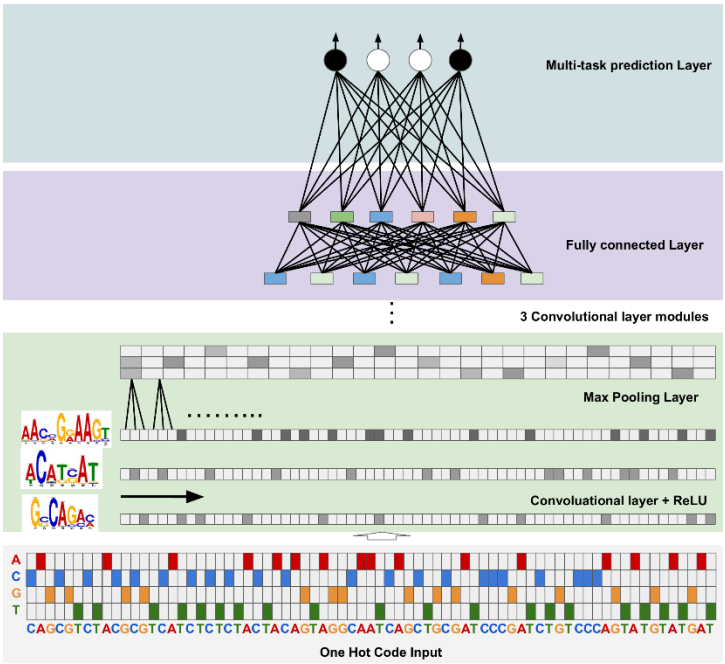

FIGURE 4-1: Architecture of the convolutional neural networks.

4.2.5. Model training, validation and evaluation

We split a dataset into a training dataset, a validation dataset and a test dataset with a ratio about 2:1:1, and the objective function is binary cross entropy. We apply a stochastic gradient descent to minimize the objective function by updating all model parameters using RMSprop with a learning rate 0.001 on minibatch [220]. To avoid overfitting, we apply L1 and L2 regularization terms and the early stopping strategy. To keep the filters free to grow based on input sequences, we only apply L1 and L2 regularization terms to the fully connected layer. To quickly choose the best set of hyperparameters of the models, we use parallel random search and apply L1 and L2 as well as maximum epochs as shown in TABLE 4-1.

TABLE 4-1: Hyper-parameter configurations for training the models.

Trail	L1	L2	Patience	Max epochs	Batch size
1	1e-07	2e-08	5000	20	128
2	2e-07	4e-08	5000	20	128
3	3e-07	8e-08	5000	20	128
4	4e-07	2e-07	5000	20	128
5	5e-07	4e-07	5000	20	128
6	6e-07	8e-07	5000	20	128

Overlap is 0.5, 0.6, 0.7, 0.8, 0.9

We performed the receiver operating characteristic (ROC) curve analysis and used the area under the curve (AUC) to evaluate the performance of the models. We also define the accuracy of a model as follows:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \quad 4\text{-}4$$

where TP is true positive, TN is true negative, FN is false negative and FP is false positive.

4.2.6. Interpretation of the kernels/filters in the first convolutional layer

The first convolutional layer of the models scans the DNA sequences with its kernel or filters to capture the k-mer motifs that differentiate modified and unmodified DNA sequences. Thus these filters potentially correspond to the binding motifs of TFs or chromatin remodeling proteins whose interactions with the motifs may lead to the specific modifications at the loci. To reveal such these motifs, we construct a position weight matrix (PWM) for each filter by extracting k-mers in the test dataset, which has a score against the filter greater than a threshold defined as,

$$Threshold = (\alpha_{max} - \alpha_{min}) \times \beta, \quad 4\text{-}5$$

where α_{max} and α_{min} are the maximum and minimum activations for a k-mer across all sequences in the test dataset, respectively, and β is a ratio constant. For each filter, we evaluated β

ranging from 0.3 to 0.8, and chose the resulting PWM with the highest information content. We discard the resulting PWMs with 0 information content. To evaluate the inference of a filter on the model's prediction, we nullify forward information of the filter by setting its output as its mean output over all nucleotides of all sequences in the test dataset [208], and quantify each filter's inference as the sum of square of the difference of the prediction probability in the test dataset before and after the nullification as follows,

$$Influence(k) = \sum_{x \in D} \left(P_{pre}(x) - P_{aft}(x) \right)^2, \quad 4\text{-}6$$

where D is the test dataset and $P_{pre}(x)$ and $P_{aft}(x)$ are the prediction probabilities before and after nullifying the filter k, respectively.

4.2.7. Motif conservation analysis

We used Fimo [221] to scan sequences for binding sites of each motif as follows:

fimo –parse-genomic-coord –thresh 1e-5 –bgfile fasta file background model –oc output_folder motifs_meme target_sequences

We used a 5th-order Markov model [222] to generate the background sequences as follows:

fasta-get-markov -m 5 -dna sequences background_model

We extracted the phastCons [223] score for each position in each binding site, and calculated a conservation score for each motif as the mean the PhastCons scores of all the binding sites of each motif learned in the models. To study the relationship between the inferences of the learned motifs and their conservation levels, we computed the Pearson correlation coefficient between them, and tested the null hypothesis of the non-correlation using two-tailed p-values,

$$r = \frac{cov(I,C)}{\sqrt{var(I)}\sqrt{var(C)}}, \quad 4\text{-}7$$

where I, C are the inference and phastCons scores of motifs, respectively, and r is the Pearson correlation coefficient.

4.2.8. Merging highly similar motifs

To merge similar motifs learned in all the cell and histone mark models, we compared each motif with all other motifs using TOMTOM [224], and constructed a graph by connecting two motifs if they were a pair of bidirectional best hits with a minimum overlap of 7 bps and E value < 0.1. We then cut the network into connected components using Networkx [225]. Some components are singletons containing a single original motif, while others are formed by multiple highly similar original motifs. We consider each of these components as a unique motif. To find the PWM for the merged motifs, we performed motifs finding on the merged binding sites using ProSampler [226].

4.2.9. Prediction of interactions between cognate TFs of learned motifs

To predict possible interactions between the cognate TFs of the learned motifs, we applied a linear model to the changes in the prediction probability for random selected 2,000 sequences after the two motifs were simultaneously nullified, defined as:

$$\Delta P_{ij} = \alpha \times \Delta P_i + \beta \times \Delta P_j + \gamma \times \Delta P_i P_j, \quad 4\text{-}8$$

where ΔP_{ij} is the sum square of changes in the prediction probability after simultaneously nullifying motifs i and j, ΔP_i and ΔP_j are the sum square of changes in the prediction probabilities after nullifying motifs i and j, respectively, and α, β and γ are constants. Clearly the absolute value of γ reflects the intensity of the interaction, while its sign $(+/-)$ indicates a positive or negative

interaction. Therefore, we call γ the interaction coefficient and used it to quantify the interaction between two motifs.

4.3. Results

4.3.1. The cell type CNN models achieve highly accurate and robust performance

In the genome of a cell type, different loci are modified by the same and/or different chromatin marks in unique ways. It is the different combinations of these chromatin marks that determine the distinct chromatin states of the genomes in different cell types [227]. To learn the sequence determinants that govern the unique combinations of histone modifications in a cell type, we constructed a CNN model for the cell type for predicting the histone marks in its genome. We first evaluated the model using the data set of six histone marks collected from the four human CD_4^+ T cell types derived during T-cell differentiation [209]. Specifically, we used 459,814, 653,272, 978,543 and 2,131,540 histone modification peaks in building the models for the Tn, Tcm, Tem and Temra cells, respectively (Methods and materials). As shown in FIGURE 4-2A-D, all the models perform very well for predicting the patterns of the six histone marks in each of the four cell types, with an average accuracy and AUC (area under the receiver operating characteristic (ROC) curve) of 91.53% and 0.916, respectively, which are better than the results achieved by the earlier state-of-the-art CNN models for the same marks although their results were based on a different dataset [206].

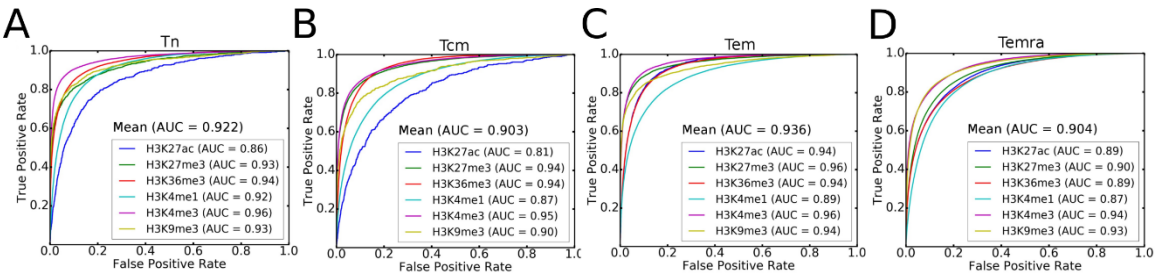

FIGURE 4-2: The ROCs of the Tn, Tcm, Tem and Temra cell models for predicting the six histone marks.

To evaluate the generality and robustness of our model, we applied it to a dataset for the six histone marks collected from the H1 human embryonic stem (H1) cells, trophoblast-like (TBL) cells, mesendoderm (ME) cells, mesenchymal (MSC) cells and neural progenitor (NPC) cells [210], which has been studied intensively [36]. In this case, we used 1,038,201, 363,349, 880,462,

1,011,252, and 315,266 histone modification peaks in building the models for the H1, TBL, ME, MSC and NPC cells. As show in FIGURE 4-3A-E, the models also perform very well for predicting the patterns of the six histone marks in the five cell types, with an average accuracy 90.6% and AUC 0.917, which are comparable to those obtained in the CD_4^+ cell dataset (91.53% and 0.916), but also are better than the results achieved by the earlier state-of-the-art CNN models for the same markers albeit on a different dataset (AUC 0.856) [206]. Our models (average accuracy 90.6% and AUC 0.917) also outperform the earlier random forest-based algorithm on the same dataset (average accuracy 79.0%, average AUC 0.837, FIGURE 4-4). The relative performance of our models on predicting the six marks also is consistent with the random forest-based method (FIGURE 4-4) except for H3K9me3, which holds the second place in our model while it was ranked fifth in the earlier study. Such consistent performance of the different methods in different datasets strongly suggests that the active enhancer marks H3K27ac (AUC 0.880) and H3K4me1 are more complicatedly used in the cell types than the other marks. Therefore, our cell type CNN models are very robust and highly accurate for predicting unique patterns of various histone marks in given different cell types.

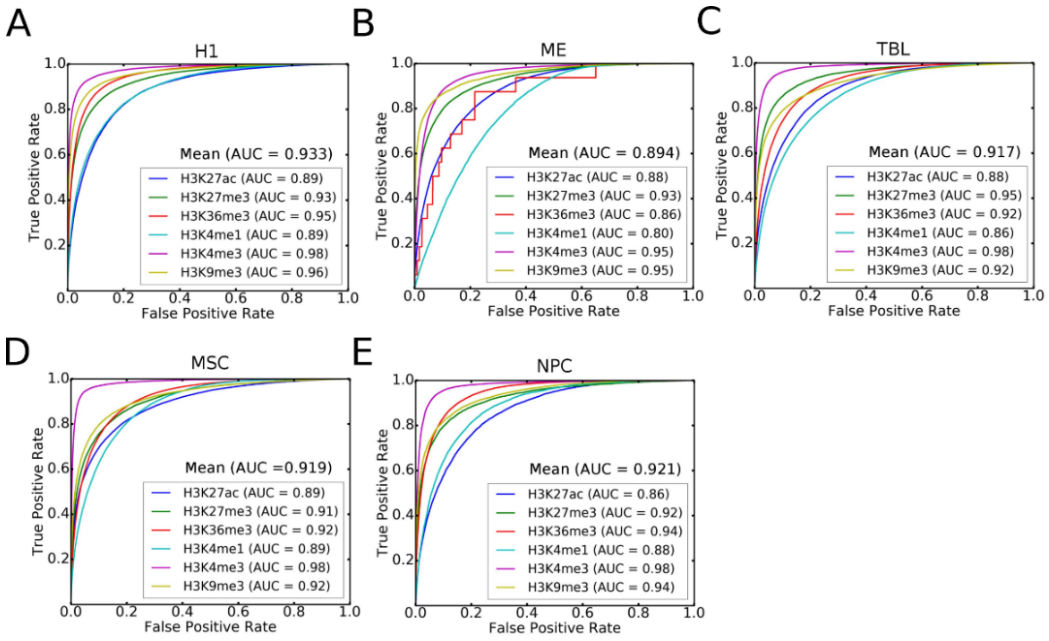

FIGURE 4-3: Performance of the CNN models of the five cell types for predicting the six histone marks. A-E. The ROCs of the H1, ME, TBL, MSE and NPC models for predicting the six histone marks.

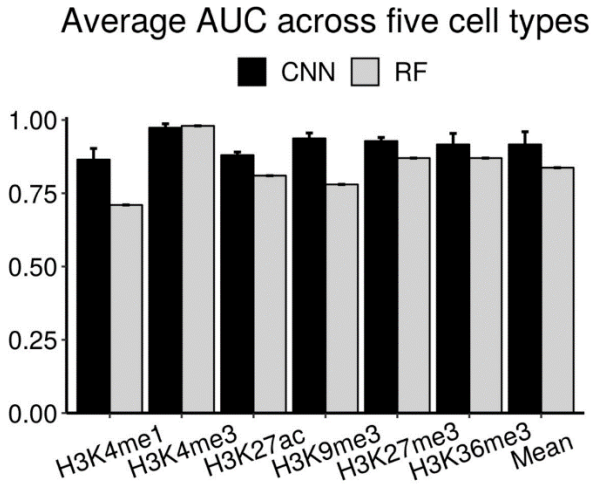

FIGURE 4-4: Comparision between CNN and Random forest. Average AUCs achieved by our CNN models and those obtained by the random forest-based models for the marks across the five cell type models. The error bars for the random forest-based models are not shown due to their unavailability.

4.3.2. The histone mark CNN models are highly accurate and robust

To reveal the determinants that specify different patterns of the same histone mark in different cell types, we constructed a CNN model for each histone mark for predicting different cell types

based on the different patterns of the same mark. We also first evaluated the accuracy of the models using the dataset collected from the four CD_4^+ T cell types [209], and employed 227,420, 691,032, 839,057, 867,398, 296,079 and 435,351 histone peaks in building the models for H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3 and H3K9me3, respectively (Methods and materials). As shown in FIGURE 4-5A-F and FIGURE 4-6, the models generally perform very well for predicting each cell type, although the models for the gene repression-related mark H3K27me3 (AUC 0.95) and the heterochromatin-related mark H3K9me3 (AUC 0.93) perform better than the models for the activation-related marks H3K36me3 (AUC 0.87), H3K27ac (AUC 0.85), H3K4me3 (AUC 0.83) and H3K4me1 (AUC 0.71).

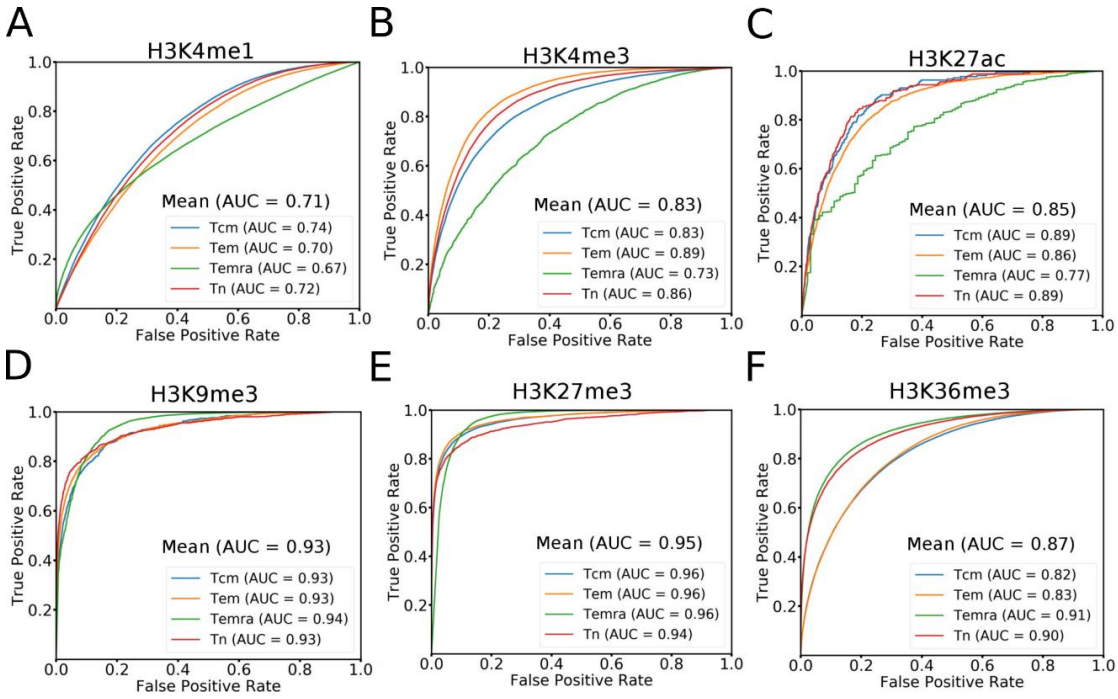

FIGURE 4-5: Performance of the CNN models of the six histone marks for predicting the four cell types. A-F, ROCs of the H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K27me3 and H3K36me3 models for predicting the four cell types.

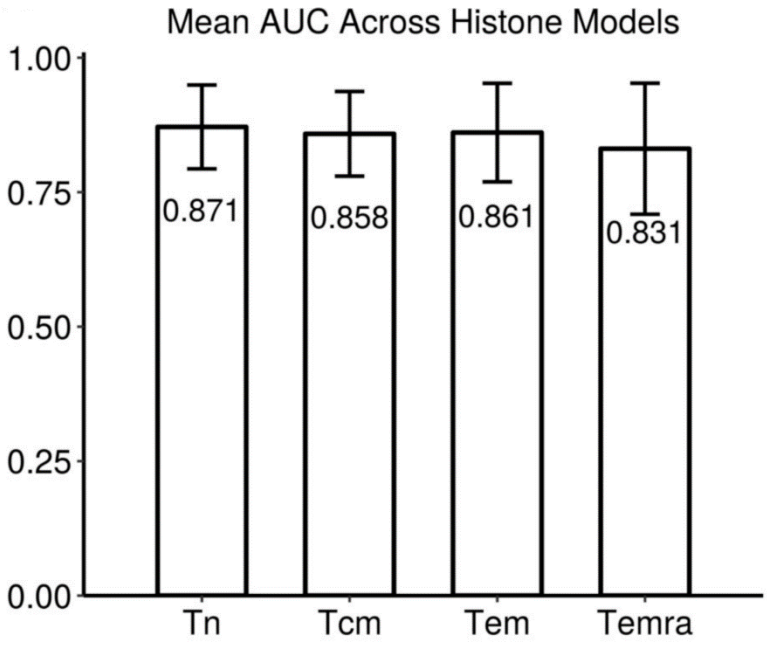

FIGURE 4-6: Mean AUC for each cell type model across the six histone mark models.

To evaluate the generality and robustness of the mark model, we also applied it to the dataset collected from the human embryonic cells H1 and four of its derived types [210], and used 332,704, 458,844, 952,615, 185,182, 253,289, 360,040 histone modification peaks in building the models for H3K4me1, H3K9me3, H3K36me3, H3K4me3, H3K27me3 and H3K27ac, respectively (Methods and materials). As shown in FIGURE 4-7 and FIGURE 4-8, similar to the results from the CD_4^+ T cell dataset (FIGURE 4-7A-F and FIGURE 4-8), the models also generally perform very well, although the models for the gene repression-related mark H3K27me3(AUC 0.909) and the heterochromatin-related mark H3K9me3(AUC 0.862) perform better than the models for the activation-related H3K4me3 (AUC 0.815), H3K4me1 (AUC 0.720), H3K27ac (AUC 0.770) and H3K36me3 (AUC 0.679). These consistent results from different datasets from different sources strongly suggest that the two repressive histone marks are more distinctly used in different cell types than the four activation-related marks. Therefore, our histone mark CNN models are highly accurate and very robust for predicting different cell types based on the pattern of single histone marks.

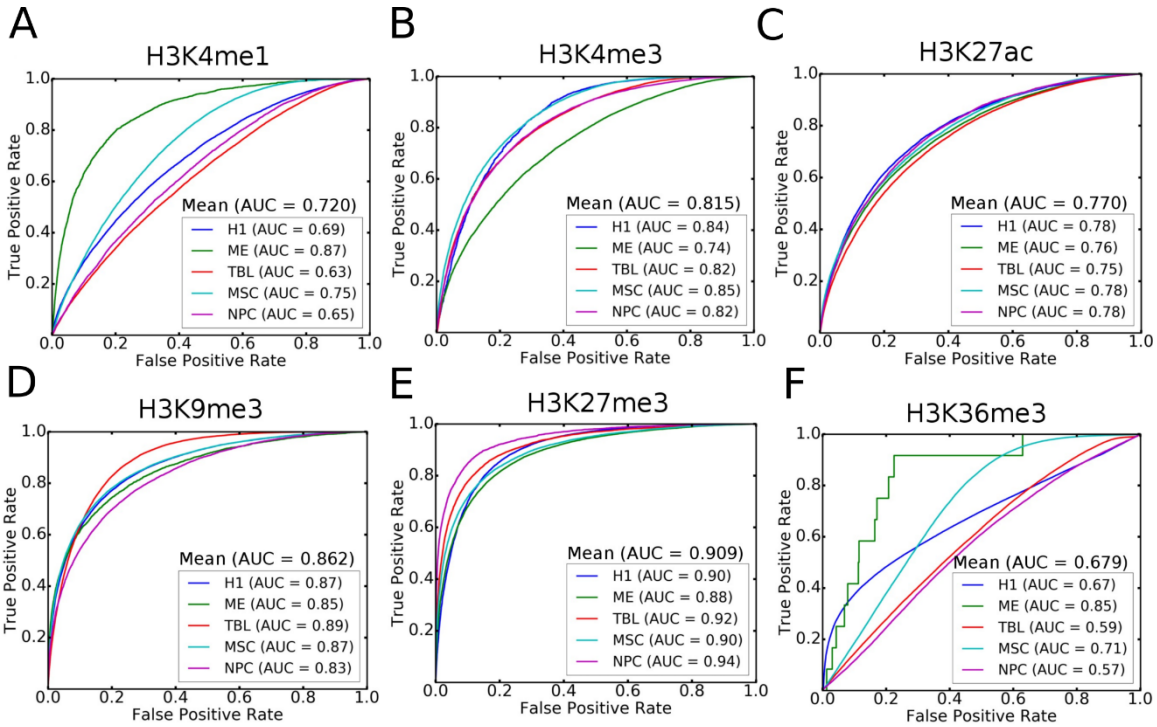

FIGURE 4-7: Performance of the CNN models of the six histone marks for predicting the five cell types. A-F, ROCs of the H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K27me3 and H3K36me3 models for predicting the five cell types.

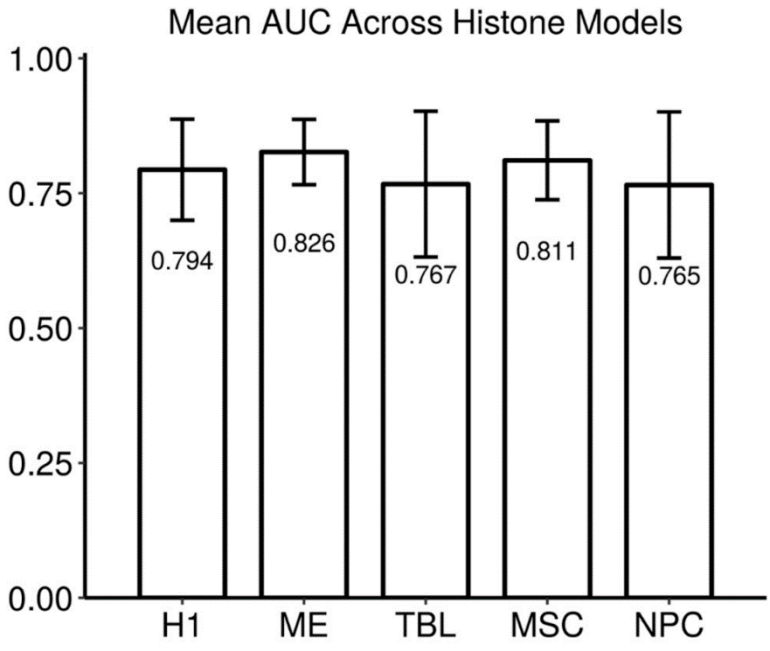

FIGURE 4-8: Mean AUC for each cell type across the six histone mark models.

