LANDSCAPE AND ARCHITECTURE OF *CIS*-REGULATORY MODULES AND PREDICTION OF THEIR FUNCTIONAL TYPES, STATES AND TARGET GENES

by

Sisi Yuan

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 and Computational Biology

Charlotte

2024

Approved by:

_____
Dr. Zhengchang Su

_____
Dr. Jun-tao Guo

_____
Dr. Abigail Leavitt LaBella

_____
Dr. Bao-Hua Song

ABSTRACT

SISI YUAN. Landscape and Architecture of *cis*-regulatory Modules and Prediction of Their Functional Types, States and Target Genes (Under the direction of DR. ZHENGCHANG SU)

Cis-regulatory modules (CRMs) can function as enhancers and/or silencers to promote and repress, respectively, the transcription of their target genes in a spatiotemporal manner, thereby playing critical roles in virtually all biological processes. However, despite recent progresses, the understanding of CRMs' precise locations, landscape and architecture in terms of transcription factor binding sites (TFBSs) in the genomes as well as their functional types (enhancer or silencer), states (active or inactive) and target genes in various cell/tissue types of organisms is still limited.

We have recently predicted comprehensive maps of CRMs and constituent TFBSs in the human and mouse genomes, enabling us to investigate the organization and architecture of the CRMs in both genomes. We reveal common rules of the organization and architecture of CRMs in the genomes. We conclude that the rules governing the organization and architecture of CRMs in the human and mouse genomes are highly conserved.

Moreover, until recently research has long been focused on enhancers, and much less is known about silencers. To fill the gap, we develop two logistic regression models for predicting the functional states of our previously predicted 1.2M CRMs as enhancers and silencers in any cell/tissue types using five epigenetic marks data. Applying the models to 56 human cell/tissue types with the required data available, we predict that 793,140 of the 1.2M CRMs are active as enhancers or/and silencers in at least one of these cell/tissue types, of which 14.8% and 28.6% of them only function as enhancers (enhancer-predominant) and silencers (silencer-predominant), respectively, while 10.6% functioned both as enhancers and silencers (dual functional). Thus, both dual functional CRMs and silencers might be more prevalent than previously assumed. Most dual functional CRMs function either as enhancers or silencers in different cell/tissue types (Type I),

while some have dual functions regulating different genes in the same cell/tissue types (Type II). Different types of CRMs display different lengths and TFBS densities, reflecting the complexity of their functions.

Furthermore, identifying their target genes of predicted or experimentally validated CRMs remains a challenge due to the low quality of the predicted CRMs and the fact that CRMs often do not regulate their closest genes. To fill this gap, we developed a method — correlation and physical proximity (CAPP) to not only predict the CRMs' target genes but also their functional types using only chromatin accessibility (CA) and RNA-seq data in a panel of cell/tissue types plus Hi-C data in a few cell types. Applying CAPP to a panel of 107 human cell/tissue types with CA and RNA-seq data available, we predict target genes for 20% of the 1.2M CRMs, of which 4.5% are predicted as both enhancers and silencers (dual functional CRMs), 95.2% as exclusive enhancers and 0.3% as exclusive silencers. Different types of CRMs as well as their target genes and regulatory links exhibit distinct properties. CAPP predicts more enhancer-gene and silencer-gene links with higher accuracy than state-of-the-art methods.

ACKNOWLEDGEMENTS

Words cannot express my gratitude to my advisor - Dr. Zhengchang Su for his invaluable guidance and instructions throughout this project. I would also like to extend my sincere appreciation to my committee members, Dr. Jun-tao Guo, Dr. Abigail Leavitt LaBella and Dr. Bao-Hua Song, for their insightful comments and suggestions, which greatly enriched the quality of this work. Special thanks to Dr. Pengyu Ni for his enlightening explanation of his prior research. Additionally, I would like to express my heartfelt thanks to other members of the Su Lab for their valuable discussions and contributions.

I would like to express my gratitude to the Encyclopedia of DNA Elements (ENCODE) project for providing the RNA-seq data, Hi-C data, epigenetic data, and transcription factor binding data. Additionally, I am thankful to the 4D Nucleome Program for providing the Hi-C interaction matrices, and to the VISTA Enhancer and FANTOM databases for providing experimentally validated *cis*-regulatory elements. I also acknowledge the contributions of the CISTROME project for sharing epigenetic data and transcription factor binding data. I thank the silencerDB database for providing validated and predicted silencers and their potential target genes.

I extend my sincere thanks to the Graduate Assistant Support Plan (GASP) of UNC Charlotte and the Graduate Assistantships of the Department of Bioinformatics and Genomics for their financial support and assistantship offers. Additionally, I would like to express my gratitude to all faculty and staff members in the Department of Bioinformatics and Genomics, the campus OneIT University Research Computing (URC) group, and the International Student and Scholar Office (ISSO) for their supportive assistance.

DEDICATION

I dedicate my dissertation work to my family and many friends. I would like to extend my heartfelt thanks to my friends: Qinglin Mei, Sampson Wang, Judy Zhu, and their lovely son Jason, Chris Wang, Qian Zhang, and their two lovely daughters Christie and Valerie for their help and companionship on this journey.

I cherish the memories of my grandmother, Meihua Wu, who passed away during my PhD studies. She raised me up and always hoped to witness my achievements, and her belief in my potential continues to inspire me.

A special feeling of gratitude goes to my loving mother, Xuannyu Lin, my father, Dequ Yuan, my stepfather, Deheng Yuan, my father-in-law, Jihua Liu, my mother-in-law, Fanqiu Zeng who passed away during my PhD studies and is always missed, and my grandmother-in-law, Jiayu Luo. Their words of encouragement and push for tenacity rang in my ears.

My husband, Shawn, my daughter, Chloe, and my son, Ian, have never left my side and have been incredibly supportive throughout this total seven-year journey, especially during the difficult times of the pandemic.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABC: Activity-by-Contact

aCRN: active *cis*-regulatory network

ATAC-seq: Assay for transposase-accessible chromatin using sequencing

AUROC: Area under the receiver operator characteristic curve

B-H: Benjamini-Hochberg

CA: Chromatin accessibility

CAPP: Correlation and Physical Proximity

cCRE: Candidate *cis*-regulatory element

ChIA-PET: Chromatin interaction analysis with paired-end-tag sequencing

ChIP-seq: Chromatin immunoprecipitation sequencing

CNA: Closest neighbor assignment

CoSVM: Correlation and SVM method

CPC: Core promoter-containing

CPL: Core promoter-lacking

CRISPR: Clustered regularly interspaced short palindromic repeats

CRISPRa: CRISPR activation

CRISPRi: CRISPR interference

CRM: *Cis*-regulatory modules

CRno: CRM-RE not overlapping

DHS: DNase I hypersensitive site

DNase-seq: DNase I hypersensitive sites sequencing

Enhancer-G: Enhancer-gene link

Enhancer-Gm: Enhancer-G link with matched gene

Enhancer-Gnm: Enhancer-G link with not matched gene

ERno-Co: Enhancer-RE not overlapping, but CRM-RE overlapping

ERno: Enhancer-RE not overlapping

ERo: Enhancer-RE overlapping

FDR: False discovery rate

gkmSVM: gapped k-mer SVM

GWAS: Genome-wide association studies

KR: Knight-Ruiz Matrix Balancing

LR: Logistic regression

MPRA: Massively parallel reporter assay

PECA: Paired expression and chromatin accessibility

RE-G: RE-gene link

RE-Gm: RE-G link with matched gene

RE-Gnm: RE-G link with not matched gene

RE: Regulatory element

ReSE: Repressive ability of silencer elements

sCRN: static *cis*-regulatory network

Silencer-G: Silencer-gene link

Silencer-Gm: Silencer-G link with matched gene

Silencer-Gnm: Silencer-G link with not matched gene

SNP: Single nucleotide polymorphism

SRno-Co: Silencer-RE not overlapping, but CRM-RE overlapping

SRno: Silencer-RE not overlapping

SRo: Silencer-RE overlapping

SSA: Simple subtractive approach

SVM: Support vector machine

TAD: Topologically associating domain

TF: Transcription factor

TFBS: Transcription factor binding site

TPM: Transcript per million

TSS: Transcription start site

VC: Vanilla-Coverage

Chapter 1
INTRODUCTION

Approximately 99.5% of genomes among human individuals are identical at the single nucleotide level(1), suggesting that the remaining 0.5% of genetic variation, mostly located in non-coding regions(2, 3), might account for the diverse phenotypes within the human populations. Consistently, genome-wide association studies (GWAS) have unveiled that nearly 90% of single nucleotide polymorphisms (SNPs) associated with complex diseases or phenotypes reside in non-coding regions(4). This underscores that disparities in phenotypes and disease susceptibilities across individuals primarily stem from variation within non-coding regions, particularly those that disrupt transcription factor (TF) binding sites (TFBSs) in *cis*-regulatory elements (CRMs) located predominantly in non-coding regions(5-8).

CRMs play critical roles in virtually all biological processes, from cell differentiation to physiological homeostasis, pathogenesis and evolution by regulating transcription of genes in various cell/tissue types, thereby rendering their types and functions(9, 10). Thus, specific binding of TFs to their cognate TFBSs within enhancers and silencers can facilitate and repress the recruitment of RNA polymerase to the promoters of target genes, resulting in upregulation and downregulation of gene transcription, respectively(11, 12), in a cell/tissue specific manner. Moreover, binding of CTCF TF to its cognate sites within insulators establishes boundaries between topologically associating domains (TADs), preventing cross-regulation of CRMs in a TAD(13). Cellular specificity of CRMs is largely established by unique epigenetic modifications of CRMs in different cell types by altering the accessibility and binding affinity of TFs to their cognate binding sites(14, 15). Therefore, to fully understand the functions of CRMs, one needs to not only precisely locate their locations in the genome and understand their landscape and architecture in terms of TFBSs, but also to characterize their functional types (mainly enhancers

and silencers), states (active or inactive) and target genes in various cell/tissue types of the organism. However, only a small portion of CRMs (mostly enhancers) in genomes have been fully characterized in all these three aspects due to the difficulty to study them using traditional molecular biology methods(16).

The landscape of CRMs and their architecture in terms of constituent TFBSs is a fundamental aspect of genomics and molecular biology. Understanding these elements is crucial for unraveling the intricacies of gene regulation, revealing how genes are turned on and off in different contexts, and how these processes contribute to various biological functions and disease mechanisms. Despite significant advancements in high-throughput techniques of chromatin immunoprecipitation sequencing (ChIP-seq), a comprehensive map of CRMs and TFBSs in genomes remains elusive. This limitation has hindered the exploration of the precise organization and distribution of CRMs across the genome. Furthermore, current methods often fail to accurately predict the TFBSs within CRMs, limiting our understanding of the mechanisms underlying transcriptional regulation. The lack of highly accurate and comprehensive CRM maps has led to contradictory findings in the literature regarding the distribution and organization of these elements. For instance, while some studies(17) suggest that candidate *cis*-regulatory elements (cCREs) tend to cluster in regulatory islands, others(18) propose that enhancers are more evenly dispersed across extensive genomic regions. Recent efforts, including our work(19-21), have made strides in predicting more comprehensive maps of CRMs and their constituent TFBSs using available TF ChIP-seq data from various cell and tissue types. These advancements position us to better analyze the landscape and organization of CRMs in genomes, uncovering the underlying rules of their organization and architecture.

CRMs, classified as enhancers, silencers, promoters and insulators based on their effects and roles in transcriptional regulation, often exhibit on-off switches in functional states across different cell/tissue types(9, 10). Particularly, specific TFs binding to cognate binding sites in an enhancer or a silencer can facilitate or prevent, respectively, the recruitment of RNA polymerase to the promoter of the target gene, thereby upregulating or downregulating transcription of the gene, respectively(11, 12). Epigenetic modifications of CRMs within diverse cell types can influence the accessibility and binding affinity of TFs to their cognate binding sites. Consequently, this dynamic interplay among TFs, TFBSs, RNA polymerase and epigenetic modification systems contributes to distinct spatiotemporal expression patterns of genes in various biological processes across different cell/tissue types(14, 15). The recent availability of enormous functional and epigenomic data in various cell/tissue types in well-studied organisms have provided an unprecedented opportunity to predict loci of enhancers and silencers in the genomes and their functional states in various cell/tissue types of the organisms. Most of these methods attempt to simultaneously predict the loci of CRMs and their functional states in a cell/tissue type using multiple epigenetic marks such as chromatin accessibility (CA) assayed by either DNase I hypersensitive sites sequencing (DNase-seq)(22-24) or assay for transposase-accessible chromatin using sequencing (ATAC-seq)(25), and histone marks assayed by ChIP-seq(26). Although conceptually appealing, these one-step methods result in high false discovery rate (FDR)(19, 27-32) since CA and histone marks, though informative, are not specific marks for active enhancers or silencers(29, 30, 33). Additionally, it has been shown that TF binding data are more informative for identifying loci of CRMs than epigenetic data(29-33), suggesting a two-step approach: first locating CRMs using TF data and then predicting functional types including silencers and states using epigenetic data.

In recent years, advancements in high-throughput techniques such as Hi-C(34) and chromatin interaction analysis with paired-end-tag sequencing (ChIA-PET)(35), along with clustered regularly interspaced short palindromic repeats (CRISPR) technologies such as CRISPR activation (CRISPRa) and CRISPR interference (CRISPRi) have provided insights into the regulatory relationships between enhancers, silencers, and their target genes across various cell and tissue types. However, identifying precise CRM-gene regulations remains challenging due to limitations in data resolution and the complexity of regulatory interactions. Therefore, experimental determination of target genes of CRMs on a genome-wide scale remains an ongoing challenge. To address this challenge, various computational methods have emerged in recent years, including score-based(36, 37), correlation-based(38-40) and machine learning methods (40-44), in past few years, aiming to predict target genes of putative enhancers and silencers. However, these methods are also limited, since in the absence of a precise and comprehensive CRM map in the genome, they aim to predict target genes for specific genomic regions marked by distinct epigenetic modifications. Besides, though these methods have provided some valuable insights, they face constraints due to the arbitrary selection of flanking regions around putative CRMs; the inconsistence of experimentally validated training sets introduces noise and potentially influence prediction accuracy for machine learning methods. In response, we proposed a new method termed correlation and physical proximity (CAPP), which leverages our predicted CRMs as well as their functional states while considering regulatory interactions within TADs.

1.1 Organization

The remainder of this dissertation is structured as follows. In Chapter 2, we analyze the landscape of CRMs and the architecture of their constituent TFBSs in both the human and mouse genomes and reveal the common rules of their organizations. In Chapter 3, we use two logistic regression

(LR) models using distinct features to predict the functional types (enhancers or silencers) and states (active or inactive) of CRMs. In Chapter 4, we use our target gene prediction method, CAPP, to predict both the target genes as well as the functional states of these CRMs. We finally conclude this dissertation in Chapter 5.

Chapter 2
COMMON RULES OF THE ORGANIZATION AND ARCHITECTURE OF CRMs IN THE HUMAN AND MOUSE GENOMES

2.1 Introduction

Annotating all CRMs in genomes is essential for understanding genome functions and its evolutionary history. Despite significant progress made in the two last decades with the development of techniques like ChIP-seq and others, this task remains incomplete(45). Indeed, the degree to which the human genome is deemed functional is actively debated within the scientific community(46, 47). The main source of contention arises from the challenge of precisely delineating the locations and thus the proportion of functional non-coding sequences in the genome, particularly CRMs, which have been estimated to span from 8% to 40% of the genome(46, 47). Additionally, estimates of the number of CRMs in the human genome vary widely, from 400,000(46) to over a million(45). Similarly, systematic annotation of CRMs in the well-studied mouse genome is still in its early stages, despite significant efforts made by large consortia such as ENCODE(48, 49) and individual research laboratories(50, 51). Furthermore, current methods often fail to accurately predict the TFBSs within CRMs, hindering the understanding of the intricate mechanisms of transcriptional regulation in vital biological processes.

Moreover, how CRMs are distributed and organized in genomes and their relationships with target genes in linear DNA are fundamental issues in understanding their functions, as such information provides crucial insights into the complexities of gene regulation and mechanisms of gene activation and repression in various biological contexts. However, our understanding of the genomic landscape of CRMs remains limited due to the lack of highly accurate and comprehensive maps of CRMs in genomes. Consequently, contradictory findings have been reported in the literature. For instance, the ENCODE project(17) observed a pronounced nonuniform distribution

of so-called cCREs within ENCODE regions, and thus proposed that cCREs tended to cluster in certain regions, forming regulatory islands, while other regions were relatively devoid of regulatory elements, forming regulatory "deserts". On the contrary, other researchers(18) have reported that enhancers are dispersed across extensive regions rather than clustering around genes.

Furthermore, at a finer resolution, a CRM consists of clusters of TFBSs interspersed with spacer sequences(52). TFBSs serve as docking sites for cognate TFs, facilitating either the activation or repression of target gene transcription in a highly context-specific manner. TFBSs of different TFs may overlap(53), forming TFBS islands within CRMs. The regions between these TFBS islands, known as inter-TFBS spacers, are recognized as playing crucial roles in gene regulation. For instance, studies have demonstrated that these spacers can influence the conformation of specific regions such as the DNA-binding surface, the "lever arm" and the dimerization interface of the rat glucocorticoid receptor DNA binding domain(54), highlighting their functional relevance in modulating TF activities. Moreover, it has been shown that modification of spacers in synthetic elements can alter gene expression, supporting the idea that the flanking regions of TFBS islands play a significant role in determining *cis*-regulatory activity(55). However, no genome scale study on the inter-TFBS spacers has been conducted due to the aforementioned reasons, to our best knowledge.

As a continued effort, we have recently predicted unprecedentedly comprehensive maps of CRMs and their constituent TFBSs from 85.5% of the human(19, 20) and 79.9% of the mouse(21) genomes using available TF ChIP-seq data from various cell/tissue types of each species. The availability of these CRM and TFBS maps well-positioned us to analyze the landscape and organization of the CRMs in the genomes as well as their architecture in terms of TFBSs, thereby uncovering the underlying rules of the organization and architecture of CRMs.

2.2 Results

2.2.1 The total lengths and numbers of CRMs on chromosomes are strongly correlated with both the numbers of genes on and the sizes of chromosomes

We first examined the distributions of the 1.2M and 0.8M CRMs, alongside 63,133 and 55,361 annotated genes, across the autosomal and sex chromosomes of the human (hg38) and mouse (mm10) genomes, respectively. Both the numbers and total lengths of CRMs on chromosomes are strongly correlated with the chromosome sizes as well as with the numbers and total lengths of genes on chromosomes for both humans (Figure 2-1A) and mice (Figure 2-2A). Expectedly, the numbers and total lengths of CRMs on chromosomes also are strongly correlated with each other (Figures 2-1A and 2-2A), as both the number and total length of CRMs on a chromosome in both species are proportional to the chromosome's size.

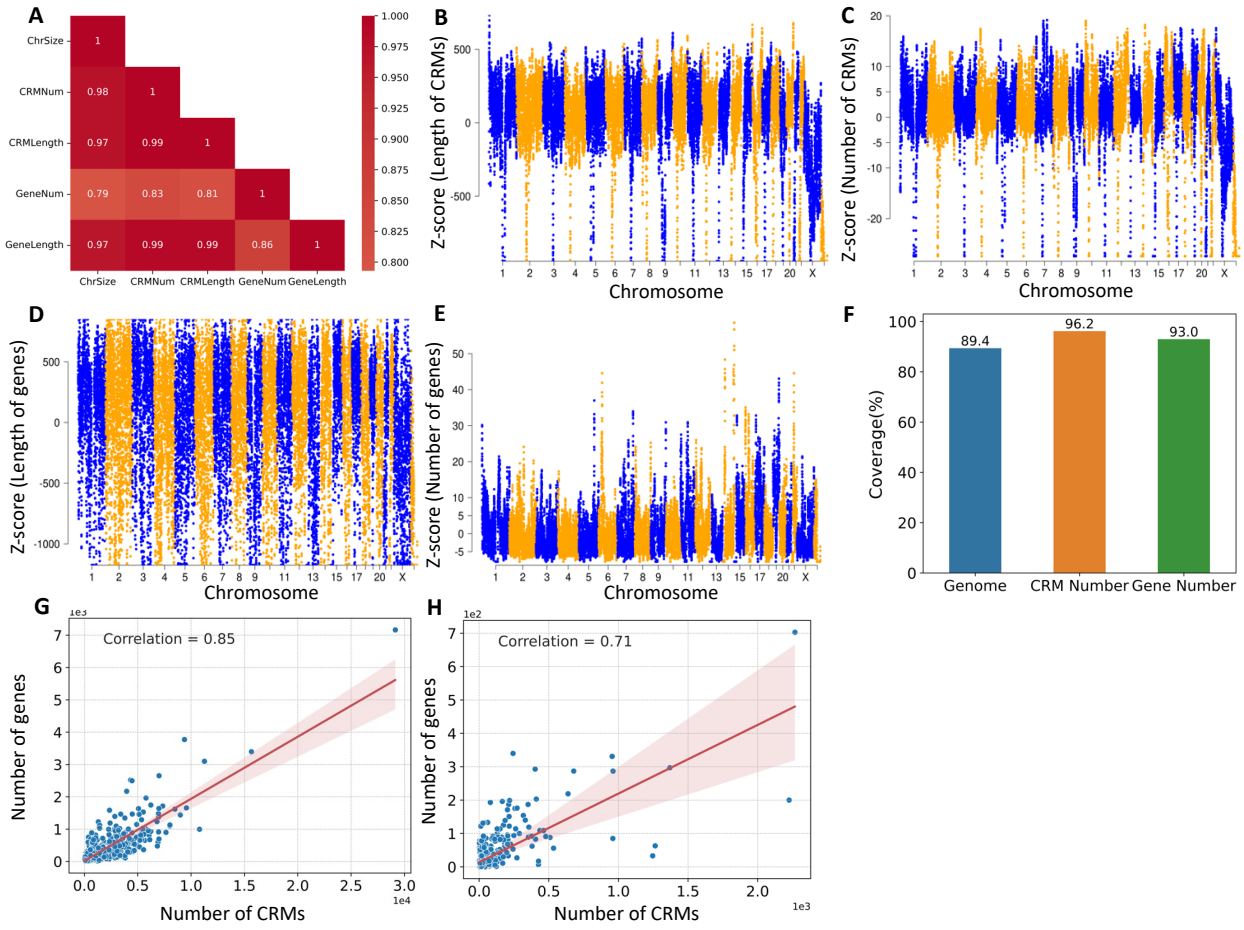


Figure 2-1. The landscape of CRMs and genes on the chromosomes, in sliding windows of 10^6 bp with a step size of 10^5 bp along the chromosomes and in TADs and non-TADs in the human genome. **A.** Heatmap of correlations among genome sizes, numbers of CRMs and genes on chromosomes, and total lengths of CRMs and genes on chromosomes. **B.** Manhattan plot of the normalized lengths of CRMs in sliding windows along the chromosomes. **C.** Manhattan plot of the normalized numbers of CRMs in sliding windows along the chromosomes. **D.** Manhattan plot of the normalized lengths of genes in sliding windows along the chromosomes. **E.** Manhattan plot of the normalized numbers of genes in sliding windows along the chromosomes. **F.** Percentages of the genome length, and numbers of CRMs and genes, covered by TADs. **G.** Correlation between the numbers of CRMs and the numbers of genes within TADs. **H.** Correlation between the numbers of CRMs and the numbers of genes within non-TAD regions.

2.2.2 CRMs and genes are unevenly but correlatedly distributed along chromosomes

To explore the distributions of CRMs and their relationships with those of genes along chromosomes, we analyzed the total lengths and numbers of CRMs as well as of genes within a sliding window of 10^6 bp with a step size of 10^5 bp along each chromosome of the human genome. We then converted the total lengths and numbers of CRMs as well as of genes in each window

into corresponding z-scores with the null hypotheses that the total lengths and numbers of CRMs as well as of genes are evenly distributed along the genome (Materials and Methods). Among the 30,678 sliding windows along the human genome for a window size of 10^6 bp, a considerable portion (20,217/65.9% and 7,081/23.1%) showed enrichment ($z > 5.2$) for the lengths and numbers of CRMs (Figures 2-1B and 2-1C, Supplementary Table S2-1), suggesting the presence of CRM-rich islands. Conversely, many windows (9,920/32.3% and 3,919/12.8%) exhibited depletion ($z < -5.2$) for the lengths and numbers of CRMs, indicating the existence of CRM deserts. For genes, a substantial number of windows are enriched for the lengths ($z > 5.2$: 17,962/58.6%) and numbers ($z > 5.2$: 4,947/16.1%), while many others are depleted of the lengths ($z < -5.2$: 12,547/40.9%) and numbers ($z < -5.2$: 4,673/15.2%) (Figures 2-1D and 2-1E, Supplementary Table S2-1), suggesting an uneven distribution of gene along chromosomes, forming islands or deserts of genes. Notably, there are more than twice as many CRM and gene islands as well as deserts measured by their lengths as measured by their numbers. The results using a sliding window of 10^7 bp with a step of 10^6 bp (Figure 2-3 and Supplementary Table S2-2) support these findings. Additionally, CRMs and genes in sliding windows are strongly correlated in their lengths and numbers (Figure 2-4), indicating the simultaneous enrichment or depletion of CRMs and genes within the same sliding windows, aligning with the previous report by ENCODE(17) that CRMs are clustered around the genes. Similar observations were made in the mouse genome (Figures 2-2B to 2-2E, 2-5, 2-6 and Supplementary Tables S2-3, S2-4).

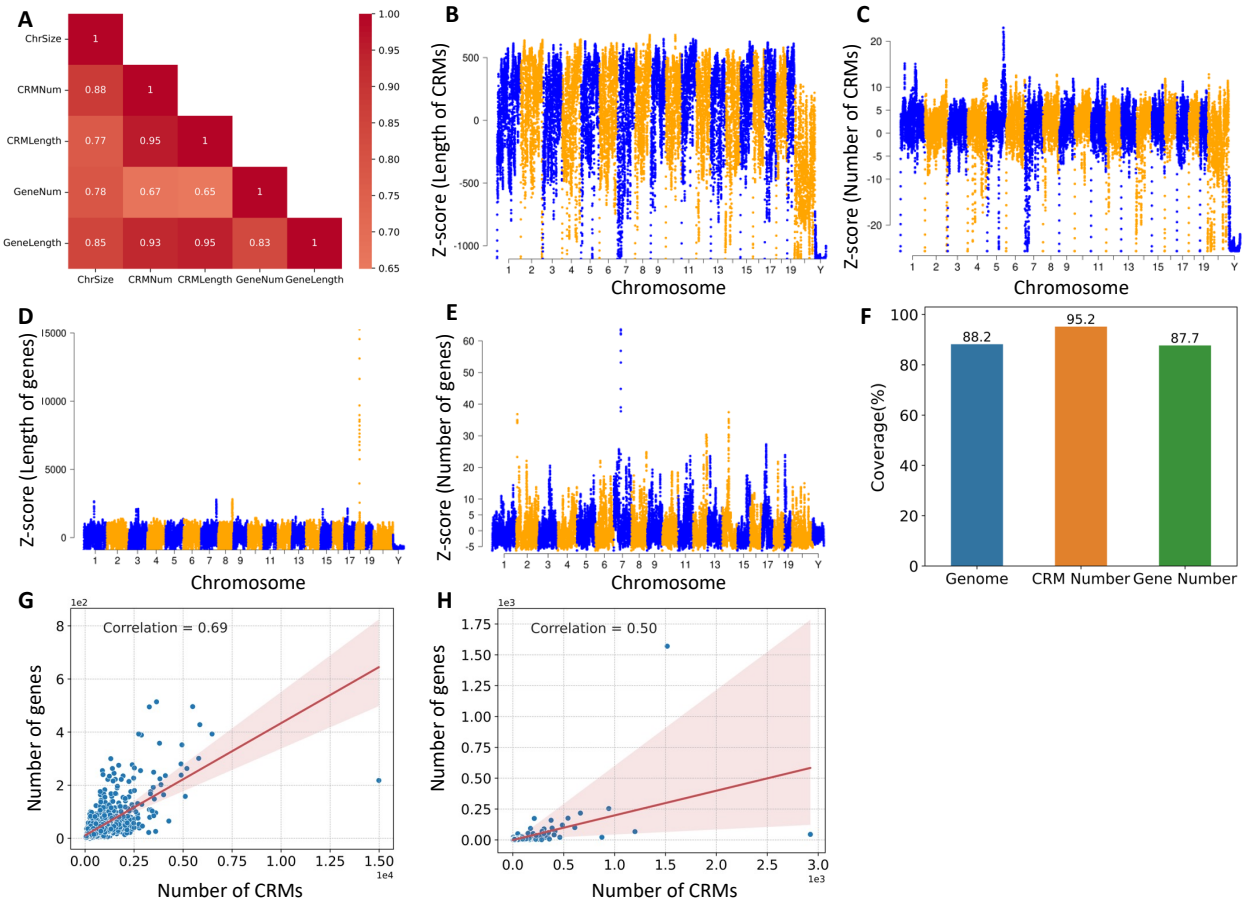


Figure 2-2. The landscape of CRMs and genes on the chromosomes, in sliding windows of 10^6 bp with a step size of 10^5 bp along the chromosomes and in TADs and non-TADs in the mouse genome. **A.** Heatmap of correlations among genome sizes, numbers of CRMs and genes on chromosomes, and total lengths of CRMs and genes on chromosomes. **B.** Manhattan plot of the normalized lengths of CRMs in sliding windows along the chromosomes. **C.** Manhattan plot of the normalized numbers of CRMs in sliding windows along the chromosomes. **D.** Manhattan plot of the normalized lengths of genes in sliding windows along the chromosomes. **E.** Manhattan plot of the normalized numbers of genes in sliding windows along the chromosomes. **F.** Percentages of the genome length, and numbers of CRMs and genes, covered by TADs. **G.** Correlation between the numbers of CRMs and the numbers of genes within TADs. **H.** Correlation between the numbers of CRMs and the numbers of genes within non-TAD regions.

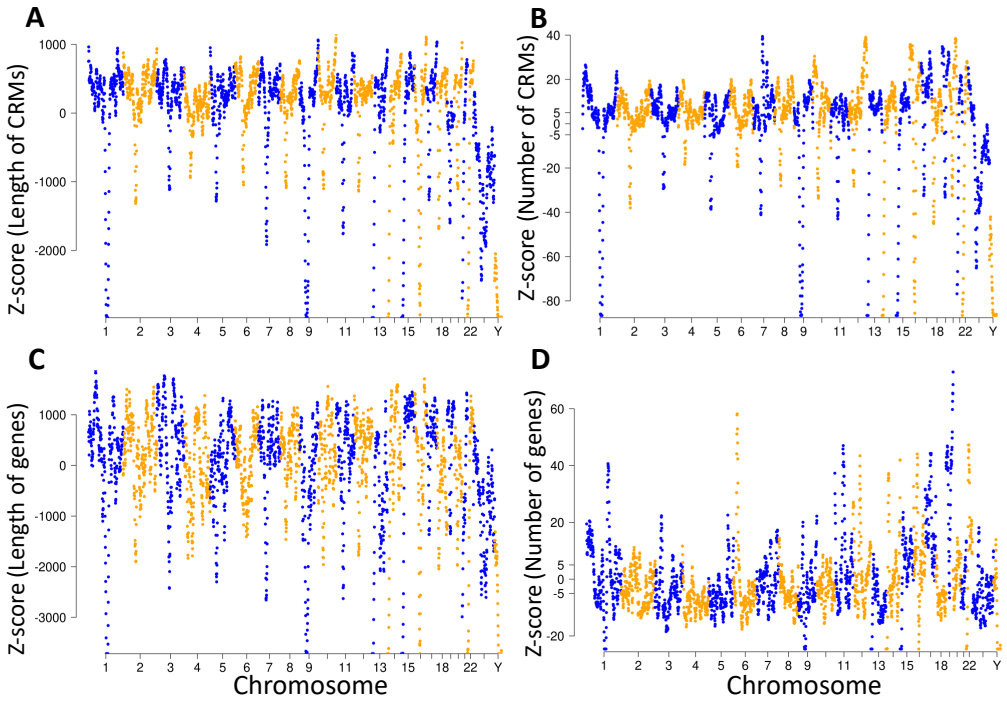


Figure 2-3. Lengths and Numbers of CRMs and genes in sliding windows of 10^7 bp with a step size of 10^6 bp along the chromosomes of the human genome. **A.** Manhattan plot of the normalized lengths of CRMs in sliding windows along the chromosomes. **B.** Manhattan plot of the normalized numbers of CRMs in sliding windows along the chromosomes. **C.** Manhattan plot of the normalized lengths of genes in sliding windows along the chromosomes. **D.** Manhattan plot of the normalized numbers of genes in sliding window along the chromosomes.

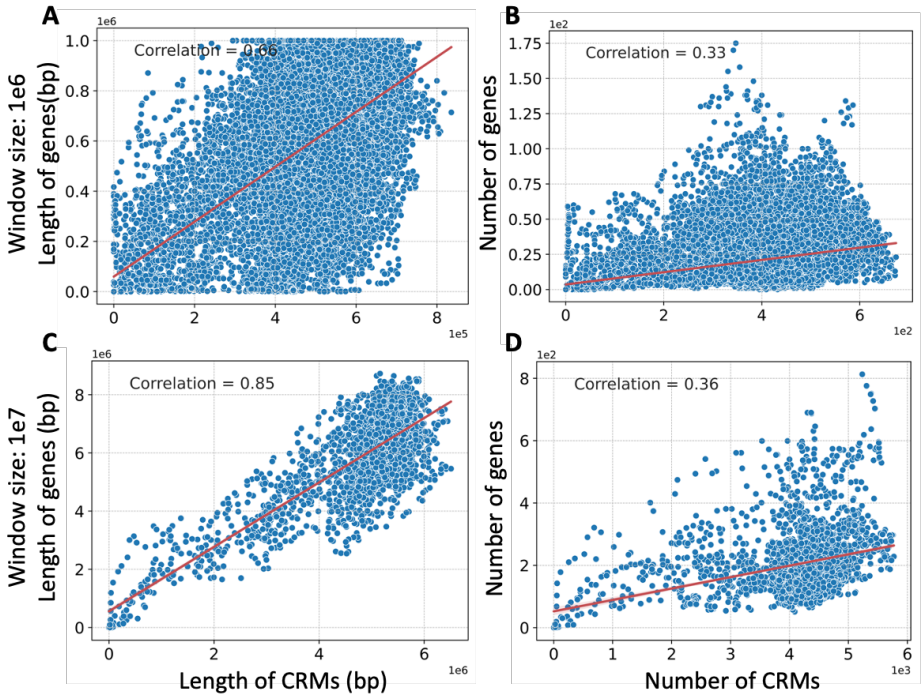


Figure 2-4. Relationships between the lengths or numbers of CRMs and those of genes in sliding windows in the human genome. **A.** Relationship between the lengths of CRMs and those of genes in sliding windows of 10^6 bp. **B.** Relationship between the numbers of CRMs and those of genes in sliding windows of 10^6 bp. **C.** Relationship between the lengths of CRMs and those of genes in sliding windows of 10^7 bp. **D.** Relationship between the numbers of CRMs and those of genes in sliding windows of 10^7 bp.

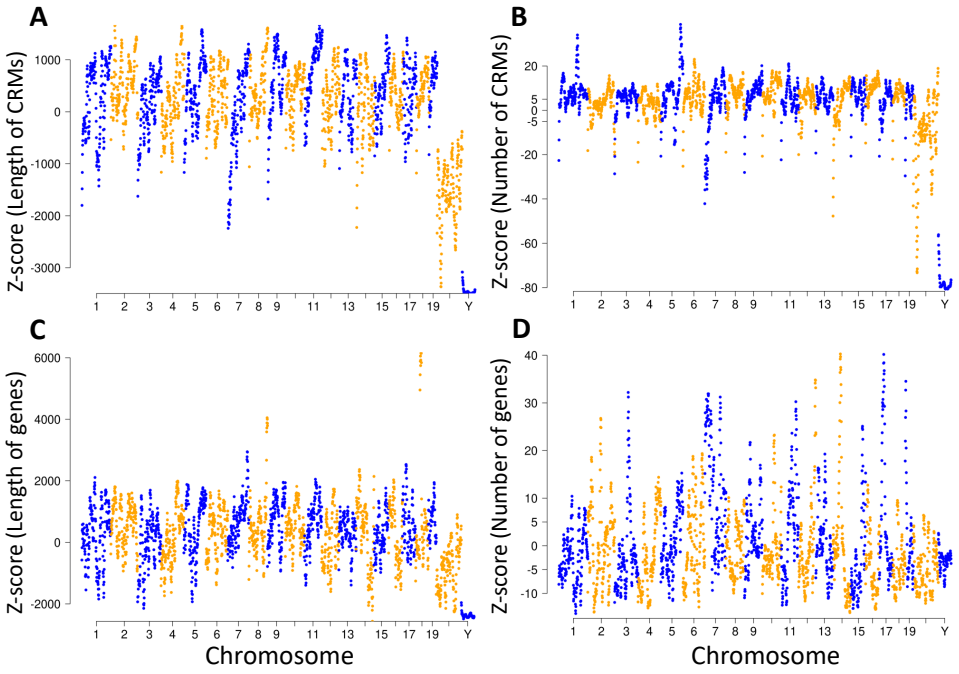


Figure 2-5. Lengths and Numbers of CRMs and genes in sliding windows of 10^7 bp with a step size of 10^6 bp along the chromosomes of the mouse genome. **A.** Manhattan plot of the normalized lengths of CRMs in sliding windows along the chromosomes. **B.** Manhattan plot of the normalized numbers of CRMs in sliding windows along the chromosomes. **C.** Manhattan plot of the normalized lengths of genes in sliding windows along the chromosomes. **D.** Manhattan plot of the normalized numbers of genes in sliding window along the chromosomes.

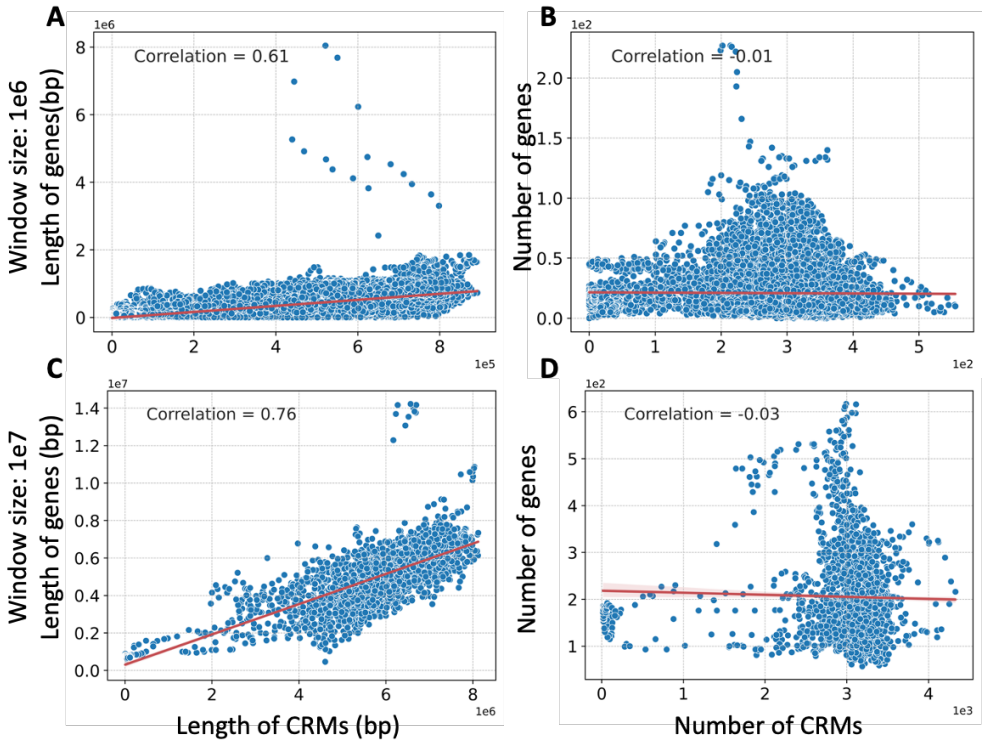


Figure 2-6. Relationships between the lengths or numbers of CRMs and those of genes in sliding windows in the mouse genome. **A.** Relationship between the lengths of CRMs and those of genes in sliding windows of 10^6 bp. **B.** Relationship between the numbers of CRMs and those of genes in sliding windows of 10^6 bp. **C.** Relationship between the lengths of CRMs and those of genes in sliding windows of 10^7 bp. **D.** Relationship between the numbers of CRMs and those of genes in sliding windows of 10^7 bp.

2.2.3 The numbers of CRMs and genes within a TAD exhibit stronger correlation than those in non-TAD regions

TADs typically manifest at the sub-megabase scale, and CRMs predominantly regulate genes within the same TADs(56). After demonstrating the correlation between the occurrences of CRMs and genes along chromosomes in both genomes, we explored the relationships between the numbers of CRMs and genes within TADs. As depicted in Figure 2-1F, TADs cover 89.4% of the human genome, yet they contain 96.2% of the CRMs and 93.0% of the genes in the genome, indicating an enrichment of CRMs and genes within TADs compared to non-TAD regions. We observed a similar trend for the numbers of CRMs, but not for the number of genes, in the mouse genome (Figure 2-2F). Interestingly, the numbers of CRMs and genes within both TADs and non-

TAD are correlated; however, the correlation in TADs is significantly higher (Fisher's z test: $p<0.001$) than that in non-TAD regions in both the human (Figures 2-1G and 2-1H) and the mouse (Figures 2-2G and 2-2H) genomes. The results indicate stronger dependency between the numbers of CRMs and genes in TADs than in non-TAD regions.

2.2.4 Both our CRMs and experimentally validated regulatory elements are slightly biasedly distributed downstream of their nearest transcription start sites (TSSs)

Although not every enhancer regulates their nearest genes, a considerable portion of them do(57). It also is unknown whether CRMs are preferentially located upstream or downstream of their nearest TSSs. We therefore analyzed the distributions of the distances between the middle points of our CRMs and their nearest TSSs (middle point distance, d_m, Materials and Methods) while taken the orientations of the TSSs in consideration, such that a negative d_m indicates an upstream middle point and a positive d_m a downstream middle point. We compared the results with those of experimentally validated FANTOM promoters(58), FANTOM enhancers(59), and VISTA enhancers(60) in both the human and mouse genomes. In both genomes, FANTOM promoters (n=184,326 for humans, Figures 2-7A1 and 2-7B1; and n=164,421 for mice, Figures 2-7C1 and 2-7D1), FANTOM enhancers (n=32,684 for humans, Figures 2-7A2 and 2-7B2; and n=49,797 for mice, Figures 2-7C2 and 2-7D2), VISTA enhancers (n=1,002 for humans, Figures 2-7A3 and 2-7B3; and n=702 for mice, Figures 2-7C3 and 2-7D3) and our predicted CRMs (n=1.2M for humans, Figures 2-7A4 and 2-7B4; and n=0.8M for mice, Figures 2-7C4 and 2-7D4) are all almost symmetrically distributed around their nearest TSSs, but all slightly biased to downstream of their nearest TSSs (Table 2-1, except for VISTA enhancers in mice), and this is particularly true for FANTOM promoters. Moreover, FANTOM promoters (Figures 2-7A1 and 2-7C1) are more closely located around TSSs than are FANTOM enhancers (Figures 2-7A2 and 2-7C2), VISTA enhancers (Figures 2-7A3 and 2-7C3) and our CRMs (Figures 2-7A4 and 2-7C4). More

specifically, most of FANTOM promoters (66.8% for humans and 59.1% for mice) have a $|d_m|$ < 1,000 bp (Figures 2-7B1 and 2-7D1). In contrast, only a small portion of FANTOM enhancers (11.8% for humans, Figure 2-7B2; and 12.5% for mice, Figure 2-7D2), VISTA enhancers (12.7% for humans, Figure 2-7B3; and 6.8% for mice, Figure 2-7D3), and our CRMs (7.2% for humans, Figure 2-7B4; and 5.5% for mice, Figure 2-7D4) have a $|d_m|$ < 1,000 bp. These results are consistent with the general understanding that promoters tend to be proximal to TSSs while enhancers tend to be distal to TSSs. However, the vast majority of FANTOM enhancers (80.8% for humans, Figure 2-7A2; and 96.7% for mice, Figure 2-7C2), VISTA enhancers (66.2% for humans, Figure 2-7A3; and 98.0% for mice, Figure 2-7C3) and our CRMs (73.4% for humans, Figure 2-7A4; and 90.2% for mice, Figure 2-7C4) have a $|d_m|$ < 0.1M bp.

