COMPUTATIONAL DISCOVERY AND ANNOTATIONS OF CELL-TYPE SPECIFIC LONG-RANGE GENE REGULATION

By

Binbin Huang

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Computational Mathematics, Science and Engineering — Doctor of Philosophy

2021

ABSTRACT

COMPUTATIONAL DISCOVERY AND ANNOTATIONS OF CELL-TYPE SPECIFIC LONG-RANGE GENE REGULATION

By

Binbin Huang

Long-range regulation by distal enhancers plays critical roles in cell-type specific transcriptional programs. Delineation of the underlying mechanisms underlying long- range enhancer regulation will improve our systems-level understandings on the gene regulatory networks and their functional impacts on human diseases. Although there are experimental approaches to infer cell-type specific long-range regulation, they suffer from the problems of low resolution or high false negative rates. Recent technological advances make it possible to have a comprehensive profile of the regulatory activities in multiple layers, bringing us to the multi-omics era. Here, we took use of the booming data resources and integrated them into machine learning models to uncover the resulting effects of long- range regulation, especially in diseases. In the first study about androgen- induced gene regulation in the ovary and its impact on female fertility, we identified a total of 190 annotated significant differentially expressed genes. The H3K27me3 histone modification level change was observed in more than half of the DEGs, highlighting the importance of complex long-range multi-enhancer regulation of androgen receptors regulated genes in the ovarian cells. However, current computational predictions of genome-wide enhancer–promoter interactions are still challenging due to limited accuracy and the lack of knowledge on the molecular mechanisms. Based on recent biological investigations, the protein–protein interactions (PPIs) between transcription factors (TFs) have been found to participate in the regulation of chromatin loops. Therefore, we developed a novel predictive model for cell-type specific enhancer– promoter interactions by leveraging the information of TF PPI signatures. Evaluated by a series of rigorous performance comparisons, the new model achieves

superior performance over other methods. In this chromatin loop prediction model, TF bindings inferred from Chromatin immunoprecipitation followed by high- throughput sequencing (ChIP-seq) make an essential contribution to the instruction to prioritize specific TF PPIs that may mediate cell-type specific long-range regulatory interactions and reveal new mechanistic understandings of enhancer regulation. When processing ChIP-seq data, we detected, on average, 25% of the ChIP-seq reads can be aligned to multiple positions in the reference genome. These reads are discarded by traditional pipeline, which causes a large loss of information. To cope with this waste, we developed a Bayesian model and designed a Gibbs sampling algorithm to properly align these reads. Evidences from a series of biological comparisons indicated a significantly better performance of this model over the competing tool. In summary, our studies took full advantage of the booming data in this multi-omics era, to provide a novel view of the cell-type specific long-range regulation by distal enhancers and its effects on diseases.

ACKNOWLEDGEMENTS

Throughout the writing of this dissertation, I have received a great deal of support and assistance.

I would like to express the appreciation to my advisor and committee chair, Professor Jianrong Wang. He led me into the area of computational biology and helped me to establish a deep understanding of algorithm development.

I also would like to thank my committee members, Professor David N. Arnosti, Professor Ming Yan and Professor Yuehua Cui, for their insightful feedback in my research and professional guidance in personal skill development.

I appreciate the other members in the lab, Wenjie Qi, Jiaxin Yang and Hao Wang for their help in my research.

Thanks to my dear friends, Yuning Hao, Jieqian He, Ningyu Sha, Hongnan Wang and Fei Long for their patience and understandings.

Most importantly, I would like to express my deepest appreciation to my parents, Chuanyin Huang and Shuhua Du, for their mental and financial supports. They always stand behind me and encourage me to keep going. Without them, I would not be able to have any achievements.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

Adamts15	ADAM metallopeptidase with thrombospondin type 1 motif 15
Adamts4	ADAM metallopeptidase with thrombospondin type 1 motif 4
AR	Androgen receptor
ARE	Androgen response element
AUC	Area under the curve
BAM	Binary version of a SAM file
Bmp4	Bone morphogenetic protein 4
bp	Base pair
Capture-C	Promoter capture Hi-C
Casp7	Caspase 7
CCDC86	Coiled-coil domain containing 86
ChIA-PET	Chromatin interaction analysis with paired-end tag sequencing
ChIP-seq	Chromatin immunoprecipitation followed by sequencing
CRISPR	Clustered regularly interspaced short palindromic repeats
CTCF	CCCTC-binding factor
Cyp19a1	Cytochrome P450 family 19 subfamily A member 1
DEG	Differentially expressed gene
DHT	Dihydrotestosterone
DNase	Deoxyribonuclease
Egfr	Epidermal growth factor receptor

ELF1	E74 like ETS transcription factor 1
ENCODE	Encyclopedia of DNA Elements
eQTL	Expression quantitative trait locus
Erbb4	Erb-B2 receptor tyrosine kinase 4
Fshr	Follicle stimulating hormone receptor
FTO	FTO alpha-ketoglutarate dependent dioxygenase
GC	Granulosa cell
GO	Gene ontology
GWAS	Genome-wide association study
H3K27ac	Acetylation of the lysine residue at N-terminal position 27 of the histone H3 protein
H3K27me3	Tri-methylation of lysine 27 on histone H3 protein
H3K4me1	Mono-methylation at the 4^{th} lysine residue of the histone H3 protein
Hi-C	High-throughput chromosome conformation capture
IRX3	Iroquois Homeobox 3
IRX5	Iroquois Homeobox 5
ISCU	Iron-sulfur cluster assembly enzyme
kb	Kilo base pair
L1	LINE 1
L2	LINE 2
Lepr	Leptin Receptor
Lhcgr	Lutropin-choriogonadotropic hormone receptor
LINE	Long interspersed nuclear element

LTR	Long terminal repeat
MIR	Mammalian-wide interspersed repeat
Mmp2	Matrix metallopeptidase 2
OR	Odds ratio
PCOS	Polycystic ovary syndrome
PE	Paired end
PPI	Protein protein interaction
RELB	RELB proto-oncogene, NF-KB subunit
RNA	Ribonucleic acid
RNA Pol II	RNA polymerase II
RNA-seq	RNA sequencing
ROC	Receiver operating characteristic
Runx1	RUNX family transcription factor 1
RUNX3	RUNX Family Transcription Factor 3
SAM	Sequence alignment map
SE	Single end
SINE	Short interspersed nuclear element
Smad3	SMAD family member 3
SNP	Single nucleotide polymorphisms
STAT4	Signal Transducer And Activator Of Transcription 4
TCGA	The Cancer Genome Atlas
TE	Transposable element

TES	Transcription end sites
TF	Transcription factor
TSS	Transcription start sites
YY1	Yin Yang 1

CHAPTER 1

OVERVIEW

Conventionally viewed as male hormone, androgens play a critical role in female fertility. Although androgen receptors (AR) are transcription factors, to date very few direct transcriptional targets of ARs have been identified in the ovary. Using mouse models, this study provides two critical insights about androgen-induced gene regulation in the ovary and its impact on female fertility. First, RNA-sequencing reveals a number of genes that were previously not known to be directly regulated by androgens in the ovary. Second, correlation analysis shows androgens may also influence gene expression by decreasing the tri-methyl mark on lysine 27 of histone3 (H3K27me3), a gene silencing epigenetic mark. ChIP-seq analyses highlight that androgen-induced modulation of H3K27me3 mark within gene bodies, promoters or distal enhancers have a much broader impact on ovarian function than the direct genomic effects of androgens. Among the list of differentially expressed genes (DEGs), more than half of them showed H3K27me3 level change in the distal enhancer regions, indicating the critical roles of long-range regulation by distal enhancers in cell-type specific transcriptional programs.

Computational predictions of genome-wide enhancer–promoter interactions are still challenging due to limited accuracy and the lack of knowledge on the molecular mechanisms. Based on recent biological investigations, protein–protein interactions (PPIs) between transcription factors (TFs) have been found to participate in the regulation of chromatin loops. Therefore, we developed a novel predictive model for cell-type specific enhancer–promoter interactions by leveraging the information of TF PPI signatures. Evaluated by a series of rigorous performance comparisons, the new model achieves superior performance over other methods. The model also identifies specific TF PPIs that may mediate long-range regulatory interactions,

revealing new mechanistic understandings of enhancer regulation. The prioritized TF PPIs are associated with genes in distinct biological pathways, and the predicted enhancer– promoter interactions are strongly enriched with cis-eQTLs. Most interestingly, the model discovers enhancer-mediated trans-regulatory links between TFs and genes, which are significantly enriched with trans-eQTLs. The new predictive model, along with the genome-wide analyses, provides a platform to systematically delineate the complex inter- play among TFs, enhancers and genes in long-range regulation. The novel predictions also lead to mechanistic interpretations of eQTLs to decode the genetic associations with gene expression. In the chromatin loop prediction model, TF bindings inferred from Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) make an essential contribution to the instruction to prioritize specific TF PPIs that may mediate cell-type specific long-range regulatory interactions and revealing new mechanistic understandings of enhancer regulation.

We detected, on average, 25% of the ChIP-seq reads can be aligned to multiple positions in the reference genome. These reads are discarded by traditional pipeline, which causes a large loss of information. To cope with this waste, we designed a Naïve Bayesian model and applied a Gibbs sampling algorithm to properly align these reads. Evidences from a series of biological comparisons indicated a significant better performance of this model over the competitor tool. Transposable elements (TE), as the most dynamic genomic units, have been suggested to contribute regulatory elements on gene expression and rewire the network topology and functions. By applying our model of ChIP-seq read allocation, we identified millions of new TE- regulatory elements. Systematic analysis of the TE-rewired networks reveals different waves of network innovation through evolution, and potential co-evolution between TE-regulatory elements and the flanking co-factors. Genetic variants disrupting these TE-regulatory networks are found to be associated with diverse phenotypes, including cancer. Our integrated analyses on TE-derived network rewiring provides mechanistic insights on the dynamics of network evolution.

CHAPTER 2

ANDROGENS AFFECT OVARIAN GENE EXPRESSION THROUGH LONG-RANGE REGULATION

A modified version of this chapter was previously published (Roy S*, Huang B* et al, 2021): Roy S*, Huang B*, Sinha N, Wang J, & Sen A. (2021) Androgens regulate ovarian gene expression by balancing Ezh2-Jmjd3 mediated H3K27me3 dynamics. PLoS Genet 17(3):e1009483.

2.1 INTRODUCTION

Androgens are traditionally considered as male hormones with well-established roles in male physiology and prostate cancer. However, in the last decade, several genetic models and *in vitro* studies have proven that androgens acting through androgen receptors (AR) are critical for ovarian function and female fertility [1–6]. While excess androgen level leads to polycystic ovary syndrome (PCOS) [7–9], a certain amount of direct androgen actions through the androgen receptor (AR) are essential for normal ovarian function [10]. Thus, it is now believed that with respect to androgen actions in the ovary, balance is key [4]. To date, in addition to the global androgen receptor knockout (ARKO) mouse models [11–13], AR has been knocked out specifically in different cell types along the hypothalamus-pituitary-gonadal (HPG) axis, namely granulosa cells (GCARKO) [14,15], theca cells (TCARKO) [16], oocyte (OoARKO) [15], pituitary (PitARKO) [17] and neurons (NeuroARKO) regulating the HPG axis [18]. All of these ARKO mouse models establish that the granulosa cells (GCs) of the ovary are the primary site of androgen actions in regulating normal follicular development and female fertility; while in hyper-androgenic conditions, neuroendocrine ARs play a major role in the development of PCOS [18]. Moreover, *ex vivo* [5,19], *in vitro* [20–23] and clinical studies [10,24–31] show that androgens are essential for follicle growth while simultaneously preventing follicular atresia. Despite these

studies, how androgens regulate these follicular endpoints is poorly understood.

Androgen actions are mediated by *"nuclear"* transcriptional signals or *"extra-nuclear"* kinase actions [32–34]. Primary AR target genes are those at which AR occupies an androgen response element (ARE) on the promoter of a gene and regulates gene transcription. However, to date, very few ovarian genes have been identified as AR-ARE target genes and, intriguingly, there are no studies on the global impact of androgens on GC gene expression under normal conditions. Here we describe androgen-induced gene expression profiles in mouse GCs and provide molecular insight into the underlying mechanism of how androgens regulate the expression of these genes.

Importantly, this study also shows that androgens can regulate gene expression in an AR-ARE independent fashion, involving membrane-initiated androgen signaling [5,35–37]. Previous study [38] reported that androgens influence gene expression through post- translational histone modifications. H3K27me3 (tri-methyl lysine 27 histone3), which is a gene silencing mark, [39] is a downstream target of androgen actions. Here using ChIP-seq studies with H3K27me3 antibody we identify the ovarian (GC-specific) genes and their enhancer regions that are regulated by androgen-induced modulation of H3K27me3 mark on the enhancers, promoters and gene bodies. This study provides a mechanistic understanding of the global impact of androgens in normal follicular development.

2.2 MATERIALS AND METHODS

2.2.1 Bioinformatics analysis for RNA-seq. Raw data quality was judged based on Illumina's Q score, which represents the error rate at each base, built on a log10 score. Thereafter, sequence reads were trimmed to remove possible adapter sequences and nucleotides with poor quality using Trimmomatic v.0.36 [40]. The trimmed reads were mapped to the reference Mus

musculus GRCm38 genome available on ENSEMBL [41] using the STAR [42] aligner v.2.5.2b. The STAR aligner uses a splice aligner that detects splice junctions and incorporates them to help align the entire read sequences. BAM files were generated as a result of this step. Unique gene hit counts were calculated by using feature Counts from the Subread [43] package v.1.5.2. Only unique reads that fell within exon regions were counted. After extraction of gene hit counts, the gene hit counts table was used for downstream differential expression analysis. Using DESeq2 [44] R package, differentially expressed genes (DEGs) were identified between control (n = 3) and DHT-treatment (n = 3). The heat maps were constructed using rlog transformed values obtained from RNA-seq data followed by z-normalization. The Wald test was used to generate p-values and Benjamini-Hochberg test for adjusted p-value. Genes with adjusted p-values ≤ 0.05 and absolute log2 fold changes > 1 were called as differentially expressed genes for each comparison.

2.2.2 Bioinformatics analysis for ChIP-seq. FastQC [45](version 0.11.7) was used for quality controls of ChIP-seq datasets, including 3 replicates for control and 3 replicates for DHT treated GCs. The ChIP-seq reads were mapped to the reference mouse genome (mm9) using Bowtie2 [46] (version 2.3.4), and only uniquely mapped reads were used for subsequent analyses. MACS2 [47] (version 2.1.1) was then applied to identify signal peaks of H3K27me3 signals, in broad peak calling mode since H3K27me3 distribution along the genome is more diffusive and demonstrate broader peaks. PCR duplicates were removed in peak calling. Peaks with q-value ≤ 0.05 were used as the significant signal peaks. Peaks in control samples and treatment samples are center-aligned separately with +/-10kb flanking regions. The flanking regions of each peak were further divided into 100bp bins and the normalized H3K27me3 ChIP-seq read densities for each bin were plotted. The genomic locations of significant H3K27me3 peaks in controls and DHT treated samples were then compared to gene annotations to identify genes overlapping with H3K27me3

peaks in gene bodies. Promoters (+/- 1kb from transcription start sites) were also compared with H3K27me3 peaks. Gene promoters overlapping with control-specific H3K27me3 peaks were center-aligned with +/- 1kb flanking regions, and the average peak densities per 50bp bins were plotted. Similar peak densities were plotted for gene promoters overlapping with treatment-specific H3K27me3 peaks. For differentially expressed genes that also contain H3K27me3 peaks in gene bodies, their gene bodies were divided into 10 bins, starting from transcription start sites (TSS) to transcription end sites (TES), where every bin represented 10% of the specific gene's body. The number of H3K27me3 peaks in each bin were then calculated for every gene, based on H3K27me3 signals from control samples and treatment samples.

A combined chromatin interaction dataset, including Hi-C and Capture-C, were used to identify candidate distal enhancers that can interact with promoters of differentially expressed genes. Hi-C datasets include: GSE81503 [48], GSE82144 [49], GSE119171 [50], GSE121753 [51], and GSE63525 [52]. Capture-C dataset includes GSE81503. For the long-range chromatin interactions profiled in these dataset, they were first compared with promoters of differentially expressed genes and a subset of chromatin interactions was then identified, if one of the interacting anchors overlapped with promoters. For each interaction in this subset, the other interacting anchor of the interaction that did not overlap with promoters were identified as candidate distal enhancers that may regulate the gene. The identified enhancer regions from different chromatin interaction datasets were then combined together. To purify false positives of enhancers, we calculated the Pearson correlations between the gene's expression levels and the enhancer's H3K27me3 signal levels across the 6 samples. The lengths of enhancers and the library sizes of different samples were normalized for H3K27me3 signals. Therefore, for each enhancer-gene pair, an activity correlation was calculated. Only enhancer-gene pairs with correlations < -/+0.4 were considered as true regulatory pairs and the corresponding enhancers were used for subsequent analysis. For each differentially expressed gene with distal interacting

enhancers, we generated the distribution of the numbers of enhancers co-regulating the same genes. We also calculated the distribution of the distances between gene promoters and distal interacting enhancers.

Motif enrichment analysis was applied on the identified enhancers using MEME [53] (version 5.0.4). Enhancers were classified into two groups: the first group of enhancers interact with promoters of up-regulated genes and the second group of enhancers interact with promoters of down-regulated genes. For each group of enhancers, the top five enriched sequence motifs were identified using MEME. TomTom [54] (version 5.0.4) was then applied on the top-enriched motifs to identify the corresponding transcription factors, based on transcription factor motif annotations in "HOCOMOCOv11_full_MOUSE_mono_meme_format.meme" from the MEME suite [55]. The matched transcription factors with E-value < 0.1 were then identified as the candidate factors associated with epigenetic changes in distal enhancers.

2.3 RESULTS

2.3.1 Effect of androgen on granulosa cell (GC) transcriptome

Global effects of androgens in GCs were elucidated by RNA-seq analysis in primary mouse GC cultures treated with media (control) or DHT. DESeq2 analysis identified a total of 190 annotated significant differentially expressed ENSEMBL genes (DEGs) (Table A.1). Out of these genes, 129 were upregulated and 61 were downregulated genes. The global transcriptional change across the two groups compared (control vs DHT) is represented by a volcano plot in Fig 2.1A and hierarchical clustering of all the significant DEGs in control vs DHT treated GCs are shown in Fig 2.1B. *In silico* analysis revealed that all of the DEGs have at least one or more ARE sequences in the promoter and/or distal (within 5Kb) region thereby suggesting that most of these genes are regulated directly by AR, rather than secondary effects of hormone exposure.

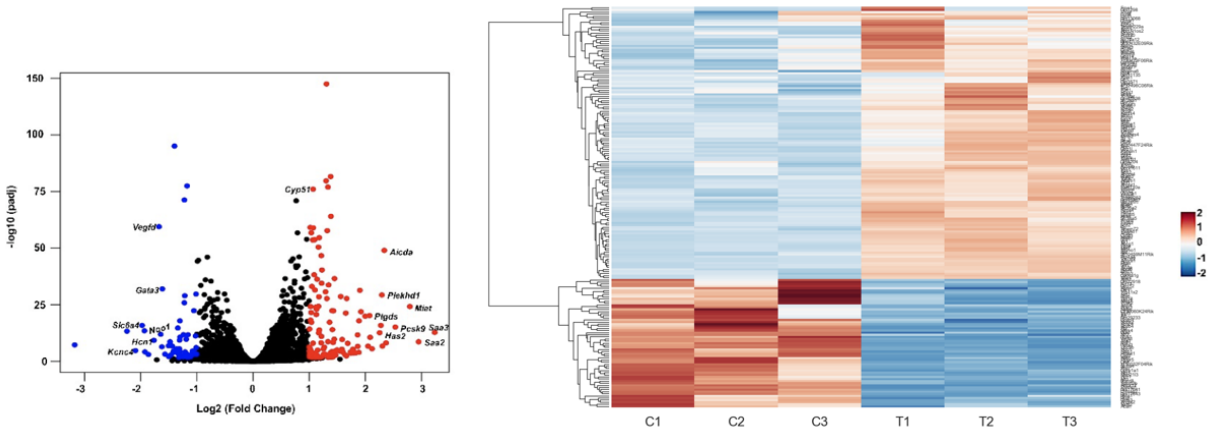


Figure 1.1. Androgen-induced transcriptome analysis in primary mouse granulosa cells (GC). A: Volcano plot: Representing the global transcriptional change across the groups compared. Each data point in the scatter plot represents a gene. Genes with an adjusted P≤0.05 and a log2 fold change≥1 are indicated by red dots and represent up-regulated genes. Genes with an adjusted P≤0.05 and a log2 fold change≤-1 are indicated by blue dots and represent downregulated genes. B: Heatmap of differentially expressed genes sorted by adjusted p-value by plotting their log2 transformed expression values in samples.

2.3.2 Androgens significantly modulate H3K27me3 mark on gene promoters and enhancers

Previous study [38] showed that androgens decrease the H3K27me3 mark in GCs. To evaluate the impact of androgens on genome-wide distribution of H3K27me3 landscape, we performed ChIP with the H3K27me3 antibody followed by high-throughput sequencing in control vs. DHT treated GCs

Total H3K27me3 peaks modulated by DHT. The analysis of sequencing reads revealed 16,345 H3K27me3 peaks in control and 3975 H3K27me3 peaks in DHT treated samples: a 75% reduction in peaks in the DHT treated samples. Fig 2.2A shows a heat map of genome wide H3K27me3 peaks in GCs from control and treatment groups. Each row in the heatmap corresponds to a H3K27me3 signal peak identified from either controls or DHT treated samples. The normalized ChIP-seq read densities around the peaks are shown with the summits of the peaks in the middle, along with +/-10kb flanking regions. Of 16,345 H3K27me3 peaks in control

samples, 5513 peaks were within gene bodies while 153 peaks were in promoter regions. In contrast, there were only 1389 H3K37me3 peaks in the gene body and 54 peaks in the promoter region in the DHT treated samples. Fig A.2 represents the number of H3K27me3 peaks overlapping different genomic annotations in the control and treatment (DHT) group.

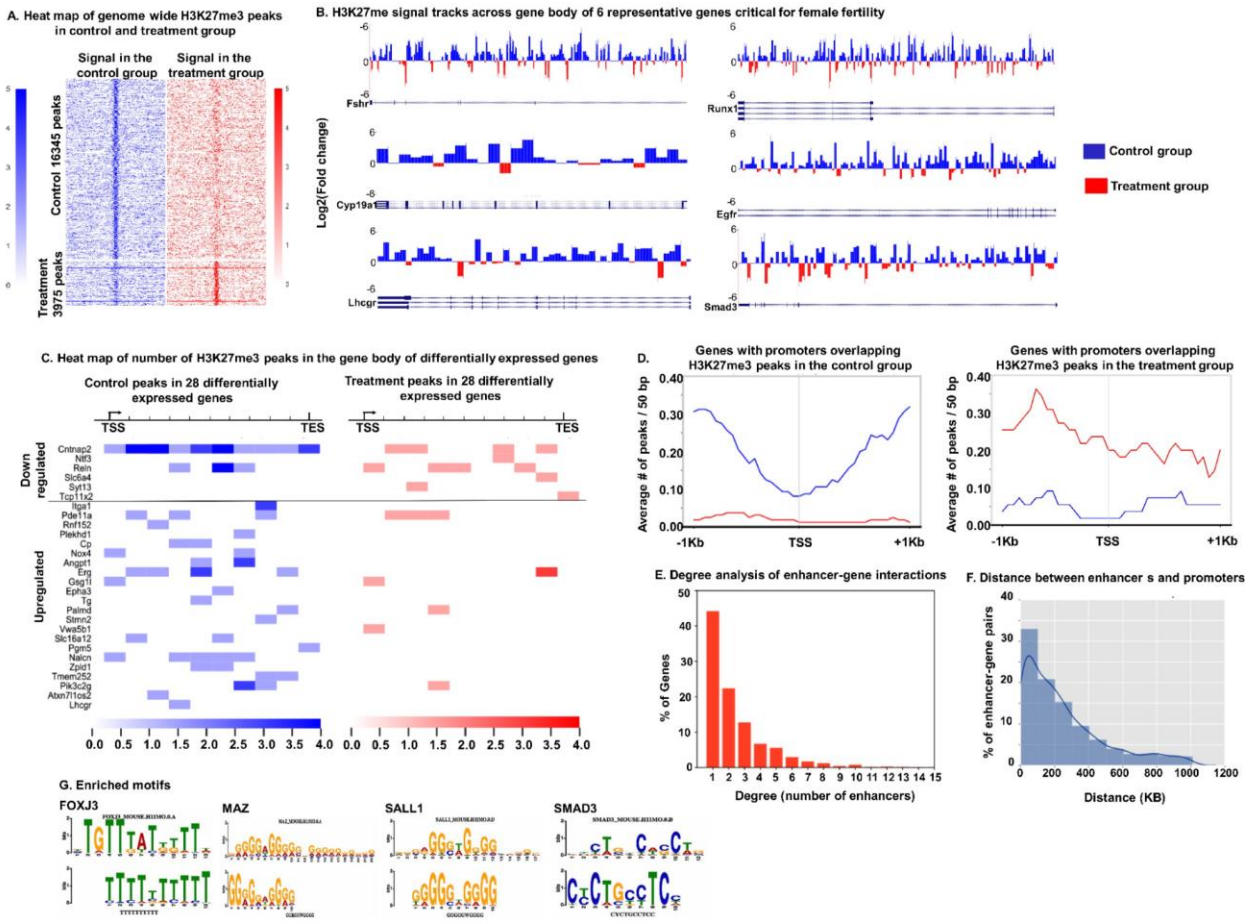


Figure 2.2. Genes associated with androgen-induced decrease in H3K27me3 mark in the gene body and overlapping with promoters and distal enhancers involving long-range regulation of gene expression. A: Heat maps showing the read density change along the peak regions for the 16345 control peaks and the 3975 treatment peaks. B: H3K27me3 signals in control and DHT treatment groups along the gene body of Fshr- follicle stimulating hormone receptor, Cyp19a1- aromatase, Lhcgr- luteinizing hormone/Choriogonadotropin receptor, Runx1- runt-related transcription factor 1, Egfr–epidermal growth factor receptor and Smad3- Mothers against decapentaplegic homolog 3. Log2 fold change was calculated as log2{(H3K27me3 control signal)/ (H3K27me3 treatment signal)}. Regions along the gene body with higher signals in the control group are represented as positive value- blue peaks and regions where signal was higher in the treatment group are shown as negative value-red peaks.

C: Heatmap showing H3K27me3 peak counts in control and DHT-treated samples for 28 differentially expressed genes identified by comparing genes containing H3K27me3 peaks in the gene body with DEGs from the RNA-seq data. D: Average number of H3K27me3 peaks per 50bp overlapping with the promoter regions of 160 genes in the control samples and 55 genes in the DHT-treated samples. E: Degree analysis of enhancer-gene interactions for genes associated with significant decrease in H3K27me3 peaks in their enhancer regions with respect to DHT treatment. F: Distance between the promoters of genes and their corresponding enhancers that have decreased H3K27me3 signal by DHT treatment. G: Enriched transcription factor binding motifs in distal enhancers of genes associated with decrease in H3K27me3 signal with DHT treatment. E-values were < 0.1 and indicate the probabilities of observing the enrichment from random control DNA sequences. For each transcription factor, the upper motif logo corresponds to the consensus motif based on HOCOMOCO database and the lower motif logo corresponds to the observed sequence motifs that are enriched in linked distal enhancers.

DHT-induced modulation of H3K27me3 peaks in gene bodies. To further analyze the influence of androgen-modulated epigenetic dynamics on genes, we examined genes with gene bodies overlapping with H3K27me3 peaks. We identified 3144 genes in control and 1146 genes in DHT treated GCs with H3K27me3 peak signal across the gene body. Comparison of these two gene sets revealed that 2462 genes exclusively had H3K27me3 peak signal across the gene body in the control but not in the DHT-treated samples. Fig 2.2B demonstrates log2-fold change of H3K27me3 peaks overlapping with gene bodies of six representative genes (*Fshr*, *Cyp19a1*, *Lhcgr*, *Runx1*, *Egfr* and *Smad3*) that are known to play critical roles in ovarian function. For each gene, the H3K27me3 mark in control and treatment signals along the gene body was calculated by dividing each gene into 1000bp windows. The number of reads falling under each 1000bp window were considered the H3K27me3 signal in that window and log2 fold change of H3K27me3 signals along the gene body was calculated. Results show that DHT-treatment significantly lowers the H3K27me3 signal in all of the genes.

Subsequently, we compared the list of genes containing H3K27me3 peaks in their gene bodies with the list of DEGs from the RNA-seq data. We found 28 genes (22-upregulated and 6-downregulated genes) that were both differentially expressed and overlapped with H3K27me3 peaks, suggesting that the condition-specific epigenetic landscape of H3K27me3 may be related

to the transcriptional variation of these genes. Fig 2.2C is a heat map representing the number of H3K27me3 peaks in the gene body (TSS to TES) for the 28 DEGs. Most of the upregulated genes had significantly lower levels of H3K27me3 marks while the downregulated genes had higher H3K27me3 marks in DHT-treated GCs than in controls. This shows that in addition to AR-ARE, the expression of these genes may also be regulated by androgen-induced H3K27me3 modulation.

DHT-induced modulation of H3K27me3 peaks in the promoter region. Further analysis revealed that there were 160 genes in control and only 55 genes in the DHT treated samples with H3K27me3 peaks located specifically in the gene promoter regions. Fig 2.2D (Left panel) represents the average number of H3K27me3 peaks per 50bp in the promoter region (TSS +/-1KB) for the 160 genes in the control group and corresponding average peaks for the same genes in the treatment group. Similarly, Fig 2.2D (right panel) represents the average number of H3K27me3 peaks per 50bp in the promoter region (TSS+/-1KB) for the 55 genes in the treatment group and corresponding average peaks for the same genes in the control group. The list of all the genes with promoters overlapping with H3K27me3 peaks in the control and treatment groups is provided in the supplemental data. Considering the complex regulatory activities in promoter regions, these androgen-induced differential H3K27me3 peaks located in the promoters may play pivotal roles in regulating the transcriptional levels of the corresponding genes.