4.3.3. Histone marks and cell types are largely determined by a unique set of motifs

The superior performance of our cell models indicates that the filters in the convolutional layers have largely learned the sequence determinants for specifying the patterns of various histone marks in the cell type; while the superior performance of our histone mark models suggest that the filters in the convolutional layers have largely learned the sequence determinants for governing different patterns of the same histone mark in different cell types. These results promoted us to reveal these sequence determinants by looking into the filters in the convolutional layers of the models. In particular, we expect that the filters in the first convolutional layer may have learned the binding motifs of TFs involved in the specification of different histone modification patterns in different cell types. In other words, these filters may correspond to position weight matrices (PWMs) of the TF binding motifs. To this end, we constructed motif models for all the filters learned in the first constitutional layers, resulting in 295, 295, 278 and 285 motifs in the Tn, Tcm, Tem and Temra cell models, respectively; and 280, 291, 271, 270, 293, 267 motifs for the H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3, H3K9me3 mark models, respectively. Some of the motifs learned in different models are highly similar to each other, thus we clustered them according to their similarity (Methods and materials), resulting in 2,474 clusters. Of these clusters, 203 are formed by more than two learned motifs, and we call each of them a Merged Motif (M-Motif), while the remaining 2,271 are singleton motifs, and we consider each of them as a cell- or mark-specific motif dependent on the type of the model by which it is learned. Interestingly, 113 (4.57%) of these 2,474 unique motifs are shared by at least a cell type model and a histone mark model, indicating that common sequence determinants were captured by the two types of models. On the other hand, the remaining 958 and 1,403 motifs are unique to the cell type models and histone mark models, respectively (FIGURE 4-9). Thus, besides the common motifs, both the cell type models and mark models captured quite different sets of motifs for predicting the patterns of different histone modifications in the cells and the cell types based on single histone marks, respectively. Furthermore, 42 (3.92%) of the 1,071 motifs learned in the cell type models and 68 (4.49%) of the

1,516 motifs learned in the histone mark models are shared by more than two cell models (42/1,071=3.92%) and two mark models (68/1,516=4.49%) (FIGURE 4-10A and FIGURE 4-11A), respectively. However, only two (0.21%) and one (0.10%) motifs are shared by all the four cell type models and all the six histone mark models, respectively. The remaining 1,029 (96.08%) and 1,448 (95.51%) motifs are unique to a single cell type model and a single mark model, respectively. These results suggest that the unique patterns of various histone marks in each cell type as well as the different patterns of the same histone mark in different cell types are largely determined by a unique set of motifs, although they may share some common ones. This conclusion agrees with the general understanding about how the unique epigenomes are established in different cells type by the interplay of TF, chromatin remodeling systems and environment cues [198-201].

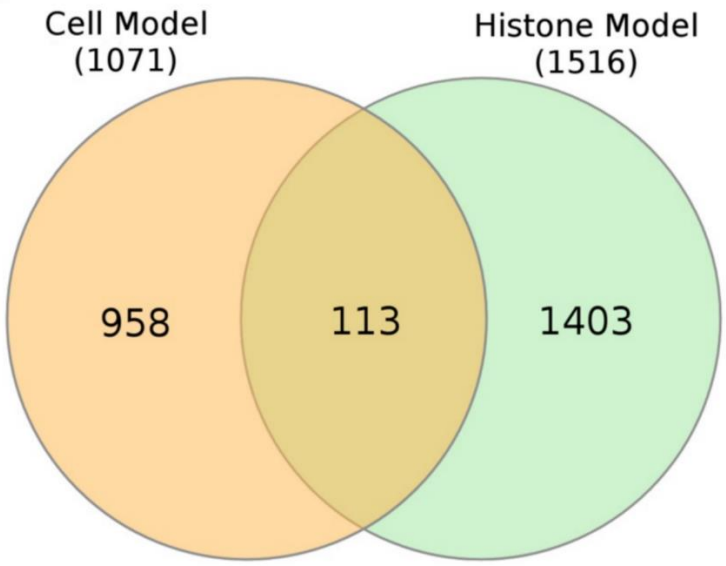

FIGURE 4-9: Overlap of motifs learned in the cell models and histone mark models.

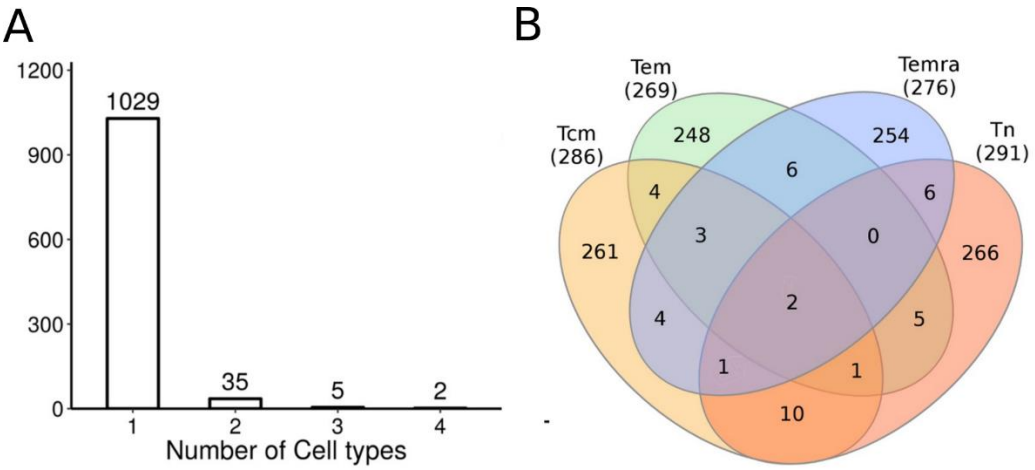

FIGURE 4-10: Shared learned mottifs in different cell types. A. Number of learned motifs shared by different number of cell models. B. Venn diagram showing the number of learned motifs shared by the cell models.

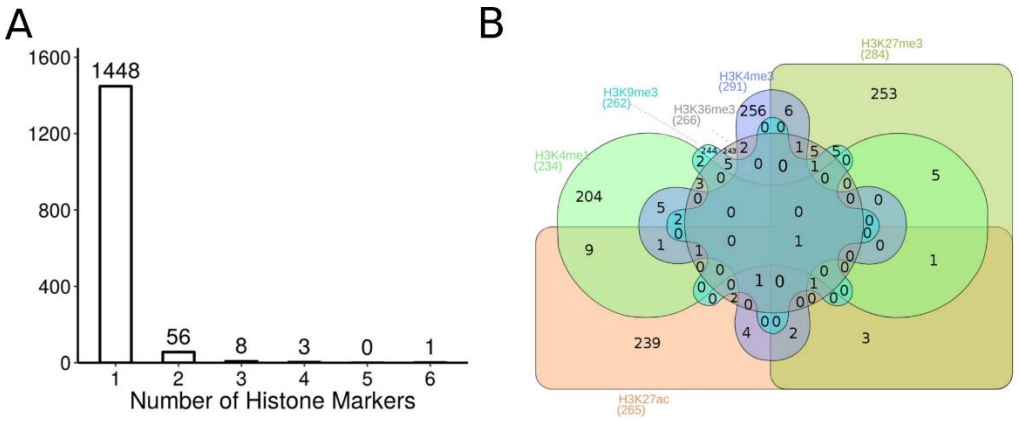

FIGURE 4-11: Shared learned mottifs in different histone marks. A. Number of learned motifs shared by different number of histone mark models. B. Venn diagram showing the number of learned motifs shared by the histone mark models.

At an E-value threshold of 0.5, 974 (39.37%) of the 2,474 motifs match known human TF binding motifs in the HOCOMOCO database [142], and many of them are known to be involved in T cell differentiation (FIGURE 4-12). We described a few examples of them. M-Motif 12 shared by all the cell type models matches that of ETS1 that controls T cell differentiation by regulating the expression of signaling molecules [228, 229] in response to external environment stimuli. M-Motif 67 shared by the H3K9me3 and Tem models matches that of ATF2 that is an histone acetyltransferase for histones H2B and H4, playing an essential role in the T cells activation in late-stage [230, 231]. Temra-Motif 117 learned in the Temra model matches that of RUNX3, which

plays a crucial role in T cell's differentiation by interacting with master regulators cooperatively[232]. M-Motif 178 shared by the Tn and H3K4me1 models resembles that of SMAD4 that cooperatively regulates interleukin 2 receptor in T cells and balances the differentiation of CD_4^+ T cells [233, 234]. H3K27ac-Motif 229 learned in the H3K27ac model matches that of ZN274 that is involved in transcription repression [235]. H3K27me3-Motif 127 learned in the H3K27me3 model resembles that of FOXP1, which is the "naive keeper" for T memory cell differentiation [209, 236]. These results suggest that at least 39.37% of the learned motifs that match known ones are likely to be authentic motifs of the cognate TFs.

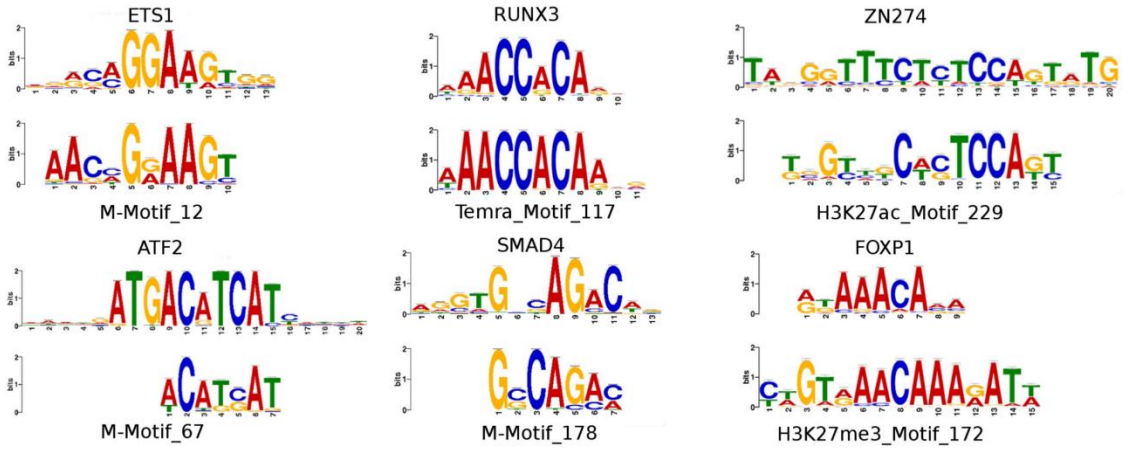

FIGURE 4-12: Examples of learned motifs matching known motifs involved in T cell functions.

4.3.4. Motifs learned in the cell type models reflect the lineage of the cells

It is now well established that along the lineage of cell differentiation, the epigenomes of cells undergo step-wise changes with each cell division through the regulation of a specific set of both common and unique TFs in the derived intermediate and terminal cell types [198-201, 237, 238]. Cells in adjacent differentiation stages possess more similar epigenomes [205], presumably because they share more TFs than those that are distal from each other along the lineage of differentiation. To see whether this is reflected in the motifs learned by the cell type models, we hierarchically

clustered the cell types based on the similarity of the learned motif profiles in the cell type models. As shown in FIGURE 4-13, Tn branches earliest in the tree while the three memory/effector T cell types form a clade, indicating that Tn is most distinct from the more developed cell types as generally believed. Tem and Temra form a clade, indicating that they are more similar to each other than to Tcm, which is in agreement with early observations[239]. These results suggest a linear lineage model of the development of these cells: Tn → Tcm → Tem → Temra, which is in line with the results derived based on changes in the DNA methylation, gene expression and DNAase accessibility in these cells [209]. Therefore, the sequence motifs learned in the cell type models indeed reflect the lineage relationships of the cells. It is highly likely that the unique motifs to a cell model account for the distinction of the cell type from the other cell types, while the shared motifs are responsible for the shared features of linearly closely-related cell types.

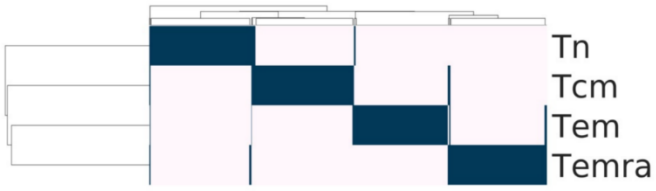

FIGURE 4-13: Two way clustering of the cells and learned motfis. Hierarchical two-way clustering of the cells, based on the similarity of the learned motifs profiles in the models using hamming distance and average linkage.

4.3.5. Motifs learned in histone mark models reflect functional relationships

It is well-known that certain types of sequences can be co-modified by different histone marks, while other types of sequences tend to be exclusively modified by a specific mark [240]. To see whether such co-modifications and exclusiveness of the marks are reflected by the learned motifs in the histone mark models, we hierarchically clustered the histone marks based on the similarity of the learned motif profiles. As shown in FIGURE 4-14, H3K4me1 and H3K27ac form a group, which is consistent with the fact that they co-mark active enhancers, thus the respective modification systems might be recruited by some common motifs or similar mechanisms. On the

other hand, H3K9me3, H3K27me3, K3K36m3 and H3K4me3 form a singleton group by themselves, which is consistent with the facts that they exclusively mark DNA domains with different epigenomic states [241]. For instance, H3K9me3 marks heterochromatins, H3K27me3 labels polycomb-associated domains, K3K36m3 marks transcribed gene body and H3K4me3 labels active promoters. Therefore, the learned motifs in the histone mark models indeed reflect the known functional relationships of the marks.

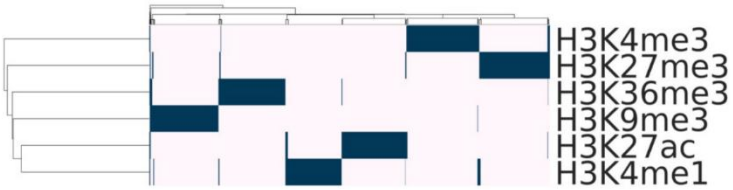

FIGURE 4-14: Two way clustering of the histone marks and learned motfis. Hierarchical two-way clustering of the histone marks, based on the similarity of the learned motifs profiles in the models using hamming distance and average linkage.

4.3.6. The learned motifs have varying inferences on the prediction

To evaluate the contribution and importance of a learned motif to the prediction of a model, we nullified the motif and then calculated its inference score on the predictions (Methods and materials). The inference scores of the motifs learned in both the cell type models (FIGURE 4-15A-D) and the histone mark models (FIGURE 4-16A-F) have bell-shape distributions with different extent of right skewness. These results suggest that most learned motifs have intermediate inferences, while a small portion have large inferences on predicting the patterns of different histone marks in a cell type or different cell types based on single histone marks. The motifs with high influences might play crucial roles in the cell differentiation process. For example, in the Tn model, the motif with the highest influence score 4.26 (FIGURE 4-15A-D) resembles that of FOXD1 that is involved in T cell proliferation [242]; in the H3K4me1 model, the motif with the highest inference score 2.74(FIGURE 4-16A-F) resembles that of SP1 that plays a role in T cell differentiation [243]. The inferences of the motifs learned in either the cell type models (FIGURE

4-15A-D) or the histone mark models (FIGURE 4-16A-F) do not significantly correlate with their information contents, suggesting that only few positions of the motifs have a strong predictive power, which is consistent with the general understanding about the mechanisms of TF-DNA interactions. The learned motifs that do not match known motifs have similar inference scores to those matching known motifs (FIGURE 4-15A-D and FIGURE 4-16A-F), indicating that they are equally likely to be true motifs, and the unmatched ones are likely to be novel motifs of unknown TFs.

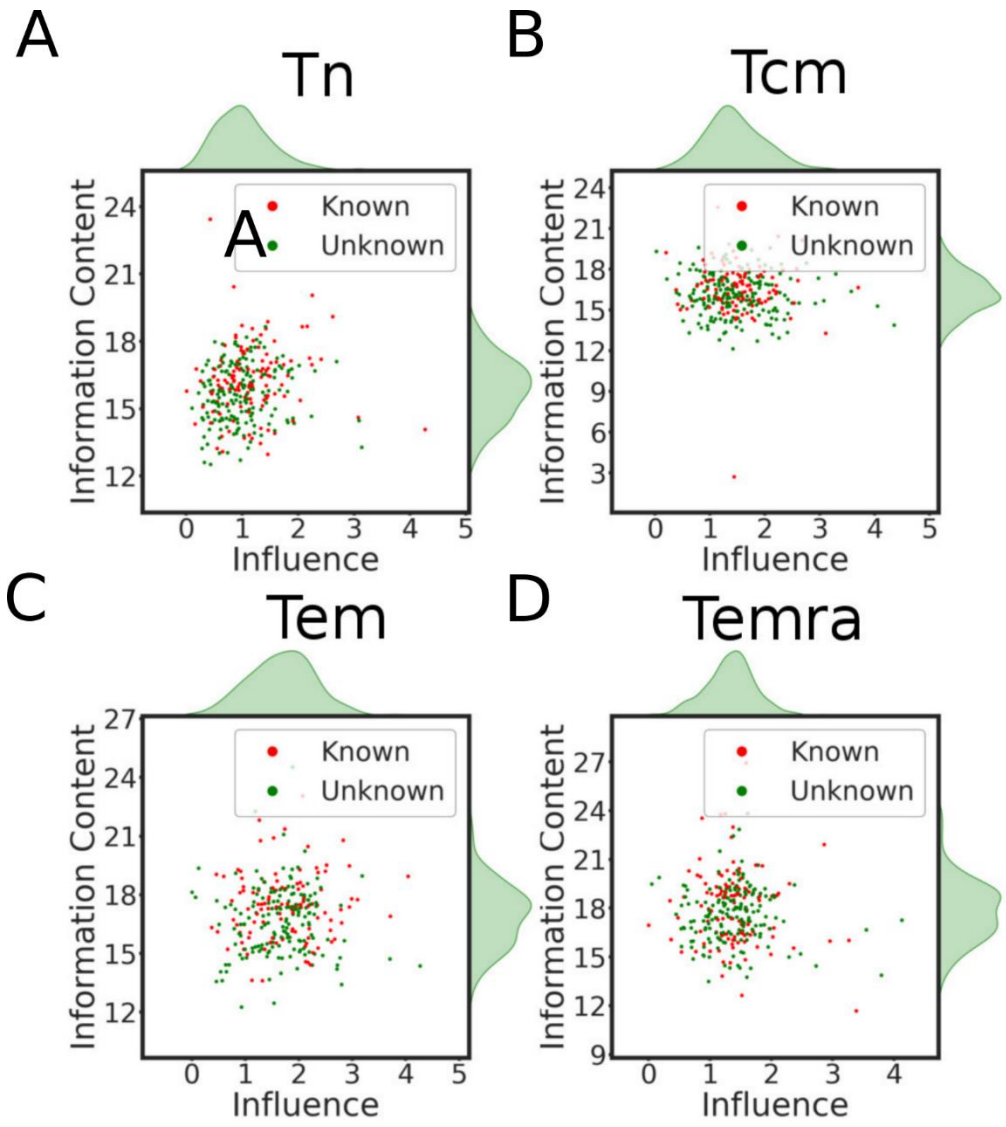

FIGURE 4-15: Relationship between IC and influence.in cells. A-D. Relationship between the inference scores and information contents of motifs learned in the four cell models.

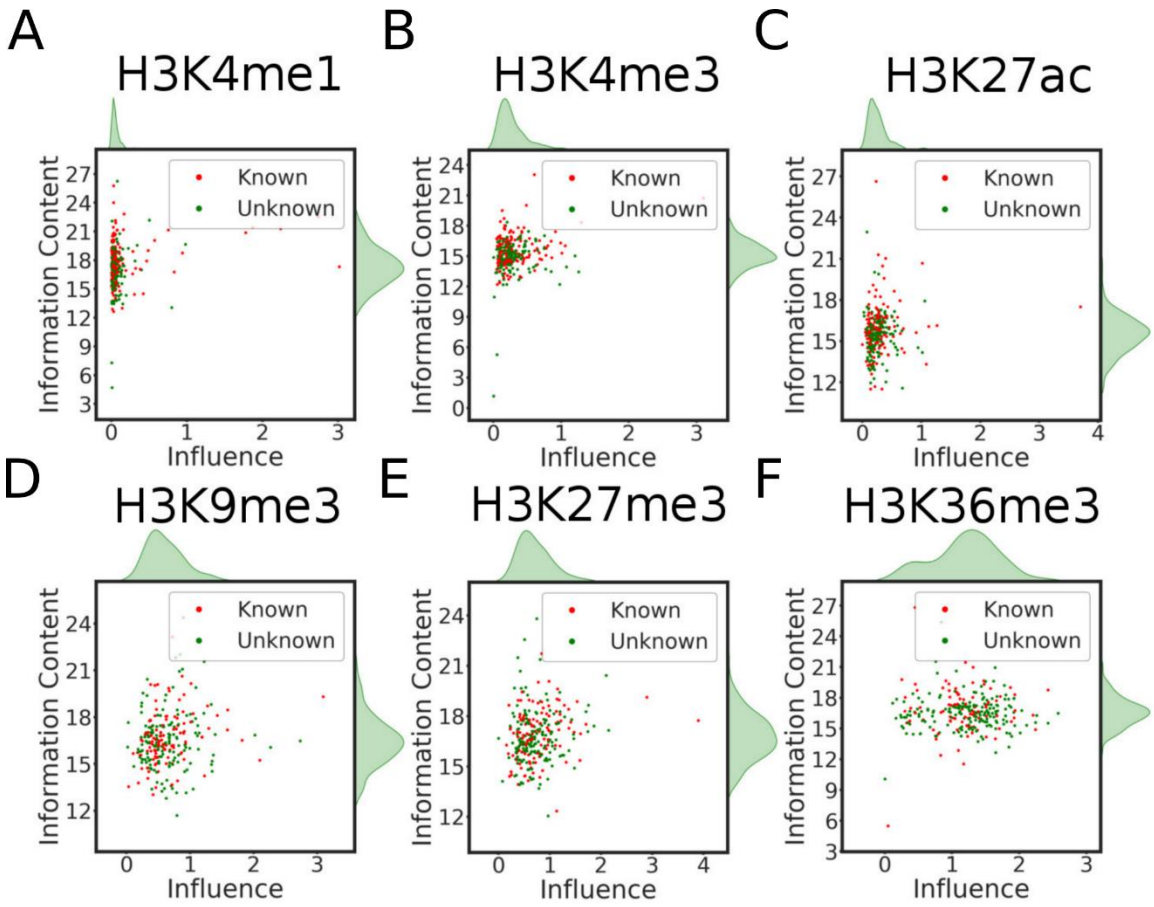

FIGURE 4-16: Relationship between IC and influence.in hisone marks. A-F. Relationship between the inference scores and information contents of motifs learned in the six histone mark models, respectively.

Interestingly, the inferences of motifs learned in Tn, Tcm, Tem cell models increased along the proposed linear cell lineage, and then decreased in the Temra cell model (FIGURE 4-17A). These results suggest that the functions of learned motifs become more and more specific in determining the patterns of various histone modifications in the cells along the differentiation lineage Tn → Tcm → Tem, and then somehow become less specific in Temra. Furthermore, the inference scores of motifs learned in the six histone mark models are also significantly different from one another (FIGURE 4-17B). Specifically, motifs learned in the models of H3K4me1, H3K27ac and H3K4me3 that mark active enhancers and promoters have the lowest inference scores, while those learned in the models of H3K9me3 and H3K27me3 that are associated with repression regions have the moderate inference scores, and those learned in the model of H3K36me3 that marks actively

transcribed regions have the highest inference scores (FIGURE 4-16A-F). These results suggest that the motifs specifying histone modifications in actively transcribed regions have the highest specificity, followed by those for determining histone modifications in repression regions, active promoters and enhancers regions.