Table 2-1. Numbers and proportions of experimentally validated regulatory elements and our CRMs located upstream and downstream of their nearest TSSs based on d_m values

	Total Number		$d_m < 0$		$d_m > 0$	
	Human	Mouse	Human	Mouse	Human	Mouse
FANTOM Promoters	184,326	164,421	79,964 (43.4%)	68,904 (41.9%)	100,858 (54.7%)	93,630 (56.9%)
FANTOM Enhancers	32,684	49,797	15,663 (47.9%)	23,644 (47.5%)	17,016 (52.1%)	26,148 (52.5%)
VISTA Enhancers	1,002	702	480 (47.9%)	354 (50.4%)	520 (51.9%)	348 (49.6%)
Our CRMs	1,225,115	798,257	583,588 (47.6%)	387,093 (48.5%)	641,408 (52.4%)	411,125 (51.5%)

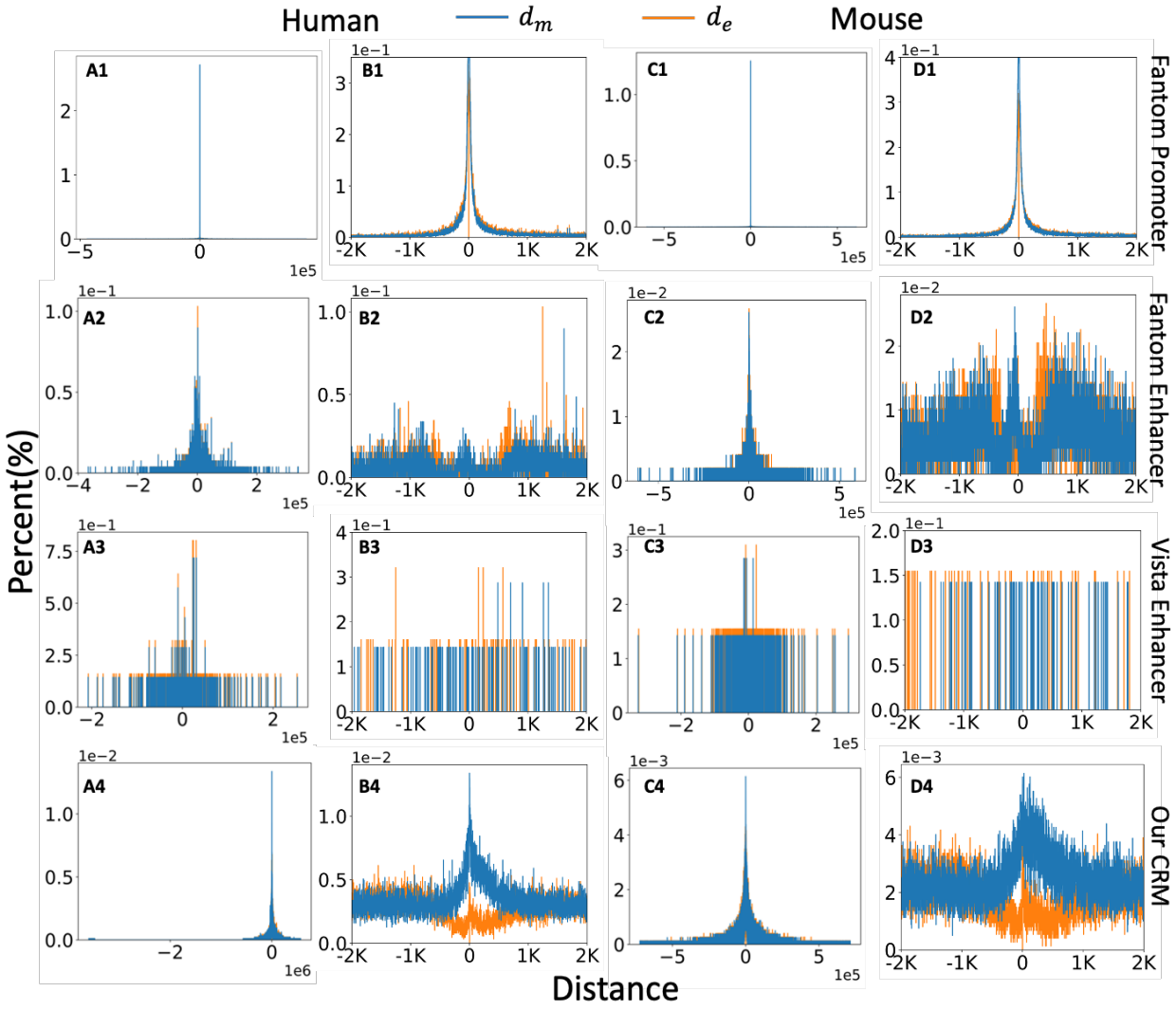


Figure 2-7. Distributions of FANTOM promoters and enhancers, VISTA enhancers and CRMs in the human and mouse genomes around the nearest TSSs. **A1-A4.** Histogram of d_m of FANTOM promoters, FANTOM enhancers, VISTA enhancers and our CRMs and histogram of d_e of their CPL category in the human genome. **B1-B4.** Zooming-in views of regions of A1-A4 indicated by the axes, respectively. **C1-C4.** Histogram of d_m of FANTOM promoters, FANTOM enhancers, VISTA enhancers and our CRMs and histogram of d_e of their CPL category in the mouse genome. **D1-D4.** Zooming-in views of regions of C1-C4 indicated by the axes, respectively.

2.2.5 Both our CRMs and experimentally validated regulatory elements can be classified in two categories based on whether they overlap TSSs or not

Notably, in both humans and mice (Figure 2-7), the histograms of d_m for FANTOM promoters, FANTOM enhancers, VISTA enhancers and our predicted CRMs all peak around the 0 distance, indicating that varying yet considerable portions of all these elements overlap their nearest TSSs.

Since the elements overlapping TSSs might contain core promoters, while those that do not might not, we thus classified these elements in two categories: 1) those that overlap their nearest TSSs as core promoter-containing (CPC) elements, and 2) those that do not overlap their nearest TSSs as core promoter-lacking (CPL) elements. In humans, as expected, a considerable portion (19.9%) of FANTOM promoters are classified as CPC promoters, while the remaining 80.1% are classified as CPL promoters (Figure 2-8A, Table 2-2). Interestingly, a considerable portion of VISTA enhancers (12.2%) also are classified as the CPC category, while the remaining 80.1% are classified as the CPL category (Figure 2-8A, Table 2-2), indicating the enhancers indeed can overlap core promoters. Moreover, smaller but still considerable portions of FANTOM enhancers (3.0%) and our CRMs (6.1%) are classified as CPC enhancers, while the remaining vast majority are classified as CPL elements (Figure 2-8A, Table 2-2). These results indicate that a considerable number of experimentally validated enhancers and our predicted CRMs contain core promoters in addition to other regulatory elements. These findings also highlight the substantial differences in the proportions of CPC and CPL categories between FANTOM promoters as well as VISTA enhancers and FANTOM enhancers as well as our CRMs. Similar results are obtained in the mouse data (Figure 2-9A, Table 2-2).

Table 2-2. Numbers and proportions of experimentally validated regulatory elements and our CRMs categorized into the CPC and CPL categories in the two genomes

	Total Number		CPC		CPL	
Species	Human	Mouse	Human	Mouse	Human	Mouse
FANTOM Promoters	184,326	164,421	36,759 (19.9%)	25,565 (15.5%)	147,567 (80.1%)	138,856 (84.5%)
FANTOM Enhancers	32,684	49,797	972 (3.0%)	1,066 (2.1%)	31,712 (97.0%)	48,731 (97.9%)
VISTA Enhancers	1,002	702	122 (12.2%)	56 (8.0%)	880 (87.8%)	646 (92.0%)
Our CRMs	1,225,115	798,257	74,329 (6.1%)	54,203 (6.8%)	1,150,786 (93.9%)	744,054 (93.2%)

To further investigate the location relationships between the CPL elements and their nearest TSSs, we analyzed the distribution of nearer end distance d_e for the CPL elements (defined as the distance between the nearer end of an element and its nearest TSS, Materials and Methods). In both humans and mice (Figure 2-7, Table 2-3), all the CPL elements are almost symmetrically distributed around their nearest TSSs, but all slightly biased to downstream of their nearest TSSs except for VISTA enhancers in mice. More than half (56.3% and 51.7% for humans and mice, respectively) of FANTOM CPL promoters are located around the nearest TSSs with a $|d_e| < 1{,}000$ bp (Figures 2-7A1, 2-7B1 and 2-7C1, 2-7D1). In contrast, only a small portion of FANTOM CPL enhancers (11.7% for humans, Figures 2-7A2, 2-7B2; and 12.9% for mice, Figures 2-7C2, 2-7D2), VISTA CPL enhancers (7.1% for humans, Figures 2-7A3, 2-7B3; and 3.6% for mice, Figures 2-7C3, 2-7D3) and our CPL CRMs (4.5% for humans, Figures 2-7A4, 2-7B4; and 2.9% for mice, Figures 2-7C4, 2-7D4) are located around the nearest TSSs with a $|d_e| < 1{,}000$ bp.

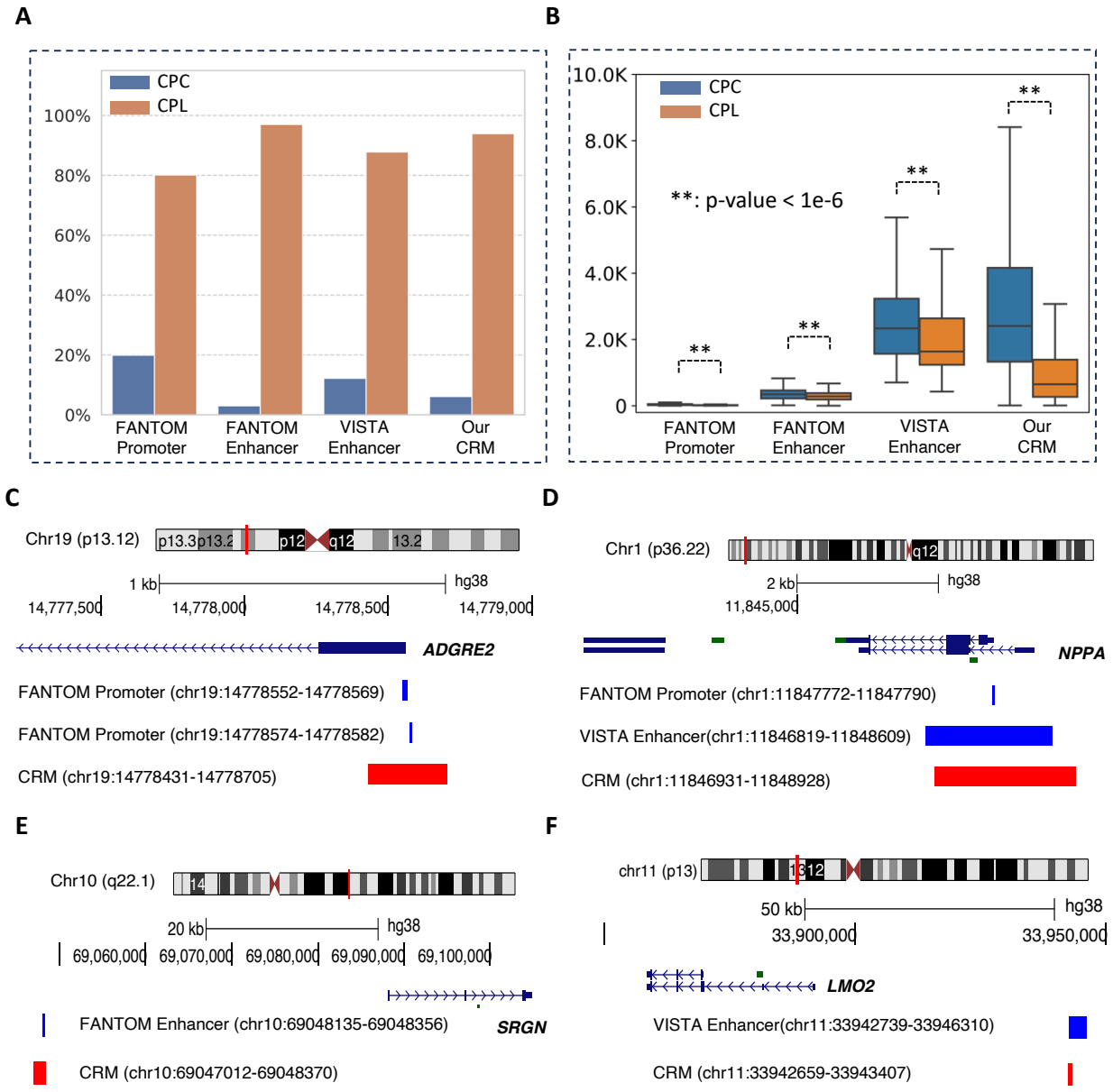


Figure 2-8. Classification of FANTOM promoters and enhancers, VISTA enhancers and our CRMs based on whether or not they overlap TSSs in the human genome. **A.** Comparison of percentages of the CPC and CPL categories in FANTOM promoters and enhancers, VISTA enhancers and our CRMs. **B.** Boxplots of the lengths of the CPC and CPL categories in FANTOM promoters and enhancers, VISTA enhancers and our CRMs. **C.** CRM (chr19:14778431-14778705) is classified to be the CPC category, containing a FANTOM CPC promoter (a core promoter of gene *ADGRE2*: chr19:14778552-14778569) and a FANTOM CPL promoter (a proximal promoter of *ADGRE2*: chr19:14778574-14778582) plus additional putative regulatory elements. **D.** CRM (chr1:11846931-11848928) is classified to be the CPC category, containing a FANTOM CPC promoter (a core promoter of gene *NPPA*: chr1:11847772-11847790) and a VISTA CPC enhancer (chr1:11846819-11848609) plus additional putative regulatory elements. **E.** CRM (chr10:69047012-69048370) located upstream of gene *SRGN* is classified to be the CPL category,

overlapping a FANTOM CPL enhancer (chr10:69048135-69048356). **F.** CRM (chr11:33942659-33943407) located upstream of gene *LMO2* is classified to be the CPL category, overlapping a VISTA CPL enhancer (chr11:33942739-33946310).

Table 2-3. Numbers and proportions of experimentally validated regulatory elements and our CRMs located upstream and downstream of their nearest TSSs based on d_e values

	Total Number		$d_e < 0$		$d_e > 0$	
	Human	Mouse	Human	Mouse	Human	Mouse
FANTOM Promoters	184,326	164,421	65,180 (35.4%)	57,540 (35.0%)	82,386 (44.7%)	81,316 (49.5%)
FANTOM Enhancers	32,684	49,797	14,953 (45.8%)	22,862 (45.9%)	16,759 (51.3%)	25,869 (51.9%)
VISTA Enhancers	1,002	702	411 (41.0%)	325 (46.3%)	469 (46.8%)	321 (45.7%)
Our CRMs	1,225,115	798,257	554,778 (45.3%)	366,184 (45.9%)	596,025 (48.7%)	377,870 (47.3%)

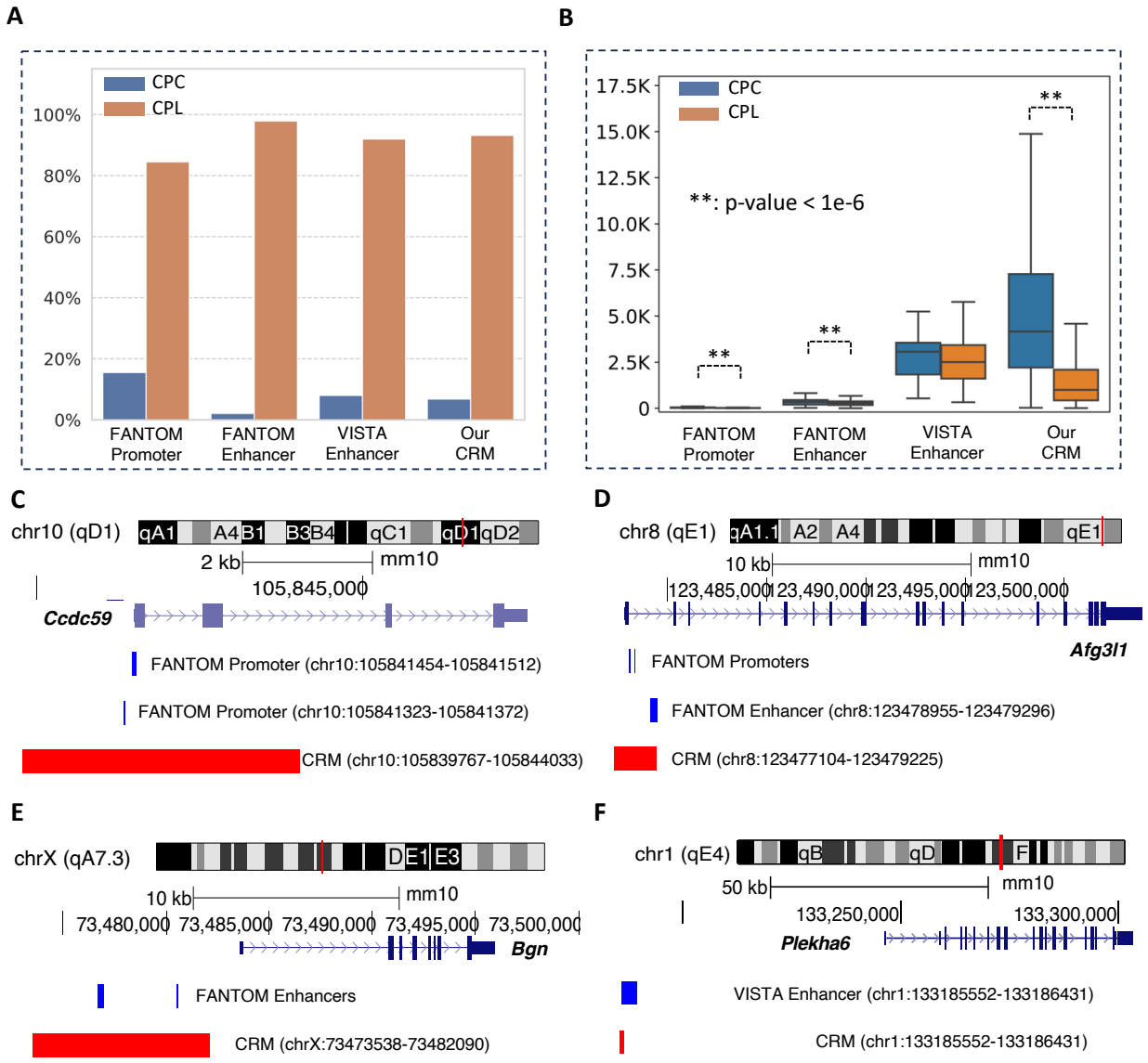


Figure 2-9. Classification of FANTOM promoters and enhancers, VISTA enhancers and our CRMs based on whether or not they overlap TSSs in the mouse genome. **A.** Comparison of percentages of the CPC and CPL categories in FANTOM promoters and enhancers, VISTA enhancers and our CRMs. **B.** Boxplots of the lengths of the CPC and CPL categories in FANTOM promoters and enhancers, VISTA enhancers and our CRMs. **C.** CRM (chr10:105839767-105844033) is classified as the CPC category, containing a FANTOM CPC promoter (a core promoter of gene *Ccdc59*: chr10: 105841454-105841512) and a FANTOM CPC promoters (a core promoter of gene *Ccdc59*: chr10: 105841323-105841372) plus other regulatory elements. **D.** CRM (chr8:123477104-123479225) is classified as the CPC category, containing a FANTOM CPL promoters (a proximal promoter of gene *Afg3l1*: chr8:123477849-123477856), two FANTOM CPC promoters (core promoters of gene *Afg3l1*: chr8:123477859-123477904, chr8:123477934-123477945), and a downstream promoter element of gene *Afg3l1*: chr8:123478144-123478181 and a FANTOM CPL enhancer (chr8:123478955-123479296) plus other regulatory elements. **E.** CRM (chrX:73473538-73482090) located upstream of gene *Bgn* is classified as the CPL category, containing two FANTOM CPL enhancers (chrX:73476667-73476958 and chrX:73480499-

73480597) plus other regulatory elements. **F.** CRM (chr1:133185552-133186431) located upstream of gene *Plekha6* is classified as the CPL category, overlapping a VISTA CPL enhancer (chr1:133185980-133189493).

2.2.6 CPC elements are generally longer than CPL elements

We proceeded to compare the lengths of FANTOM promoters, FANTOM enhancers, VISTA enhancers and our CRMs. In humans, as anticipated, FANTOM promoters are short, with a nearly uniform length distribution, a median length of 15 bp and 99.4% of them being shorter than 100 bp (Figure 2-8B). Notably, FANTOM CPC promoters have a significantly longer median length (29 bp) than FANTOM CPL promoters (15 bp). Considering the distributions of FANTOM promoters around the nearest TSSs, in addition to a core promoter, a FANTOM CPC promoter might contain an upstream proximal promoter element and/or a downstream promoter element(61), while a FANTOM CPL promoter might only contain an upstream proximal promoter element or a downstream promoter element. Although the vast majority (99.5%) of FANTOM enhancers are shorter than 1,000 bp, with a median length of 288 bp, they are generally longer than FANTOM promoters (Figure 2-8B). Interestingly, FANTOM CPC enhancers have a significantly longer median length (346 bp) than FANTOM CPL enhancers (286 bp) (Figure 2-8B). VISTA enhancers have a length ranging from 428 to 11,051 bp with a median length of 1,688 bp, and thus are much longer than FANTOM enhancers (Figure 2-8B). Interestingly, as in the case of FANTOM enhancers, VISTA CPC enhancers also have a significantly longer median length (2,337 bp) than VISTA CPL enhancers (1,638 bp) (Figure 2-8B). It is likely that in addition to an enhancer element, a FANTOM CPC enhancer or a VISTA CPC enhancer contains a core promoter, while a FANTOM CPL enhancer or a VISTA CPL enhancer only contains enhancer elements.

With a median length of 707 bp, our CRMs are more similar to VISTA enhancers than to FANTOM enhancers in length (Figure 2-8B). Interestingly, as in the cases of FANTOM and VISTA enhancers, our CPC CRMs also have a significantly longer median length (2,408 bp) than

our CPL CRMs (650 bp) (Figure 2-8B). A small number (103 or 0.1%) of our CPC CRMs are shorter than 100 bp that is the typical length of core promoters (Figure 2-8B), suggesting that these CPC CRMs might be core promoters. The remaining vast majority (99.9%) of our CPC CRMs are longer than 100 bp, with 83.6% of them being longer than 1,000 bp. Thus, like FANTOM CPC enhancers and VISTA CPC enhancers, the CPC CRMs that are longer than 100 bp might contain core promoters in addition to enhancer or silencer elements. The analysis conducted on the mouse genome yielded similar results and conclusions (Figure 2-9B). Moreover, of the 213,882 and 125,827 annotated unique TSSs in the human and mouse genome regions from which we were able to predict CRMs and constituent TFBSs(19-21), 192,735 (90.1%) and 118,896 (94.5%) overlap our CPC CRMs in the human and mouse genome regions, respectively, indicating that core promoters rarely exist alone.

2.2.7 Overlaps among our CRMs, FANTOM promoters, FANTOM enhancers and VISTA enhancers

We note that FANTOM enhancers and VISTA enhancers in both humans (Supplementary Table S2-5) and mice (Supplementary Table S2-6) rarely overlap each other. In humans, only 131 FANTOM enhancers overlap with 102 VISTA enhancers (Supplementary Table S2-5) while the numbers for mice are 317 FANTOM enhancers and 220 VISTA enhancers (Supplementary Table S2-6). Therefore, FANTOM enhancers and VISTA enhancers are two quite different sets of experimentally validated enhancers determined using different techniques(59, 60). Moreover, FANTOM enhancers that overlap VISTA enhancers are shorter than the counterpart VISTA enhancers, and multiple such FANTOM enhancers overlap the different parts of the same VISTA enhancers. This suggests that some of the FANTOM enhancers might be components of long enhancers, likely due to the limitations of the eRNA-seq techniques used for their determination.

Although FANTOM enhancers and VISTA enhancers do not have extensive overlaps, we

have previously shown that our predicted CRMs overlap the vast majority of both sets in the human(19) and mouse(21) genomes. We show a few examples of overlaps in light of our classification of the elements in the CPC and CPL categories in humans. First, CPC CRM (chr19:14778431-14778705) contains a FANTOM CPC promoter (a core promoter of gene *ADGRE2*: chr19:14778552-14778569) and a FANTOM CPL promoter (a proximal promoter of *ADGRE2*: chr19:14778574-14778582), in addition to other putative enhancer or silencer elements (Figure 2-8C). Second, CPC CRM (chr1:11846931-11848928) contains a FANTOM CPC promoter (a core promoter of gene *NPPA*: chr1:11847772-11847790) and the most part of a VISTA CPC enhancer (chr1:11846819-11848609) plus additional putative regulatory elements (Figure 2-8D). Third, CPL CRM (chr10:69047012-69048370) located upstream of the nearest TSS of gene *SRGN* overlaps a FANTOM CPL enhancer (chr10:69048135-69048356) (Figure 2-8E). Finally, CPL CRM (chr11:33942659-33943407) located upstream of gene *LMO2* overlaps a VISTA CPL enhancer (chr11:33942739-33946310) (Figure 2-8F). A few examples of overlaps in the mouse genome are shown in Figures 2-9C to 2-9F.

2.2.8 Inter-TFBS spacers in CRMs are under similar evolutionary constraints as TFBS islands

We next investigated the architecture of the CRMs (Figure 2-10A) by analyzing the landscape and properties of the 125M TFBSs within our predicted 1.2M CRMs in the human genome. Our predicted TFBSs within these putative CRMs have a length ranging from 10 to 21 bp, with the majority being 10 bp in length (Figure 2-10B). On average, each CRM contains around 102 TFBSs. By examining the arrangement of TFBSs within CRMs (Figure 2-10A), we found that adjacent TFBSs often overlapped each other by 1 to 10 bp, with 10, 9, and 8 bp overlaps being the most common (Figure 2-10C). On the other hand, it is relatively rare for two adjacent TFBSs to be separated by more than 100 bp (Figure 2-10C). We thus merge two adjacent TFs if they overlap

by at least one bp, and refer to the resulting sequences as TFBS islands, which range from 10 bp to 1,255 bp with a median length of 12 bp (Figure 2-10B). The distance between adjacent TFBS islands within the CRMs ranges from 1 to 2,238 bp, with a median distance of 13 bp, indicating variability in the spacers between the TFBS islands (Figure 2-10C). The analysis conducted on the 165M putative TFBSs in the 0.9M predicted CRMs in the mouse genome yielded similar conclusions (Figures 2-11A and 2-11B).

Of the human genome regions (85.5%) from which we were able to predict CRMs and constituent TFBSs(19, 20), 20.7%, 34.6% and 44.7% consist of TFBS islands, inter-TFBS spacers and non-CRMs, respectively (Figure 2-10D). These proportions are 29.9%, 39.6% and 30.5% for TFBS islands, inter-TFBS spacers and non-CRMs, respectively, in the mouse genome regions (79.9%) from which we were able to predict CRMs and constituent TFBSs(21) (Figure 2-11C). To assess possible functionality of inter-TFBS spacers, we compared the distribution of phyloP conservation scores(62) of their nucleotide positions with those of TFBS islands and non-CRMs. As expected, the distribution of phyloP scores of TFBS islands have a much lower peak around 0 than that of non-CRMs and is right-shifted relative to that of non-CRMs (Figures 2-10E and 2-11D), indicating that TFBS islands are more likely under evolutionary constraints than non-CRMs. Surprisingly, the distribution of phyloP scores of inter-TFBS spacers differs only slightly from that of TFBS islands, with a slightly lower peak of 0 and slightly less right shift (Figures 2-10E and 2-11D), indicating that inter-TFBS spacers are under almost the same evolutionary constraints as TFBS islands, which are much stronger than those on non-CRMs. Moreover, TFBS islands and inter-TFBS spacers in the human genome have similar median LINSIGHT scores that were computed to measure functionality of nucleotide positions in the human genome(63), both are significantly higher than that of non-CRMs (Figure 2-12). These results strongly suggest that inter-

TFBS spacers also play critical roles in CRM functions.

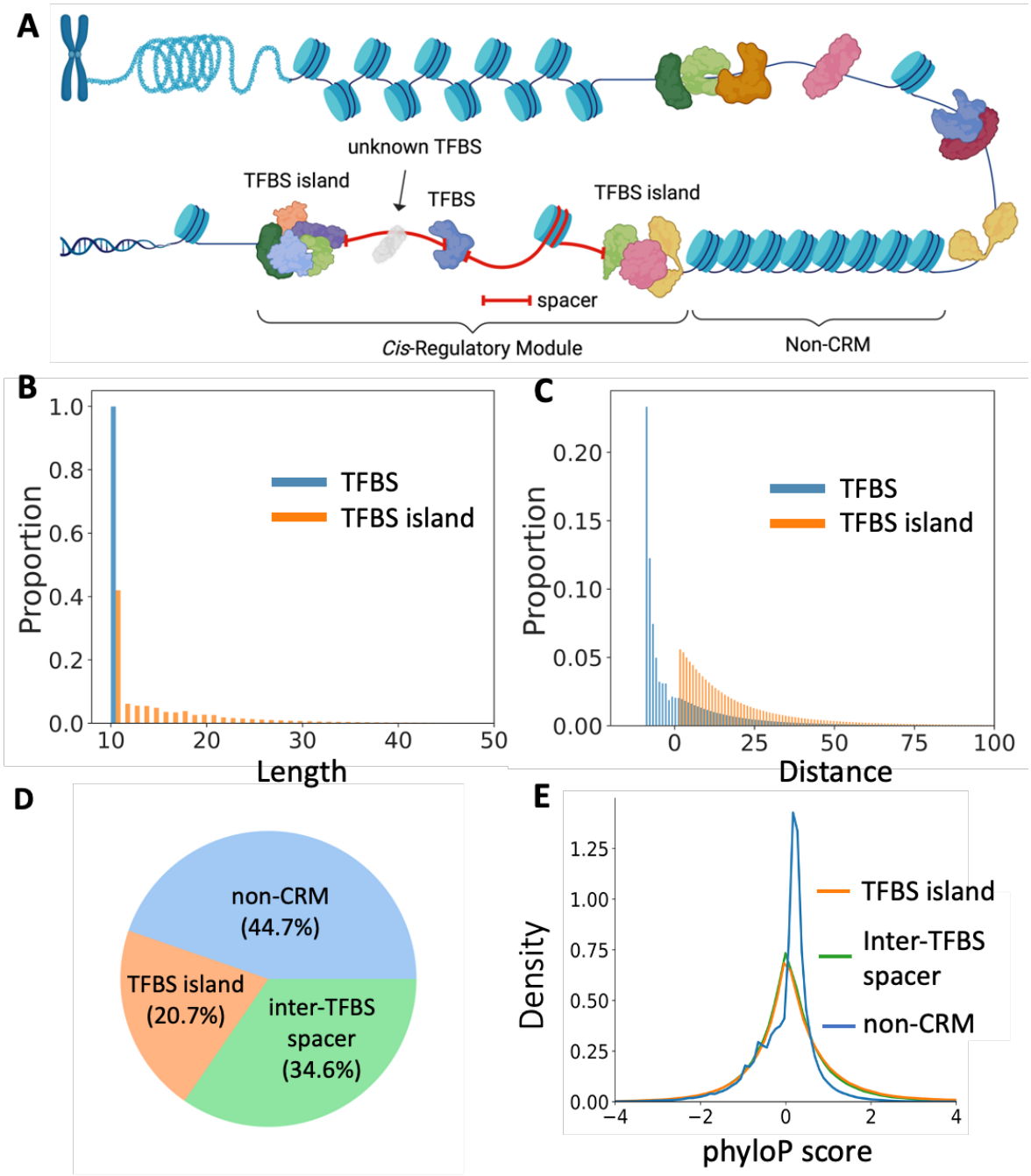


Figure 2-10. Properties of putative TFBSs and inter-TFBS islands in the predicted CRMs in the human genome. **A.** Cartoon showing the arrangement of the TFBSs, TFBS islands and inter-TFBS spacers in a CRM on a chromosome. **B.** Distribution of the lengths of putative TFBSs and TFBS islands in the predicted CRMs (only the length region from 10 to 50 bp is shown). **C.** Distribution of the distance between adjacent of putative TFBSs and TFBS islands in the predicted CRMs (only the distance region from -20 to 100 bp is shown). **D.** Coverage of putative TFBS islands, inter-

TFBS spacers and non-CRMs in the genome regions from which we were able to predict CRMs and constituent TFBSs. **E.** Distribution of phyloP scores of TFBS islands, inter-TFBS spacers in the predicted CRMs in comparison with that of the non-CRMCs (only the score region from -4 to 4 is shown).

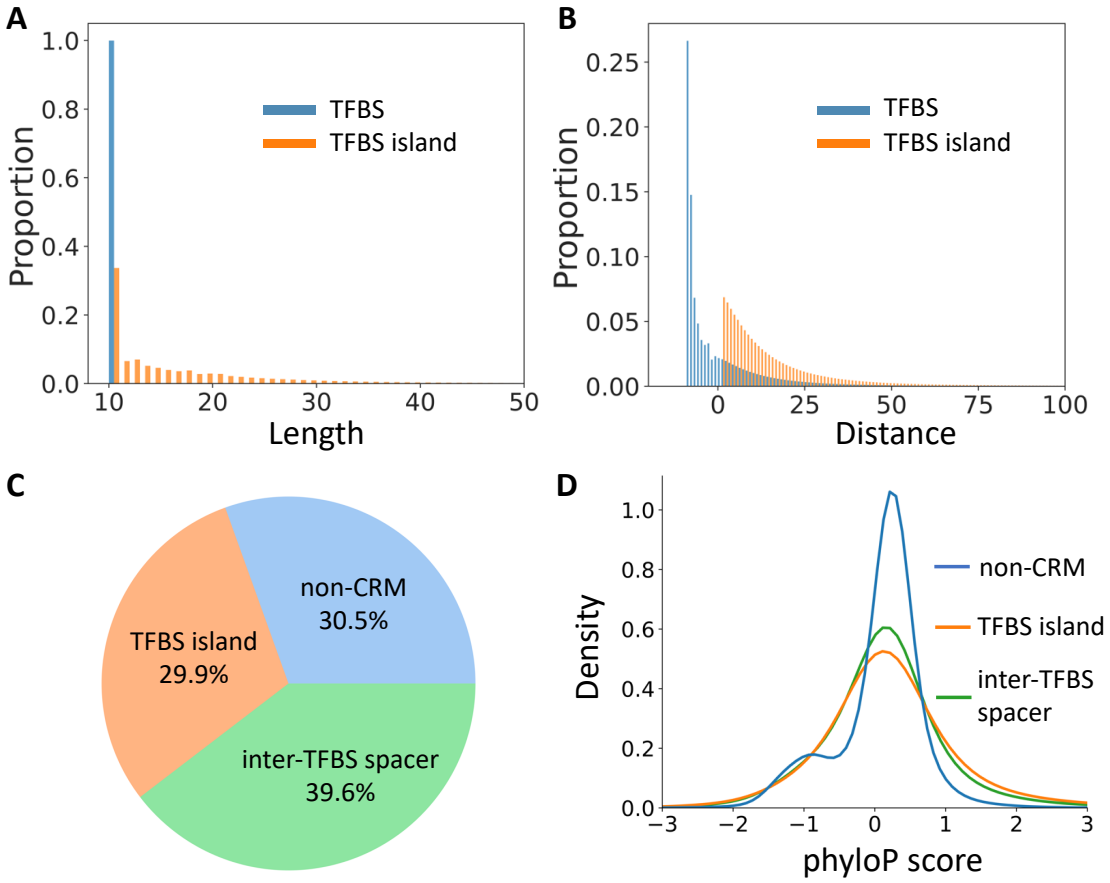


Figure 2-11. Properties of putative TFBSs and inter-TFBS islands in the predicted CRMs in the mouse genome. **A.** Distribution of the lengths of putative TFBSs and TFBS islands in the predicted CRMs (only the length region from 10 to 50 bp is shown). **B.** Distribution of the distance between adjacent of putative TFBSs and TFBS islands in the predicted CRMs (only the distance region from -20 to 100 bp is shown). **D.** Coverage of putative TFBS islands, inter-TFBS spacers and non-CRMs in the genomic regions from which we were able to predict CRMs and constituent TFBSs. **E.** Distribution of phyloP scores of TFBS islands, inter-TFBS spacers in the predicted CRMs in comparison with that of the non-CRMCs (only the score region from -3 to 3 is shown).

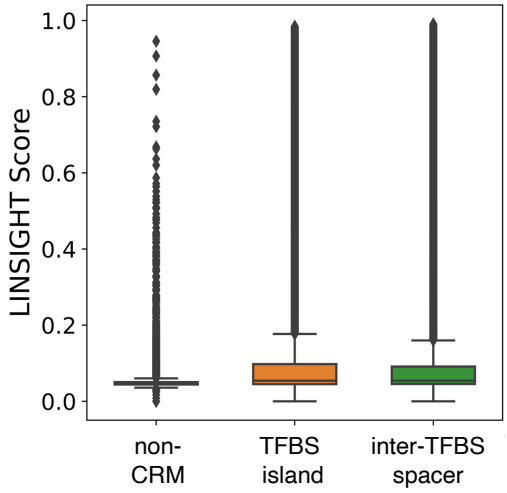


Figure 2-12. Boxplots of LINSIGHT scores of non-CRMs, TFBS islands and inter-TFBS spacers in the human genome.

2.2.9 Inter-TFBS spacers might have functional roles other than direct TF binding in transcriptional regulation

TFBS islands bound by their cognate TFs and inter-TFBS spacers wrapped around histones in a CRM are resistant to DNase I cleavage, while TFBS islands and inter-TFBS spacers free of binding by both TFs and histones can be cut by DNase I (Figure 2-10A). To see whether or not inter-TFBS spacers are directly involved in TF binding, we compared the DNase I cleavage profiles of TFBS islands, inter-TFBS spacers and non-CRMs in the human genome using maps of 3.6 million DNase I hypersensitive site (DHS) cores(64) and 4.6 million TF footprints(65) produced by ENCODE via aggregating DNase-seq data from hundreds of human bio-samples spanning hundreds of cell/tissue types. With an average length of 55 bp, the 3.6 million DHS cores occupying 6.4% of the genome are consensus regions that could be cut by DNase I at multiple positions in different cell/tissue types(64). As summarized in Table 2-4, TFBS islands have the highest proportion (14.4%) overlapping DHS cores, followed by inter-TFBS spacers (8.3%) and non-CRMs (3.6%). Thus, both TFBS islands and inter-TFBS spacers are enriched for DHS cores, while non-CRMs are depleted of DHS cores. This result suggests that compared with TFBS islands, inter-TFBS spacers might be more likely wrapped around histones (Figure 2-10A) and inaccessible to regulatory

proteins, while compared with non-CRMs, they might be less likely so (Figure 2-10A). The similar results are seen from the same analysis in the mouse genome using full-length DHSs (Table 2-5).

Table 2-4. Summary of overlaps between DHS cores and TFBS islands, inter-TFBS spacers as well as non-CRMs in the human genome

	Total Numbers	Total Length (bp)	Overlapping length (bp)
TFBS islands	35,576,559	545,014,253	78,653,450 (14.4%)
inter-TFBS spacers	35,154,776	910,694,944	75,657,276 (8.3%)
non-CRMs	1,755,876	1,177,572,135	42,087,044 (3.6%)

Table 2-5. Summary of overlaps between full-length DHSs and TFBS islands, inter-TFBS spacers and non-CRMs in the mouse genome

	Total Numbers	Total Length (bp)	Overlapping length (bp)
TFBS islands	38,350,446	650,376,376	277,206,030 (42.6%)
inter-TFBS spacers	38,014,062	861,299,538	273,550,233 (31.8%)
non-CRMs	1,270,937	664,133,141	145,840,695 (22.0%)

Moreover, with an average length of 16 bp, the 4.5 million TF footprints comprising 2.1% of the genome are regions around the summit of DHSs, which are bounded by TFs and thus are protected from DNase I cleavage(65). As summarized in Table 2-6, TFBS islands have the highest proportion (5.6%) overlapping TF footprints, followed by inter-TFBS spacers (2.8%) and non-CRMs (1.3%). Thus, TFBS islands are highly (5.6% vs 2.1%) but inter-TFBS spacers are only slightly (2.8% vs 2.1%) enriched for TF footprints, while non-CRMs are (1.3% vs 2.1%) depleted of TF footprints. Taken together, these results suggest that some inter-TFBS spacers might be in the nucleosome form, and thus cannot be cut by DNase I or bounded by TFs; some other inter-TFBS spacers might be both nucleosome-free and TF-free, and thus can be cut by DNase I; and few inter-TFBS spacers might be bounded by TFs, and thus cannot be cut by DNase I (Figure 2-10A). Therefore, functions of inter-TFBSs, if any, might not mainly be carried out by direct

interactions with TFs.

Table 2-6. Summary of overlaps between TF footprints and TFBS islands, inter-TFBS spacers as well as non-CRMs in the human genome

	Total Numbers	Total Length (bp)	Overlapping length (bp)
TFBS islands	35,576,559	545,014,253	30,783,710 (5.6%)
Inter-TFBS spacers	35,154,776	910,694,944	25,706,710 (2.8%)
non-CRMs	1,755,876	1,177,572,135	15,126,676 (1.3%)

2.3 Discussion

In this study, we investigated the landscapes and organizations of CRMs in the human and mouse genomes as well as the architecture of constituent TFBSs within the CRMs. We reveal a few common rules for the organization of CRMs in the two genomes. First, like genes, the numbers and lengths of CRMs on chromosomes are correlated with the sizes of chromosomes. Second, CRMs are unevenly but correlatedly distributed with genes along chromosomes, forming mega-base-sized CRMs islands and deserts. Third, the numbers of CRMs and genes in TADs have stronger correlation than those in non-TAD regions. Fourth, like FANTOM enhancers and VISTA enhancers, CRMs are slightly biasedly distributed downstream of their nearest TSSs. Fifth, like FANTOM promoters and enhancers as well as VISTA enhancers, a small yet considerable portion (7%) of CRMs overlap TSSs, while the remaining vast majority do not. Based on this observation, we categorize the regulatory elements into two categories, i.e., CPC and CPL.