DHT-induced modulation of H3K27me3 peaks in the enhancer region. Since large portions of the mouse genome are non-coding regions with numerous enhancers widely spread [56], we further extended our analysis to non-coding regions and focused on distal enhancers that have long-range chromatin interactions with promoters and may play a crucial role in controlling the expression of genes. While enhancers act through the binding of transcription factors just like promoters, their locations greatly vary from the transcription start site (TSS) of the gene they regulate. Moreover, while a single enhancer can influence the expression of multiple genes, a

single gene can be regulated by multiple enhancers. Thus, we determined enhancer-gene pairs for the 2462 genes in which the H3K27me3 signal peaks were significantly decreased by DHT treatment (genes with exclusive H3K27me3 peaks in the control group). For these genes, the chromatin interaction data including Hi-C and Capture-C were used to find potential enhancers. For each gene, the H3K27me3 in the gene body and H3K27me3 level in each of its potential enhancers were calculated. We found 1380 genes where DHT treatment lowered the H3K27me3 signal. For enhancer–gene pair, the correlation between the gene body H3K27me3 level and enhancer H3K27me3 level was calculated and only positively correlated enhancer–gene pair (Pearson correlation > 0.4) were selected. Results show 3447 enhancer–gene pairs. Next, we determined the number of enhancers that regulate the same gene (Fig 2.2E). Results show that 45% of these genes are regulated by only 1 enhancer region while 21% of the genes are regulated by 2 enhancers. Furthermore, we calculated the distance between the promoters and their corresponding enhancers that have decreased H3K27me3 signal by DHT treatment. Fig 2.2F shows the distance analysis of the enhancer–gene interactions for the genes that show DHT-induced decrease in H3K27me3 levels. 33% of the enhancer–gene pairs that show decreased H3K27me3 with DHT treatment have 0 to 100KB distance between the gene and the enhancer. These analyses show that androgens may modulate gene expression by reducing the H3K27me3 mark in the promoter region of genes or in distal enhancers.

Moreover, comparing the chromatin contact maps (ChIP-seq data) with the androgen-induced DEGS revealed 186 enhancers whose H3K27me3 levels were significantly negatively correlated (Pearson correlation < -0.4) with the expression of 99 DEGs, out of which 66 were upregulated and 33 were downregulated genes across samples of controls and DHT treated GCs. This highlights the importance of complex long-range multi-enhancer regulation of AR regulated genes in the ovarian GCs.

Motif analysis of the enhancers. Given that H3K27me3 is a gene repressive mark, it is

likely that the androgen-induced decrease of H3K27me3 allows specific transcription factors to bind to these enhancer regions. We analyzed the enhancer regions linked with H3K27me3 peaks for motif enrichment using MEME-ChIP as described in bioinformatics analysis for ChIP-seq. Four transcription factors (TFs), FOXJ3 (forkhead box j3; *p*-value 1.44e-04 and *q*-value 1.49e-01), MAZ (MYC associated zinc finger protein; *p*-value 6.85e-08 and *q*-value 7.05e-05), SALL1 (spalt like transcription factor 1; *p*-value 2.20e-04 and *q*-value 9.90e-02) and SMAD3 (*p*-value 3.20e-05 and *q*-value 3.32e-02) with E-value < 0.1 were identified as candidate factors associated with epigenetic changes in distal enhancers (Fig 2.2G).

2.4 DISCUSSION

In the ovary, androgens are not merely a substrate for estrogen synthesis, but direct androgen actions through the ARs are critical for normal follicular development and female fertility [2,4]. However, there is a dearth of knowledge about the genes and biological pathways regulated by androgens, which is a significant limitation towards understanding how androgens regulate follicular growth and function. This study for the first time provides three critical insights about androgen actions in the ovary.

First, we have identified a large number of genes and biological processes that were not formerly known to be regulated directly by androgens in GCs. Results show that genes like *Bmp4* [57], *Lhcgr* [58], *Adamts4* [59], *Ptgds4* and *Mmp2* [60–62], that are critical for follicular function are AR-induced genes. Previously studies have reported that androgens primarily maintain normal follicular development by regulating pre-antral to antral follicle transition by increasing FSH receptor levels and prevent follicular atresia. However, our gene expression data now clearly show that androgens have a much far-reaching impact on follicular function. Moreover, androgen treatment resulted in downregulation of 61 genes. Interestingly, comparison of the RNA-seq and ChIP-seq dataset revealed that out of all the downregulated genes, only 6 genes had higher

H3K27me3 level. This suggests that androgen-induced downregulation of genes may be a secondary effect of androgen treatment. For example, a previous study has shown [38] that androgens induce the expression of *miR-101* that in turn downregulates the expression of *Ezh2*. Another example is androgen-induced expression of *miR-125b* that decreases the expression of pro-apoptotic proteins [5].

Second, we have performed ChIP-seq analysis in GCs to determine the DHT-induced changes in H3K27me3, which is a gene silencing mark. Control of gene expression is exerted at a number of levels, one of which is the accessibility of genes and their controlling elements to the transcription machinery. Accessibility is dictated broadly by the degree of chromatin compaction, which is influenced in part by post-translational histone modifications. Our results highlight an important concept: in GCs, in addition to the genomic actions of androgens through the "classical" AR-ARE binding, androgen-induced decrease in H3K27me3 mark is another avenue through which androgens can regulate gene expression. The fact that we identified only 190 DEGs (from RNA-seq study) that are directly regulated by androgens in contrast to 2462 genes (from ChIP-seq) that specifically had lower H3K27me3 mark in DHT- treated GCs, clearly shows that androgen-induced modulation of H3K27me3 mark has a much broader impact than the direct effects of androgens on GC function. On the basis of our present study, we propose that in GCs, androgens prime the promoter and/or enhancer regions of genes by lowering the H3K27me3 mark, that enables other transcription factors to induce the expression of these genes. In fact, we have reported previously [38] that *Runx1*, a gene critical for ovulation, is one such downstream target and androgens remove the H3K27me3-repressive mark from the *Runx*1 promoter. This enables the hCG-induced transcription machinery to access the *Runx1* promoter region leading to increased expression of *Runx1*. Intriguingly, previous studies have reported that androgen treatment increases the expression of genes like *Fshr* [6,65–68] and *Cyp19a1* [1,63,64] that are critical for follicular development. However, there was no evidence that these genes are direct

targets of AR-ARE mediated actions. Prior to this study, it was not known how androgens regulated the expression of these critical ovarian genes to promote ovarian function and female fertility, in general. We now show that androgen treatment significantly lowers the H3K27me3 mark in the gene body of *Fshr* and *Cyp19a1* which provides a mechanistic explanation of how androgens, independent of AR-ARE interaction, through H3K27me3 modulation may influence the expression of these genes. Moreover, some of the genes with lower H3K27me3 mark on the gene body/enhancer regions following DHT treatment, like *Cyp19a1* [69], *Adamts15* [70] *Casp7* [71], *Erbb4* [72] and *Lepr* [73] have been reported to be elevated and/or associated with PCOS.

In summary, given the role of androgens in female fertility and women's health in general, results of this study provide a global perception of androgen effects in follicular function and insights into the androgen-induced molecular mechanisms responsible for normal ovarian physiology as well as for disease conditions like PCOS.

CHAPTER 3

PREDICT LONG-RANGE ENHANCER REGULATION BASED ON PROTEIN-PROTEIN INTERACTIONS BETWEEN TRANSCRIPTION FACTORS

A modified version of this chapter was previously published (Wang H*, Huang B* et al, 2021): Wang H*, Huang B*, Wang J. (2021) Predict long-range enhancer regulation based on protein-protein interactions between transcription factors. Nucleic Acids Res 49(18): 10347-10368.

3.1 INTRODUCTION

Cell-type specific transcriptional regulation plays important roles in differentiation and development [74–86]. In addition to proximal regulatory elements, e.g. promoters, which are located around transcriptional start sites (TSS) of genes, distal enhancers provide complex and precise controls on gene expression through long-range regulation [87,88]. Based on recent genome-wide enhancer annotations from ENCODE and Roadmap Epigenomics projects [89,90], hundreds of thousands of putative enhancers across the whole human genome have been identified, especially in non-coding regions, highlighting the biological impacts of enhancer regulation. Although a series of computational algorithms have been developed to predict the genomic locations of cell-type specific enhancers [91,92], it remains challenging to identify the specific target genes regulated by enhancers in different cell-types or tissues. Unlike promoters, enhancers are usually located far away from their target genes along the genome [93] and the nearest genes may not be regulated by a proximal enhancer [94]. In three-dimensional (3D) space, an enhancer and its target genes are placed close to each other through long-range chromatin interactions, *i.e.* enhancer-promoter interactions [95].

The discoveries of tissue-specific long-range enhancer regulation have the potential to

enable novel insights in a wide range of different biological studies. As one of the canonical examples, long-range regulation by distal enhancers play pivotal roles in controlling the tissue and condition-specific expression of the mouse β-globin (*Hbb*) gene expression [74,78,79]. As another well-known example, the expression of the *Shh* gene in mouse limb bud is precisely regulated by a distal enhancer located 850 kb away, which is critical for the proper limb development [80–82,96]. In addition to normal tissue development, the annotation of long-range enhancer regulation has also facilitated the interpretation of genetic variants underlying complex diseases. A non-coding genetic variant associated with obesity is located in an intron of the *FTO* gene but regulates the *IRX3* and *IRX5* genes that are located >400 kb away [75,83,97]. Similar examples of long-range interactions linking disease-associated genetic variants to distal genes have also been found in studies of autoimmune diseases [76,77,84–86].

Given the functional importance of long-range enhancer regulation, experimental techniques have been developed to identify chromatin interactions linking distal enhancers to promoters of their target genes. Based on the pioneering chromosome conformation capture (3C) technology [98], along with its derivatives of 4C and 5C [99,100], the genome-wide version, i.e. Hi-C [101], has been applied to several human cell-types and tissues [89,102,103]. Furthermore, the promoter-enriched genome conformation assay, Capture Hi-C [104], improves the resolution and cell-type specificity of the identified chromatin interactions for gene promoters [105]. On the other hand, the method of chromatin interaction analysis with paired-end-tag sequencing (ChIA-PET) [106] was developed to capture long-range chromatin interactions associated with a protein of interest, such as a specific transcription factor (TF), with high-resolution and cell-type specificity [107]. These cutting-edge technologies have generated large-scale chromatin contact maps for a number of cell-types or tissues in the human genome and other model species [89,102,103,107].

Although experimental techniques have substantially expanded the catalog of annotations for long-range chromatin interactions, there are several limitations that hinder in-depth analysis on

cell-type specific enhancer–promoter interactions. First, the resolution of interacting genomic anchors profiled by Hi-C and Capture Hi-C is relatively low (~5–10 kb genomic fragments) [102,104], which makes it difficult to pinpoint the specific enhancers involved in long-range regulation. Second, while Capture Hi-C and ChIA-PET experiments can discover cell-type or tissue-specific enhancer regulation, data generated by Hi-C experiments have been found to be largely invariant across different cell-types or tissues [108]. Third, the background noise levels of Hi-C and Capture Hi-C datasets are high, leading to many false positive discoveries [109]. Fourth, due to the dependency on specific protein antibodies, such as CTCF or RNA Pol II [107], each ChIA-PET experiment can only profile a subset of long-range interactions, resulting in large numbers of false negative interactions that are not identified [110].

Because of these limitations, computational models are needed to predict cell-type specific long-range enhancer regulation, based on integration of multi-omics signatures, e.g. genomics, transcriptomics, and epigenomics. Large-scale multi-omics data resources collected by the ENCODE and Roadmap Epigenomics projects contain the multi-view information of gene regulation [89], including gene expression, transcription factor binding and histone modifications. They can help to overcome the limitations of experimental techniques because they are cell-type or tissue specific [111], provide high-resolution signal landscape along the genome [112,113], have high signal-to-noise ratio [113], and cover the genomic binding sites for diverse transcription factors [89]. The existing computational models of long-range enhancer–promoter interaction prediction can be grouped into two classes. For the first class, i.e. supervised algorithms, 3D chromatin interactions profiled by experimental techniques are used as labels for enhancer–promoter pairs. The commonly used features include: (i) cell-type specific gene expression based on RNA-seq data; (ii) enhancer activity based on specific epigenetic signals, such as H3K4me1, H3K27ac or DNase hypersensitivity; (iii) genomic separation distance between enhancers and gene promoters and (iv) correlations between gene expression and enhancer activity. Supervised

methods incorporating some or all of these features include RIPPLE [114], FOCS [115], EAGLE [116] and JEME [117]. As one of the most recently developed supervised methods, JEME [117] employs a combined approach of regression and random forest to predict long-range regulatory links between enhancers and genes. But it requires multi-omics datasets from a large panel of diverse cell-types and tissues as inputs, which is usually not available for users. The other two top-performing methods are IM-PET [118] and TargetFinder [119]. These two algorithms not only integrate the features described above but also leverage additional features of transcription factor binding in promoters, enhancers, or genomic windows between enhancers and promoters. With respect to machine learning techniques, IM-PET employs a random forest model, and TargetFinder implements a boosting tree approach. For the second class, i.e. unsupervised algorithms, every enhancer–promoter pair is assigned with a score and then ranked based on the scores. Top-ranking enhancer– promoter pairs are predicted to interact with each other. The scores are generally based on genomic separation distance and co-activity patterns, e.g. correlations, between enhancers and genes [120–122]. Based on a systematic performance evaluation analysis [123], supervised methods overall demonstrate better performance than unsupervised methods, but many of the supervised methods suffer from over-fitting issues due to high model complexity [123] or excessively high-dimensional features that are often shared across training and testing sets [124]. Furthermore, existing methods provide limited mechanistic insights on how specific long-range chromatin interactions are established to link distal enhancers with promoters of target genes [125].

Interestingly, as shown by recent experimental studies [75,126–131], in addition to the binding of individual TFs on enhancers or promoters, the protein–protein interactions (PPIs) between TFs have been found to participate in the process of long-range chromatin interaction formation and thus, mediate distal enhancer to the proximity of target gene promoters (Fig 3.1A–D). For example, the PPI between the enhancer-binding and promoter-binding YY1s (i.e. YY1

dimerization) has been found to mediate enhancer–promoter contacts [132]. The ChIA-PET data from mESCs suggests that the YY1–YY1 interactions largely participate in the connections between active enhancers and gene promoters [132]. In a chromatin structure engineering study, based on a CRISPR-dCas9 system, two proteins (PYL1 and ABL1) are fused to dCas9 and are guided to bind on different genomic locations [133]. Remarkably, the PYL1–ABL1 dimerization can establish novel long-range chromatin interactions, highlighting the mechanistic importance of PPIs in orchestrating chromatin loops. In addition, a couple of genome-wide analyses have also found that specific groups of transcription factors are enriched in cell-type specific long-range chromatin interactions [134–136]. Within each group, some TF members can interact with each other and form protein complexes. As a representative example, a group of CTCF, RAD21, SMC3 and ZNF143 is found to be enriched in chromatin interactions [134], consistent with the chromatin loop extrusion model that CTCF and cohesin can interact with each other and regulate chromatin loops [137,138].

These observations strongly support the mechanistic hypothesis that specific TF PPIs, except intratypic dimerizations where TFs can only co-bind locally to DNA instead of across long-range distances, may mediate long-range enhancer regulation. Therefore, incorporation of TF PPIs as a new set of features into a machine learning model is expected to improve the accuracy of long-range enhancer–promoter interaction predictions. Moreover, the prioritized TF PPIs from the predictive model can further indicate the important transcription factors that facilitate long-range enhancer regulation, leading to novel understandings of enhancer biology. However, unlike basic enrichment analysis of candidate TF–TF pairs that are over-represented in enhancer–promoter interactions [134–136], building a predictive model based on TF PPI features is computationally challenging. First, the number of candidate TF PPIs is large (~200 000). By filtering the features using cell-type specific TF expression, there are still large amounts of potential TF PPI features. Take the human GM12878 cell-line as an example, by only considering

TFs that are expressed [90], the number of PPIs between expressed TFs is ~1900. The

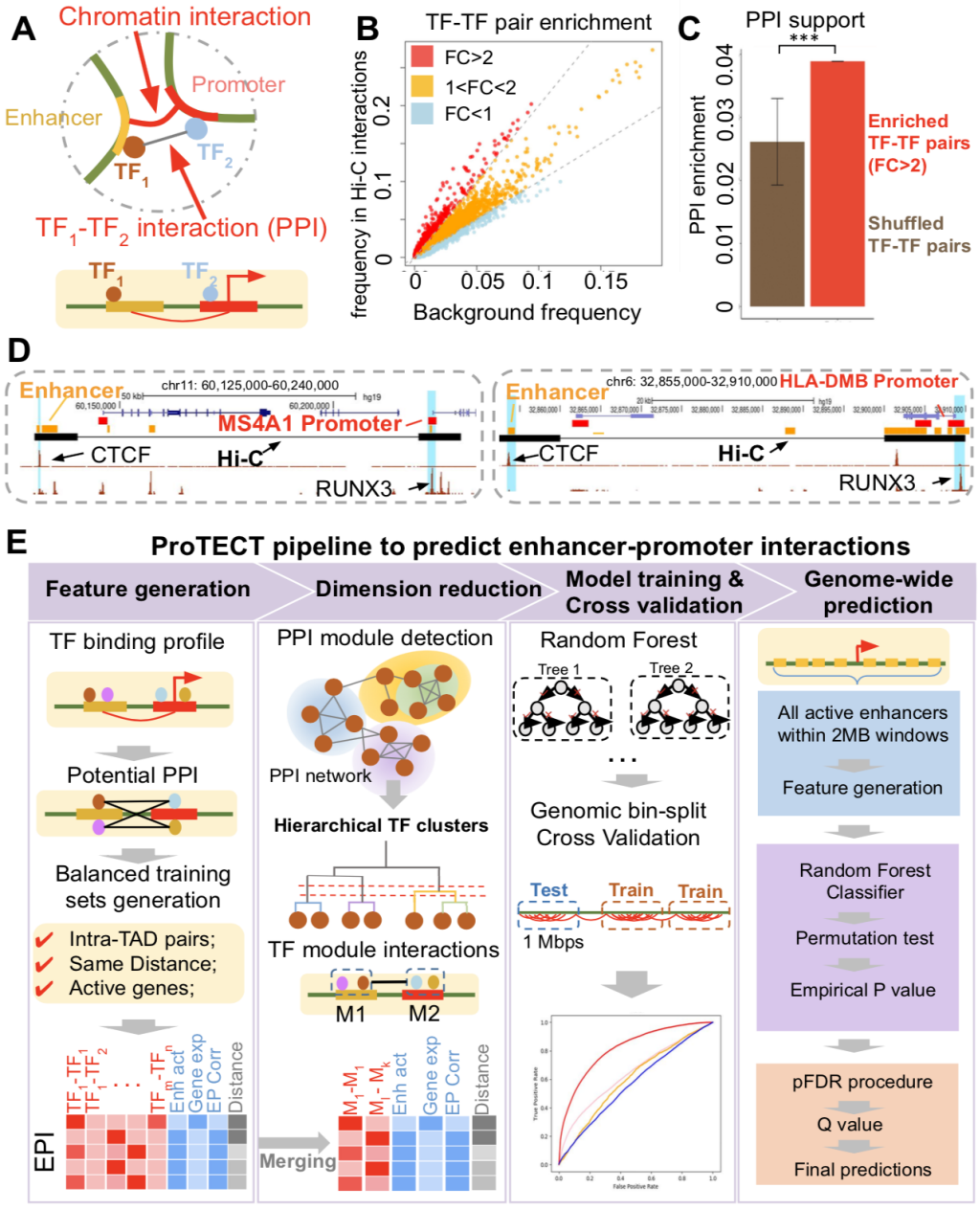


Figure 3.1. ProTECT infers long-range enhancer–promoter interactions based on TF PPI features. (A)The enhancer–promoter interactions are regulated by PPIs between enhancer-binding TFs (brown) and promoter-binding TFs (blue), which link distal enhancers (orange) to the proximity of promoters (red) in 3D chromatin structure. (B) Enrichment of TF–TF pairs in Hi-C interactions (y-axis) compared to background (x-axis). Points represent TF–TF pairs. Frequency is calculated as the fraction of enhancer-gene pairs containing the specific TF–TF

pairs. Fold-change (FC) is the ratio of the frequency in Hi-C interactions over the frequency in background. TF–TF pairs are colored by the FC (red: FC > 2; orange: 1 < FC < 2; blue: FC < 1). (C) Enriched TF–TF pairs are supported by PPIs. The fraction of pairs supported by PPIs are calculated for the set of enriched TF–TF pairs (red). As controls, the TF members from the enriched TF–TF pairs are randomly paired (brown). Statistical test is done based on 1000 random repeats of controls (***P-value = 10–3). Error bar represents sd. (D) Examples of Hi-C interactions linking enhancers (orange) and promoters (red) showing enhancer-binding CTCF ChIP-seq peaks and promoter-binding RUNX3 ChIP-seq peaks in GM12878 cells. (E) The workflow of ProTECT algorithm. A balanced training dataset is generated with confounding factors controlled. A feature matrix summarizing cell-type specific TF PPI features, activity-based features (enhancer activity, gene expression, enhancer-gene activity correlation), and genomic distances is then constructed. A novel hierarchical network community detection-based approach is applied for feature dimension reduction. Based on the reduced feature matrix, a random forest model is trained, and rigorous genomic-bin split cross- validations are used for performance evaluations and comparisons. Using the trained predictive model, genome-wide high-confidence enhancer–promoter interactions are predicted based on stringent permutation statistical tests.

excessively high-dimensional TF PPI features easily render predictive models with high overfitting risks. Second, individual TF PPIs are not independent features because of (i) co-binding TF modules along the 1D genome [89] and (ii) protein complexes consisting of multiple interacting TFs [139,140]. Both challenges require advanced feature dimension reduction approaches to efficiently handle the non-linear dependencies in features. In addition, as highlighted by recent benchmark studies [123,124], rigorous settings of cross-validation need to be designed for unbiased performance evaluation and interpretation.

In this study, we developed a new predictive model, ProTECT, to infer long-range enhancer–promoter interactions with substantially improved accuracy. A unique novelty of the model is designing a graph-based dimension reduction algorithm, which can efficiently incorporate combinatorial TF PPI features into the model and, in the meantime, control the overfitting risks. By setting rigorous genomic bin-split cross-validations and controlling various confounding factors, we systematically demonstrated the superior performance of our model compared to existing algorithms. Furthermore, we analyzed the relative importance of TF PPI features in different cell-types and prioritized the key TF PPIs that may participate in the regulation

of long-range enhancer–promoter interactions, leading to new mechanistic insights on enhancer regulation. Accordingly, we further classified genes into specific subsets, where enhancer-gene interactions are predicted to be mediated by different TF PPIs. Interestingly, genes in different subsets are enriched with distinct biological pathways, suggesting the specific functional impacts of TF PPIs. Genome-wide implementation of ProTECT in human GM12878 and K562 cell-lines results in 134 792 long-range enhancer–promoter interactions, which are significantly enriched with cis-eQTLs. In addition, by analyzing enhancer–promoter interactions mediated by different TF PPIs, we were able to assign specific TFs as upstream trans-factors to downstream target genes through distal enhancers. Strikingly, the prioritized TF–gene pairs are significantly supported by *trans*-eQTLs, leading to new mechanistic interpretations of trans-genetic effects propagated through the combined regulatory path-ways of TF bindings, TF PPIs and long-range chromatin interactions.

3.2 MATERIALS AND METHODS

To predict cell-type specific long-range enhancer–promoter interactions and obtain understandings of the under-lying mechanisms, we have developed a new algorithm ProTECT (i.e. PROtein-protein interactions of Transcription factors predicting Enhancer Contacts with Target genes). In addition to cell-type specific multi-omics data, ProTECT (https://github.com/wangjr03/PPI-based_prediction_enh_gene_links) further integrates the information of PPIs between transcription factors as new features, because TF PPIs have been found to be functionally associated with the regulation of chromatin loops [74–78,83,85,86,96]. The major steps of ProTECT are summarized in Fig 3.1E. By creating balanced training sets with confounding factors systematically controlled, ProTECT is trained on cell-type specific chromatin interactions linking distal enhancers and gene promoters. The high-dimensional TF PPI features are hierarchically grouped into feature modules based on a novel graph-based dimension

reduction approach. This approach can simultaneously control the overfitting risk and also reveal the cooperative complexes of TF interactions. Our model demonstrated substantially improved accuracy based on a series of rigorous performance evaluations. Along with genome-wide enhancer–promoter interaction predictions, ProTECT also identifies the key TF PPIs involved in chromatin interaction mediation and prioritizes specific gene sets whose expressions are regulated by distinct TF PPIs.

3.2.1 Chromatin contact maps and multi-omics datasets. ProTECT can take different types of chromatin contact maps as input data (Fig 3.1E), such as Hi-C [102], Capture Hi-C [103] and ChIA-PET [106]. In this study, we used the significant high-resolution Hi-C interactions from human GM12878 and K562 (GEO: GSE63525) [102] to train models for the two cell-lines separately. Enhancer-promoter pairs are labeled as positive samples if overlapping with Hi-C interactions, or are labeled as negative samples otherwise.

Enhancer coordinates are based on Roadmap and ENCODE enhancer annotations [89,90]. Cell-type specific enhancer activities in GM12878 and K562 cell-lines are quantified using the cell-type specific DNase-seq signals [90]. Other enhancer-associated histone marks, such as H3K27ac or H3K4me1 ChIP-seq data, can also be used to represent enhancer activities and have been found to produce similar predictions in our testing (see Results). Promoters of genes are defined as ±1 kb around transcriptional start sites (TSS), based on gene annotations from GENCODE v17 [141]. Cell-type specific gene expressions are measured by RPKM values of RNA-seq dataset from Roadmap Epigenomics project [90]. Correlation coefficients are calculated for enhancer-gene pairs across diverse cell-types [89,90] based on the same set of RNA-seq data for genes and DNase-seq data for enhancers.

The ChIP-seq datasets of transcription factor (TF) bindings in GM12878 and K562 are collected from ENCODE separately [89]. For each TF, if multiple datasets exist, one ChIP-seq

dataset is selected based on data quality evaluations. In total, 129 TFs in GM12878 and 270 TFs in K562 cell-lines are included in the analysis (Fig B.1A). The significant narrow peaks identified by MACS2 [142] are used to label whether a TF binds to a specific genomic location (Fig 3.1E).

The protein–protein interaction dataset is collected from the STRING database v11 [140]. To remove low-quality PPIs, only PPIs with confidence scores greater than 100 in the 'Experiments' category are included into the analysis. Multiple PPI confidence score thresholds (e.g. 200 and 300) are also tested, which produce similar predictive performance (see Results). The high-quality PPIs are then summarized into a matrix and represented as a PPI network, where every node corresponds to a protein and every edge corresponds to a protein–protein interaction. To account for the intratypic dimerizations of TFs from the Nuclear Receptor (NR), bHLH and bZIP families, these PPI edges are removed from the PPI network [143], because they can only bind locally as dimers. The nodes are further classified into two types: (i) TF protein nodes and (ii) non-TF protein nodes. For edges connecting two TF nodes, i.e. TF–TF PPIs, if both TFs are expressed in the specific cell-type, then the TF–TF PPI is considered as active. Therefore, cell-type specificity is assigned for every TF–TF PPI. non-TF protein nodes are maintained in the PPI network because they are useful to identify indirect TF–TF interactions mediated by non-TF proteins, leading to the discovery of TF PPI modules in subsequent steps.

3.2.2 Generation of the training dataset and the matrix of features. In a specific cell-type, enhancer–promoter pairs that overlap with significant Hi-C interactions [102], i.e. the enhancer of the pair overlaps with one of the Hi-C interaction anchors and the promoter overlaps with the other anchor, are labeled as positive samples of enhancer–promoter interactions. As reported by previous studies [108,144,145], the data quality of Hi-C interactions whose anchors are located in different topologically associated domains (TADs) are substantially reduced. Therefore, we remove cross-TAD interactions from the analysis, and only use intra-TAD enhancer–promoter

interactions, i.e. the interacting enhancer and promoter are located in the same TAD, to train the model.

To avoid biased model training and inflated performance evaluations, we generate a balanced negative set of training samples by randomly selecting the same number of enhancer–promoter pairs that do not overlap with Hi-C interactions. In addition, as pointed out by recent bench-mark studies [123], predictions of enhancer–promoter interactions can be substantially biased due to uncontrolled confounding factors. Thus, in the process of generating the balanced random set of negative samples, we strictly control three key confounding factors that have been found to influence the model (Fig 3.1E): (i) the negative samples of enhancer–promoter pairs should be intra-TAD pairs (Fig B.1B); (ii) the genomic separation distances between the enhancers and promoters follow the same distance distribution of the positive training set. Uncontrolled genomic distances have been found to substantially dominate the models and result in simple short-range predictions, leading to inflated performance [123,124]. Using the positive training set of enhancer–promoter pairs, we group them into different genomic distance bins. For each distance bin (bin-size = 50 kb), we sample the same number of negative enhancer–promoter pairs as observed from the positive set. Therefore, the genomic distance is controlled and the final predictions will not be driven by genomic distances alone (Fig B.1C, 1D). (iii) The negative enhancer–promoter pairs are sampled for genes which are actively transcribed (Fig B.1E, F). As demonstrated by previous studies [146], the false negative rates of Hi-C datasets are substantially lower in actively transcribed genomic regions, i.e. more enhancer– promoter interactions can be mapped by Hi-C in active regions compared to repressive genomic regions. To account for this intrinsic bias of Hi-C data, we restrict the sampling of negative enhancer–promoter pairs only from genes whose cell-type specific expression is nonzero (RPKM > 0). By controlling these three key sets of confounding factors, we thus construct the rigorous balanced training dataset for robust model training and performance evaluation. In total, the balanced training dataset contains 5348

enhancer–promoter pairs in GM12878 and 8650 enhancer–promoter pairs in K562.

Based on the cell-type specific multi-omics datasets, the matrix of features are then constructed for enhancer– promoter pairs in the training dataset (Fig 3.1E). There are three types of features incorporated into the model: (i) activity-based features; (ii) genomic distance and (iii) TF PPI features. Activity-based features include (i) cell-type specific enhancer activity measured by DNase-seq signals as described above [90]; (ii) cell-type specific gene expression measured by RNA-seq [90] and (iii) the activity correlations between enhancers and their paired genes calculated from diverse cell-types profiled in the ENCODE and Roadmap Epigenomics projects [89,90]. All these activity based features are differentially distributed across positive and negative training sets, suggesting they are informative to make predictions (Fig B.2A-C). For each enhancer-gene pair, the genomic distance is calculated as the distance between the center of the enhancer and the gene's TSS. Although they have been controlled in the positive and negative training sets based on genomic bins, there might be residue distance bias within bins. Therefore, the inclusion of genomic distances into the feature matrix captures the residue effects of genomic distances, leading to robust feature prioritization in subsequent analyses.