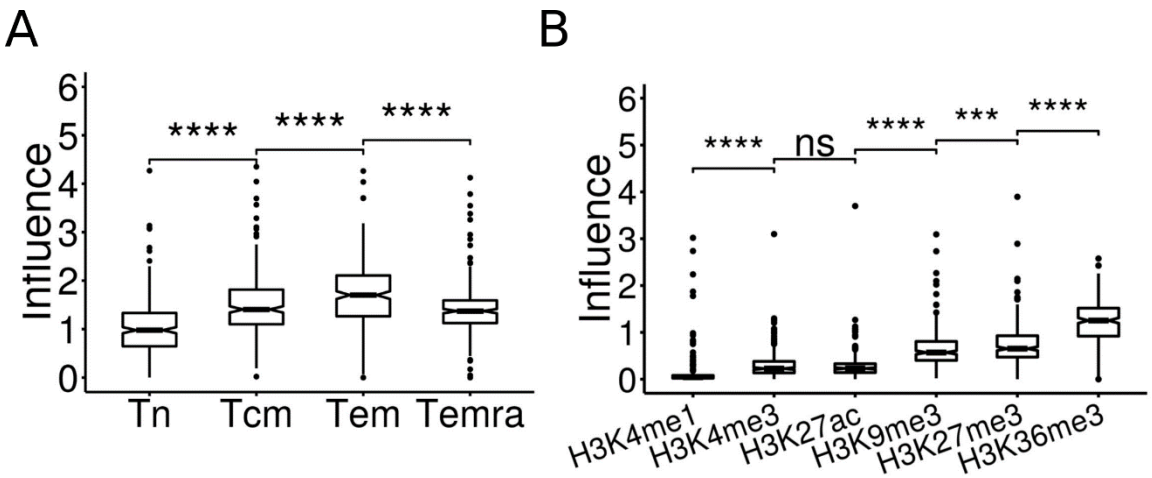

FIGURE 4-17: Influence of the motifs in cell models and histone mark models. Boxplots of the inference scores of the motifs learned in the cell models and histone mark models, respectively (***, p<0.001; ****p<0.0001; Wilcoxon test).

4.3.7. The motifs have highly variable inferences on different histone marks

An important question in epigenomics study is to understand how different histone marks are placed at specific domains of the genome in a cell type. Our cell models might provide an easy way to address this question by simply finding out the learned motifs that impose a high inference on the prediction of each histone mark in the models. More specifically, we computed an inference score of each learned motif on each histone mark in a cell type model. Shown in FIGURE 4-18 are the results for the learned motifs that are ranked top 100 for their inferences on predicting at least one histone mark in the cell type models. Clearly, the motifs learned in each cell type model have highly variable inferences on different histone marks. For example, in all the four cell type models, H3K36me3 and H3K27me3 are highly impacted by a large number of the learned motifs, while H3K4me3 is only highly impacted by a few learned motifs, such as Tn-26:FOXD1, Tn-106:HXB4,

TN-21 and Tn- 294 in Tn (FIGURE 4-18). H3K27ac is highly impacted by a large number of learned motifs in Tcm, but is highly impacted by only a few learned motifs in Tn, Tem and Temra. H3K4me1 is highly impacted by a larger number of learned motifs in Tcm, Tem and Temra, but is highly impacted by a few learned motifs in Tn. H3K9me3 is highly impacted by an intermediate number of learned motifs in all the four cell types. Moreover, in all the four cell models, only a few learned motifs have high inferences on all the histone marks, while most motifs have a high inference only on 1-3 histone marks (FIGURE 4-18). For instance, in Tn model, only motifs Tn-26:FOXD1, Tn-106:HXB4, Tn-21 and Tn-294 have high inferences on all the six histone marks, while most of other motifs have high inferences only on one or two histone marks. Thus, each histone mark is impacted by a unique combination of motifs that may have inferences on more than two histone marks. These results suggest that the cognate TFs of most learned motifs exerting more specific inferences on one or two histone marks might play crucial roles in specifying the unique patterns of different histone marks in the cell type, while the cognate TFs of a few learned motifs having high inferences on multiple histone marks might be involved in the establishment of multiple histone marks, probably by playing roles in the common mechanisms of different histone modifications such as opening up of DNA domains.

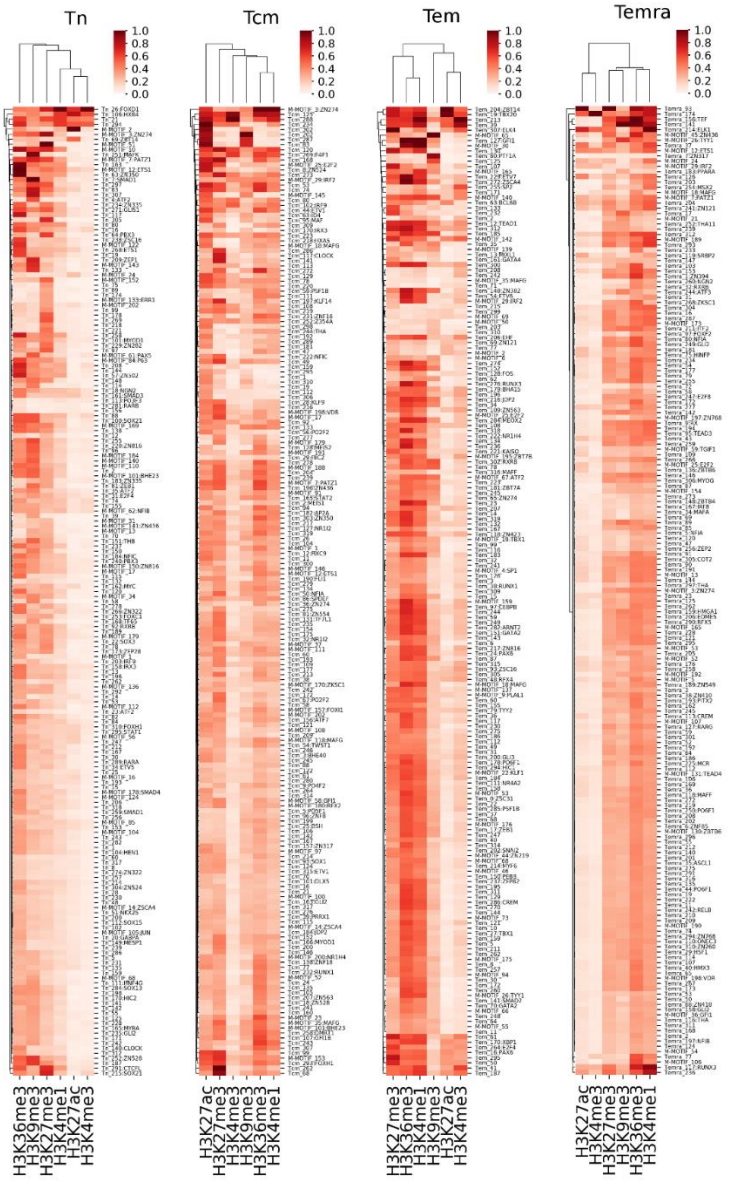

FIGURE 4-18: Influence of the top 100 learned motifs in cell type models. The heatmaps show the influence scores of the top 100 learned motifs on predicting the six histone marks in the indicated cell type models. The scale bar shows range of the inference score of a motif on a histone mark.

4.3.8. The motifs have highly variable inferences on different cell types

Another important question in epigenomics study is to understand how the same histone mark is differentially placed in the genomes of different cell types. Our histone mark models might provide a convenient way to tackle this question by simply identifying the learned motifs that

impose a high inference on the prediction of each cell type by the models. More specifically, we calculated an inference score of each learned motif on the prediction of each cell type by a histone mark model. Shown in FIGURE 4-19 is the result of the motifs that are ranked top 100 for their inferences on predicting at least one cell type by the six histone mark models. Interestingly, motifs learned in each histone mark model have highly variable inferences on different cell types. For instance, in the H3K4me1 model, most of the learned motifs have similarly small inferences on all the four cell types, only few have high inferences on at least one cell type. However, the latter set of motifs exert high inferences only on one or two cell types with the exception that motif H3K4me1-236:HXC10 has high inferences on all the four cell types. Thus, it seems that H3K4me1 in each cell type is specified by a small set of motifs with unique combinations. In both the H3K4me3 and H3K27ac models, most of the learned motif have similarly small inferences on the Tem, Tcm and Tn cell types, only few have high inferences on at least one of these three cells types, suggesting that these two histone marks are specified by a small set of motifs with unique combination in these three cell types. However, most of the motifs learned in the H3K4me1 and H3K27ac models impose high inferences on the Temra cells, suggesting that these cells might have more complex H3K4me3 and H3K27ac modifications than the other three cell types, which is in line with the fact that Temra is the terminally differentiated cells with more activated enhancers and promoters. In the H3K9me3 and H3K27me3 models, each cell type is impacted by a large number of learned motifs with few having high inferences on more than three cell types, suggesting these two histone modifications in each cell type are specified by a large set of motifs with unique combinations. This result might be related to the functions of H3K9me3 that marks heterochromatins and of H3K27me3 that labels polycomb-associated domains. In the H3K36me3 model, the numbers of learned motifs having high inferences in the cells increase along their linear lineage: Tn → Tcm → Tem → Temra. Each cell type is highly impacted by a large number of the learned motifs that impact adjacent cells along the lineage. These results reflect the similarity of the transcriptomes of these adjacent cell types [36], and thus are in excellent agreement with the

functions of H3K36me3 that marks actively transcribed genes. Taken together, the cognate TFs of few learned motifs that exert high inferences on multiple cell types might account for the similar patterns of a histone mark and the common mechanisms of the histone modification in different cell types, while the cognate TFs of the motifs that have more specific inferences might play crucial roles in specifying the different patterns of the histone modification in different cell types.

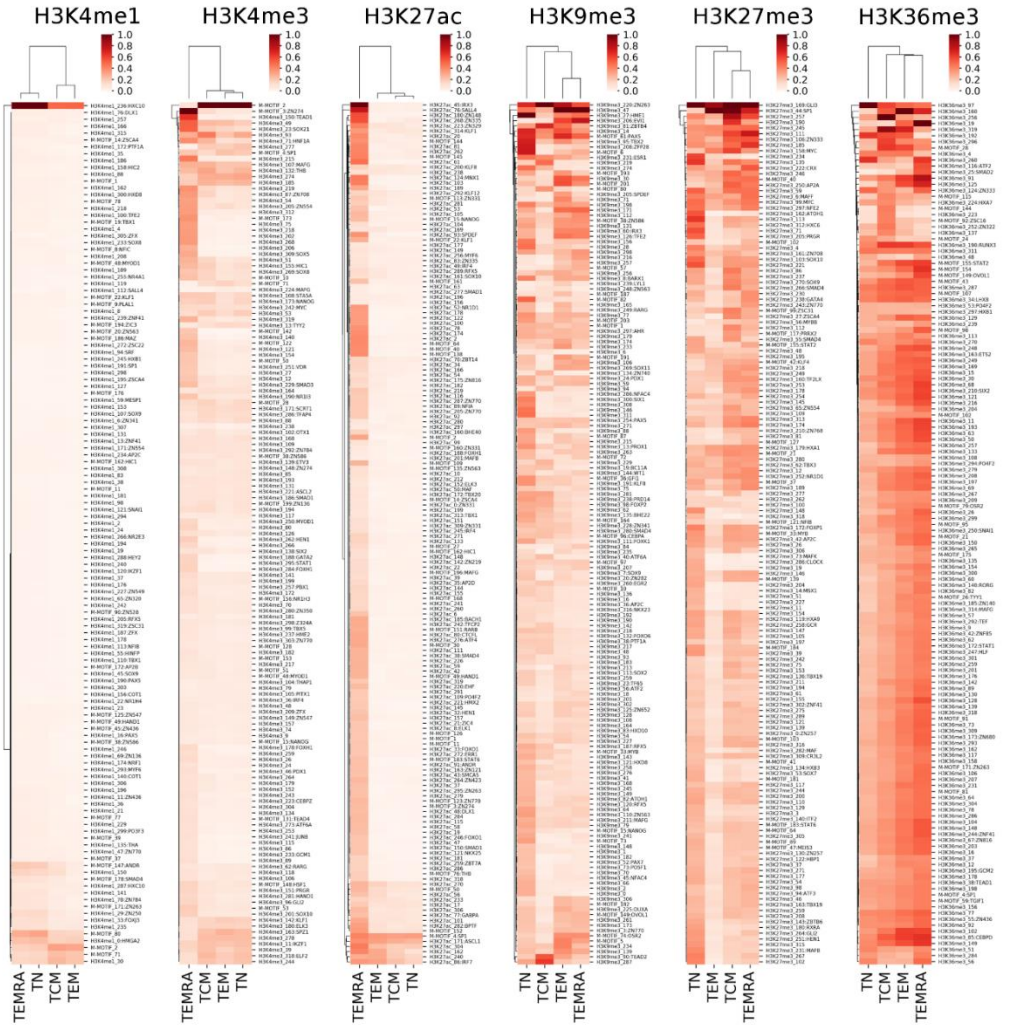

FIGURE 4-19: Influencec of the learned motifs on the prediction of each cell type by the histone mark models. In each histone model, we show the influences for four cell types of top 100 learned motifs with highest impacts. Dark red corresponds strong influences and light red corresponds weak influences.

4.3.9. Conserved learned motifs tend to have higher inferences on the predictions

We also examined the relationships between the inference scores and the conservation levels of the motifs learned in the cell and histone mark models. As shown in FIGURE 4-20A-D, there is positive correlation between the inference scores and the conservation levels of motifs learned in all the cell models (Tn: γ=0.15, p=0.011; Tcm: γ=0.11, p=0.052; Tem: γ=0.079, p=0.19; and Temra: γ=0.17, p=0.003), though with varying levels of significance. Moreover, as shown in FIGURE 4-21A-F there is a positive correlation between the inference scores and the conservation levels of motifs learned in the models of the four activation-related histone marks H3K4me1 (r=0.43, p=2.3e-13), H3K4me3 (r=0.17, p=0.0043), H3K27ac (r=0.35, p=2.6e-9) and H3K36me3 (r=0.23, p=0.00016). However, there is negative or no significant correlation between the inference scores and the conservation levels of motifs learned in the models of the two repression-related marks H3K9me3 (r=-0.13, p=0.036) and H3K27me3 (r=0.063, p=0.29). These results indicate that more conserved motifs learned in either the cell or histone mark models generally have higher inferences on the respective predictions than less conserved ones, with the exception that rapidly evolving motifs in the H3K9me3 mark peaks (heterochromatins) tend to have higher inferences on the prediction of cell types than more conserved ones. These observations are in line with the general understanding of the evolution of DNA sequences that functionally important sequences tend to be either more conserved due to purifying selection or evolved more rapidly due to positive selection. Thus, these results further corroborate our predicted motifs.

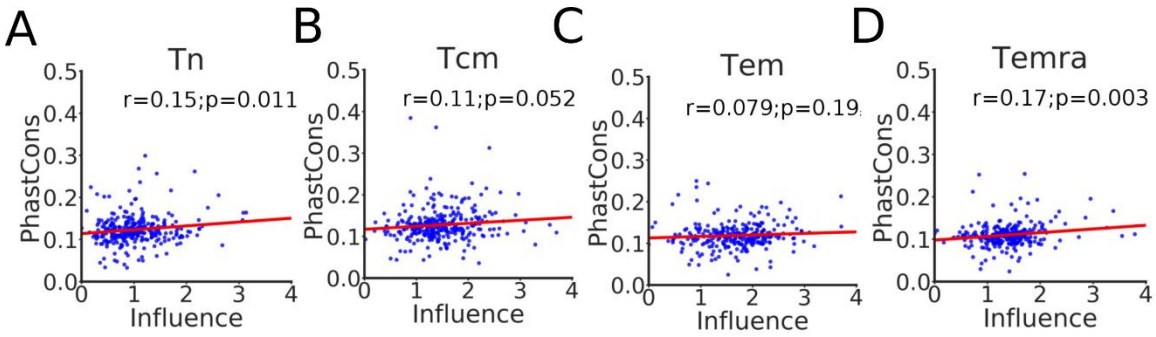

FIGURE 4-20: Relationship between the inference and PhastCons in cell models. A-D Relationship between the inference scores and PhastCons scores of the learned motifs in the cell models. The red line is the linear regression between the inference scores and PhastCons scores.

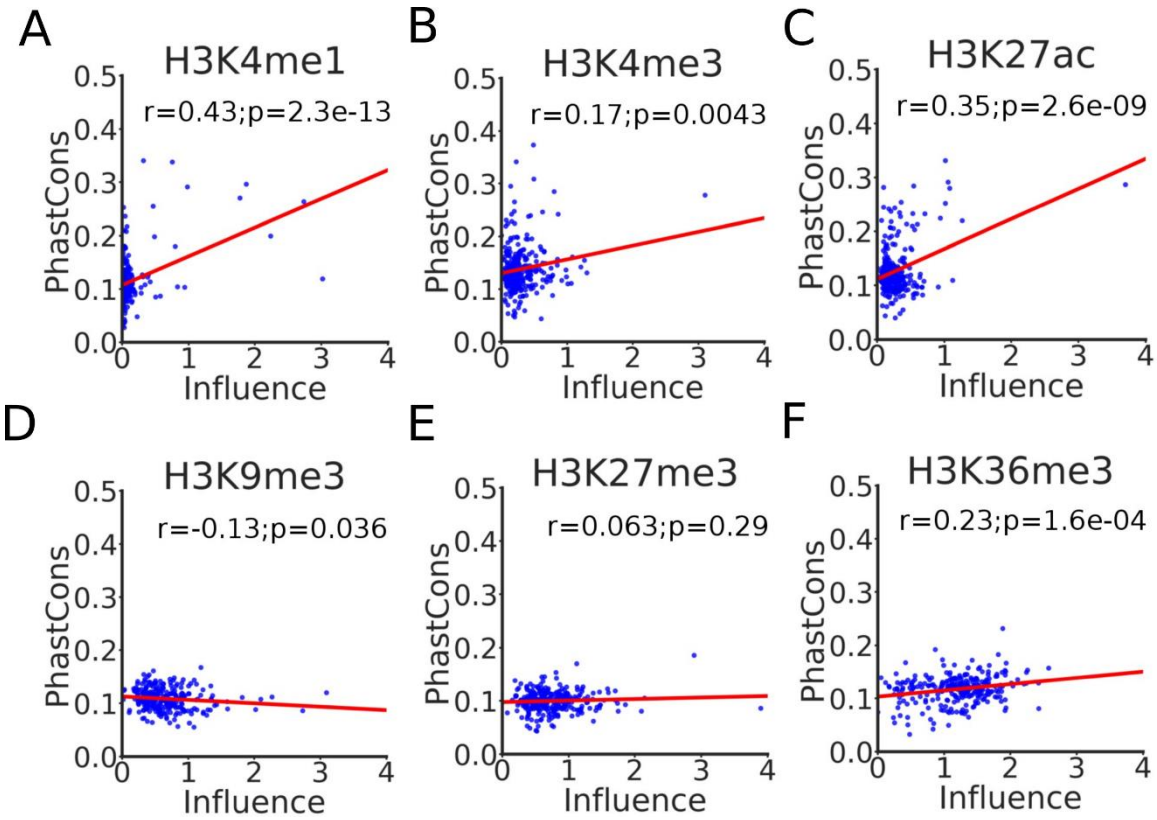

FIGURE 4-21: Relationship between the inference and PhastCons in histone models. A-F Relationship between the inference scores and PhastCons scores of the learned motifs in the histone mark models, respectively. The red line is the linear regression between the inference scores and PhastCons scores.

Interestingly, motifs learned in the Tn and Tcm models tend to be more conserved than those learned in the Tem and Temra models, and the motifs learned in the Temra model are least conserved (FIGURE 4-22A). Thus, there is a trend that the more differentiated the cells, the less conserved the motifs learned from the corresponding models, suggesting that more conserved mechanisms might be used in the cells at the earlier stages of differentiation to specify their histone modification patterns than in the cells in the later stages of differentiation. This conclusion is consistent with the general understanding about the development of animals during embryogenesis [244]. Moreover, motifs learned in the models of gene activation-related marks H3K4me3, H3K27ac, H3K4me1 and H3K36me3 are more conserved than those learned in the models of repression-related marks H3K9me3 and H3K27me3 (FIGURE 4-22B). This result suggests that more conserved mechanisms might be used to specify the patterns of the four activation-related marks than those used to govern the patterns of the two repression-related marks.

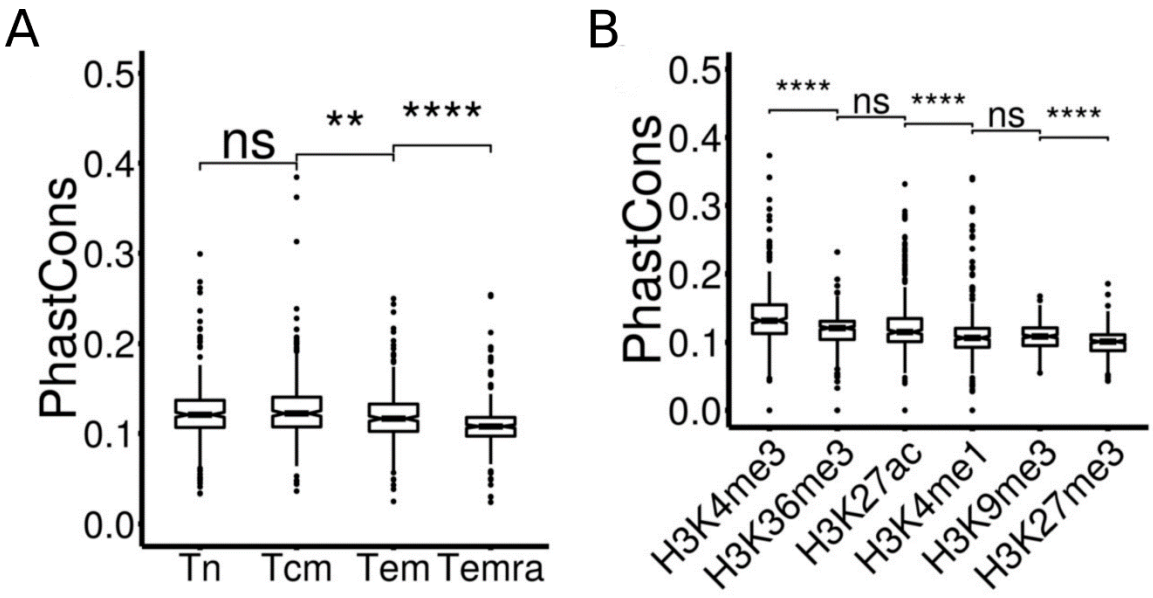

FIGURE 4-22: Distributions of the PhastCons scores in cell and histone models. A,B Boxplots of the PhastCons scores of the motifs learned in cells models and histone models, respectively (**, $p<0.01$; ****, $p<0.0001$; Wilcoxon test).