Sixth, promoters are traditionally classified into core promoters, proximal promoter elements and downstream promoter elements based on whether or not they overlap TSSs and their relative location to TSS, while enhancers are traditionally classified into distal enhancers and proximal enhancers based on their distances to the target TSSs(37, 46, 66). However, we find that only few (0.1%) of our CPC CRMs have a typical length of FANTOM promoters (<100 bp), and thus might be simple core promoters, while most (99.9%) of them are longer than 100 bp, and thus

might contain other regulatory elements in addition to core elements. On the other hand, most (>90.0%) of annotated TSSs overlap our CRMs. Thus, it appears that only a small portion (<10%) of core promoters exist alone, while most of them prefer to cluster with nearby regulatory elements to form longer CPC CRMs. Our classification of CRMs into the CPC and CPL categories does not differentiate traditional promoters and enhancers and their subtypes, but rather treats them as the same type of *cis*-regulatory sequences. This is consistent with growing evidence that enhancers and promoters share common molecular attributes, and are not functionally distinguishable in line of function and structure(67, 68). In this study, we also show that 12.2% of human and 8.0% of mouse VISTA enhancers contain core promoters, and thus are of the CPC category (Table 2-2). According to our classification, a CPC enhancer may contain a core promoter, and other regulatory element such as a proximal promoter element, a downstream promoter element and an enhancer element, and a CPL enhancer may also include a proximal promoter element or downstream promoter element and an enhancer element when it is close to a TSS. Finally, we find that CPC elements tend to be longer than CPL elements due at least partially to the fact that CPC elements contain core promoters in addition to other regulatory elements while CPL elements lack core promoters. However, containing a core promoter that has a mean length of ~100 bp cannot account for the difference in the mean lengths of CPC (2,408 bp) and CPL (650 bp) CRMs, thus other unknown reasons should exist and need to be elucidated in the future.

We also reveal a few common rules for the organization and architecture of TFBSs within the CRMs in the two genomes. First, adjacent TFBSs in a CRMs tend to overlap with each other, forming longer TFBS islands. This result is consistent with earlier reports in the fly(69, 70) and mammals including humans based on extensive overlaps of binding peaks of various TF ChIP-seq data(60). Besides, it provides valuable insights into the organization and characteristics of TFBSs

within putative CRMs and their potential roles in gene transcriptional regulation. In agreement with this, it has been shown that different TFs can compete for partially overlapping binding sites(71) or bind synergistically to the opposite faces of the DNA duplex(72). The adjacent putative TFBSs with a small portion of overlaps might be binding sites of different TFs for competitive binding, while those with a large portion of overlaps might be parts of a long TFBS, which our algorithm was unable to merge to form a long one. Second, TFBS islands comprise less than half (37.4% for human and 43.1% for mice) of CRMs in length, while the remaining majority positions in the CRMs are inter-TFBS spacers. Finally, inter-TFBS spacers within CRMs are under almost the same evolutionary constraints as are TFBS islands, suggesting that the same portion of positions in the spacers might also be functional. Although it is likely that some inter-TFBS spacers contain unknown TFBSs, most positions in the spacers do not appear to be TFBSs, since we find that the spacers are much less likely to overlap DHS cores than are TFBS islands, implying that a larger percentage of the spacers might be wrapped around histones than might TFBS islands. Moreover, we found that inter-TFBS spacers are much less likely to overlap TF footprints than TFBS islands, thus, a smaller percentage of the spacers might be bounded by TFBSs. While interactions between TFBSs and cognate TFs are critical in transcriptional regulation, adjacent spacers of TFBSs might have other functional roles other than direct TF binding. Indeed, it has been shown that inter-TFBS spacers could influence the conformation of adjacent TFBSs(54) and the interactions between adjacent TFBSs(55) (i.e. the regulatory grammar). That inter-TFBSs are as conserved as TFBSs suggests that the functions of inter-TFBS spacers, if any, depend on their sequences. Although many details of the functions of inter-TFBS spacers in CRMs remain to be elucidated, acknowledging their potential functions in transcriptional regulation beyond direct TF binding expands our understanding of the intricate transcriptional regulatory landscapes encoded

in genomes.

2.4 Conclusion

By analyzing our recently predicted unprecedentedly complete maps of CRMs and constituent TFBSs in the human and mouse genomes, we reveal common rules for the landscape and organization of CRMs on the human and mouse chromosomes as well as for the architecture inside of CRMs in terms of their constituent TFBSs and inter-TFBSs spacers. These findings will significantly advance the understanding of regulatory genomes and have a profound impact on the field of gene transcriptional regulation research.

2.5 Materials and Methods

2.5.1 The datasets

For the analysis in the human genome, we obtained 1,225,115 predicted CRMs and 124,923,659 constituent TFBSs at p-value = 0.05 and 1,755,876 predicted non-CRMs from our PCRM database(73). We downloaded the Hi-C interaction matrix in the K562 cell line from the ENCODE portal(74) with the accession ID ENCFF080DPJ. We downloaded 1,002 experimentally verified enhancers from the VISTA Enhancer database(60), as well as 32,684 enhancers and 184,326 promoters from the FANTOM project website(58, 59). We downloaded the precomputed 20,971,9847 LINSIGHT entries from the CshlSiepelLab GitHub repository at https://github.com/CshlSiepelLab/LINSIGHT?tab=readme-ov-filePrior. We obtained 3,591,899 unique DHS cores from ENCODE portal(74) with the accession ID ENCSR857UZV. We downloaded 4,465,728 TF footprints from https://www.vierstra.org/resources/dgf#appendix-file-format-descriptions.

For the analysis in the mouse genome, we obtained 798,257 predicted CRMs and 164,866,277 constituent TFBSs at p-value = 0.05 and 1,270,937 non-CRMs from our PCRM

database(21, 73). We downloaded the Hi-C interaction matrix in the CH12F3 cell line from ENCODE portal(74) with the accession ID ENCFF909ODS. We downloaded 702 experimentally verified enhancers from the VISTA Enhancer database(60), as well as 49,797 enhancers and 164,421 promoters from the FANTOM project website(58, 59). We downloaded a total of 14,532,289 DHSs from ENCODE portal(74) (Supplementary Table S2-7).

2.5.2 Generation of Manhattan plots

We assume that the total lengths and numbers of CRMs and genes within sliding windows in a genome are evenly distributed — our null hypotheses. To test that all these random variables are unevenly distributed in a genome — our alternative hypotheses, we computed z -values for each sliding window: $z = \frac{x-\mu}{\sigma}$, where x denotes the total length or number of CRMs or genes within the window, μ the mean and σ the standard deviation of the length or number of CRMs or genes in the windows in the genome assuming that our null hypotheses are true. Under the null hypotheses, the total lengths and numbers of CRMs and genes within sliding windows follow binomial distributions. Therefore, the means and standard deviations of the total lengths or numbers of CRMs and genes within sliding windows can be computed as $\mu = np$ and $\sigma = \sqrt{np(1-p)}$. For the total lengths of CRMs and genes in windows, n is the window size in bp, p the coverage defined as the total length of CRMs or genes in the genome divided by the genome size, which is 0.47 or 0.58, respectively in the human genome, and 0.55 or 0.45, respectively in the mouse genome. For the total number of CRMs or genes in windows, n is the number of CRMs or genes that can fit into a window, estimated as the size of the window divided by the mean length of CRMs or genes, i.e., $n = \frac{Window\ Size}{Mean\ Length}$. Then, $p = \frac{N}{Genome\ Size} \times Mean\ Length$, where N denotes the number of CRMs (1,225,115 and 798,257 for the human and mouse genome, respectively) or genes (63,313 and 55,364 for the human and mouse genome, respectively). The

size of the human and mouse genome is 3,088,269,832 and 2,725,521,370 bp, respectively. Subsequently, we generated Manhattan plots of these z-scores using the "qqman" library within R version 4.2.2. We defined thresholds of z>|5.2| for window size 1M bp and z>|4.7| for window size 10M bp as cutoffs for identifying CRM and gene islands and deserts for humans, corresponding to a p-value 0.01 to reject the null hypotheses after Bonferroni correction(75) for multiple hypothesis tests. The cutoffs for mice are z>|5.1| for window size 1M bp and z>|4.7| for window size 10M bp for identifying CRM and gene islands and deserts, corresponding to a p-value 0.01 to reject the null hypotheses after Bonferroni correction(75) for multiple hypothesis tests.

2.5.3 Generation of TADs

To generate TADs in the human genomes, we applied the Arrowhead algorithm of the Juicer tools(76) with the Knight-Ruiz Matrix Balancing (KR)(77) normalization method on the Hi-C interaction matrix in the K562 cell line at different resolutions: 5K bp, 10K bp, 25K bp, 50K bp and 100K bp. We then merged overlapping TADs into larger domains, resulting in a total of 944 merged TADs. To generate TADs in the mouse genome, we followed a similar procedure and generated 1,018 TADs using the Hi-C interaction matrix in the CH12F3 cell line. However, due to the unavailability of KR normalization for the Hi-C data in the CH12F3 cell line, we opted for the Vanilla-Coverage (VC) normalization method(78) instead.

2.5.4 Middle and nearer end distance between a CRM and its nearest TSS

2.5.4.1 The middle distance: The middle distance between a CRM and its nearest TSS is defined as $d_m = c_m(CRM) - c(TSS)$, if the nearest TSS is in the forward orientation, or $d_m = c(TSS) - c_m(CRM)$, if the nearest TSS is in the reverse orientation; where $c(TSS)$ and $c_m(CRM)$ are the coordinates of the nearest TSS and the middle point of the CRM, respectively.

2.5.4.2 The nearer end distance: The nearer end distance between a CRM and its nearest TSS (when they do not overlap) is defined as $d_e = c_e(\mathrm{CRM}) - c(\mathrm{TSS})$, if the nearest TSS is in the forward orientation, or $d_e = c(\mathrm{TSS}) - c_e(\mathrm{CRM})$, if the nearest TSS is in the reverse orientation; where $c(\mathrm{TSS})$ is the coordinate of the nearest TSS, and $c_e(\mathrm{CRM})$ is the coordinate of the nearer end of the CRM to the TSS.

2.5.5 Distance between adjacent TFBSs or TFBS islands in a CRM

For each CRM, we first arranged the constituent TFBSs or their TFBS islands by sorting them according to their starting coordinates. We then computed the distance d between two adjacent TFBSs or TFBS islands by subtracting the end coordinate of the current TFBS or TFBS island from the starting coordinate of the downstream TFBS or TFBS island. If $d < 0$, then the two adjacent sequences overlap each other by d bp. It is evident that for two adjacent TFBS islands $d > 0$, as they do not overlap each other.

2.5.6 LINSIGHT scores

As a LINSIGHT entry may cover multiple nucleotide positions, we assigned each of the positions within the loci the pre-computed LINSIGHT score. Since these entries may overlap, some positions may have multiple LINSIGHT scores. We calculated the average score for each position as its final LINSIGHT score if the position has more than one assigned score.

2.5.7 Overlaps between DHSs or TF footprints and TFBS islands, inter-TFBS spacers, and non-CRMs

For the analysis in the human genome, we obtained overlaps between DHS cores or TF footprints and TFBS islands, inter-TFBS spacers, and non-CRMs using bedtools2 version 2.29.0. However, for the analysis in the mouse genome, due to unavailability of DHS core or TF footprint data, we only aggregated the original DHSs from various cell/tissue types, and then merged these regions

to a set of non-redundant DHSs. We then obtained overlaps between the non-redundant DHSs and TFBS islands, inter-TFBS spacers, and non-CRMs using bedtools2 version 2.29.0.

2.6 Availability of data and materials

The datasets supporting the conclusions of this chapter are available at https://osf.io/7t8nm/ and are included within the chapter and its supplementary tables at https://github.com/sisyyuan/CRM_Dissertation.

Chapter 3

SIMULTANEOUS PREDICTION OF FUNCTIONAL STATES AND TYPES OF CRMs REVEALS THEIR PREVALENT DUAL USES AS ENHANCERS AND SILENCERS

3.1 Introduction

CRMs have long been characterized by using low throughput laborious molecular biology methods. However, recent advancements in a plethora of omics techniques have revolutionized our capacity to investigate CRMs. These techniques encompass: 1) ChIP-seq for identifying TFBSs(79-81) and regions modified by histone marks(26) in the genome; 2) DNase-seq(22-24) and ATAC-seq(25) for probing CA of genome regions; 3) Hi-C technology for measuring physical proximity between genomic loci in the nucleus(78, 82); and 4) RNA-seq for quantifying transcriptomes in cells/tissues(83). The wide adaptations of these techniques have resulted in vast volumes of data, originating from large consortia as well as individual laboratories worldwide(84-91). This wealth of data presents an unparalleled opportunity to reliably predict the location of CRMs in genomes, along with their functional states (on/active or off/inactive), types (predominantly enhancers or silencers), and target genes across diverse cell/tissue types(92, 93). Most existing methods attempted to simultaneously predict locations and functional states of enhancers in a given cell/tissue type by integrating multiple epigenetic marks including CA and various histone modifications(94-98). Although conceptually appealing, these one-step methods are limited for their high FDRs(19, 27-32). This is due to the fact that the presence of CA and histone marks, while informative, is not exclusively indicative for enhancers and their functional states, as these marks are also present in non-CRM sequences(29, 30, 33).

On the other hand, it has been shown that TF binding data are more informative for identifying loci of CRMs than CA and histone modification data(29-33). Furthermore, it has been established that accurate anchoring of a CRM's location through the binding of key TFs renders

epigenetic marks on the CRM a reliable predictor of its functional states(29-33). In light of these findings, we have recently introduced a two-step approach to predict CRMs and their functional states sequentially(19, 20). Firstly, we predict an accurate and more complete map of CRMs in the genome using TF binding data. Secondly, we predict the functional states of all the predicted CRMs in any given cell/tissue type of the organism using few epigenetic marks. For the first step, we have developed the dePCRM2 algorithm(19), which predicts loci of CRMs by integrating putative TF binding motifs identified in a large number of diverse TF ChIP-seq datasets using our ultra-fast motif finder ProSampler(99, 100). dePCRM2 is able to effectively segregate genomic regions covered by TF binding peaks into two exclusive sets: the CRM candidates and the non-CRMs(19). While dePCRM2 can predict a CRM's functional state in a specific cell/tissue type based on its overlaps with TF binding peaks available in the very cell/tissue type, this predictive capacity is often limited due to the scarcity of available TF binding datasets in most cell/tissue types. Therefore, similar to the cCREs identified recently by the ENCODE project(101), our predicted CRMs are generally cell/tissue type agnostic. For the second step, we have developed a machine learning model that can accurately predict the functional states of all the predicted CRMs as enhancers in diverse human and mouse cell/tissue types using only four epigenetic marks as features(20). This two-step approach significantly surpasses existing one-step methods in terms of sensitivity and specificity for predicting active enhancers in various cell/tissue types(20).

However, recent investigations have found that silencers are more prevalent than initially believed(102-104), and that an active enhancer in a cellular context could be an active silencer in another cellular context(9, 104). Thus, it is interesting to also predict functional states of CRMs as silencers. Indeed, a few diverse computational tools have been developed to predict active silencers in specific cell/tissue types(103). However, as in the case of enhancers, these methods attempted

to simultaneously predict the locations and functional states of silencers using epigenetic marks on candidate DNA segments(105-107). For example, a correlation-based method correlated putative active silencer mark (e.g., H3K27me3 and DHS) signals with the expression levels of neighboring genes across different cell/tissue types(105). A support vector machine (SVM) model was then trained using sequence features as well as features derived from the aforementioned correlation-based method to predict silencers(105). Additionally, a simple subtractive approach (SSA) excluded genome regions with enhancer chromatin signatures to be putative silencers and a gapped k-mer SVM (gkmSVM) was trained on massively parallel reporter assay (MPRA) data and sequence patterns to predict silencers(107). However, the accuracy of these one-step methods is quite low (see later), due to similar reasons for predicting enhancers and their functional states using one-step methods.

Building upon the success of our two-step strategy in predicting enhancers and their functional states in cell/tissue types, we now extend our machine learning model to simultaneously predict the functional types (enhancer or silencer) and states (on/active or off/inactive) in various cell/tissue types in a genome-wide fashion. Our methods achieve an area under the receiver operator characteristic curve (AUROC) > 0.96 and show superior performance to state-of-the-art methods. Using the tools, we predicted functional types and states of 1.2M CRMs in 107 human cell/tissue types. Our results indicate that silencers and dual functional CRMs are more prevalent than previously thought and that various types of CRMs display distinct properties in terms of their lengths and TFBS densities, reflecting their functional complexity.

3.2 Result

3.2.1 Functional states of CRMs as silencers and enhancers can be accurately predicted using three epigenetic marks

We have previously developed an LR model to predict functional states in a cell/tissue type of our 1.2 M predicted CRMs in the human genome(19) using signals of few epigenetic marks on the CRMs as features that are more or less associated with active enhancers. We employed a similar LR model (Figure 3-1A) to predict functional states of the CRMs as silencers in a cell/tissue type using three epigenetic marks (CA, H3K9me3 and H3K27me3) on the CRMs as features. We pooled positive and negative silencer sets compiled in each of 40 of the 67 human cell/tissue types with the required data available (Materials and Methods), resulting in a positive set containing a total 256,766 positive silencers and a negative set with the same number of negative control sequences. As shown in the UpSet plot in Figure 3-1B, H3K27me3 peaks pooled from the 40 cell/tissue types have the highest coverage of the human genome, followed by CA and H3K9me3 peaks, and around 100 Mb of the genome are covered by the peaks of all the three marks. To evaluate the ability of these three marks to predict the functional states of the CRMs as silencers in cell/tissue types, we trained and evaluated the seven models using all the seven possible combinations of one, two and three of the three marks as features by 10-fold cross validation (Figure 3-1B). Of the three models using only one mark, model 2 using CA had the highest median AUROC of 0.948, followed by model 1 using H3K27me3 (median AUROC=0.826) and model 3 using H3K9me3 (median AUROC=0.685). Thus, CA alone has quite high prediction accuracy, while H3K27me3 alone and particularly, H3K9me3 alone have only intermediate prediction accuracy. Of the three models using two marks, model 4 using CA and H3K27me3 (CA&H3K27me3) obtained the highest median AUROC of 0.960, followed by model 6 (CA&H3K9me3, median AUROC=0.951) and model 5 (H3K9me3&H3K27me3, median AUROC= 0.919). Model 7 using all the three marks (CA&H3K9me3&H3K27me3) achieved the highest median AUROC of 0.962 (Figures 3-1B, 3-1C), which is significantly higher than the other

six models (p value < 0.01, Mann-Whitney U test). Consistently, CA in model 7 had a much higher weight (102.7) than H3K27me3 (16.5) and H3K9me3 (10.9) (Figure 3-1D). We thus selected model 7 as our silencer functional state predictor in the subsequent predictions. The numbers of predicted active silencers in these 40 cell/tissue types were greater than those of positive silencers compiled in them (Figure 3-1E), suggesting that the positive silencers used for training and testing consist of only a small portion of active silencers in these cell/tissue types.

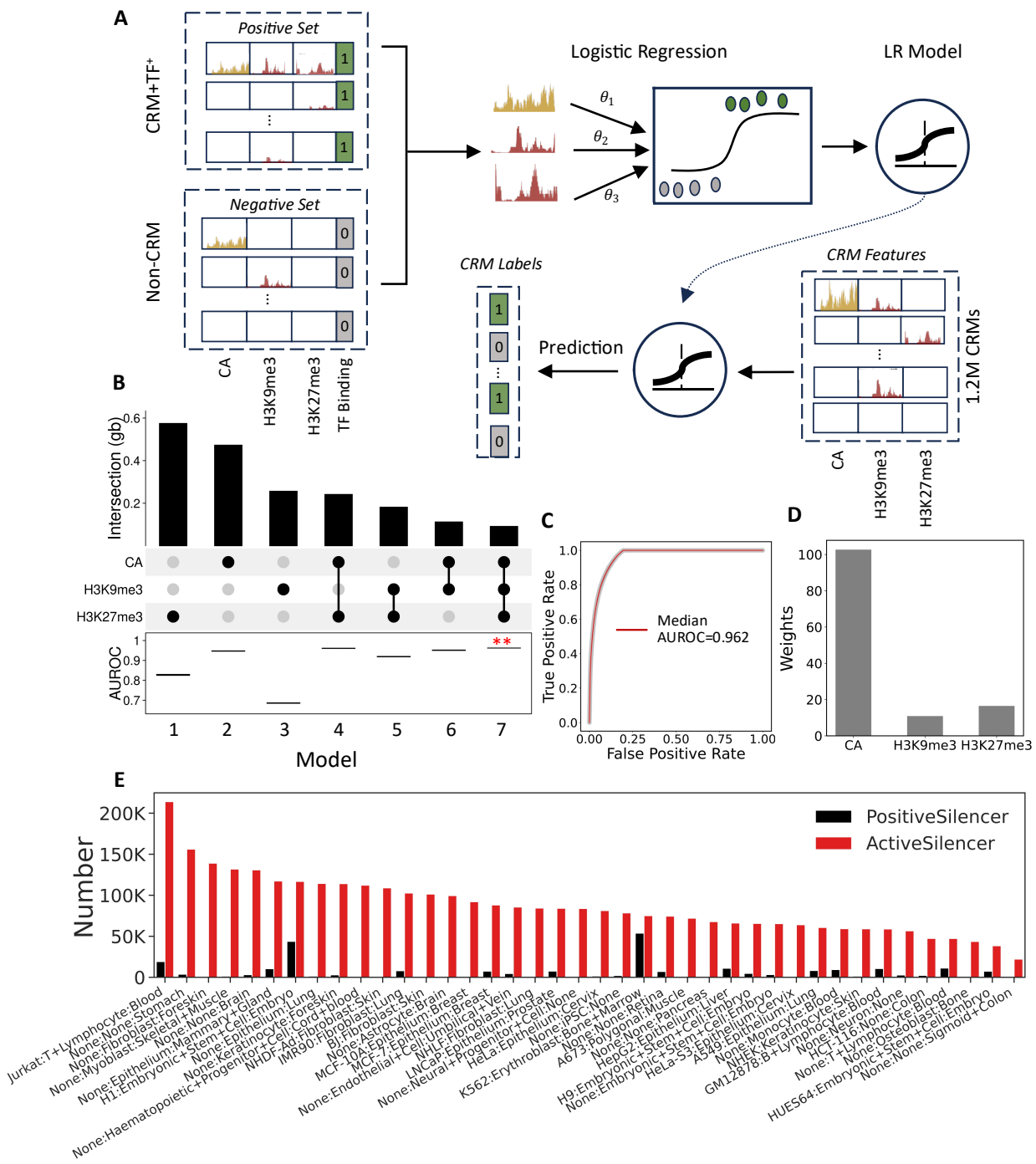


Figure 3-1. The epigenetic marks can accurately predict the functional states of putative silencers. **A.** A cartoon illustrating the workflow of our LR model using CA, H3K9me3 and H3K27me3 signals as the features. **B.** The UpSet plot showing intersection sizes (Gb) of mark peaks (upper bar graph) and the boxplot of AUROCs of the seven LR models using all possible combinations of the three epigenetic marks with 10-fold cross validation (lower boxplots). **: p value < 0.01, Mann-Whitney U test. **C.** ROC curves of model 7 using all the three marks. The red curve is the median ROC curve from the results of 10-fold cross validation. The AUROC curve of each fold is

invisible since these curves have almost the same shape as the median curve. **D.** Bar graph of the weights of CA, H3K9me3 and H3K27me3 in model 7. **E.** Bar graph of the numbers of positive and active silencers compiled and predicted, respectively, in each of the 40 cell/tissue types.

Although we have successfully used four epigenetic marks (CA, H3K4me1, H3K4me3 and H3K27ac) as features to predict functional states of our CRMs as enhancers(20), in this study we only used three of them (CA, H3K4me1 and H3K27ac) in our LR model. We excluded H3K4me3, since it is more likely associated with promoters than to enhancers(108, 109). We pooled positive and negative sets compiled in each of the 67 human cell/tissue types used in our previous study(20) (Materials and Methods), resulting in a positive set containing a total 1,415,796 positive enhancers and a negative set with the same number of negative control sequences. We trained and evaluated the seven LR models using all the seven possible combinations of the three epigenetic marks as features by 10-fold cross validation. Of the three models using only one mark, model 4 using CA had the highest median AUROC 0.913, followed by model 1 using H3K4me1 (median AUROC=0.897) and model 2 using H3K27ac (median AUROC=0.866). Of the three models using two marks, model 3 using H3K4me1 and H3K27ac (H3K4me1&H3K27ac) obtained the highest median AUROC of 0.971, followed by model 5 (CA&H3K4me1, median AUROC=0.963) and model 6 (CA&H3K27ac, median AUROC=0.952). Model 7 using all of the three marks achieved the highest median AUROC of 0.977 (Figures 3-2A and 3-2B), which is significantly higher than the other six models (p value < 0.01, Mann-Whitney U test). Consistently, CA has a higher weight (92.0) in the model than H3K4me1 (30.0) and H3K27ac (17.1) (Figure 3-2C). The median AUROC value achieved by model 7 (0.977) is comparable with our previous model (0.986) using four epigenetic marks, which substantially outperforms five existing state-of-the-art methods(20). We thus selected model 7 as our enhancer functional state predictor (enhancer predictor) for the subsequent predictions. As expected, the numbers of predicted active enhancers in 65 of the 67 cell/tissue types are greater than those of positive enhancers compiled in them (Figure 3-2D),

suggesting that the positive enhancers used for training and testing consist of only a small portion of active enhancers in most of the cell/tissue types. However, both the positive sets and predicted active enhancers in each of these cell/tissue type are smaller than those compiled and predicted in our earlier study(20), due to the more stringent criterion used to compile the positive sets to ensure the CRMs in the positive sets are true active enhancers.

In summary, CA alone is a more effective predictor for both active silencers (Figure 3-1B) and active enhancers (Figure 3-2A) than the other two marks (H3K27me3 and H3K9me3 for silencers and H3K4me1 and H3K27ac for enhancers) alone. Using additional two histone marks could moderately improve the enhancer prediction accuracy (mean AUROC 0.977 vs 0.913), but only slightly increase the silencer prediction accuracy (mean AUROC 0.962 vs 0.948), over that obtained by using CA alone. In both predictors CA has overwhelmingly higher weights than the other two marks, making the other two marks weaker predictors.

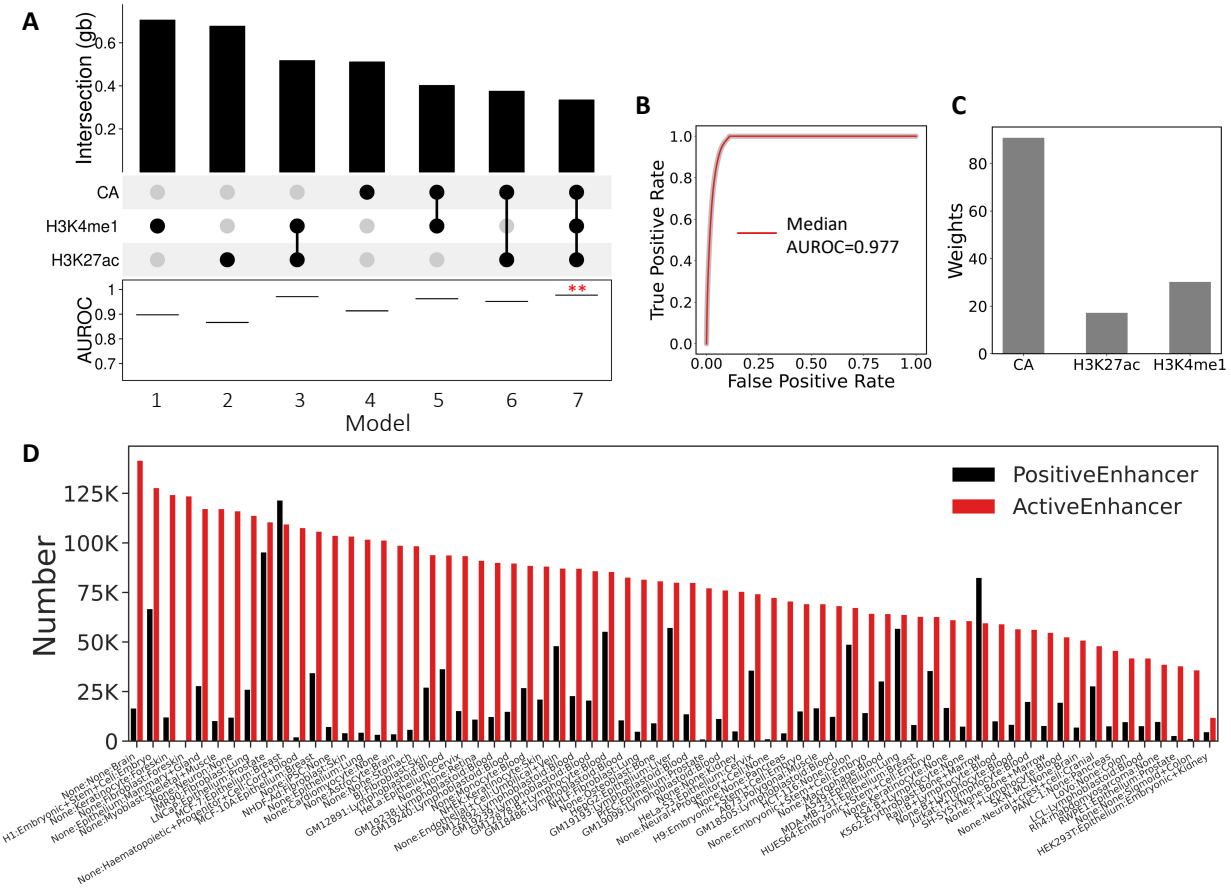


Figure 3-2. The epigenetic marks can accurately predict the functional states of enhancers. **A.** UpSet plot showing intersection sizes (Gb) of the three types of epigenetic mark peaks (upper bar graph) and the boxplot of AUROCs of the seven LR models using all possible combination of the three epigenetic marks as features with 10-fold cross validation (lower boxplots). **: p value < 0.01, Mann-Whitney U test. **B.** ROC curves of model 7 using CA, H3K4me1 and H3K27ac as features. The red curve is the median ROC curve from the results of 10-fold cross validation. The AUROC curve of each fold is invisible since these curves have almost the same shapes as the median curve. **C.** Bar plots of the weights of CA, H3K27ac and H3K4me1 in model 7. **D.** Bar plots of numbers of positive and active enhancers compiled and predicted, respectively, in each of the 67 cell/tissue types (Materials and Methods).

3.2.2 Varying portions of the 1.2M CRMs are active as enhancers or silencers in various cell/tissue types

Using our enhancer and silencer predictors trained on the pooled positive and negative sets in the 67 and 40 (Supplementary Tables S3-1 to S3-3) cell/tissue types, respectively, with the required epigenetic data available from Cistrome (110, 111), we predicted the functional states of our 1.2M predicted CRMs(19) as enhancers in 105 cell/tissue types and as silencers in 58 cell/tissue types,

respectively, with the required epigenetic data available from ENCODE (74) (Figure 3-3A). This yielded highly varying numbers of active enhancers in the 105 cell/tissue types ranging from 31,947 (2.7% of the 1.2 M CRMs) in tibial nerve cells to 168,471 (14.3%) in motor neuron cells, with a median of 95,496 (8.1%) in a cell/tissue type (Supplementary Table S3-4). We predicted a total of 10,068,782 active enhancers in the 105 cell/tissue types. After removing the redundancy, we ended up with a total of 695,507 (59.0%) non-redundant active enhancers from the 105 cell types. Moreover, we also predicted highly varying numbers of active silencers in the 58 cell/tissue types ranging from 27,843 (2.4%) in tibial nerve cells to 197,133 (16.7%) in HepG2 cells, with a median of 86,668 (7.4%) in a cell/tissue (Supplementary Table S3-5). We predicted a total of 5,096,269 active silencers in the 58 cell/tissue types. After removing the redundancy, we ended up with a total of 677,840 (57.5%) non-redundant active silencers from the 58 cell types. Thus, we predicted a slightly higher median number of active enhancers than active silencers (95,496 vs 86,668). As expected, most (78.0%~97.2%) of the CRMs were not active either as enhancers or as silencers in a cell/tissue type (Supplementary Table S3-6). In total, we predicted the functional types and states for 868,944 (73.8%) of the 1.2M CRMs as active enhancers or active silencers in at least one of the cell/tissue types.

3.2.3 Predicted functional types and states of CRMs are reflected by their epigenetic mark signals

Notably, in each of the 49 cell/tissue types with only enhancer marks (CA, H3K4me1 and H3K27ac) data available (Materials and Methods), we were only able to predict each CRM either as an active enhancer or as an inactive enhancer (Figure 3-3A, Supplementary Table S3-7). For example, in the A549 cells, we predicted 128,601 (10.9%) CRMs to be active enhancers and the remaining 89.1% to be inactive enhancers (Supplementary Table S3-7). The predictions in each of these 49 cell/tissue types are reflected by the signal patterns of all the three epigenetic marks on

the CRMs. Figure 3-3B shows the case in the A549 cells from donor ENCDO000AAZ as an example. Specifically, CRMs that were predicted to be active enhancers in a cell/tissue type such as the A549 cells were enriched in the active enhancer marks (CA, H3K4me1 and H3K27ac), while those that were not, were depleted of these signals (Figure 3-3B). Similarly, in each of the two cell/tissue types with only putative active silencer marks (CA, H3K27me3 and H3K9me3) data available (Materials and Methods), we were only able to predict each CRM either as an active silencer or as an inactive silencer (Figure 3-3A, Supplementary Table S3-8). For example, in the heart left ventricle cells from donor ENCDO039RUH, we predicted 108,302 (9.2%) CRMs to be active silencers and the remaining 90.8% to be inactive silencers (Supplementary Table S3-8). The predictions in these two cell/tissue types also are reflected by the signal patterns of all the three epigenetic marks on the CRMs. Figure 3-3C shows the case for the heart left ventricle cells. Specifically, CRMs that were predicted to function as active silencers in a cell/tissue type such as heart left ventricle cells were enriched in putative active silencer marks (CA, H3K27me3 and H3K9me3), while those that were not, were depleted of the signals (Figure 3-3C).

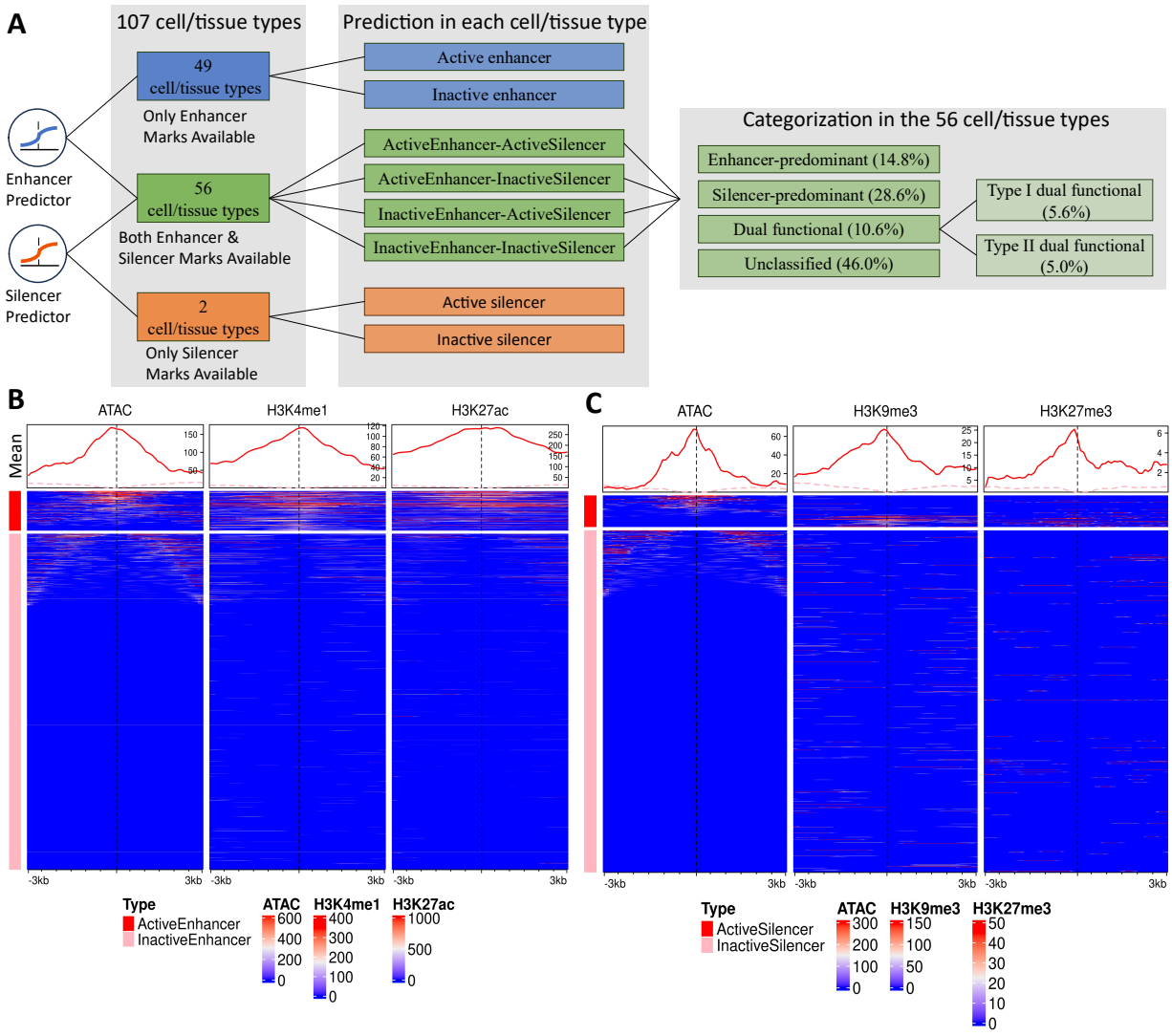


Figure 3-3. Prediction of functional types and states of CRMs in the 107 cell/tissue types**. A.** Prediction of functional types and states of CRMs in various cell/tissue types with different available data and categorization of CRMs based on predictions in the 56 cell/tissue types with both active enhancer and silencer marks data available. **B.** Heatmaps of signals of three active enhancer marks in a 6 kb window centering on the middle points of the predicted active enhancers and inactive enhancers in A549 cells from donor ENCDO000AAZ. **C.** Heatmaps of signals of three putative active silencer marks in a 6 kb window centering on the middle points of the predicted active silencers and inactive silencers in the heart left ventricle cells from donor ENCDO039RUH. The heatmaps show the mean signals of the epigenetic marks in each 100 bp sliding window along each sequence; the line plot shows the mean signal of each window at a position in all the sequences in the set (Materials and Methods). The color code for the types in the line plot above each column is the same as the left legends of the heatmaps.

Furthermore, in each of the 56 cell/tissue types with both enhancer and putative silencer marks data available (Materials and Methods), we have four possible predictions about the functional types and states of each of the 1.2M CRM (Figure 3-3A): i) both the enhancer predictor and the silencer predictor predict it to be active (ActiveEnhancer-ActiveSilencer); ii) the enhancer predictor predicts it to be active, but the silencer predictor predicts it to be inactive (ActiveEnhancer-InactiveSilencer); iii) the enhancer predictor predicts it to be inactive, but the silencer predictor predicts it to be active (InactiveEnhancer-ActiveSilencer); and iv) both the enhancer predictor and the silencer predictor predict it to be inactive (InactiveEnhancer-InactiveSilencer). The numbers of predicted CRMs in each of the categories are shown in Figure 3-4A (Supplementary Table S3-9). For example, in the MCF-7 cells from donor ENCDO000AAE, we predicted 69,327 (5.9%) CRMs to be "ActiveEnhancer-ActiveSilencer", 45,278 (3.8%) to be "ActiveEnhancer-InactiveSilencer", 45,862 (3.9%) to be "InactiveEnhancer-ActiveSilencer", and the remaining 1.02M (86.4%) to be "InactiveEnhancer-InactiveSilencer" (Figure 3-4A, Supplementary Table S3-9). The predictions in each cell/tissue type also are reflected by the relevant epigenetic marks on the CRMs. Figures 3-4B and 3-4C show the cases in the MCF-7 cells as examples. Specifically, "ActiveEnhancer-ActiveSilencer" CRMs were enriched in both the marks of active enhancers (CA, H3K4me1 and H3K27ac) and marks of putative active silencers (CA, H3K9me3 and H3K27me3) (Figure 3-4B). "ActiveEnhancer-InactiveSilencer" CRMs were enriched in marks of active enhancers H3K4me1 and H3K27ac but depleted of marks of putative active silencer H3K9me3 and H3K27me3 (Figure 3-4B). Interestingly, the CA signals on these "ActiveEnhancer-InactiveSilencer" CRMs were weak in the middle but strong at the two flanking regions, while the H3K4me1 signals were narrowly peaked at the middle, suggesting that the middle of these CRMs might not be nucleosome free, and therefore could not be cut by transposase.

"InactiveEnhancer-ActiveSilencer" CRMs were enriched in putative active silencer marks but depleted of active enhancer marks H3K4me1 and H3K27ac (Figure 3-4C). "InactiveEnhancer-InactiveSilencer" CRMs had weak signals of all the five epigenetic marks (Figure 3-4C). In summary, by predicting the functional states of the 1.2M CRMs as enhancers or silencers in a cell/tissue type, we are able to simultaneously predict the functional states and types of the CRMs using only five epigenetic marks data in the very cell/tissue type.