TF PPIs are the most important set of features for the model because of both the mechanistic relationship with long-range regulation [131,132,147] and their significant enrichment in enhancer–promoter interactions (Fig 3.1B, C and Fig B.2D). In each specific cell-type (i.e. GM12878 or K562 cells), all TFs with available ChIP-seq datasets are collected as described above and compared with the PPI database [140]. From the pool of all candidate pairs, the TF–TF pairs that are capable of forming direct PPIs are considered as TF PPIs. Considering the differences of binding sites in enhancers or promoters, each TF PPI pair is allocated with two directional features. For example, TF_a –TF_b represents the PPI between enhancer-binding TF_a and promoter-binding TF_b , while TF_b –TF_a represents the PPI between enhancer-binding TF_b and promoter-binding TF_a. Thus, a set of directional TF PPI features is generated. Because the

features are generated only for TFs with cell-type specific ChIP-seq signals, PPIs between TFs that are not active in the specific cell-type do not participate in the predictions. Enhancer-promoter pairs are scanned for TF binding peaks in enhancers and promoters. For each enhancer–promoter pair, if TF_a binds to the enhancer and TF_b binds to the promoter, then the directional PPI feature TF_a–TF_b is labeled as 1. Therefore, a matrix of TF PPI features is constructed for all enhancer–promoter pairs. Combining with the activity-based features and genomic distances, the full matrix of features is then built (Fig 3.1E).

3.2.3 Hierarchical TF community detection on the PPI network. Due to the large number of TF PPI features, dimension reduction is fundamentally important for the construction of robust predictive models. Without dimension reduction, there are 1888 TF PPI features in GM12878 and 7066 TF PPI features in K562 cells. Although a number of TF PPIs are enriched in enhancer–promoter interactions (Fig 3.1B and C), direct incorporation of these TF PPI features makes the model to be over-complicated, leading to poor generalization of predictions. To illustrate the significant overfitting issues of direct incorporation of high-dimensional TF PPI features, a basic random forest model is used to test the performance in GM12878 [102]. The features include the activity correlations between enhancers and genes, genomic distances and 1888 active TF PPI features. Although the regular 5-fold cross-validation shows an AUC of 0.89, a rigorous genomic-bin split cross-validation (see subsequent sections on cross-validation) shows the unbiased AUC as 0.55, suggesting strong overfitting problems without advanced feature dimension reductions (Fig B.3). Thus, a novel predictive model is needed for predicting long-range enhancer–promoter interactions based on PPI features among transcription factors.

To address the over-fitting problem, we substantially reduce the feature dimensions by hierarchically grouping individual TF PPIs into TF PPI modules based on the topology of the PPI network, while maintaining the predictability of the model (Fig 3.1E). TF PPI modules represent

densely connected groups of TFs in the PPI network, and they are hierarchically organized where smaller PPI modules merge together to form larger modules (Fig B.4). Biologically, using TF PPI modules as features is consistent with the regulatory mechanisms of long-range chromatin loops, because multiple TFs usually interact with each other as protein complexes. Empirically, the biological relevance of TF PPI modules is also supported by the data. As can be seen in Fig B.5, similar to individual TF–TF pairs, a specific subset of TF modules are strongly enriched in enhancer– promoter Hi-C interactions and are strongly supported by PPI connections (*P*-value = 1.39×10^{-2}, permutation test).

TF PPI modules are computationally identified from the PPI network [140] using a random-walk based network-community detection approach. The PPI network, including non-TF protein nodes, is modeled as an undirected weighted graph, where the weights on edges are the 'Experiment' PPI scores from the STRING database [140]. Define W as the adjacency matrix of the PPI network, and define the diagonal degree matrix D as $D_{ii} = \sum_j W_{ij}$. Hence, based on the stochastic model of random-walks on graphs [148], the 1-step transition probability from node i to node j is $\frac{W_{ij}}{D_{ii}}$, and the p-step transition matrix $Trans_p$ can be calculated as $Trans_p = (D^{-1} * W)^P$. Based on the p-step transition matrix, the pairwise distance matrix between TFs (denoted as R) can be further calculated as: $R = diag(G)^t * 1 + 1^t * diag(G) - 2G$, where $G = Trans_p * Trans_p^t$. Each entry in the matrix R quantifies the distance between a pair of TFs based on the PPI network structure. Hierarchical clustering is then applied to the pair-wise distance matrix R to identify hierarchical PPI modules of TFs (Fig 3.1E). 'wald' method is used in the hierarchical clustering as suggested by previous studies of network-community detections [149]. By testing multiple values (Fig B.4A and B.4B), p is set to be 20 in order to balance the detection of both local (i.e. small-size) and global (i.e. large-size) modules.

In the constructed hierarchical clustering tree, the leaf nodes are individual TF PPIs. By

applying the bottom-up merging strategy on the tree, individual TF PPIs are first grouped into small-size PPI modules, *i.e*. S-modules, with the maximum size of S_{max}. *S*-modules represent densely connected TFs in the PPI network, corresponding to candidate protein complexes. S-modules are further merged to form large-size PPI modules, i.e. *L*-modules, with the maximum size of L_{max}. *L*-modules represent larger PPI network components that cover multiple densely connected S-modules. Biologically, *L*-modules represent candidate groups of highly interacting protein complexes. The maximum sizes for S-modules (S_{max}) and *L*-modules (L_{max}) are selected based on the modularity score of the clustering [150] (Fig B.4). The modularity score Q is defined as $Q = \frac{1}{2m} * \sum_{ij}\left(W_{ij} - \frac{k_i k_j}{2m}\right) * \delta(c_i, c_j)$ where W is the adjacency matrix, k_i is the degree of node I, m is the total number of edges in the PPI network ($m = \frac{1}{2}\sum_i k_i$), and c_i is the membership assignment to modules for node i. Modularity scores are extensively calculated for different choices of maximum module sizes (Fig B.4C and D), because the choice of specific maximum module sizes automatically determines the total number of modules and results in the final module membership assignments. The optimal size of *S*-modules is selected as the one yielding the maximum modularity score, which guarantees that the generated *S*-modules represent densely connected TF groups. The optimal size of *L*-modules is selected as the one corresponding to the elbow point of modularity score curves, leading to the delineation of large-scale PPI components without significant loss of modularity. Compared to Markov Cluster Algorithm, the PPI modules from our approach demonstrate higher modularity scores and larger module sizes (Fig B.6), which is desired for feature dimension reductions. Using this procedure, a two-layer hierarchical modular structure is finally built and each individual TF PPI is assigned with the memberships belonging to a specific *S*-module and a specific L-module.

Based on the TF PPI module assignments, individual TF PPI features (i.e. direct TF–TF PPIs) are merged into module-level PPI features, and, therefore, the feature matrix of TF PPIs are restructured accordingly (Fig 3.1E). There are two types of module-level PPI features: (i)

intra-module features, which include all *S*-modules and *L*-modules. The intra-module features cover PPIs between TFs within the same modules. (ii) inter-module features, which include inter *S*-module features and inter *L*-module features. The inter-module features cover PPIs linking TFs from two different modules. Given a pair of *S*-modules, e.g. *S*-module *a* and S-module *b*, if there exists a TF member from *S*-module *a* that has PPI with a TF member from *S*-module *b*, then the pair of *S*-modules *a* and *b* is included into the feature matrix as one inter *S*-module PPI feature. The inter L-module PPI features are defined in the same way by checking PPIs of TF members from two L-modules. Each inter-module feature is further split into two directional features, depending on the binding sites of TF members in enhancers and promoters. Using this approach, the PPI features are substantially reduced. For example, the 1,888 individual TF PPI features are reduced to only 78 module-level PPI features in GM12878 and the 7066 individual TF PPI features are reduced to only 238 module-level PPI features in K562 cells.

The training set of enhancer–promoter pairs are then scanned for module-level PPI features. For each specific enhancer–promoter pair, based on the counts of individual TF PPI features calculated in the previous step, the counts of module-level PPI features are generated depending on the module memberships of TFs (Fig 3.1E). For each module-level PPI feature, if multiple TF PPI features are found for an enhancer–promoter pair, the maximum count is used for the module-level feature. Although the number of features is substantially reduced after using module-level PPIs, the specific PPI information is still maintained in this procedure, as shown in Fig B.5. It suggests that the module-based dimension reduction does not cause the loss of information, while substantially reducing the risk of over-fitting

3.2.4 Predictive model of long-range enhancer–promoter interactions. Random forest model is used to predict cell-type specific long-range enhancer–promoter interactions based on the feature matrix constructed above, after module-based dimension reduction (Fig 3.1E). Random

forest model is selected due to its superior performance of handling non-linear feature dependency and its capability of prioritizing the key set of important features for subsequent biological interpretations. As a free model parameter, the number of decision trees in the model is extensively tested with different values, and the accuracy of predictions is found to be robust (Fig B.7).

Additionally, to quantitatively demonstrate the contributions from TF PPIs, we train random forest models based on two versions of input features: (i) the model is trained using only activity-based features and genomic distances; and (ii) the full set of features including module-level TF PPI features. The Area Under Curve (AUC) values of cross-validations are calculated for the two versions. The increased AUC from version 2 is the quantitative measurement of the additional information contributed from TF PPIs that is not encoded in activity-based or genomic distance features.

3.2.5 Feature selection. In the random forest model, the backward feature elimination approach is used to select useful module-level TF PPI features, where the features with the minimum importance are recursively eliminated from the model. Furthermore, the statistical significance of the directions of TF PPI features are evaluated. As described in the previous section, every module-level PPI feature is split into a pair of two directional features, based on the binding sites of TFs in enhancers or promoters. For example, the feature *module a– module b* represents the PPI between an enhancer-binding TF member from *module a* and a promoter-binding TF member from *module b*. Reversely, the feature *module b – module a* represents the PPI between an enhancer-binding TF member from *module b* and a promoter-binding TF member from *module a*. Based on the statistical evaluation of the feature directions, insignificant directional features are merged into un-directional features. This feature merging procedure not only reduces the number of features but also reveals the biological roles of TF bindings in the

context of different binding orientations.

The determination of whether a pair of directional TF PPI features to be merged into an un-directional feature is a model selection problem. While Akaike Information Criterion (AIC) has been a widely-used metric for parametric models, it can not be applied to random forest models, which are non-parametric. Instead, we use the Generalized Degrees of Freedom (GDF) method to calculate a relaxed AIC [151] for the random forest model. GDF is a metric to evaluate the degrees of freedom for Bernoulli distributed data, *e.g.* the binary labels for enhancer–promoter interactions. And it is defined as $GDF \approx \sum_i(\widehat{y_i'} - \widehat{y_i})/(y_i' - y_i)$, where y_i is the observed label for data point i, y_i' is the perturbed label by inverting y_i, i.e. $y_i' = 1 - y_i$, $\widehat{y_i}$ is the predicted label from the model using the unperturbed y_i, and $\widehat{y_i'}$ is the predicted label from the model using the perturbed y_i'. As suggested by previous studies [151], to calculate GDF, 20% samples are simultaneously perturbed. The relaxed AIC of random forest models are then estimated as $AIC = -2l_m + 2GDF + GDF(GDF + 1)/(N - GDF - 1)$, where N represents the total number of data points and l_m represents the goodness-of-fit of the random forest model. As suggested by previous analyses [151], l_m is calculated as the averaged R^2 value from 5-fold cross-validations.

For each pair of directional TF PPI features, the relaxed AIC metrics are calculated before and after they are merged into an un-directional feature. If a smaller AIC is observed by merging the two directional features, the model with the merged un-directional feature is then selected, because the reduced AIC suggests the directions of the pair are not statistically important. This procedure is conducted for all pairs of directional TF PPI features, and a final random forest model with the selected features is built. In GM12878 cells, the number of module-level TF PPI features is reduced to 53 from 78. In K562 cells, the number is reduced to 139 from 238. This feature selection process further boosts the generalizability of our model and improves the biological interpretations of the learned TF PPI features (i.e. directional or un-directional).

3.2.6 Cross-validation and performance comparison. To evaluate the performance of our model, i.e. area under curve (AUC), we designed a stringent strategy of 5-fold cross-validation. As highlighted by previous studies [123,124], multiple factors have been found to substantially inflate the performance evaluations and cause overfitting problems. First, the confounding factors (i.e. TAD domain structures, genomic distances between enhancers and promoters, and gene expression levels) need to be controlled. Otherwise, the performance will be biased and dominated by confounding factors. We addressed this issue in the step of data generation as described in previous sections. Negative samples are randomly generated with the con-founding factors controlled to have the same distributions as seen from the positive samples. Second, inflated cross-validation AUC can be found due to the spatially proximal enhancer–promoter pairs across the training and testing datasets [123,124]. Because TF binding profiles are highly correlated among enhancers and promoters in neighboring genomic regions, proximal enhancer–promoter interactions that are allocated in the testing set will substantially inflate the accuracy. Therefore, random splits of samples based on typical cross-validation may suffer from the dependency of spatially proximal samples allocated in both training and testing sets, as has been noted in previous studies [123,124]. To address this issue, we developed a genomic bin-split cross-validation approach (Fig 3.1E). In this approach, the human genome is first divided into consecutive 1Mb bins. In each of the 5-fold cross-validation steps, 80% of the genomic bins are selected as training bins. And the balanced and con-founding factor controlled samples of enhancer–promoter pairs from the training bins are used to train the random forest model. The remaining 20% bins are selected as testing bins, and the samples of enhancer–promoter pairs from the testing bins are used to test the model. Using this genomic bin-split cross-validation method, the dependency between training and testing samples are broken and the model performance can be rigorously quantified.

The performance of our model, ProTECT, is compared with two most recent supervised methods that also leverage TF information: IM-PET [118] and TargetFinder [119]. In addition to activity-based features and genomic distances, IM-PET and TargetFinder also includes the TF binding features in enhancers and promoters, while TargetFinder further incorporates TF binding information in the genomic windows between enhancers and promoters. By comparing with these two algorithms, we can further demonstrate the improved accuracy is obtained purely from the unique features of our model, i.e. the PPIs between TFs.

The stand-alone package of IM-PET (https://github.com/tanlabcode/IM-PET) is applied to the same dataset. Since IM-PET automatically makes predictions for all enhancer-gene pairs with distances <2 Mb, only the enhancer-gene pairs overlapping with the dataset are used for performance evaluation, leading to a fair comparison for IM-PET. The TargetFinder software (https://github.com/shwhalen/targetfinder) is also implemented to the same training and testing dataset. The same set of TF ChIP-seq peaks are used to generate the window related features for TargetFinder. 5-fold cross-validation with the same genomic bin-split strategy is applied to remove the potential issues of inflated performance evaluations.

In addition, to quantitatively demonstrate that the improved accuracy of ProTECT is indeed contributed by TF PPI features, we randomly permute the PPIs between TFs, with the degree of each TF in the PPI network unchanged. Furthermore, for every TF, the specific binding sites in enhancers and promoters are also maintained. Therefore, only the TF PPI features are shuffled across enhancer–promoter pairs. The same model training and evaluation procedure are then applied on the permuted dataset. The resulting AUC is then compared to the model trained on the original dataset. This comparison provides direct evidence on the contributions of TF PPIs to chromatin interaction regulation.

3.2.7 Genome-wide prediction of long-range enhancer–promoter interactions. The trained

ProTECT algorithm is applied to all enhancer– promoter pairs with genomic distances <2 Mb across the whole human genome to make genome-wide predictions of cell-type specific enhancer–promoter interactions (Fig 3.1E). The features for each candidate enhancer–promoter pair are generated in the same way as described in previous sections. By applying the trained random forest classifier, every candidate enhancer–promoter pair is assigned with a predicted score of interacting with each other. To derive un-biased estimates of the statistical significance for the scores, i.e. *P*-values, a null distribution of the scores is generated by permuting the feature matrix across enhancer–promoter pairs. This permutation approach effectively maintains the overall abundances of different features in the shuffled dataset. Based on the null distribution, the *P*-value for each enhancer–promoter pair is then calculated.

Unlike the phase of model training, where the genomic distances are controlled in order to learn specific TF PPI signatures, the phase of genome-wide predictions requires the incorporation of genomic distance information. As shown by chromatin contact maps, e.g. Hi-C datasets, enhancer– promoter pairs with shorter genomic separation distances have higher probability to interact and the probabilities decay as the distances increase (Fig B.1C). To statistically incorporate the genomic distances based on this prior knowledge, we use the pFDR algorithm [152] to transform *P*-values into distance-aware *q*-values. In pFDR, the distribution of distances between Hi-C linked enhancers and promoters is treated as prior probabilities of interactions for enhancer–promoter pairs. Based on Hi-C data, ProTECT divides the range of distances into consecutive 20 kb bins, and the prior probability of interactions for each distance bin is calculated as: , where is the prior probability for distance-bin . The prior probability for bin 1 (i.e. the shortest distance bin) is set to be the default 0.05. The pFDR under rejection region in distance-bin is then calculated as , where represents the *P*-value for each enhancer–promoter interaction. *P* follows the uniform distribution under the null hypothesis, i.e. H=0, so that can be estimated by , where is the *P*-value for the enhancer–promoter interaction *j*, *N* represents the

total number of *P*-values, and equals to 1 if x is true and equals to 0 otherwise. Therefore, the *q*-values can be calculated as , which combines the information from both the distance-aware prior probabilities and the *P*-values from the random forest model (*P*). Based on the *q*-value threshold of 0.05, the final genome-wide predictions of significant enhancer–promoter interactions are obtained.

3.2.8 Feature interpretation for mechanistic insights. Using the trained random forest model of ProTECT, we evaluate and rank the importance of features, i.e. the module-level PPI features in the model. The top-ranking module-level PPIs are considered as important features, which represent putative protein complexes that may regulate chromatin interactions. Furthermore, in order to obtain detailed mechanistic understandings of important PPIs between specific TFs, we decode the module-level PPI feature importance into TF-level PPI feature importance. For each prioritized module-level PPI feature, we decompose it into individual TF–TF PPI features, i.e. specific PPIs between an individual enhancer-binding TF and an individual promoter-binding TF. Then the genome-wide predictions of enhancer–promoter interactions are scanned, and the fractions of predictions that contain the specific TF-level PPI features are calculated. The fractions scanned from genome-wide predictions are highly correlated with the fractions calculated from the Hi-C training samples in model training, and are more robust, given the larger pool of genome-wide enhancer–promoter pairs (see Results). Using the fractions, the top-ranking TF-level PPI features are thus identified for each important module-level PPI feature. The prioritized features, both module-level and TF-level, shed light on new biological insights on long-range enhancer regulation.

3.2.9 Pathway enrichment analysis for genes regulated by specific TF PPIs. To investigate whether chromatin interactions mediated by different TF PPIs may participate in distinct biological

pathways, we classify genes based on the specific TF PPI features involved in their interactions with enhancers. For each top-ranking module-level PPI feature, we first identify the top five TF-level PPI features using the method described above. Then, we scan the genome-wide predictions of enhancer–promoter interactions and collect the subset of interactions that contain at least one of the top five TF-level PPI features. Finally, the subset of interactions are ranked by their *q*-values, and the top 1000 genes regulated by these interactions are selected. In this way, the prioritized subset of genes represent strong targets of long-range enhancer regulation mediated by the important TF PPIs. Gene Ontology enrichment analyses are performed on different gene sets using DAVID [153] to check whether they are enriched with specific biological pathways.

3.2.10 cis-eQTL enrichment analysis for predicted long-range enhancer–promoter interactions. As the orthogonal information to validate the accuracy of genome-wide predictions made by ProTECT, *cis*-eQTL datasets from the matched human tissues and cell-types are compared with the predicted enhancer–promoter interactions. Because our genome-wide predictions are made in human GM12878 and K562 cells, we selected four eQTL datasets [154–157] which were profiled from either whole blood tissues or lymphoblastoid cells. A predicted enhancer–promoter interaction is considered to be supported by a *cis*-eQTL (i.e. a significantly associated SNP-gene pair), if the enhancer contains the SNP and the promoter matches with the gene. For each eQTL dataset, the fraction of predicted enhancer–promoter interactions that are supported by *cis*-eQTLs is calculated, and is compared to two versions of negative controls. The first version of negative control is based on random pairing enhancers with promoters that are within 2 Mb distances. The second version of negative control further requires the genomic distances of random enhancer–promoter pairs follow the same distribution from our predicted enhancer–promoter interactions. Therefore, the second version is a more stringent control. For

each version, 1000 random samples are generated. And the statistical significance, i.e. *P*-values, of the observed overlapping fractions from our predictions is calculated as the portion of random samples showing a higher overlapping fraction than the real observed one.

In addition to cis-eQTLs, we also use *cis*-hQTLs, i.e. histone QTLs, to evaluate the accuracy of our predictions. The hQTL dataset was also profiled from the human GM12878 cells [158]. Similarly, a predicted enhancer–promoter interaction is considered to be supported by a cis-hQTL (i.e. a significantly associated SNP-histone pair), if the enhancer contains the SNP and the promoter overlaps with the histone modification peak. The overlapping fraction is also compared with the two versions of negative controls to justify the enrichment of *cis*-hQTLs in support of our predictions.

3.2.11 cis-eQTL enrichment around TF binding sites. For *cis*-eQTLs that overlap with predicted enhancer– promoter interactions, the genomic locations of the SNPs from *cis*-eQTLs are further compared with TF binding sites within enhancers. Here, the TF binding sites are defined as the ChIP-seq peak summits. For each enhancer included in this analysis, the TFs involved in important PPI features prioritized from the previous steps are selected. The genomic distances between the SNPs and the binding sites of these TFs are calculated. To statistically test whether the SNPs are closer to these important PPI-related TFs, two versions of random controls are generated. The first version is generated by randomly sampling binding sites of any TFs within the same set of enhancers. And the second version is generated by randomly sampling binding sites of TFs that are members of bottom-ranking PPI features, based on feature importance calculations from the previous sections. For each version of negative controls, *P*-values are calculated using Kolmogorov–Smirnov tests by comparing the cumulative distributions of distances.

3.2.12 trans-eQTL enrichment analysis for enhancer-mediated TF– gene pairs. Compared to *cis*-eQTLs, *trans*-eQTLs can provide additional evidence to support the functional associations between the prioritized TFs and specific genes, where the TF's PPIs are predicted to mediate enhancer–promoter interactions of the target genes. For enhancer-binding TFs that are members of the important PPI features, we first collect the predicted enhancer–promoter interactions mediated by the corresponding PPI features. Genes regulated by these predicted interactions are thus considered as the downstream target genes of the specific enhancer-binding TFs. We define this relationship as enhancer-mediated TF–gene pairs. To exclude the possibility of promoter-mediated effects, we remove the genes whose promoters are also bound by the specific TF.

Using the trans-eQTLs from the published database [159], we identify a subset of trans-eQTLs whose SNPs are located within TF's gene bodies (plus –10 kb from TSS) and target genes are covered in our input dataset. For this specific subset of *trans*-eQTLs, the SNPs are likely to disrupt the transcription of the TF genes, which in turn affects the TF's regulation on the downstream target gene's expression.

Hypergeometric test is used to statistically test whether the enhancer-mediated TF–gene pairs significantly overlap with the subset of trans-eQTLs described above. A TF–gene pair is considered to overlap with a *trans*-eQTL if the SNP is located within the TF's gene body and the gene is the same as the *trans*-eQTL's target gene. As comparisons, two versions of controls are generated based on the same set of TFs and enhancers. The first version uses the nearest genes to the enhancers as target genes, instead of using ProTECT's predictions. The second version randomly selects genes within 2 Mb distances as target genes. In each version, the same number of enhancer–promoter interactions are generated as seen from the foreground for each sample, and totally 1000 random samples are created, along with the hypergeometric *P*-values.

3.3 RESULTS

3.3.1 Long-range enhancer–promoter interaction prediction based on PPIs among TFs

As discovered by recent experimental studies [77–79,81–86,131,132], the protein–protein interactions between specific transcription factors have been found to participate in the regulation of long-range chromatin loops, where the TFs bind to enhancers and promoters respectively (Fig 3.1A). The PPIs between the enhancer-binding TFs and promoter-binding TFs facilitate the 3D proximity of enhancers and the target gene's promoters. By analyzing the Hi-C interactions between enhancers and promoters in human GM12878 cells, a specific set of TF–TF pairs are found to be enriched in enhancer–promoter interactions (Fig 3.1B), compared to their frequencies in distance-controlled random enhancer–promoter pairs. Interestingly, these TF–TF pairs are also enriched with known PPIs (Fig 3.1C, *P*-value = 10^{-3}), suggesting that the TFs within each pair can establish interactions at the protein level. Fig 3.1D shows two examples, where both enhancer–promoter Hi-C interactions contain enhancer-binding CTCF peaks and promoter-binding RUNX3 peaks. And the physical interaction between RUNX3 and CTCF is validated by the PPI database STRING [140], suggesting the RUNX3-CTCF interaction as a putative mechanism linking the enhancers with specific promoters. These observed enrichments strongly indicate the functional importance of TF PPIs in long-range chromatin loops and the possibility of predicting cell-type specific enhancer–promoter interactions using TF PPI features.

Due to the large number of TF PPI features, i.e. PPIs between enhancer-binding TFs and promoter-binding TFs, basic predictive models significantly suffer from overfitting problems, as shown in Fig B.3. Therefore, to efficiently leverage the information of TF PPIs from the high-dimensional feature space and overcome the overfitting risks, we developed a new machine learning classifier, ProTECT, to predict cell-type specific long-range enhancer– promoter interactions (Fig 3.1E). Detailed algorithmic designs have been described in Materials and Methods. Overall, there are four main steps to achieve the final predictions: (i) generation of the

balanced Hi-C based training dataset, along with cell-type specific TF PPI features; (ii) dimension reduction of features based on hierarchical network community detection; (iii) predictive model construction using random forest and (iv) Genome-wide predictions of cell-type specific enhancer–promoter interactions.

As a new predictive model, here we highlight a series of key novelties of ProTECT (see Materials and Methods for details). First, a rigorous method of controlling confounding factors, such as TAD domains, genomic separation distances and gene expression levels, is designed in the steps of data and feature generations. This method efficiently removes the impacts of confounding factors, which are fundamentally important to control as discussed by recent benchmark analyses [123,124]. Second, the graph-based dimension reduction approach not only addresses the potential risk of overfitting but also facilitates the prioritization of functionally important TF PPIs and TF complexes. Third, a generalized degree of freedom (GDF) technique [151] is incorporated to improve feature selections, leading to new biological understandings of specific TFs. Fourth, a stringent genomic bin-split cross-validation strategy is developed for unbiased and robust performance evaluation. This stringent strategy thoroughly breaks the dependency between the training and testing datasets and avoids the inflated performance estimations that have been commonly found in existing methods [123,124]. Fifth, a genomic distance-aware pFDR procedure [152] is implemented to identify statistically significant enhancer–promoter interactions along the whole human genome.

We trained ProTECT using the high-resolution Hi-C datasets from the human GM12878 and K562 cell-lines separately [102]. The balanced and confounding factor-controlled training dataset contains 5,348 long-range enhancer–promoter interactions in GM12878 and 8650 interactions in K562 cells. The trained classifiers were further applied to make genome-wide cell-type specific predictions of enhancer–promoter interactions. As shown in subsequent sections, the ProTECT algorithm not only improves the prediction accuracy substantially, but also reveals

novel mechanistic insights on the functional roles of TF PPIs in the regulation of long-range

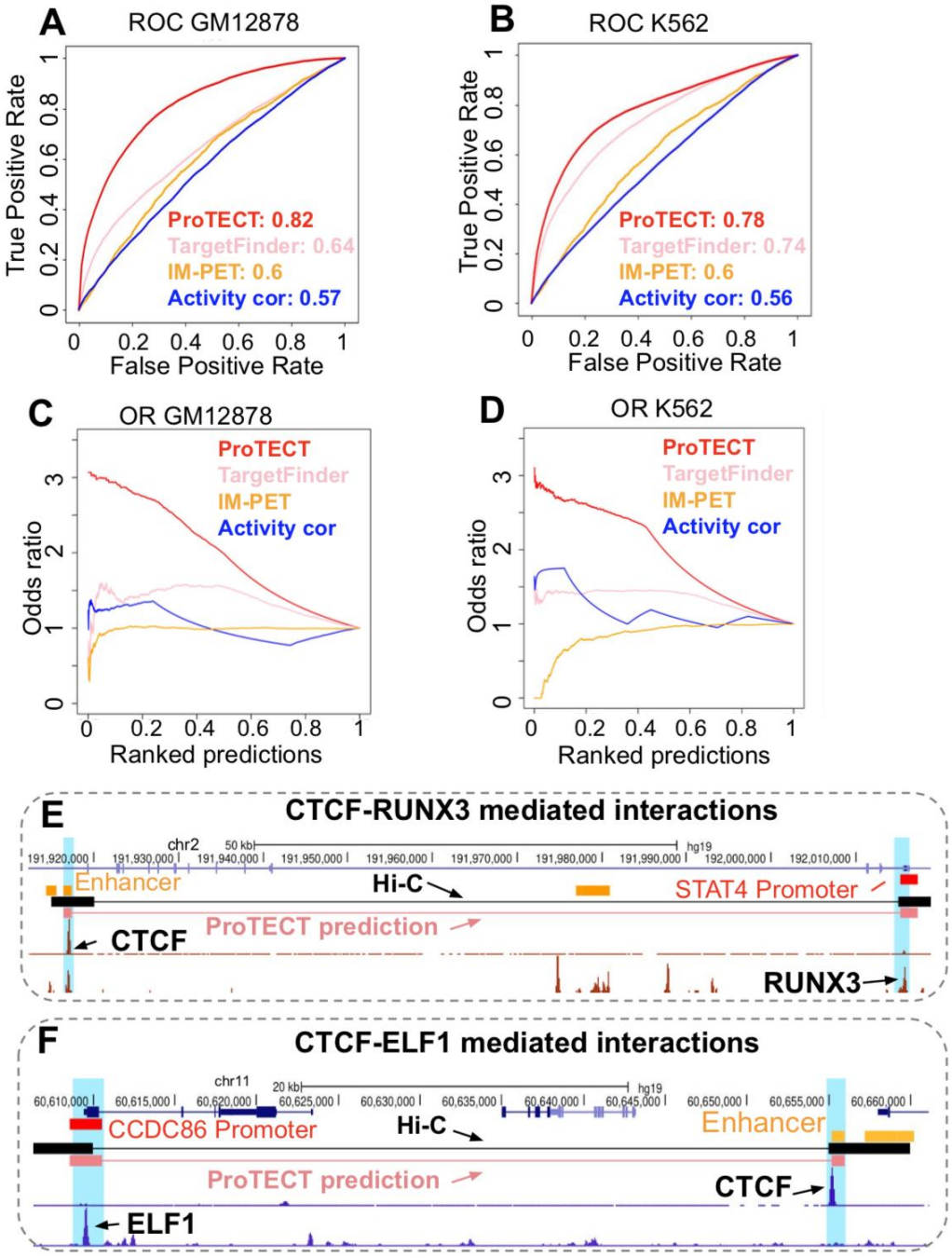


Figure 3.2. Performance comparison in GM12878 and K562 cells. ProTECT, TargetFinder, and IM-PET are applied on the same input datasets and are evaluated based on the averaged performance of 5-fold genomic-bin split cross-validation. As a baseline comparison, a random forest model using only enhancer-gene activity correlations is also included in the analysis. (A,

B) ROC curves in GM12878 (A) and K562 (B). (C, D) The enrichment of Hi-C interactions in top-ranking predictions. Cumulative odds ratios of true positives (y-axis), i.e. overlapping Hi-C interactions, are calculated across the ranked lists of predictions where predictions with stronger scores are ranked at the top (x-axis), in GM12878 (C) and K562 (D). (E, F) Examples of enhancer– promoter interactions predicted by ProTECT (pink paired lines) in GM12878 (E) and K562 (F). In each example, the highlighted enhancer (orange) is predicted to interact with the highlighted promoter (red) by ProTECT. Both predictions are supported by cell-type specific Hi-C interactions (black paired lines). The prioritized TF PPIs mediating the interactions are CTCF-RUNX3 (E) and CTCF-ELF1 (F) respectively, both of which are top-ranking PPI features from the random forest model.

chromatin loops. The prioritized TFs and their specific PPIs provide a new platform to understand the complex interplay among TFs, enhancers and genes, and remarkably, open a new avenue to systematically interpret both cis- and trans-eQTLs in human genetics analyses.