4.3.10. The CNN models can predict cooperative TFs

To see if the models can be used to identify cooperative TFs that define the histone modification patterns in the T cells, we quantified the interactions between each pair of learned motifs using a linear regression model where a positive or negative interaction coefficient indicates positive or negative interaction (Methods and materials). To reduce the computational time, we only focused on the top 50 of learned unique motifs with the highest inference scores for both the cell models and histone mark models. Shown in FIGURE 4-23A-F are the results for the Temra cell model. Clearly, there are different patterns of positive and negative interactions between the learned motifs for predicting different histone marks in the cell type. Interestingly, the motifs can be clustered into groups based on the patterns of their interactions in predicting the histone modifications. For example, in the case of predicting H3K4me1 modifications, learned motifs matching those of RUNX3, ETS1 and PATZ1 form a group with positive interactions among them; learned motifs matching those of EOMES, NFIA, ELK1, HINFP and ITF2 form a group with many putative novel motifs with largely positive interactions among them; learned motifs matching those of TEAD3, ZN121, HMGA1, ZN436, GLI1, ZN274, COT2, RX, TEF, ZN394 and TYY1 form a group with many putative novel motifs with largely negative interactions among them. Some of the predicted interactions are supported by experimental evidence. For example, we predicted ITF2 (also named T cell specific transcription factor 4 (TCF4)) had significant interactions with ETS1 for predicting histone marks H3K27ac (γ=1.27, p=3.69e-65), H3K27me3 (γ=0.18, p=0.01), H3K36me3 (γ=0.21, p=0.00077), H3K4me3 (γ=1.15, p=8.54e-57) and H3K9me3 (γ=-0.39, p=6.70e-06). In agreement with these predictions, it has been shown that ITF2 might be involved in histone acetyltransferase CBP recruitment by interacting with ETS1 [245]. Furthermore, we predicted that ITF2 had a positive interaction with RUNX3 for determining histone marks H3K27ac (γ=1.40, p=4.29e-49), H3K27me3 (γ=-0.20, p=0.013), H3K36me3 (γ=-0.59, p=6.40e-25), H3K4me1 (γ=0.32, p=8.91e-05), H3K4me3 (γ=-1.13, p=4.00e-40), and H3K9me3 (γ=-0.18, p=0.03), which is in line with the earlier finding that RUNX3 involves in regulating Wnt signaling activity by interacting with ITF2

(TCF4) in a ternary complex manner [246]. The predicted interactions between known and unknown motifs as well as between unknown motifs are likely to be novel interactions, in particular those with strong and highly significant interactions, such as the interactions for predicting the H3K27ac mark, between GLI2 and Temra 146 (γ=2.137, p=4.95e-43), between TEAD3 and Temra 54 (γ=1.97, p=4.43e-50), and between Temra 141 and Temra 146 (γ=1.99, p=4.41e-43), etc. Similar patterns of interactions were observed in the models of the other three T cell types (FIGURE 4-24 to FIGURE 4-26).

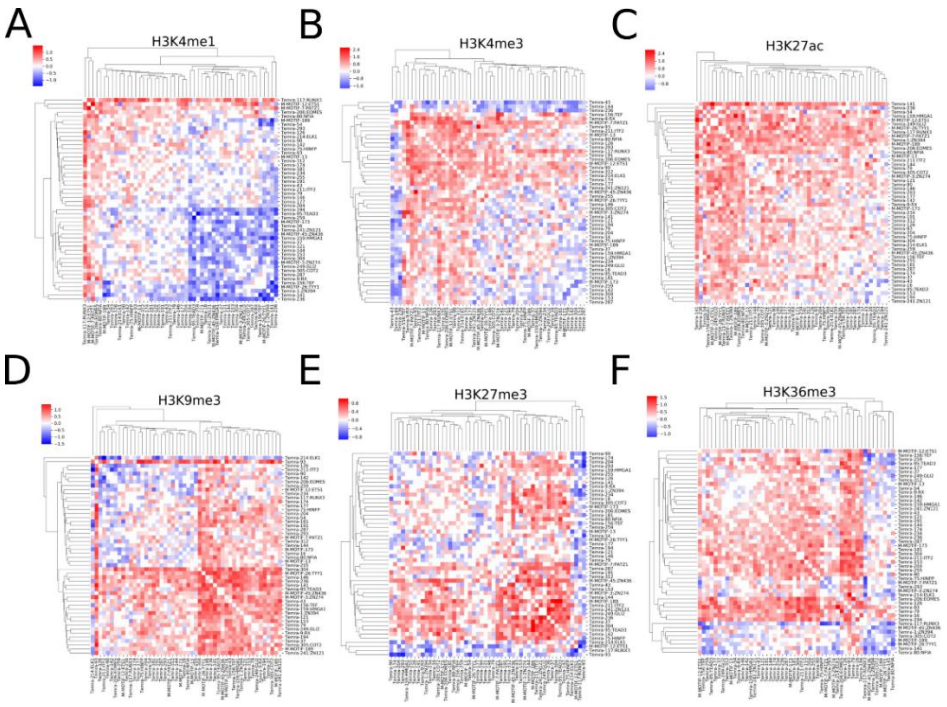

FIGURE 4-23: Interaction coefficient γ between the top 50 learned motifs in Temra cell. A-F The heatmaps show the values of interaction coefficient γ between the top 50 learned motifs on predicting the indicated histone marks in the Temra cell model. The scale bar shows the range of interaction coefficient γ. A negative value indicates a negative interaction while a positive value indicates a positive interaction between the pair of motifs.

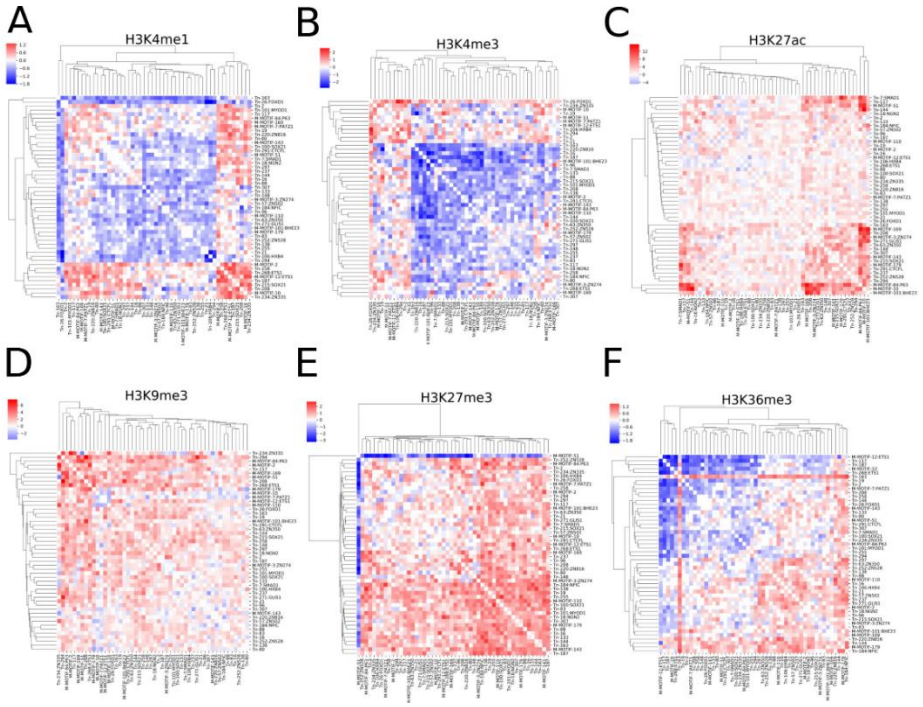

FIGURE 4-24: Interaction coefficient γ between the top 50 learned motifs in Tn cell. A-F Interactions between each pair of the top 50 learned motifs on the prediction of the six marks by the Tn cell model.

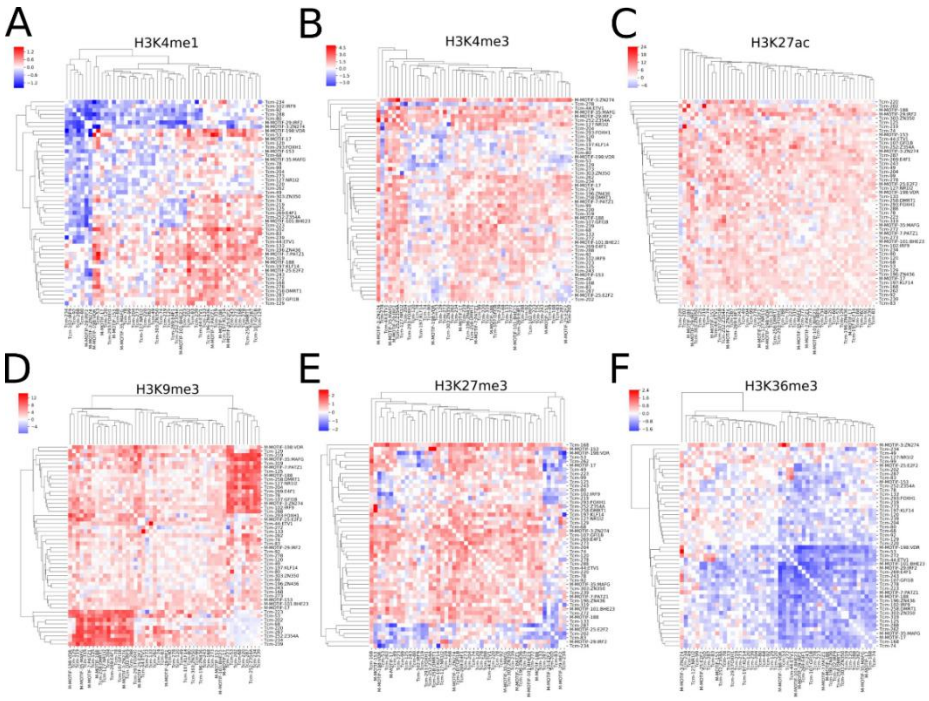

FIGURE 4-25: Interaction coefficient γ between the top 50 learned motifs in Tcm cell. A-F Interactions between each pair of the top 50 learned motifs on the prediction of the six marks by the Tcm cell model.

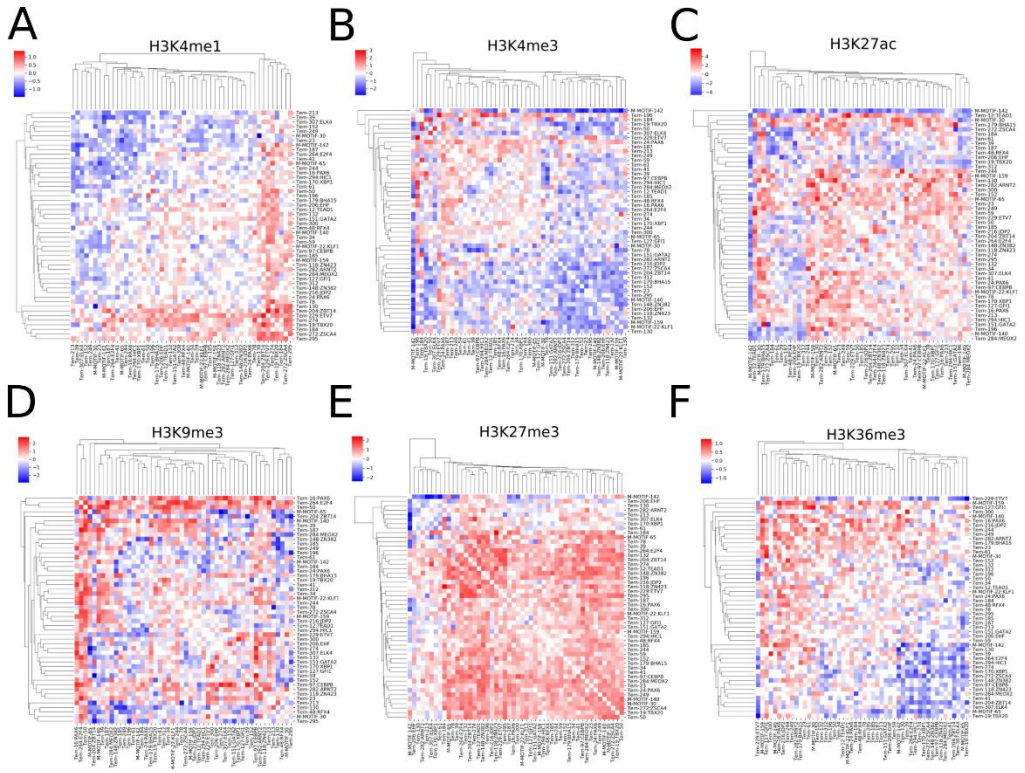

FIGURE 4-26: Interaction coefficient γ between the top 50 learned motifs in Tem cell. A-F Interactions between each pair of the top 50 learned motifs on the prediction of the six marks by the Tem cell model.

Shown in FIGURE 4-27A-D. are the results for the H3K4me1 model. Again, there are distinct patterns of positive and negative interactions between the motifs for predicting different cell types by the model. As in the cases of cell models, the motifs can be clustered into groups based on the patterns of their interactions for predicting the cell types. For instance, in the case of predicting the Tn cells, the putative novel motifs M-Motif-71 and H3K4me1-30 form a group with a negative interaction; learned motifs matching those of HIC2, HXD2, TFE2, ZN547, HAND1, COT1, SMAD4, TBX1, ANDR, ZN263, THA, ZN784, ZSCA4, ZN436, PTF1A and ZN770 form a group with many putative novel motifs with largely positive interactions among them; learned motifs matching those of HXC10, PO3F3, POXJ3, HMGA2, HXC10, DLX1 and ZN250 form a group with many putative novel motifs with largely negative interactions among them. Some of the predicted interactions are supported by experimental evidences. For example, we predicted that TFE2 interacted with HAND1 for predicting Tn (γ=5.38, p=1.84e-137), Tcm (γ=4.00, p=6.94e-

115), Tem (γ=2.82, p=1e-70) in Temra (γ=-7.61, p=1.97e-45), while it is has been reported that TFE2 (also named E47) directly interacts with HAND1 [247]. We predicted that SMAD4 interacted with ANDR for predicting Tn (γ=2.91, p=2.68e-47), Tcm (γ=3.49, p=1.86e-77), Tem (γ=2.99, p=3.47e-79) and Temra (γ=-0.93, p=0.0002), while SMAD4 is known to interact with ANDR, which might be involved in differential regulation of the androgen receptor gene transactivation [248]. We predicted that TFE2 interacted with PTF1A for predicting Tn (γ=5.22, p=3.69e-20), Tcm (γ=3.247, p=2.84e-29), Tem(γ=2.40, p=1.86e-13), and Temra (γ=-5.54, p=1.89e-68), while it has been reported that SMAD4 physically interacted with PTF1A and plays a crucial role in regulating signal pathways[249]. We predicted that HMGA2 interacted with SMAD4 for predicting Tn (γ=-0.41, p=0.026), Tcm (γ=-2.24, p=2.84e-13) and Temra (γ=0.90, p=8.77e-05), while it is known that HMGA2 interacts with SMAD3/SMAD4 to regulate SNAIL1 gene expression [250]. The predicted interactions between known and unknown motifs as well as those between unknown motifs are likely to be novel interactions, in particular those with strong and highly significant interactions, such as the negative interaction between M-Motif-71 and H3K4me1-30 for predicting Tn (γ=-4.17, p=2.75e-274), Tcm(γ=-2.88, p=3.40e-115) and Tem (γ=-2.28, p=2.36e-100), and a positive interaction for predicting Temra(γ=3.78, p=2.09e-63). Similar patterns of interactions are seen in the models of the other five histone marks (FIGURE 4-28A-D to FIGURE 4-32A-D).

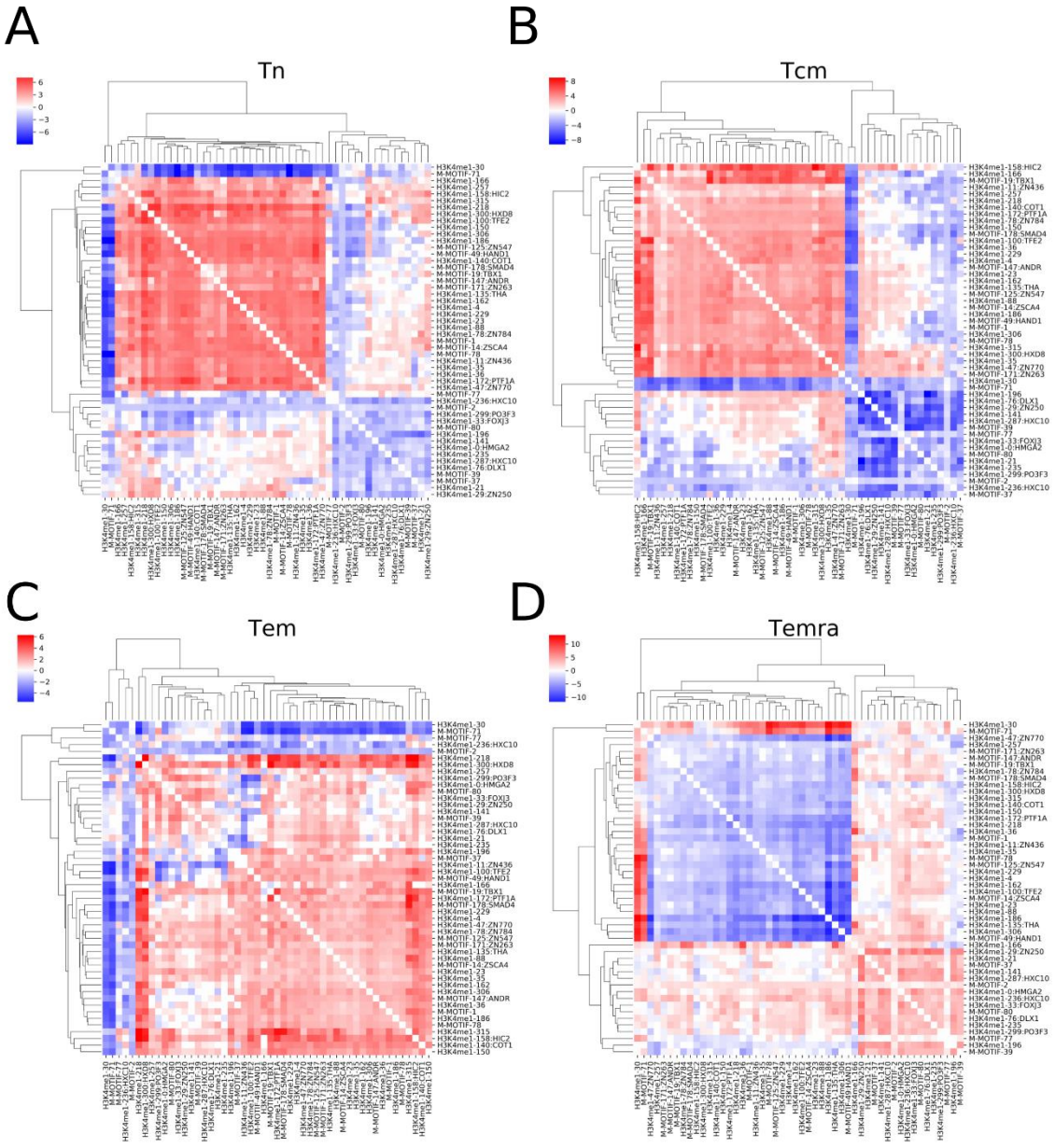

FIGURE 4-27: Interaction between the learned motifs in H3K4me3 model. A-D Interactions between each pair of the top 50 learned motifs on the prediction of the four cell types by the H3K4me1 model.

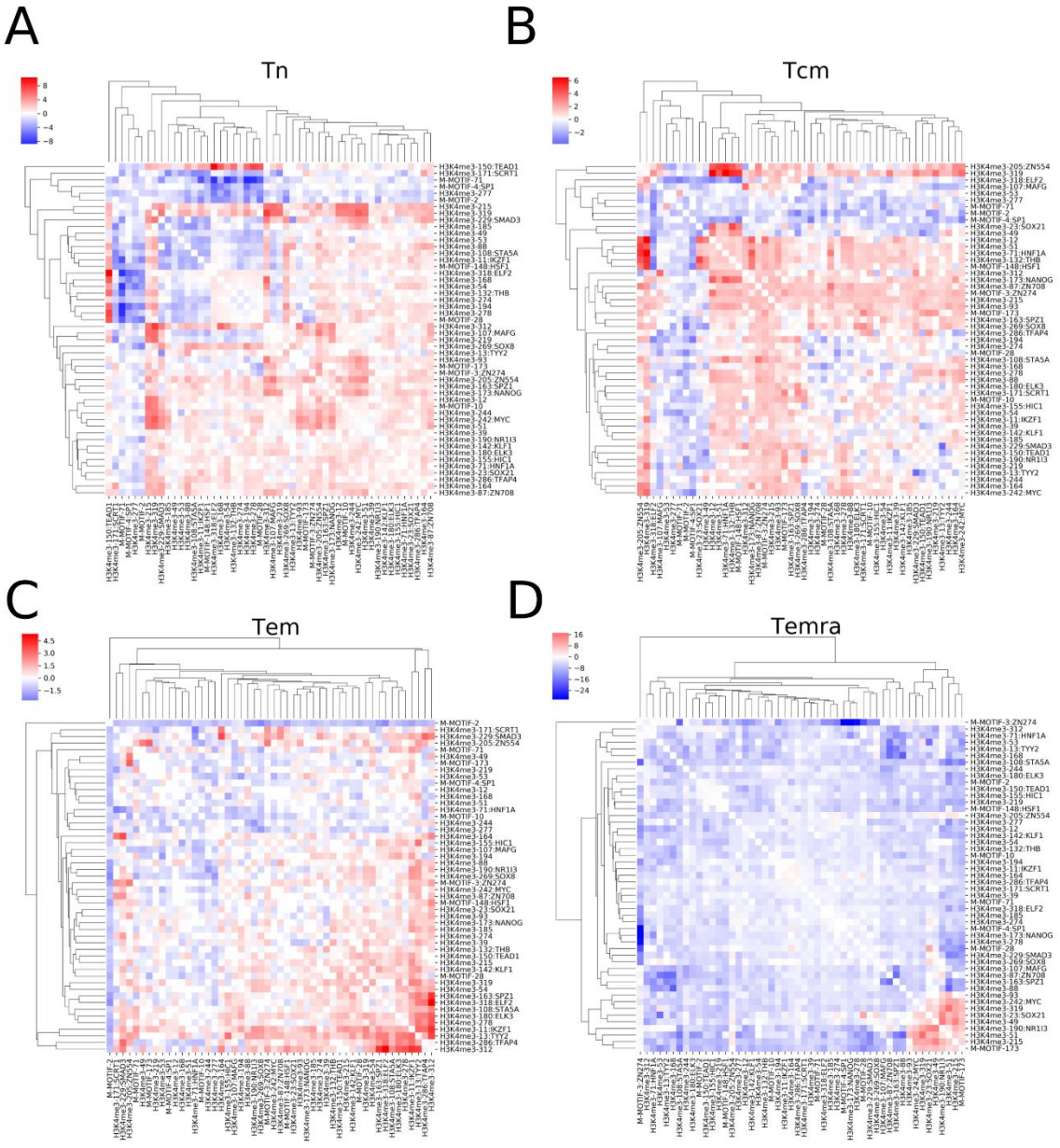

FIGURE 4-28: Interaction between the learned motifs in H3K4me3 model. A-D Interactions between each pair of the top 50 learned motifs on the prediction of the four cell types by the H3K4me3 model.