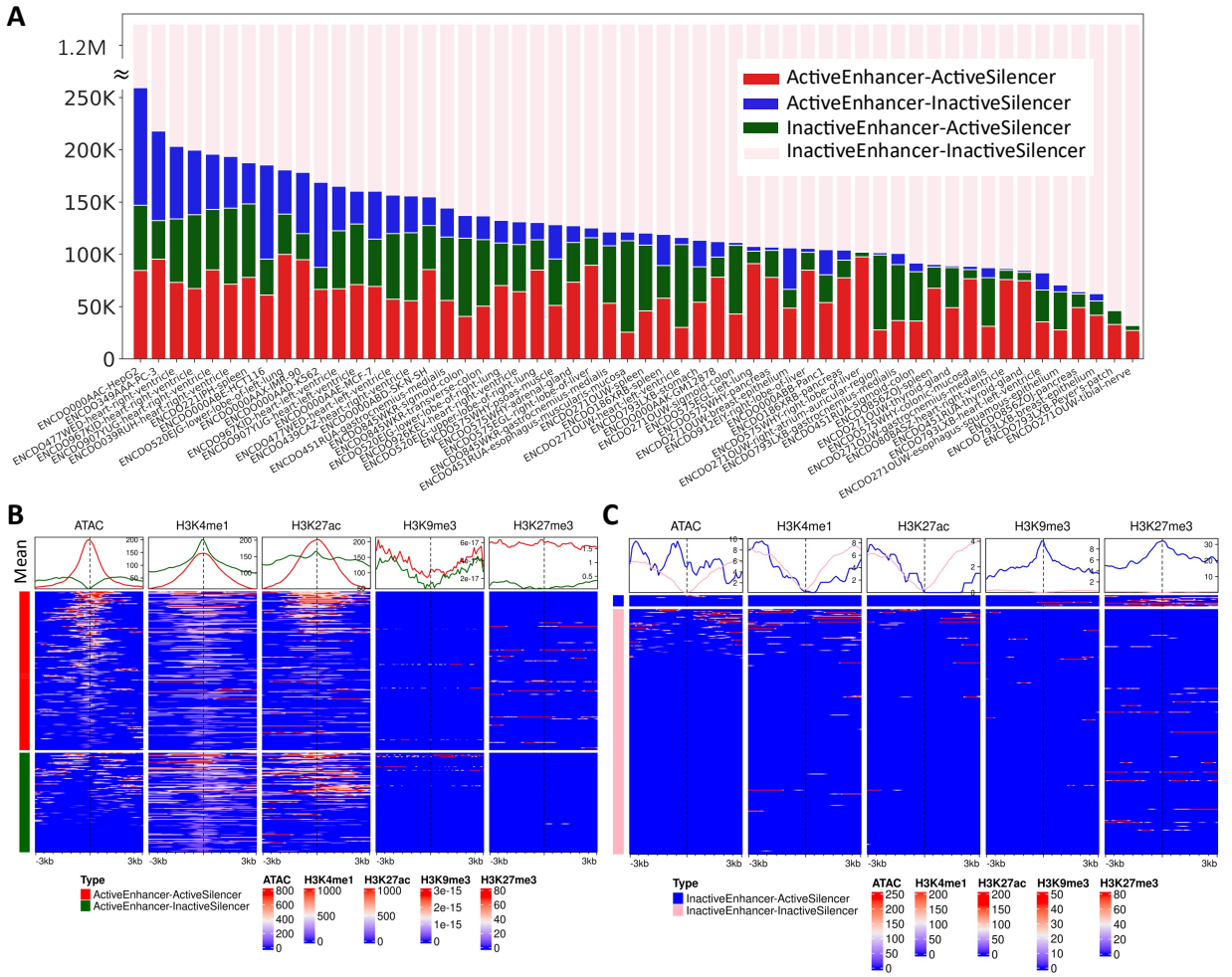


Figure 3-4. Four possible combinations of predictions of the functional types and states of the CRMs in the 56 cell/tissue types with both active enhancer and putative active silencer marks data available. **A.** Bar plots of the numbers of the CRMs with the four possible combinations of predicted functional types and states in each of the 56 cell/tissue types. **B.** Heatmaps of signals of the five epigenetic marks in a 6 kb window centering on the middle of the predicted ActiveEnhancer-ActiveSilencer and ActiveEnhancer-InactiveSilencer CRMs in MCF-7 cells. **C.** Heatmaps of signals of the five marks in a 6 kb window centering on the middle of the predicted

InactiveEnhancer-ActiveSilencer and InactiveEnhancer-InactiveSilencer CRMs in MCF-7 cells. The heatmaps show the mean signals of the epigenetic marks in each 100 bp sliding window of each sequence; the line plot shows the mean signal of each window at a position of all the sequences in the set (Materials and Methods). The color code for the types in the line plot above each column is the same for the heatmaps.

3.2.4 At least 10% of the CRMs are dual functional

We predicted a total of 793,140 (67.3%) CRMs to be active as enhancers and/or silencers in at least one of the 56 cell/tissue types (Figure 3-3A, Supplementary Table S3-9) with both active enhancer and putative silencer marks data available. These predictions provide us an opportunity to investigate the predominant roles of the CRMs used as enhancers, silencers, or both in these cell/tissue types. Of these 793,140 CRMs, 117,646 (14.8%) were predicted to be active only as enhancers across all the cell/tissue types (Enhancer-predominant), 227,211 (28.6%) were predicted to be active only as silencers across the cell/tissue types (Silencer-predominant), and 448,283 (56.6%) were predicted to be active both as enhancers and silencers in the 56 cell/tissue types (Dual functional CRMs) (Figure 3-3A). Of the 448,283 dual functional CRMs, 408,451 (91.1%) were predicted to be both as active enhancers and active silencers in the same cell/tissue types (denoted as type I dual functional CRMs), while the remaining 39,832 (8.9%) were predicted to be as active enhancers in some cell/tissue types and as active silencers in other cell/tissue types (denoted as type II functional CRMs). As we indicated earlier, CA signals have much higher weights than the two other markers in both our enhancer (Figure 3-2C) and silencer (Figure 3-1D) predictors, thus a CRM with a very strong CA signal but relatively weak signals of the two other marks can be predicted both as active enhancer and as active silencer in the same cell/tissue type. To reduce possible false positives, for type I dual functional CRMs, we only consider those that have at least one active enhancer mark and at least one putative active silencer mark for further analysis, yielding a total of 44,153 (10.8%) more stringent type I dual functional CRMs, while the remaining 89.2% were categorized as unclassified CRMs (Figure 3-3A). Our subsequent analyses

will mainly focus on a total of 83,985 (10.6%) dual functional CRMs (44,153 type I and 39,832 type II) (Figure 3-3A).

3.2.5 Dual functional CRMs can switch their roles in different cellular contexts

It has been shown in previous reports that CRMs may switch their roles between active enhancers and active silencers in different cellular contexts(9, 10, 104). Consistently, our type II dual functional CRMs functioned as active enhancers in some cell/tissue types while as active silencers in other cell/tissue types. On the other hand, all of the 44,153 type I dual functional CRMs could function both as active enhancers and as active silencers in at least one of the 56 cell/tissue types, most (89.1%) of which could also only function as active enhancers or active silencers in at least one of the 56 cell/tissue types. Moreover, more than half (56%) of the type I dual functional CRMs have dual functions in the same cell/tissue type in only one cell/tissue type, and only a small portion (3.8%) of them did so in more than half (28) of the 56 cell/tissue types (Figure 3-5A). Thus, most type I dual functional CRMs also were able to switch their roles in different cell/tissue types. For example, the CRM at chr7:1,428,212-1,430,677 was dual functional in spleen (ENCDO221IPH) and HepG2 (ENCDO000AAC) cells, but only functioned as an active enhancer in Panc1 (ENCDO000ABB) cells, and only functioned as an active silencer in heart left ventricle (ENCDO477WED), SK-N-SH (ENCDO000ABD) and breast epithelium (ENCDO271OUW) cells, as indicated by the patterns of epigenetic marks on the CRM in relevant cell/tissue types. More specifically, as shown in Figure 3-5B as examples, in spleen cells the CRM was heavily marked by both active enhancer marks (H3K4me1) and active silencer marks (H3K27me3) in addition to CA, while in Panc1 cells it was heavily marked by the active enhancer marks (H3K4me1 and H3K27ac), but depleted of active silencer marks, and in heart left ventricle cells it

was heavily marked by the active silencer marks (H3K27me3), but with weak active enhancer marks.

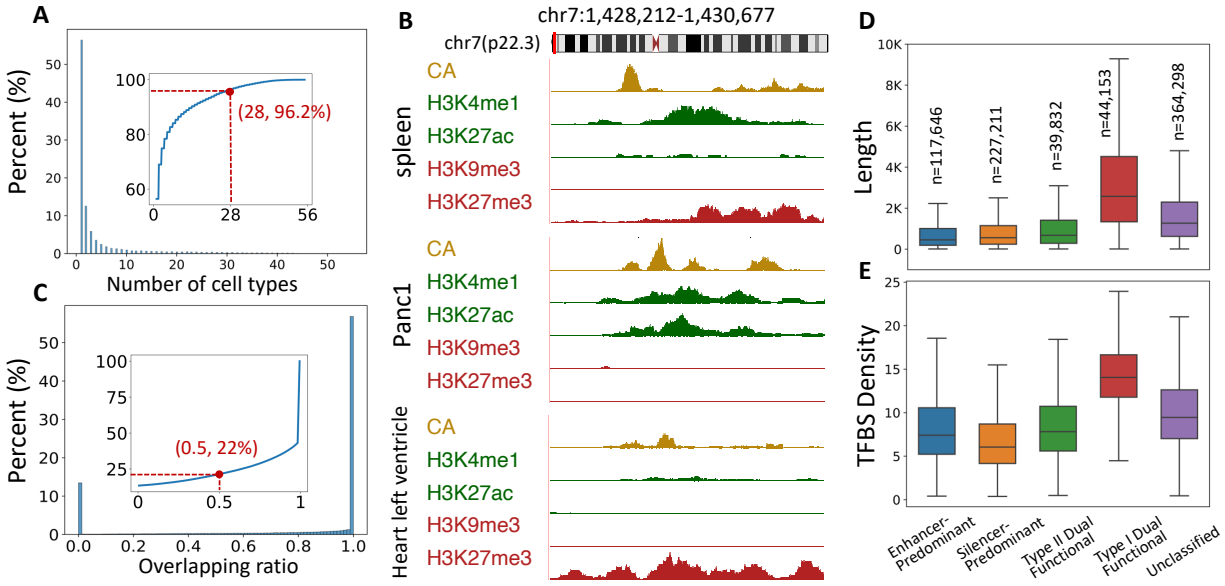


Figure 3-5. Analysis of different types of active CRMs in the 56 cell/tissue types. **A.** Histogram of percentages of type I dual functional CRMs that are active both as enhancers and silencers in the indicated number of cell/tissue types. The inset plot is cumulative percentage of type I dual functional CRM that are active both as enhancers and silencers in at least the indicated numbers of cell/tissue types. **B.** Epigenetic marks on a type I dual functional CRM at chr7:1,428,212-1,430,677 in spleen cells where it functions both as an active enhancer and as an active silencer (upper panel); in Panc1 where it functions only as an active enhancer (middle panel); and in heart left ventricle cells where it functions only as an active silencer (lower panel). **C.** Histogram of percentages of type I dual functional CRMs with the indicated overlapping ratio between their enhancer mark regions and silencer mark regions. The inset plot is the cumulative percentage of overlapping ratio less than the indicated numbers. **D.** Boxplots of the lengths of the four types of CRMs and unclassified CRMs. **E.** Boxplots of the density of TFBSs (number of TFBSs per 100 bp) in the four types of CRMs and unclassified CRMs.

3.2.6 Enhancer and silencer mark peaks on type I dual functional CRMs overlap each other

As we predicted CRMs by stitching adjacent TFBSs(19), it is conceivable that type I dual-functional CRMs may be the result of simply merging distinct enhancers and silencers. To test this possibility, we assessed the extent to which active enhancer mark peaks and putative active silencer mark peaks along type I dual functional CRMs overlap each other (Materials and Methods). As shown in Figure 3-5C, of type I dual functional CRMs, 55% have their active enhancer and silencer

mark peaks completely overlapping each other (overlapping ratio = 1), and more than 78% have an overlapping ratio larger than 0.5, indicating that type I dual functional CRMs are unlikely formed by incorrectly stitching adjacent enhancers and silencers. For example, active enhancer and silencer marks interdigitate and overlap one another along the CRM chr7:1,428,212-1,430,677 in the spleen cells where it functions both as an active enhancer and as an active silencer (Figure 3-5B). These results indicate that the dual functions of a CRM might be achieved by the collaboration between different parts of the CRM, but not by two non-overlapping parts each conferring the CRM a different role. It is likely that such collaboration renders the dual functional CRMs to be longer than enhancer-predominant and silencer-predominant CRMs.

3.2.7 Length and TFBS density of a CRM reflect the complexity of its functional type

To see how the length of a CRM is related to its predicted functional type, we plotted the distributions of the lengths of the different types of predicted CRMs. Interestingly, as shown in Figure 3-5D, different types of CRMs show distinct length distributions. Specifically, type I dual functional CRMs have the longest median length (2,577 bp), followed by type II dual functional CRMs (678 bp), silencer-predominant CRMs (558 bp), and enhancer-predominant CRMs (454 bp). Unclassified CRMs are shorter than type I dual functional CRMs, yet longer than the other three types, suggesting that they might be a blend of type I dual functional CRMs and other three types of CRMs. The longer lengths of dual functional CRMs might be related to their more complex functions. Moreover, it has been demonstrated that enhancers and silencers in mouse retinal photoreceptor cells (cones and rods) possess different information content in terms of TFBS composition(112). In light of this, we analyzed TFBS densities (number of TFBSs in 100 pb length, Materials and Methods) of enhancer-predominant CRMs, silencer-predominant CRMs, type I and type II dual functional CRMs, and unclassified CRMs. As shown in Figure 3-5E, type I dual

functional CRMs have the highest TFBS densities, followed by type II dual functional CRMs, enhancer-predominant CRMs and silencer-predominant CRMs. Likewise, unclassified CRMs exhibit a lower TFBS density than type I dual functional CRMs, but a higher TFBS density than the other three types (Figure 3-5E), suggesting again that unclassified CRMs might be a combination of type I dual functional CRMs and other three types of CRMs.

3.2.8 Type I dual functional CRMs might execute dual functions by regulating different genes in the same cell type

Of the 56 cell/tissue types that we used to predict type I dual functional CRMs (Figure 3-3A), 9 are cell lines, each nominally contains a single cell type; while the remaining 47 are primary tissues, each might contain multiple cell types. Thus, we compared the numbers of type I dual functional CRMs that functioned as both active enhancers and active silencers in the cell lines (n=9) with those in the primary tissues (n=47). As shown in Figure 3-6A, the numbers of dual active CRMs observed in the cell lines were not significantly different (p-value>0.26) from those in the primary tissues. This suggests that the dual functions of CRMs may not necessarily be attributed to various cell types in a primary tissue, rather, CRMs can be dual functional in the same cell type.

To investigate how dual functional CRMs could possibly exert their enhancer and silencer functions, we identified genes whose promoters were in close physical proximity to a dual functional CRMs from the Hi-C interaction map and compared expression levels of putative target genes in different cell/tissue types according to the CRM's predicted functional states as an enhancer or a silencer of the genes. For instance, the Hi-C interaction map shows that CRM chr7:1,428,212-1,430,677 is in close physical proximity with the promoters of genes *MICALL2, INST1, MAFK,* and *PSMG3* (Figure 3-6B). Figure 3-6C shows the expression levels of these genes across diverse cell/tissue types based on the CRM's functional states as an enhancer for these genes. *INST1* and *PSMG3* had significantly higher expression levels in cell/tissue types where the CRM

was active as an enhancer than in cell/tissue types where the CRM was inactive as an enhancer, while *MICALL2* and *MAFK* did not. This leads us to conclude that the CRM might function as an enhancer for *INST1* and *PSMG3*, but not for *MICALL2* and *MAFK*. Similarly, Figure 3-6D shows the expression levels of the four genes across different cell/tissue types based on the CRM's functional states as a silencer for the genes. *MICALL2* had significantly lower expression levels in cell/tissue types where the CRM was active as a silencer than in cell/tissue types where the CRM was inactive as a silencer, while the other three genes did not. We therefore conclude that the CRM might function as a silencer for *MICALL2*, but not for the other three genes. Although the CRM interacts with gene *MAFK*, our result suggests that it may not regulate this gene as either an enhancer or a silencer in the cell/tissue types that we examined.

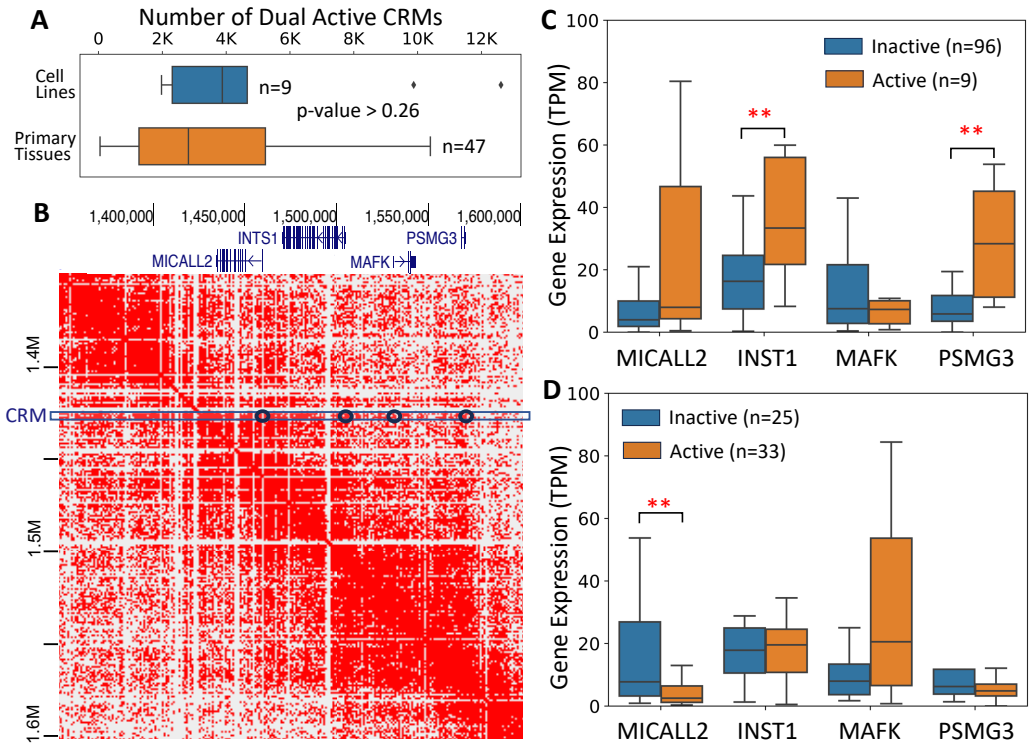


Figure 3-6. Comparison of numbers of type I dual functional CRMs in cell lines and primary tissues and an analysis of expression levels of putative target genes of a type I dual functional CRM. **A.** Boxplots of numbers of type I dual functional CRMs predicted in 9 cell lines and 47 primary tissues. p-value>0.26 (Mann-Whitney U-test). **B.** A Hi-C interaction map for the region of 1,350,000 to 1,60,000 bp on chromosome 7. CRM chr7:1428212-1430677 indicated by the box interacts with the promoters of genes *MICALL2*, *INST1*, *MAFK* and *PSMG3*, highlighted by the

circles. **C.** Boxplots of expression levels of genes *MICALL2*, *INST1*, *MAFK* and *PSMG3* across different cell/tissue types based on the CRM's functional states as an enhancer in these cell/tissue types. **: p-value <0.01 (Mann-Whitney U-test). **D.** Boxplots of expression levels of genes *MICALL2*, *INST1*, *MAFK* and *PSMG3* across different cell/tissue types based on the CRM's functional states as a silencer in these cell/tissue types. **: p-value <0.01 (Mann-Whitney U-test).

3.2.9 The "validated" silencers from silencerDB may contain false positives

We have previously shown that our CRM functional state predictor that used four active enhancer marks substantially outperformed five state-of-the-art methods(20). As our enhancer predictor trained on a more stringent positive set using three active enhancer marks achieved comparable AUROC value (0.980) to that of our previous predictor (0.986) in the same dataset, to avoid repetition, here we only evaluated the performance of our silencer predictor. To this end, we first compared our 677,840 (57.5%) predicted silencers pooled from the 58 cell/tissue types (Figure 3-3A) with the "validated" silencers from the silencerDB database(113). There were 8,588 "validated" unique silencers in silencerDB, which were identified by two recent studies using MPRA(106) or its variant called repressive ability of silencer elements (ReSE) screen(107). Our predicted silencers overlap 4,661 (54.3%) of the "validated" silencers by at least a 1bp, while only 5,525 (64.3%) of the "validated" silencers overlap 5,434 (0.5%) of our predicted 1.2 M CRMs. To see whether the rest 3,063 (35.7%) that do not overlap our CRMs are really functional, we analyzed their evolutionary behaviors using the phyloP scores. As shown in Figure 3-7A, like our predicted silencers the "validated" silencers that at least partially overlap our CRMs are under strong selection as indicated by their broadly distributed phyloP scores(62). By contrast, the remaining "validated" silencers that do not overlap our CRMs are more selectively neutral as indicated by their narrowly distributed phyloP scores around 0 (Figure 3-7A)(62). We thus posit that the "validated" silencers that do not overlap our CRMs might represent false positives. If we exclude these false negative "validated" silencers (35.7%) and only consider the 5,525 (64.3%) of the validated silencers that overlap our predicted 1.2M CRMs, then we recall 84.4% (4,661) of them.

Thus, our method has achieved 84.4% sensitivity for recalling validated silencers overlapping our 1.2M CRMs, substantially higher than by chance (0.3%=57.5% x 0.5%).

3.2.10 Comparison of our predicted silencers with predicted silencers from two existing methods

Next, we compared our 677,840 silencers with the predicted silencers from silencerDB, primarily by two previous studies(105, 107). Specifically, Huang et al.(105) predicted a set of silencers by correlating putative active silencer epigenetic marks (H3K27me3) signals on DHSs with mRNA levels of neighboring genes across 25 different cell/tissue types, and then predicted additional silencers using an SVM model trained on the set using a combination of sequence features, epigenetic marks and gene expression profiles (CoSVM). Hawkins et al.(107) predicted silencers using a gkmSVM model by employing a simple subtractive strategy to obtain uncharacterized regulatory elements as potential silencers. After removing the redundancy in different cell/tissue types, we ended up with 157,813 and 982,985 non-redundant putative silencers predicted by CoSVM and gkmSVM, respectively.

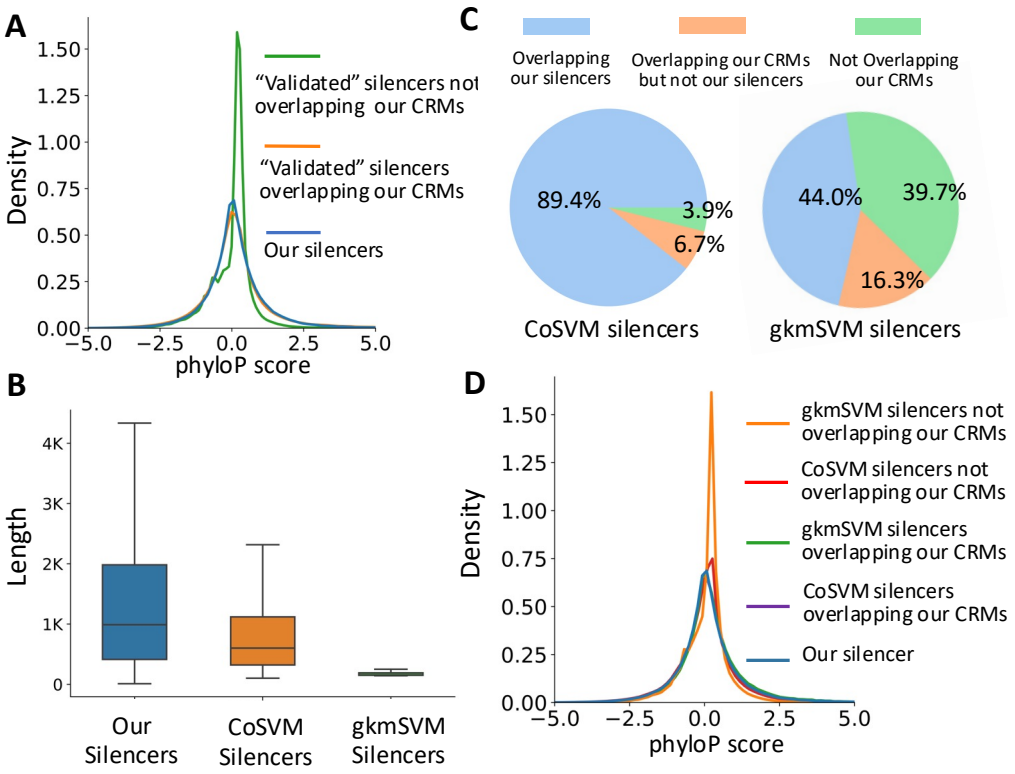


Figure 3-7. Comparison of our predicted silencers with the "validated" silencers and predicted silencers by CoSVM and gkmSVM. **A.** Distributions of phyloP scores of "validated" silencers that overlap or do not overlap our CRMs, as well as of our silencers. **B.** Boxplots of lengths of our predicted silencers, CoSVM-predicted silencers and gkmSVM-predicted silencers. **C.** Pie charts of CoSVM-predicted silencers (left) and gkmSVM-predicted silencers (right), which overlap our silencers, do not overlap our silencer but overlap our CRMs and do not overlap our CRMs. **D.** Distributions of phyloP scores of gkmSVM-predicted silencers and CoSVM-predicted silencers, which overlap and do not overlap our CRMs, as well as of our silencers.

As summarized in Table 3-1, gkmSVM predicted the highest number (982,985) of silencers, followed by our method (677,840) and CoSVM (157,813). However, our predicted silencers with a median length of 989 bp are longer than those predicted by CoSVM (601 bp) and gkmSVM (150 bp) (Figure 3-7B) and covers a greater proportion (33.3%) of the genome than those by CoSVM (4.8%) and gkmSVM (5.7%) (Table 3-1). Only 38,525 (24.4%) of the CoSVM-predicted silencers could be mapped to 52,539 (5.3%) gkmSVM-predicted silencers, indicating that the two methods predicted quite different sets of sequences as silencers. On the other hand, of the CoSVM-predicted silencers, 151,706 (96.1%) overlap 142,136 (12.1%) of our predicted 1.2M CRMs, of which

141,062 (89.4%) overlap 127,380 (10.8%) of our predicted silencers, while 10,644 (6.7%) overlap our CRMs but do not overlap our silencers, and 6,107 (3.9%) do not overlap our predicted CRMs (Figure 3-7C). As expected, the 151,706 CoSVM-predicted silencers that overlap our CRMs have similar evolutionary constraints as our predicted silencers (Figure 3-7D), suggesting that they might be true silencers. In this regard, our predicted silencer recalled most (93.0%) of CoSVM-predicted silencers, substantially higher by chance (7%=57.5% x 12.1%). Interestingly, the 6,107 CoSVM-predicted silencer that do not overlap our CRMs also have similar evolutionary constraints as our predicted silencers (Figure 3-7D), suggesting that they might be true silencers, but our CRM predictor dePCRM2 failed to predict them to be CRMs. This could be explained by the fact that dePCRM2 is not able to touch about 15% of the genome because of unavailability of TF binding data in these regions, therefore missing those CRMs that can function as silencers.

Moreover, of the gkmSVM-predicted silencers, 592,502 (60.3%) overlap 342,813 (29.1%) of our 1.2 M predicted CRMs, 432,572 (44.0%) overlap 232,960 (19.8%) of our predicted silencers, 159,930 (16.3%) overlap our CRMs but do not overlap our silencers, and 390,483 (39.7%) do not overlap our predicted CRMs (Figure 3-7C). As expected, the 592,502 gkmSVM-predicted silencers that overlap our CRMs have similar evolutionary constraints as our predicted silencers (Figure 3-7D), suggesting that they might be true silencers. In this regard, our method recalled most (73.0%) of them, substantially higher by chance (16.7%=57.5% x 29.1%). In contrast, the gkmSVM-predicted silencers that do not overlap our CRMs are largely selective neutral or nearly so, suggesting that they are more likely false positives (Figure 3-7D). In summary, CoSVM-predicted silencers appear highly accurate, and our predicted silencers recall 93.0% of them. Of the gkmSVM-predicted silencers, 44.0% appear to be authentic, and our predicted silencers recall

73.0% of them, while 39.7% appear to be false positives. Our method is comparable to CoSVM but superior to gkmSVM in accuracy. However, our method predicts more silencers than CoSVM.

Table 3-1. Summary of silencers predicted by the three methods

Method	# of unique silencers	Genome coverage	Median length
CoSVM(105)	157,813	4.75%	601 bp
gkmSVM(ENCODE)(107)	982,985	5.67%	150 bp
Our prediction	677,840	33.3%	989 bp

3.3 Discussion

In this study, we introduce two LR models to separately predict the functional states of our previously predicted 1.2M CRMs(19) as enhancers and silencers. The enhancer predictor uses signals of three epigenetic marks (CA, H3K4me1 and H3K27ac) on the CRMs as features. We choose these marks since they have been shown to be associated with active enhancers(114). The silencer predictor also employs signals of three epigenetic marks (CA, H3K9me3 and H3K27me3) on the CRMs as features. We choose these three marks based on the following reasons: CA is a hallmark of TF binding on CRMs including silencers(115, 116), both H3K9me3(117, 118) and H3K27me3(119) have been shown to be associated with repressive DNA sequences, and CA as well as H3K27me3 have been used as features for predicting silencers(105). As many cell/tissue types only have one set of these epigenetic marks, we build two independent predictors for their wider applicability as demonstrated in this study. Our enhancer predictor achieves comparable AUROC (0.977 vs 0.986) as our previous predictor that used four marks, which substantially outperforms five state-of-the-art methods(120). Our silencer predictor also achieves a high AUROC of 0.962, albeit slightly smaller than that (0.977) of the enhancer predictor. Although none of the three epigenetic marks alone or their combinations are specific for either active enhancers(29, 30, 33) or silencers(105), each of the three marks alone and their combinations achieve from moderate (0.685) to high (>0.95) AUROC values. We attribute the high accuracy of

the models to our two-step approach(120), i.e., we predict the functional types and states (second step) of our CRMs that were predicted (first step) using TF binding data. This conclusion is in good agreement with earlier reports that once the loci of CRMs are accurately anchored by the binding of key TFs, epigenetic marks can be accurate predictors of the functional states of the CRMs(29, 30, 33). Applying the enhancer model to the 105 cell/tissue types with data of the three active enhancer marks available, and the silencer model to the 58 cell/tissue types with data of the three putative active silencer marks available, we predict 868,944 (73.8%) of our 1.2 M CRMs in the human genome(19) to be active as enhancers or silencers in at least one of the 107 cells/tissue types.

Particularly, in the 56 of the 107 cell/tissue types, with both active enhancer and silencer marks data available, we predict 793,140 (67.3%) CRMs to be active enhancers, active silencers, or both in at least one of the cell/tissue types. We classify the 793,140 CRMs in four types: enhancer-predominant, silencer-predominant, dual functional, and unclassified. Moreover, we further classified dual functional CRMs into type I and type II based on whether or not they can be both active enhancers and active silencers in the same cell/tissue type. Moreover, since CA has much higher weights than the two other marks in both the enhancer and silencer predictors, they may predict some CRMs with very high CA signals but low signals of the two other marks both as active enhancers and silencers in the same cell/tissue type, thereby overestimating dual functional CRMs. To reduce possible such false positive predictions, we only consider the CRMs that are predicted to be active by both predictors as type I dual functional CRMs only if they are also labeled by at least one of the two other enhancer marks and one of the two other silencer marks, and classify the remaining CRMs that do not meet this criterion as unclassified. This gives a lower bounder 10.6% of the 793,140 CRMs to be dual functional. Consistent with earlier

reports(9, 102, 104), type II dual functional CRMs may switch their roles in different cell/tissue types, presumably by binding two different sets of TFs in different cellular contexts. To the best of our knowledge, we for the first time find that type I dual functional CRMs can function both as enhancers and silencers in the same cell/tissue type.

We show that the four types of CRMs (enhancer-predominant, silencer-predominant, type I dual functional and type II dual functional) possess distinct properties in terms of their lengths and TFBS densities, which reflect their functional complexity. Specifically, the longer lengths and the higher TFBS densities of type I dual functional CRMs might also be related to their dual functions in the same cell/tissue type, necessitating longer length and denser TF bindings. The shorter lengths and lower TFBS densities of type II dual functional CRMs than those of type I dual functional CRMs might be due to the fact that TFBSs in the former type can be shared for different functions across different cell/tissue types, since they only serve a single function in each cell/tissue type, while this might not be the case for the latter type. Such sharing might result in reduced TFBS densities and shorter lengths of type II dual functional CRMs. The higher TFBS density of type II dual functional CRMs than those of enhancer-predominant and silencer-predominant CRMs suggest that the former type might need denser TFBSs than the latter two types to execute both enhancer and silencer functions in different cell/tissue types by interacting with different sets of TFs. Although enhancer-predominant CRMs tend to be shorter than silencer-predominant CRMs (Figure 3-5D), the higher TFBS density of the former type suggests that activating genes might be a more intricate process than repressing genes. Unclassified CRMs with intermediate lengths and TFBS densities might be a mixture of the four types of CRMs. In the future, we need to determine the types of unclassified CRMs. One possible approach is to consider the positive or negative correlation between the predicted activation probabilities of a CRM and

the expression levels of its potential genes in a TAD across multiple cell/tissue type, as well as the physical proximity between the CRM and the transcription start sites of the potential genes as we demonstrated for the CRM shown in Figure 3-6B. Alternatively, we may use more specific epigenetic marks for active enhancers and silencers as features in machine-learning models when such marks are available in the future.

The substantial or complete overlaps between enhancer and silencer epigenetic marks peaks along type I dual functional CRMs strongly suggest that they are not artifacts by erroneously concatenating enhancers/silencers with their adjacent silencers/enhancers. A type I dual functional CRM might accomplish its dual roles in the same cell by simultaneously binding two sets of TFs via their cognate binding sites that are interdigitated along the CRM. Alternatively, it might accomplish its dual roles in different individual cells in a cell population of the same type by separately binding two sets of TFs in different individual cells. However, before relevant single-cell data are available, we could not differentiate these two possibilities. In either scenario, the high overlaps between the epigenetic mark peaks of enhancers and silencers along type I dual functional CRMs suggest that they might fulfill their dual roles by collaborative bindings of two different sets of TFs to their cognate binding sites along the CRMs as recently suggested for enhancers(121). Although type I dual functional CRMs can function both as enhancers and silencers in the same cell types, they more often only function as enhancers or as silencers in different cell/tissue types, presumably by binding one of the two sets of TFs, indicating the highly dynamic and cellular context dependent nature of their usage.

Our predicted silencers recall 84.4% of MPRA-validated silencers falling in our predicted CRMs, while missing the remaining 15.6%, indicating that we might need data from more cell/tissue types to predict them. On the other hand, we find that the “validated” silencers that do

not overlap our predicted CRMs might be false positives as they are largely selective neutral. This result is consistent with our recent finding(122) and other reports(123-127) that a considerable proportion of "regulatory sequences" identified by expression vector-based methods such as MPRA and its variants might be false positives. The two sets of previously predicted silencers cover a similar proportion (4.75% vs 5.67%) of the genome, but differ largely in their numbers (157,813 vs. 982,984) and have few overlaps. We find that CoSVM-predicted silencers are rather accurate, while gkmSVM-predicted silencers might have a false positive rate at least 39.7%. Our predicted silencers recall 93.0% CoSVM-predicted and 73.0% of gkmSVM-predicted silencers falling in our CRMs. Clearly, to recall the missed 7.0% of CoSVM-predicted and 27% of gkmSVM-predicted silencers falling in our CRMs, we might need more data from more diverse cell/tissue types. On the other hand, both CoSVM and gkmSVM miss 89.2% and 79.2% of our predicted silencers. Therefore, our method predicts much more silencers than CoSVM and gkmSVM while achieving accuracy comparable to that of CoSVM and superior to that of gkmSVM.

With the ongoing expansion of epigenetic and TF binding data available in a wide spectrum of cell/tissue types, our two-step approach holds the potential to uncover a more comprehensive map of CRMs in the genome, and then predict their functional types and states within these cell/tissue types. Based on the accurately predicted functional states of the CRMs and the expression levels of genes across a large number of cell/tissue types, as well as physical proximity between the CRMs and genes in TADs, it is possible to predict the target genes of the CRMs. This forward-looking perspective underscores the adaptive nature of our approach and its ability to yield deeper insights into the regulatory genomes and transcriptional machineries as datasets continue to grow and diversify in the future.

3.4 Conclusion

In this study, we extend our two-step approach(20) so that it can simultaneously predict the functional states and types of our previously predicted 1.2M CRMs(19) using data of only five epigenetic marks in a cell/tissue type. Applying the method to cell/tissue types with the data available, we classified the CRMs into four types (enhancer-predominant, silencer-predominant, type I dual functional and type II dual functional CRMs) with distinct properties reflecting their functional complexity. Dual functional CRMs and silencers might be more prevalent than previously assumed. The accurate prediction of functional types and states of CRMs opens avenues for identifying their target genes.

3.5 Materials and Methods

3.5.1 The datasets

We obtained a set of 1,178,229 predicted CRMs in the human genome from our prior work(19). We downloaded histone mark ChIP-seq, DNase-seq, ATAC-seq and TF ChIP-seq data in 67 human cell/tissue types from Cistrome Datasets Browser(110, 111) (Supplementary Table S3-1, S3-2, S3-3). All these 67 cell/tissue types have data for the three active enhancer marks (CA, H3K4me1 and H3K27ac), and 40 of them also have data for the three active silencer marks (CA, H3K27me3 and H3K9me3). We downloaded ATAC-seq and histone mark ChIP-seq data and RNA-seq data in 107 cell/tissue types from ENCODE data portal(74), of which 105 cell types have data for active enhancer marks (CA, H3K4me1 and H3K27ac), 58 have data for active silencer marks (CA, H3K27me3 and H3K9me3), 56 have data for both the active enhancer and active silencer marks, 49 only have data for active enhancers, and 2 only have data for active silencer marks (Figure 3-3A, Supplementary Table S3-10, S3-11). We downloaded the Hi-C contact matrix of the K562 cell line from the ENCODE(74) portal with the session ID ENCFF080DPJ.

3.5.2 Epigenetic mark feature scores

For each epigenetic mark e and a sequence c, which can be a CRM or non-CRM, we define a raw feature score as:

$$F_{raw}(c,e) = \sum_{i=1}^{N_e} r_{i,e} s_{i,e} \quad (3-1)$$

where N_e is the number of peaks of e mapping to c at least 50% of the length of either one, $r_{i,e}$ the ratio of overlapping length between c and the i_{th} peak of e over the length of the i_{th} peak of e, $s_{i,e}$ the signal of the i_{th} peak of e quantified by MACS2(128, 129). We then normalized each raw epigenetic feature score in each cell/tissue type by the min-max normalization, i.e.,

$$F(c,e) = \frac{F_{raw}(c,e) - min(F_{raw}(C,e))}{max\big(F_{raw}(C,e)\big) - min(F_{raw}(C,e))} \quad (3-2)$$

where C denotes all candidate sequences in the genome, $min\ (F_{raw}(C,e))$ and $max\ (F_{raw}(C,e))$ the minimum and maximum raw score of the epigenetic mark e over C in the cell/tissue type.

3.5.3 Prediction of functional states of CMRs

Since a CRM can function both as an enhancer and a silencer in different cellular contexts, we use two separate models to predict the activation probability of a candidate CRM to be an enhancer and a silencer in a cell/tissue type. Specifically, we used an LR model to predict the activation probability of a candidate CRM functioning as an enhancer using signals of three active enhancer markers, i.e., CA, H3K4me1 and K3K27ac. Meanwhile, we used a similar LR model to predict the activation probability of a candidate CRM functioning as a silencer using signals of three active silencer markers, i.e., CA, H3K9me3 and K3K27me3.

3.5.3.1 Construction of positive and negative sets: In each of 67 cell/tissue types with the required data available, we selected the CRMs that overlap TF binding peaks and at least one of active enhancer marks H3K4m1 and K3K27ac, or of silencer marks K3K27me3 and K3K9me3 in

the cell/tissue type as the positive enhancer or silencer set, to ensure the high quality of the positive set in a cell/tissue type. At the same time, we randomly selected predicted non-CRMs with matched numbers of the positive sets as the negative sets. We pooled the positive and negative sets in all the relevant cell/tissue types separately to construct a comprehensive positive set and a negative set for enhancers and silencers, resulting in 1,415,796 positive enhancers and 256,766 positive silencers and the same numbers of respective negative sets. Thus, the positive sets and negative sets for both enhancers and silencers are well-balanced.

3.5.3.2 Model training and evaluation: Ten-fold cross-validation was conducted to train and assess the performance of seven models using all the seven possible combinations of three marks as the features. The models were implemented using sci-kit learn v.0.24.2 and the code is available at https://github.com/zhengchangsulab/EnhancerSilencerPrediction.

3.5.3.3 Prediction: We applied both trained enhancer model and silencer model to the 1,178,229 CRMs in each of the 107 cell/tissue types with the required data available. We predict a CRM to be an active enhancer or silencer if its activation probability as an enhancer or a silencer is greater than 0.5.

3.5.4 Heat maps of epigenetic marks

We used the "EnrichedHeatmap" package(130) version 4.2.2 to generate heat maps of signal intensities of epigenetic marks in a 6 kb region centered on a CRM. We computed the mean signal value for a mark in each100 bp sliding window in each of the 6 kb sequences, using "w0" as the "mean_mode". The line plot on the top of the heat map is the mean signals of each window at a position across all the sequences of a set of CRMs. The CRMs within each set are ranked based on their CA signal strengths in descending order.

3.5.5 Overlapping ratio of the enhancer and silencer epigenetic marks along dual functional CRMs

We merged the peak regions of the two enhancer marks H3K4me1 and H3K27ac on a CRM to form a unified region E_{mark}, and those of the two silencer marks H3K9me3 and H3K27me3 on the CRM to form another unified region S_{mark}. We computed the overlapping ratio between enhancer and silencer marks on the CRM as:

$$overlapping\ ratio = \frac{overlap(E_{mark},\ S_{mark})}{min(len(E_{mark}), len(S_{mark}))} \qquad (3-3)$$

where $overlap(E\_mark,\ S\_mark)$ denotes the length of the overlapping part between E_{mark} and S_{mark}, and $len(E\_mark)$ and $len(S\_mark)$ the lengths of S_{mark} and S_{mark}, respectively.

3.5.6 Heat map of Hi-C contact matrix

We used the Juicebox(131) version 2.17.00 to generate the heat map of the Hi-C contact matrix using the Hi-C data from K562 cells with default settings at a resolution of 1 kb on the region from 1,350,000 to 1,600,000 bp on chromosome 7.

3.6 Availability of data and materials

The datasets and code supporting the conclusions of this chapter are available at https://github.com/zhengchangsulab/EnhancerSilencerPrediction and are included within the chapter and its supplementary tables at https://github.com/sisyyuan/CRM_Dissertation.

Chapter 4
PREDICTION OF TARGET GENES OF CRMs IN THE HUMAN GENOME REVEALS THEIR DISTINCT PROPERTIES

4.1 Introduction

High throughput methods like Hi-C(34) and chromatin interaction analysis with ChIA-PET(35), which examine physical proximity between two linearly distant loci, have been used to map regulatory relationships between enhancers and target genes(132-134) in various cell/tissue types. Hi-C technologies have led to the identification of distinct genome compartments at the mega-base scale. Compartment A resides in open, gene-rich euchromatin, while compartment B is composed of closed, gene-poor heterochromatin(78). Interactions are more frequent within the same compartment than across different compartments(135). Within the compartments, TADs are formed at the sub-mega-base scale, where interactions are highly enriched within TADs relative to between TADs(56). At a higher resolution, chromatin loops can establish spatial proximity between specific distant genomic loci through a process called loop extrusion(136-138). However, it is difficult to identify CRM-gene regulations precisely due to the often-low genomic resolution of such data and that genomic contacts may not guarantee regulation relationships. More recently, CRISPR technology has been used to identify target genes of putative CRMs in various ways(139). For example, CRISPRa was used to probe putative enhancers for their potential to regulate nearby genes(140). Additionally, CRISPRi was employed to identify regulatory relationships by targeting putative enhancers and assessing the effects on the neighboring genes(141-144). Nonetheless, these methods are limited to a few CRMs and genes. Consequently, experimental determination of target genes of CRMs on a genome-wide scale remains an ongoing challenge.

In the past few years, several computational methods have been developed for predicting target genes of putative enhancers and silencers. However, these methods are limited, since in the

absence of a precise and comprehensive CRM map in the genome, they aim to predict target genes for specific genomic regions marked by distinct epigenetic modifications. These methods attempt to predict both CRMs and their target genes, along with their functional types if applicable, all in a single step. The predominant one-step approaches for predicting target genes of enhancers include score-based(36, 37), correlation-based(38-40) and machine learning methods(40-44).