3.3.2 Boosted performance based on features of TF PPIs

Using the genomic bin-split cross-validation strategy (see Materials and Methods), we rigorously tested the accuracy of ProTECT and compared with the other two supervised methods, i.e. IM-PET(45) and TargetFinder [119]. In both GM12878 and K562 cell-lines, ProTECT achieves the highest performance (Fig 3.2A and B): AUC = 0.82 in GM12878 and AUC = 0.78 in K562 cells.

And the accuracy of ProTECT is robust with respect to the number of trees used in the random forest models (Fig B.7). As comparison, TargetFinder is ranked as the second algorithm with AUC values below 0.74, while the AUC metrics of IM-PET is around 0.6. As a baseline comparison, a random forest model using only activity correlations between enhancers and genes, without using TF PPI features, shows AUC values around 0.57. Because we systematically controlled confounding factors in the training dataset, the AUC estimates are not dominated or biased by those factors, especially the genomic separation distances. Therefore, these comparisons strongly support that the ProTECT model substantially boosts the prediction accuracy over existing algorithms.

In addition to the overall AUC metrics, to demonstrate that ProTECT has better capabilities of pinpointing true enhancer–promoter interactions in top-ranking predictions, we calculated the cumulative odds ratio (OR) of true positives along the ranked list of predictions. As shown in Fig 3.2C and 3.2D, ProTECT achieves much higher OR curves than other algorithms, especially in the zone of top-ranking predictions. Because top-ranking predictions are the main *de novo* discoveries used for experimental studies in practice, this observation further exemplifies the superior precision of ProTECT.

Moreover, we further evaluated the robustness of ProTECT's superior performance with respect to different settings of input features and data. As shown in Fig B.8, by setting different confidence score cut-offs on PPIs to be included as input features (i.e. 100, 200 and 300), ProTECT robustly achieves the highest accuracy (AUC > 0.78) compared to other methods. In addition, using different epigenetic signals to represent cell-type specific enhancer activity levels, such as DNase-seq, H3K27ac and H3K4me1, ProTECT demonstrates highly similar accuracy, with DNase-seq and H3K27ac based versions slightly better than the H3K4me1 based version (Fig B.8). Furthermore, we also tested the performance on imbalanced dataset, where the ratio of positive-to-negative samples is 0.1, as suggested by previous studies [118,119]. ProTECT consistently shows the best ROC and Precision-Recall curves (Fig B.9). To obtain orthogonal evidence on ProTECT's accuracy, we also used a diverse panel of Hi-ChIP [94,160,161] and ChIA-PET [89] datasets from the matched cell-types as gold-standards for enhancer–promoter interactions. Remarkably, ProTECT maintains the highest accuracy across all comparisons based on different gold-standard datasets (Fig B.10 and 11). Across the five Hi-ChIP evaluations, ProTECT achieves AUC >0.78, while TargetFinder and IM-PET only show AUC <0.66. Using ChIP-PET datasets as gold-standards, ProTECT achieves AUC >0.84 while other methods demonstrate AUC <0.76. These tests systematically support the robustness of ProTECT's performance advantages.

Fig 3.2E shows one example predicted by ProTECT in human GM12878 cells. The distal enhancer is located 99.4 kb from the predicted target gene's promoter, and this long-range prediction is supported by a cell-type specific Hi-C interaction [102]. Based on the trained random forest model, this enhancer–promoter interaction is mediated by the PPI between the enhancer-

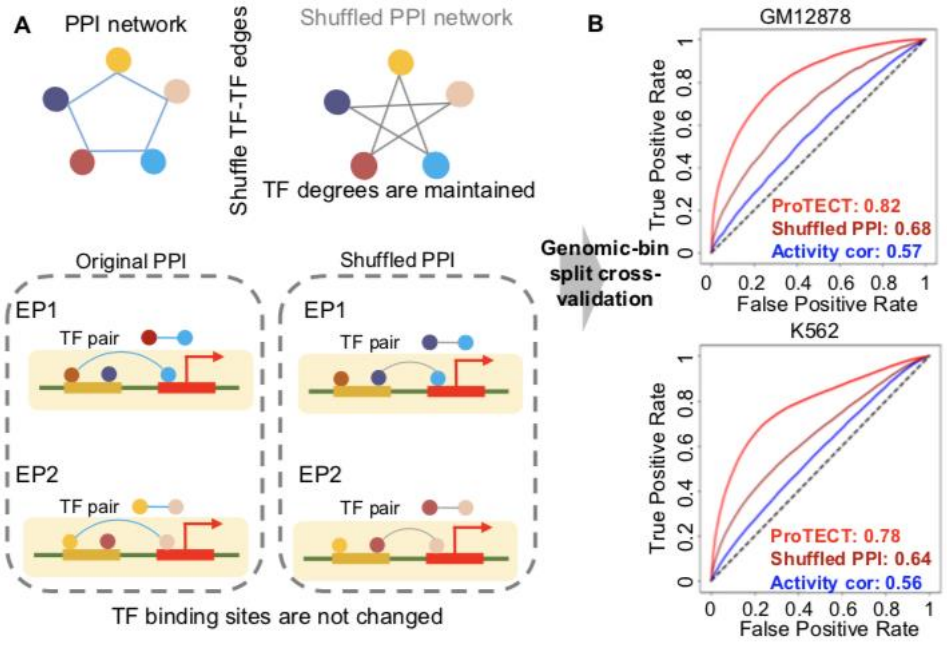


Figure 3.3. TF PPI features provide additional information beyond TF bindings and activity-based features. (A) Schematic figure of the permutation test on TF PPI features. The shuffled PPIs are generated by randomly pairing two interacting TFs from the original pool of TF PPIs, while the degrees of PPI partners and TF binding sites in enhancers and promoters are maintained. Based on the shuffled PPI features, a new random forest model is trained and then evaluated by the same cross-validation procedure. (B) ROC plots for the models based on the original TF PPI features (red), the models based on the shuffled TF PPI features (salmon), and the baseline models based on activity-correlation features alone (blue), in GM12878 and K562 cells.

binding CTCF and the promoter-binding RUNX3 (Fig 3.2E). Interestingly, the correlation between the enhancer's activity and the target gene's expression across different cell-types is only 0.28, which strongly suggests the importance of incorporating TF PPI features in predicting enhancer–

promoter inter-actions. A similar example from K562 is shown in Fig 3.2F, where the distal enhancer is located 46kb from the predicted target gene's promoter, and is also supported by a cell-type specific Hi-C interaction (Fig 3.2F). This enhancer–promoter interaction, which only shows an activity correlation of 0.261, is successfully predicted based on the PPI between enhancer-binding CTCF and promoter-binding ELF1. Overall, these results demonstrate that TF PPI features can improve the delineation of specific interacting enhancer–promoter pairs from neighboring non-interacting pairs, beyond the information of activity-related features. In addition, specific hypotheses of the mechanisms mediating chromatin interactions, i.e. the functional TF PPIs linking enhancers and promoters, are derived from the model simultaneously.

To further justify that the superior performance of ProTECT is indeed due to the information from TF PPI features, we randomly shuffled the TF–TF connections in the PPI network (Fig 3.3A). Therefore, the specific TF binding sites in enhancers and promoters are strictly maintained (see Materials and Methods), while the PPI features across enhancer–promoter pairs are randomized. This shuffling strategy also controls the degree of PPI partners for each TF, i.e. the number of protein neighbors in the PPI network. By training the ProTECT model on the shuffled data, we found that the accuracy is substantially reduced. The AUC based on PPI-shuffled data is only 0.68, while the original AUC of ProTECT is 0.82 in human GM12878 cells (Fig 3.3B). Similar decrease of performance is also observed in human K562 cells (Fig 3.3B). The striking differences of prediction accuracy suggest that the performance improvement of ProTECT is mainly induced by TF PPI features, instead of TF binding information, consistent with previous biological studies of the functional roles of PPIs in chromatin loop regulation [137].

To evaluate the model's dependence on the cell-type specificity of TF bindings, we swapped the TF ChIP-seq data across GM12878 and K562, and run ProTECT based on the swapped data. As expected, the prediction accuracy decreased in both cell-types (Fig B.12A and B), suggesting the necessity of using TF datasets from the matched cell-types. Interestingly,

ProTECT still maintains the highest prediction accuracy when other algorithms are also trained on the swapped TF data, suggesting reasonable generalizability of ProTECT. In addition, to test the model's dependence on the number of TFs included as features, we obtained the intersection subset of TFs whose ChIP-seq are available in both GM12878 and K562, and trained ProTECT based on features derived from this subset. The cell-type specific predictions in GM12878 and K562 demonstrate similar accuracy (AUC = 0.74 and 0.70, Fig B.12C), suggesting additional TFs are needed in each cell-type beyond the intersection subset.

3.3.3 Genome-wide prediction of long-range enhancer–promoter interactions

The trained random forest model is then applied to the genome-wide dataset in GM12878 and K562 cell-lines separately to predict novel enhancer–promoter interactions (Fig B.13A-D). All enhancer–promoter pairs within 2Mb distance windows are included into genome-wide predictions (see Materials and Methods), as suggested by observations from experimental Hi-C datasets [102]. For each enhancer–promoter pair, a *P*-value from the permutation test is generated, which is further used to derive a *q*-value based on the pFDR approach [152] (see Materials and Methods). Using the *q*-value threshold of 0.05, there are totally 60 016 significant enhancer–promoter interactions predicted in GM12878, and 80 591 significant enhancer–promoter interactions predicted in K562 (Fig 3.4A). The median separation genomic distance between linked enhancers and promoters is 243 kb in GM12878 (Fig B.13E), consistent with enhancer's function of long-range regulation. In the predicted GM12878 enhancer–promoter network, >37% of enhancers regulate multiple genes (Fig B.13F), whose accuracy is consistent with the overall performance (Fig B.14) and 24% of these multi-gene enhancer links are supported by experimental chromatin interactions. On average, every gene is regulated by 6.9 enhancers (Fig B.13G), suggesting combinations of multiple enhancers are recruited for precise transcriptional regulation. Similar patterns are also observed in the predicted K562 enhancer–

promoter network (Fig B.13H-J). Furthermore, the predicted enhancer–promoter interactions are highly cell-type specific. By comparing the predictions in GM12878 and K562, only 5815 (~4.2%) enhancer–promoter interactions are shared by the two cell-types (Fig 3.4A). Compared to the recent activity-by-contact (ABC) model [162], our genome-wide predictions demonstrate higher accuracy, as quantified by both ROC and Precision-Recall curves, using Hi-ChIP data as gold-standards (Fig B.15).

3.3.4 Important protein–protein interactions regulating chromatin interactions

To gain insights of the underlying mechanisms of linking distal enhancers to target gene's promoters, we analyzed the feature importance of module-level PPI features inferred by the random forest model and further prioritize the representative TF-level PPI features. We first identified the top-ranking module-level PPI features, which represent the protein complexes of interacting TFs involved in chromatin loops (Fig 3.4B and C). For example, in GM12878 cells, module(CTCF)-module(POLR2A) is ranked as the top third feature (here the module-level features are named by the most abundant TF-level PPIs linking the modules). Interestingly, this is consistent with a recent experimental study [163], which also found that the enhancer-binding CTCF interacts with the promoter-binding Pol II and participates in the regulation of long-range chromatin loops. As another interesting example, the module-level PPI feature module(IKZF1)–module(RB1) is one of the top-ranking features in K562, consistent with their critical functions in leukemia cells and their impacts on chromatin structure [164,165]. Additional examples of the prioritized module-level TF PPIs are visualized as PPI networks in Fig B.16, showing the complex PPI connectivity between TF modules binding to enhancers and promoters.

In order to characterize the key PPI features between individual TFs, instead of TF modules, we further decode the module-level PPI features into ranked TF-level PPI features (Fig 3.4D), based on their occurrences across genome-wide predictions of enhancer–promoter

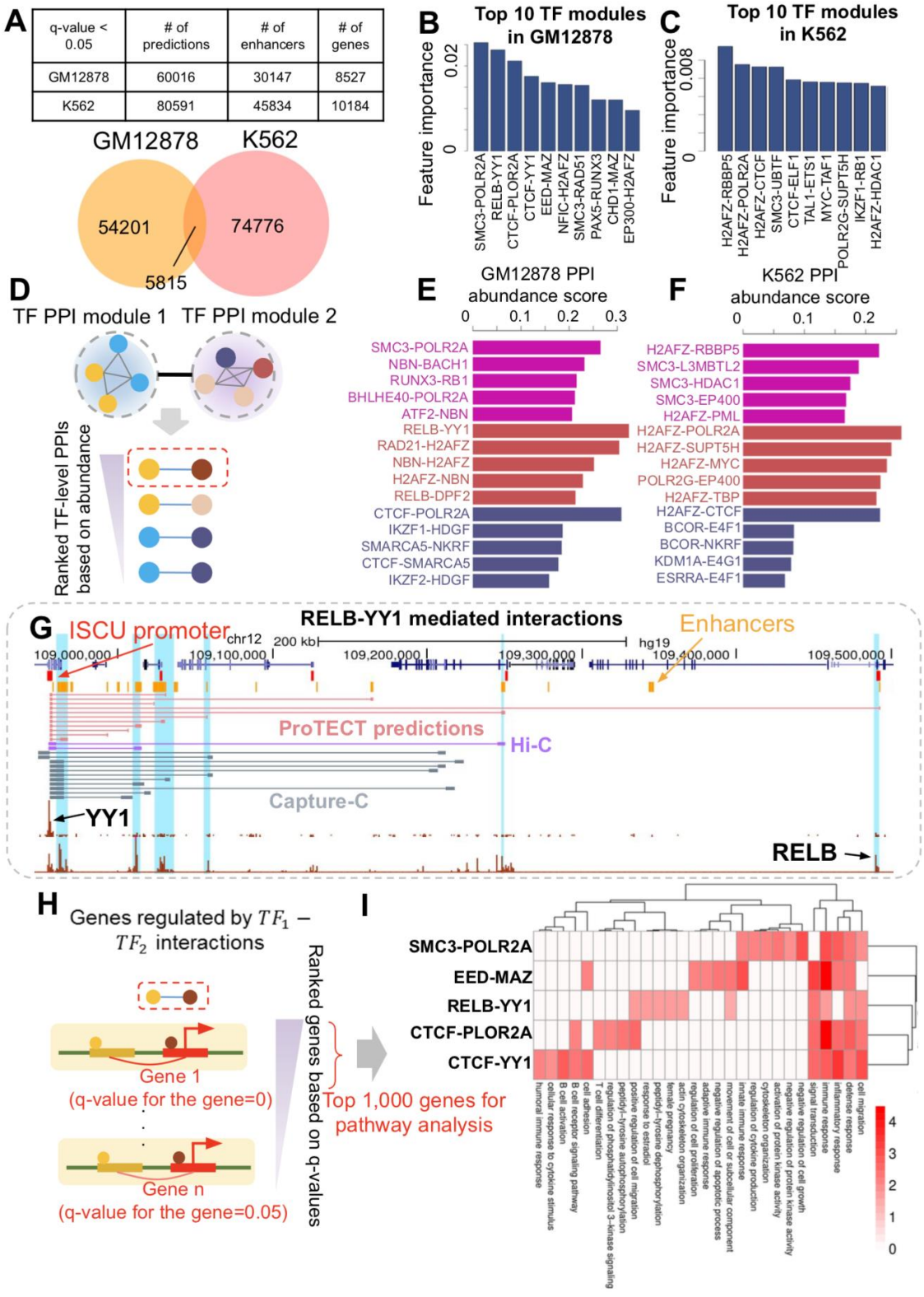


Figure 3.4. Genome-wide prediction of enhancer–promoter interactions reveals functional roles of TF PPIs in gene regulation. (A) Summary of genome- wide predictions in GM12878 and K562. The venn-diagram shows the overlap between predicted enhancer–

promoter interactions in GM12878 (yellow) and K562 (salmon). (B, C) Feature importance (y-axis) of top 10 module-level TF PPI features based on the random forest models in GM12878 (B) and K562 (C). Each module-level PPI feature is named by the most abundant TF-level PPIs between the modules as axis-labels (x-axis). (D) Schematic figure of ranking specific TF-level PPIs in each PPI module. For each module-level PPI feature, all TF-level PPIs linking two TFs from the pair of two modules (the pair of modules can be the same to represent intra-module TF-level PPIs) are ranked by their occurrences in the predicted long-range enhancer–promoter interactions (abundance scores). (E, F) Examples of top 5 TF-level PPIs for three representative module-level features in GM12878 (E) and K562 (F). (G) Examples of predicted enhancer–promoter interactions regulated by RELB-YY1 in the ISCU locus. Predicted enhancer–promoter interactions for the ISCU gene are shown as the pink paired lines. Totally 11 enhancers are predicted to interact with the promoter of ISCU, and five predictions are supported by Hi-C (purple paired lines) or capture Hi-C (grey paired lines). ChIP-seq signal tracks of RELB and YY1 (brown signal peaks) are consistent with predictions. (H) Schematic figure of ranking enhancer–promoter interactions regulated by specific TF PPIs. For each prioritized TF PPI feature, enhancer–promoter interactions are ranked based on the q-values inferred by ProTECT. Top 1000 genes are then selected by following the ranked list of interactions for pathway enrichment analysis. (I) Pathway enrichments of genes regulated by five different TF PPIs in GM12878. The top 10 most enriched pathways for each TF PPI feature are shown. The heatmap is colored based on the $-\log_{10}$(P value) of pathway enrichments.

interactions (see Materials and Methods). Genome-wide predictions are used to calculate the abundance scores for TF level PPIs because they provide a large pool of enhancer– promoter links, and the abundance scores are found to be highly correlated with the observations from Hi-C training samples (Fig B.17, Spearman correlation = 0.95). For each module-level feature, the top 5 most abundant PPI features between specific enhancer-binding and promoter-binding TFs are identified. For example (Fig 3.4E), RELB-YY1 is predicted to be a key TF-level PPI feature in long-range enhancer regulation. In support of this new discovery, RELB has recently been found to promote gene expression by interacting with YY1 [166]. As another example, SMC3-HDAC1 is one of the top-ranking features in K562 (Fig 3.4F), consistent with the reported regulatory roles of HDAC1 on chromatin structure by interacting with SMC3 [167]. The discoveries of these key TFs and their PPIs as candidate functional factors in chromatin loop formation may lead to new biological hypotheses of enhancer regulation for in-depth experimental investigations.

As a demonstration of the potential importance of TF PPIs in linking distal enhancers to promoters, Fig 3.4G shows the predicted long-range enhancer–promoter interactions for the gene

ISCU. There are totally 11 enhancers predicted by ProTECT to interact with ISCU's promoter, and five of them are supported by experimental data of chromatin interactions based on Hi-C or Capture Hi-C (Fig 3.4G), indicating the high accuracy of the predictive model. The inferred top-ranking feature is the PPI between enhancer-binding RELB and promoter-binding YY1. Consistent with this prediction, YY1 has a strong ChIP-seq binding site at the promoter of ISCU, and almost all linked enhancers have ChIP-seq signals of RELB binding. Importantly, four out of the five validated enhancers show the strongest RELB ChIP-seq binding signals (Fig 3.4G), indicating the shared mechanism of these enhancer–promoter interactions for the gene ISCU. In this region, the longest interaction predicted by ProTECT is from a distal enhancer located >547 kb from ISCU's promoter. Although not captured by chromatin contact map experiments, this specific enhancer contains a sharp ChIP-seq peak of RELB binding (Fig 3.4G), suggesting this novel prediction as a strong candidate of enhancer–promoter interactions. It also implies the capability of ProTECT to discover long-range enhancer regulation that might be missed by experimental approaches.

To investigate whether the orientations of PPI features between enhancer-binding and promoter-binding TFs have impacts in chromatin interactions, we designed a systematic model selection strategy to test whether a pair of two TF PPI features with opposite directions can be merged into one un-directional PPI feature without reducing the predictive accuracy (see Materials and Methods). Using this approach, 32 pairs of directional PPI features in GM12878 are merged into 16 un-directional features, suggesting there is no statistical preference of binding sites (i.e. enhancers versus promoters) between interacting TFs involved in these PPIs. For example, the features ATF2-SMARCA5 and SMARCA5-ATF2 are merged into an un-directional feature by the model, consistent with the observation that the two directional PPI features have similar abundance in enhancer–promoter interactions (Fig B.18A). A similar example involves the merge of IKZF1-CREM and CREM-IKZF1 features (Fig B.18A). In

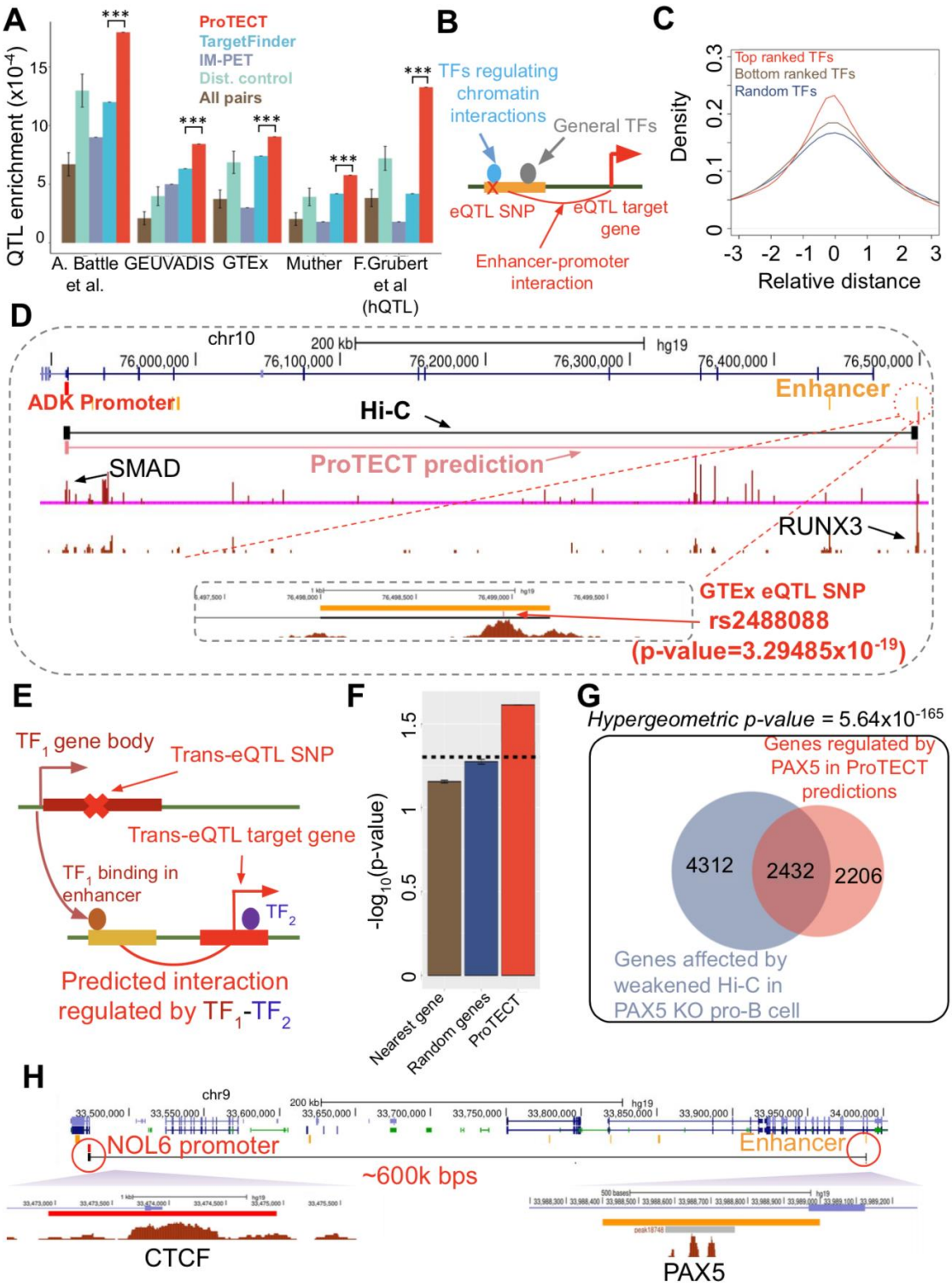


Figure 3.5. Predicted enhancer–promoter interactions are enriched with cis-QTLs and trans-QTLs. (A) cis-eQTLs and cis-hQTLs from multiple datasets (x- axis) are significantly enriched in predicted enhancer–promoter interactions in GM12878 (red). The fractions of enhancer–promoter interactions overlap- ping with cis-QTLs (y-axis) are compared with other methods and two versions of controls: (1) random enhancer–promoter pairs (brown) and (2)

distance- controlled random enhancer–promoter pairs (blue). 1,000 samples are generated for both versions to calculate P-values (***: P-value < 1.04×10^{-4}). Error bars represent sd. (B) Schematic figure of cis-eQTL SNPs located in the binding sites of functionally important TFs (blue) of chromatin interactions, com- pared to general enhancer-binding TFs (grey), as a mechanistic hypothesis of cis-regulatory effects on target gene expression. (C) Distributions of relative distances between cis-eQTL SNPs and binding sites of different enhancer-binding TFs. Relative distances (x-axis) are genomic distances between SNPs and TF ChIP-seq peak summits normalized by the sizes of TF peaks. Binding sites of top-ranking TFs inferred by ProTECT (red) significantly overlap with cis-eQTL SNPs, compared with bottom-ranking TFs (grey, P-value = 3.02×10^{-4}) and random enhancer-binding TFs (blue, P-value = 4.17×10^{-18}). (D) Example of a cis-eQTL, i.e. the rs2488088-ADK pair, overlapping with a predicted enhancer–promoter interaction (pink paired lines). The predicted interaction is supported by Hi-C (black paired lines). The prioritized PPI feature is RUNX3-SMAD, consistent with the ChIP-seq signal tracks (brown signals). Zoom-in view of the distal enhancer (orange) shows the cis-eQTL SNP rs2488088 is located at the peak summit of RUNX3 binding site. (E) Schematic figure of trans-eQTL SNPs located in specific TF genes, whose binding to enhancers are predicted to mediate long-range enhancer–promoter interactions of trans-eQTL target genes. (F) Hypergeometric test on the overlaps between trans-eQTLs (i.e. trans- SNP-gene pairs) and enhancer-mediated TF–gene pairs, if the SNP is located in the TF's gene body and the trans-eQTL's target gene is the same as the TF's target gene (red, P-value = 0.014). The $-\log_{10}$(P-value) (y-axis) from the hypergeometric test is compared to two versions of controls: 1) nearest genes to the enhancers (brown); and 2) random target genes (blue). Each control is generated 1000 times and the error bars show the sd. The black dash line corresponds to $-\log_{10}(0.05)$. (G) Venn diagram comparing genes affected by weakened Hi-C interactions in PAX5 KO pro-B cells and genes regulated by PAX5 in ProTECT predictions (Hypergeometric test, P-value = 5.64×10^{-165}). (H) Example of a trans-eQTL, i.e. rs10973104-NOL6 pair, supported by the predicted enhancer-mediated PAX5-NOL6 pair. The predicted enhancer–promoter interaction for NOL6 (black paired lines) is based on the prioritized TF PPI feature PAX5-CTCF. ChIP-seq signals (brown signal tracks) show a strong CTCF peak in the NOL6 promoter (red) and strong PAX5 peaks in the linked enhancer (orange). The trans-eQTL SNP rs10973104 is located in the gene body of PAX5, which is 3.6 Mb away from this locus.

spite of these un-directional PPI features, there are 37 features remaining to be directional in GM12878. For example, there is a significant preference of SMC3-MXI1 feature over the MXI1-SMC3 feature (fold-enrichment = 7.80, Fig B.18B). This is an interesting observation considering the function of SMC3 (a subunit of cohesin [168]) in chromatin structural maintenance, and the reported regulatory function of MXI1 binding in promoter regions [169]. Another example corresponds to the preference of EP300-POL2R2A over POL2R2A-EP300 (fold-enrichment = 9.19, Fig B.18B), consistent with the well-known enhancer binding activities of EP300 [170] and the transcriptional initiation function of POL2R2A [171]. Similarly, 184 pairs of directional PPI

features in K562 are merged into 92 un-directional features, while 47 PPI features remain to be directional.

3.3.5 Genes regulated by different TF PPIs are enriched in distinct pathways

To evaluate the downstream impacts of chromatin interactions mediated by different TF PPIs, we focused on the top 5 module-level PPI features (Fig 3.4B and C). We identified the strongest enhancer–promoter interactions mediated by each feature separately based on the ranked q-values of predictions (see Materials and Methods). Genes that are regulated by the top-ranking enhancer–promoter interactions are therefore collected for pathway enrichment analysis (Fig 3.4H). Overall, these prioritized genes are enriched with immune-related or B-cell-related pathways (Fig B.19A and B), which is expected since the predictions are inferred from GM12878 and K562 cell-lines. Strikingly, for each specific PPI feature, the gene sets are strongly enriched with distinct groups of pathways (Fig B.19A and B). Fig 3.4I shows the most enriched pathways for each TF PPI feature discovered in the GM12878 cell-line. Clearly, the enhancer–promoter interactions mediated by different TF PPIs are enriched with diverse biological processes. For example, the CTCF-YY1 feature is found to be associated with long-range regulation of genes in the B cell receptor signaling pathway, while the SMC3-POLR2A feature is associated with genes of the innate immune response pathway (Fig 3.4I). To exclude the potential bias caused by gene background, we carried out pathway enrichment analysis based on two additional gene backgrounds, respectively: (i) genes with the same set of promoter-binding TFs and (ii) genes with the same set of enhancer-binding TFs (Fig B.19C and D). Based on these two rigorous gene backgrounds, the majority (>67%) of enriched pathways are still discovered. These differentially enriched pathways further highlight the functional roles of TF PPIs in regulating gene expression and maintaining the specific cellular states.