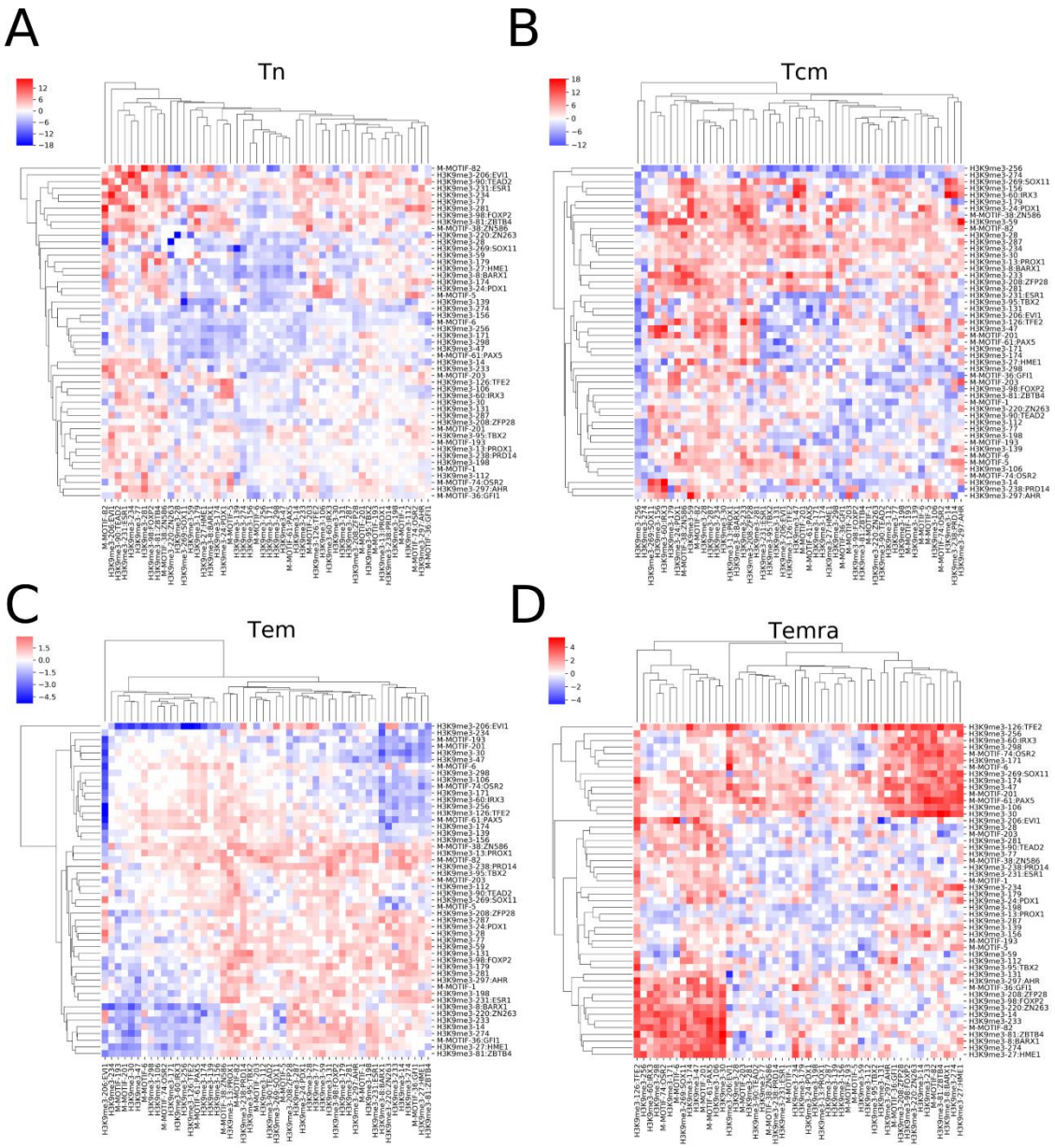

FIGURE 4-29: Interaction between the learned motifs in H3K9me3 model. A-D Interactions between each pair of the top 50 learned motifs on the prediction of the four cell types by the H3K9me3 model.

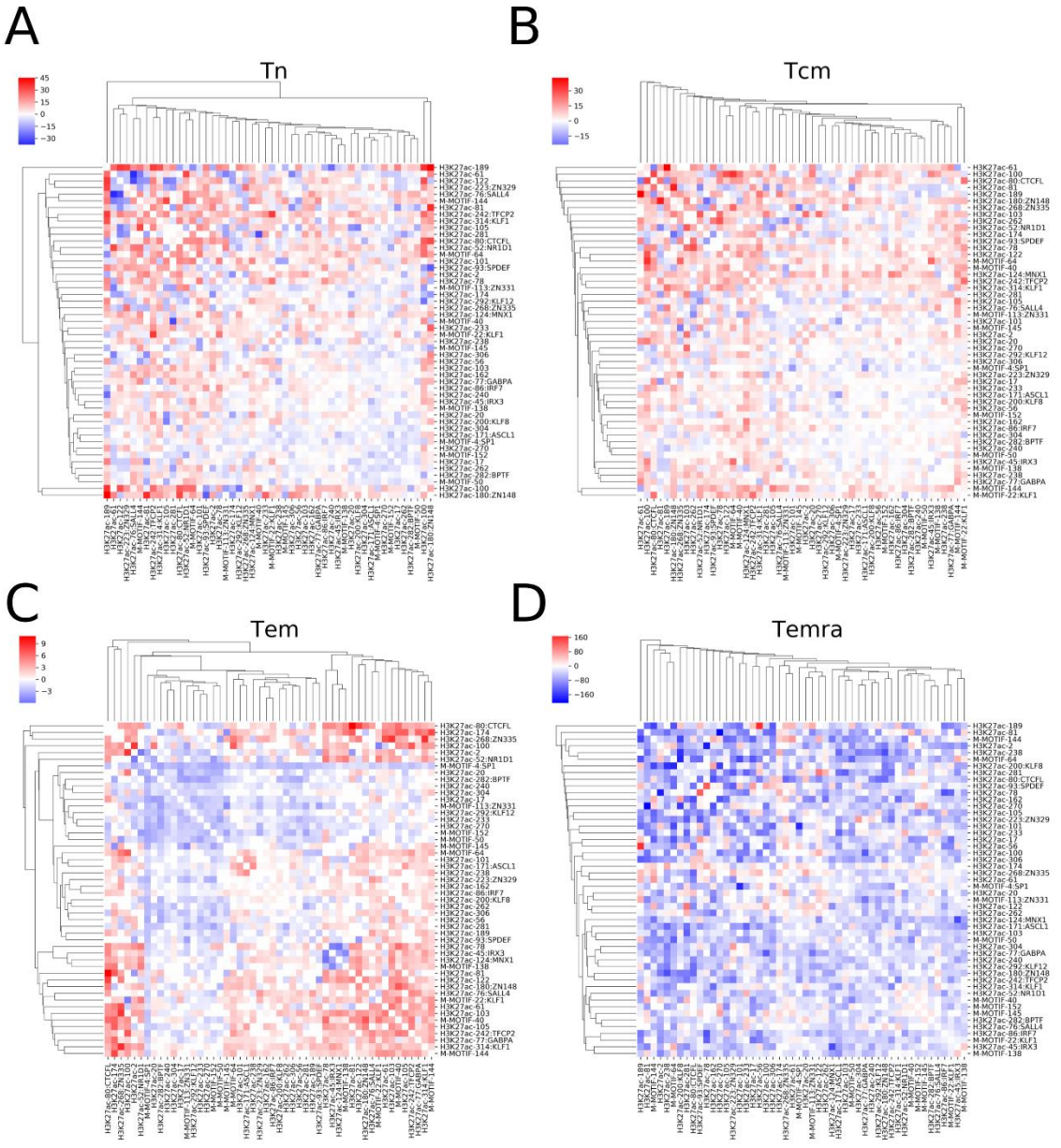

FIGURE 4-30: Interaction between the learned motifs in H3K27ac model. A-D Interactions between each pair of the top 50 learned motifs on the prediction of the four cell types by the H3K27ac model.

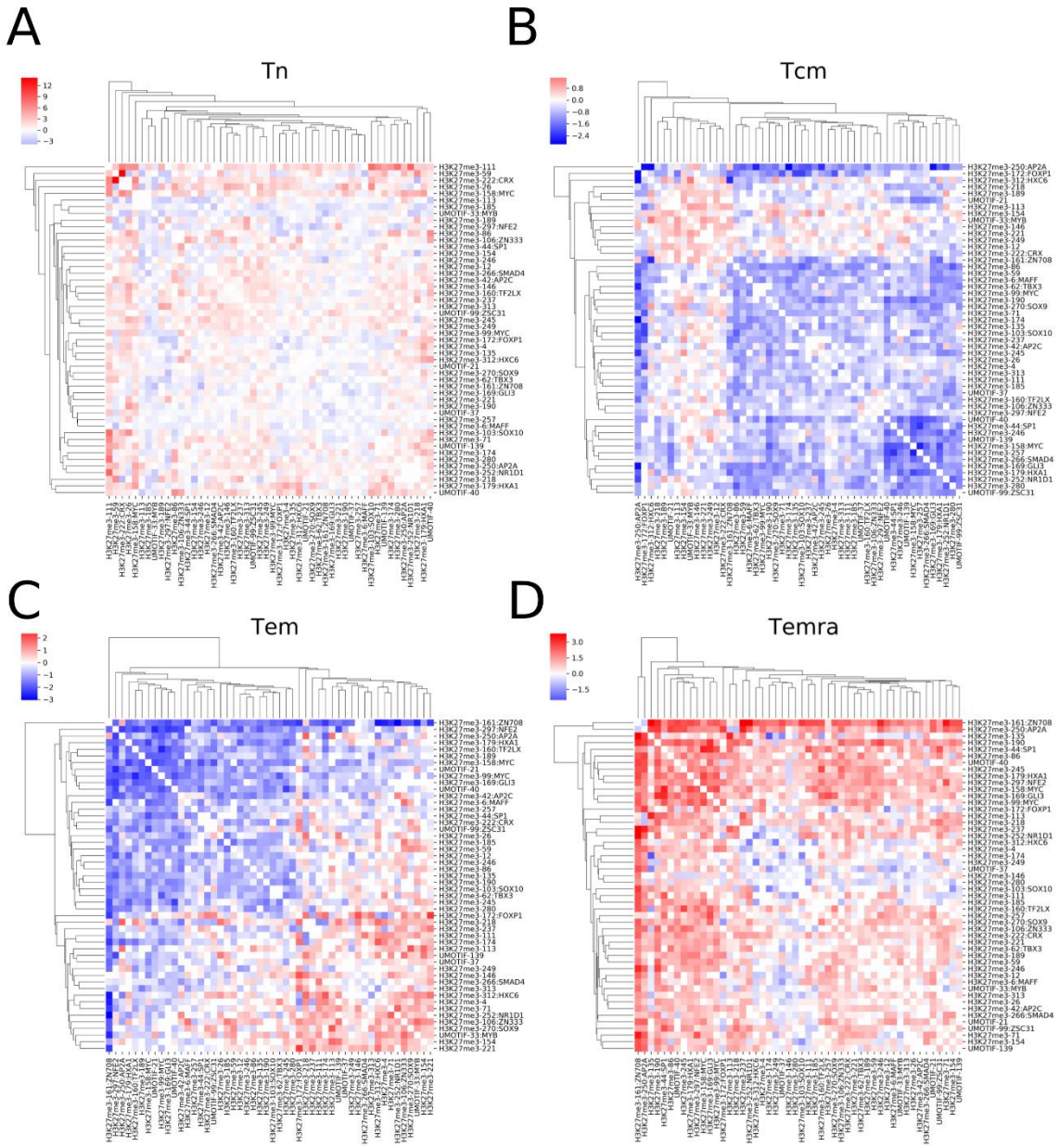

FIGURE 4-31: Interaction between the learned motifs in H3K27me3 model. A-D Interactions between each pair of the top 50 learned motifs on the prediction of the four cell types by the H3K27me3 model.

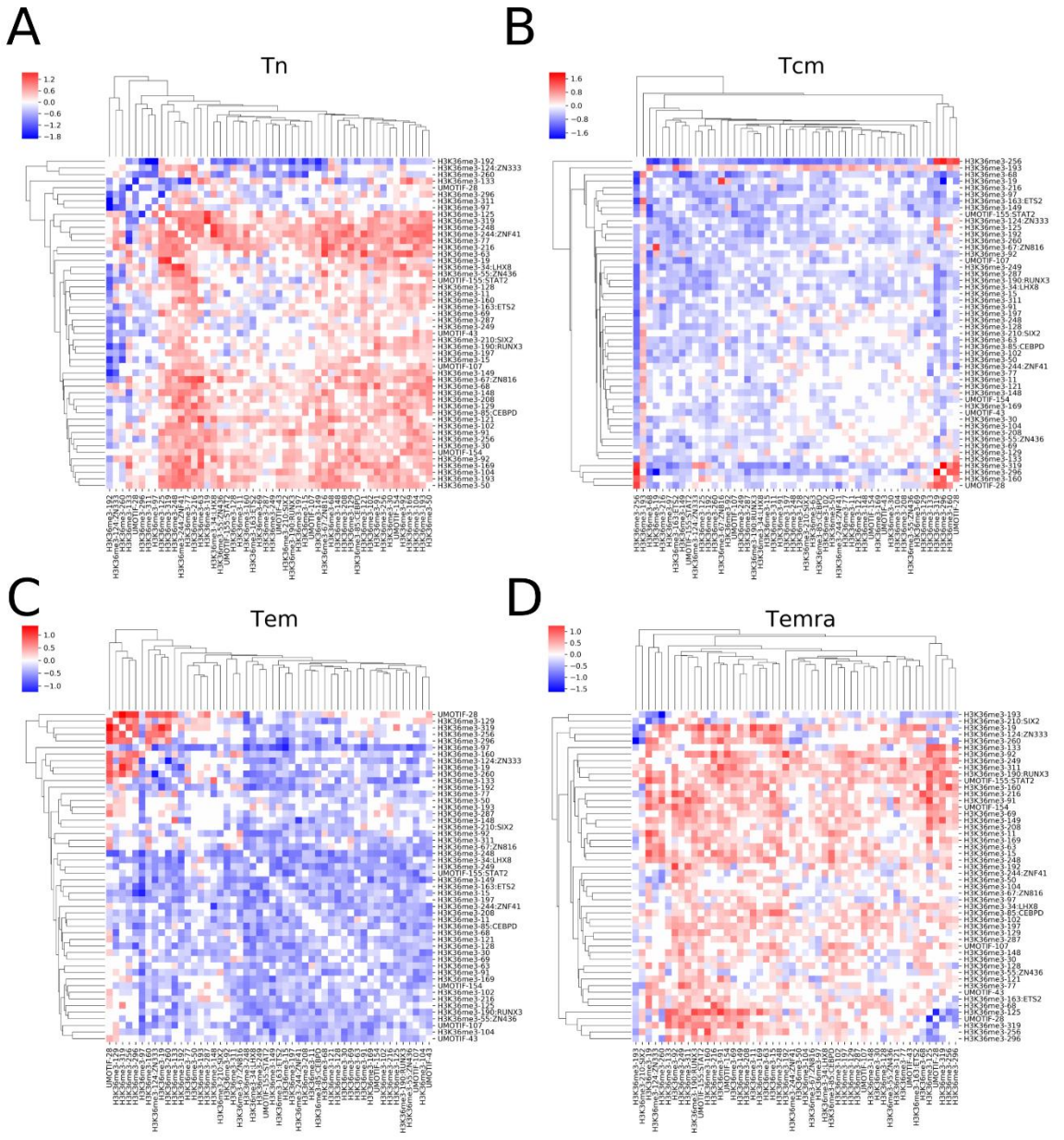

FIGURE 4-32: Interaction between the learned motifs in H3K36me3 model. A-D Interactions between each pair of the top 50 learned motifs on the prediction of the four cell types by the H3K36me3 model.

4.4. Discussion

DNA sequence plays a crucial role in determining its epigenomic state through interacting with the TFs and epigenome remodeling systems. However, our current understanding of these sequence determinants is still limited, and thus new methods are needed to reveal them. Recently, Whitaker and colleagues [36] trained a random forest classifier based on a set of pre-specified DNA motifs to predict six histone marks in H1 and its derived cell types with high accuracy. The results strongly support the pivotal roles of these motifs in specifying the unique epigenomes in the cells. However, this method could not discover sequence determinants ab initio, therefore, new methods are needed to gain a better understanding of the sequence determinants of epigenomes of cell types. CNNs have been proved to be a powerful approach to predict epigenomic features including TF binding [205], DNase I accessibility [208], DNA methylation [206, 251] and histone modifications [206]. And one of the advantages of CNNs, which other machine-learning methods often lack, is their ability to automatically learn the features of the objects through the filters in the convolutional layers [252]. In the case of epigenomic analysis, these features include sequence determinants that define the unique patterns of epigenetic modifications in different cell types produced during embryogenesis and development. Thus, CNNs can be a powerful approach to real the epigenomic sequence determinants.

Indeed, efforts have been made to interpret the sequence features learned by CNN models for predicting epigenomic marks [205-208]. However, these studies used a single mixed model to predict a combination of multiple epigenetic marks with multiple cell types, thus lack the power of comparative analyses for the learned sequence features. To overcome this limitation and facilitate interpreting CNN models which can be otherwise highly challenging [253], we developed two types of CNN models to capture the sequence features for various histone modifications in different cell types: 1) the cell type model for predicting patterns of various histone modifications in a cell type, and 2) the histone mark model for predicting various cell types based on a histone mark. In

this way, by comparing the motifs earned in different cell type models, we could identify the common and unique motifs that specify unique patterns of various histone modifications in a cell type; and by comparing the motifs learned in different histone mark models, we could detect the common and unique motifs that determine different patterns of the same histone mark in different cell types. Furthermore, the models enable us to evaluate the inferences of learned motifs and their interactions on the prediction accuracy, thereby predicting roles of each motif in specifying the epigenome and the type of cells.

To validate this strategy, we applied it to a dataset of six histone marks derived from four well-studied CD_4^+ T cell types in humans, i.e., Tn, Tcm, Tem and Temra. Both our histone mark models and cell type models achieved very high accuracy and were highly robust when tested on the dataset for H1 and its derived cell types, suggesting that our models have largely learned the relevant sequence features in determining the unique histone mark patterns in these cells. Not surprisingly, a large portion of the learned motifs in the first convolutional layers in the models resemble those of TFs that are known to play crucial roles in T cell development, while the remaining ones could be novel motifs of unknown TFs participating in T cell differentiation. By comparing the motifs learned in different cell models, we predicted that the unique patterns of various histone modifications in each cell type were largely determined by a unique set of motifs (FIGURE 4-10A and FIGURE 4-10B) and at the same time, the number of common motifs shared by two cell models reflected the linear lineage relationships of the four CD4+ T cell types (FIGURE 4-13), which is consistent with the results based on DNA methylation, DNase hypersensitivity and transcription patterns in the earlier study that produced the datasets used in our analysis. Furthermore, by comparing the motifs learned in different histone mark models, we predicted that different patterns of the same histone marks in different cell types were largely determined by a unique set of motifs (FIGURE 4-10A and B), while the number of common motifs shared by two histone mark models reflected their co-modification and exclusiveness natures (FIGURE 4-14). All these results suggest

that at least most of the learned motifs are likely to be authentic and play roles in T cell differentiation. Moreover, by computing the inference scores of the learned motifs, we further predicted the specific roles of each learned motif in determining the patterns of various histone modifications in a cell (FIGURE 4-15A-D and FIGURE 4-17A), or different patterns of the same histone modification in different cells (FIGURE 4-16A-F and FIGURE 4-17B). Finally, by computing an interaction score, we predicted the interactions of the cognate TFs of the learned motifs in either the cell models or histone mark models. Some of these predictions have experimental supports. Thus, our results support the hypothesis that sequences ultimately determine the unique epigenomes of different cell types through their interactions with TFs, epigenome remodeling system and extracellular cues during cell differentiation in a stepwise manner. Therefore, the motifs learned in our CNN models are highly interpretable and may provide insights into the underlying molecular mechanisms of establishing the unique histone modifications in different cell types.

4.5. Conclusion

We have developed two types of highly accurate CNNs constructed for cell types and for histone marks to predict the different histone marks in a cell type and different patterns of same mark in different cells, respectively. We showed that both the unique histone modification patterns in a cell type and the different patterns of the same histone mark in different cell types are determined by a set of motifs with unique combinations. The level of sharing motifs learned in the different cell models reflects the lineage relationships of the cells, while the level of sharing motifs learned in different histone mark models reflects their functional relationships. The models enable the prediction of the importance of the learned motifs and their interactions in determining specific histone mark patterns in the cell types. Therefore, the motifs learned in the models are highly interpretable and may provide insights into the underlying molecular mechanisms of establishing the unique histone modifications in different cell types. Our results suggest the hypothesis that DNA sequences ultimately determine the unique epigenomes of different cell types through their interactions with TFs, epigenome remodeling system and extracellular cues during cell differentiation in a stepwise manner.

CHAPTER 5: DEPCRMS DATABASE

5.1. Background

Cis-regulatory modules (CRMs), such as enhancers, promoters, silencers and insulators, are composed of clusters of short DNA sequences where transcriptional factors (TFs) can bind to regulate the expressions of the target genes in many biology processes. Recent studies have showed that most complex trait-associated single nucleotide polymorphisms (SNPs) often disrupt transcriptional factors binding sites (TFBSs) in CRMs. This might affect TF binding and gene transcription, which leads to complex diseases [7, 8]. In general, SNPs that disrupt TFBSs could affect the affinity of TF binding, resulting in the changes of chromatin characteristics and gene expression in specific cell types[10-13, 15]. Finally, the alternations of phenotypes in the molecular level could contribute to the changes of the phenotypes in the cellular or organ level in species[30, 254]. Therefore, categorization of the CRMs and their constituent TFBSs in sequenced genomes can facilitate characterizing the functions of the regulatory sequences and their roles in diseases.

Recently, multiple next-generation sequencing (NGS)-based technologies have been developed to characterize different features of the CRMs, such as chromatin immunoprecipitation followed by sequencing (ChIP-seq) [255] to profile the regions of various TF bindings or histone modifications, DNase I hypersensitive sites sequencing (DNase-seq) [256], assay for transposase-accessible chromatin using sequencing (ATAC-seq)[257], formaldehyde-assisted isolation of regulatory elements sequencing (FAIRE-seq) [258], and micrococcal nuclease digestion with deep sequencing (MNase-seq) [259] to identify the chromatin accessibility. An exponentially increasing number of datasets have been generated by consortia such as ENCODE[41, 260], Epigenomics Roadmap [43, 44] and Genotype-Tissue Expression (GTEx)[45]. Based on different data types that capture different aspects of the CRMs, many computational strategies have been developed to

predict the CRMs. For instance, based on the TF ChIP-seq data, methods such as SpaMo[73], CPModule[76], COPS[77], and INSECT[78] have been developed to identify regions of binding peaks, which contain closely located TFBSs, as putative CRMs. Based on multiple histone marks and chromatin accessibility datasets, hidden Markov models[87, 89] and dynamics Bayesian models[90] have been developed to predict CRMs in different cell types. Based on the bidirectional pairs of capped RNAs, the FANTOM project identified enhancers across genomes [120]. By integrating multiple tracks of epigenetics marks, TF binding, predicted and experimentally validated enhancers, several groups have developed CRM/enhancer databases, such as dbSUPER[166], SEdb[261], DENdb[97], EPDnew promoters[167], UCNEbase[262], CraniofacialAtlas[263], GeneHancer[98], HACER[264], RAEdb[265], HEDD[266], DiseaseEnhancer [267], SEA[268] and EnhancerAtlas (ref). However, none of them provides the de novo predicted constituent TFBSs information in the CRMs, which is critical to understand the mechanisms of the transcriptional regulation as well as to pinpoint causal variants of phenotype diversity and diseases.

Using a highly accurate CRM and TFBS prediction tool dePCRM2 that we developed recently, we predicted the CRMs and their constituent TFBSs in *Homo sapiens*, *Mus musculus* and *Caenorhabditis elegans*. We now constructed a database dePCRMS to facilitate the community to use these predictions for various purposes. The database currently contains 1,155,151, 777,409 and 19,515 predicted CRMs as well as 89948206, 103718473, and 3758557 TSBSs for 201, 210, 61 unique motif (UM) families in *H. sapiens*, *M. musculus* and *C. elegans*, respectively. The web interface of the dePCRMS database can quickly browse and visualize the contents of the database and provide three functional analysis modules. Using these modules, the user can find the closest CRMs to a gene, search the CRMs that are located in a specified range around a gene, and retrieve all CRMs that contain TFBSs of a specific TF. The interface also provides copy, export and

download functions of the selected CRMs or all predicted CRMs in a BED format. We will update the database when new datasets are available and include prediction in other organisms in the future.

5.2. Methods and materials

5.2.1. Datasets

We downloaded 6092 TF ChIP-seq datasets for 779 TFs in 2631 cells/tissues/organs of human, and 4,786 TF ChIP-seq datasets for 501 TFs in 1,560 cells/tissues/organs of mouse from the Cistrome database[47], and 212 TF ChIP-seq datasets for 91 TFs of C. elegans from the modEncode database[269]. After filtering out peaks with low quality, for each left peak, we extracted 1,000bp genome sequence centering on the middle point of the binding peaks, thereby extending majority of the binding peaks.