As the most intuitive and widely used score-based method, the distance-based method uses the genomic distance or some function of the distance, usually defined as the number of bp between the potential regulatory region and TSS of the candidate gene(36). The simplest distance-based method assigns the gene whose TSS is closest to either end of a CRM as the CRM's target gene. While this closest neighbor assignment (CNA) method suits some predictions well, it can overlook distant regulations. This is crucial because a CRM can target genes from hundreds of thousands to a million bp away(145), and a CRM can regulate multiple target genes(146). Other score-based methods incorporate additional geometric or functional data, consolidating multiple metrics into a composite score that quantifies a putative regulatory region's potential in regulating a candidate gene. For example, GeneHancer(37) integrates five features associated with enhancers and/or target genes — RNA and eRNA levels, and eQTL, cHi-C and distance data — to predict target genes of candidate enhancers compiled from four databases.

The correlation-based methods evaluate correlation between the activities of a potential regulatory region and the activities of possible target genes across a panel of cell/tissue types. DHSs, indicative of DNA accessibility, have been employed as potential regulatory regions, and correlations between DHS signals in these regions and in promoters across a panel of 79 tissues have been used to predict enhancer-gene links(147). However, it turns out that the correction between DHS signals alone is inadequate to indicate enhancer-gene regulations(38). Thus, a

variant method was proposed that quantified correlations between DHS signals at potential regulatory regions with the expression levels of possible target genes(38). Moreover, correlation between DNA methylation levels at potential regulatory regions and the expression levels of possible target genes has also been utilized to pinpoint enhancer-gene regulations(39, 148).

Supervised learning models trained on "golden standard" enhancer-gene links, along with an equivalent number of negative links, have been used to predict unknown links using features such as gene expression levels(42), DHS signals(42, 43), histone marks(40, 42), correlations of these signals(40), sequence compositions(40, 44), and the distance(40, 43) between candidate enhancers and the TSSs of potential target genes. However, a significant drawback of these methods is the lack of sufficient experimentally validated "golden standard" positive and negative sets(16), leading different groups to define training sets differently. This divergence in defining training sets can introduce noise and potentially influence prediction accuracy. Moreover, in many of these methods(132, 133, 144), the candidate genes were typically screened within an arbitrarily selected flanking region around putative CRMs, although the true target genes can be located outside of the region.

To overcome the limitation of these existing one-step methods for predicting the loci of enhancers and silencers, we have recently proposed a two-step approach based on the following two observations: i) TF binding is more informative for locating CRMs than CA and histone marks; and ii) once the loci of a CRM's is accurately located, epigenetic marks on the enhancers are good predictors of its functional state(29-33). Specifically, we first predict a highly accurate and more complete map of CRMs in the genome, and then predict the functional states of all the CRMs in various cell/tissue types of the organism. For the first step, we have developed the dePCRM2 algorithm(19) that predicts the locations of all possible CRMs in the genome by integrating all

available TF ChIP-seq datasets in various cell/tissue types of the organism(19). Applying dePCRM2 to more than 11k TF ChIP-seq datasets in various human cell/tissue types, we predicted 1.2M CRMs in the human genome. For the second step, we have developed two LR models that can accurately predict functional states of the many CRMs as enhancers and silencers in any cell/tissue types using CA in combination with two active enhancer or silencer histone marks, thereby simultaneously predicting the functional types and states of the CRMs. This two-step approach substantially outperforms existing one-step methods for predicting the loci of CRMs in the genome as well as their functional types and states in cell/tissue types of the organism(20). Nonetheless, since CA has overwhelmingly higher weights than do the two enhancer or silencer histone marks in both the LR models, they are unable to unambiguously distinguish enhancers and silencers that have high CA signals but weak enhancer or silencer histone marks signals(20, 149). Moreover, unlike CA data that are widely available in a broader range of cell/tissue types, data of the two enhancer or silencer histone marks are only available in a small number of even well-studied human and mouse cell/tissue types, limiting the application and statistical power of the models.

Building upon the success of the two-step strategy and by overcoming the limitation of the LR models, we now introduce a method, CAPP for the third step of our strategy to predict the target genes of the predicted CRMs in a genome-wide fashion. CAPP predicts enhancer-gene and silencer-gene links based on the correlation between functional states of the CRMs and the expression levels of potential target genes within the same TAD across a panel of cell/tissue types as well as physical proximity between the CRMs and potential target genes. In this way, CAPP is able to not only predict the CRMs' target genes, but also to more accurately predict their functional

types. We show that CAPP out-performs state-of-the-art methods in accuracy, while predicting substantially more CRM-gene links.

4.2 Results

4.2.1 Most of the genome and our predicted CRMs are covered by consensus TADs in various cell types

Since CRM-gene regulations are believed to occur mainly within a TAD(18, 150-152), to reduce the search space we only consider CRMs and genes in the same TAD for potential regulation relationships. Utilizing Hi-C interaction data in six cell/tissue types, we identified their TADs at five resolutions (5,000 bp~100,000 bp). As expected, TADs identified at different resolutions cover varying proportions of the human genome, genes and our predicted CRMs in each cell/tissue type (Figure 4-1A shows the case in the K562 cells). As TADs predicted at a higher resolution tend to be shorter and are often nested inside of larger ones predicted at a lower resolution, we thus merged overlapping TADs predicted at different resolutions to form certain number of merged TADs (e.g., there are 944 merged TADs in K562 cells), which cover higher proportions of the genome, genes and CRMs than TADs predicted at a single resolution (Figure 4-1A for the K562 cells). Particularly, the vast majority (1,178,225, or 96.2%) of our 1,225K predicted CRMs in the genome are located within the merged TADs (Figure 4-1A). Similar results were obtained in other five cell/tissue types with Hi-C data available (Figure 4-2). As shown in Figure 4-1B, 81.12% of the genome can be covered by the merged TADs in all the six cell/tissue types, and 6.48% of the genome remains uncovered by any of the merged TADs in the six cell/tissue types, while 91.44% of the genome are covered by the merged TADs in at least two cell/tissue types, suggesting that TADs from different cell/tissue types are largely invariable, consistent with previous reports(153-155). Therefore, in cases where Hi-C data is not yet available in a cell/tissue, we use TADs from

these six cell/tissue types as a substitute. Particularly, if not otherwise noted, the analysis was conducted using TADs derived from the K562 cells.

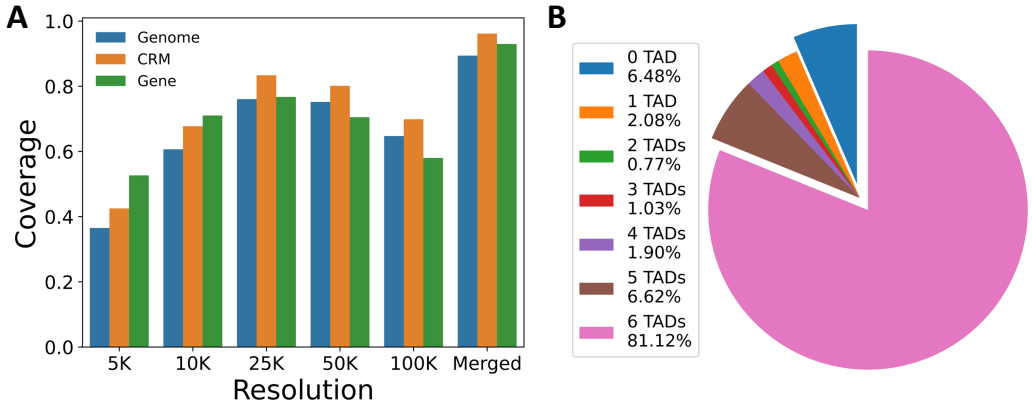


Figure 4-1. Coverage analyses on predicted TADs. **A.** Coverage of the genome, CRMs and genes by TADs identified at various resolutions and by merged TADs in the K562 cells. **B.** Proportion of the genome covered by the merged TADs from different numbers of cell lines. The legends such as "6 TADs 81.12%" means 81.12% of the genome is covered by the merged TADs from the six cell lines.

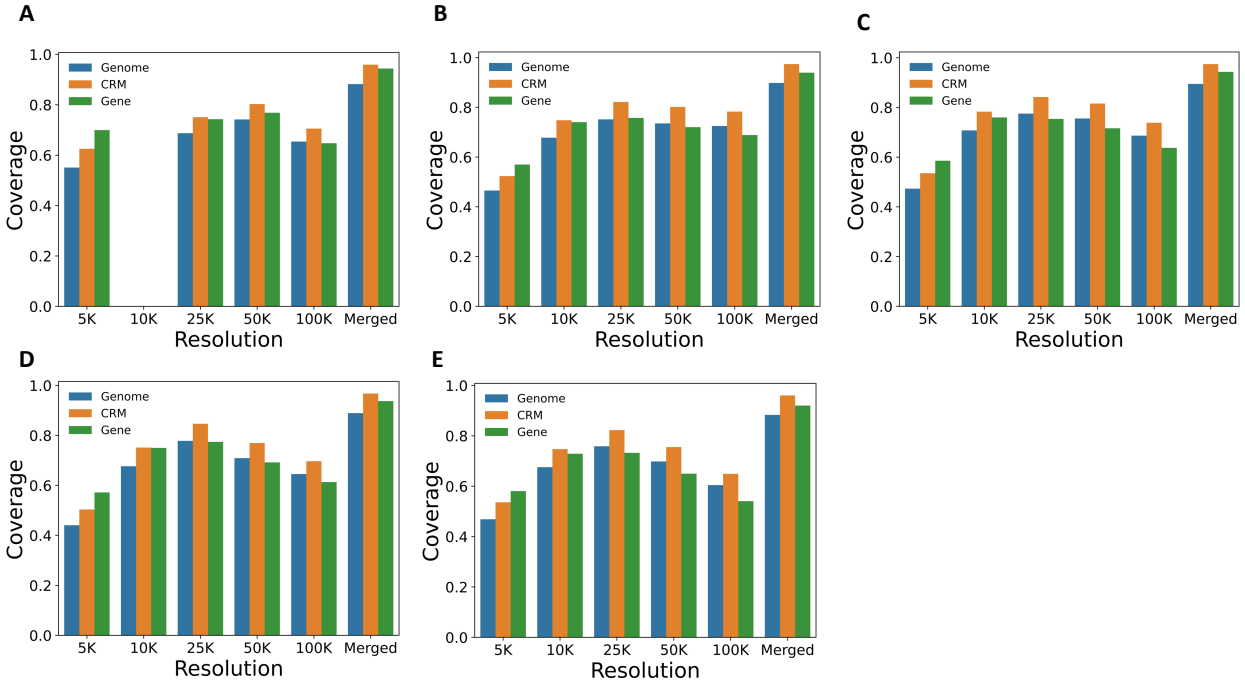


Figure 4-2. Coverage of the genome, CRMs and genes by TADs identified at various resolutions and by merged TADs in different cell lines: **A.** GM12878 (B Lymphocyte in Blood). Hi-C contact matrix cannot be generated at resolution 10K bp using Arrowhead, resulting in missing bars. **B.** H1 (Embryonic Stem Cell in Embryo). **C.** HeLa-S3 (Epithelium in Cervix). **D.** HepG2 (Epithelium in Liver). **E.** IMR90 (Fibroblast in Lung).

4.2.2 CA alone can accurately predict the functional states of CRMs

As the first step of the CAPP method, it predicts functional states of CRMs in as many as possible cell/tissue types of the organism. To this end, unlike our previous method that used two LR models with CA and two histone marks as the features(20, 149), CAPP employs a single LR model with CA as the sole feature to predict functional states of CRM without first differentiating their types (enhancers and silencers). When trained and evaluated on the 67 human cell/tissue types (Materials and Methods) with ten-fold cross-validation, the LR model achieved a median AUROC of 0.93 (Figure 4-3), which was only slightly lower than those (AUROC=0.98) of LR models using two additional histone marks(20, 149). Hence, the LR model using CA as the sole feature is able to accurately predict the functional states of CRMs in a given cell/tissue type, although it cannot differentiate their functional types (enhancers and silencers). Applying the model to the 107 human cell/tissue types, we predicted highly varying numbers of active CRMs ranging from 18,995 (1.6%) in SJSA1 cells to 166,236 (14.1%) in motor-neuron cells, with a median of 64,476 (5.5%) in the cell/tissue types (Supplementary Table S4-1). We predicted a total of 7,363,163 active CRMs in the 107 cell/tissue types. After removing the redundancy, we ended up with a total of 547,695 (46.5%) non-redundant active CRMs in the 107 cell types.

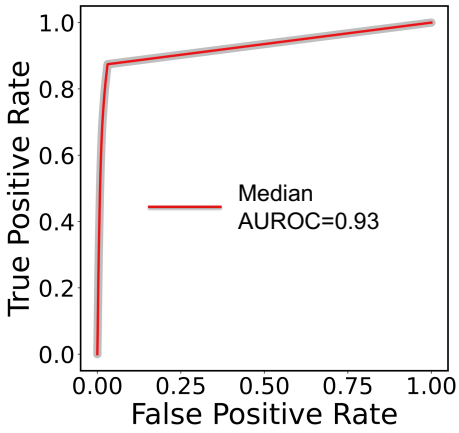


Figure 4-3. ROC curves of the LR model with CA as the sole feature. The red curve is the median ROC curve from the results of 10-fold cross-validation using positive and negative data in 67 human cell/tissue types.

4.2.3 Target genes of one fifth of CRMs can be predicted using currently available datasets

After the functional states of the CRMs are predicted, CAPP predicts the target genes of the CRMs in TADs by examining the correlation across a panel of cell/tissue types between their functional states and expression levels of genes located within the same TADs, followed by validation of physical proximity between the CRMs and the genes using Hi-C interaction data available in the six cell/tissue types (Materials and Methods). To ensure high statistical power of the predictions, we only considered the CRMs that were predicted to be active and inactive in at least five different cell/tissue types, resulting in 260,220 (47.8%) CRMs out of the 547,695 CRMs that were predicted to be active in at least one of the 107 cell/tissue types. At an FDR of 0.1, CAPP predicted 240,024 (92.2%) CRMs that enhanced the expression of 47,765 genes via 4,399,244 enhancer-gene regulations. Similarly, at the same FDR, CAPP predicted 11,592 (4.5%) CRMs repressed 10,400 genes via 31,477 silencer-gene regulations. In total, CAPP predicted 47,775 (82.0% out of 58,261) target genes for 240,680 (92.5% out of 260,220) CRMs. Thus, CAPP was able to predict target genes for one fifth (20.4%) of the 1,178K CRMs within TADs using data available in only 107 cell/tissue types. Of the 240,680 CRMs with predicted target genes, 10,936 (4.5%) functioned as

both enhancers and silencers (dual CRMs) for different genes, while 229,088 (95.2%) exclusively acted as enhancers (exclusive enhancers), and the remaining 656 (0.3%) exclusively acted as silencers (exclusive silencers). The fewer predicted silencers than predicted enhancers can be attributed to the less prevalence of silencers than enhancers in transcriptional regulation. It is evident that obtaining more data from more and diverse cell/tissue types is crucial to predict target genes for more CRMs.

4.2.4 Dual functional CRMs tend to regulate the largest number of genes, followed by exclusive enhancers and exclusive silencers

We first analyzed the number of genes regulated by the three different types of CRMs, i.e., dual CRMs, exclusive enhancers and exclusive silencers. As shown in Figure 4-4A, dual CRMs tend to regulate more genes than exclusive CRMs. Furthermore, exclusive enhancers tend to regulate more genes than exclusive silencers (Figure 4-4A). Specifically, 33.9% of dual CRMs regulate no more than 10 genes, contrasting with 52.4% for exclusive enhancers and 95.3% for exclusive silencers. These results indicate that exclusive silencers tend to have narrower effects by regulating few genes, and exclusive enhancers tend to have broader effects by regulating larger numbers of genes, while dual CRMs can function as both enhancers and silencers, and thus, regulate the largest number of genes.

4.2.5 Enhancers are more cooperative than silencers to regulate target genes

We next analyzed the numbers of enhancers or silencers that regulate a gene. As shown in Figure 4-4B, 95.6% of genes are regulated by no more than 10 silencers and only 4.4% of genes are regulated by more than ten silencers. In contrast, only 9.8% of genes are regulated by no more than 10 enhancers and 90.2% of genes are regulated by more than ten enhancers. These results suggest that enhancers are more likely to cooperate with one another to regulate a gene than silencers. In other words, multiple enhancers are required to up-regulate a gene, while only few silencers are

needed to down-regulate a gene. This result is in line with a previous report that an assembly of multiple enhancers, often located tens to hundreds of thousands of bp away from their target gene, tend to collaboratively regulate the common target gene by looping to the gene's promoter(156).

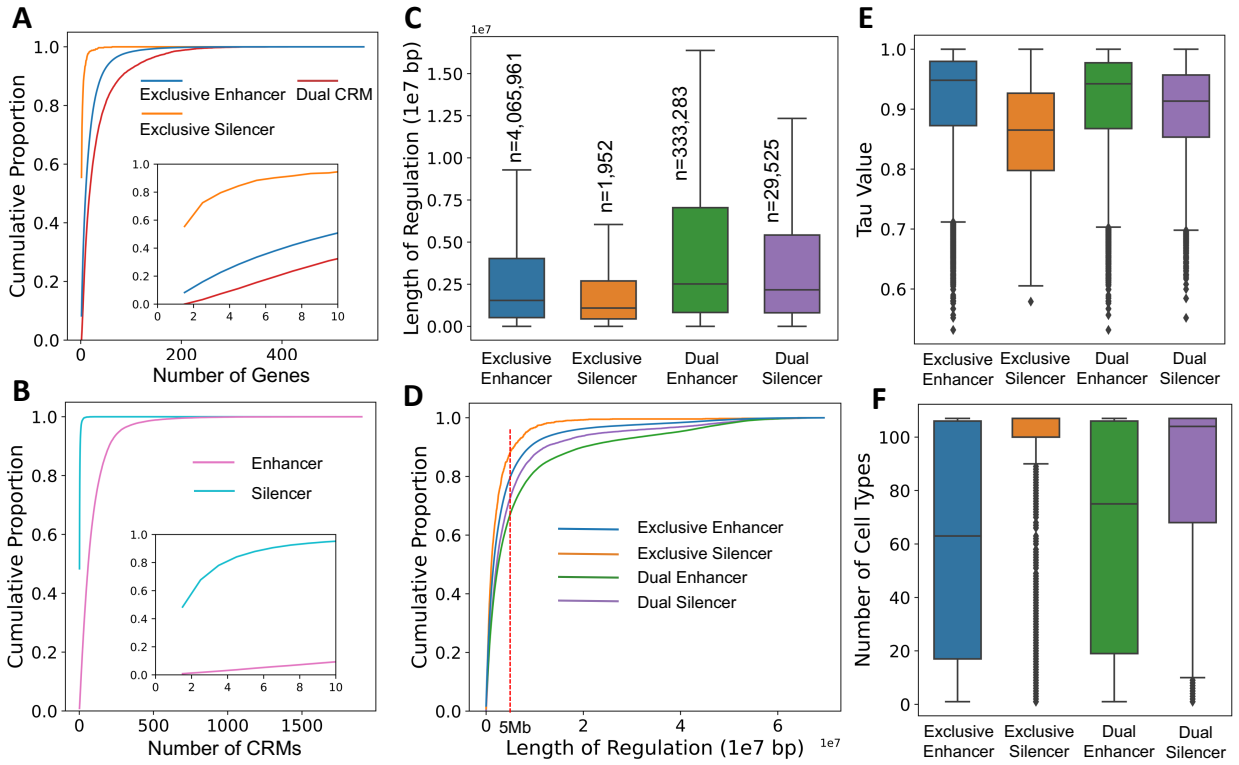


Figure 4-4. Comparisons of different types of CRMs for their predicted target genes and regulation lengths. **A.** Cumulative probability of CRMs regulating the indicated numbers of genes. The inset is a zooming-in view of the region with the number of genes not larger than 10. **B.** Cumulative probability of genes regulated by the indicated numbers of CRMs. The inset is a zooming-in view of the region with the number of CRMs not larger than 10. **C.** Boxplots of regulation lengths (in bp) of the predicted exclusive enhancer-gene, exclusive silencer-gene, dual enhancer-gene and dual silencer-gene regulations. **D.** Cumulative probability of regulation lengths (in bp) of the predicted exclusive enhancer-gene, exclusive silencer-gene, dual enhancer-gene and dual silencer-gene regulations. **E.** Boxplots of τ values of target genes of exclusive enhancers, exclusive silencers, dual enhancers and dual silencers. **F.** Boxplots of the number of cell/tissue types where target genes of exclusive enhancers, exclusive silencers, dual enhancers and dual silencers are expressed.

4.2.6 Dual functional CRMs tend to regulate more distant genes

We ask whether different types of CRMs have preference to regulate nearby or distant genes. To this end, we compared the distribution of regulation lengths of our predicted CRM-gene

regulations. Here the regulation length is defined as the distance in bp between the nearer end of a CRM and the TSS of its target gene. If a target gene's TSS falls within the body of a CRM, we consider the regulation length to be 0. For each dual CRM, we call it a dual enhancer when it functions as an enhancer, and a dual silencer when it functions as a silencer. As illustrated in Figure 4-4C, the regulation lengths of dual enhancers and silencers are significantly longer than those of exclusive CRM-gene links, while those of exclusive enhancers are significantly longer than those of exclusive silencers. Interestingly, dual enhancers tend to regulate more distant genes than dual silencers, mirroring the behaviors of exclusive enhancers and exclusive silencers. Remarkably, approximately 32.7%, 27.1%, 20.1%, and 11.6% of the dual enhancer-gene, dual silencer-gene, exclusive enhancer-gene, and exclusive silencer-gene links, respectively, exhibit a regulation length exceeding 5M bp (Figure 4-4D). This exceeds the fixed 1~5M bp flanking regions used in conventional approaches, which can potentially miss such regulations. Similarly, the number of intervening genes between dual functional CRMs and their target genes are greater than those between exclusive CRMs and their target genes (Figure 4-5A and 4-5B).

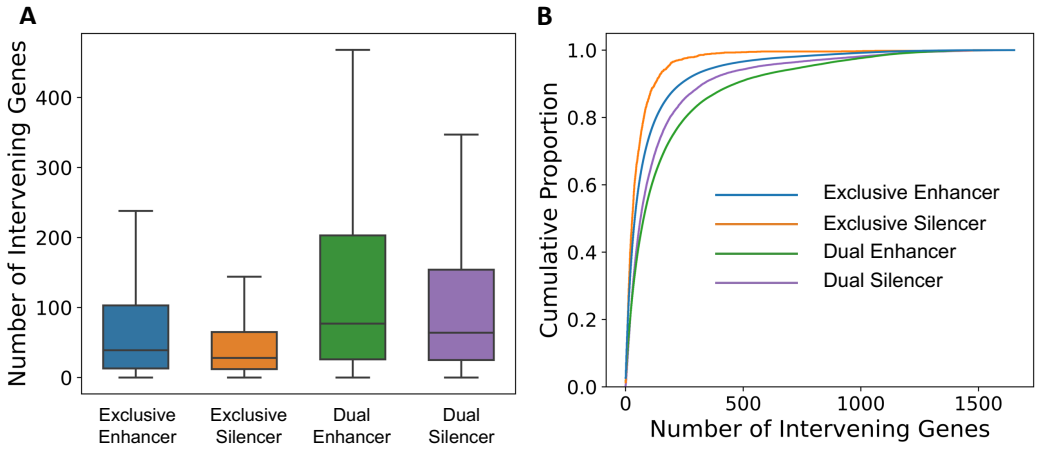


Figure 4-5. Numbers of Intervening genes between CRMs and their target genes**. A.** Boxplots of numbers of intervening genes between target genes and exclusive enhancers, exclusive silencers, dual enhancers and dual silencers. **B.** Cumulative probability of numbers of intervening genes between target genes and exclusive enhancers, exclusive silencers, dual enhancers and dual silencers.

4.2.7 Enhancers tend to regulate more narrowly expressed genes while silencers tend to regulate more broadly expressed genes

To explore whether enhancers and silencers exhibit preferences for regulating specific types of genes, we analyzed the expression patterns of their predicted target genes. Specifically, we first calculated the τ index value(157) for each gene within TADs that were expressed (transcript per million (TPM) > 0) in at least one of the 107 distinct cell/tissue types. The τ index measures gene expression specificity, with a value of 0 indicating ubiquitous expression and a value of 1 indicating expression in a single cell/tissue type(157). Genes regulated by enhancers tend to have greater τ values than those regulated by silencers (Figure 4-4E), implying that enhancers tend to regulate more narrowly expressed genes, while silencers tend to regulate more broadly expressed genes. Exclusive enhancers tend to have the highest τ values, indicating its preference in regulating narrowly expressed genes, while exclusive silencers tend to have the lowest τ values, indicating its preference in regulating broadly expressed genes (Figure 4-4E). Consistently, genes regulated by enhancers tend to exhibit expression in fewer cell/tissue types than those regulated by silencers (Figure 4-4F). Furthermore, exclusive enhancers have a propensity to regulate genes expressed in fewer cell/tissue types, while exclusive silencers tend to regulate genes expressed in a broader array of cell/tissue types (Figure 4-4F). These observations underscore the distinct strategies that enhancers and silencers take to regulate different gene types based on the prevalence of their usages.

4.2.8 Static and active *cis*-regulatory networks can be built by the predicted CRM-gene links

Based on our 47,775 predicted target genes of the 240,680 CRMs, we constructed static *cis*-regulatory networks (sCRNs) whose nodes are the CRMs and their target genes, and the edges are the CRM-gene links between them, regardless of the functional states of the CRMs in cell/tissue types. The sCRNs are made of 753 connected components distributed across the 23 pair of

chromosomes (Supplementary Table S4-2). Most of the connected components correspond to a TAD since the predicted regulations are primarily within the same TAD. However, some components may correspond to more than one TAD due to CRMs or genes crossing the border between two TADs, resulting in regulation across adjacent TADs, or one TAD might break into several components due to the hierarchical nature of TAD structures. The active *cis*-regulatory networks (aCRNs) in a cell/tissue type can be induced from the sCRNs by the active enhancer-gene and silencer-gene links in the cell/tissue type. As an example, Figures 4-6A and 4-6B show the sub-sCRNs on chromosome 10, composed of 33 connected components and the sub-aCRNs on chromosome 10 in the K562 cells embedded in the sub-sCRNs with the active exclusive enhancer-gene, exclusive silencer-gene and dual CRM-gene links highlighted in green, red and orange, respectively. More network examples can be found in the APPENDIX B.

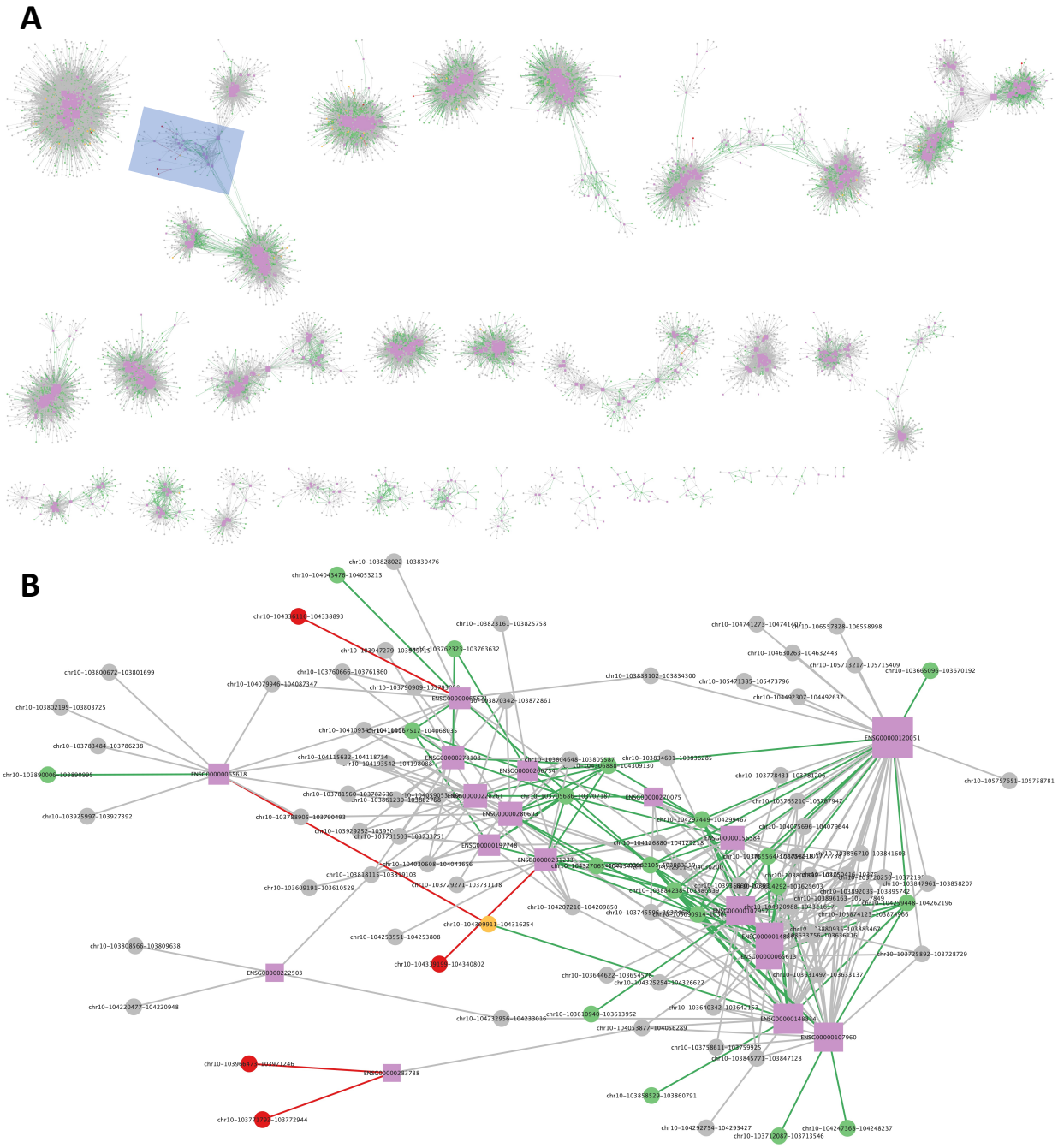


Figure 4-6. Examples of sub-static and sub-active *cis*-regulatory networks on chromosome 10. **A.** The sub-static *cis*-regulatory networks on chromosome 10 are made up of 33 connected components. The sub-active *cis*-regulatory networks on the chromosome in the K562 cells are embedded in the static *cis*-regulatory networks and can be induced by active exclusive enhancers, active exclusive silencers and active dual CRMs in the cells. The circles represent CRMs, including inactive CRMs (gray), active exclusive enhancers (green), active exclusive silencers (red) and active dual CRMs (yellow) in the K562 cells. The purple squares represent genes and their size are proportional to the degrees, i.e., the number of their regulating CRMs. The edges represent regulation relationships between CRMs (circles) and genes (squares). The edges linked to active

CRMs are colored by the same colors as the type of active CRMs. **B.** The zooming-in view of the connected component from Figure 4-6A in light blue shadow rotated counterclockwise by 30 degrees.

4.2.9 Our model outperforms the distance-based method CNA

We compared our method CAPP with a commonly used distance-based method, CNA. To do so, we considered the 260,220 CRMs that were predicted to be active and inactive in at least five of the 107 cell/tissue types. For each of these 260,220 CRMs, we assigned the gene(s) whose TSS(s) was(were) closest to the CRM as its target gene(s), resulting in 287,946 CRM-gene regulations (a CRM may have multiple closest genes). First, we evaluated the possibility of these CRMs being enhancers of the assigned target genes. To this end, we tested whether the expression levels of the putative target genes of each CRM were significantly higher in the cell/tissue types where the CRM was active than in the other cell/tissue types where the CRM was inactive. At an FDR of 0.1, 117,319 (40.7%) of the 287,946 CRM-gene regulations predicted by CNA exhibited significantly positive regulation, while the rest 59.3% did not show enhancer-gene regulations (Figure 4-7A). In contrast, all of our predicted 4,399,244 enhancer-gene regulations exhibited significantly positive regulations (Figure 4-7A). Consistently, only 94,988 (33.0%) of the 287,946 CRM-gene regulations predicted by CNA coincided with our 4,399,244 predicted enhancer-gene links (Figure 4-7B), implying that 33.7% of our enhancers are able to regulate their closest genes, while the remaining 66.3% only regulate distant genes. Figure 4-7C shows an example of enhancer chrY:20574642-20577538 regulating its closest gene *EIF1AY*. On average, our predicted enhancers regulated 18.3 genes while those predicted by CNA only targeted 1.1 genes. Hence, CNA might overlook distant regulations, missing out on the crucial one-to-many regulatory mechanism for enhancers. This is significant as a CRM can target genes located hundreds of thousands to a million bp away(145), and a single CRM might regulate multiple target genes(146).

Next, we evaluated the possibility of these CRMs being silencers of the assigned target genes. To this end, we tested whether the expression levels of the putative target genes of each CRM were significantly lower in the cell/tissue types where the CRM was active than in the other cell/tissue types where the CRM was inactive. Of the 287,946 CRM-gene links, at the same FDR of 0.1, only 43 (0.01%) regulations exhibited significantly negative regulations, while the remaining 99.99% of the CRM-gene regulation did not show significantly negative regulations (Figure 4-7D). In contrast, all of our predicted 31,477 silencer-gene regulations exhibited significantly negative regulations (Figure 4-7D). Consistently, only 123 (0.04%) of the 287,946 CRM-gene links predicted by CNA overlapped our 31,477 predicted silencer-gene links (Figure 4-7E), implying that only about 0.39% of silencers are able to regulate their closest genes, while the remaining 99.61% only regulate distant genes. This result suggests that silencers, compared with enhancers, tend not to regulate nearby genes. Figure 4-7F illustrates an example of a silencer chr5:67798877-67801081 regulating its closest gene *lnc-CD180-6*. On average, our predicted silencers regulated 2.7 genes while those predicted by CNA only targeted 1.1 genes. As in the case of enhancers, CNA might neglect distant regulations, leading to a potential oversight of a one-to-many regulatory mechanism for silencers. In summary, CAPP demonstrates superior performance over CNA in predicting the target genes for both enhancers and silencers.

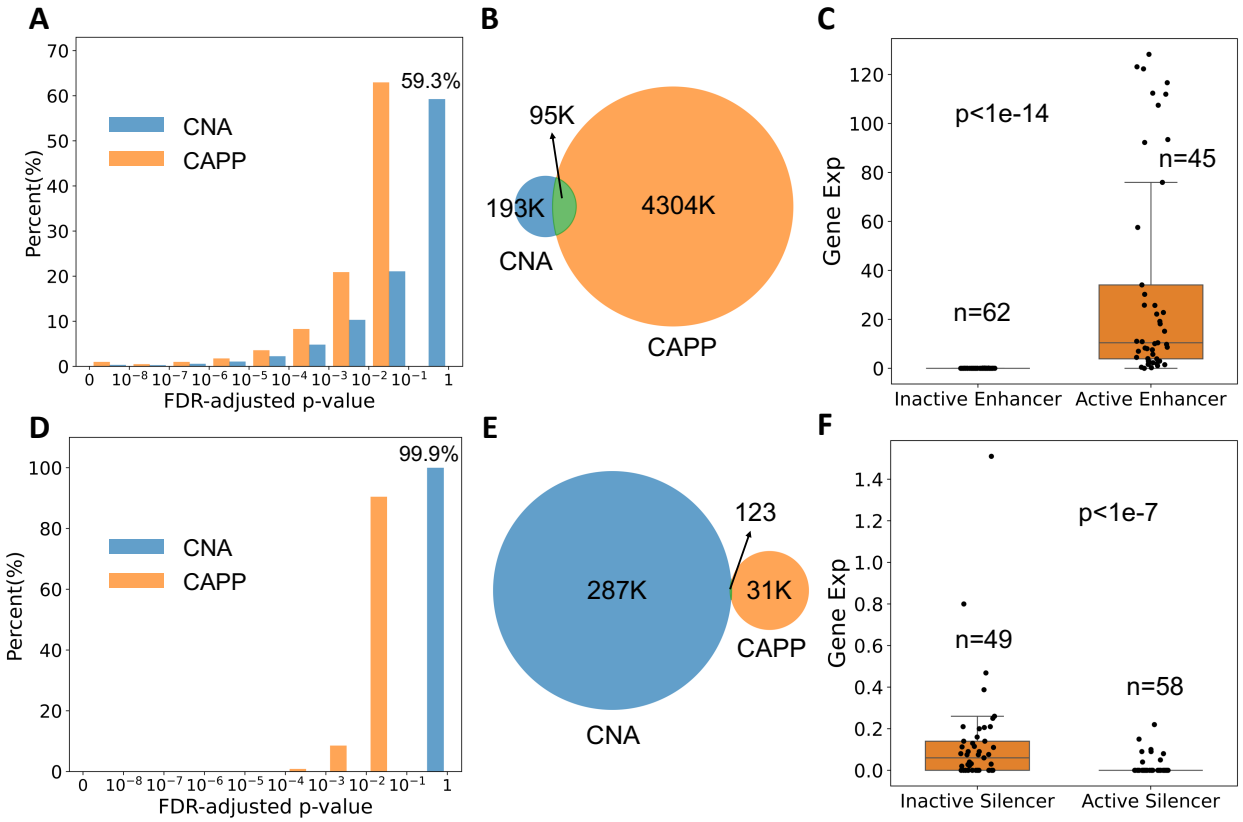


Figure 4-7. Comparison of CNA and our method CAPP for predicting target genes of CRMs. **A.** Distribution of FDR-adjusted p-values of enhancer-gene regulations assigned by CNA or predicted by CAPP. The p-values were calculated by one tailed Mann-Whitney U test. **B.** Venn diagram of enhancer-gene regulations assigned by CNA and enhancer-gene regulations predicted by CAPP. **C.** Boxplots of expression levels of gene *EIF1AY* against the functional states of its closest CRM chrY:20574642-20577538 in the 107 cell/tissue types. p < 1e-14, one tailed Mann-Whitney U test. The CRM chrY:20574642-20577538 is predicted by both CNA and CAPP to enhance the transcription of gene *EIF1AY*. **D.** Distribution of FDR-adjusted p-values of silencer-gene regulations assigned by CNA or predicted by CAPP. The p-values were calculated by one tailed Mann-Whitney U test. **E.** Venn diagram of silencer-gene regulations assigned by CNA and silencer-gene regulations predicted by CAPP. **F.** Boxplots of expression levels of *lnc-CD180-6* against the functional states of its closest CRM chr5:67798877-67801081 in the 107 cell/tissue types. p < 1e-7, one tailed Mann-Whitney U test. The CRM chr5:67798877-67801081 is predicted by both CNA and CAPP to repress the transcription of gene *lnc-CD180-6*.

4.2.10 Comparison of our method with the activity-by-contact (ABC) model predictions

We next compared our predicted enhancer-gene regulations with those predicted by the ABC model(144). The ABC model considers DNA sequences with DNase-seq and/or H3K27ac ChIP–seq signals as regulatory elements (REs), and calculates a score for each RE to regulate a gene within the two 5 Mb flanking regions around the RE. The score is defined as the product of the

strength of these epigenetic marks signals (Activity) in the RE and the Hi-C interaction frequency (Contact) between the RE and the gene. Using a predefined threshold of the score, the ABC model predicts a total of 175,860 non-redundant RE-gene links, involving 89,248 REs and 18,390 genes in five human tissue/cell types (GM12878, K562, liver, LNCAP and NCCIT). CAPP predicted 4,399,244 enhancer-gene links involving 240,024 enhancers and 47,765 genes in these five cell/tissue types. Therefore, CAPP predicted a much larger number of enhancer-gene links than did the ABC model in the five cell types.

We found that 51,501 (57.7%) of the 89,248 REs overlap 32,855 (13.7%) of our 240,024 enhancers by at least one bp (Figure 4-8A). We noted that multiple REs may overlap one our predicted enhancer, as REs are generally shorter than our enhancers (data not shown). We refer these overlapping REs and enhancers as ERo (enhancer-RE overlapping) REs and enhancers, respectively. The ERo REs and enhancers are involved in 99,241 (56.4%) RE-G (RE-gene) links and 638,789 (14.5%) enhancer-G (enhancer-gene) links, respectively. If a pair of overlapping ERo RE and enhancer regulate the same gene, we refer their respective link as an ERo RE-Gm (gene match) link and an ERo enhancer-Gm link (Figure 4-8A); otherwise, we refer their respective link as an ERo RE-Gnm (gene not match) link and an ERo enhancer-Gnm link (Figure 4-8A). This classification yielded 30,448 (17.3%) ERo RE-Gm links, 68,793 (39.1%) ERo RE-Gnm links, 21,521 (0.5%) ERo enhancer-Gm links and 617,268 (14.0%) ERo enhancer-Gnm links (Figure 4-8A). Thus, the ABC model predicted more RE-Gm links than enhancer-Gm links by CAPP, while CAPP predicted more enhancer-Gnm links than RE-Gnm links by the ABC model. Moreover, we found that 28,306 (31.7%) REs overlap our CRMs for which we were unable to predict their target genes using the data available to us, and we refer them as ERno-Co (Enhancer-RE not overlapping, but CRM-RE overlapping) REs, which are involved in 59,186 (33.7%) RE-G links (Figure 4-8A).

The remaining 9,441 (10.6%) REs do not overlap any of the CRMs within TADs, we refer them as CRno (CRM-RE not overlapping) REs, which are involved in 17,433 (9.9%) CRno RE-G links (Figure 4-8A). To see whether these CRno REs are functional, we plotted the distribution of their nucleotides' phyloP scores(62). As shown in Figure 4-8B, the density curve of CRno REs is more narrowly distributed with a high peak around 0, compared to those for the ERo REs and ERno-Co REs, indicating that the CRno REs are more not under natural selection, and thus, at least some of them might not be even parts of authentic enhancers (Figure 4-8B). Finally, 207,169 (86.3%) of our 240,024 enhancers do not overlap the REs, and we refer them as ERno (Enhancer-RE not overlapping) enhancers, which are involved in 3,760,455 (85.5%) enhancer-G links (Figure 4-8A).

Since H3K27ac is generally considered as a mark for active enhancers, the ABC model is actually aimed to predict target genes of active enhancers in a cell type(144). However, we noted that some REs from the five cell types overlapped our predicted inactive enhancers in the cell types, although the REs were identified by their H3K27ac signals. We thus further investigated the issue by applying our LR state predictor to the REs in the five cell types. Of the 20,982 REs in the K562 cells, while the majority (16,786 or 80.0%) were predicted to be active, the remaining considerable 4,196 (20.0%) were predicted to be inactive. By contrast, of the 1,178,225 CRMs in TADs, we predicted 84,083 (7.1%) to be active enhancers and the remaining 1,094,142 (92.9%) to be inactive enhancers in the K562 cells. Moreover, we predicted target genes for 60,164 (71.6%) of the 84,083 predicted active enhancers and for 179,860 (16.4%) of the 1,094,142 predicted inactive enhancers in the K562 cells. Though our LR predictor only used CA as the feature, our predictions are supported by the signals of the three active enhancer marks (CA, H3K27ac and H3K4me1) on the predicted i) active and inactive REs with either ERo RE-Gm links or ERo RE-Gnm links (Figure 4-8C), ii) active and inactive enhancers with either ERo enhancer-Gm links or ERo enhancer-Gnm

links (Figure 4-9A), iii) active and inactive REs with either ERno-Co RE-G links or CRno RE-G links (Figure 4-8D), and iv) active and inactive enhancers with ERno enhancer-G links (Figure 4-9B). Specifically, in all these cases, the three epigenetic marks are enriched around the centers of the predicted active RE or enhancers and are depleted around the centers of predicted inactive REs (Figures 4-8C and 4-8D) and enhancers (Figures 4-9A and 4-9B). Hence, our prediction of functional states for the REs are as accurate as for the enhancers using CA as the sole feature.