3.3.6 Predicted enhancer–promoter interactions are enriched with cis-eQTLs

Because the predictive model is trained on Hi-C datasets, we use cis-eQTLs as orthogonal evidence to quantitatively evaluate the accuracy of the genome-wide predictions of enhancer–promoter interactions. By comparing the predictions with the SNP-gene pairs of significant eQTLs, we calculated the overlapping enrichment scores (see Materials and Methods). Using four eQTL datasets generated from matched cell-types or tissues (e.g. whole blood tissues or lymphoblastoid cell-lines) [154–157], the predicted enhancer– promoter interactions in GM12878 cell-line show significantly higher fractions overlapping with eQTLs, compared to stringent distance-controlled random interactions and other algorithms (P-value < 1.04×10^{-4}, Fig 3.5A). Similar, but relatively weaker, enrichment with eQTLs is found for predictions in K562 cell-line (Fig B.20A). In addition to *cis*-eQTLs, we compared our predictions in GM12878 with histone-QTLs from the same cell-line [158] and also observed strong enrichment (P-value = 3.27×10^{-5}) compared to distance-controlled random samples and other algorithms (Fig 3.5A). These observations not only support the high accuracy of genome-wide predictions but also suggest the putative mechanisms of *cis*-eQTLs mediated by chromatin interactions between regulatory elements and target genes.

3.3.7 *cis*-eQTLs are enriched in binding sites of prioritized TFs

The prioritized TF PPI features by the ProTECT model provides a new metric of delineating functionally important TFs for enhancer regulation against general enhancer-binding TFs, which is complicated due to the large array of TFs binding to enhancers. For a typical enhancer, it contains 10 different TF binding sites on average, based on the counts of TF ChIP-seq peaks in GM12878 from the ENCODE project [89]. However, binding itself is not sufficient to assign functional importance for TFs. As found by previous studies, TFs binding in enhancer regions are not equally important for the function of enhancers, with many enhancer-binding TFs

lacking evidence of regulatory impacts on gene expression [172]. This ambiguity hinders the understanding of enhancer activation and downstream effects. We hypothesized the TFs involved with top prioritized PPI features are more likely to be functional for enhancers. We tested this hypothesis by checking the enrichment of *cis*-eQTL SNPs within the binding sites of the prioritized TFs in enhancers (Fig 3.5B, see Materials and Methods). The *cis*-eQTLs are called in whole blood tissues from the GTEx project [154]. Interestingly, the SNPs of *cis*-eQTLs are located significantly closer to the binding sites of prioritized TFs in GM12878 (P-value = 4.17×10^{-18}, Kolmogorov–Smirnov test), compared to the binding sites of other adjacent enhancer-binding TFs (Fig 3.5C). To control the potential bias caused by data availability, we also generated a more stringent background only using TFs included in the model but inferred with low feature importance (see Materials and Methods). Compared with this new background, the prioritized TFs are still significantly enriched with cis-eQTL SNPs (P-value = 3.02×10^{-4}, Kolmogorov-Smirnov test, Fig 3.5C). In the K562 cell-line, cis-eQTL SNPs are also closer to the binding sites of the prioritized TFs but not statistically significant (Fig B.20B). Overall, this analysis supports the stronger regulatory effects of prioritized TFs whose PPIs may mediate long-range enhancer–promoter interactions. Additionally, the prioritized TF binding sites provide a new layer of information to pinpoint regulatory SNPs at a higher resolution, by dissecting the ambiguity of numerous TF bindings within enhancers.

As a representative example, a distal enhancer located > 589kb away is predicted by ProTECT to interact with the promoter of the ADK gene in GM12878 (Fig 3.5D), which is supported by experimental Hi-C data [102]. This long-range interaction is also supported by a significant eQTL, i.e. rs2488088-ADK (P-value = 3.29×10^{-19}) [154]. The prioritized TF PPI feature for this interaction is RUNX3-SMAD, where RUNX3 binds to the enhancer and SMAD binds to the promoter. By zooming into the enhancer element, which is 1.2 kb long and contains binding sites of five different TFs, the SNP rs2488088 is found to be precisely located at the ChIP-

seq peak summit of RUNX3 (Fig 3.5D), consistent with our prioritization of RUNX3 as the important TF for this enhancer. This observation also implies the mechanistic interpretation of this non-coding SNP, whose disruptive effect on the RUNX3 binding causes the loss of RUNX3-SMAD mediated long-range interaction to ADK.

3.3.8 *trans*-eQTLs are enriched in enhancer-mediated TF–gene pairs

As one of the advantages of the ProTECT algorithm, both *cis*-regulatory elements (i.e. enhancers) and *trans*-regulatory factors (i.e. TFs) are jointly modeled in long-range chromatin interactions. In traditional studies of *trans*-regulation of gene expression, analyses have been mainly limited to promoter-binding TFs as candidate *trans*-regulatory factors [173,174]. Based on the functional impacts of the predicted important TF PPI features (Fig 3.4B-I) and the observed enrichment of *cis*-eQTL SNPs in prioritized enhancer-binding TFs (Fig 3.5B–D), we hypothesized that there is an enhancer-mediated pathway of *trans*-regulation, i.e. the enhancer-binding TFs associated with top-ranking PPI features for long-range chromatin interactions are *trans*-regulatory factors for the expression of distal target genes (Fig 3.5E). To quantitatively validate this hypothesis, we compared the enhancer-mediated TF–gene pairs with significant *trans*-eQTLs [159], and the significance of overlaps are statistically tested using Hypergeometric tests (see Materials and Methods). Interestingly, the enhancer-mediated TF–gene pairs are found to be strongly supported by *trans*-eQTLs (*P*-value = 0.014, Fig 3.5F, Fig B.20C), suggesting that the SNPs of *trans*-eQTLs are associated with target gene's expression via the disruption of the TF gene's activity (Fig 3.5E), although the SNPs may be located far away from the target genes or even located in different chromosomes. The observed statistical significance is also stronger than two versions of controls, excluding the potential confounding effects of biased enhancer activity and genomic distances (Fig 3.5F, see Materials and Methods).

To obtain additional experimental evidence on the predicted enhancer-mediated TF–gene

regulation, we leveraged a differential Hi-C interaction dataset in mouse pro-B cells where 7810 weakened Hi-C interactions were identified following PAX5 knock-out [175]. The top-ranking PAX5 related PPI feature predicted by ProTECT is PAX5-CTCF, consistent with their collaborative roles in B cells [176,177]. Based on our genome-wide predictions in GM12878, we identified the subset of PAX5-CTCF mediated enhancer– promoter interactions (see Materials and Methods), and thus collected the enhancer-mediated target genes of PAX5. To purify the subsequent analysis, genes whose promoters are also bound by PAX5 are removed from the list. If PAX5 is a true trans-regulatory factor for these genes, the genes are expected to be targeted by the weakened long-range interactions following PAX5 knock-out. By mapping the genes to their homology in the mouse genome [178], 6,744 enhancer-mediated target genes of PAX5 are conserved. Strikingly, these genes are found to significantly overlap with the genes of weakened Hi-C interactions in PAX5-/- pro-B cells [175] (hypergeometric *P*-value = 5.64×10^{-165}, Fig 3.5G). To control the potentially biased enhancer activity and TF bindings, we generated two versions of controls. The first version randomly selects genes as enhancer-mediated target genes of PAX5. And the second version randomly chooses target genes of other TFs. 1000 random samples are generated for each version and the same number of genes are selected for each sample. Both versions of negative controls show decreased overlap with genes of weakened Hi-C interactions in PAX5–/– pro-B cells (*P*-value = 10^{-3}), supporting the predicted *trans*-regulatory links between PAX5 and target genes by ProTECT. Fig 3.5H shows one representative example of PAX5-CTCF mediated long-range enhancer–promoter interaction (~600 kb), where the enhancer contains multiple PAX5 binding sites and the promoter of the target gene, *i.e.* NOL6, contains a strong CTCF binding site. Interestingly, NOL6 is linked with weakened Hi-C interactions in PAX5–/– pro-B cells. These strong experimental validations, along with the enrichment of *trans*- eQTLs, suggest the biological validity of the predicted enhancer-mediated TF–gene pairs, and provide a new regulatory mechanism to discover and interpret *trans*-regulatory genetic variants.

3.4 DISCUSSION

In this study, we have developed a novel supervised algorithm, ProTECT, to predict long-range enhancer–promoter interactions. By incorporating new features of protein–protein interactions among transcription factors, the algorithm achieves superior performance compared to other methods, based on a rigorously designed genomic bin-split cross-validation procedure. Considering the overfitting risk of high-dimensional inter-dependent TF PPI features, a novel network-community based dimension reduction strategy is used to hierarchically organize TF PPIs into module-level features. This approach efficiently improves the generalizability of the predictive model to make robust predictions based on complex TF PPI patterns, while maintaining the detailed ranking of TF-level PPI features for specific mechanistic understandings of long-range enhancer regulation. With the impacts of confounding factors strictly controlled, the relative contributions of different features are systematically evaluated, which shows that TF PPIs contain substantially additional information beyond activity-based features of enhancers and genes.

The genome-wide implementation of ProTECT in GM12878 and K562 cell-lines generated 60 016 and 80 591 new predictions of significant enhancer–promoter interactions, which will be useful resources of cell-type specific enhancer regulation for biologists. In addition, a set of prioritized TF PPIs, in both module-level and TF-level, are identified as the key PPIs mediating long-range chromatin loops. Different TF PPIs are found to mediate enhancer regulation for genes in distinct biological pathways, implying specific functional roles of complex TF cooperation. The TF members participating in these prioritized PPI features can be used as candidate targets for knock-out to investigate the changes of specific enhancer–promoter interactions, which will expand the insights on the underlying mechanisms of chromatin loop formation and long-range gene regulation.

To gain orthogonal evidence of the validity of genome-wide predictions, *cis*- and *trans*-eQTLs are compared with the predicted enhancer–promoter interactions in three ways, each of

which supports one aspect of the interplay among TFs, enhancers and genes. First, the enrichment of overlaps between *cis*-eQTLs and enhancer–promoter interactions suggests the accuracy of predicted long-range *cis*- regulation by distal enhancers. Second, the enrichment of *cis*-eQTL SNPs located within the binding sites of prioritized TFs underscores the precise delineation of functionally important TFs for enhancer activities against other general enhancer-binding TFs. Third, the enrichment of overlaps between *trans*-eQTLs and enhancer-mediated TF–gene pairs highlights the novel identification of *trans*-regulatory pathways from upstream TFs to downstream genes via distal enhancers. The promising enrichment analyses further indicate that the predictions from ProTECT can be used as a platform to interpret *cis*- and *trans*-eQTLs, i.e. characterize the non-coding SNP's disruptive effects propagated through long-range enhancer regulation on gene expression. Therefore, combined with eQTL datasets, the ProTECT model can also be a useful tool to generate testable hypotheses in statistical genetics studies.

To control the model complexity, only direct PPIs between TFs are included as features, while indirect PPIs between TFs may also participate in the regulation of chromatin loops. For example, an enhancer-binding TF and a promoter-binding TF may not be able to interact with each other but they both can interact with a third protein. The incorporation of module-level TF PPI features helps to capture the potential indirect PPIs to some degree, but does not explicitly address this problem. Due to the large number of indirect PPI features and the limited number of labeled samples for model training, more advanced designs of feature selection will be needed to achieve a balance between predictive accuracy and model generalizability.

As a major novelty of the ProTECT model, the efficient inclusion of TF PPIs as features not only improves the predictions but also reveals mechanistic insights on long-range enhancer regulation. In the meantime, the algorithm requires the availability of large panels of TF ChIP-seq data for the specific cell-types under study, which may be a practical challenge for users. As one of the directions to extend the ProTECT model, it is possible to leverage the combined information

of chromatin accessibility data, e.g. DNase-seq or ATAC-seq data, and TF binding motif annotation datasets as approximations for cell-type specific TF bindings. Several recent studies have demonstrated the reasonable accuracy of this approximation [89,90]. Furthermore, multiple imputation algorithms have been recently developed for ENCODE cell-types or tissues to impute cell-type specific TF binding ChIP-seq signals [179,180]. The imputed TF binding signals can be used as alternative inputs for the model to make cell-type specific predictions of enhancer–promoter interactions, for cell-types lacking ChIP-seq datasets. As an evaluation of this possibility, we generated the imputed TF bindings by overlapping TF motifs with cell-type specific DNase-seq peaks, and then derived TF PPI features based on the imputed data. Remarkably, applied on the imputation-based input features, Pro-TECT is able to achieve high accuracy (Fig B.21). This evaluation strongly supports the wide applicability of ProTECT on diverse cell-types even if TF ChIP-seq data is not directly available.

CHAPTER 4

CHIP-SEQ MULTI-MAPPING READ ALLOCATION ALGORITHM REVEALS REGULATORY ELEMENT ACTIVITIES IN HUMAN GENOME

4.1 INTRODUCTION

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is a technology to map DNA-binding proteins and histone modifications in a genome-wide manner [181]. It has the advantages of high resolution, less noise and high genome coverage and becomes the preferred approach to profile cell-type/ tissue-specific regulatory elements. In ChIP-seq experiment, proteins first cross link with specific DNA sequences. The whole DNA sequence is then sheared into small fragments. These protein-DNA complexes are extracted through antibody enrichment and sequenced by high-throughput sequencing platform [182]. After mapping these sequenced DNA fragments (reads) back to the genome, we can infer the location of the regulatory elements in the genome. The sequence length in most platforms is short, often varying from 20bp to 100bp. As significant fraction of human genome is composed of repetitive regions and a large fraction of ChIP-seq reads are from repetitive region. In general, about 20-25% of the reads in ChIP-seq can be aligned to multiple locations on the genome. We call these multi-mapping reads ambiguous reads. Traditional pipeline [183] just discards these ambiguous reads, which results in a large loss of information.

Although there have been some computational algorithms that can allocate the ambiguous reads to one of its possible mapping positions, they suffer from either high false positive and high false negative rate [184, 185] or require the use of additional data [186]. In order to cope with the challenge of ambiguous read allocation and improve the mapping specificity, we developed a statistical model (RegisTER-ME) to iteratively infer the binding positions of ambiguous reads. RegisTER-ME takes use of the neighborhood read counts and local sequence information to

formulate the probability of an ambiguous read binding to one possible location. Gibbs sampling algorithm is applied to update these probabilities and infer the read allocation simultaneously. We showed that RegisTER-ME is able to identify numerous novel regulatory elements with higher accuracy than existing tool. Furthermore, we applied RegisTER-ME to all the single-end (SE) ChIP-seq data in the ENCODE database and achieved around 3 million new ChIP-seq peaks. These new peaks enable us to study the waves of transposable elemen(TE)t-derived transcription factor (TF) binding position creation, the co-evolution between TFs and the corresponding co-factors and the co-regulation among TE-derived TFs. In addition, by identifying the TF binding region an eQTL located in, we were able to reveal the functional mechanism of eQTLs.

4.2 MATERIALS AND METHODS

4.2.1 RegisTER-ME Algorithm. RegisTER-ME is a software for assigning each multi-mapping read in human single-end ChIP-seq data to one of its candidate mapping positions with the highest posterior probability. RegisTER-ME contains five major steps:

(i) ChIP-seq data cleaning.

Only read assignments in autosome and chromosome X, Y are kept for later calculation. If the total number of nucleotide mismatches, insertions or deletions is more than 2, the ChIP-seq read will also be removed.

For a genome position, if the number of canonical reads mapped to it exceeds a threshold, these reads will be considered as PCR duplications. Only canonical reads with higher mapping scores will be left among the PCR duplications.

To calculate the read count threshold for each ChIP-seq dataset, we fit a Poisson distribution, where

$$\lambda = \frac{total\ canonical\ read\ counts}{total\ length\ of\ autosome\ and\ chromosome\ X,Y\ (bp)}$$

By setting p-value =$\frac{0.001}{total\ length\ of\ autosomes\ and\ chromosome\ X,Y}$, we got the maximum number of reads mapped to the same genome position.

(ii) genome segmentation

The whole genome is segmented into a large number of small intervals (sites) to speed up the later calculation. If there are few ChIP-seq reads falling into a region and the concatenation of reads is less than 100bp, then this region is divided into multiple sites with length 100bp. If there are numerous reads falling into a region, to avoid the signal diffusion, we set a length threshold for each site in this region. In other cases, the site length is between 100bp and the length limitation.

To obtain a reasonable site length limitation, we run MACS1 (v1.4.2; default settings) [187] to gain a sense of binding region width of the target protein. Half of the median width of these binding sites is used as the site length limitation so that there will be around two to three sites within a true protein binding region and the read signals will not be largely dispersed by the ChIP-seq peak tails. In addition, all the genome sites are classified into two groups: within MACS1 peak regions and outside peak regions. These two groups are given label 1 and 0 respectively.

(iii) likelihood fitting

Likelihood value here represents the confidence of a site being the protein binding region or background when we observe a certain number of canonical reads and multi-mapping reads.

We use P_s to denote the likelihood distribution of site densities when these sites are within protein binding regions, and P_n to denote the likelihood when the sites are just genome

background. There are two groups of P_s, P_n distributions, one for only canonical reads within the site and the other for canonical reads and multi-mapping reads within the site.

As one multi-mapping read can be assigned to multiple sites, if we consider it the same as the canonical reads, these multi-mapping reads will be over-represented. To solve this problem, we give each multi-mapping read a weight 0.2 and each canonical read a weight 1. Then we use canonical read density to estimate the P_s, P_n. for canonical reads and weighted read density to estimate the P_s, P_n for all the reads (canonical reads and multi-mapping reads).

To smooth the P_s, P_n distribution, we compare them with some known distributions and find Gamma distribution achieves the best fit for both P_s and P_n.

(iv) prior estimation

Prior value in the model reflects the target protein's preference on a given sequence. Since the prior value only depends on its genome sequence, we use $P(D)$ to represent the prior value of a site k_i. D here denotes the original ChIP-seq data.

To quantify the target protein's preference, we down-sample the same number of sites with label 1 and sites with label 0. Next, we calculate the 6-mer frequencies in both groups. Using fold-change > 1.5, adjusted P-value < 0.001 in binomial test and 6-mer frequency larger than the median frequency of all the 6-mers, we get enriched 6-mers and depleted 6-mers in protein binding regions. To reduce the feature dimension, we further use K-means to create 6-mer clusters using their sequence similarity. For each site, we calculate its 6-mer cluster features using $n_i \times log10$(6-mer group ith fold-change). n_i refers to the occurrence of 6-mers in cluster i. Fitting the samples using logistic regression and applying the fitting results to all the sites, we obtain the prior value for each site.

(v) Gibbs sampling optimization

We use a_i to represent the ith multi-mapping reads. s_i =s_{i1}, S_{i2}, …, S_{in_i} are all the sites a_i can be assigned to, and k_{ij} is the sequence information of site s_{ij}. Let D represents the original ChIP-seq data, U stands for the assignment of all the genome sites, M$[-i]$ refers to the mapping of all the multi-mapping reads except ai and rij is the number of canonical reads in site sij, we can formulate the posterior probability of a_i being assigned to s_{ij} as

$$P\big(M[-i], D, k_{ij} \big| a_i \sim s_{ij}\big) \sim P(a_i \sim s_{ij} | M[-i], D, k_{ij}) \cdot P(M[-i], D, k_{ij})$$

$$\sim P\big(a_i \sim s_{ij} \big| M[-i], D, k_{ij}\big) P(k_{ij}|D)$$

$$= \frac{P_s(r_j+1)\prod_{m\in s_{i/j}} P_n(r_m)\cdot P(U\backslash s_i)}{\sum_{\tau\in s_i}[P_s(r_\tau+1)\prod_{m\in s_{i/\tau}} P_n(r_m)]\cdot P(U\backslash s_i)}$$

$$= \frac{\frac{P_s(r_j+1)}{P_n(r_\tau)}\cdot\frac{P(k_{ij}|D)}{1-P(k_{ij}|D)}\cdot}{\sum_{\tau\in s_i}[\frac{P_s(r_\tau+1)}{P_n(r_\tau)}\cdot\frac{P(k_{i\tau}|D)}{1-P(k_{i\tau}|D)}]}$$

Then we can apply Gibbs sampling method to sample a mapping site for each multi-mapping reads using this posterior probability formula.

In each iteration, we go through all the multi-mapping reads and calculate the posterior probabilities of a multi-mapping read assigned to its candidate mapping sites, and sample one site to be the mapping position of this multi-mapping read in this iteration.

In the initialization step, we use the P_s, P_n fitted from canonical reads. However, after the initialization, each multi-mapping read can be assigned to only one site. In some sense, the multi-mapping reads become canonical reads, and we use the P_s, P_n fitted from weighted reads in the following iterations.

The optimization procedure will end once the whole system converges.

4.2.2 ChIP-seq Data Source and Data Processing. We downloaded human ChIP-seq datasets from ENCODE data portal [188]. Only single-end ChIP-seq raw reads (fastq file) with at least two isogenic replicates and without any treatments were used. We also downloaded the corresponding control ChIP-seq files and applied the same processing procedure to the control files.

ChIP-seq raw reads were mapped to human reference genome assembly GRCh37 (hg19) by Bowtie2 (V2.3.4; -k 11) [189]. Reads mapped to only one genome position are named 'canonical reads' while reads mapped to multiple positions are called 'multi-mapping reads'. Multi-mapping reads are restricted to be assigned to at most 11 genome positions by Bowtie2.

To control the ChIP-seq data quality, we also calculated the FRiP value [190]. If either of the isogenic replicates has FRiP smaller than 1%, this ChIP-seq dataset was removed.

Multi-mapping read assignment tools, RegisTER-ME (V1.0.0; --weight 0.2 –flanking 0.5site –max_iter 35) and CSEM(V2.4; default settings) [191], were applied to the read files (SAM file) respectively. Every multi-mapping read will be assigned to only one genome position after this procedure.

We used the SPP peak caller(v1.10.1; -npeak=300000) [192] to identify potential protein binding regions (peaks) and IDR(V2.0.3; --idr-threshold 0.02) [193] to filter high confident peaks. The peaks from canonical reads are regarded as 'canonical peaks' and newly discovered peaks after assigning the multi-mapping reads are named after the multi-mapping read assignment tools (RegisTER-ME peaks or CSEM peaks).

4.2.3 Processing of RNA-seq Data. To get expressed genes in specific cell lines, we downloaded gene-level expression data from ENCODE Portal for 7 cell lines (GM12878: ENCFF300QDV, ENCFF313RNF. K562: ENCFF186TXT, ENCFF728TIT. MCF-7:

ENCFF057ITQ, ENCFF456OYZ. HepG2: ENCFF974MUO, ENCFF649AHO. IMR-90: ENCFF480FTB, ENCFF588AJG. Liver: ENCFF804QWF, ENCFF418BVF. H1: ENCFF045NOY, ENCFF334LZL.). The expression values (TPM) from the 2 isogenic replicates were merged and genes with the lowest 10% expressions in the files were removed.

4.2.4 Promoter-enriched TFs Selection. To get the list of TFs more enriched in promoters than other genomic regions, we performed TF enrichment analysis in cell-type specific promoters and cell-type specific enhancers. Cell-type specific promoters were defined as +/- 2.5kb of expressed genes' RefSeq [194] transcription start sites and cell-type specific enhancers were downloaded from Roadmap Epigenomics and ENCODE data portal. Enrichment is calculated as the fraction of regions with target TF binding (canonical peaks). We down-sampled the same number of enhancers as the promoters for 1000 times, and calculated the enrichments of these 1000 enhancer sets. The TF enrichment in promoter regions was compared with these 1000 enrichments using one-tailed binomial test. TF ChIP-seq datasets with P-value < 0.001 were considered to be more enriched in promoter regions.

4.2.5 Processing of Chromatin interaction data. We obtained the chromatin interaction data used in this study from publicly available datasets. GM12878 Capture-C data and H1 Capture-C data was obtained from GSE86189 [195] using P-value < 0.1 and P-value < 0.05 as the corresponding filters. K562 ChIA-PET data was obtained from GSE33664 [196]. In long-range chromatin interaction analysis, these chromatin interaction datasets were further filtered. Only chromatin interactions in the datasets that anchored at Refseq transcription start sites were kept.

4.2.6 Paired-end ChIP-seq Analysis. We downloaded peak files of Paired-end ChIP-seq datasets from ENCODE Portal for 3 cell lines (GM12878: ENCFF568FPC, ENCFF360OXD, ENCFF809OOE, ENCFF431XRU, ENCFF356UKE, ENCFF006MIL. K562: ENCFF845UBO, ENCFF235UZG, ENCFF631IAC, ENCFF661CJD, ENCFF837QOK, ENCFF120MVT, ENCFF988FFD, ENCFF529GVQ, ENCFF168KBY, ENCFF231HJU, ENCFF906BNB, ENCFF401FXT, ENCFF067ZUO, ENCFF724TVS, ENCFF591UOR, ENCFF836FKF. HepG2: ENCFF304CJQ, ENCFF459HJZ, ENCFF324OAJ.). These peaks were used as the gold standard to examine the confidence of newly discovered peaks.

4.2.7 Processing of Genomic Domain Data. Genomic domain lists for GM12878 and K562 cell lines were obtained from GSE63525 [197]. By extending upstream and downstream a certain distance of each domain boundary, we got the domain boundary region lists for GM12878 and K562.

4.2.8 Gene Ontology (GO) Enrichment Analysis. GO enrichment analysis was performed using GOATOOLS (--pval=0.1) [198]. GOATOOLS is a computational tool to identify enriched biological themes of the input gene list against the background gene set. In this study, we used all the UniProt IDs in the GOATOOLS database as the background gene set.

In method comparison, we generated the genes in GO analysis by including the nearest expressed RefSeq genes within 50kb of the tool specific peaks. Q-values in cell-type specific Biological Pathways (BP) from RegisTER-ME peaks and CSEM peaks were extracted and were compared across all the ChIP-seq datasets in the cell line. In co-evolution analysis, we increased the expanding region to include enough genes for GO analysis. The nearest genes within 100kb of the peaks were added to the gene set. Since the size of the gene set was small, the q-values

for all the BPs were not significant. Therefore, we calculated the fraction of genes within a pathway and compared with the fraction of background gene sets. The pathways with the highest fold-change values were selected in the study. In TE-derived long-range interaction analysis, the gene set was derived from the target genes inferred from chromatin interaction data. We also used fold-change to select the top pathways for the study.

4.2.9 Motif Analysis. The genome sequences in motif analysis were derived from canonical peaks, RegisTER-ME peaks and RegisTER-ME competitor regions. RegisTER-ME competitor regions are defined as regions that compete with RegisTER-ME peaks for most of the multi-mapping reads. We extended upstream and downstream of the peak summit (RegisTER-ME competitor regions used the region middle point) to half of the mean canonical peak and got the sequences. FIMO (V5.0.5; --thresh 1.0E-4) [199] was used to scan these sequences to identify the positions of TF motifs within the sequences. The position weighted matrices of the TFs were downloaded from HOCOMOCO database (Version 11) [200]. The TF motifs that did not overlap the ChIP-seq target TF's motif positions were ranked on their occurrences among all the canonical peak sequences. Top TF motifs were selected as the potential co-factors.

In co-evolution analysis, we also used FIMO to scan the TE sequences. Scores from FIMO results were used to describe the motif matching level.

4.2.10 Protein-Protein Interaction Analysis. The protein-protein interactions (PPI) were downloaded from STRING database (Version 11.0) [201]. We filtered the PPIs using experiment score > 0.

4.2.11 Gene Expression Correlation Analysis. We downloaded the normalized gene expression file from ENCODE data portal and Roadmap Epigenomics. Pearson's correlation coefficient between every two genes was calculated using their expression values across the 56 cell lines.

4.2.12 Genome Annotation. The gene position file of hg19 was downloaded from the RefSeq database. We extended upstream and downstream 2.5 kb of each gene's transcription start site (TSS) to obtain the promoter regions. The genic regions were defined as 10 kb upstream of a gene's TSS to downstream 10 kb of a gene's transcription end site (TES). We obtained segmental duplication data (Segmental Dups track), repeat annotations (RepeatMasker track) and centromere and telomere data (Gap track) from the UCSC Table Browser [202]. Only TEs in 'LINE', 'LINE?', 'LTR', 'LTR?', 'SINE', 'SINE?' families were kept. Per-centromere was defined as upstream 1 MB to the downstream 1 MB of a centromere. And peri-telomere was defined as the first 2 MB or the last 2 MB of a chromosome. Repressive regions were retrieved from Roadmap Epigenomics**.**

4.2.13 Enrichment of TF Binding Sites in TEs. To compare the contribution of different TE families in TF binding sites, we calculated an enrichment value of the TF binding sites in each TE family. The enrichment value was defined as the fraction of TF binding sites derived from a TE family over this TE family's length fraction among the whole genome.

The TE position file was downloaded from UCSC Genome Browser database. If the summit region (+/- 100bp of the peak summit) of a TF ChIP-seq peak overlapped a TE, then we considered it as TE-derived.

4.2.14 Co-evolution Analysis. The consensus sequences for all the TE subfamilies were retrieved from the Dfam database [203]. We directly used the 'milliDiv' value from the repeat annotation data as an estimation of the TE age.

TEs overlapping ChIP-seq peak summits were considered to contribute to TF binding sites. We extracted the sequences of these TEs and used FIMO to detect the motif positions of target TF and the potential co-factors (see Motif Analysis in Method). TEs from the same sub-family and contributing to TF binding sites were extracted. We then did multiple sequence alignment on these TEs and their corresponding consensus sequence using MUSCLE (V3.8.31; default setting) [204]. The multiple sequence alignment results were visualized on Jalview (V2.11.1.4) [205] and the positions with target TF motif or co-factor motif were colored based on nucleotides.

4.2.15 Co-regulation Analysis. We obtained TE-derived long-range interactions by overlapping TE-derived TF binding sites and cell-type specific chromatin interaction data (see Processing of Chromatin Interaction Data in Method). If the two anchors of a chromatin interaction were located within a TF binding site and at a gene TSS respectively, we considered it as one TE-derived long-range interaction.

For each gene, we calculated the fraction of enhancers from different TE families. Then we averaged these enhancer fractions across all the genes for each TE family respectively. We called these average numbers as the contributions of the corresponding TE families in co-regulation analysis. Genes regulated by only one enhancer were not considered in the calculation, but the enhancers were included in later permutation to generate a shuffled background.

We shuffled the TE labels of all the enhancers 1000 times, and used the contributions of a specific TE family in these 1000 shuffles as the null distribution. Kernel density estimation was

used to estimate the probability of observing a contribution value equal or larger than the true contribution value under this null distribution.

4.2.16 SNP Enrichment Analysis. We downloaded three types of SNPs publicly available datasets: GWAS SNPs, eQTLs and TCGA mutations. GWAS SNPs were retrieved from GWAS Catalog [206] and somatic mutations in acute myeloid leukaemia (LAML) were downloaded from TCGA program. We combined eQTLs from 3 studies [207-209]. These eQTLs were from either blood tissues or lymphoblastoid and thus related to GM12878 and K562 cell lines.