5.2.2. Prediction of CRMs and constituent TFBSs

To predict CRMs and TFBSs, we apply dePCRM2 to the datasets with extended binding peaks from each species using the default paraments. Briefly, dePCRM2 first finds overrepresented motifs and co-occurring motifs pairs (CPs) in each dataset. It then constructs a similarity network of highly similar motifs in CPs across all the datasets and identifies unique motifs (UMs). dePCRM2 constructs an interaction network of the UMs, where UMs are the nodes, and two nodes are connected by a weighted edge with their interaction score being the weight, which is defined as follows,

$$S_{\mathrm{INTER}}(U_i, U_j)\frac{1}{|D(U_i, U_j)|}\sum_{d\in D(U_i,U_j)}\left(\frac{1}{|U_i(d)|}+\frac{1}{|U_j(d)|}\sum_{s\in S(U_i(d),U_j(d))}\frac{150}{r(s)}\right), \qquad 5\text{-}1$$

where $D(U_i, U_j)$ is the datasets which contain the TFBSs of UMs U_i and U_j, $U_k(d)$ is the peaks containing at least one TFBS of U_k in dataset d, $S(U_i(d), U_j(d))$ is the peaks which contain at least one TFBS of both U_i and U_j, and $r(s)$ is the shortest distance between a TFBS of U_i and a TFBS of U_j in a sequence s. dePCRM2 connects any two adjacent TFBSs of the UMs if their distance $d \leq$ 300bp and considers each resulting connected DNA segment as a CRM candidate (CRMC),

thereby partitions the genome regions covered the extended peaks in a CRMC set and a non-CRMC set. dePCRM2 evaluates each CRMC containing $b_1, b_2 \cdots, b_n$ TFBSs by computing a score defined as follows,

$$S_{CRM}(b_1, b_2 \cdots, b_n) = \frac{2}{n-1} \times \sum_{i=1}^{n} \sum_{j>i} W\left[U(b_i), U(b_j)\right] \times \left[S(b_i) + S(b_j)\right], \quad 5\text{-}2$$

where $U(b_k)$ is the UM of TFBS b_k, $W\left[U(b_i), U\left(b_j\right)\right]$ is the interaction score between $U(b_i)$ and $U\left(b_j\right)$, $S(b_k)$ is the binding score of b_k based on the position weight matrix (PWM) of $U(b_k)$. Only TFBSs with a positive score are considered. dePCRM2 also computes a p-value for each CRMC as follows. For each predicted CRMC, dePCRM2 generates a Null CRMC that has the same length and 4-mer nucleotide frequencies as the CRMC using a third order Markov chain model [117], and computes a S_{CRM} score for each Null CRMC based on a random interaction network which is generated by randomly rewiring the nodes of the UM interaction network. Then, an empirical p-value for a CRMC with a S_{CRM}=s is computed based on the distribution of S_{CRM} score for Null CRMCs,

$$p = \frac{n(s)}{N}, \quad 5\text{-}3$$

where $n(s)$ is the number of Null CRMCs with a S_{CRM} score greater than s, and N is the total number of CRMCs.

5.2.3. Technical implementation

The current version of PCRMv2 was developed using MySQL 5.7.17 (http://www.mysql.com) and runs on a Linux-based Apache2 server (http://www.apache.org). The PHP 7.2 (http://www.php.net) was used for back-end scripting. The interactive interface and responsive feature were implemented using Bootstrap 4 (https://getbootstrap.com/), JQuery (http://jquery.com)

and dataTables (https://datatables.net), and NCBI sequence viewer 3.38.0 (https://www.ncbi.nlm.nih.gov/projects/sviewer) was used for visualization.

5.3. Results

5.3.1. Web interface to the database

We provide a user-friendly web interface to the PCRMv2 database for inquiring and browsing predicted CRMs at different significant levels for each organism. The user can conduct gene centric, CRM centric and TFBS centric queries through three functional analysis modules, (i) to search CRMs at a p-value in a given upstream and/or downstream regions of a gene of interest, (ii) to search the closest or bracket genes to a given CRM, and (iii) to search the TFBSs of a TF on one or more chromosomes. The user can filter, export, and download the returned query results (FIGURE 5-1).

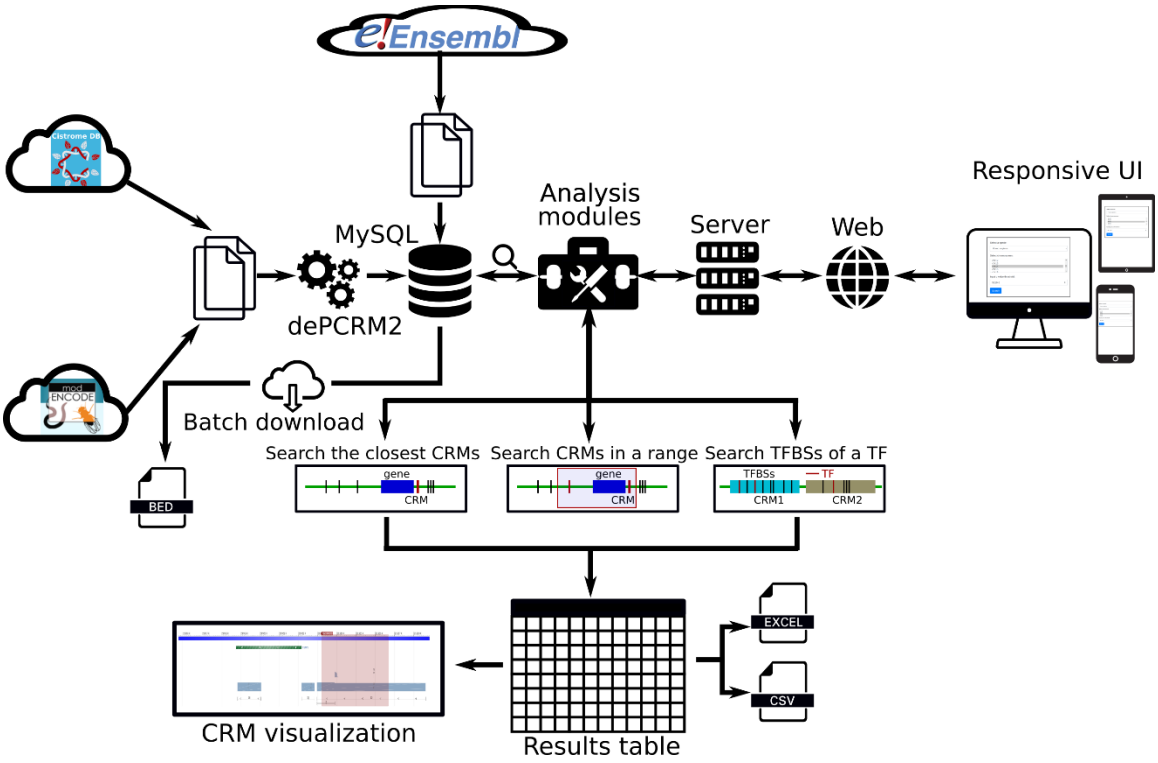

FIGURE 5-1: Overview of dePCRM webserver. it includes the database, data integration pipeline, analysis modules and features.

5.3.2. Quick browse of database contents

We provide a quick search function by which the user can browse the CRMs predicted at a selected p-value on selected one or multiple chromosomes in a selected organism (FIGURE 5-2A, panel 1). The results are displayed in an interactive manner (FIGURE 5-2A, panel 2), so that the user can change the number of entries shown in a page, sort results based on different criteria/columns, filter the results using the search box, and set visible columns. The user can copy or export the selected results in a file in the CSV or Excel formats, or export all records if no CRM is selected by default (FIGURE 5-2A, panel 3). The coordinates of a selected CRM can be visualized in the NCBI genome viewer with a red rectangle (FIGURE 5-2A, panel 4), and the TFBSs in the CRM can be displayed in a responsive table by clicking the CRM ID alongside the NCBI viewer panel (data not shown). The user can download all the CRMs in each specie in the BED format in the download menu.

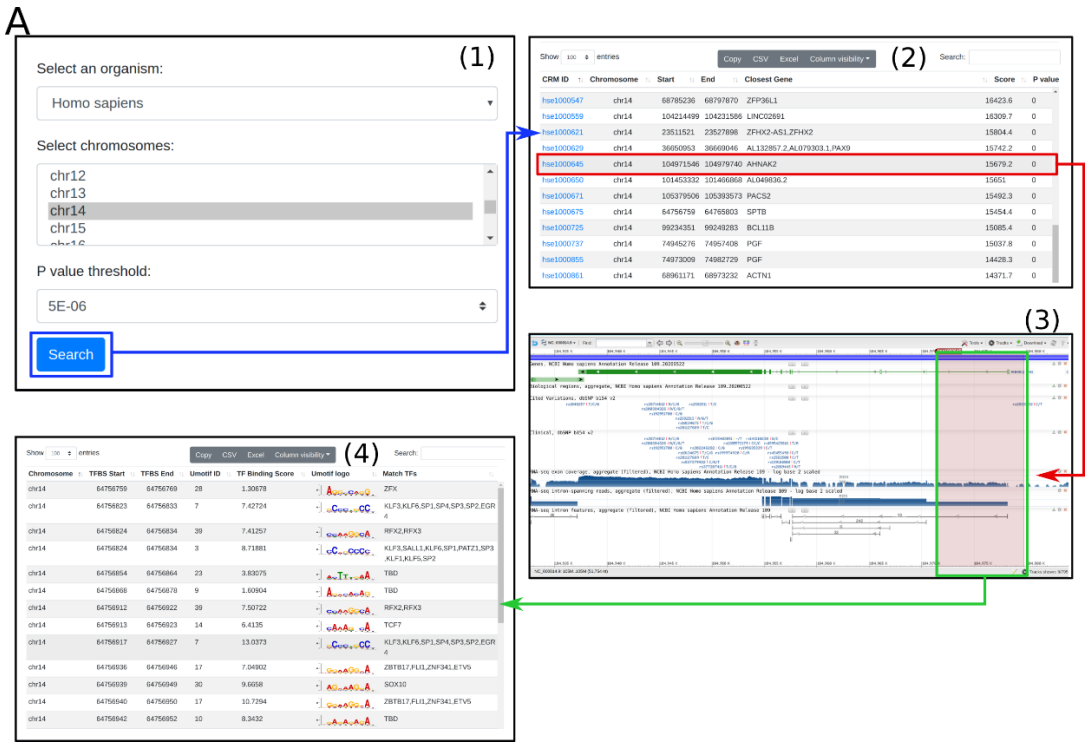

FIGURE 5-2: The quick browsering interface of CRMs in dePCRMS. (A) Quick search panel. (1) CRMs search form, (2) search results table, (3) CRM visualization in the NCBI genome viewer, (4) detailed TFBSs information.

5.3.3. Functional analysis modules

To facilitate analyzing potential gene-CRM relationship and TFBS landscape of specific TFs, we provide three functional analysis modules. First, the user can search the closest CRMs to a gene in an organism at different p-values (e.g., 0.05, 0.01, 5×10^{-6}, 1×10^{-6}) in the dropdown list (FIGURE 5-3A). Second, the user can search CRMs located in specified upstream and/or downstream ranges of a gene from its either end (FIGURE 5-4A). In both modules, the returned CRMs are displayed in an interactive table. The user can sort, filter and export the results. Each CRM can be visualized in a new page in the NCBI genome viewer by clicking on the CRM ID. The user can view TFBSs in the CRM by clicking on the CRM ID in the left panel in the CRM table (FIGURE not shown). Each TFBS record displayed includes its coordinates, UM ID, Binding score, UM logo and Matched TFs. The user can sort the results based on CRM IDs, coordinates of TFBSs and their cognate TFs. CRMs can also be filtered by filling their p-value in the search box. Last, the user can retrieve all TFBSs of a selected set of TFs in one or more selected chromosomes in an organism in an interactive table. Each record of a TFBS includes its host CRM's ID, coordinates, cognate TF, TF binding score, and UM ID and Logo (FIGURE 5-5A). Furthermore, the user can visualize the CRMs containing the TFBSs in the same manner as above described.

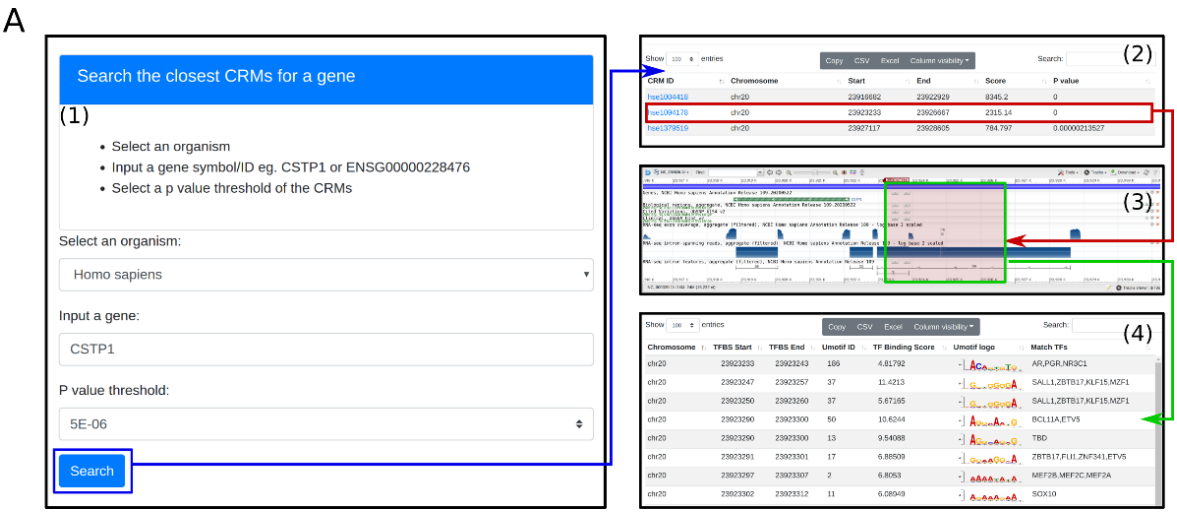

FIGURE 5-3: Searching closest CRM(s) to a gene module.

A

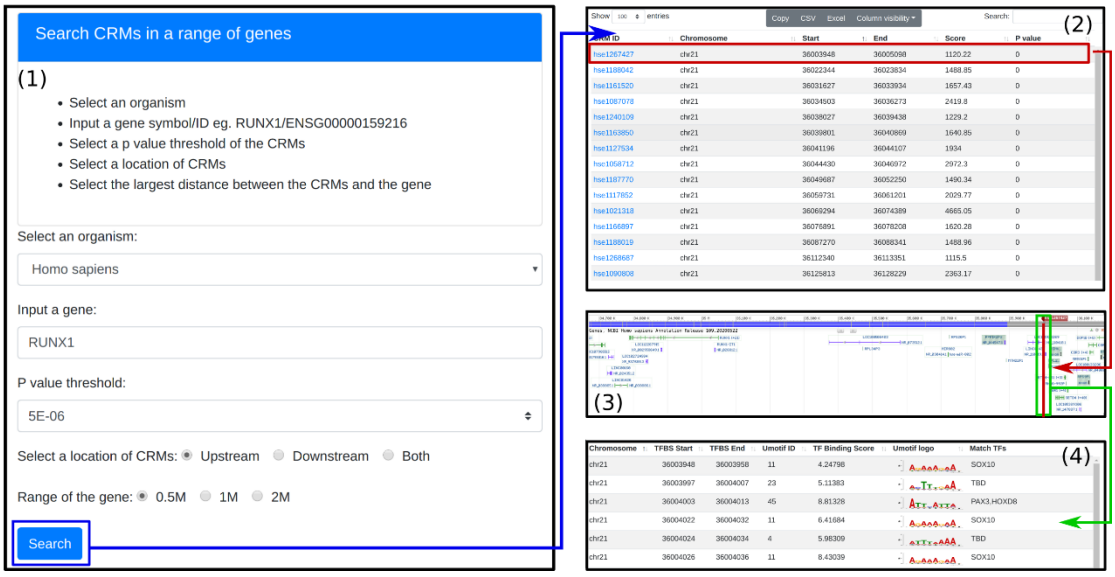

FIGURE 5-4: Searching CRM(s) in a range around a gene module.

A

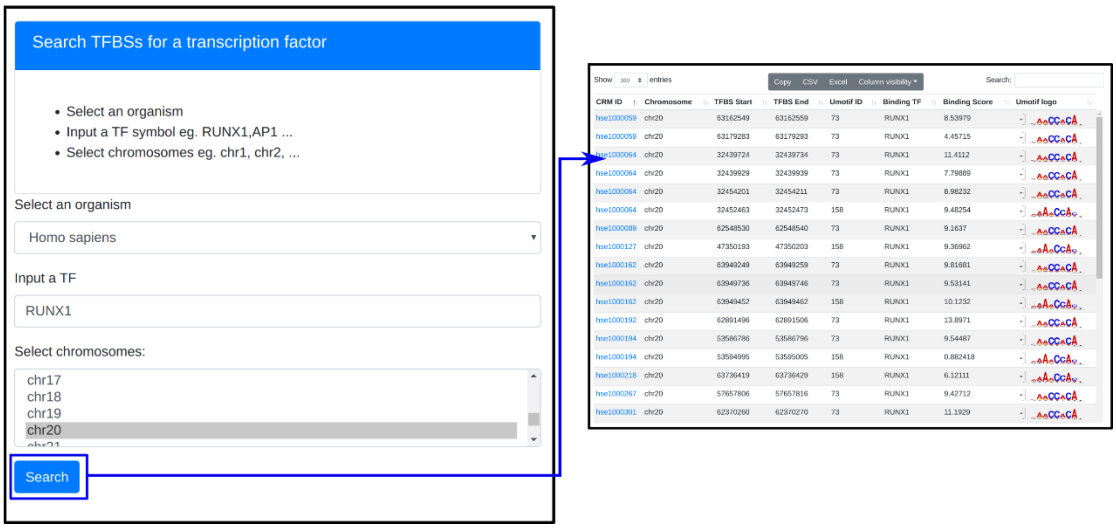

FIGURE 5-5: Searching all TFBSs of a TF module.

5.4. Conclusion

The dePCRMS database contains 1,155,151, 777,409 and 19,515 CRMs as well as 201, 210 and 61 unique motif families for *Homo sapiens*, *Mus musculus* and *Caenorhabditis elegans*, respectively. The web interface to dePCRMS database provides quick browsing, searching and visualizing the CRMs and TFBSs. It also provides three functional analysis modules to search closest CRM(s) to a gene, CRM(s) in a range around a gene, and landscape of TFBSs of a specific TF. It helps the users to select the CRMs and export into files in CSV or EXCEL formats, or to batch download the whole CRMs datasets in the BED format. In the future development, we will add predictions in other important organism, and update the prediction when more data are available in Cistrome or other databases. We will also add more functional analysis modules to support analyses such as target genes of CRMs and causal SNPs of traits and diseases by integrating more data sources. To our knowledge, dePCRMS is the first comprehensive CRMs database with de novo predicted TFBSs at a single nucleotide resolution in multiple important genomes, and we hope it will facilitate the research community to characterize the regulatory genomes in important organisms.

CHAPTER 6: CONCLUSION

In this dissertation, we have developed a pipeline called dePCRM2 to predict CRMs and their constituent TFBSs in genomes by integrating multiple ChIP-seq datasets for various TFs from different cells or tissues. We predicted an unprecedentedly complete map of the CRMs and their constituent TFBSs in 77.47% of the human genome using more than 6,000 datasets. Both the evolutionary constraints and the experiment validated enhancers indicate that dePCRM2 might achieve high sensitivity and specificity. It is possible to predict a more complete map of the CRMs and their constituent TFBSs with more diverse and balanced data in the future. With a static map of CRMs and their constituent TFBSs, the next question we could ask is what the influence of a mutation in a specific TFBS is on the underlying molecular mechanisms in different cell types. To address this question, at least partially, we developed two types of CNNs to predict the cell types in the same histone mark and to predict the histone marks based on the same cell type, respectively. We indicated that a unique combination of motifs could determine the unique histone mark patterns in one cell type or the occurrence of a histone mark in different cell types. The degrees of the sharing motifs learned in various cell models reflect the lineage relationships of the cells, while the degrees of the sharing motifs learned in various histone mark models reflect their functional relationships. By manipulating the forward propagation information of the learned motifs and then measuring the changes between the predictions, we found that the learned motifs might interpret the underlying molecular mechanisms of the unique histone mark combination in different cell types. The results suggest that DNA sequences, more specifically, TFBSs ultimately determine the unique epigenomes in different cell types via their interactions with TFs, as well as the epigenome remodeling during cell differentiation. To facilitate the research community to characterize the regulatory modules in human and import the model organisms, we developed the dePCRMS database. The dePCRMS database includes 1,155,151, 777,409 and 19,515 CRMs as well as 201,

210 and 61 unique motif families for *Homo sapiens, Mus musculus* and *Caenorhabditis elegans*, respectively. The web interface helps the users quickly browse, search and visualize the CRMs and TFBSs. And three types of functional analyses can be conducted: searching the closest CRM(s) to a gene, searching CRM(s) in a range around a gene, and searching TFBSs landscape of a specific TF.

To summarize, in this dissertation, we tried to address three questions: firstly, we predicted a map of the CRMs and their constituent TFBSs. Secondly, we pinpointed the influence of the motifs on the underlying molecular mechanisms of the histone mark formation and the cell differentiation. Finally, we built a database for holding the maps of the CRMs and their constituent TFBSs in three genomes, *Homo sapiens, Mus musculus* and *Caenorhabditis elegans*.

REFERENCES

[1] M. King and A. Wilson, "Evolution at two levels in humans and chimpanzees," Science vol. 188, pp. 107-116, 1975.

[2] P. J. Wittkopp and G. Kalay, "Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence," Nat Rev Genet, vol. 13, no. 1, pp. 59-69, 2012.

[3] R. R. Haraksingh and M. P. Snyder, "Impacts of variation in the human genome on gene regulation," Journal of molecular biology, vol. 425, no. 21, pp. 3970-7, 2013.

[4] M. Rubinstein and F. S. de Souza, "Evolution of transcriptional enhancers and animal diversity," Philos Trans R Soc Lond B Biol Sci, vol. 368, no. 1632, p. 20130017, 2013.

[5] A. Siepel and L. Arbiza, "Cis-regulatory elements and human evolution," Curr Opin Genet Dev, vol. 29, pp. 81-9, 2014.

[6] S. K. Reilly *et al.*, "Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis," Science., vol. 347, no. 6226, pp. 1155-9, 2015.

[7] L. A. Hindorff *et al.*, "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits," Proc Natl Acad Sci U S A, vol. 106, no. 23, pp. 9362-7, 2009.

[8] E. M. Ramos *et al.*, "Phenotype-Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources," Eur J Hum Genet, vol. 22, no. 1, pp. 144-7, 2014.

[9] M. T. Maurano, H. Wang, T. Kutyavin, and J. A. Stamatoyannopoulos, "Widespread site-dependent buffering of human regulatory polymorphism," PLoS genetics, vol. 8, no. 3, p. e1002599, 2012.

[10] M. T. Maurano *et al.*, "Systematic Localization of Common Disease-Associated Variation in Regulatory DNA," Science, vol. 337, no. 6099, pp. 1190-1195, 2012.

[11] M. Kasowski *et al.*, "Extensive variation in chromatin states across humans," Science, vol. 342, no. 6159, pp. 750-2, 2013.

[12] H. Kilpinen *et al.*, "Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription," Science, vol. 342, no. 6159, pp. 744-7, 2013.

[13] G. McVicker *et al.*, "Identification of genetic variants that affect histone modifications in human cells," Science, vol. 342, no. 6159, pp. 747-9, 2013.

[14] L. Wu *et al.*, "Variation and genetic control of protein abundance in humans," Nature, vol. 499, no. 7456, pp. 79-82, 2013.

[15] D. Huang and I. Ovcharenko, "Identifying causal regulatory SNPs in ChIP-seq enhancers," Nucleic Acids Res, vol. 43, no. 1, pp. 225-36, 2015.

[16] J. Majewski and T. Pastinen, "The study of eQTL variations by RNA-seq: from SNPs to phenotypes," Trends in genetics : TIG, vol. 27, no. 2, pp. 72-9, 2011.

[17] C. Attanasio *et al.*, "Fine tuning of craniofacial morphology by distant-acting enhancers," Science, vol. 342, no. 6157, p. 1241006, 2013.

[18] W. Fu, T. D. O'Connor, and J. M. Akey, "Genetic architecture of quantitative traits and complex diseases," Curr Opin Genet Dev, vol. 23, no. 6, pp. 678-83, 2013.