We reason that if the RE-G links or our enhancer-G links are correctly predicted, then the expression levels of the target genes of REs or enhancers should be significantly higher in certain cell/tissue types where the REs or enhancers are active than in other certain cell/tissue types where the REs or enhancers are inactive. To ensure statistical power in evaluating the RE-G links, we only consider REs that are active and inactive in at least five cell/tissue types. Our reasoning holds for all our enhancer-G links (Figure 4-7A). By contrast, it holds for only 77.8% of the ERo RE-Gm links (Figure 4-8E), suggesting that 22.2% of ERo RE-Gm links might be false positives. The discrepancy between ERo RE-Gm links and ERo enhancer-Gm arises due to our requirement of a single-bp overlap between REs and silencers, potentially leading to differing predictions of their functional states. Moreover, of the ERo RE-Gnm links, only 19.7% were identified as significant positive regulations at an FDR of 0.1, while the remaining majority (80.3%) might be false positives. Furthermore, our ERo enhancers were linked to an average of 19.4 target genes, while the overlapping REs were only linked to an average of 1.9 genes. Therefore, it is highly likely that the ABC model might overlook some regulations, given the fact that the human genome encodes at least 20 times as many CRMs as the genes(19, 48). In addition, only 32.7% of the ERno-Co RE-G links and 36.7% of the CRno RE-G links (Figure 4-8E), exhibit significant positive regulations (FDR of 0.1). Thus, our predictions might have missed some positive regulations predicted by the

ABC model, particularly, the significant predictions in the ERno-Co RE-G links and the CRno RE-G links. However, the ABC model might have disregarded a much larger number of ERno enhancers and their target genes predicted by our method CAPP. Interestingly, we predicted more enhancers to regulate a gene than the ABC model (91.2 vs 9.6). This might be due to an intrinsic limitation of the ABC model, as it may overlook some regulations due to the reduction in the weighted score for each regulation when a large number of enhancers regulate a gene. Considering that different methods only take into account a subset of features of up-regulations, both our approach and the ABC model merely scratch the surface of capturing the extensive landscape of true enhancer-G regulations.

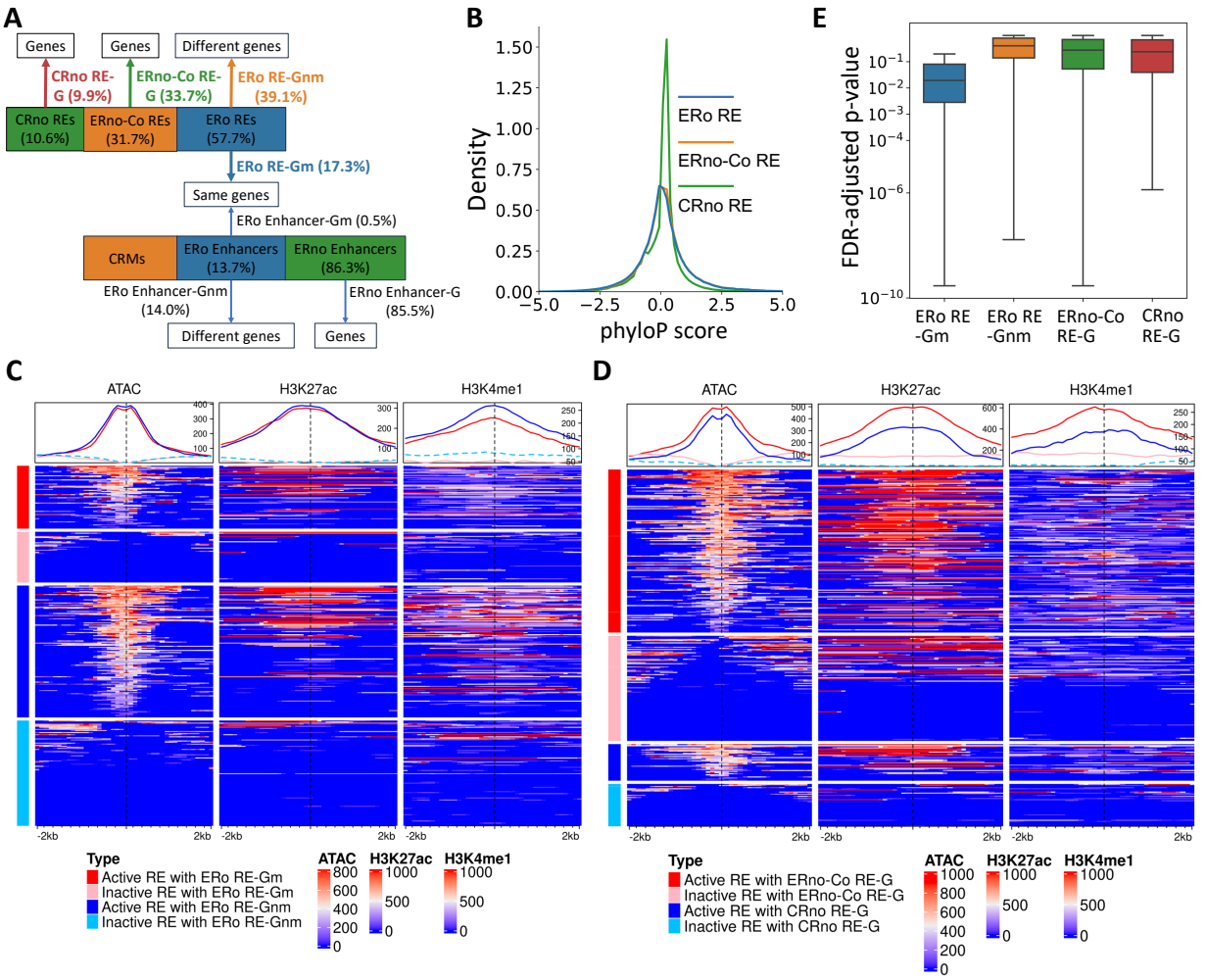


Figure 4-8. Comparison of our predicted enhancer-gene links with the RE-gene links predicted by the ABC model. **A.** A cartoon showing the overlaps between the RE-G links and the enhancer-G links. Based on the overlaps between the REs and our enhancers, we divide them into different categories: ERo enhancers and REs that overlap each other; ERno-Co REs that do not overlap the enhancers but overlap our other CRMs; CRno REs that do not overlap any CRMs; and ERno enhancers that do not overlap any REs. Accordingly, we also divide the RE-G links and enhancer-G links into different categories: ERo RE-Gm and ERo enhancer-Gm links are those with overlapping enhancers and REs and matched target genes; ERo RE-Gnm and ERo enhancer-Gnm links are those with overlapping enhancers and REs but different target genes; ERno-Co RE-G links are those for REs not overlapping enhancers but overlapping other CRMs; ERno enhancer-G links are those for enhancers that do not overlap REs; and CRno RE-G links are those for REs that do not overlap any CRMs. **B.** Density curves of PhyloP scores of the three categories of REs: ERo, ERno-Co and CRno. **C.** Heat maps of ATAC, H3K27ac and H3K4me1 signals of active and inactive REs with ERo RE-Gm links or ERo RE-Gnm links in the K562 cells. **D.** Heat maps of ATAC, H3K27ac and H3K4me1 signals of active and inactive REs with ERno-Co RE-G links or CRno RE-G links in the K562 cells. The heat maps show the mean signal of each 100 bp window in each sequence and the plot above each column of heat maps shows the mean signal of each window position across the sequences sampled in the same categories (Materials and Methods). The color code for the categories in the density plot above each column is the same with the heat

map legends. **E.** Boxplots of FDR-adjusted p-values for comparing the expression levels of putative target genes of the REs with different categories of regulation links in cell/tissue types where the REs are active with those in other cell/tissue types where the REs are inactive. The p-values were calculated by one tailed Mann-Whitney U test as used in our method (CAPP).

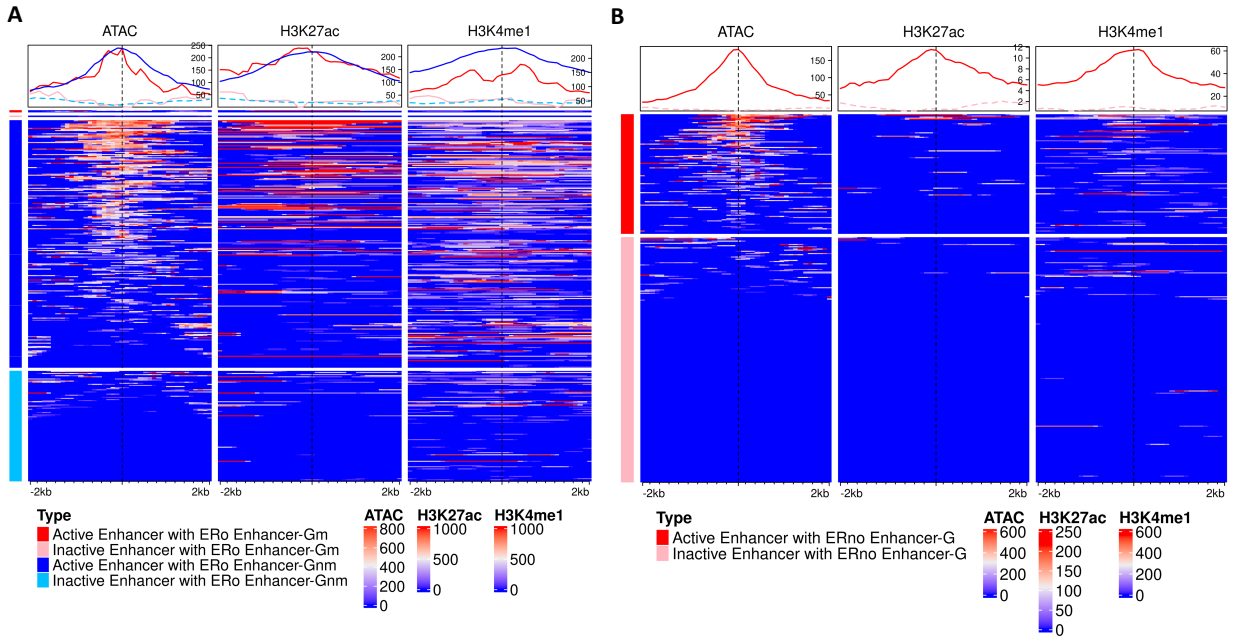


Figure 4-9. Heat maps of the three epigenetic marks around our enhancers. **A.** ATAC, H3K27ac and H3K4me1 signals of active and inactive enhancers with ERo enhancer-Gm links or ERo enhancer-Gnm links. **B.** ATAC, H3K27ac and H3K4me1 signals of active and inactive enhancers with ERno enhancer-G links. The heat maps show the mean signal of each window in each sequence and the density plot shows the mean signal of each window position across the sequences sampled in the same categories (Materials and Methods). The color code for the categories in the density plot above each column is the same with the heat map legends.

4.2.11 Comparison of our predicted silencer-gene links with those compiled in the silencerDB database

SilencerDB(113) documents a total of 33,060 validated and 5,045,547 predicted silencers (hereafter referred as REs). The predicted REs were collected from the CoSVM(105), the gkmSVM method(107) and a variant of gkmSVM called deepSilencer(158). SilencerDB also contains non-redundant 86,035 RE-G links predicted by paired expression and chromatin accessibility (PECA)(159), involving 79,604 REs and 10,195 target genes. We compared our predicted silencer-G (silencer-gene) links with those (hereafter referred as RE-G links) predicted

by PECA. Among the RE-gene links, 2,084 have their 1,932 (2.4%) REs overlapping 1,086 (9.4%) of our 11,592 silencers with predicted target genes. Like enhancers, we refer these overlapping silencers and REs as SRo (Silencer-RE overlapping) REs and SRo silencers, respectively (Figure 4-10A). Out of the 2,084 SRo RE-G links, 97 (0.1%) also target the same gene as overlapping silencers (SRo RE-Gm links), while 1,987 (2.3%) target different genes than overlapping silencers (SRo RE-Gnm links). The 1,086 (9.4%) SRo silencers are involved in 3,520 (11.2%) SRo silencer-G links, of which 55 target the same gene as overlapping REs involved in 55 (0.2%) SRo silencer-Gm links, while 1,079 target different genes than overlapping REs involved in 3,465 (11.0%) SRo silencer-Gnm links. It is important to note that an SRo REs and an SRo silencer can be involved in both types of links. Among the RE-G links whose REs do not overlap our silencers, 56,342 (65.4%) have their REs overlapping our CRMs for which we were unable to predict their target genes, and we refer them as SRno-Co REs; the remaining 27,609 (32.2%) have their REs not overlapping with our CRMs, and we refer them as CRno REs (Figure 4-10A). As expected, in contrast to the REs that overlap with our silencers or other CRMs, the CRno REs are more likely selectively neutral (Figure 4-10B), suggesting that more of CRno REs might be false positives (Figure 4-10B). Finally, 10,506 (90.6%) of our 11,592 silencers do not overlap the REs, and we refer them as SRno (Silencer-RE not overlapping) silencers, which are involved in 27,957 (88.8%) silencer-G links (Figure 4-10A).

As the functional states of the 79,604 REs with target genes predicted by PECA are unknown in the cell types, we predicted them using our LR model using CA as the sole feature. In K562 cells, 13,832 (17.4%) and 65,772 (82.6%) of the REs are predicted to be active and inactive, respectively. Among our 11,592 silencers with predicted target genes, 2,820 (24.3%) are predicted to be active, while the remaining 8,772 (75.7%) are deemed inactive in the K562 cells. As expected,

predicted active REs (Figures 4-10C and 4-10D) and silencers (Figures 4-11A and 4-11B) are enriched for the CA mark, while predicted inactive REs (Figures 4-10C and 4-10D) and silencers (Figures 4-11A and 4-11B) are depleted of the CA mark in the cells. However, unlike enhancers, there are no discernible patterns between active REs or silencers and inactive REs or silencers in other epigenetic marks such as H3K27me3. This suggests that H3K27me3 may not serve as a specific active silencer mark, as previously reported in a separate study(105).

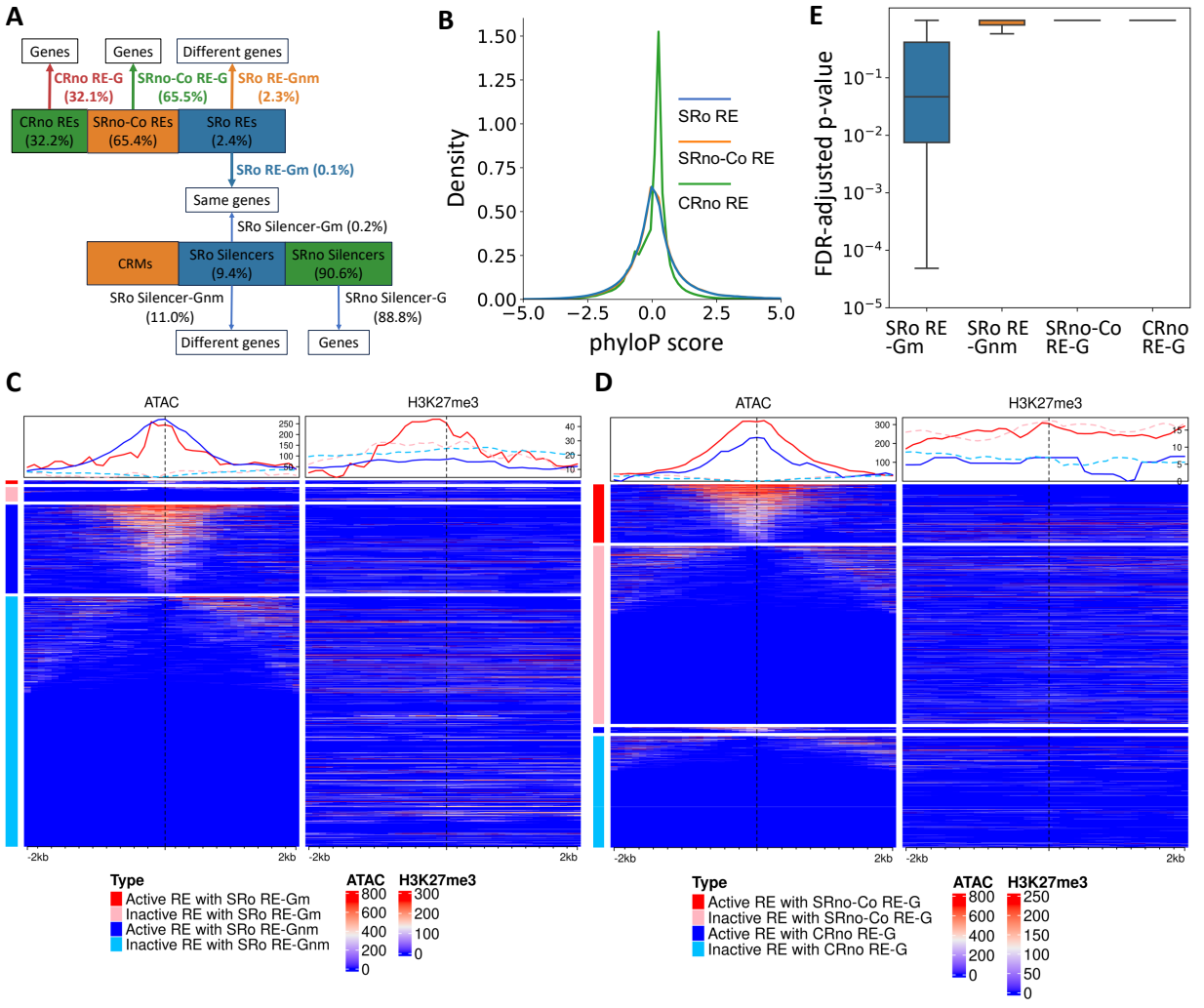


Figure 4-10. Comparison of our predicted silencer-gene links with the RE-gene links predicted by the PECA from silencerDB. **A.** A cartoon showing the overlaps between the RE-G links and our silencer-G links. Based on the overlaps between the REs and our silencers, we divide them into different categories: SRo silencers and REs that overlap each other; SRno-Co REs that do no overlap our silencers but overlap our other CRMs; CRno REs that do not overlap with any CRMs;

and SRno silencers that do not overlap REs. Accordingly, we also divide the RE-G links and silencer-G links into different categories: SRo silencer-Gm and SRo RE-Gm links are those with overlapping silencers and REs and matched target genes; SRo RE-Gnm and SRo silencer-Gnm links are those with overlapping silencers and REs but different target genes; SRno-Co RE-G links are those for REs not overlapping silencers but overlapping other CRMs, and SRno silencer-G links are those for silencers not overlapping REs; CRno RE-G links are those for REs that do not overlap any CRMs. **B.** Density curves of PhyloP scores of the three categories of REs: SRo, SRno-Co and CRno. **C.** Heat maps of ATAC and H3K27me3 signals of active and inactive REs with SRo RE-Gm links and SRo RE-Gnm links in the K562 cells. **D.** Heat maps of ATAC and H3K27me3 signals of active and inactive REs with SRno-Co RE-G links or CRno RE-G links in the K562 cells. The heat maps show the mean signal of each 100 bp window in each sequence and the plot above each column of heat maps shows the mean signal of each window position across the sequences sampled in the same categories (Materials and Methods). The color code for the categories in the density plot above each column is the same with the heat map legends. **E.** Boxplots of FDR-adjusted p-values for comparing the expression levels of putative target genes of the REs with different categories of regulation links in cell/tissue types where the REs are active with those in cell/tissue types where the REs are inactive. The p-values were calculated by one tailed Mann-Whitney U test as used in our method (CAPP).

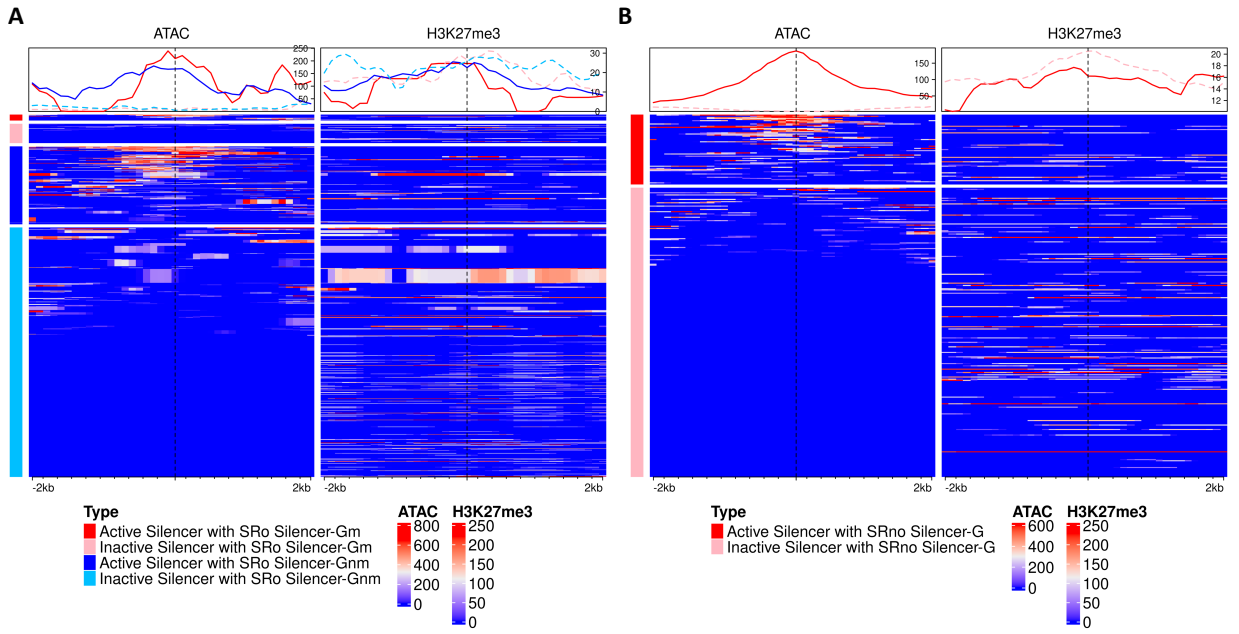


Figure 4-11. Heat maps of the two epigenetic marks around our silencers. **A.** ATAC and H3K27me3 signals of active and inactive silencers with SRo silencer-Gm links or SRo silencer-Gnm links. **B.** ATAC and H3K27me3 signals of active and inactive silencers with SRno silencer-G links. The heat map shows the mean signal in each window in each sequence and the density plot shows the mean signal of each window position across the sampled sequences in the same categories (Materials and Methods). The color code for the categories in the density plot above each column is the same with the heat map legends.

We reason that if the RE-G links or our silencer-G links were correctly predicted, then the expression levels of the target genes should be lower in some cell/tissue types where the REs or silencers are active than in other some cell/tissue types where the REs or silencers are inactive. To ensure statistical power in evaluating the RE-G links, we only consider REs that are active and inactive in at least 5 cell/tissue types. Our reasoning holds for all our silencer-G links (Figure 4-7D). In contrast, it holds for 59.7% of the SRo RE-Gm links (Figure 4-10E), while the remaining 40.3% might not be correct links. The discrepancy between SRo RE-Gm links and SRo silencer-Gm links arises due to our requirement of a single-bp overlap between REs and silencers, potentially leading to differing predictions of their functional states. Moreover, all the SRo RE-Gnm, SRno-Co RE-G, and CRno RE-G links do not exhibit significant negative regulation at an FDR of 0.1(Figure 4-10E). This suggests that none of these RE-G links are true regulations. Notably, our SRo silencers demonstrate an average regulation to 3.2 target genes, while overlapping REs are linked to just 1.1 genes on average. It is likely that PECA might overlook certain regulations, as a single silencer, on average, tends to regulate multiple target genes(160). Despite PECA predicting more regulations than our approach, an overwhelming majority (97.6%) diverges from the expected behavior of silencers, and thus, might be false positives.

4.3 Discussion

In this chapter, we introduced a new method CAPP to predict the target genes of CRMs. Leveraging on our previously predicted map of 1.2M CRMs in the human genome and functional states (active and inactive) of the CRMs, CAPP identifies target genes of the predicted CRMs within TADs by finding genes within the same TADs, whose expression levels are correlated with the functional states of the CRMs across a panel of cell/tissue types, followed by checking whether the CRMs and the genes are in close physical proximity as measured by Hi-C data in multiple

cell/tissue types. There are a few merits of our method. Firstly, CAPP can potentially enable us to predict the target genes of all CRMs in the genome when data are available from a sufficiently large number of diverse cell/ tissue types. Even using data in only 107 cell/tissue types in this study, we are able to predict target genes for around 20% of the 1.2M CRMs. Clearly, with the required data in a larger number of diverse cell/tissue types becoming available, we will be able to predict target genes for a higher proportion of the CRMs. Secondly, in addition to Hi-C data in a few cell types, CAPP only needs CA and RNA-seq data in a panel of cell types for the prediction, thus is highly cost-effective. Thirdly, in addition to target genes of CRMs, CAPP also is able to predict the functional types of CRMs as enhancers and/or silencers, the first of its kind, to the best of our knowledge. Fourthly, CAPP predicts the functional types of CRMs based on whether their functional states exhibit positive or negative correlation with the expression level of their target genes, respectively, in a panel of cell/tissue types, thus, overcoming a drawback of our previous LR models using three epigenetic marks(149). Fifthly, CAPP evaluates every pair of CRM and gene within the same TAD without a fixed distance constraint, allowing it to predict target genes located more than 5M bp away from the regulating CRM. This is the first of its kind, to our best knowledge, as previous methods predict target genes of candidate enhancers or silencers only in a fixed flanking genomic range. In fact, about 20% of our predicted enhancer-gene regulations have a regulation distance longer than 5M bp, which could be missed by existing methods that use a fixed distance constraint, typically 1~5M bp. Sixthly, although CAPP is aimed to predict cell type agnostic sCRNs encoded in the genome, an aCRN in any cell/tissue type can be readily induced from the sCRNs by the active CRMs predicted using only CA data in the cell/tissue type. Finally, CAPP predicts more CRM-gene links with higher accuracy than existing methods. The expression levels of our predicted target genes are significantly positively and negatively correlated with the

functional states of the predicted regulating enhancers and silencers, respectively. However, such correlations do not hold for a considerable proportion of target genes of enhancers and silencers predicted by the state-of-the-art methods, namely the ABC model and PECA, respectively. However, in the absence of a large gold standard set for enhancer-gene and silencer-gene regulations, it is still difficult to calculate the sensitivity and specificity of each method.

Based on our predicted enhancer-gene and silencer-gene regulations, we revealed distinct properties of various CRM types. Firstly, dual functional CRMs tend to regulate the greatest number of genes, followed by exclusive enhancers and exclusive silencers. Secondly, enhancers display a higher degree of cooperation in gene regulation than silencers. Thirdly, dual functional CRMs tend to regulate more distant genes than exclusive enhancers and exclusive silencers. Finally, enhancers prefer to regulate narrowly expressed genes, whereas silencers tend to regulate more broadly expressed genes.

4.4 Conclusion

In this study, we used a correlation and physical proximity method to predict both the functional types and target genes of CRMs simultaneously. Through this approach, we have identified millions of enhancer-gene regulations and tens of thousands of silencer-gene regulations. These findings highlight the diverse characteristics of different types of CRMs, their target genes, and their regulation links, providing insights into the distinct regulatory behaviors of exclusive enhancers, exclusive silencers, and dual functional CRMs. Importantly, our method surpasses traditional closest gene assignments and existing state-of-the-art methods, marking a notable advancement in predicting target genes and CRM functional types.

4.5 Materials and Methods

4.5.1 The Datasets

We obtained a comprehensive set of 1,225,115 predicted CRMs in the human genome from our prior work(19). We downloaded Hi-C data of six cell lines from the ENCODE(74) or 4D Nucleome(161) data portals (Supplementary Table S4-3). We downloaded DNase-seq, ATAC-seq and TF ChIP-seq data of the 67 human cell/tissue types for training from Cistrome Datasets Browser(110, 111) (Supplementary Table S4-4 and S4-5). We downloaded ATAC-seq data and RNA-seq data of the 107 cell/tissue types for predicting from ENCODE data portal(74) (Supplementary Table S4-6).

4.5.2 Generation of TADs

TADs were generated in each cell line at various resolutions (5K bp, 10K bp, 15K bp, 25K bp, 50K bp and 100K bp) using the Arrowhead algorithm of Juicer tools(76) version 2.17.00 with KR normalization method. We subsequently merged overlapping TADs from different resolutions into larger domains in each cell line using the bedtools2/2.29.0 merge command.

4.5.3 Identifications of CRMs within TADs

We identified the CRMs that overlap at least one bp with the merged TADs using the bedtools2/2.29.0 intersect command and ended up with 1,178,225 CRMs for further analysis.

4.5.4 CA feature score

For a sequence q, we define its raw CA feature score as:

$$F_{raw}(q) = \sum_{i=1}^{\mathrm{N}} r_i s_i \qquad (4-1)$$

where N is the number of peaks of CA mapping to q at least 50% of either one, r_i the ratio of overlapping length between q and the i_{th} peak of CA over the length of the i_{th} peak of CA, s_i the signal of the i_{th} peak of CA quantified by MACS2(128, 129). We then normalized the raw feature score in each cell/tissue type by the min-max normalization, i.e.,

$$F(q) = \frac{F_{raw}(q) - min\ (F_{raw}(Q))}{max\big(F_{raw}(Q)\big)\ - min\ (F_{raw}(Q))} \qquad (4-2)$$

where Q denotes all candidate sequences in the genome, $min\ (F_{raw}(Q))$ and $max\ (F_{raw}(Q))$ the minimum and maximum raw score of CA over Q in the cell/tissue type.

4.5.5 Quantification of gene expression levels

We computed log(TPM+1) for a gene as its expression level in a cell/tissue type. If multiple RNA-seq datasets were available for a cell/tissue type, we first computed the average expression level (mean(TPM)) of the gene across all the datasets and then computed the log(mean(TPM)+1).

4.5.6 Prediction of functional states of sequences

We employed a simple LR model using CA as the only feature to predict the functional state of the 1.2M CRMs within the TADs in each of the 107 cell/tissue types. A CRM in a cell/tissue type is considered active if its activation probability exceeds 0.5.

4.5.6.1 Construction of positive and negative sets: In each of the 67 cell/tissue types with the required data available, we selected the CRMs as the positive set that overlap TF binding peaks. At the same time, we randomly selected predicted non-CRM candidates with matched numbers of the positive set as the negative set in the cell/tissue type. We pooled the positive and negative sets in all the cell/tissue types to construct a comprehensive positive set and a negative set. The resulting positive set contains 1,784,345 CRMs and the negative set contains the same numbers of non-CRM candidates. Thus, the positive sets and negative sets are well-balanced.

4.5.6.2 Model training and evaluation: Ten-fold cross-validation was conducted to train and assess model performance. The models were implemented using sci-kit learn v.0.24.2 and the code is available at https://github.com/sisyyuan/Target-Gene-Prediction.

4.5.7 Prediction of target genes for CRMs

We predict target genes for the CRMs and at the same time identify their functional types as enhancers and/or silencers by using a one-tailed Mann-Whitney U test, with Benjamini-Hochberg (B-H) multiple hypothesis correction, followed by validation of physical proximity between the CRMs and the target genes using Hi-C interaction data in any of the six cell/tissue types (Supplementary Table S4-3). Assuming that the regulation of a gene by a CRM remains consistent across various cell/tissue types, then when the CRM is activated in a cell/tissue type, the expression level of the gene will elevate if the CRM functions as an enhancer, or will decrease if the CRM functions as a silencer. We adopted a statistical approach to assess the significance of such correlation.

4.5.7.1 Step 1: Test correlation between CRM activity and gene expression using Mann-Whitney U Test

For each pair of a CRM c and a gene g in a TAD D, we perform two one-tailed Mann-Whitney U Tests based on two datasets $Exp_{\text{active}}(g,c)$ and $Exp_{\text{inactive}}(g,c)$, where Exp_{active} is the expression levels of g in a minimum of t cell/tissue types where c is active (group A), and $Exp_{\text{in}active}$ is the expression levels of g in a minimum of t cell/tissue types where c is inactive (group B). The first test is to evaluate whether c function as an enhancer of g. Thus, the null hypothesis H_0 is: the median of group A is the same as or smaller than the median of group B, and the alternative hypothesis H_1 is: the median of group A is greater than the median of group B. The second test is to evaluate whether c function as a silencer of g. Thus, the null hypothesis H_0 is: the median of group A is the same as or greater than that of group B, and the alternative hypothesis H_1 is: the median of group A is smaller than the median of group B. The Mann-Whitney U Test function “mannwhitneyu” from the scipy.stat library in python3 was used to conduct the tests. Here, to ensure robust statistical analysis, we only consider the CRMs that are predicted to be

active and inactive in at least t cell/tissue types. We applied the B-H procedure "fdr_bh" from the "statsmodels.stats.multitest" library in python3 to correct the p-values with an FDR of 0.1.

In this study, we performed the tests based on the predicted functional states of the CRMs and experimentally measured expression levels of the genes in the 107 cell/tissue types, and we set t=5 to ensure robust statistical analysis. In other words, we only considered the CRMs that were active and inactive in a minimum of five cell/tissue types.

4.5.7.2 Step 2: Test for physical contact

As correlation does not guarantee causal regulation relationship, for each significant CRM-gene correlation, we checked whether the CRM and the gene have physical contact. We define a CRM and a gene to have physical contact, if both the CRM and target gene can be mapped to the respective ends of at least one pair of Hi-C reads from any of the six cell/tissue types (Supplementary Table S4-3), with at least one bp overlap. The Hi-C contact matrices were generated using the Straw algorithm(162) from the hicstraw library in python3 with SCALE normalization method at a resolution of 2000 bp.

4.5.8 Closest neighbor assignment to CRMs

For each CRM under consideration, we assign the gene whose TSS is linearly closest to either end of the CRM as its CNA target gene. In cases where a CRM overlaps a gene's TSS, we consider the gene as the CRM's target genes. Obviously, when TSSs of multiple genes are located within a CRM, multiple target genes will be assigned to the CRM.

4.5.9 The τ index

We used the τ index(157, 163) to measure the cell/tissue specificity of a gene based on its expression profiles across a panel of cell/tissue types, defined as:

$$\tau(g) = \frac{\sum_{i=1}^{N}(1 - e'_i)}{N - 1}, e'_i = \frac{e_i}{\max\limits_{1 \le j \le N}(e_j)} \qquad (4-3)$$

where e_i denotes the expression level of gene g in cell type i , e'_i the expression level of g normalized by the maximum of the expression levels of g in N cell/tissue types.

4.5.10 Heat maps of epigenetic marks

Heat maps of epigenetic mark signals were generated using the "EnrichedHeatmap" package(130) within R version 4.2.2. For a mark on each element in a set of sequences, we considered a 2K bp extension on either side of the element, and computed the mean signal of the mark in a 100 bp sliding window along the element (lower heat maps). For a mark in a set of sequences, we also computed the mean of the mean signal of the mark in the 100-bp sliding windows across all the sequences in the set (upper line annotations). The elements in each set of sequences were organized based on their CA signal in descending order.

4.6 Availability of data and materials

The datasets and code supporting the conclusions of this chapter are available at https://github.com/sisyyuan/Target-Gene-Prediction and are included within the chapter and its supplementary tables at https://github.com/sisyyuan/CRM_Dissertation.

Chapter 5
CONCLUSION AND FUTURE WORK

In this dissertation, we analyzed the organization and architecture of predicted CRMs in the human and mouse genomes, and developed computational methods for predicting functional types, states, and target genes of CRMs using few functional genomics data.

We revealed common rules for the organization and architecture of CRMs in the human and mouse genomes. CRM abundance on a chromosome correlates with the size and gene abundance on the chromosome. Like genes, CRMs also are highly unevenly distributed along chromosomes, forming “islands” and “deserts”. CRMs can be classified into two categories CPC and CPL depending on whether they overlap TSSs or not. CPC CRMs are generally longer than CPL CRMs. Within CRMs, TFBSs have extensive overlaps, forming islands, suggesting potential competitive or cooperative binding of different TFs. Finally, the spacers between TFBS islands exhibit similar evolutionary constraints to TFBS islands, indicating their alternative functional roles beyond direct TF binding in transcriptional regulation.

Our RL models are able to simultaneously predict the functional states and types of CRMs in any cell/tissue types using only five epigenetic marks data. Applying the models to 56 human cell/tissue types with the required data available, we revealed different types of CRMs: predominant enhancers, predominant silencers and dual functional CRMs. Different types of CRMs display distinct properties in lengths and TFBS densities, reflecting the complexity of their functions. Moreover, we found that both dual functional CRMs and silencers might be more prevalent than previously assumed.

Our target gene prediction method CAPP is able to not only predict target genes, but also more accurately predict functional types of CRMs using only CA and RNA-seq data in a panel of cell/tissue types plus Hi-C data in few cell lines. Applying CAPP to 107 human cell/tissue types,

we predict target genes for 20% of the 1.2M CRMs, of which 4.5% function as both enhancers and silencers (dual functional CRMs), 95.2% as exclusive enhancers and 0.3% as exclusive silencers. The different types of these CRMs show distinct properties in the numbers and expression patterns of their target genes as well as regulatory lengths. CAPP outperforms state-of-the-art methods, and thus represents a significant advancement in predicting target genes and functional types of CRMs.

In the future, we will further our research by focusing on the following issues. Firstly, while predicting functional types and states of CRMs, we utilized cell/tissue type data, which aggregated signals from a population of not necessarily a homogeneous cell type, and thus lacked single-cell resolution. Integrating single-cell multi-omics data in future analyses promises more precise CRM functional predictions across diverse cellular contexts. Secondly, although we have identified CRMs as enhancers and silencers, understanding how they cooperatively regulate their target genes remains elusive. Thus, further studies are needed to unveil intricate gene regulatory networks. Thirdly, as complex traits including diseases are mainly caused by variation in CRMs, integration of GWAS data with CRMs and their functional states and target genes presents an opportunity to identify causal non-coding variants of diseases. Revealing the chains of events linking causal non-coding variants to diseases within CRM-gene regulatory networks would uncover key players that could be therapeutically targeted, offering novel disease treatments. Lastly, establishing publicly accessible databases housing CRM maps as well as their functional types, states, and target genes would foster data sharing and collaboration within the research community.

REFERENCE

1. Robert F, Pelletier J. Exploring the Impact of Single-Nucleotide Polymorphisms on Translation. Frontiers in Genetics. 2018;9.
2. Zhang F, Lupski JR. Non-coding genetic variants in human disease. Hum Mol Genet. 2015;24(R1):R102-10.
3. Kumar V, Westra HJ, Karjalainen J, Zhernakova DV, Esko T, Hrdlickova B, et al. Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. PLoS Genet. 2013;9(1):e1003201.
4. Giral H, Landmesser U, Kratzer A. Into the Wild: GWAS Exploration of Non-coding RNAs. Front Cardiovasc Med. 2018;5:181.
5. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009;106(23):9362-7.
6. Ramos EM, Hoffman D, Junkins HA, Maglott D, Phan L, Sherry ST, et al. Phenotype-Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. Eur J Hum Genet. 2014;22(1):144-7.
7. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science. 2012;337(6099):1190-5.
8. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet. 2017;101(1):5-22.
9. Huang D, Ovcharenko I. Enhancer-silencer transitions in the human genome. Genome Res. 2022;32(3):437-48.
10. Erceg J, Pakozdi T, Marco-Ferreres R, Ghavi-Helm Y, Girardot C, Bracken AP, et al. Dual functionality of cis-regulatory elements as developmental enhancers and Polycomb response elements. Gene Dev. 2017;31(6):590-602.
11. Hardison RC, Taylor J. Genomic approaches towards finding cis-regulatory modules in animals. Nat Rev Genet. 2012;13(7):469-83.
12. Suryamohan K, Halfon MS. Identifying transcriptional cis-regulatory modules in animal genomes. Wiley Interdiscip Rev Dev Biol. 2015;4(2):59-84.
13. Chen D, Lei EP. Function and regulation of chromatin insulators in dynamic genome organization. Curr Opin Cell Biol. 2019;58:61-8.
14. Liu L, Jin G, Zhou X. Modeling the relationship of epigenetic modifications to transcription factor binding. Nucleic Acids Res. 2015;43(8):3873-85.
15. Wang M, Zhang K, Ngo V, Liu C, Fan S, Whitaker JW, et al. Identification of DNA motifs that regulate DNA methylation. Nucleic Acids Res. 2019;47(13):6753-68.
16. Hoellinger T, Mestre C, Aschard H, Le Goff W, Foissac S, Faraut T, et al. Enhancer/gene relationships: Need for more reliable genome-wide reference sets. Front Bioinform. 2023;3:1092853.
17. Zhang ZD, Paccanaro A, Fu Y, Weissman S, Weng Z, Chang J, et al. Statistical analysis of the genomic distribution and correlation of regulatory elements in the ENCODE regions. Genome Res. 2007;17(6):787-97.
18. Symmons O, Uslu VV, Tsujimura T, Ruf S, Nassari S, Schwarzer W, et al. Functional and topological characteristics of mammalian regulatory domains. Genome Res. 2014;24(3):390-400.
19. Ni P, Su Z. Accurate prediction of cis-regulatory modules reveals a prevalent regulatory genome of humans. NAR Genom Bioinform. 2021;3(2):lqab052.

20. Ni P, Moe J, Su Z. Accurate prediction of functional states of cis-regulatory modules reveals common epigenetic rules in humans and mice. BMC Biol. 2022;20(1):221.
21. Ni P, Wilson D, Su Z. A map of cis-regulatory modules and constituent transcription factor binding sites in 80% of the mouse genome. BMC Genomics. 2022;23(1):714.
22. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, et al. High-resolution mapping and characterization of open chromatin across the genome. Cell. 2008;132(2):311-22.
23. Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee BK, et al. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res. 2011;21(10):1757-67.
24. Crawford GE, Holt IE, Whittle J, Webb BD, Tai D, Davis S, et al. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Res. 2006;16(1):123-31.
25. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 2013;10(12):1213-8.
26. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, et al. High-resolution profiling of histone methylations in the human genome. Cell. 2007;129(4):823-37.
27. Catarino RR, Stark A. Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation. Genes Dev. 2018;32(3-4):202-23.
28. Kheradpour P, Ernst J, Melnikov A, Rogov P, Wang L, Zhang X, et al. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. Genome Res. 2013;23(5):800-11.
29. Dogan N, Wu W, Morrissey CS, Chen KB, Stonestrom A, Long M, et al. Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility. Epigenetics & chromatin. 2015;8:16.
30. Arbel H, Basu S, Fisher WW, Hammonds AS, Wan KH, Park S, et al. Exploiting regulatory heterogeneity to systematically identify enhancers with high accuracy. Proc Natl Acad Sci U S A. 2019;116(3):900-8.
31. Kwasnieski JC, Fiore C, Chaudhari HG, Cohen BA. High-throughput functional testing of ENCODE segmentation predictions. Genome Res. 2014;24(10):1595-602.
32. Kleftogiannis D, Kalnis P, Bajic VB. DEEP: a general computational framework for predicting enhancers. Nucleic Acids Res. 2015;43(1):e6.
33. Podsiadlo A, Wrzesien M, Paja W, Rudnicki W, Wilczynski B. Active enhancer positions can be accurately predicted from chromatin marks and collective sequence motif data. BMC Syst Biol. 2013;7 Suppl 6:S16.
34. Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi-C: a comprehensive technique to capture the conformation of genomes. Methods. 2012;58(3):268-76.
35. Li G, Cai L, Chang H, Hong P, Zhou Q, Kulakova EV, et al. Chromatin Interaction Analysis with Paired-End Tag (ChIA-PET) sequencing technology and application. BMC Genomics. 2014;15 Suppl 12(Suppl 12):S11.
36. Sikora-Wohlfeld W, Ackermann M, Christodoulou EG, Singaravelu K, Beyer A. Assessing computational methods for transcription factor target gene identification based on ChIP-seq data. PLoS Comput Biol. 2013;9(11):e1003342.
37. Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database (Oxford). 2017;2017.