We first calculated the fraction of TE-derived RegisTER-ME peaks with the SNPs across all the ChIP-seq datasets in GM12878 and K562 cell lines. Then we generated permutation peaks for each dataset using the 'shuffle' function from bedtools (V2.27.1; default settings) [210] and calculated the fraction of permutation peaks with the SNPs. The permutation was done 1000 times. We did the two-tailed Wilcoxon signed-rank test to compare the SNP enrichment in RegisTER-ME peaks with the enrichment in the permutation peaks from the same ChIP-seq dataset.

4.2.17 Statistical Analysis and Figure Generation. We did all the statistical analysis and associated figures by Python. Examples were generated with UCSC Genome Browser screenshots [211].

4.3 RESULTS

4.3.1 ChIP-seq multi-mapping reads alignment based on neighborhood read counts and flanking sequences

Previous studies [184,185] have proved the local genomic ChIP-seq read context (neighborhood reads) can help to guide the allocation of multi-mapping reads. However, in some

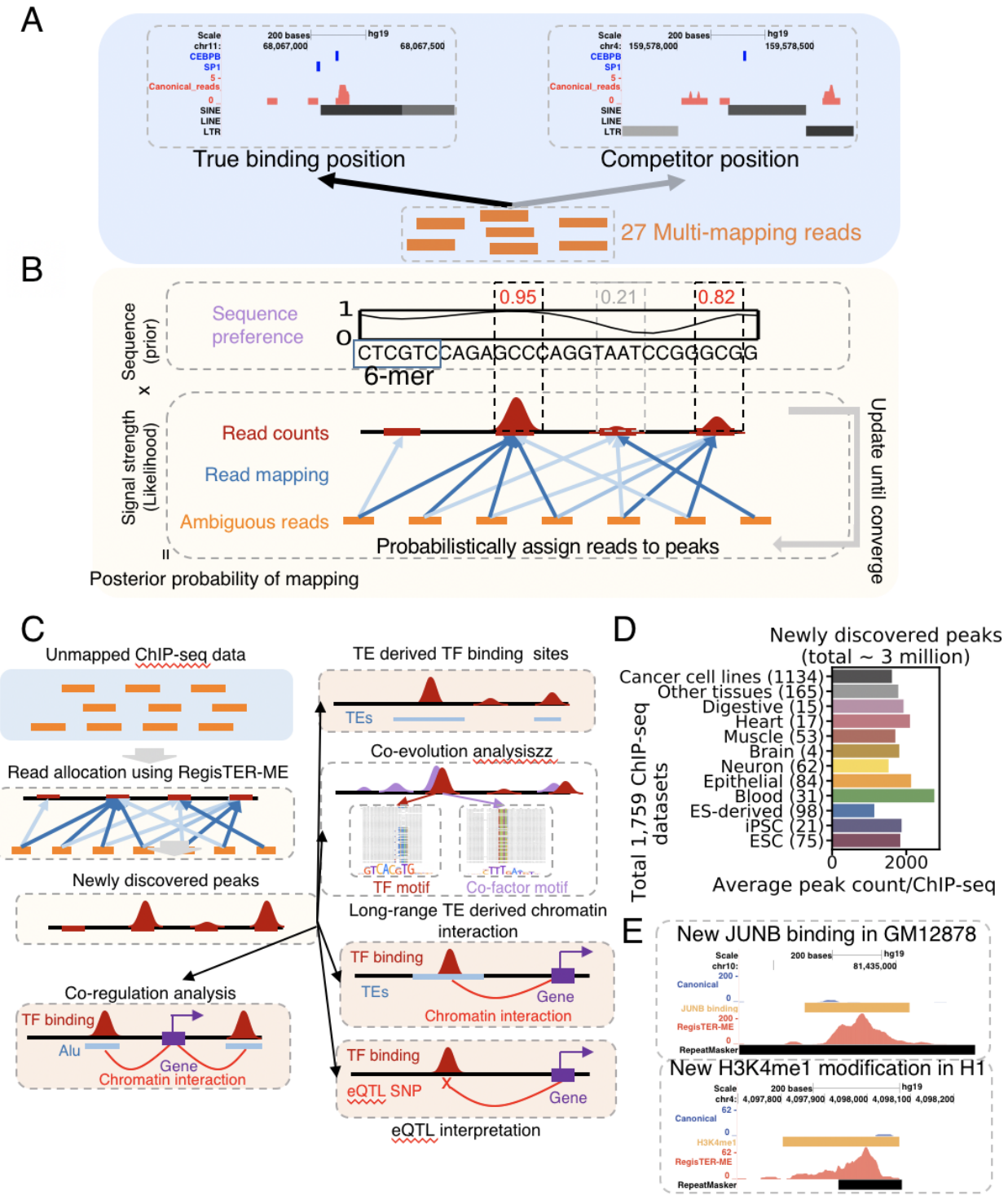


Figure 4.1. ChIP-seq multi-mapping reads alignment based on neighborhood read counts and flanking sequences. (A) A real example of a multi-mapping read from a CEBPB ChIP-seq in UCSC genome browser. This read can be mapped to 2 positions with the same read count but different flanking sequences. (B) The schematic figure of RegisTER-ME. (C) The pipeline of downstream analysis. (D) The summary of newly discovered peaks for ChIP-seq data in ENCODE project. (E) Real examples of newly discovered peaks after applying RegisTER-ME.

cases, the potential mapping positions of a multi-mapping read share similar local genomic read contexts. Fig 4.1A is a real example from a CEBPB ChIP-seq dataset. There are totally 27 multi-mapping reads which can be mapped to two regions in the genome. Since both regions are within TEs, the canonical read in these two regions are rare. Therefore, previous computational algorithms may not be able to distinguish the true binding region.

Inspired by the binding activity of transcription factors and the coordinated binding of cofactors [212,213], we exanimated the sequences around the exact mapping coordinates. The binding position in the left (Fig 4.1A) has higher probability to be a true binding position, as it is near a gene promoter. In this true binding position, we observed both TF motif (CEBPB) and the motif of SP1, while there is only CEBPB motif in the other mapping position. CEBPB is shown to cross-talk with SP1 to modulate downstream gene production [214]. This example illustrated the benefits of integrating flanking sequence information when there are similar neighborhood reads in competing binding regions. The TF motifs and co-factor motifs in the flanking sequences of the true binding position should help us to discriminate it from other regions.

We then designed a Bayesian model, RegisTER-ME (METHOD), which considers both local genomic ChIP-seq read context and the flanking sequence, to compute the probability for each multi-mapping read mapping to any potential binding positions (Fig 4.1B). Since there are millions of multi-mapping reads in a ChIP-seq dataset, we also took use of the Gibbs sampling algorithm to optimize the mappings. The combinational use of ChIP-seq read context and flanking sequence boosted the performance of the model an allowed the model to converge with only a few interactions (Fig C.1).

We applied RegisTER-ME to the whole ENCODE database (Fig 4.1D), and achieve around 3 million new binding positions (RegisTER-ME peaks). Fig 4.1E shows real examples in 2 ChIP-seq datasets (JUNB in GM12878 and H3K4me1 in H1). The genomic regions in the examples are near gene promoters and are very possible to be true binding positions. However,

these regions are within TEs, making them to have low mapping abilities [215]. RegisTER-ME overcame the limitation and re-discovered the binding positions. The large repository of newly discovered TE-derived TF-binding positions allows us to have a comprehensive study of the activity of TEs in regulation network evolution (Fig 4.1C).

4.3.2 Boosted performance through sequence information integration

We conducted four method comparisons with an existing read mapping algorithm, CSEM [5], to systematically exanimate the performance of RegisTER-ME in accurately allocating the multi-mapping ChIP-seq reads. We applied both methods to ChIP-seq datasets and called the newly discovered peaks RegisTER-ME peaks or CSEM peaks respectively.

(i) Promoter region enrichment

Some TFs are especially more enriched in promoter regions [216]. For those TFs, we exanimated the promoter enrichment of RegisTER-ME peaks and CSEM peaks. Fig 4.2A shows the enrichment differences of these algorithms across promoter enriched TF ChIP-seq datasets in K562. RegisTER-ME peaks showed higher percentage in promoter regions in the majority of the datasets. The results in other cell lines (Fig C.2) reflected a similar pattern.

(ii) Domain boundary enrichment of CTCF peaks

CTCF is an insulator-binding protein and previous study identified significant binding of CTCFs in domain boundaries [217]. Therefore, we used domain boundary enrichment as a standard to determine CTCF peaks quality. Fig 4.2B shows the domain boundary enrichments for CTCF peaks in K562 and GM12878. In all these CTCF ChIP-seq datasets, RegisTER-ME peaks always have higher fraction to near domain boundaries.

(iii) Chromatin interaction activity enrichment

Chromatin interaction data connects regulatory elements to the target genes [218]. We

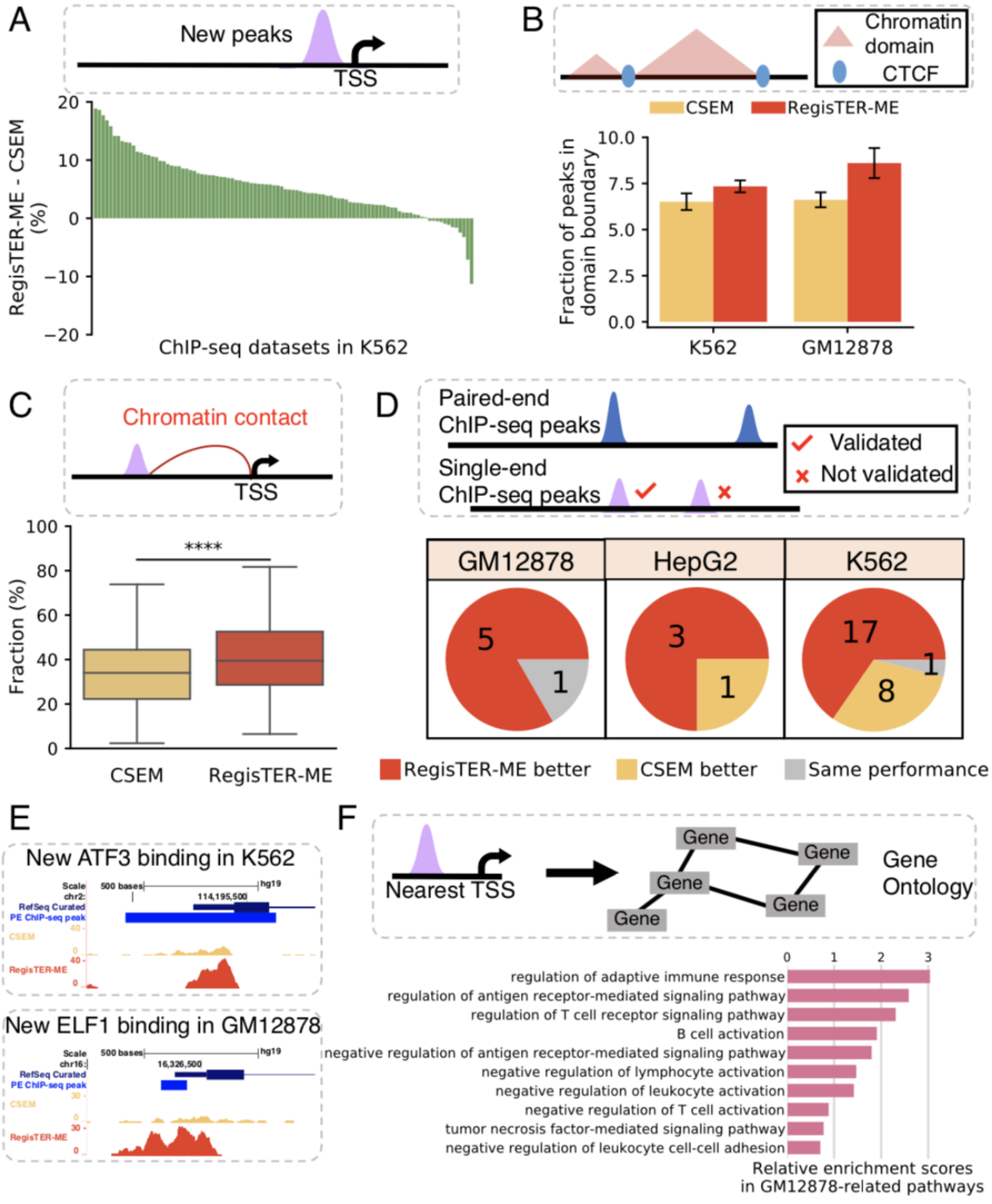


Figure 4. 2. Boosted performance through sequence information integration. (A) Method comparison using fraction of peaks within promoter regions. (B) Method comparison using fraction of CTCF peaks in domain boudaries. (C) Method comparison using fraction of peaks involved in enhancer-promoter interactions. (D) Method comparison using fraction of peaks validated by PE ChIP-seq data. (E) Real examples of newly discovered peaks from RegisTER-

ME but missed by CSEM. (F) Method comparison using enriched pathways from the nearest genes.

extracted anchors from K562 ChIA-PET data as the potential TF binding positions to check the overlapping of newly discovered peaks with these potential bindings. Fig 4.2C shows the overlapping fraction of RegisTER-ME peaks or CSEM peaks in K562 ChIP-seq datasets with chromatin interaction anchors. A statistical test indicated there was a significant higher overlapping fraction in RegisTER-ME peaks than CSEM peaks. Similar results were observed in other cell lines using other type of chromatin interaction data (Fig C.3).

(iv) Paired-end (PE) ChIP-seq validation

PE ChIP-seq technology sequences both ends of a DNA fragments and captures more information than SE ChIP-seq data [219]. We considered the peaks from PE ChIP-seq as gold standards to determine if a peak in SE ChIP-seq has high confidence or not. Fig 4.2D shows the comparison results between RegisTER-ME and CSEM. In all 3 cell lines, there are more ChIP-seq datasets where RegisTER-ME has higher fraction of high confidence peaks.

Except for the overall performance evaluation, we also dig into real examples. Fig 4.2E shows genomic regions with high TF binding confidence. RegisTER-ME was able to identify them as true binding positions while CSEM failed.

In addition, pathway analysis for nearby genes of newly discovered peaks shows RegisTER-ME peaks tend to be around genes with cell type specific functions. GM12878 is a lymphoblastoid cell line. Pathways related to GM12878 involve in multiple immune functions. We checked the enrichment of nearby genes from RegisTER-ME peaks and CSEM peaks, and found RegisTER-ME peaks show higher fraction to near genes involve in these pathways (Fig 4.2F).

Both overall performance comparisons and examples checking indicate a better performance of

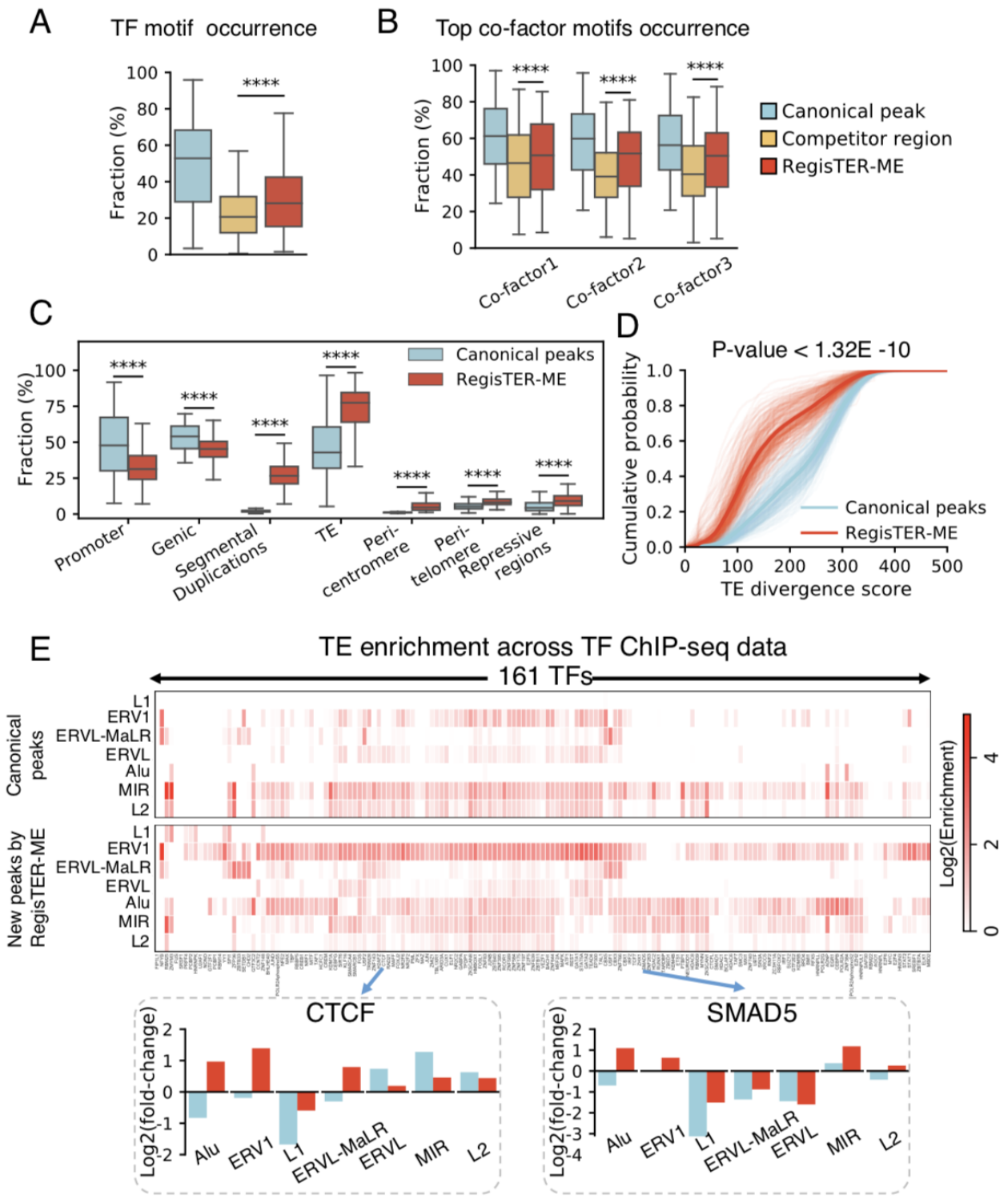


Figure 4.3. Transposable element (TE) activities from RegisTER-ME peaks. (A) Fractions of peaks with the corresponding TF motifs for canonical peaks, RegisTER-ME peaks and the competitor regions of these RegisTER-ME peaks. (B) Fractions of peaks with the top 3 co-factor motifs from canonical peaks for the 3 groups of peaks. (C) Fractions of peaks overlapping

different genome annotations for canonical peaks and RegisTER-ME peaks. (D) TE divergence score cumulative distributions for canonical peaks and RegisTER-ME peaks across K562 ChIP-seq data. (E) Enrichment of different TE families across different TF ChIP-seq data. Bar plots show the enrichment of different TE families in CTCF ChIP-seq data (left panel) and SMAD5 ChIP-seq data (right panel).

RegisTER-ME algorithm over CSEM.

4.3.3 Transposable element (TE) activities from RegisTER-ME peaks

The integration of flanking sequences boosted the performance of RegisTER-ME in capturing the true binding positions. To explore the benefits of using flanking sequences and reveal the potential mechanism of TF binding, we checked the occurrence of TF motifs and co-factor motifs in three groups of peaks: canonical peaks, RegisTER-ME peaks and the competitor regions of RegisTER-ME peaks. Fig 4.3A and 4.3B shows the overall occurrence of TF motifs or co-factor motifs in these three groups of peaks across all K562 ChIP-seq datasets. Significant higher fractions of occurrences can be observed through the figures or through statistical tests, which illustrates the ability of RegisTER-ME in capturing useful motif information from the flanking sequences.

We also checked the genome annotations of canonical peaks and RegisTER-ME peaks (Fig 4.3C). The result shows that RegisTER-ME peaks tend to locate in repeat regions, such as TEs, where the ChIP-seq read mappability is low. Digging deeper into those TE-derived peaks, we found RegisTER-ME peaks were originated from younger TEs (Fig 4.3D), indicating the ability of RegisTER-ME in capturing the recent activities of TEs in regulatory network evolution.

Figure 4.3E shows the enrichments of TE families in different TF ChIP-seq datasets. Previous studies uncovered a pervasive phenomenon where TEs expanded the TF regulatory networks [220,221]. Once there is a TF binding site in a TE, it can be spread to the genome through the copy and insertion of the TE. The enrichment heatmaps provided a good source to

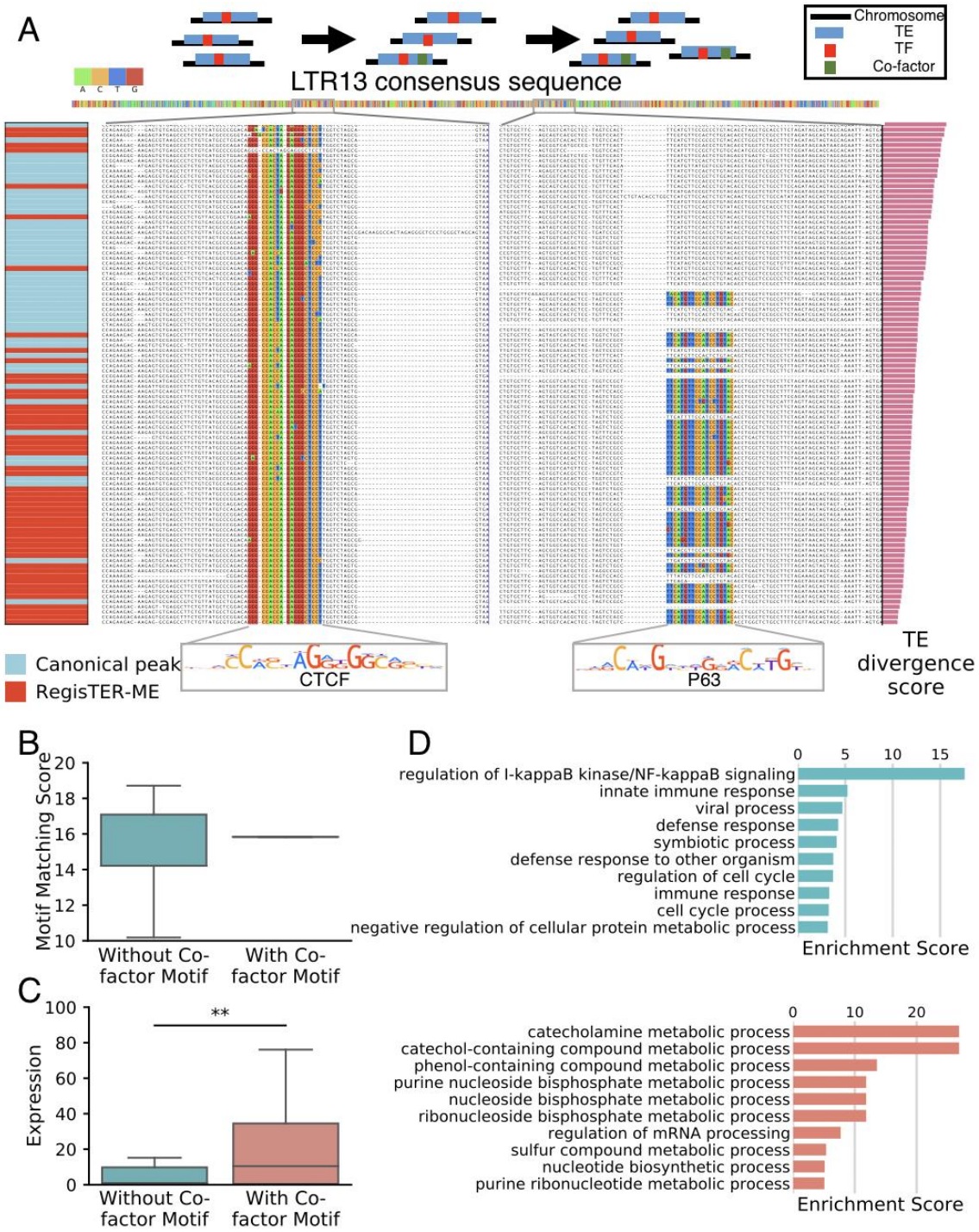


Figure 4.4. Co-evolution of transcription factors and co-factors in TE-derived TF binding sites. (A) A sequence alignment for LTR13-derived CTCF binding regions. The left panel shows the identification of each binding site (blue: from canonical peaks, red: from RegisTER-ME peaks). The positions with motifs in the sequence alignment result are colored based on the nucleotides. The right panel shows the divergence score of each LTR13 sequence. (B) The motif matching scores for CTCF binding sites. The CTCF binding sites are divided into 2 groups: before having p63 motifs in the flanking regions and after having p63 motifs in the flanking

regions. (C) The expression level of the nearest genes for the 2 groups of CTCF binding sites. (D) Enriched pathways for the 2 groups of nearest genes.

study different waves of TFBS creation from TEs in evolution. Using CTCF as an example, the enrichments from canonical peaks illustrated that long before, TEs in MIR family or L2 family contributed to the creation of CTCF binding sites, while in recent time, Alu elements and ERV1 elements began to participate in this creation procedure. The pattern in SMAD5 binding sites are quite different, as the majority of the TE creation activities were observed from RegisTER-ME peaks.

4.3.4 Co-evolution of transcription factors and co-factors in TE-derived TF binding sites

Although TEs can expand a TF's binding sites, not all the TEs of the same family (or sub-family) contain this TF binding site. Some TF binding sites are fixed while others decay with evolution. We then raise the question, what is the potential underlying mechanism of this phenomenon? To answer this question, we first collected peaks (both canonical and RegisTER-ME) derived from the same TE subfamily. By scanning the sequences with all the motifs from public available database [200], we identified a list of TF - co-factor pairs in a few ChIP-seq datasets (Fig 4.4A and Fig C.4-C.9).

In Fig 4.4A, we profiled the motifs in LTR13-derived peaks and sorted the peaks on TE divergence score in a CTCF ChIP-seq in HepG2. While the majority of the peaks have CTCF motifs, p63 motif appeared only in peaks with smaller TE divergence scores. After the occurrence of this p63 motif in the peak, we observed a fixed high motif matching score of the TF, compared to previous fluctuated scores (Fig 4.4B), and an overall increased expression level in the nearest genes (Fig 4.4C). We did not have these observations for p53 motif or p73 motif.

Previous studies showed p63 cooperates with CTCF to modulate chromatin architecture

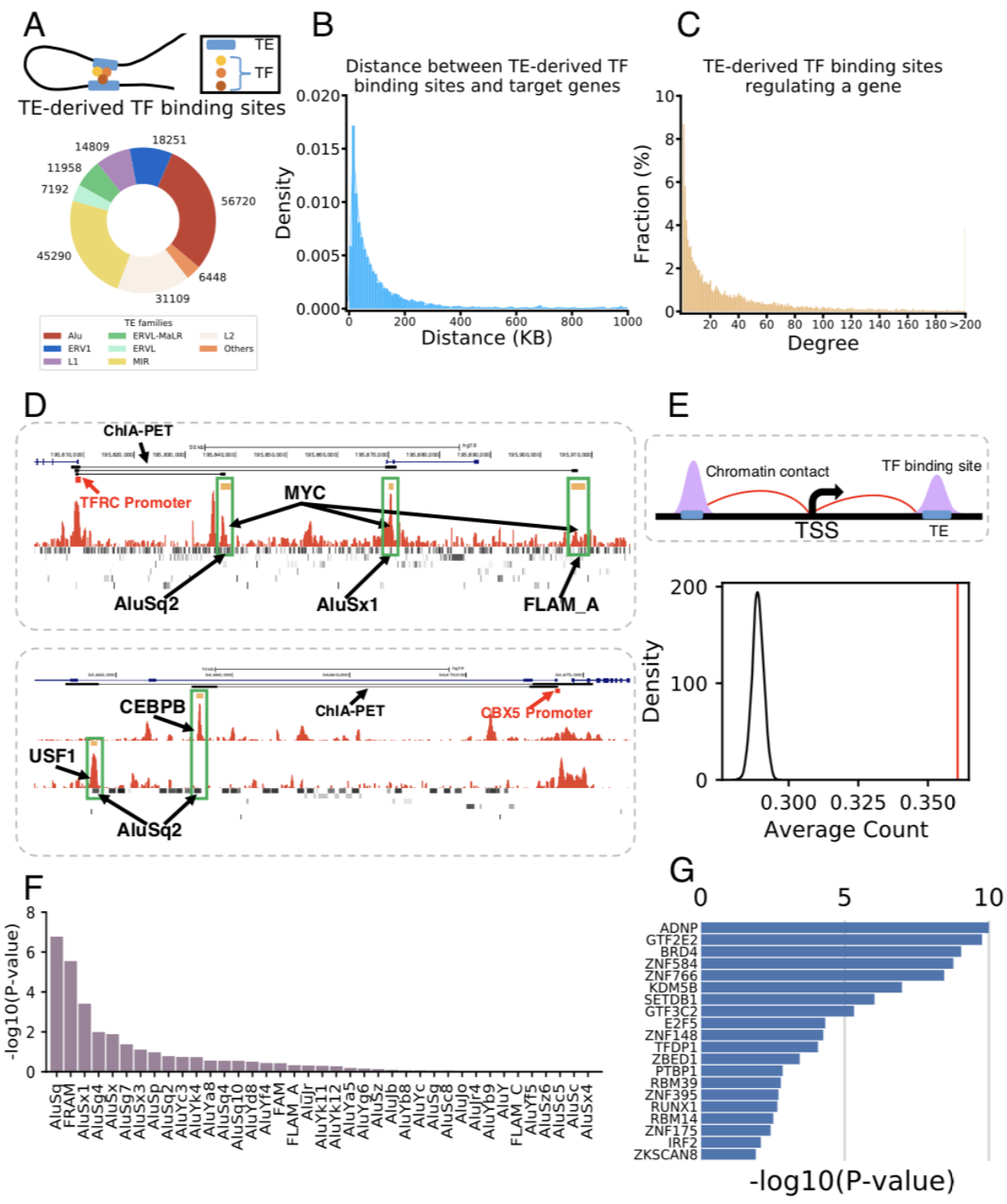


Figure 4.5. Co-regulation of TE-derived TF binding sites in long-range chromatin interactions. (A) Number of TEs participating in long-range interactions. (B) Distance distribution for long-range interactions. (C) Degree distribution for long-range interactions. (D) 2 examples of Alu-derived co-regulation. The example in the upper panel shows 3 Alu-derived MYC binding sites that regulate the TFRC gene. The example in the lower panel shows 2 Alu-

derived binding sites (one for CEBPB and one for USF1) that regulate the CBX5 gene. (E) Average number of Alu-derived enhancers that regulate a gene among shuffled long-range interactions. (F) –log10(P-value) of different Alu sub-families in Alu-derived long-range interaction co-regulations. (G) –log10(P-value) of top 20 TFs in Alu-derived long-range interaction co-regulations.