[19] M. Spielmann and S. Mundlos, "Structural variations, the regulatory landscape of the genome and their alteration in human disease," Bioessays, vol. 35, no. 6, pp. 533-43, 2013.

[20] E. Smith and A. Shilatifard, "Enhancer biology and enhanceropathies," Nat Struct Mol Biol., vol. 21, no. 3, pp. 210-9, 2014.

[21] A. Mathelier, W. Shi, and W. W. Wasserman, "Identification of altered cis-regulatory elements in human disease," Trends Genet, vol. 31, no. 2, pp. 67-76, 2015.

[22] H. M. Herz, D. Hu, and A. Shilatifard, "Enhancer malfunction in cancer," Mol Cell, vol. 53, no. 6, pp. 859-66, 2014.
[23] H. Ongen *et al.*, "Putative cis-regulatory drivers in colorectal cancer," Nature., vol. 512, no. 7512, pp. 87-90, 2014.
[24] L. HD, "- The Maternal-to-Zygotic Transition. Preface," Current topics in developmental biology, pp. 00076-9, 2015.
[25] H. Heyn, "Quantitative Trait Loci Identify Functional Noncoding Variation in Cancer," PLoS genetics, vol. 12, no. 3, p. e1005826, 2016.
[26] H. Heyn *et al.*, "Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer," Genome biology, vol. 17, p. 11, 2016.
[27] E. Khurana, Y. Fu, D. Chakravarty, F. Demichelis, M. A. Rubin, and M. Gerstein, "Role of non-coding sequence variants in cancer," Nat Rev Genet, vol. 17, no. 2, pp. 93-108, 2016.
[28] S. Zhou, A. E. Treloar, and M. Lupien, "Emergence of the Noncoding Cancer Genome: A Target of Genetic and Epigenetic Alterations," Cancer discovery, vol. 6, no. 11, pp. 1215-1229, 2016.
[29] X. Li *et al.*, "OncoBase: a platform for decoding regulatory somatic mutations in human cancers," Nucleic Acids Res, 2018.
[30] L. D. Ward and M. Kellis, "Interpreting noncoding genetic variation in complex traits and human disease," Nat Biotechnol, vol. 30, no. 11, pp. 1095-106, 2012.
[31] A. A. Pai, J. K. Pritchard, and Y. Gilad, "The genetic and mechanistic basis for variation in gene regulation," PLoS Genet., vol. 11, no. 1, p. e1004857, 2015.
[32] F. Collins, "Has the revolution arrived?," Nature, vol. 464, no. 7289, pp. 674-5, 2010.
[33] A. Burga and B. Lehner, "Beyond genotype to phenotype: why the phenotype of an individual cannot always be predicted from their genome sequence and the environment that they experience," The FEBS journal, vol. 279, no. 20, pp. 3765-75, 2012.
[34] D. S. Paul, N. Soranzo, and S. Beck, "Functional interpretation of non-coding sequence variation: concepts and challenges," BioEssays : news and reviews in molecular, cellular and developmental biology, vol. 36, no. 2, pp. 191-9., 2014.
[35] F. W. Albert and L. Kruglyak, "The role of regulatory variation in complex traits and disease," Nat Rev Genet, vol. 16, no. 4, pp. 197-212, 2015.
[36] J. W. Whitaker, Z. Chen, and W. Wang, "Predicting the human epigenome from DNA motifs," Nature methods, vol. 12, no. 3, pp. 265-72, 7 p following 272, 2015.
[37] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold, "Genome-wide mapping of in vivo protein-DNA interactions," Science, vol. 316, no. 5830, pp. 1497-502, 2007.
[38] G. Robertson *et al.*, "Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing," Nat Methods, vol. 4, no. 8, pp. 651-7, 2007.
[39] X. Chen, L. Guo, Z. Fan, and T. Jiang, "W-AlignACE: an improved Gibbs sampling algorithm based on more accurate position weight matrices learned from sequence and gene expression/ChIP-chip data," Bioinformatics, vol. 24, no. 9, pp. 1121-8, 2008.
[40] T. E. P. Consortium, "The ENCODE (ENCyclopedia Of DNA Elements) Project," Science, vol. 306, no. 5696, pp. 636-640, 2004.
[41] E. P. Consortium, "A user's guide to the encyclopedia of DNA elements (ENCODE)," PLoS Biol, vol. 9, no. 4, p. e1001046, 2011.
[42] J. A. Stamatoyannopoulos *et al.*, "An encyclopedia of mouse DNA elements (Mouse ENCODE)," Genome biology, vol. 13, no. 8, p. 418, 2012.
[43] B. E. Bernstein *et al.*, "The NIH Roadmap Epigenomics Mapping Consortium," Nat Biotechnol, vol. 28, no. 10, pp. 1045-8, 2010.
[44] A. Kundaje *et al.*, "Integrative analysis of 111 reference human epigenomes," Nature., vol. 518, no. 7539, pp. 317-30, 2015.

[45] G. T. Consortium, "The Genotype-Tissue Expression (GTEx) project," Nat Genet, vol. 45, no. 6, pp. 580-5, 2013.
[46] G. Consortium, "Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans," Science, vol. 348, no. 6235, pp. 648-60, 2015.
[47] S. Mei *et al.*, "Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse," Nucleic Acids Res, vol. 45, no. D1, pp. D658-d662, 2017.
[48] D. Kleftogiannis, P. Kalnis, and V. B. Bajic, "DEEP: a general computational framework for predicting enhancers," Nucleic Acids Res, vol. 43, no. 1, p. e6, 2015.
[49] T. L. Bailey and C. Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers," Proc Int Conf Intell Syst Mol Biol, vol. 2, pp. 28-36, 1994.
[50] G. Pavesi, G. Mauri, and G. Pesole, "An algorithm for finding signals of unknown length in DNA sequences," Bioinformatics, vol. 17 Suppl 1, pp. S207-14, 2001.
[51] G. Pavesi, P. Mereghetti, G. Mauri, and G. Pesole, "Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes," Nucleic Acids Res, vol. 32, no. Web Server issue, pp. W199-203, 2004.
[52] F. Fauteux, M. Blanchette, and M. V. Stromvik, "Seeder: discriminative seeding DNA motif discovery," Bioinformatics, vol. 24, no. 20, pp. 2303-7, 2008.
[53] X. Liu, D. L. Brutlag, and J. S. Liu, "BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes," Pac Symp Biocomput, pp. 127-38, 2001.
[54] L. Ettwiller, B. Paten, M. Ramialison, E. Birney, and J. Wittbrodt, "Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation," Nat Methods, vol. 4, no. 7, pp. 563-5, 2007.
[55] I. V. Kulakovskiy, V. A. Boeva, A. V. Favorov, and V. J. Makeev, "Deep and wide digging for binding motifs in ChIP-Seq data," Bioinformatics, vol. 26, no. 20, pp. 2622-3, 2010.
[56] M. Hu, J. Yu, J. M. Taylor, A. M. Chinnaiyan, and Z. S. Qin, "On the detection and refinement of transcription factor binding sites using ChIP-Seq data," Nucleic Acids Res, vol. 38, no. 7, pp. 2154-67, 2010.
[57] M. J. Mason, K. Plath, and Q. Zhou, "Identification of context-dependent motifs by contrasting ChIP binding data," Bioinformatics, vol. 26, no. 22, pp. 2826-32, 2010.
[58] J. E. Reid and L. Wernisch, "STEME: efficient EM to find motifs in large data sets," Nucleic Acids Res, vol. 39, no. 18, p. e126, 2011.
[59] T. L. Bailey, "DREME: motif discovery in transcription factor ChIP-seq data," Bioinformatics, vol. 27, no. 12, pp. 1653-9, 2011.
[60] P. Machanick and T. L. Bailey, "MEME-ChIP: motif analysis of large DNA datasets," Bioinformatics, vol. 27, no. 12, pp. 1696-7, 2011.
[61] V. Boeva *et al.*, "De novo motif identification improves the accuracy of predicting transcription factor binding sites in ChIP-Seq data analysis," Nucleic Acids Res, vol. 38, no. 11, p. e126, 2010.
[62] P. Huggins *et al.*, "DECOD: fast and accurate discriminative DNA motif finding," Bioinformatics, vol. 27, no. 17, pp. 2361-7, 2011.
[63] M. Thomas-Chollier, C. Herrmann, M. Defrance, O. Sand, D. Thieffry, and J. van Helden, "RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets," Nucleic Acids Res, vol. 40, no. 4, p. e31, 2012.
[64] X. Ma, A. Kulkarni, Z. Zhang, Z. Xuan, R. Serfling, and M. Q. Zhang, "A highly efficient and effective motif discovery method for ChIP-seq/ChIP-chip data using positional information," Nucleic Acids Res, vol. 40, no. 7, p. e50, 2012.

[65] H. Hartmann, E. W. Guthohrlein, M. Siebert, S. Luehr, and J. Soding, "P-value-based regulatory motif discovery using positional weight matrices," Genome Res, vol. 23, no. 1, pp. 181-94, 2013.
[66] D. Quang and X. Xie, "EXTREME: an online EM algorithm for motif discovery," Bioinformatics, vol. 30, no. 12, pp. 1667-73, 2014.
[67] N. Colombo and N. Vlassis, "FastMotif: spectral sequence motif discovery," Bioinformatics, vol. 31, no. 16, pp. 2623-31, 2015.
[68] S. Heinz *et al.*, "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities," Mol Cell, vol. 38, no. 4, pp. 576-89, 2010.
[69] S. Sinha and M. Tompa, "YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation," Nucleic Acids Res, vol. 31, no. 13, pp. 3586-8, 2003.
[70] M. B. Gerstein *et al.*, "Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project," Science, vol. 330, no. 6012, pp. 1775-87, 2010.
[71] G. Chen and Q. Zhou, "Searching ChIP-seq genomic islands for combinatorial regulatory codes in mouse embryonic stem cells," BMC Genomics, vol. 12, p. 515, 2011.
[72] N. Negre *et al.*, "A cis-regulatory map of the Drosophila genome," Nature, vol. 471, no. 7339, pp. 527-31, 2011.
[73] T. Whitington, M. C. Frith, J. Johnson, and T. L. Bailey, "Inferring transcription factor complexes from ChIP-seq data," Nucleic Acids Res, vol. 39, no. 15, p. e98, 2011.
[74] S. Zhang, S. Li, M. Niu, P. T. Pham, and Z. Su, "MotifClick: prediction of cis-regulatory binding sites via merging cliques," BMC bioinformatics, vol. 12, no. 1, p. 238, 2011.
[75] T. L. Bailey and P. Machanick, "Inferring direct DNA binding from ChIP-seq," Nucleic Acids Res, vol. 40, no. 17, p. e128, 2012.
[76] H. Sun, T. Guns, A. C. Fierro, L. Thorrez, S. Nijssen, and K. Marchal, "Unveiling combinatorial regulation through the combination of ChIP information and in silico cis-regulatory module detection," Nucleic Acids Res, vol. 40, no. 12, p. e90, 2012.
[77] N. Ha, M. Polychronidou, and I. Lohmann, "COPS: detecting co-occurrence and spatial arrangement of transcription factor binding motifs in genome-wide datasets," PloS one, vol. 7, no. 12, p. e52055, 2012.
[78] C. O. Rohr, R. G. Parra, P. Yankilevich, and C. Perez-Castro, "INSECT: IN-silico SEarch for Co-occurring Transcription factors," Bioinformatics, vol. 29, no. 22, pp. 2852-8, 2013.
[79] P. Jiang and M. Singh, "CCAT: Combinatorial Code Analysis Tool for transcriptional regulation," Nucleic Acids Res, vol. 42, no. 5, pp. 2833-47, 2014.
[80] A. J. Muller-Molina, H. R. Scholer, and M. J. Arauzo-Bravo, "Comprehensive human transcription factor binding site map for combinatory binding motifs discovery," PLoS One, vol. 7, no. 11, p. e49086, 2012.
[81] A. Vandenbon, Y. Kumagai, S. Akira, and D. M. Standley, "A novel unbiased measure for motif co-occurrence predicts combinatorial regulation of transcription," BMC Genomics, vol. 13 Suppl 7, p. S11, 2012.
[82] A. Mathelier and W. W. Wasserman, "The next generation of transcription factor binding site prediction," PLoS Comput Biol, vol. 9, no. 9, p. e1003214, 2013.
[83] E. Wingender *et al.*, "The TRANSFAC system on gene expression regulation," Nucleic Acids Res., vol. 29, pp. 281-283, 2001.
[84] D. Vlieghe *et al.*, "A new generation of JASPAR, the open-access repository for transcription factor binding site profiles," Nucleic Acids Res, vol. 34, no. Database issue, pp. D95-7, 2006.

[85] K. J. Won, S. Agarwal, L. Shen, R. Shoemaker, B. Ren, and W. Wang, "An integrated approach to identifying cis-regulatory modules in the human genome," PLoS One, vol. 4, no. 5, p. e5501, 2009.
[86] K. J. Won, I. Chepelev, B. Ren, and W. Wang, "Prediction of regulatory elements in mammalian genomes using chromatin signatures," BMC Bioinformatics, vol. 9, p. 547, 2008.
[87] J. Ernst and M. Kellis, "ChromHMM: automating chromatin-state discovery and characterization," Nature methods, vol. 9, no. 3, pp. 215-6, 2012.
[88] J. Ernst and M. Kellis, "Discovery and characterization of chromatin states for systematic annotation of the human genome," Nat Biotechnol, vol. 28, no. 8, pp. 817-25, 2010.
[89] J. Ernst *et al.*, "Mapping and analysis of chromatin state dynamics in nine human cell types," Nature, vol. 473, no. 7345, pp. 43-9, 2011.
[90] M. M. Hoffman, O. J. Buske, J. Wang, Z. Weng, J. A. Bilmes, and W. S. Noble, "Unsupervised pattern discovery in human chromatin structure through genomic segmentation," Nat Methods, vol. 9, no. 5, pp. 473-6, 2012.
[91] M. M. Hoffman *et al.*, "Integrative annotation of chromatin elements from ENCODE data," Nucleic Acids Res, vol. 41, no. 2, pp. 827-41, 2013.
[92] H. A. Firpi, D. Ucar, and K. Tan, "Discover regulatory DNA elements using chromatin signatures and artificial neural network," Bioinformatics, vol. 26, no. 13, pp. 1579-86, 2010.
[93] N. Rajagopal *et al.*, "RFECS: a random-forest based algorithm for enhancer identification from chromatin state," PLoS Comput Biol, vol. 9, no. 3, p. e1002968, 2013.
[94] M. C. Villarroel *et al.*, "Personalizing cancer treatment in the age of global genomic analyses: PALB2 gene mutations and the response to DNA damaging agents in pancreatic cancer," Mol Cancer Ther, vol. 10, no. 1, pp. 3-8, 2011.
[95] M. Ghandi, D. Lee, M. Mohammad-Noori, and M. A. Beer, "Enhanced regulatory sequence prediction using gapped k-mer features," PLoS Comput Biol, vol. 10, no. 7, p. e1003711, 2014.
[96] D. R. Zerbino, S. P. Wilder, N. Johnson, T. Juettemann, and P. R. Flicek, "The ensembl regulatory build," Genome Biol., vol. 16:56., no. doi, p. 56, 2015.
[97] H. Ashoor, D. Kleftogiannis, A. Radovanovic, and V. B. Bajic, "DENdb: database of integrated human enhancers," Database : the journal of biological databases and curation, vol. 2015, 2015.
[98] S. Fishilevich *et al.*, "GeneHancer: genome-wide integration of enhancers and target genes in GeneCards," Database (Oxford), vol. 2017, 2017.
[99] C. Chen *et al.*, "SEA version 3.0: a comprehensive extension and update of the Super-Enhancer archive," Nucleic Acids Res, vol. 48, no. D1, pp. D198-d203, 2020.
[100] R. Kang *et al.*, "EnhancerDB: a resource of transcriptional regulation in the context of enhancers," Database : the journal of biological databases and curation, vol. 2019, 2019.
[101] G. Zhang *et al.*, "DiseaseEnhancer: a resource of human disease-associated enhancer catalog," Nucleic Acids Res, vol. 46, no. D1, pp. D78-d84, 2018.
[102] T. Gao, B. He, S. Liu, H. Zhu, K. Tan, and J. Qian, "EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types," Bioinformatics, vol. 32, no. 23, pp. 3543-3551, 2016.
[103] T. Gao and J. Qian, "EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species," Nucleic Acids Res, vol. 48, no. D1, pp. D58-d64, 2020.
[104] J. Cheneby, M. Gheorghe, M. Artufel, A. Mathelier, and B. Ballester, "ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments," Nucleic Acids Res, vol. 46, no. D1, pp. D267-d275, 2018.

[105] J. E. Moore *et al.*, "Expanded encyclopaedias of DNA elements in the human and mouse genomes," Nature, vol. 583, no. 7818, pp. 699-710, 2020.
[106] J. C. Kwasnieski, C. Fiore, H. G. Chaudhari, and B. A. Cohen, "High-throughput functional testing of ENCODE segmentation predictions," Genome Res, vol. 24, no. 10, pp. 1595-602, 2014.
[107] N. Dogan *et al.*, "Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility," Epigenetics Chromatin, vol. 8, p. 16, 2015.
[108] R. R. Catarino and A. Stark, "Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation," Genes Dev, vol. 32, no. 3-4, pp. 202-223, 2018.
[109] H. Arbel *et al.*, "Exploiting regulatory heterogeneity to systematically identify enhancers with high accuracy," Proc Natl Acad Sci U S A, vol. 116, no. 3, pp. 900-908, 2019.
[110] D. Wang *et al.*, "Comprehensive functional genomic resource and integrative model for the human brain," Science, vol. 362, no. 6420, 2018.
[111] A. E. Handel, G. Gallone, M. Zameel Cader, and C. P. Ponting, "Most brain disease-associated and eQTL haplotypes are not located within transcription factor DNase-seq footprints in brain," Hum Mol Genet, vol. 26, no. 1, pp. 79-89, 2017.
[112] M. Niu, E. Tabari, P. Ni, and Z. Su, "Towards a map of cis-regulatory sequences in the human genome," Nucleic Acids Res, vol. 46, no. 11, pp. 5395-5409, 2018.
[113] M. Niu, E. S. Tabari, and Z. Su, "De novo prediction of cis-regulatory elements and modules through integrative analysis of a large number of ChIP datasets," BMC Genomics., vol. 15, no. 1, p. 1047, 2014.
[114] K. Y. Yip *et al.*, "Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors," Genome biology, vol. 13, no. 9, p. R48, 2012.
[115] P. Kheradpour and M. Kellis, "Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments," Nucleic Acids Res, vol. 42, no. 5, pp. 2976-87, 2014.
[116] J. E. Allen, M. Pertea, and S. L. Salzberg, "Computational gene prediction using multiple sources of evidence," Genome Res, vol. 14, no. 1, pp. 142-8, 2004.
[117] Y. Li, P. Ni, S. Zhang, G. Li, and Z. Su, "ProSampler: an ultra-fast and accurate motif finder in large ChIP-seq datasets for combinatory motif discovery," Bioinformatics, 2019.
[118] A. Visel, S. Minovitsky, I. Dubchak, and L. A. Pennacchio, "VISTA Enhancer Browser--a database of tissue-specific human enhancers," Nucleic Acids Res, vol. 35, no. Database issue, pp. D88-92, 2007.
[119] F. Consortium *et al.*, "A promoter-level mammalian expression atlas," Nature, vol. 507, no. 7493, pp. 462-70, 2014.
[120] R. Andersson *et al.*, "An atlas of active enhancers across human cell types and tissues," Nature, vol. 507, no. 7493, pp. 455-461, 2014.
[121] M. J. Landrum *et al.*, "ClinVar: public archive of interpretations of clinically relevant variants," Nucleic Acids Res, vol. 44, no. D1, pp. D862-8, 2016.
[122] J. MacArthur *et al.*, "The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)," Nucleic Acids Res, vol. 45, no. D1, pp. D896-D901, 2017.
[123] A. Buniello *et al.*, "The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019," Nucleic Acids Res, vol. 47, no. D1, pp. D1005-D1012, 2019.
[124] M. T. Weirauch *et al.*, "Determination and inference of eukaryotic transcription factor sequence specificity," Cell, vol. 158, no. 6, pp. 1431-1443, 2014.

[125] S. Zhang, M. Xu, S. Li, and Z. Su, "Genome-wide de novo prediction of cis-regulatory binding sites in prokaryotes," Nucleic Acids Res, vol. 37, no. 10, p. e72, 2009.
[126] S. van Dongen and C. Abreu-Goodger, "Using MCL to extract clusters from networks," Methods in molecular biology (Clifton, N.J.), vol. 804, pp. 281-95, 2012.
[127] D. N. Arnosti and M. M. Kulkarni, "Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards?," J Cell Biochem., vol. 94, no. 5, pp. 890-8., 2005.
[128] J. O. Yanez-Cuna, E. Z. Kvon, and A. Stark, "Deciphering the transcriptional cis-regulatory code," Trends Genet., vol. 29, no. 1, pp. 11-22, 2013.
[129] C. M. Vockley, I. C. McDowell, A. M. D'Ippolito, and T. E. Reddy, "A long-range flexible billboard model of gene activation," Transcription, vol. 8, no. 4, pp. 261-267, 2017.
[130] E. H. Davidson *et al.*, "A genomic regulatory network for development," Science, vol. 295, no. 5560, pp. 1669-78, 2002.
[131] E. H. Davidson, *The Regulatory Genome: Gene Regulatory Networks In Development And Evolution*. Academic Press, 2006.
[132] J. Zuin *et al.*, "Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells," Proceedings of the National Academy of Sciences, vol. 111, no. 3, pp. 996-1001, 2014.
[133] S. Orchard *et al.*, "The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases," Nucleic acids research, vol. 42, no. D1, pp. D358-D363, 2014.
[134] R. Oughtred *et al.*, "The BioGRID interaction database: 2019 update," Nucleic acids research, vol. 47, no. D1, pp. D529-D541, 2019.
[135] E. Labbé, A. Letamendia, and L. Attisano, "Association of Smads with lymphoid enhancer binding factor 1/T cell-specific factor mediates cooperative signaling by the transforming growth factor-β and Wnt pathways," Proceedings of the National Academy of Sciences, vol. 97, no. 15, pp. 8358-8363, 2000.
[136] N. Martin *et al.*, "Physical and functional interaction between PML and TBX2 in the establishment of cellular senescence," The EMBO journal, vol. 31, no. 1, pp. 95-109, 2012.
[137] X. TPLi *et al.*, "Proteomic analyses reveal distinct chromatin‐associated and soluble transcription factor complexes," Molecular systems biology, vol. 11, no. 1, 2015.
[138] M. J. Landrum and B. L. Kattman, "ClinVar at five years: delivering on the promise," Human mutation, vol. 39, no. 11, pp. 1623-1630, 2018.
[139] J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf, "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position," Nature methods, vol. 10, no. 12, pp. 1213-8, 2013.
[140] M. P. Creyghton *et al.*, "Histone H3K27ac separates active from poised enhancers and predicts developmental state," Proc Natl Acad Sci U S A, vol. 107, no. 50, pp. 21931-6, 2010.
[141] A. W. Aday, L. J. Zhu, A. Lakshmanan, J. Wang, and N. D. Lawson, "Identification of cis regulatory features in the embryonic zebrafish genome through large-scale profiling of H3K4me1 and H3K4me3 binding sites," Developmental biology, vol. 357, no. 2, pp. 450-62, 2011.
[142] I. V. Kulakovskiy *et al.*, "HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis," Nucleic Acids Res, vol. 46, no. D1, pp. D252-d259, 2018.
[143] A. Mathelier *et al.*, "JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles," Nucleic Acids Res, vol. 44, no. D1, pp. D110-5, 2016.