38. Sheffield NC, Thurman RE, Song L, Safi A, Stamatoyannopoulos JA, Lenhard B, et al. Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. Genome Res. 2013;23(5):777-88.
39. Silva TC, Coetzee SG, Gull N, Yao L, Hazelett DJ, Noushmehr H, et al. ELMER v.2: an R/Bioconductor package to reconstruct gene regulatory networks from DNA methylation and transcriptome profiles. Bioinformatics. 2019;35(11):1974-7.
40. He B, Chen C, Teng L, Tan K. Global view of enhancer-promoter interactome in human cells. Proc Natl Acad Sci U S A. 2014;111(21):E2191-9.
41. Hariprakash JM, Ferrari F. Computational Biology Solutions to Identify Enhancers-target Gene Pairs. Comput Struct Biotechnol J. 2019;17:821-31.
42. Kielbasa SM, Bluthgen N, Fahling M, Mrowka R. Targetfinder.org: a resource for systematic discovery of transcription factor target genes. Nucleic Acids Res. 2010;38(Web Server issue):W233-8.
43. Zhao C, Li X, Hu H. PETModule: a motif module based approach for enhancer target gene prediction. Sci Rep. 2016;6:30043.
44. Hafez D, Karabacak A, Krueger S, Hwang YC, Wang LS, Zinzen RP, et al. McEnhancer: predicting gene expression via semi-supervised assignment of enhancers to target genes. Genome Biol. 2017;18(1):199.
45. Gasperini M, Tome JM, Shendure J. Towards a comprehensive catalogue of validated and target-linked human enhancers. Nat Rev Genet. 2020;21(5):292-310.
46. Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57-74.
47. Stamatoyannopoulos JA. What does our genome encode? Genome Res. 2012;22(9):1602-11.
48. Consortium EP, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020;583(7818):699-710.
49. Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, et al. A comparative encyclopedia of DNA elements in the mouse genome. Nature. 2014;515(7527):355-64.
50. Sharov AA, Dudekula DB, Ko MS. CisView: a browser and database of cis-regulatory modules predicted in the mouse genome. DNA Res. 2006;13(3):123-34.
51. Vierstra J, Rynes E, Sandstrom R, Zhang M, Canfield T, Hansen RS, et al. Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. Science. 2014;346(6212):1007-12.
52. Davidson EH. The Regulatory Genome: Gene Regulatory Networks In Development And Evolution. Amsterdam: Academic Press; 2006.
53. Biggin MD. Animal transcription networks as highly connected, quantitative continua. Dev Cell. 2011;21(4):611-26.
54. Watson LC, Kuchenbecker KM, Schiller BJ, Gross JD, Pufall MA, Yamamoto KR. The glucocorticoid receptor dimer interface allosterically transmits sequence-specific DNA signals. Nat Struct Mol Biol. 2013;20(7):876-83.
55. King DM, Hong CKY, Shepherdson JL, Granas DM, Maricque BB, Cohen BA. Synthetic and genomic regulatory elements reveal aspects of cis-regulatory grammar in mouse embryonic stem cells. Elife. 2020;9.
56. Szabo Q, Bantignies F, Cavalli G. Principles of genome folding into topologically associating domains. Sci Adv. 2019;5(4).

57. Karnuta JM, Scacheri PC. Enhancers: bridging the gap between gene control and human disease. Hum Mol Genet. 2018;27(R2):R219-r27.
58. Consortium F, the RP, Clst, Forrest AR, Kawaji H, Rehli M, et al. A promoter-level mammalian expression atlas. Nature. 2014;507(7493):462-70.
59. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014;507(7493):455-61.
60. Visel A, Minovitsky S, Dubchak I, Pennacchio LA. VISTA Enhancer Browser--a database of tissue-specific human enhancers. Nucleic Acids Res. 2007;35(Database issue):D88-92.
61. Levine M, Tjian R. Transcription regulation and animal diversity. Nature. 2003;424(6945):147-51.
62. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010;20(1):110-21.
63. Huang YF, Gulko B, Siepel A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nat Genet. 2017;49(4):618-24.
64. Meuleman W, Muratov A, Rynes E, Halow J, Lee K, Bates D, et al. Index and biological spectrum of human DNase I hypersensitive sites. Nature. 2020;584(7820):244-51.
65. Vierstra J, Lazar J, Sandstrom R, Halow J, Lee K, Bates D, et al. Global reference mapping of human transcription factor footprints. Nature. 2020;583(7818):729-36.
66. Zerbino DR, Wilder SP, Johnson N, Juettemann T, Flicek PR. The ensembl regulatory build. Genome Biol. 2015;16:56.(doi):10.1186/s13059-015-0621-5.
67. Andersson R, Sandelin A, Danko CG. A unified architecture of transcriptional regulatory elements. Trends Genet. 2015;31(8):426-33.
68. Kim TK, Shiekhattar R. Architectural and Functional Commonalities between Enhancers and Promoters. Cell. 2015;162(5):948-59.
69. Li XY, Thomas S, Sabo PJ, Eisen MB, Stamatoyannopoulos JA, Biggin MD. The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding. Genome Biol. 2011;12(4):R34.
70. Thurmond J, Goodman JL, Strelets VB, Attrill H, Gramates LS, Marygold SJ, et al. FlyBase 2.0: the next generation. Nucleic acids research. 2019;47(D1):D759-D65.
71. Kamar RI, Banigan EJ, Erbas A, Giuntoli RD, De La Cruz MO, Johnson RC, et al. Facilitated dissociation of transcription factors from single DNA binding sites. Proceedings of the National Academy of Sciences. 2017;114(16):E3251-E7.
72. Panne D, Maniatis T, Harrison SC. Crystal structure of ATF-2/c-Jun and IRF-3 bound to the interferon-β enhancer. The EMBO journal. 2004;23(22):4384-93.
73. Ni P, Su Z. PCRMS: a database of predicted cis-regulatory modules and constituent transcription factor binding sites in genomes. Database : the journal of biological databases and curation. 2022;2022:baac024.
74. Luo Y, Hitz BC, Gabdank I, Hilton JA, Kagda MS, Lam B, et al. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. Nucleic Acids Res. 2020;48(D1):D882-D9.
75. Bonferroni C. Teoria statistica delle classi e calcolo delle probabilita. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze. 1936;8: 3-62.
76. Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, et al. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Syst. 2016;3(1):95-8.

77. Knight PA, Ruiz D. A fast algorithm for matrix balancing. Ima J Numer Anal. 2013;33(3):1029-47.
78. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326(5950):289-93.
79. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science. 2007;316(5830):1497-502.
80. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods. 2007;4(8):651-7.
81. Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB, et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell. 2008;133(6):1106-17.
82. Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi-C: A comprehensive technique to capture the conformation of genomes. Methods. 2012.
83. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008;320(5881):1344-9.
84. Consortium TEP. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science. 2004;306(5696):636-40.
85. Stamatoyannopoulos JA, Snyder M, Hardison R, Ren B, Gingeras T, Gilbert DM, et al. An encyclopedia of mouse DNA elements (Mouse ENCODE). Genome Biol. 2012;13(8):418.
86. Snyder MP, Gingeras TR, Moore JE, Weng Z, Gerstein MB, Ren B, et al. Perspectives on ENCODE. Nature. 2020;583(7818):693-8.
87. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, et al. The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol. 2010;28(10):1045-8.
88. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518(7539):317-30. doi: 10.1038/nature14248.
89. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45(6):580-5. doi: 10.1038/ng.2653.
90. Forrest AR, Kawaji H, Rehli M, Baillie JK, de Hoon MJ, Haberle V, et al. A promoter-level mammalian expression atlas. Nature. 2014;507(7493):462-70.
91. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014;507(7493):455-61.
92. Paul DS, Soranzo N, Beck S. Functional interpretation of non-coding sequence variation: concepts and challenges. Bioessays. 2014;36(2):191-9. doi: 10.1002/bies.201300126. Epub 2013 Dec 5.
93. Kleftogiannis D, Kalnis P, Bajic VB. Progress and challenges in bioinformatics approaches for enhancer identification. BriefBioinform. 2016;17(6):967-79.
94. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57-74.
95. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473(7345):43-9.

96. He Y, Gorkin DU, Dickel DE, Nery JR, Castanon RG, Lee AY, et al. Improved regulatory element prediction based on tissue-specific local epigenomic signatures. Proc Natl Acad Sci U S A. 2017;114(9):E1633-e40.
97. Rajagopal N, Xie W, Li Y, Wagner U, Wang W, Stamatoyannopoulos J, et al. RFECS: a random-forest based algorithm for enhancer identification from chromatin state. PLoS Comput Biol. 2013;9(3):e1002968.
98. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518(7539):317-30.
99. Bailey TL. STREME: Accurate and versatile sequence motif discovery. Bioinformatics. 2021;37(18):2834-40.
100. Li Y, Ni P, Zhang S, Li G, Su Z. ProSampler: an ultrafast and accurate motif finder in large ChIP-seq datasets for combinatory motif discovery. Bioinformatics. 2019;35(22):4632-9.
101. Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, Adrian J, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020;583(7818):699-710.
102. Halfon MS. Silencers, Enhancers, and the Multifunctional Regulatory Genome. Trends Genet. 2020;36(3):149-51.
103. Pang BX, van Weerd JH, Hamoen FL, Snyder MP. Identification of non-coding silencer elements and their regulation of gene expression. Nat Rev Mol Cell Bio. 2023;24(6):383-95.
104. Gisselbrecht SS, Palagi A, Kurland JV, Rogers JM, Ozadam H, Zhan Y, et al. Transcriptional Silencers in Drosophila Serve a Dual Role as Transcriptional Enhancers in Alternate Cellular Contexts. Mol Cell. 2020;77(2):324-37 e8.
105. Huang D, Petrykowska HM, Miller BF, Elnitski L, Ovcharenko I. Identification of human silencers by correlating cross-tissue epigenetic profiles and gene expression. Genome Res. 2019;29(4):657-67.
106. Pang B, Snyder MP. Systematic identification of silencers in human cells. Nat Genet. 2020;52(3):254-63.
107. Doni Jayavelu N, Jajodia A, Mishra A, Hawkins RD. Candidate silencer elements for the human and mouse genomes. Nat Commun. 2020;11(1):1061.
108. Soares LM, He PC, Chun Y, Suh H, Kim T, Buratowski S. Determinants of Histone H3K4 Methylation Patterns. Mol Cell. 2017;68(4):773-85 e6.
109. Wang H, Fan Z, Shliaha PV, Miele M, Hendrickson RC, Jiang X, et al. H3K4me3 regulates RNA polymerase II promoter-proximal pause-release. Nature. 2023;615(7951):339-48.
110. Mei S, Qin Q, Wu Q, Sun H, Zheng R, Zang C, et al. Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. Nucleic Acids Res. 2017;45(D1):D658-d62.
111. Zheng R, Wan C, Mei S, Qin Q, Wu Q, Sun H, et al. Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. Nucleic Acids Res. 2019;47(D1):D729-D35.
112. Friedman RZ, Granas DM, Myers CA, Corbo JC, Cohen BA, White MA. Information content differentiates enhancers from silencers in mouse photoreceptors. Elife. 2021;10.
113. Zeng W, Chen S, Cui X, Chen X, Gao Z, Jiang R. SilencerDB: a comprehensive database of silencers. Nucleic Acids Res. 2021;49(D1):D221-D8.
114. Zentner GE, Tesar PJ, Scacheri PC. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. Genome Res. 2011;21(8):1273-83.

115. Yaragatti M, Basilico C, Dailey L. Identification of active transcriptional regulatory modules by the functional assay of DNA from nucleosome-free regions. Genome Res. 2008;18(6):930-8.
116. Hansen TJ, Hodges E. ATAC-STARR-seq reveals transcription factor-bound activators and silencers across the chromatin accessible human genome. Genome Res. 2022;32(8):1529-41.
117. Ninova M, Fejes Tóth K, Aravin AA. The control of gene expression and cell identity by H3K9 trimethylation. Development. 2019;146(19).
118. Padeken J, Methot SP, Gasser SM. Establishment of H3K9-methylated heterochromatin and its functions in tissue differentiation and maintenance. Nat Rev Mol Cell Biol. 2022;23(9):623-40.
119. Cai Y, Zhang Y, Loh YP, Tng JQ, Lim MC, Cao Z, et al. H3K27me3-rich genomic regions can function as silencers to repress gene expression via chromatin interactions. Nat Commun. 2021;12(1):719.
120. Sethi A, Gu M, Gumusgoz E, Chan L, Yan KK, Rozowsky J, et al. Supervised enhancer prediction with epigenetic pattern recognition and targeted validation. Nat Methods. 2020;17(8):807-14.
121. Rao S, Ahmad K, Ramachandran S. Cooperative binding between distant transcription factors is a hallmark of active enhancers. Mol Cell. 2021;81(8):1651-65.e4.
122. Ni P, Wu S, Su Z. Underlying causes for prevalent false positives and false negatives in STARR-seq data. NAR Genom Bioinform. 2023;5(3):lqad085.
123. Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science. 2013;339(6123):1074-7.
124. Wang X, He L, Goggin SM, Saadat A, Wang L, Sinnott-Armstrong N, et al. High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. Nat Commun. 2018;9(1):5380.
125. Liu Y, Yu S, Dhiman VK, Brunetti T, Eckart H, White KP. Functional assessment of human enhancer activities using whole-genome STARR-sequencing. Genome Biol. 2017;18(1):219.
126. Inoue F, Kircher M, Martin B, Cooper GM, Witten DM, McManus MT, et al. A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. Genome Res. 2017;27(1):38-52.
127. Klein JC, Agarwal V, Inoue F, Keith A, Martin B, Kircher M, et al. A systematic evaluation of the design and context dependencies of massively parallel reporter assays. Nat Methods. 2020;17(11):1083-91.
128. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137.
129. Feng J, Liu T, Qin B, Zhang Y, Liu XS. Identifying ChIP-seq enrichment using MACS. Nat Protoc. 2012;7(9):1728-40.
130. Gu Z, Eils R, Schlesner M, Ishaque N. EnrichedHeatmap: an R/Bioconductor package for comprehensive visualization of genomic signal associations. BMC Genomics. 2018;19(1):234.
131. Robinson JT, Turner D, Durand NC, Thorvaldsdóttir H, Mesirov JP, Aiden EL. Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data. Cell Syst. 2018;6(2):256-+.
132. Moore JE, Pratt HE, Purcaro MJ, Weng Z. A curated benchmark of enhancer-gene interactions for evaluating enhancer-target gene prediction methods. Genome Biol. 2020;21(1):17.
133. O'Connor T, Grant CE, Boden M, Bailey TL. T-Gene: improved target gene prediction. Bioinformatics. 2020;36(12):3902-4.

134. Ron G, Globerson Y, Moran D, Kaplan T. Promoter-enhancer interactions identified from Hi-C data using probabilistic models and hierarchical topological domains. Nat Commun. 2017;8(1):2237.
135. Fortin J-P, Hansen KD. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. Genome Biology. 2015;16(1):180.
136. Davidson IF, Bauer B, Goetz D, Tang W, Wutz G, Peters JM. DNA loop extrusion by human cohesin. Science. 2019;366(6471):1338-45.
137. Fudenberg G, Abdennur N, Imakaev M, Goloborodko A, Mirny LA. Emerging Evidence of Chromosome Folding by Loop Extrusion. Cold Spring Harb Symp Quant Biol. 2017;82:45-55.
138. Roayaei Ardakany A, Gezer HT, Lonardi S, Ay F. Mustache: multi-scale detection of chromatin loops from Hi-C and Micro-C maps using scale-space representation. Genome Biol. 2020;21(1):256.
139. Canver MC, Bauer DE, Orkin SH. Functional interrogation of non-coding DNA through CRISPR genome editing. Methods. 2017;121-122:118-29.
140. Joung J, Engreitz JM, Konermann S, Abudayyeh OO, Verdine VK, Aguet F, et al. Genome-scale activation screen identifies a lncRNA locus regulating a gene neighbourhood. Nature. 2017;548(7667):343-6.
141. Liu SJ, Horlbeck MA, Cho SW, Birk HS, Malatesta M, He D, et al. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. Science. 2017;355(6320).
142. Stuart WD, Guo M, Fink-Baldauf IM, Coleman AM, Clancy JP, Mall MA, et al. CRISPRi-mediated functional analysis of lung disease-associated loci at non-coding regions. NAR Genom Bioinform. 2020;2(2):lqaa036.
143. Tian R, Gachechiladze MA, Ludwig CH, Laurie MT, Hong JY, Nathaniel D, et al. CRISPR Interference-Based Platform for Multimodal Genetic Screens in Human iPSC-Derived Neurons. Neuron. 2019;104(2):239-55 e12.
144. Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, et al. Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. NatGenet. 2019;51(12):1664-9.
145. Pennacchio LA, Bickmore W, Dean A, Nobrega MA, Bejerano G. Enhancers: five essential questions. Nat Rev Genet. 2013;14(4):288-95.
146. Ferretti E, Cambronero F, Tumpel S, Longobardi E, Wiedemann LM, Blasi F, et al. Hoxb1 enhancer and control of rhombomere 4 expression: Complex interplay between PREP1-PBX1-HOXB1 binding sites. Mol Cell Biol. 2005;25(19):8541-52.
147. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature. 2012;489(7414):75-82.
148. Aran D, Sabato S, Hellman A. DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. Genome Biol. 2013;14(3):R21.
149. Yuan S, Ni P, Su Z. Simultaneous Prediction of Functional States and Types of cis-regulatory Modules Reveals Their Prevalent Dual Uses as Enhancers and Silencers. bioRxiv. 2024:2024.05.07.592879.
150. Acemel RD, Maeso I, Gomez-Skarmeta JL. Topologically associated domains: a successful scaffold for the evolution of gene regulation in animals. Wiley Interdiscip Rev Dev Biol. 2017;6(3).
151. Bolt CC, Duboule D. The regulatory landscapes of developmental genes. Development. 2020;147(3).

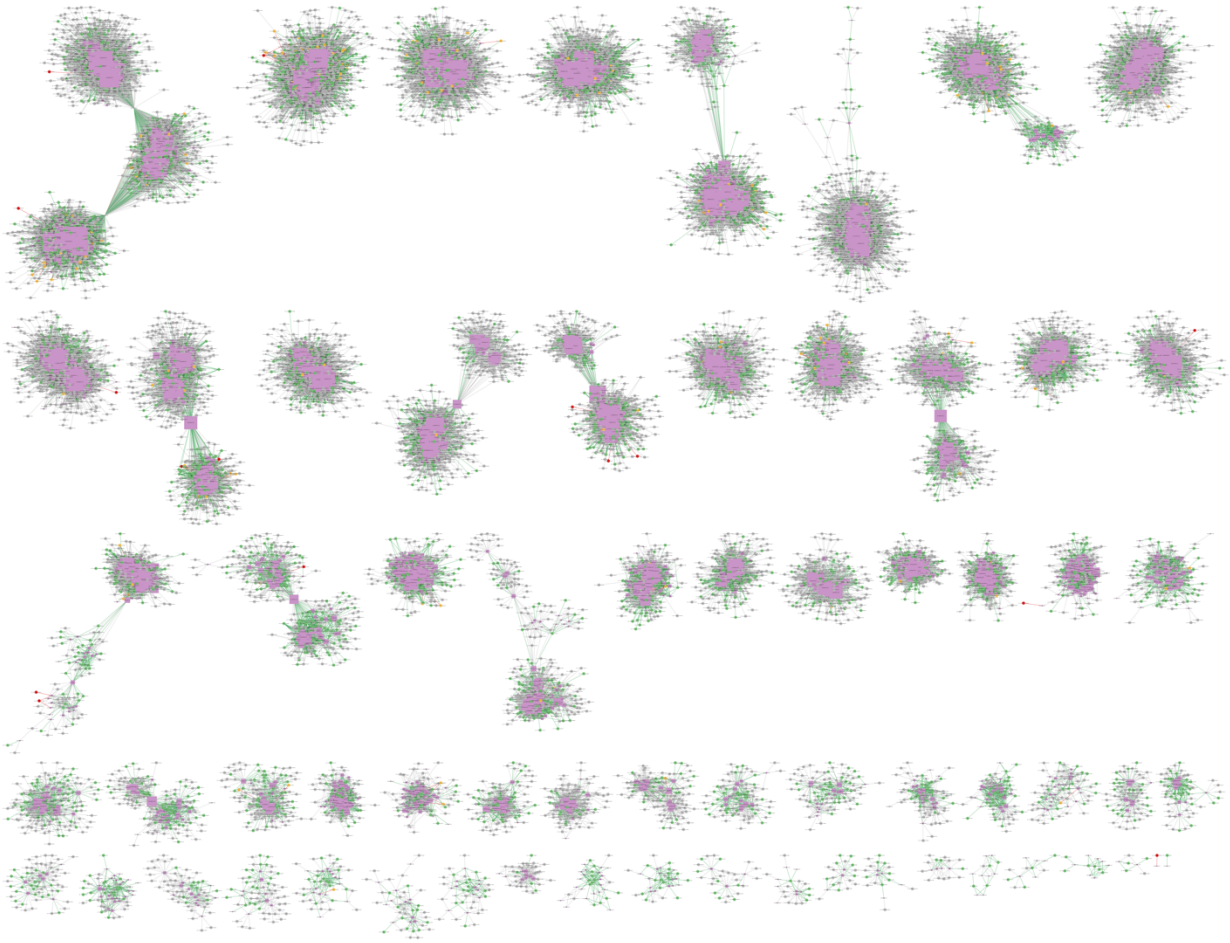
152. Furlong EEM, Levine M. Developmental enhancers and chromosome topology. Science. 2018;361(6409):1341-5.
153. Krefting J, Andrade-Navarro MA, Ibn-Salem J. Evolutionary stability of topologically associating domains is associated with conserved gene regulation. Bmc Biology. 2018;16.
154. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485(7398):376-80.
155. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping (vol 159, pg 1665, 2014). Cell. 2015;162(3):687-8.
156. Schoenfelder S, Fraser P. Long-range enhancer-promoter contacts in gene expression control. Nat Rev Genet. 2019;20(8):437-55.
157. Kryuchkova-Mostacci N, Robinson-Rechavi M. A benchmark of gene expression tissue-specificity metrics. Brief Bioinform. 2017;18(2):205-14.
158. xy-chen16. DeepSilencer: A deep convolutional neural network for the accurate prediction of silencers 2020 [Available from: https://github.com/xy-chen16/DeepSilencer.
159. Duren Z, Chen X, Jiang R, Wang Y, Wong WH. Modeling gene regulation from paired expression and chromatin accessibility data. Proc Natl Acad Sci U S A. 2017;114(25):E4914-E23.
160. Riethoven JJ. Regulatory regions in DNA: promoters, enhancers, silencers, and insulators. Methods Mol Biol. 2010;674:33-42.
161. Dekker J, Belmont AS, Guttman M, Leshyk VO, Lis JT, Lomvardas S, et al. The 4D nucleome project. Nature. 2017;549(7671):219-26.
162. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, et al. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. Cell Syst. 2016;3(1):99-101.
163. Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, et al. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics. 2005;21(5):650-9.

APPENDIX A: Link of supplementary materials

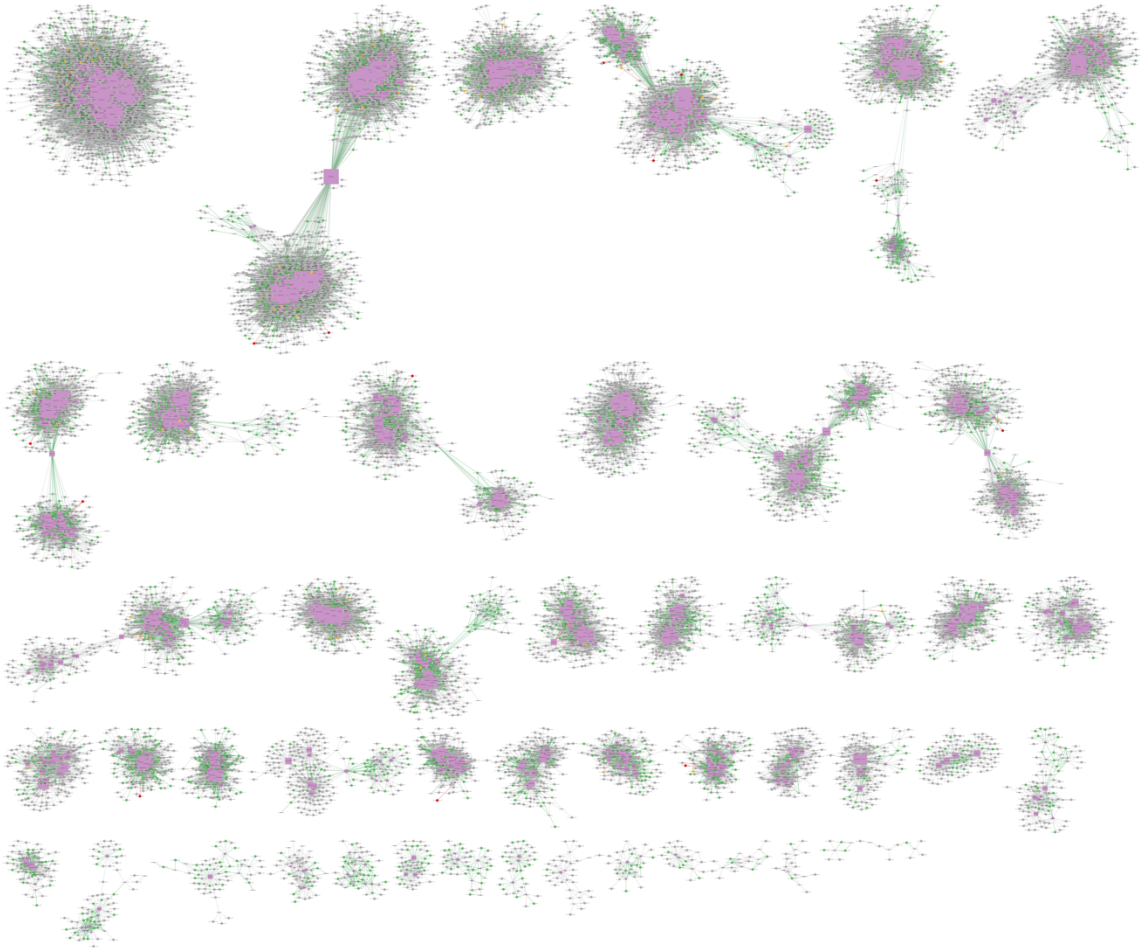
Supplementary materials are available at https://github.com/sisyyuan/CRM_Dissertation.

APPENDIX B: Regulatory Networks of Prediction of Target Genes of Enhancers and Silencers cross different chromosomes