[222] and p63 itself plays an important role in regulating metabolism of cancer cells [223], therefore we conducted pathway analysis for the nearest genes before and after the appearance of p63 motif to explore whether p63 is a potential co-factor of CTCF and the occurrence of p63 will result in functional results. In Fig 4.4D, the two groups of nearest genes showed totally different gene ontology enrichments. While the nearest genes of peaks without p63 motif tend to participate in immune-related pathways, the nearest genes of peaks with p63 motif are more enriched in metabolic processes. We thus came to the conclusion that the occurrence of co-factor motifs in the flanking region of the TF binding site may stabilize the TF motif and increase the activity of this TF binding site in cell-type specific functions.

4.3.5 Co-regulation of TE-derived TF binding sites in long-range chromatin interactions

Long-range chromatin interactions play an essential role in regulatory networks [224]. To uncover the contribution of transposable elements in long-range interaction, we analyzed TE-derived TF binding sites, including those from canonical pipeline or from RegisTER-ME application. Fig 4.5A illustrates the scales of different TE families that participate in long-range interaction. 3 TE families, Alu, MIR and L2, contribute the majority of the TF binding sites in long-range interaction. Next, we studied the distance distribution (Fig 4.5B) and found most of TE-derived TF binding sites are within 100kb of their target genes. Degree analysis showed many genes can be regulated by multiple TE-derived TF binding sites (Fig 4.5C), with most of the genes regulated by less than 100 TF binding sites.

We checked those genes regulated by multiple TE-derived TF binding sites, especially

those TF binding sites from the same TE family. These TE-derived TF binding sites and gene pairs can be classified into two groups: binding sites for the same TF and binding sites for different TFs. We showed real examples for these two groups respectively in Fig 4.5D. In Fig 4.5D. The example in the top shows three MYC binding sites regulating the same gene, TFRC. The example at the bottom shows two TF binding sites, CEBPB and USF1, regulating CBX5 gene. We called the regulation for TF binding sites derived from the same TE family co-regulation. To test which TE family engages more in co-regulation events, we conducted a perturbation on the TE label of each TE-derived TF binding site. The results for each TE family is listed in Fig C.10. Among all the TE families, only Alu shows higher participation in co-regulation than the shuffled backgrounds in K562 (Fig 4.5E). Further subdivision on Alu subfamilies and TF ChIP-seq datasets prioritized AluSg, FRAM, AluSx1 and ADNP, GTF2E2, BRD4 to be the top subfamilies or TFs involved in co-regulation activities.

4.3.6 eQTL interpretation from the identification of RegisTER-ME peaks

The application of RegisTER-ME to the whole ENCODE database resulted in around three million newly discovered peaks. We further explored the usage of these RegisTER-ME peaks in interpreting eQTL mechanisms. We first checked the enrichment of functional elements in RegisTER-ME peaks compared to shuffled backgrounds. Fig 4.6A shows RegisTER-ME peaks from K562 ChIP-seq datasets were more enriched with eQTLs in blood-related diseases, leukemia SNPs and GWAS SNPs.

In addition, we discovered a few real examples of explaining eQTL using RegisTER-ME peaks. In Fig 4.6B, there is an SNP (rs1165696) from GEUVADIS project. This SNP is an eQTL located in a TE in blood cells, with target gene C12orf73 (P-value=1.27539×10^{-10}). However, the mechanism of how this eQTL affects C12orf73 expression is understudied. By applying RegisTER-ME, we identified a CEBPB binding site around this eQTL. Since CEBPB has shown

to be involves in immune and inflammatory responses [225], a SNP within it may block this enhancer-gene interaction and thus resulting in target gene expression change. This provides a potential mechanical explanation for the functional results of this SNP.

In total, we discovered 10 cases (Fig C.11-C.19) in K562 ChIP-seq datasets where an eQTL was understudied due to their location in repeat regions.

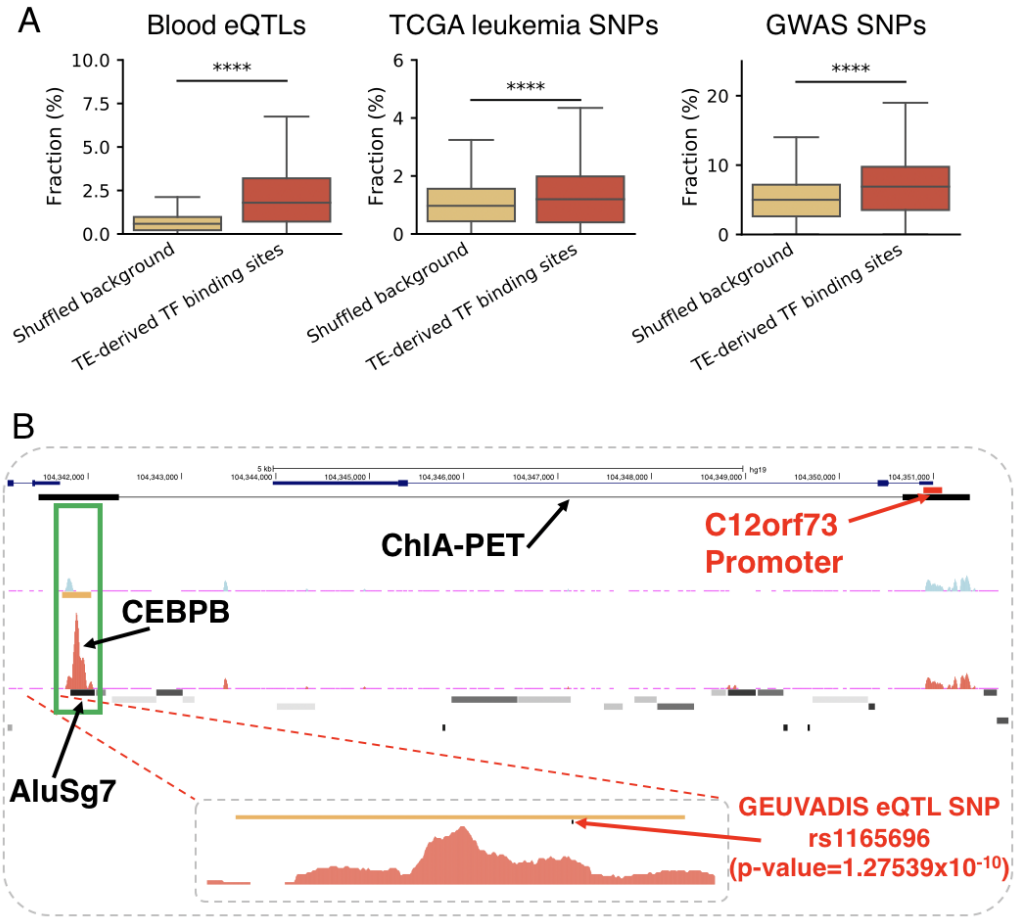


Figure 4.6. eQTL mechanism revealing from the identification of RegisTER-ME peaks. (A) Compared to shuffled background, newly discovered peaks from RegisTER-ME have higher fraction to overlap 3 types of SNPs: eQTLs in blood diseases (left), SNPs in leukemia from TCGA (middle), GWAS SNPs (right). (B) A real example shows how to utilize newly discovered peak from RegisTER-ME to interpret eQTL mechanism.

4.4 DISCUSSION

In this study, we developed a statistical model, RegisTER-ME, to allocate multi-mapping reads in ChIP-seq. This model not only takes use of the local genomic context, but also integrates flanking sequence information to help discriminate true binding regions in repetitive regions with low mappabilities. The fractions of RegisTER-ME peaks with TF motifs and top co-factor motifs indicate that the flanking sequence information used in the model can capture the TF binding activities. In addition, the inclusion of sequence information allows for the quick convergence of the model. We proved the superior performance of RegisTER-ME through conducting four overall method comparisons with an existing read mapping algorithm, CSEM [5], as well as real example checking and pathway analysis.

We applied RegisTER-ME to the whole ENCODE database and achieved around three million newly discovered ChIP-seq peaks. These peaks tend to locate in repeat regions, such as TEs, and serve as a rich resource to study the evolution of young TEs in regulatory network evolution.

Long-range chromatin interactions play an essential role in regulatory networks [224]. In the study of long-range chromatin interactions from transposable elements, we defined a term, co-regulation, which refers to the regulation events from TF binding sites derived from the same TE family. Perturbation tests shows Alu is the only TE family in K562 that is enriched in the co-regulation activities. Next step, we will expand our analysis to the collaborations among TE families in co-regulation, and the functional results from these collaborations.

Finally, we explored the capacity of RegisTER-ME in revealing eQTL mechanism. Some eQTLs are located in non-coding regions, where the mappabilities are low. Traditional ways of processing ChIP-seq data often result in peak missing in such regions. By applying RegisTER-ME, we re-discovered the protein binding events in these regions and were able to uncover the mechanism of how eQTLs affect target gene expression. In Fig 4.6B, we discussed an example,

where an eQTL is within a missing CEBPB peak to affect the expression of C12orf73 gene through long-range interactions. In total, we discovered 10 cases in K562 ChIP-seq datasets where an eQTL was understudied due to their locations in repeat regions.

CHAPTER 5

FUTURE DIRECTIONS

The protein-protein interactions between TFs on enhancers or promoters have been found to participate in the process of long-range chromatin interactions and mediate distal enhancers to the proximity of target gene promoters. There are in-direct PPIs between TFs on enhancers and TFs on promoters. Another group of TF may not directly bind to genomic regions. They have PPIs with TFs on both enhancers and promoters, thus mediating in-direct interactions between TFs on enhancers and TFs on promoters. These in-direct interactions can be partially caught by TF modules in ProTECT. Next step, we will build a weighted multi-step in-direct PPI network to capture in-direct TF interactions and further improve the performance of ProTECT.

Although unsupervised enhancer-promoter prediction algorithms demonstrate an overall worse performance than supervised methods, their performances are not constrained by experimental techniques. We are considering to adapt the features in ProTECT into an unsupervised model. In this way, the new ProTECT model can overcome the high false positive rate or high false negative rate in experimental techniques and the predicted enhancer-promoter pairs will have reduced bias, while we can still gain mechanistic insights on how specific long-range chromatin interactions are established.

PE ChIP-seq technology sequences both ends of a DNA fragments and captures more information than SE ChIP-seq. With the decrease of PE ChIP-seq cost, more PE ChIP-seq datasets are now available. Since multi-mapping reads still account for a substantial part of PE ChIP-seq, there is an urgent need to improve RegieTER-ME so that it can be applied to PE ChIP-seq.

In the study of long-range chromatin interactions from transposable elements, we defined a term, co-regulation, which refers to the collaborative regulation events from TE-derived TF binding sites. Enriched co-regulations were observed from Alu-derived enhancers. We plan to expand our analysis to the co-regulations among enhancers from different TE families, to explore the functional outcome from collaborations among TE families.

APPENDICES

APPENDIX A

Supplementary materials for Chapter 2

Table A.1. List of Differentially expressed genes (control vs DHT)

ID	Gene name	log2(FoldChange)	padj
ENSMUSG00000040026	Saa3	3.232645808	6.66E-11
ENSMUSG00000057465	Saa2	2.948129216	1.31E-07
ENSMUSG00000097767	Miat	2.791957628	4.51E-20
ENSMUSG00000044254	Pcsk9	2.531884636	8.37E-13
ENSMUSG00000026227	2810459M11Rik	2.366871294	3.33E-07
ENSMUSG00000040627	Aicda	2.335567409	6.75E-40
ENSMUSG00000102224	4930447F24Rik	2.302135849	4.11E-06
ENSMUSG00000066438	Plekhd1	2.29319525	3.85E-24
ENSMUSG00000048572	Tmem252	2.274213716	2.08E-13
ENSMUSG00000022367	Has2	2.253576087	8.32E-11
ENSMUSG00000040181	Fmo1	2.118140893	3.91E-05
ENSMUSG00000025491	Ifitm1	2.083852535	3.64E-09
ENSMUSG00000015090	Ptgds	2.074159236	7.31E-17

Table A.1 (cont'd)

ENSMUSG00000033377	Palmd	2.047156175	4.44E-06
ENSMUSG00000040170	Fmo2	1.99665727	1.18E-16
ENSMUSG00000048108	Tmem72	1.921339475	2.98E-18
ENSMUSG00000041731	Pgm5	1.902586561	8.39E-26
ENSMUSG00000040276	Pacsin1	1.886299113	5.05E-15
ENSMUSG00000025194	Abcc2	1.883052665	0.000718345
ENSMUSG00000051727	Kctd14	1.863800792	5.11E-06
ENSMUSG00000029275	Gfi1	1.861459796	4.85E-10
ENSMUSG00000059852	Kcng2	1.827329868	0.001091998
ENSMUSG00000040732	Erg	1.787673692	0.001345005
ENSMUSG00000036067	Slc2a6	1.777850079	5.22E-05
ENSMUSG00000041831	Sytl3	1.74518592	0.001483168
ENSMUSG00000043811	Rtn4r	1.737659814	0.000512981
ENSMUSG00000030228	Pik3c2g	1.684749791	4.60E-07
ENSMUSG00000037071	Scd1	1.66702247	6.96E-23
ENSMUSG00000020787	P2rx1	1.661067032	0.013374431

Table A.1 (cont'd)

ENSMUSG00000001435	Col18a1	1.650641718	9.68E-18
ENSMUSG00000028753	Vwa5b1	1.640786008	0.00551037
ENSMUSG00000026535	Ifi202b	1.601341035	0.000632483
ENSMUSG00000053469	Tg	1.585868099	2.12E-08
ENSMUSG00000064310	Zpld1	1.580805284	5.34E-05
ENSMUSG00000025020	Slit1	1.56004132	1.31E-23
ENSMUSG00000020866	Cacna1g	1.538872791	0.000268186
ENSMUSG00000024124	Prss30	1.49510662	9.26E-05
ENSMUSG00000057615	Ldoc1	1.467712441	0.007714176
ENSMUSG00000026959	Grin1	1.451093966	0.000327071
ENSMUSG00000024030	Abcg1	1.450491667	5.79E-30
ENSMUSG00000021835	Bmp4	1.445764915	1.68E-08
ENSMUSG00000029371	Cxcl5	1.44348941	0.000309589
ENSMUSG00000045281	Gpr20	1.431277922	0.000405488
ENSMUSG00000003617	Cp	1.429595418	1.47E-18
ENSMUSG00000075270	Pde11a	1.425886484	2.29E-07

Table A.1 (cont'd)

ENSMUSG00000009378	Slc16a12	1.419791994	0.02934806
ENSMUSG00000035493	Tgfbi	1.412417685	6.47E-12
ENSMUSG00000033576	Apol6	1.407447859	1.34E-18
ENSMUSG00000015950	Ncf1	1.39897161	0.005941624
ENSMUSG00000031636	Pdlim3	1.396592218	0.000477026
ENSMUSG00000035226	Rims4	1.387428306	5.71E-52
ENSMUSG00000090610	Gm3571	1.385630311	0.038014637
ENSMUSG00000059743	Fdps	1.382081539	5.14E-66
ENSMUSG00000000197	Nalcn	1.379624502	0.001284748
ENSMUSG00000020010	Vnn3	1.368341677	0.011805643
ENSMUSG00000027442	Cst8	1.362656254	0.019111214
ENSMUSG00000020264	Slc36a2	1.362423795	1.64E-14
ENSMUSG00000034452	Slc24a1	1.349514808	0.018595761
ENSMUSG00000046182	Gsg1l	1.345261258	1.06E-27
ENSMUSG00000060807	Serpina6	1.339428344	0.047426531
ENSMUSG00000014846	Tppp3	1.334558932	1.22E-07

Table A.1 (cont'd)

ENSMUSG00000042429	Adora1	1.33333995	2.66E-62
ENSMUSG00000024799	Tm7sf2	1.320636003	7.02E-47
ENSMUSG00000022309	Angpt1	1.313581442	7.21E-23
ENSMUSG00000023832	Acat2	1.307843953	9.65E-99
ENSMUSG00000031349	Nsdhl	1.300080882	1.64E-64
ENSMUSG00000004031	Brinp2	1.290549003	2.82E-25
ENSMUSG00000097471	5830432E09Rik	1.281348669	0.003149409
ENSMUSG00000030041	M1ap	1.26800181	9.89E-06
ENSMUSG00000027500	Stmn2	1.253686327	0.003644936
ENSMUSG00000074768	Bhmt	1.252135434	4.53E-05
ENSMUSG00000108022	Gm7298	1.226191284	0.028326201
ENSMUSG00000031170	Slc38a5	1.226148147	5.52E-33
ENSMUSG00000044716	Dok7	1.219641601	2.66E-28
ENSMUSG00000038295	Atg9b	1.216862215	1.68E-09
ENSMUSG00000020388	Pdlim4	1.21409607	4.53E-38
ENSMUSG00000031740	Mmp2	1.211829905	3.09E-07

Table A.1 (cont'd)

ENSMUSG00000099884	Gm8204	1.198275608	0.042096463
ENSMUSG00000032327	Stra6	1.179022794	9.06E-12
ENSMUSG00000040907	Atp1a3	1.177776882	2.08E-44
ENSMUSG00000034353	Ramp1	1.172786352	1.65E-23
ENSMUSG00000027460	Angpt4	1.169968692	9.66E-11
ENSMUSG00000029121	Crmp1	1.169663257	0.040245001
ENSMUSG00000078627	Marchf10	1.167228441	0.027691879
ENSMUSG00000026255	Efhd1	1.164338295	3.71E-20
ENSMUSG00000020262	Adarb1	1.164202495	4.65E-41
ENSMUSG00000048022	Tmem229a	1.150303917	0.000111156
ENSMUSG00000109127	Gm31135	1.148494426	0.023103526
ENSMUSG00000029168	Dpysl5	1.140837153	1.06E-29
ENSMUSG00000015093	Clic3	1.138175042	0.007403659
ENSMUSG00000024107	Lhcgr	1.132216745	0.038850278
ENSMUSG00000030088	Aldh1l1	1.130149599	3.97E-19
ENSMUSG00000047496	Rnf152	1.125091067	2.95E-17

Table A.1 (cont'd)

ENSMUSG00000030562	Nox4	1.119480441	7.26E-20
ENSMUSG00000085664	Atxn7l1os2	1.115950907	0.049939487
ENSMUSG00000085403	Gm13068	1.111291427	0.035327459
ENSMUSG00000105870	7330423F06Rik	1.102392118	0.002277307
ENSMUSG00000024665	Fads2	1.102356377	6.75E-26
ENSMUSG00000025203	Scd2	1.096572492	1.10E-43
ENSMUSG00000017969	Ptgis	1.088273402	2.70E-31
ENSMUSG00000084979	Gm16267	1.085952855	0.003474604
ENSMUSG00000024697	Gna14	1.071172568	0.001223465
ENSMUSG00000020275	Rel	1.070970176	6.74E-48
ENSMUSG00000001467	Cyp51	1.068673265	1.54E-61
ENSMUSG00000038623	Tm6sf1	1.067893051	4.15E-07
ENSMUSG00000006517	Mvd	1.062210849	1.32E-43
ENSMUSG00000032066	Bco2	1.055436112	0.047639175
ENSMUSG00000074676	Foxs1	1.050418422	0.043531869
ENSMUSG00000079654	Prrt4	1.046750317	3.65E-27

Table A.1 (cont'd)

ENSMUSG00000029380	Cxcl1	1.03981259	0.006424397
ENSMUSG00000006403	Adamts4	1.038981977	1.71E-14
ENSMUSG00000042761	Mrap2	1.038813908	0.012243783
ENSMUSG00000058258	Idi1	1.036867263	4.17E-46
ENSMUSG00000055809	Dnaaf3	1.034012926	0.000654305
ENSMUSG00000031169	Porcn	1.03313075	5.93E-18
ENSMUSG00000087042	Gm11611	1.025716389	0.028152442
ENSMUSG00000027459	Fam110a	1.02374762	4.56E-20
ENSMUSG00000020826	Nos2	1.023665657	2.98E-15
ENSMUSG00000042116	Vwa1	1.01944436	7.48E-05
ENSMUSG00000031604	Msmo1	1.018192168	4.92E-48
ENSMUSG00000108825	Gm45838	1.014362481	4.67E-05
ENSMUSG00000043953	Ccrl2	1.011263096	0.0149534
ENSMUSG00000027360	Hdc	1.009906987	0.003260173
ENSMUSG00000021678	F2rl1	1.005672859	0.027004401
ENSMUSG00000042284	Itga1	1.00515774	3.96E-12

Table A.1 (cont'd)

ENSMUSG00000073608	Gal3st2c	1.004561136	9.69E-06
ENSMUSG00000108456	4732496C06Rik	1.004194728	0.007027598
ENSMUSG00000052504	Epha3	1.003992581	0.0004508
ENSMUSG00000038173	Enpp6	1.000379706	8.78E-06
ENSMUSG00000034450	Gulo	-1.002542898	0.003162637
ENSMUSG00000033910	Gucy1a1	-1.011772779	1.44E-05
ENSMUSG00000020467	Efemp1	-1.014829693	1.84E-24
ENSMUSG00000076441	Ass1	-1.023276864	1.13E-09
ENSMUSG00000097451	Rian	-1.036267552	0.003266465
ENSMUSG00000070867	Trabd2b	-1.053062374	1.33E-18
ENSMUSG00000006538	Ihh	-1.057559256	0.001762436
ENSMUSG00000021268	Meg3	-1.063057702	0.000271158
ENSMUSG00000021396	Nxnl2	-1.071824774	0.004202448
ENSMUSG00000032899	Styk1	-1.077512677	0.000246695
ENSMUSG00000075334	Rprm	-1.090742714	0.00461596
ENSMUSG00000044337	Ackr3	-1.104640699	0.012565919

Table A.1 (cont'd)

ENSMUSG00000030607	Acan	-1.111625779	0.023745298
ENSMUSG00000025777	Gdap1	-1.124532224	0.029720475
ENSMUSG00000021640	Naip1	-1.157295562	4.37E-10
ENSMUSG00000096586	Gm22918	-1.163812831	0.004119923
ENSMUSG00000074183	Gsta1	-1.167756095	0.032930109
ENSMUSG00000022054	Nefm	-1.171375777	0.041654816
ENSMUSG00000020099	Unc5b	-1.173123864	1.07E-62
ENSMUSG00000042453	Reln	-1.188156285	1.76E-08
ENSMUSG00000026602	Nphs2	-1.189681413	0.019540185
ENSMUSG00000027419	Pcsk2	-1.199314632	0.00200391
ENSMUSG00000021373	Cap2	-1.208094708	0.00049739
ENSMUSG00000039419	Cntnap2	-1.214593399	7.00E-24
ENSMUSG00000005413	Hmox1	-1.221386276	9.93E-58
ENSMUSG00000016918	Sulf1	-1.222159623	2.12E-21
ENSMUSG00000031465	Angpt2	-1.239146624	4.10E-10
ENSMUSG00000055301	Adh7	-1.244391021	0.029440687

Table A.1 (cont'd)

ENSMUSG00000000435	Myf5	-1.296820117	0.005717626
ENSMUSG00000049107	Ntf3	-1.298559462	2.90E-09
ENSMUSG00000021806	Nid2	-1.301206732	0.000356361
ENSMUSG00000082676	Gm11843	-1.301296198	5.06E-15
ENSMUSG00000064036	Mro	-1.307659644	3.61E-08
ENSMUSG00000033342	Plppr5	-1.309653699	4.04E-05
ENSMUSG00000061353	Cxcl12	-1.319144969	4.21E-05
ENSMUSG00000033007	Asic4	-1.319314742	0.047433484
ENSMUSG00000101859	Gm29233	-1.328498354	0.002503608
ENSMUSG00000031216	Stard8	-1.339077201	1.77E-12
ENSMUSG00000025934	Gsta3	-1.357998501	0.000645456
ENSMUSG00000025127	Gcgr	-1.363476036	4.32E-05
ENSMUSG00000046699	Slitrk4	-1.398743797	8.90E-77
ENSMUSG00000020902	Ntn1	-1.415616681	0.00029717
ENSMUSG00000032419	Tbx18	-1.424888285	0.000936063
ENSMUSG00000044313	Mab21l3	-1.439242208	4.01E-06

Table A.1 (cont'd)

ENSMUSG00000074665	Bpifb4	-1.466550727	0.006972841
ENSMUSG00000027220	Syt13	-1.477151576	1.33E-07
ENSMUSG00000031870	Pgr	-1.499841059	0.021940689
ENSMUSG00000021822	Plau	-1.503772993	3.25E-06
ENSMUSG00000071766	Rhox12	-1.567693589	0.002987725
ENSMUSG00000015619	Gata3	-1.613808564	2.60E-26
ENSMUSG00000062372	Otof	-1.623276637	7.61E-06
ENSMUSG00000040856	Dlk1	-1.644044335	3.13E-10
ENSMUSG00000031380	Vegfd	-1.673597832	2.69E-48
ENSMUSG00000021730	Hcn1	-1.763970618	4.06E-08
ENSMUSG00000058252	Tcp11x2	-1.86169062	0.00393985
ENSMUSG00000029917	C130060K24Rik	-1.923897495	0.000484132
ENSMUSG00000003849	Nqo1	-1.932631264	1.71E-11
ENSMUSG00000020838	Slc6a4	-1.973826888	2.08E-13
ENSMUSG00000027895	Kcnc4	-2.092342463	0.000168556
ENSMUSG00000046999	1110032F04Rik	-2.244200074	2.37E-11

Table A.1 (cont'd)

ENSMUSG00000091272	Gm17641	-3.174375896	1.61E-06

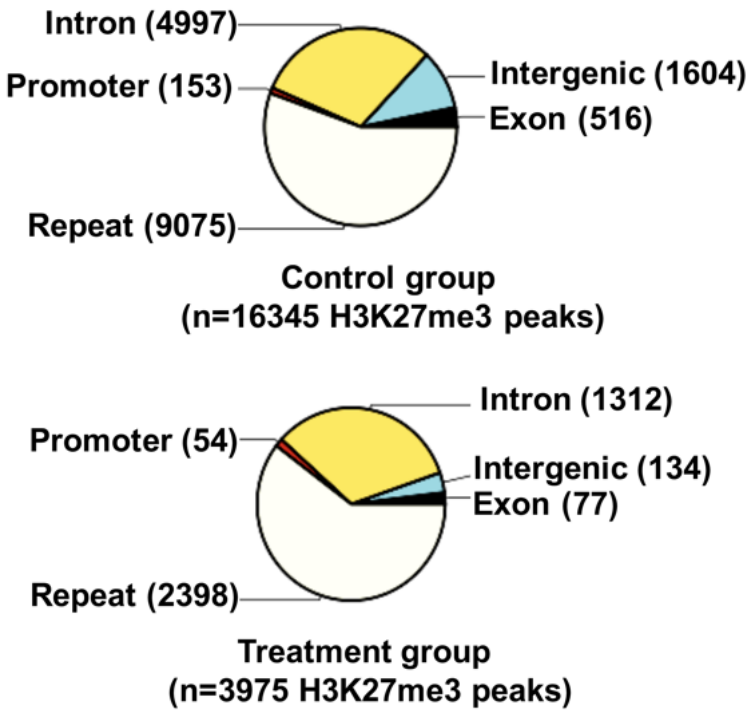


Figure A.1 Genome wide H3K27me3 peaks with respect to different genomic annotations

APPENDIX B

Supplementary materials for Chapter 3

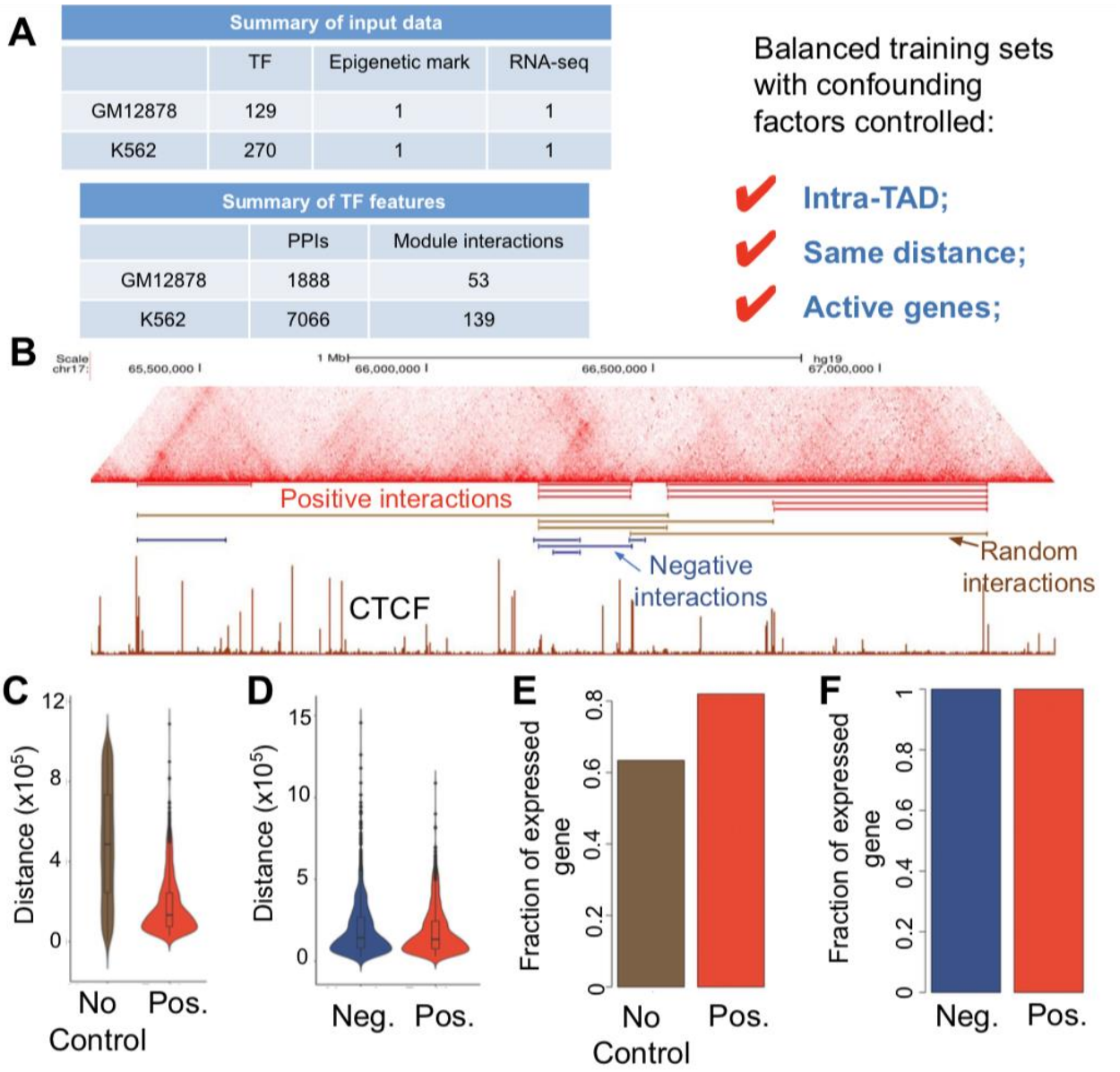


Figure B.1. Summary of training dataset generation and confounding factor controls. (A) Summary of the multi-omics datasets (upper panel) and the numbers of TF PPI features (lower panel). The number of TF features are reduced by applying the hierarchical network community detection on the PPI network. (B-F) A balanced training dataset is generated by controlling three sets of confounding factors. (B) Inter-TAD enhancer-promoter interactions are removed. Compared with Hi-C interactions (red lines), randomly generated enhancer-promoter interactions (brown lines) are enriched with inter-TAD pairs and cover domain boundaries.