[144] S. A. Lambert *et al.*, "The Human Transcription Factors," Cell, vol. 175, no. 2, pp. 598-599, 2018.
[145] G. Ambrosini *et al.*, "Insights gained from a comprehensive all-against-all transcription factor binding motif benchmarking study," Genome Biol, vol. 21, no. 1, p. 114, 2020.
[146] C. Y. Perrot, C. Gilbert, V. Marsaud, A. Postigo, D. Javelaud, and A. Mauviel, "GLI 2 cooperates with ZEB 1 for transcriptional repression of CDH1 expression in human melanoma cells," Pigment cell & melanoma research, vol. 26, no. 6, pp. 861-873, 2013.
[147] Y. Koyabu, K. Nakata, K. Mizugishi, J. Aruga, and K. Mikoshiba, "Physical and functional interactions between Zic and Gli proteins," J Biol Chem, vol. 276, no. 10, pp. 6889-6892, 2001.
[148] E. Sánchez-Tilló, O. De Barrios, E. Valls, D. Darling, A. Castells, and A. Postigo, "ZEB1 and TCF4 reciprocally modulate their transcriptional activities to regulate Wnt target gene expression," Oncogene, vol. 34, no. 46, pp. 5760-5770, 2015.
[149] M. A. Mendoza-Parra, W. Van Gool, M. A. Mohamed Saleem, D. G. Ceschin, and H. Gronemeyer, "A quality control system for profiles obtained by ChIP sequencing," Nucleic Acids Res, vol. 41, no. 21, p. e196, 2013.
[150] G. K. Marinov, A. Kundaje, P. J. Park, and B. J. Wold, "Large-scale quality analysis of published ChIP-seq data," G3 (Bethesda). vol. 4, no. 2, pp. 209-23, 2014.
[151] G. Devailly, A. Mantsoki, T. Michoel, and A. Joshi, "Variable reproducibility in genome-scale public data: A case study using ENCODE ChIP sequencing resource," FEBS Lett, vol. 589, no. 24 Pt B, pp. 3866-70, 2015.
[152] G. M. Cooper *et al.*, "Single-nucleotide evolutionary constraint scores highlight disease-causing mutations," Nat Methods., vol. 7, no. 4, pp. 250-1, 2010.
[153] K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, and A. Siepel, "Detection of nonneutral substitution rates on mammalian phylogenies," Genome Res, vol. 20, no. 1, pp. 110-21, 2010.
[154] A. Visel *et al.*, "Ultraconservation identifies a small subset of extremely constrained developmental enhancers," Nat Genet, vol. 40, no. 2, pp. 158-60, 2008.
[155] G. Bejerano *et al.*, "Ultraconserved elements in the human genome," Science, vol. 304, no. 5675, pp. 1321-5, 2004.
[156] S. Katzman *et al.*, "Human genome ultraconserved elements are ultraselected," Science, vol. 317, no. 5840, p. 915, 2007.
[157] M. Gasperini, J. M. Tome, and J. Shendure, "Towards a comprehensive catalogue of validated and target-linked human enhancers," Nat Rev Genet, vol. 21, no. 5, pp. 292-310, 2020.
[158] M. J. Landrum *et al.*, "ClinVar: improving access to variant interpretations and supporting evidence," Nucleic Acids Res, vol. 46, no. D1, pp. D1062-d1067, 2018.
[159] A. R. Forrest *et al.*, "A promoter-level mammalian expression atlas," Nature, vol. 507, no. 7493, pp. 462-70, 2014.
[160] R. Andersson *et al.*, "An atlas of active enhancers across human cell types and tissues," Nature, vol. 507, no. 7493, pp. 455-61, 2014.
[161] R. E. Thurman *et al.*, "The accessible chromatin landscape of the human genome," Nature, vol. 489, no. 7414, pp. 75-82, 2012.
[162] D. Villar *et al.*, "Enhancer evolution across 20 mammalian species," Cell, vol. 160, no. 3, pp. 554-66, 2015.
[163] R. S. Young, Y. Kumar, W. A. Bickmore, and M. S. Taylor, "Bidirectional transcription initiation marks accessible chromatin and is not specific to enhancers," Genome Biol, vol. 18, no. 1, p. 242, 2017.
[164] R. V. Chereji, P. R. Eriksson, J. Ocampo, H. K. Prajapati, and D. J. Clark, "Accessibility of promoter DNA is not the primary determinant of chromatin-mediated gene regulation," Genome Res, vol. 29, no. 12, pp. 1985-1995, 2019.

[165] B. E. Bernstein, E. Birney, I. Dunham, E. D. Green, C. Gunter, and M. Snyder, "An integrated encyclopedia of DNA elements in the human genome," Nature, vol. 489, no. 7414, pp. 57-74, 2012.
[166] A. Khan and X. Zhang, "dbSUPER: a database of super-enhancers in mouse and human genome," Nucleic Acids Res, 2015.
[167] R. Dreos, G. Ambrosini, R. Cavin Perier, and P. Bucher, "EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era," Nucleic Acids Res, vol. 41, no. Database issue, pp. D157-64, 2013.
[168] S. Dimitrieva and P. Bucher, "UCNEbase--a database of ultraconserved non-coding elements and genomic regulatory blocks," Nucleic Acids Res, vol. 41, no. Database issue, pp. D101-9, 2013.
[169] A. Wilderman, J. VanOudenhove, J. Kron, J. P. Noonan, and J. Cotney, "High-Resolution Epigenomic Atlas of Human Embryonic Craniofacial Development," Cell Rep, vol. 23, no. 5, pp. 1581-1597, 2018.
[170] L. A. Pennacchio, W. Bickmore, A. Dean, M. A. Nobrega, and G. Bejerano, "Enhancers: five essential questions," Nat Rev Genet, vol. 14, no. 4, pp. 288-95, 2013.
[171] M. Kellis *et al.*, "Defining functional DNA elements in the human genome," Proc Natl Acad Sci U S A., vol. 111, no. 17, pp. 6131-8, 2014.
[172] M. P. Snyder *et al.*, "Perspectives on ENCODE," Nature, vol. 583, no. 7818, pp. 693-698, 2020.
[173] X. Wang *et al.*, "High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human," Nat Commun, vol. 9, no. 1, p. 5380, 2018.
[174] V. A. Schneider *et al.*, "Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly," Genome Res, vol. 27, no. 5, pp. 849-864, 2017.
[175] J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, and N. M. Luscombe, "A census of human transcription factors: function, expression and evolution," Nat Rev Genet, vol. 10, no. 4, pp. 252-63, 2009.
[176] A. Jolma *et al.*, "DNA-binding specificities of human transcription factors," Cell, vol. 152, no. 1-2, pp. 327-39, 2013.
[177] J. A. Stamatoyannopoulos, "What does our genome encode?," Genome Res, vol. 22, no. 9, pp. 1602-11, 2012.
[178] D. C. King, J. Taylor, L. Elnitski, F. Chiaromonte, W. Miller, and R. C. Hardison, "Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences," Genome Res, vol. 15, no. 8, pp. 1051-60, 2005.
[179] D. N. Arnosti and M. M. Kulkarni, "Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards?," Journal of cellular biochemistry, vol. 94, no. 5, pp. 890-898, 2005.
[180] M. M. Kulkarni and D. N. Arnosti, "Information display by transcriptional enhancers," Development (Cambridge, England), vol. 130, no. 26, pp. 6569-6575, 2003.
[181] D. Panne, T. Maniatis, and S. C. Harrison, "An atomic model of the interferon-β enhanceosome," Cell, vol. 129, no. 6, pp. 1111-1123, 2007.
[182] G. Junion *et al.*, "A transcription factor collective defines cardiac cell fate and reflects lineage history," Cell, vol. 148, no. 3, pp. 473-86, 2012.
[183] S. MacArthur *et al.*, "Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions," Genome biology, vol. 10, no. 7, p. R80, 2009.
[184] C. Moorman *et al.*, "Hotspots of transcription factor colocalization in the genome of Drosophila melanogaster," Proceedings of the National Academy of Sciences, vol. 103, no. 32, pp. 12027-12032, 2006.

[185] H. K. Long, S. L. Prescott, and J. Wysocka, "Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution," Cell, vol. 167, no. 5, pp. 1170-1187, 2016.
[186] A. D. Yates *et al.*, "Ensembl 2020," Nucleic acids research, vol. 48, no. D1, pp. D682-D688, 2020.
[187] A. R. Quinlan and I. M. Hall, "BEDTools: a flexible suite of utilities for comparing genomic features," Bioinformatics, vol. 26, no. 6, pp. 841-2, 2010.
[188] R. I. Kamar *et al.*, "Facilitated dissociation of transcription factors from single DNA binding sites," Proceedings of the National Academy of Sciences, vol. 114, no. 16, pp. E3251-E3257, 2017.
[189] D. Panne, T. Maniatis, and S. C. Harrison, "Crystal structure of ATF-2/c-Jun and IRF-3 bound to the interferon-β enhancer," The EMBO journal, vol. 23, no. 22, pp. 4384-4393, 2004.
[190] M. Safran *et al.*, "GeneCards Version 3: the human gene integrator," Database, vol. 2010, 2010.
[191] M. D. Biggin, "Animal transcription networks as highly connected, quantitative continua," Dev Cell, vol. 21, no. 4, pp. 611-26, 2011.
[192] X. Y. Li, S. Thomas, P. J. Sabo, M. B. Eisen, J. A. Stamatoyannopoulos, and M. D. Biggin, "The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding," Genome biology, vol. 12, no. 4, p. R34, 2011.
[193] J. Thurmond *et al.*, "FlyBase 2.0: the next generation," Nucleic acids research, vol. 47, no. D1, pp. D759-D765, 2019.
[194] D. Villar, P. Flicek, and D. T. Odom, "Evolution of transcription factor binding in metazoans - mechanisms and functional implications," Nat Rev Genet, vol. 15, no. 4, pp. 221-33, 2014.
[195] I. Brouwer and T. L. Lenstra, "Visualizing transcription: key to understanding gene expression dynamics," Current opinion in chemical biology, vol. 51, pp. 122-129, 2019.
[196] S. R. Grossman *et al.*, "Systematic dissection of genomic features determining transcription factor binding and enhancer function," Proc Natl Acad Sci U S A, vol. 114, no. 7, pp. E1291-e1300, 2017.
[197] B. D. Strahl and C. D. Allis, "The language of covalent histone modifications," Nature, vol. 403, no. 6765, pp. 41-5, 2000.
[198] R. M. Rodriguez *et al.*, "Epigenetic Networks Regulate the Transcriptional Program in Memory and Terminally Differentiated CD8+ T Cells," Journal of immunology (Baltimore, Md. : 1950), vol. 198, no. 2, pp. 937-949, 2017.
[199] B. E. Russ *et al.*, "Distinct epigenetic signatures delineate transcriptional programs during virus-specific CD8(+) T cell differentiation," Immunity, vol. 41, no. 5, pp. 853-65, 2014.
[200] T. Juelich *et al.*, "Interplay between Chromatin Remodeling and Epigenetic Changes during Lineage-Specific Commitment to Granzyme B Expression," Journal of Immunology, vol. 183, no. 11, pp. 7063-7072, 2009.
[201] J. Zhu *et al.*, "Genome-wide chromatin state transitions associated with developmental and environmental cues," Cell, vol. 152, no. 3, pp. 642-54, 2013.
[202] J. P. Thomson *et al.*, "CpG islands influence chromatin structure via the CpG-binding protein Cfp1," Nature, vol. 464, no. 7291, pp. 1082-6, 2010.
[203] D. Benveniste, H. J. Sonntag, G. Sanguinetti, and D. Sproul, "Transcription factor binding predicts histone modifications in human cell lines," Proc Natl Acad Sci U S A, vol. 111, no. 37, pp. 13367-72, 2014.
[204] J. Z. Liu *et al.*, "Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations," Nat Genet, vol. 47, no. 9, pp. 979-986, 2015.

[205] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning," Nat Biotechnol, vol. 33, no. 8, pp. 831-8, 2015.
[206] J. Zhou and O. G. Troyanskaya, "Predicting effects of noncoding variants with deep learning-based sequence model," Nature methods, vol. 12, no. 10, pp. 931-934, 2015.
[207] H. Zeng and D. K. Gifford, "Predicting the impact of non-coding variants on DNA methylation," Nucleic Acids Res, vol. 45, no. 11, p. e99, 2017.
[208] D. R. Kelley, J. Snoek, and J. L. Rinn, "Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks," Genome Res, vol. 26, no. 7, pp. 990-9, 2016.
[209] P. Durek *et al.*, "Epigenomic Profiling of Human CD4(+) T Cells Supports a Linear Differentiation Model and Highlights Molecular Regulators of Memory Development," Immunity, vol. 45, no. 5, pp. 1148-1161, 2016.
[210] C. Roadmap Epigenomics *et al.*, "Integrative analysis of 111 reference human epigenomes," Nature, vol. 518, no. 7539, pp. 317-30, 2015.
[211] Y. Zhang *et al.*, "Model-based analysis of ChIP-Seq (MACS)," Genome biology, vol. 9, no. 9, p. R137, 2008.
[212] S. G. Landt *et al.*, "ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia," Genome research, vol. 22, no. 9, pp. 1813-1831, 2012.
[213] P. V. Kharchenko, M. Y. Tolstorukov, and P. J. Park, "Design and analysis of ChIP-seq experiments for DNA-binding proteins," Nat Biotechnol, vol. 26, no. 12, pp. 1351-9, 2008.
[214] T. E. P. Consortium, "An integrated encyclopedia of DNA elements in the human genome," Nature, vol. 489, no. 7414, pp. 57-74, 2012.
[215] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," P Ieee, vol. 86, no. 11, pp. 2278-2324, 1998.
[216] V. Nair, G. H.-P. o. t. t. i. E. H.-P. o. t. t. i. Conference, U. , and G. E. H.-P. o. t. t. international Conference, "Rectified linear units improve restricted boltzmann machines," cs.toronto.edu,
[217] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," ed: JMLR.org, 2015, pp. 448-456.
[218] T. D. Team, "Theano: A Python framework for fast computation of mathematical expressions," arXiv e-prints, vol. abs/1605.0, 2016.
[219] S. Dieleman *et al.*, "Lasagne: First release," ed, 2015.
[220] G. Hinton, N. Srivastava, and K. Swersky, "Neural networks for machine learning lecture 6a overview of mini-batch gradient descent,"
[221] C. E. Grant, T. L. Bailey, and W. S. Noble, "FIMO: scanning for occurrences of a given motif," Bioinformatics, vol. 27, no. 7, pp. 1017-8, 2011.
[222] T. L. Bailey *et al.*, "MEME SUITE: tools for motif discovery and searching," Nucleic Acids Res, vol. 37, no. Web Server issue, pp. W202-8, 2009.
[223] A. Siepel *et al.*, "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes," Genome Res, vol. 15, no. 8, pp. 1034-50, 2005.
[224] S. Gupta, J. A. Stamatoyannopoulos, T. L. Bailey, and W. S. Noble, "Quantifying similarity between motifs," Genome biology, vol. 8, no. 2, p. 24, 2007.
[225] A. Hagberg, P. Swart, and D. S Chult, "Exploring network structure, dynamics, and function using NetworkX," Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.
[226] Y. Li, P. Ni, S. Zhang, G. Li, and Z. Su, "ProSampler: an ultrafast and accurate motif finder in large ChIP-seq datasets for combinatory motif discovery," Bioinformatics, 2019.

[227] J. Ernst and M. Kellis, "Interplay between chromatin state, regulator binding, and regulatory motifs in six human cell types," Genome Research, vol. 23, no. 7, pp. 1142-1154, 2013.
[228] R. Li, H. Pei, D. K. Watson, and T. S. Papas, "EAP1/Daxx interacts with ETS1 and represses transcriptional activation of ETS1 target genes," Oncogene, vol. 19, no. 6, pp. 745-53, 2000.
[229] C. Wasylyk, S. E. Schlumberger, P. Criqui-Filipe, and B. Wasylyk, "Sp100 interacts with ETS-1 and stimulates its transcriptional activity," Mol Cell Biol, vol. 22, no. 8, pp. 2687-702, 2002.
[230] N. Feuerstein, R. Firestein, N. Aiyar, X. He, D. Murasko, and V. Cristofalo, "Late induction of CREB/ATF binding and a concomitant increase in cAMP levels in T and B lymphocytes stimulated via the antigen receptor," Journal of immunology (Baltimore, Md. : 1950), vol. 156, no. 12, pp. 4582-93, 1996.
[231] H. Kawasaki *et al.*, "ATF-2 has intrinsic histone acetyltransferase activity which is modulated by phosphorylation," Nature, vol. 405, no. 6783, pp. 195-200, 2000.
[232] W. F. Wong, K. Kohu, T. Chiba, T. Sato, and M. Satake, "Interplay of transcription factors in T-cell differentiation and function: the role of Runx," Immunology, vol. 132, no. 2, pp. 157-64, 2011.
[233] H. P. Kim, B. G. Kim, J. Letterio, and W. J. Leonard, "Smad-dependent cooperative regulation of interleukin 2 receptor alpha chain gene expression by T cell receptor and transforming growth factor-beta," J Biol Chem, vol. 280, no. 40, pp. 34042-7, 2005.
[234] N. Malhotra and J. Kang, "SMAD regulatory networks construct a balanced immune system," Immunology, vol. 139, no. 1, pp. 1-10, 2013.
[235] D. Valle-García *et al.*, "ATRX binds to atypical chromatin domains at the 3â€ ? exons of zinc finger genes to preserve H3K9me3 enrichment," Epigenetics, vol. 11, no. 6, pp. 398-414, 2016.
[236] S. M. Hedrick, R. Hess Michelini, A. L. Doedens, A. W. Goldrath, and E. L. Stone, "FOXO transcription factors throughout T cell biology," Nat Rev Immunol, vol. 12, no. 9, pp. 649-61, 2012.
[237] B. He *et al.*, "CD8(+) T Cells Utilize Highly Dynamic Enhancer Repertoires and Regulatory Circuitry in Response to Infections," Immunity, vol. 45, no. 6, pp. 1341-1354, 2016.
[238] J. G. Crompton *et al.*, "Lineage relationship of CD8(+) T cell subsets is revealed by progressive changes in the epigenetic landscape," Cellular & Molecular Immunology, vol. 13, no. 4, pp. 502-513, 2016.
[239] S. M. Henson, N. E. Riddell, and A. N. Akbar, "Properties of end-stage human T cells defined by CD45RA re-expression," Current opinion in immunology, vol. 24, no. 4, pp. 476-81, 2012.
[240] Z. Wang and H. F. Willard, "Evidence for sequence biases associated with patterns of histone methylation," BMC Genomics, vol. 13, no. 1, p. 367, 2012.
[241] J. W. Ho *et al.*, "Comparative analysis of metazoan chromatin organization," Nature, vol. 512, no. 7515, pp. 449-52, 2014.
[242] L. Lin and S. L. Peng, "Coordination of NF-κB and NFAT antagonism by the forkhead transcription factor Foxd1," The Journal of Immunology, vol. 176, no. 8, pp. 4793-4803, 2006.
[243] D. M. Moskowitz *et al.*, "Epigenomics of human CD8 T cell differentiation and aging," Sci Immunol, vol. 2, no. 8, p. 0192, 2017.
[244] S. F. Gilbert, *Developmental biology*, 6th ed. Sinauer Associates, 2000.
[245] J. S. Tushir and C. D'Souza-Schorey, "ARF6-dependent activation of ERK and Rac1 modulates epithelial tubule development," The EMBO journal, vol. 26, no. 7, pp. 1806-19, 2007.

[246] K. Ito *et al.*, "RUNX3 attenuates beta-catenin/T cell factors in intestinal tumorigenesis," Cancer Cell, vol. 14, no. 3, pp. 226-37, 2008.
[247] S. Morin, G. Pozzulo, L. Robitaille, J. Cross, and M. Nemer, "MEF2-dependent recruitment of the HAND1 transcription factor results in synergistic activation of target promoters," J Biol Chem, vol. 280, no. 37, pp. 32272-8, 2005.
[248] H. Y. Kang, K. E. Huang, S. Y. Chang, W. L. Ma, W. J. Lin, and C. Chang, "Differential modulation of androgen receptor-mediated transactivation by Smad3 and tumor suppressor Smad4," J Biol Chem, vol. 277, no. 46, pp. 43749-56, 2002.
[249] T. Shimamoto, S. Nakamura, J. Bollekens, F. H. Ruddle, and K. Takeshita, "Inhibition of DLX-7 homeobox gene causes decreased expression of GATA-1 and c-myc genes and apoptosis," Proc Natl Acad Sci U S A, vol. 94, no. 7, pp. 3245-9, 1997.
[250] S. Thuault, E. J. Tan, H. Peinado, A. Cano, C. H. Heldin, and A. Moustakas, "HMGA2 and Smads co-regulate SNAIL1 expression during induction of epithelial-to-mesenchymal transition," J Biol Chem, vol. 283, no. 48, pp. 33437-46, 2008.
[251] C. Angermueller, H. J. Lee, W. Reik, and O. Stegle, "DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning," Genome biology, vol. 18, no. 1, p. 67, 2017.
[252] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning Important Features Through Propagating Activation Differences," 2017.
[253] M. Wainberg, D. Merico, A. Delong, and B. J. Frey, "Deep learning in biomedicine," Nat Biotechnol, vol. 36, no. 9, pp. 829-838, 2018.
[254] A. A. Pai, J. K. Pritchard, and Y. Gilad, "The genetic and mechanistic basis for variation in gene regulation," PLoS genetics, vol. 11, no. 1, p. e1004857, 2015.
[255] D. Schmidt, M. D. Wilson, C. Spyrou, G. D. Brown, J. Hadfield, and D. T. Odom, "ChIP-seq: using high-throughput sequencing to discover protein–DNA interactions," Methods, vol. 48, no. 3, pp. 240-248, 2009.
[256] L. Song and G. E. Crawford, "DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells," Cold Spring Harbor Protocols, vol. 2010, no. 2, pp. pdb-prot5384, 2010.
[257] J. D. Buenrostro, B. Wu, H. Y. Chang, and W. J. Greenleaf, "ATAC-seq: a method for assaying chromatin accessibility genome-wide," Current protocols in molecular biology, vol. 109, no. 1, pp. 21-29, 2015.
[258] J. M. Simon, P. G. Giresi, I. J. Davis, and J. D. Lieb, "Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA," Nature protocols, vol. 7, no. 2, p. 256, 2012.
[259] D. E. Schones *et al.*, "Dynamic regulation of nucleosome positioning in the human genome," Cell, vol. 132, no. 5, pp. 887-898, 2008.
[260] E. P. Consortium, "The ENCODE (ENCyclopedia Of DNA Elements) Project," Science, vol. 306, no. 5696, pp. 636-40, 2004.
[261] Y. Jiang *et al.*, "SEdb: a comprehensive human super-enhancer database," Nucleic Acids Res, vol. 47, no. D1, pp. D235-d243, 2019.
[262] S. Dimitrieva and P. Bucher, "UCNEbase—a database of ultraconserved non-coding elements and genomic regulatory blocks," Nucleic acids research, vol. 41, no. D1, pp. D101-D109, 2013.
[263] A. Visel *et al.*, "A high-resolution enhancer atlas of the developing telencephalon," Cell, vol. 152, no. 4, pp. 895-908, 2013.
[264] J. Wang, X. Dai, L. D. Berry, J. D. Cogan, Q. Liu, and Y. Shyr, "HACER: an atlas of human active enhancers to interpret regulatory variants," Nucleic Acids Res, vol. 47, no. D1, pp. D106-d112, 2019.

[265] Z. Cai *et al.*, "RAEdb: a database of enhancers identified by high-throughput reporter assays," Database, vol. 2019, 2019.
[266] Z. Wang *et al.*, "HEDD: human enhancer disease database," Nucleic acids research, vol. 46, no. D1, pp. D113-D120, 2018.
[267] G. Zhang *et al.*, "DiseaseEnhancer: a resource of human disease-associated enhancer catalog," Nucleic acids research, vol. 46, no. D1, pp. D78-D84, 2018.
[268] Y. Wei *et al.*, "SEA: a super-enhancer archive," Nucleic acids research, vol. 44, no. D1, pp. D172-D179, 2016.
[269] L. W. Hillier, V. Reinke, P. Green, M. Hirst, M. A. Marra, and R. H. Waterston, "Massively parallel sequencing of the polyadenylated transcriptome of C. elegans," Genome Res, vol. 19, no. 4, pp. 657-66, 2009.

APPENDIX A: LINK OF SUPPLEMENTARY FILES

A.1. Supplementary files

This supplemental tables includes the information of all raw datasets and the unique motifs and their TF families

SUPPLEMENTARY_TABLES.xlsx