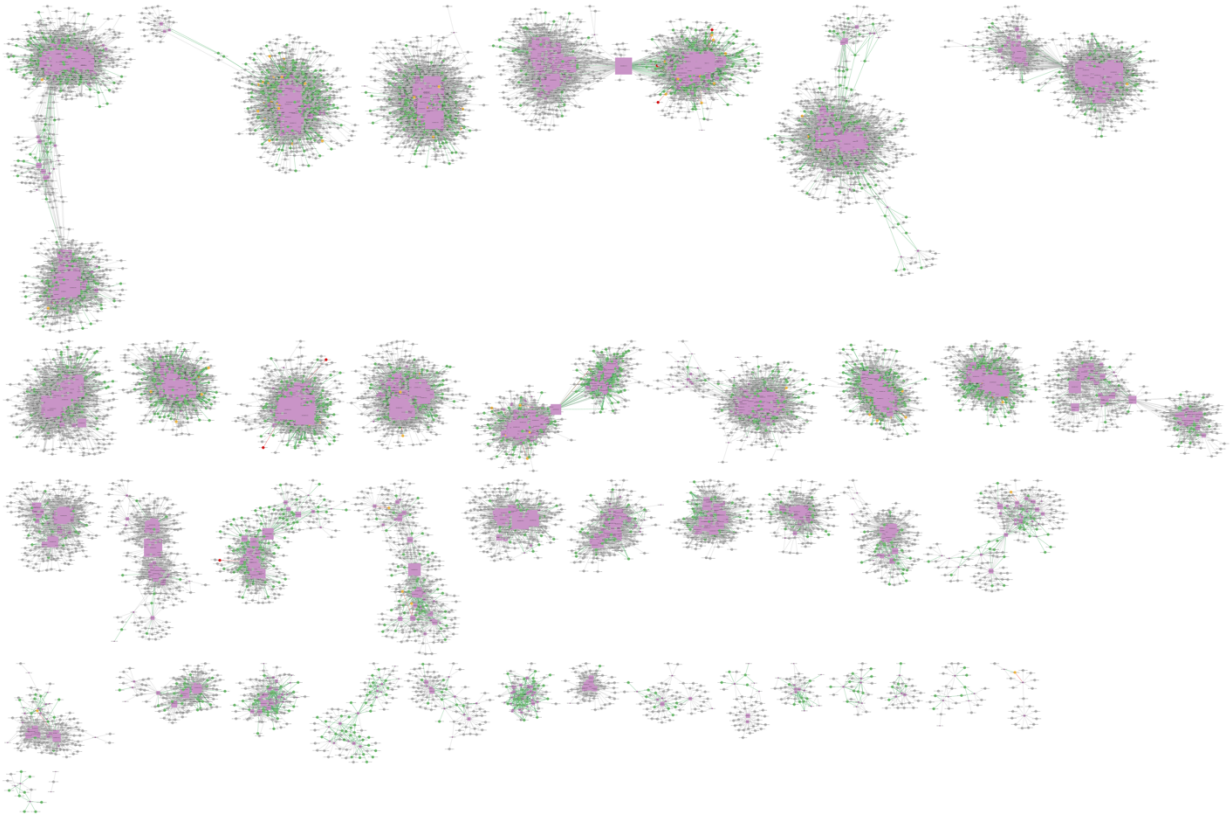
Chr 1



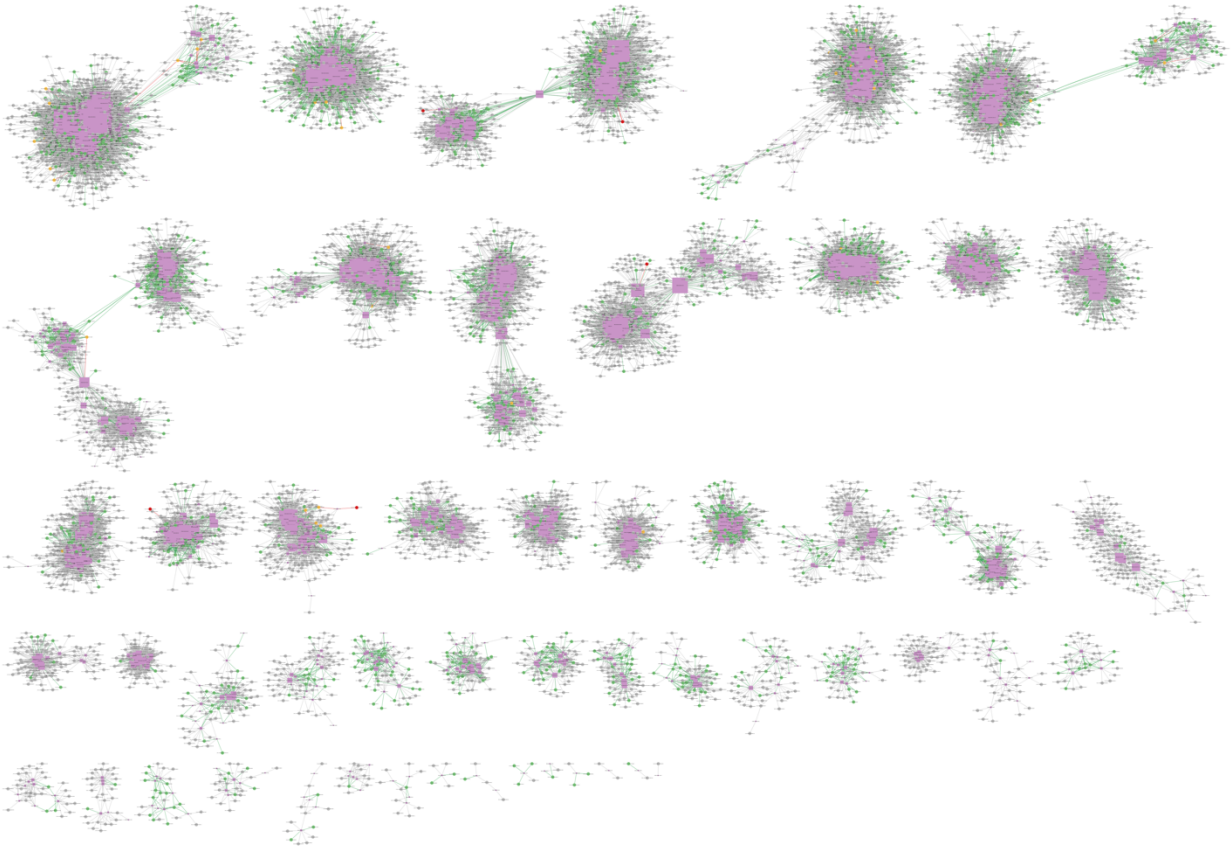
Chr 2



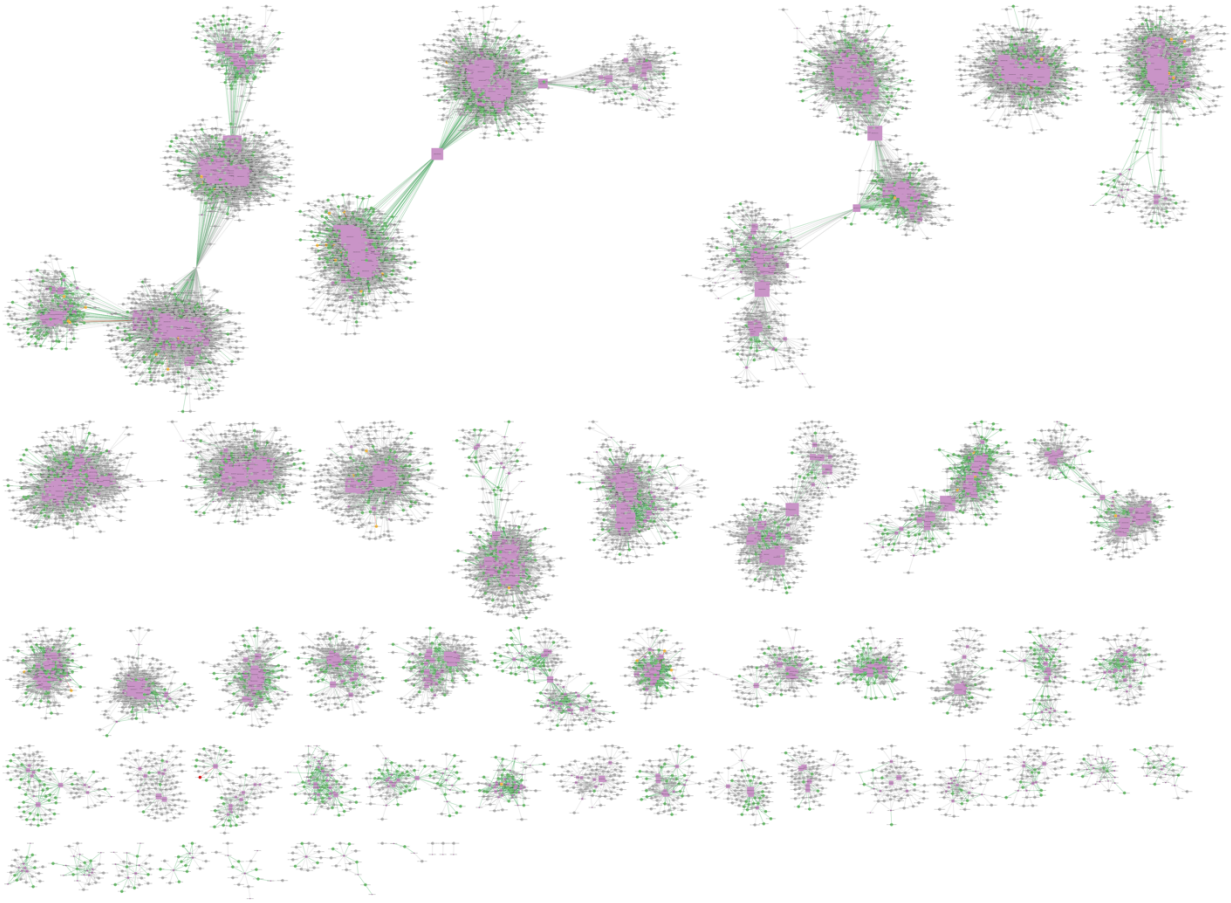
Chr 3



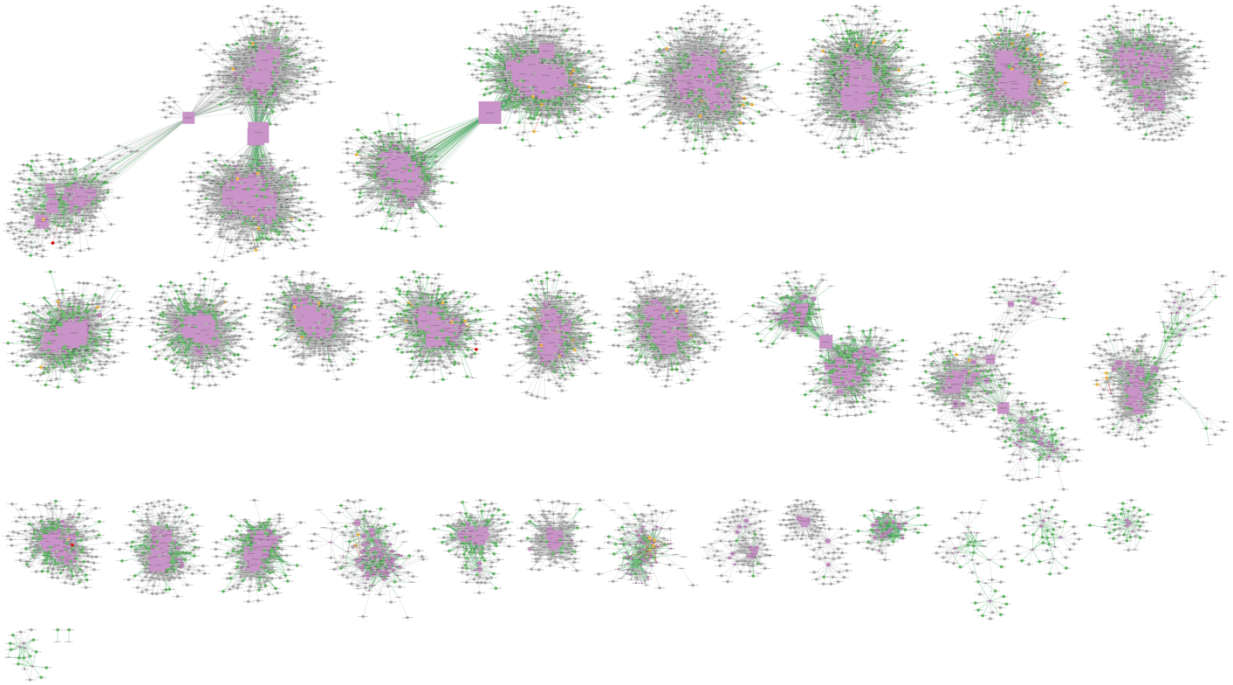
Chr 4



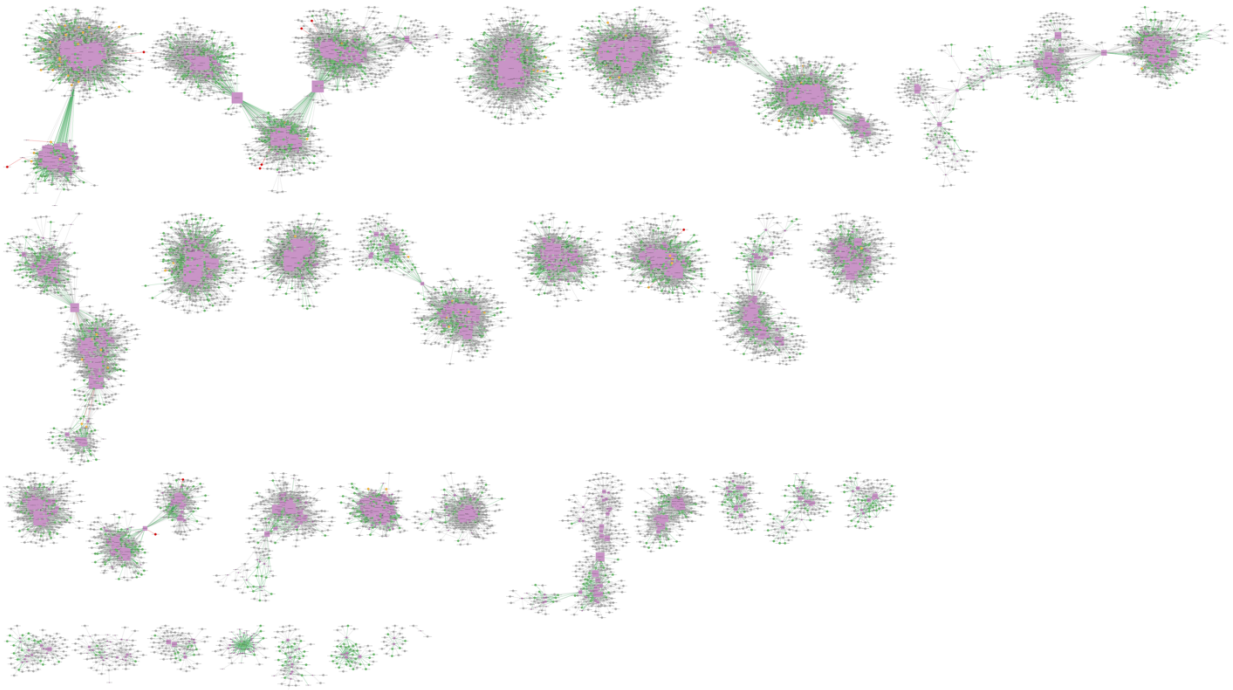
Chr 5



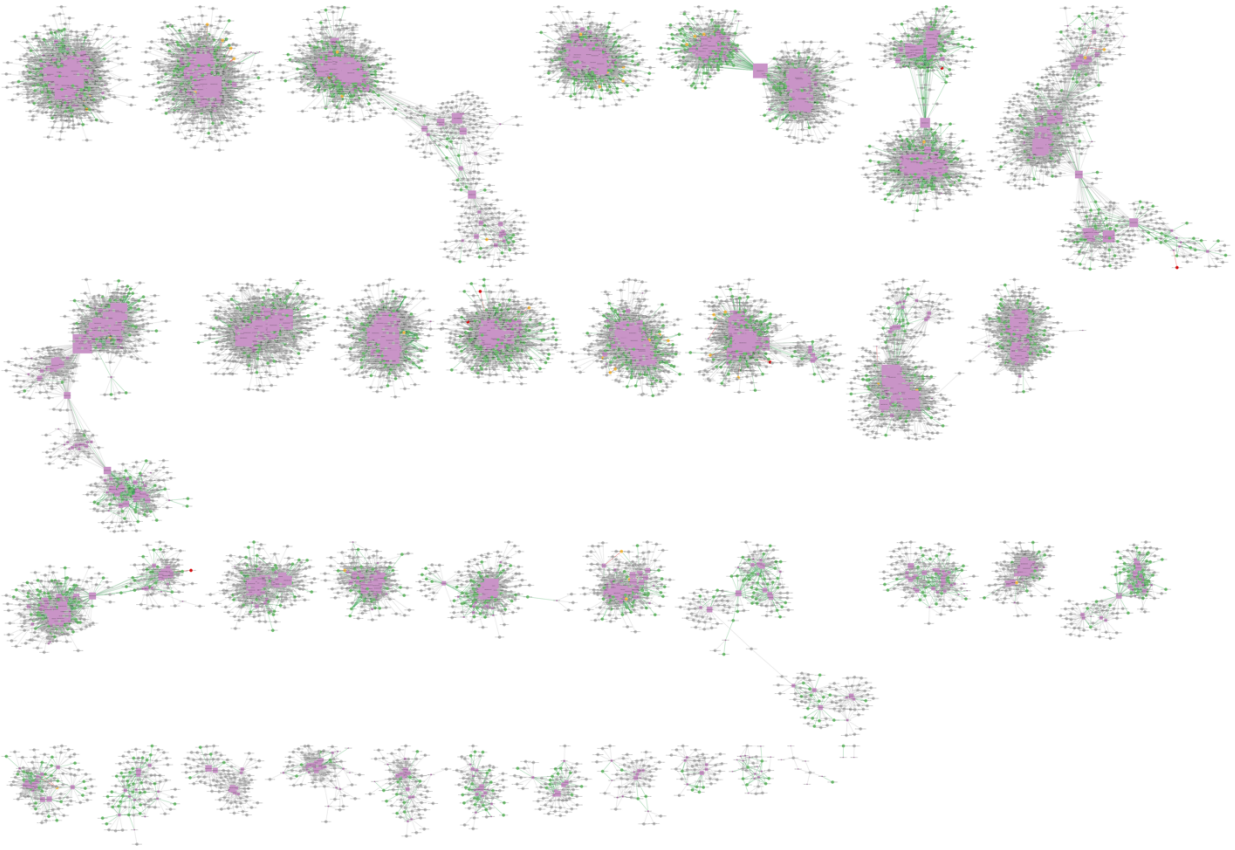
Chr 6



Chr 7

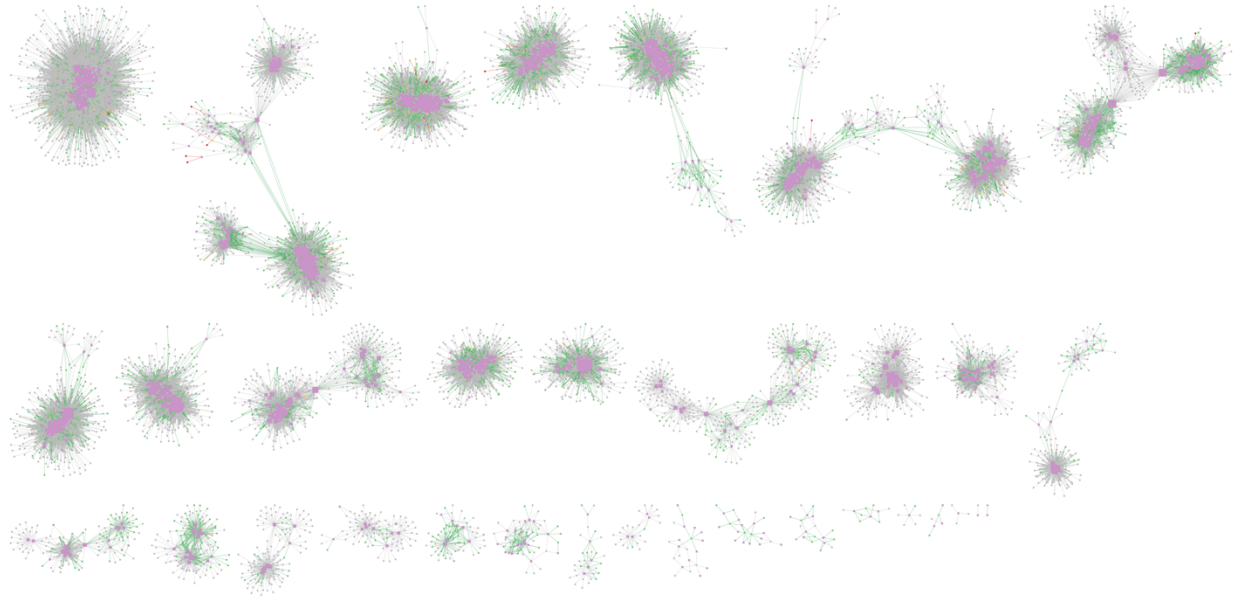


Chr 8

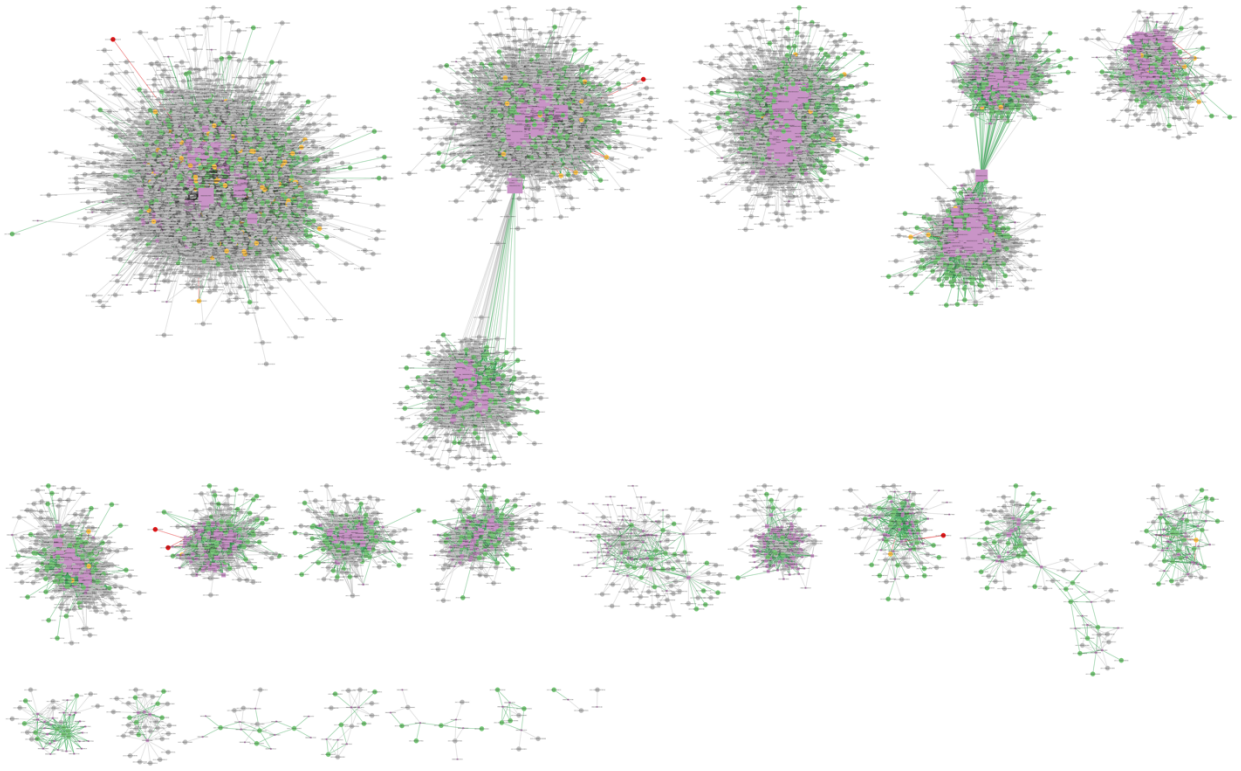


Chr 9

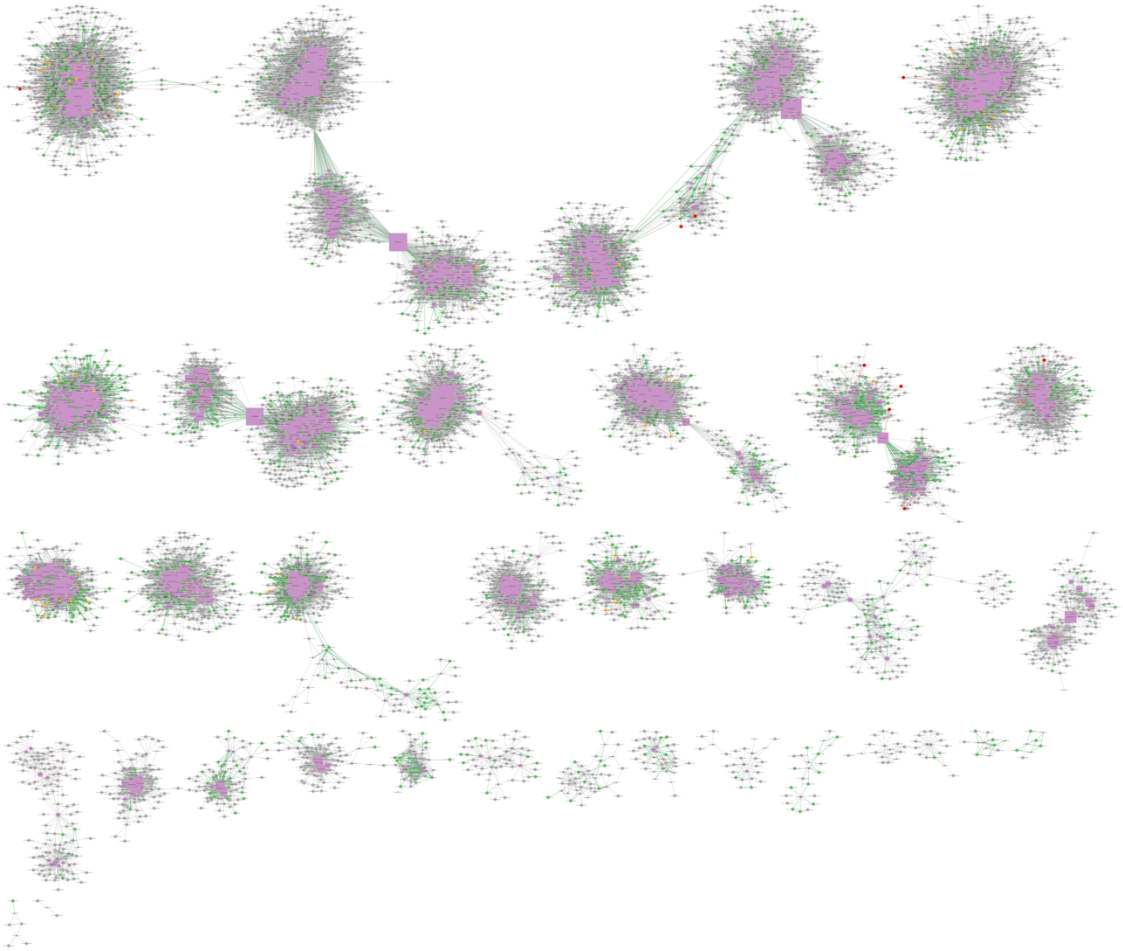
Chr 10



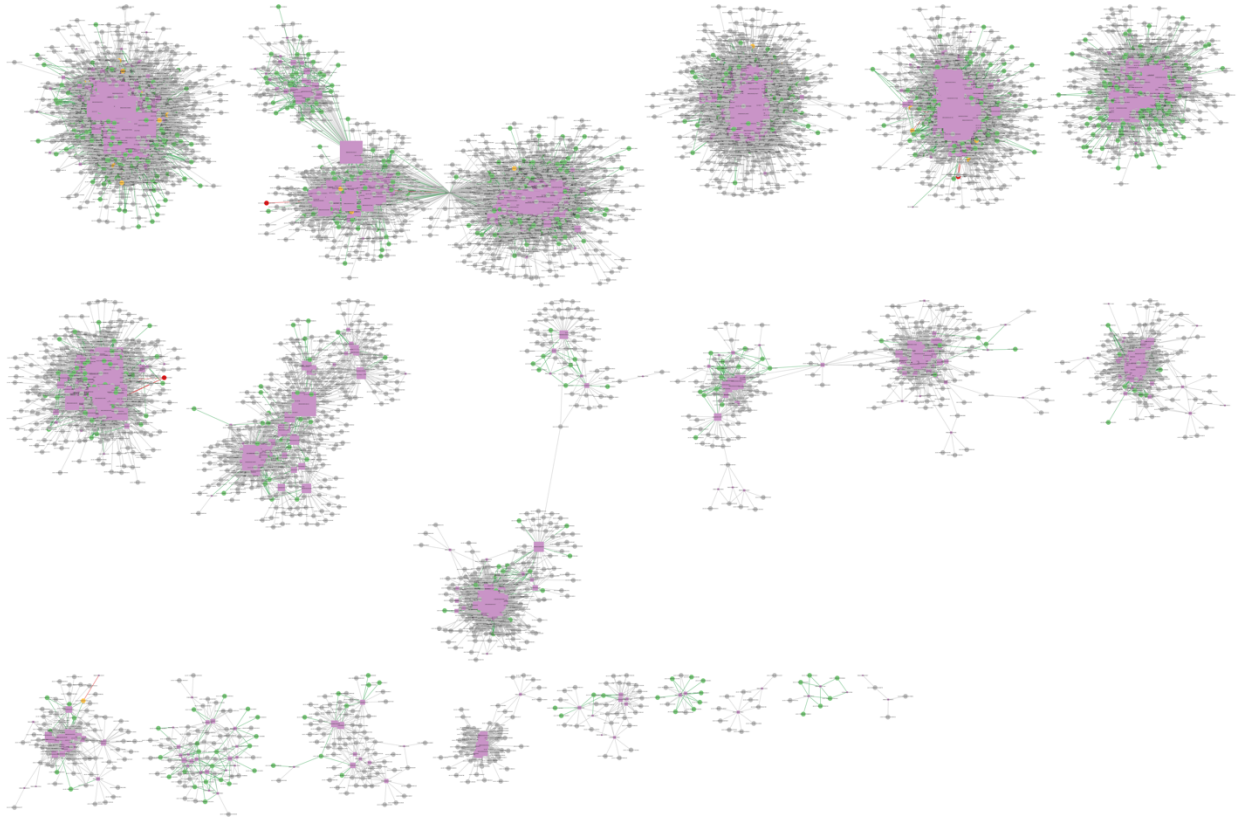
Chr 11



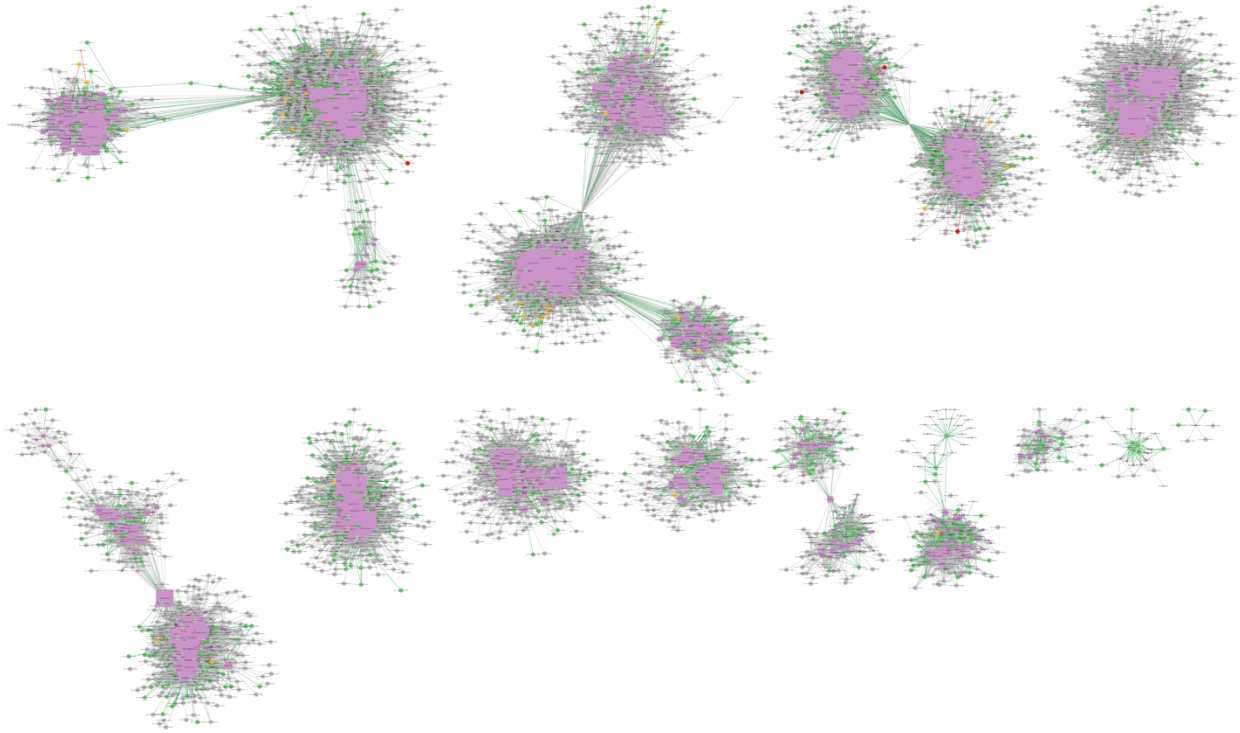
Chr 12



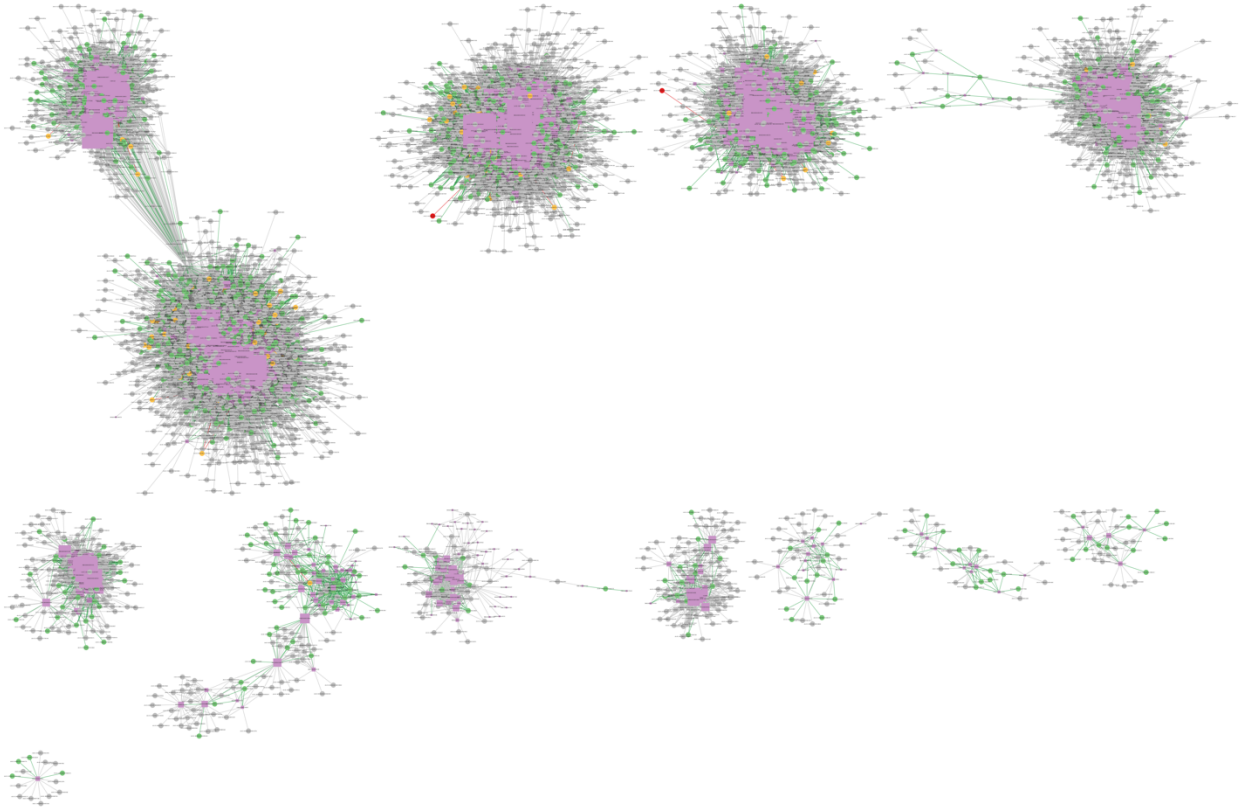
Chr 13



Chr 14

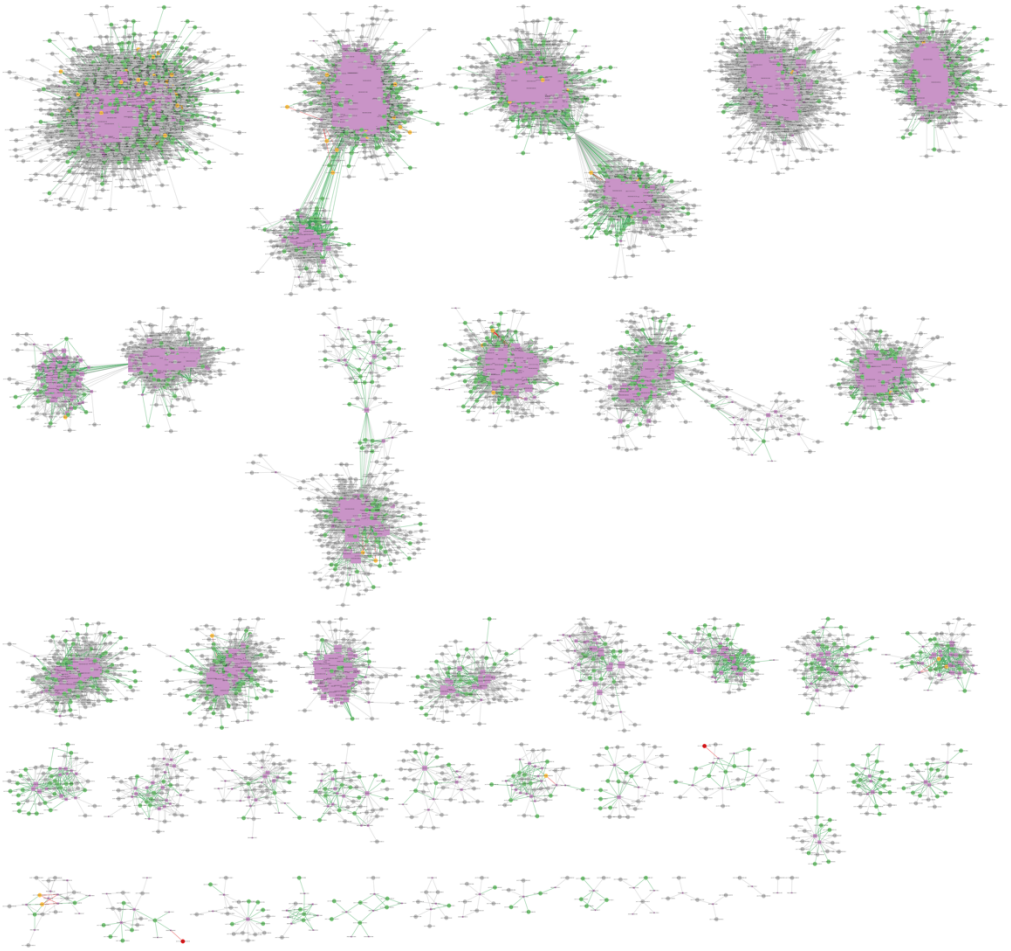


Chr 15

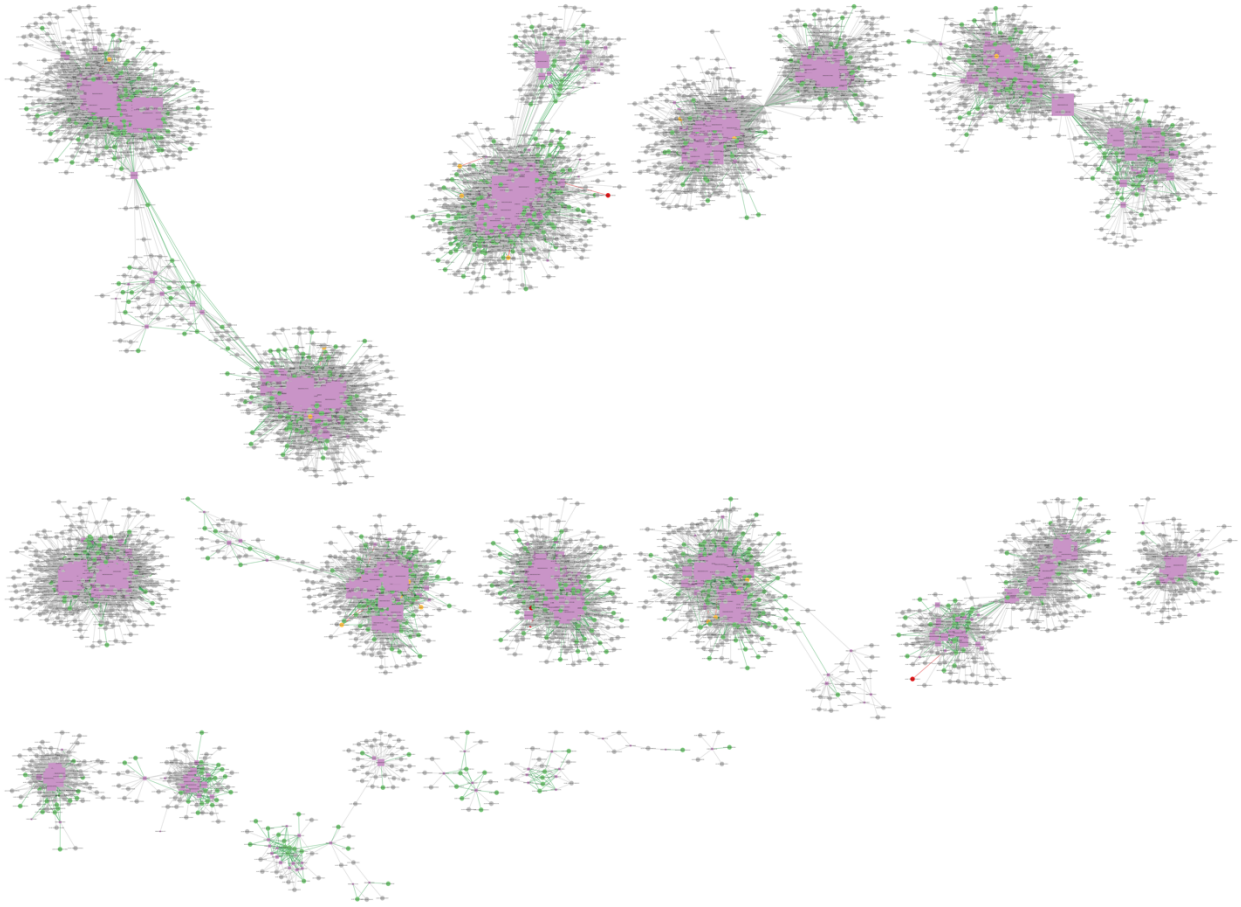


Chr 16

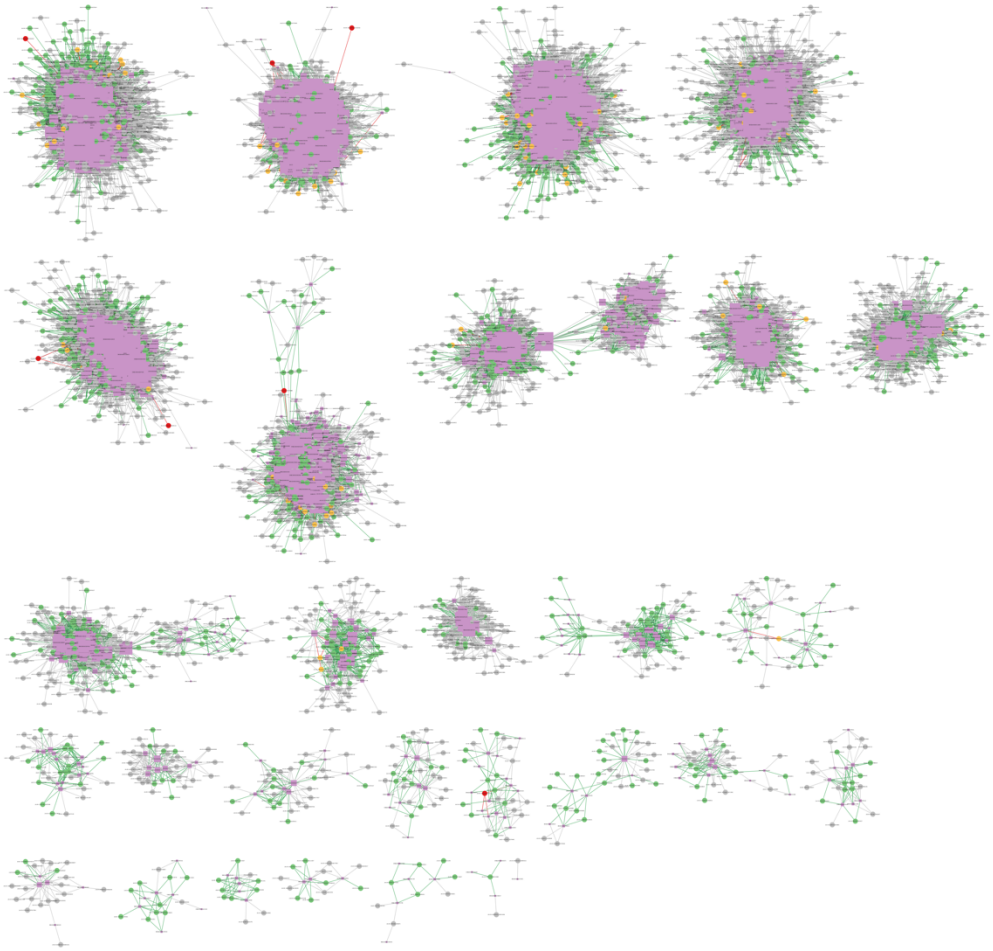
Chr 17



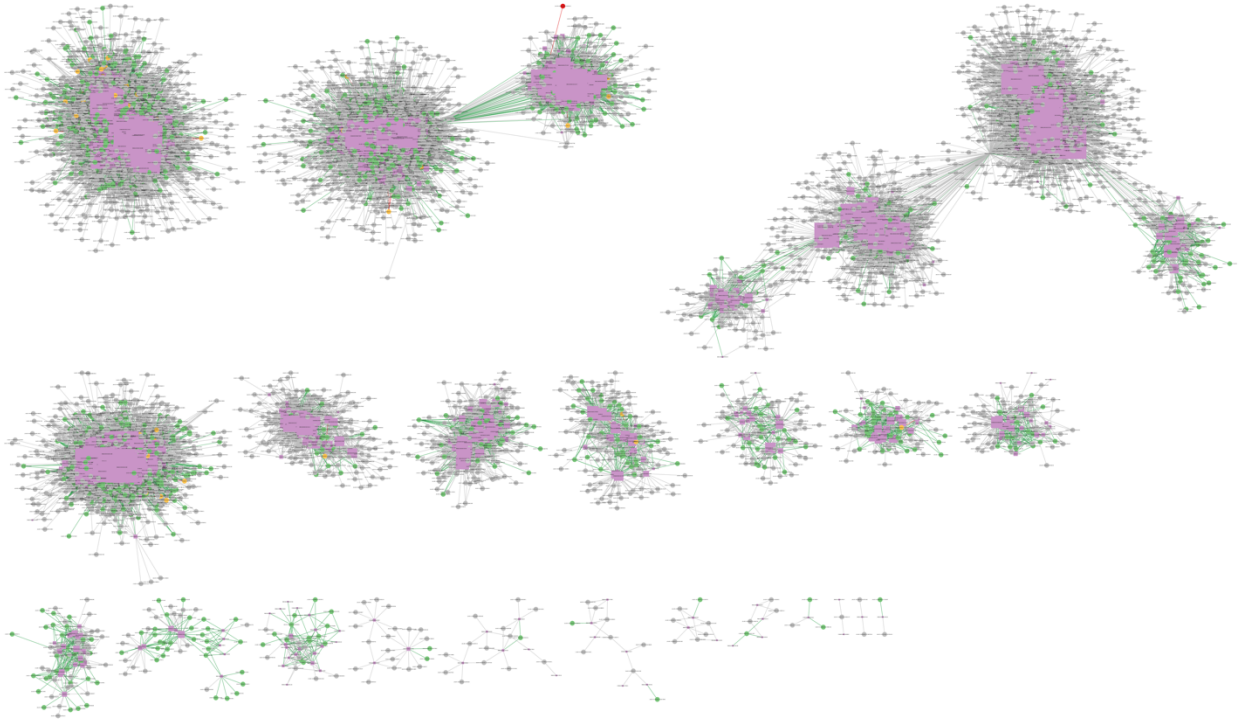
Chr 18



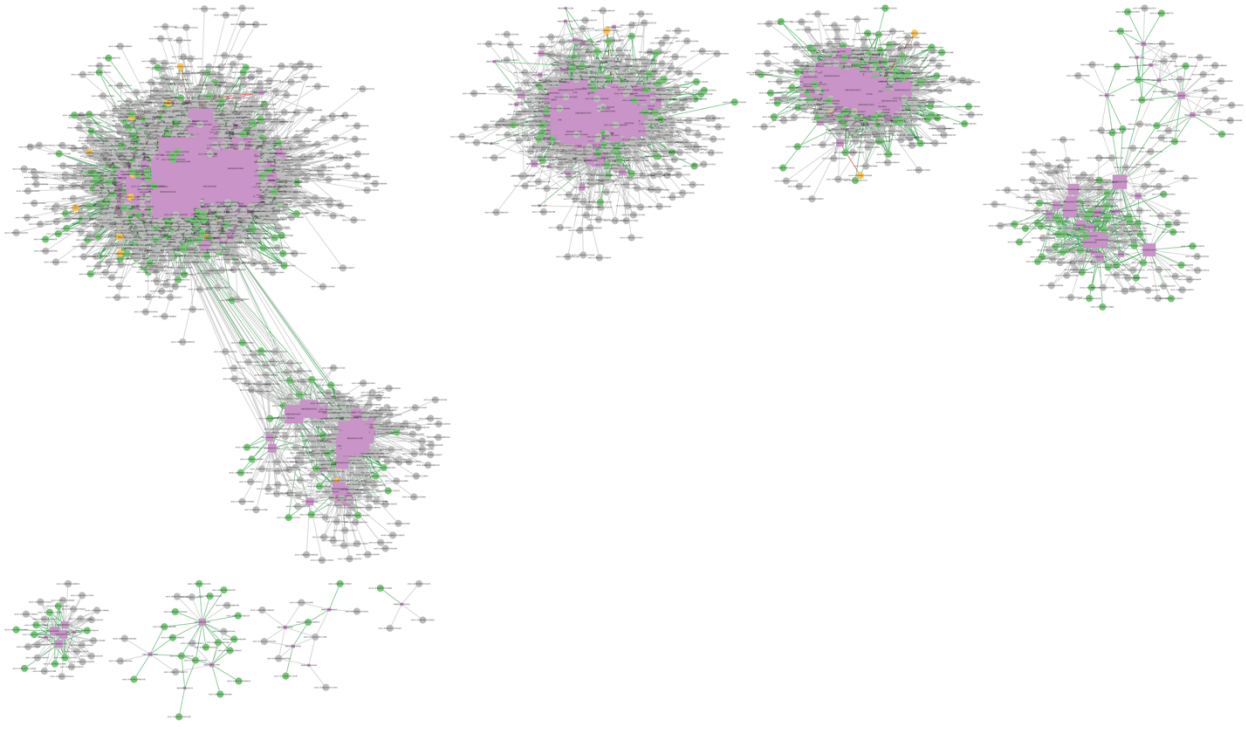
Chr 19



Chr 20

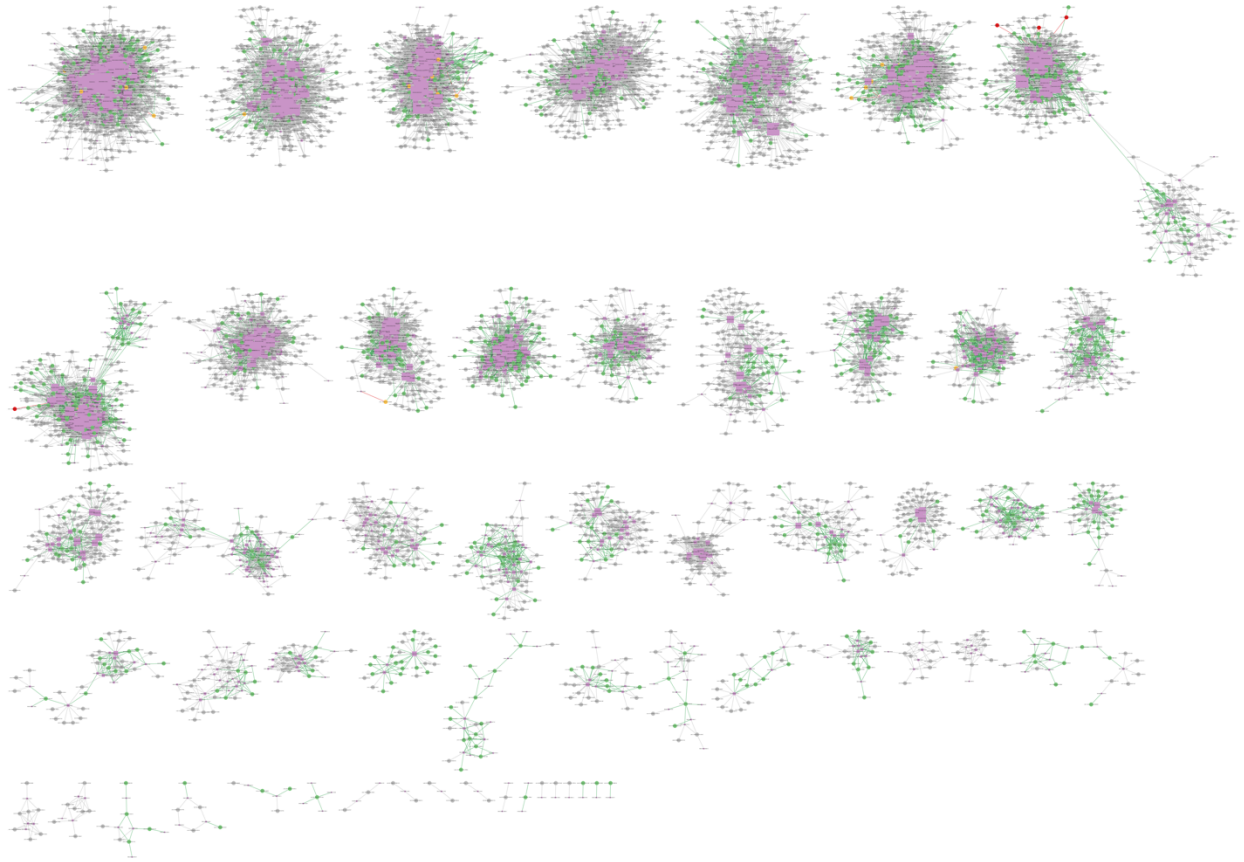


Chr 21

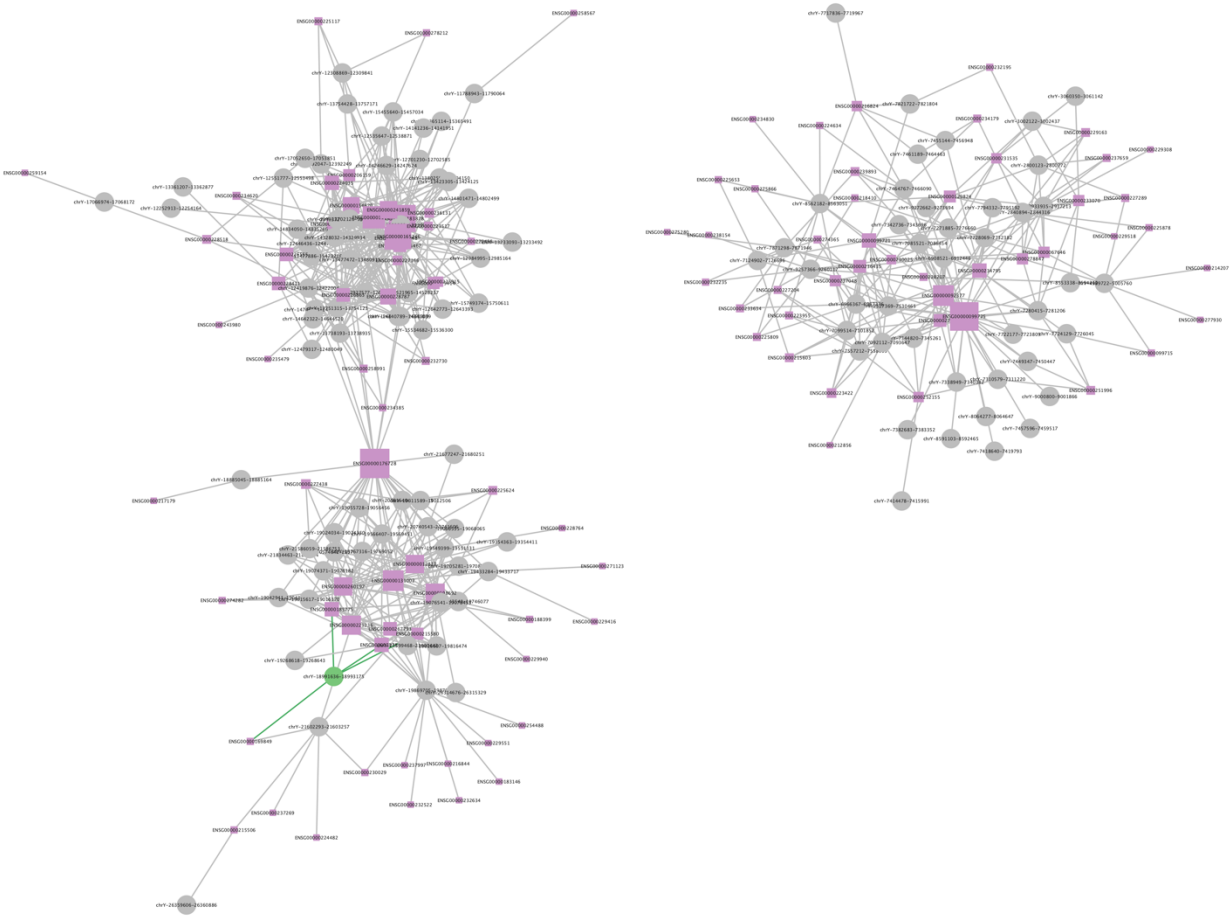


Chr 22

Chr X



Chr Y



Regulation Networks. The sub-static and sub-active *cis*-regulatory networks on each chromosome. The sub-active *cis*-regulatory networks on each chromosome in the K562 cells are embedded in the static *cis*-regulatory networks and can be induced by active exclusive enhancers, active exclusive silencers and active dual CRMs in the cells. The circles represent CRMs, including inactive CRMs (gray), active exclusive enhancers (green), active exclusive silencers (red) and active dual CRMs (yellow) in the K562 cells. The purple squares represent genes, and their size are proportional to the degrees, i.e., the number of their regulating CRMs. The edges represent regulation relationships between CRMs (circles) and genes (squares). The edges linked to active CRMs are colored by the same colors as the type of active CRMs.