Consequently, the stronger binding strength of boundary-enriched TFs, e.g. CTCF, between enhancers and genes are observed for negative interactions. The inter-TAD pairs are removed after confounding factor control (blue lines). (C-D) The genomic distance is controlled. Compared with the genomic distance of positive interactions (red), (C) Genomic distances of randomly paired negative enhancer-promoter interactions (brown) are longer than positive interactions observed from Hi-C (red), before confounding factor controls. (D) The genomic distance distribution of negative sets (blue) is consistent with positive interactions (red), after confounding factor control. (E) The fraction of linking to expressed genes in random enhancer-gene interactions (brown) is lower than the fraction in the positive set based on Hi-C (red), before confounding factor control. (F) All target genes are expressed in both negative interactions (blue) and positive interactions (red), after confounding factor control.

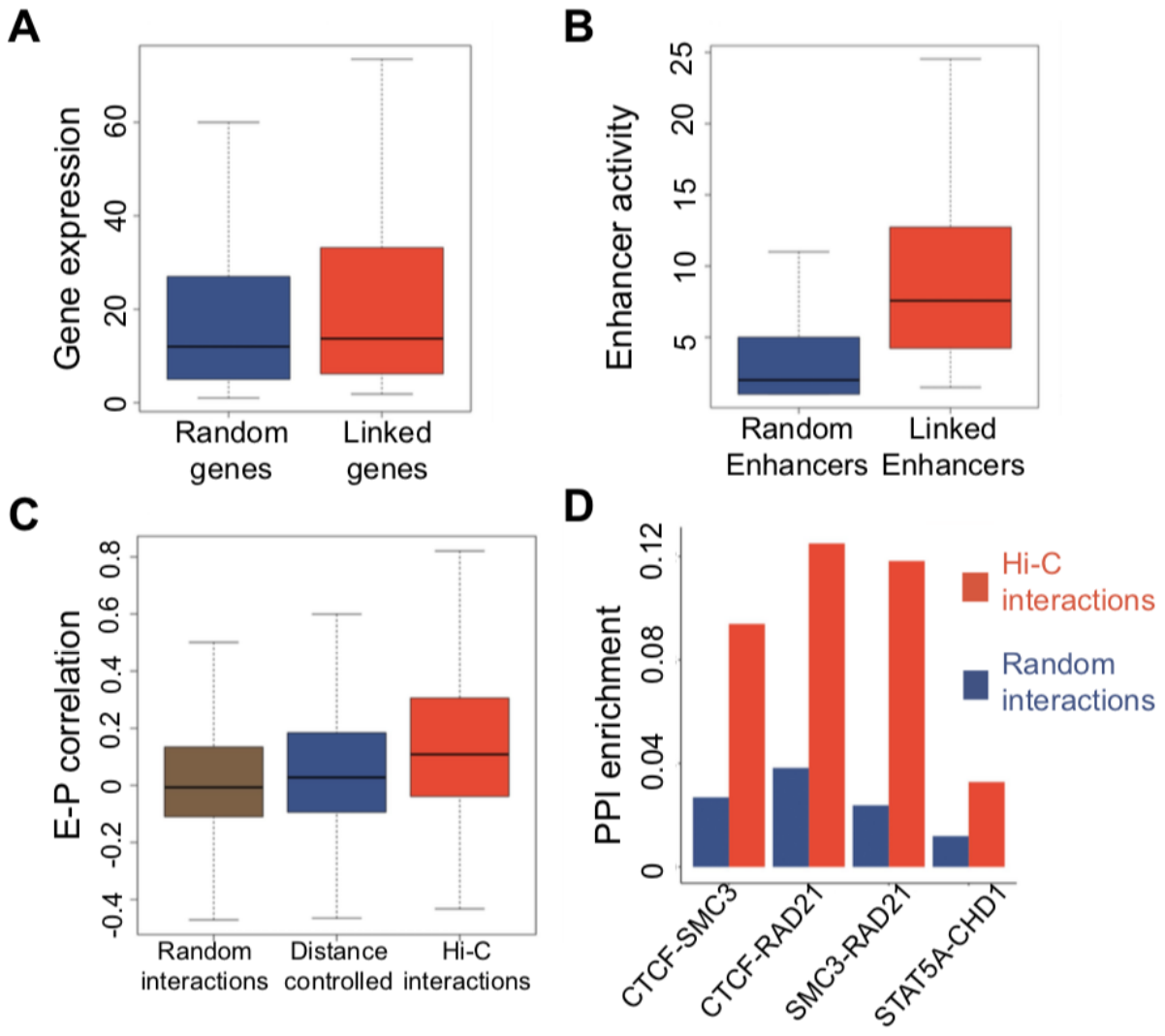


Figure B.2. Predictive powers of features are supported by the differential distributions of features in Hi-C based positive interactions and random negative interactions. (A) The genes linked by Hi-C interactions (red) have a higher expression level than random genes (blue). (B) The enhancers linked by Hi-C interactions (red) are more active than random enhancers (blue). (C) The enhancer-promoter interactions overlapping with Hi-C interactions (red) have higher activity correlations between linked enhancers and genes across 56 cell-types. As comparisons, two versions of controls are generated: 1) Distance controlled random enhancer-promoter pairs (blue): randomly paired enhancer-promoter interactions which follow the same distance distribution as Hi-C interactions. 2) Random pairs (brown): randomly paired enhancer-promoter interactions without controlling any confounding factors. For each version, the same number of enhancer-promoter interactions are generated and the correlations between enhancer activity and gene expression across cell types are calculated. (D) Examples of TF PPIs showing differential enrichments in Hi-C interactions (red) vs. random interactions (blue). The PPI enrichment is calculated as the fraction of enhancer-promoter interactions containing the specific TF PPI features

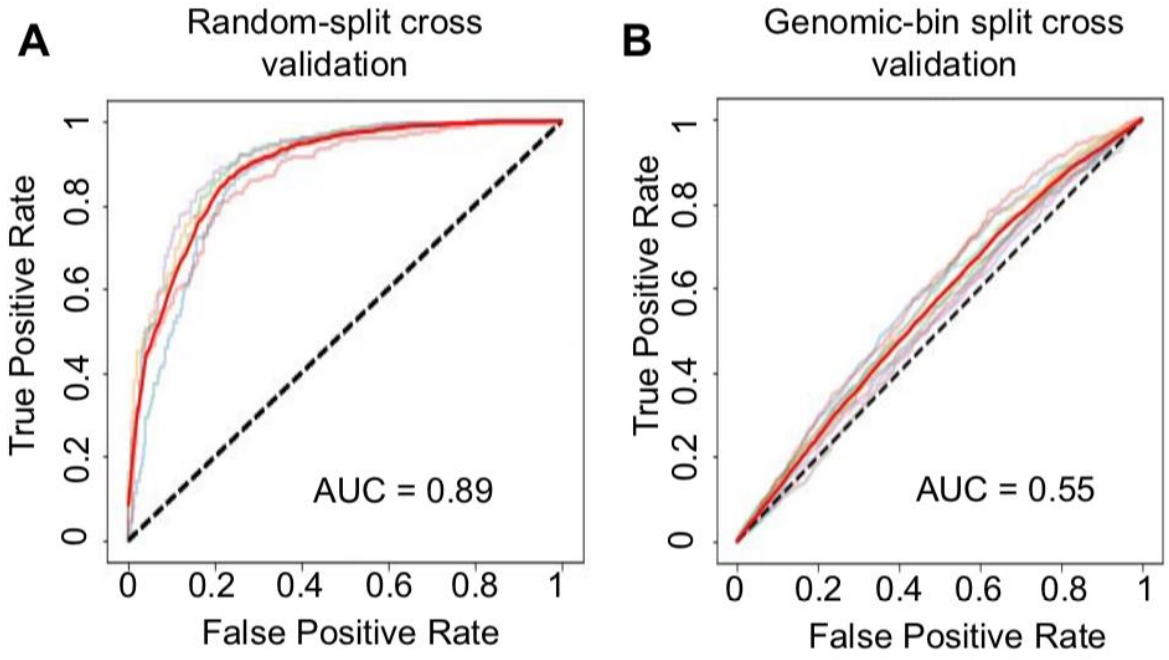


Figure B.3. Advanced feature dimension reduction is needed due to the risk of overfitting. (A) The ROC curves of a random forest predictive model using high-dimensional TF PPI features, based on typical cross-validation. (B) The ROC curves of the same predictive model, based on the rigorous genomic-bin split cross-validation, where the dependency between the training and testing datasets are strictly broken. The significantly decreased AUC is due to the large number of TF PPI features, suggesting advanced feature dimension reduction approach is required to construct a robust predictive model.

A

p	Module count (GM12878)	Module count (K562)
10	831	1083
20	401	526
30	134	419

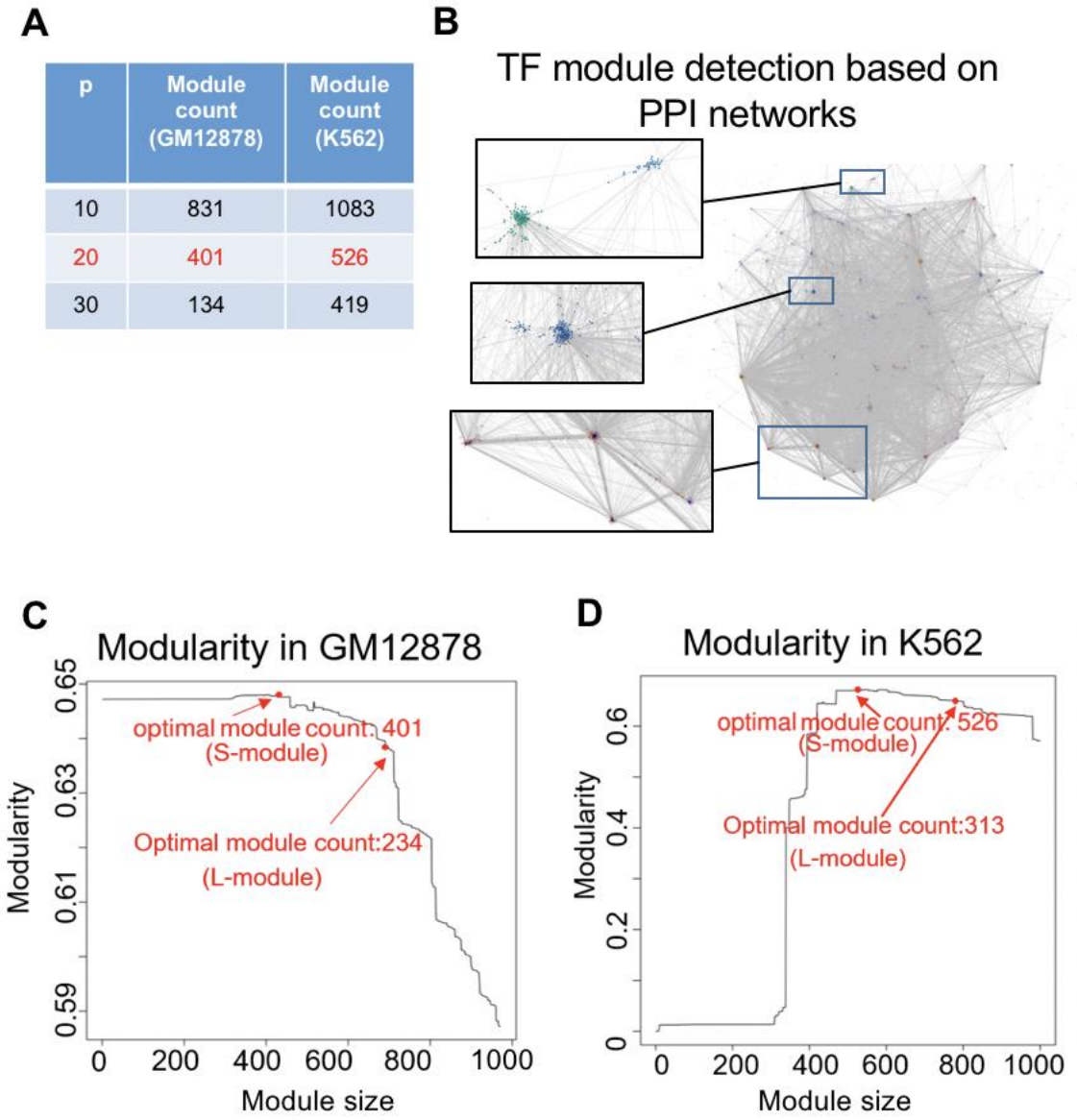


Figure B.4. Hierarchical network-community detection based on the PPI network to construct module-level TF PPI features. (A) The number of TF PPI modules as a function of different random-walk step-sizes (P). step-size=20 is selected to balance the detection of local and global modules. (B) Examples of detected TF PPI modules. Nodes represent proteins and edges represent PPIs. TFs belonging to the same module are annotated with the same color. (C-D) Modularity scores of hierarchical network-community detections with different maximum module sizes. The optimal number of S-modules is selected based on the highest modularity score. The optimal number of the L-module is selected based on the elbow points of the modularity score curves.

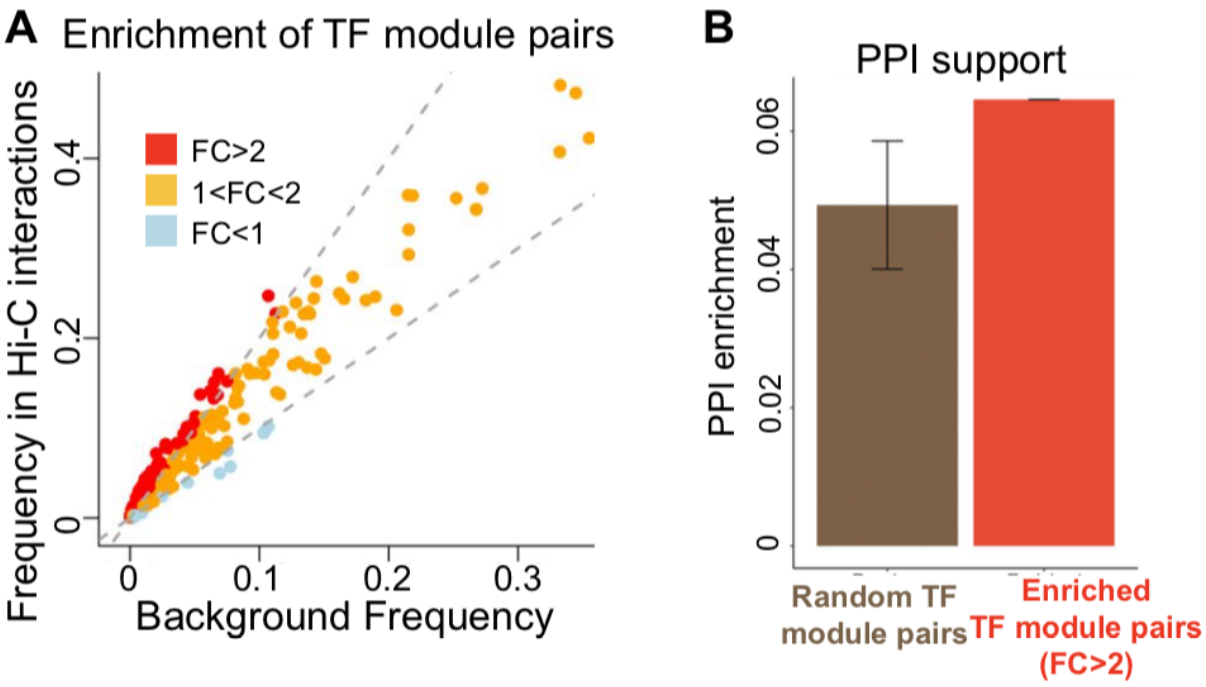


Figure B.5. Enrichment analysis and PPI support analysis for TF module pairs. (A) The enrichment of TF module pairs in Hi-C interactions (y-axis) compared to background random interactions (x-axis). Points represent TF module pairs. Frequency is calculated as the fraction of enhancer-gene interactions containing the specific TF module pairs, one on the enhancer side and one on the linked promoter side. Fold-change (FC) is the ratio of the frequency in Hi-C interactions over the frequency in backgrounds. TF module pairs are colored by the FC (red: FC>2; orange: 1<FC<2; blue: FC<1). (B) Enriched TF module pairs are supported by inter-module PPIs. The fraction of pairs supported by inter-module PPIs is calculated for the set of enriched TF module pairs (red). As controls, the TF members from the enriched module pairs are randomly paired (brown). An empirical statistical test is done based on 1,000 random repeats of controls (p-value=1.39×10^{-2}).

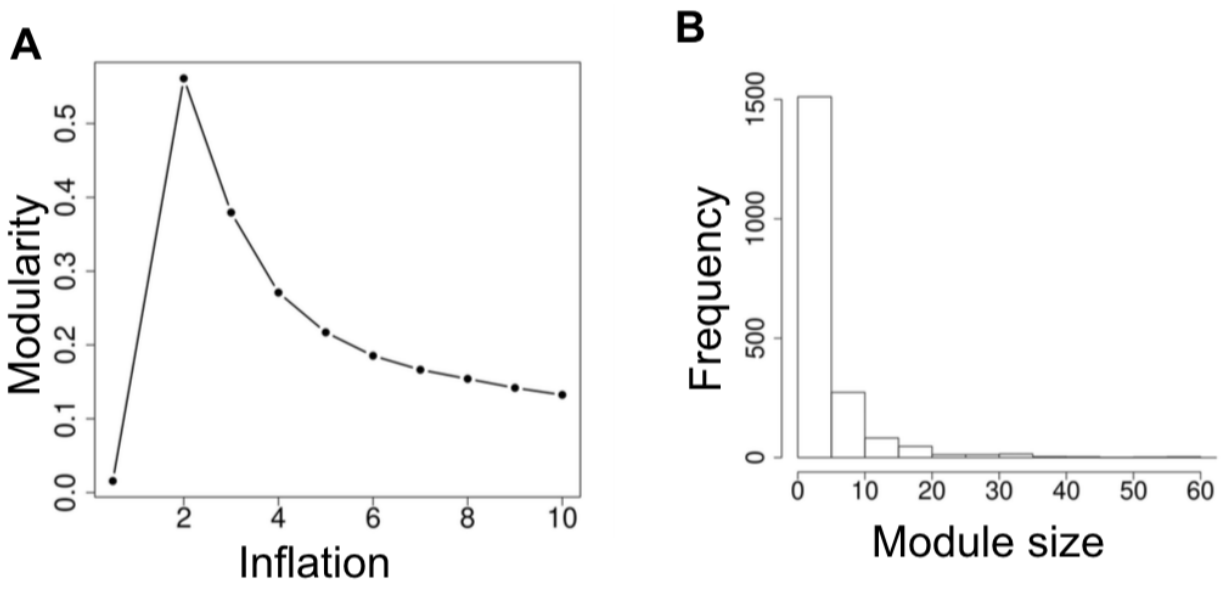


Figure B.6. PPI community detection based on the Markov Cluster Algorithm (MCL). (A) Modularity scores of PPI communities using MCL with different inflation values. (B) Distribution of the module sizes based on the MCL prediction using the inflation value with the highest modularity score.

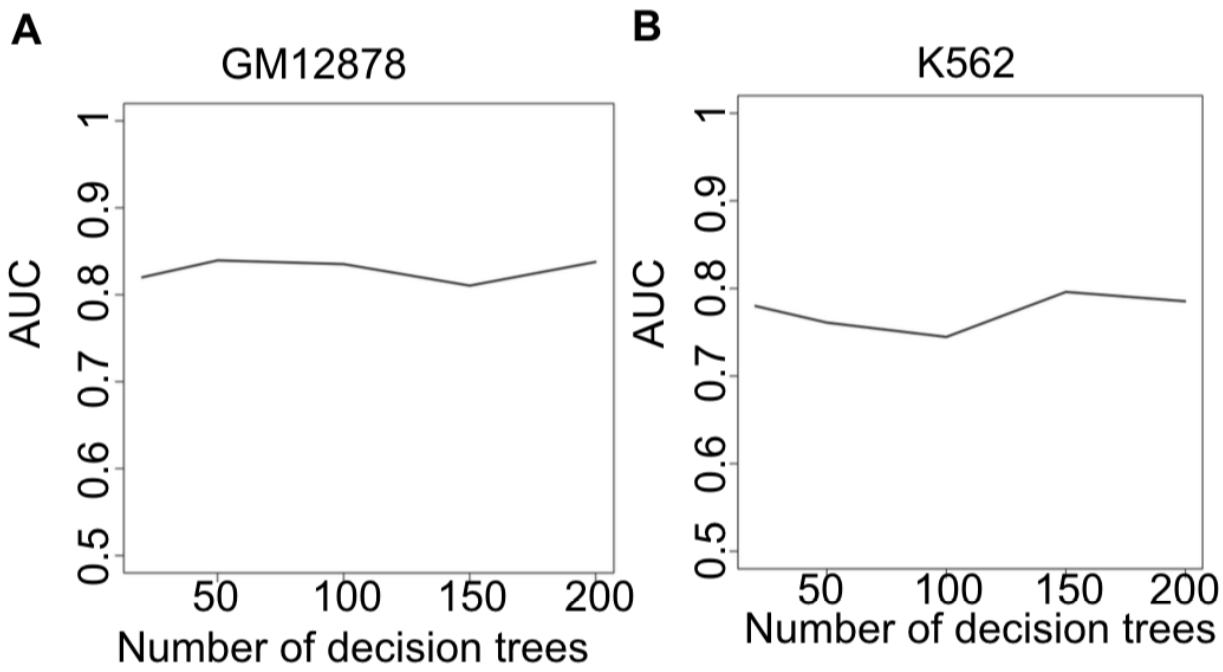


Figure B.7. Model performance (y-axis) as a function of the number of decision trees (x-axis) used in the random forest model. The AUCs are calculated based on the same data and the same cross-validation procedure. The averaged AUC of cross-validations is shown. Robust performances are observed in (A) GM12878 and (B) K562.

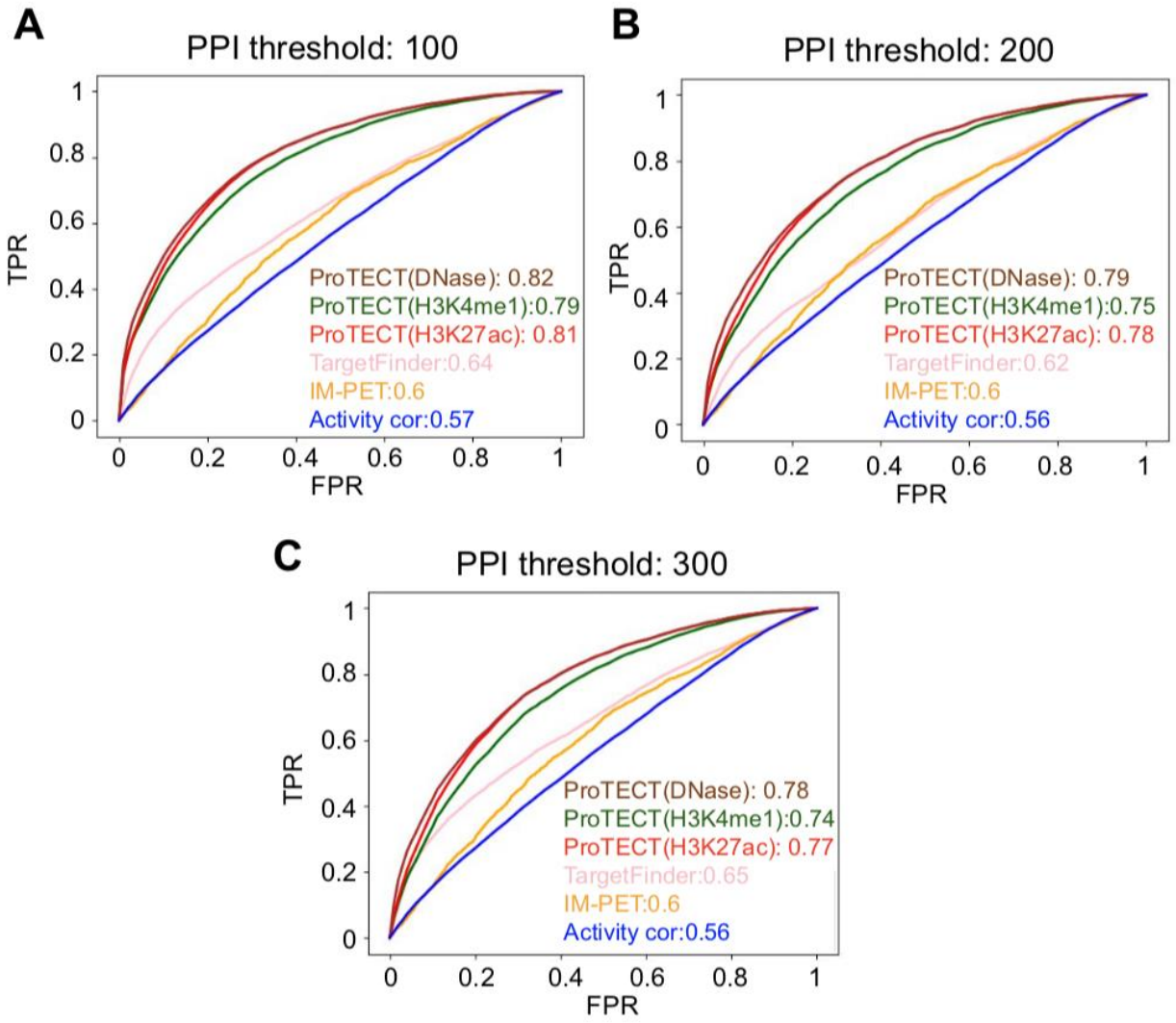


Figure B.8. Performance of ProTECT using different epigenetic signals for enhancers and thresholds for the PPI confidence scores in GM12878. The enhancer activity is quantified by DNase-seq (brown), H3K4me1 (green) and H3K27ac (red), respectively. The threshold of PPI confidence scores is set to (A) 100, (B) 200 and (C) 300. Only PPIs with 'Experimental' confidence scores greater than the thresholds are used as TF-related features in ProTECT and TargetFinder. In K562. ProTECT achieves AUC=0.8, 0.78 and 0.74 with threshold 100, 200 and 300 respectively. As comparisons, TargetFinder achieves AUC=0.71, 0.69 and 0.69.

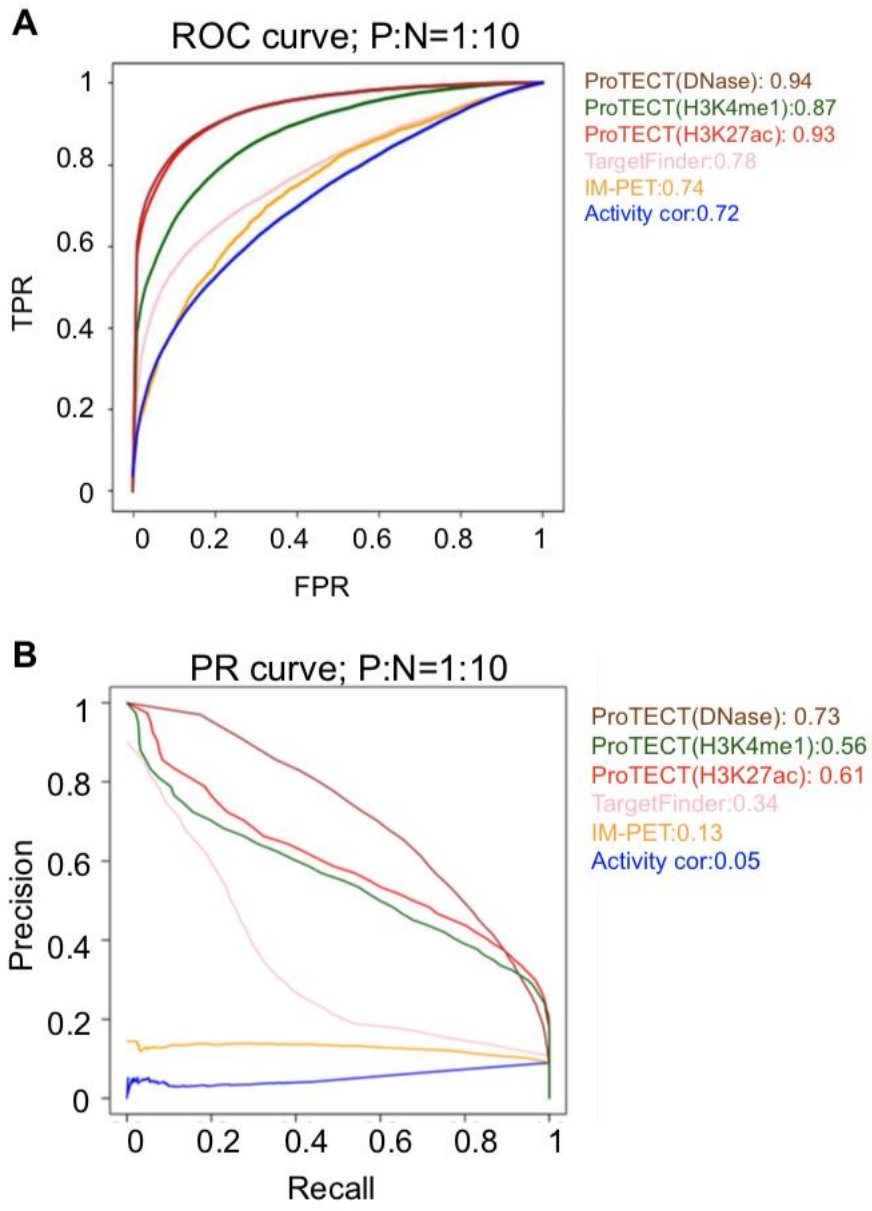


Figure B.9: Performance comparison based on imbalanced training data, using the genomic bin-split cross-validation procedure. The positive to negative ratio is set to 0.1. The model performance is quantified by (A) ROC curves and (B) PR curves.

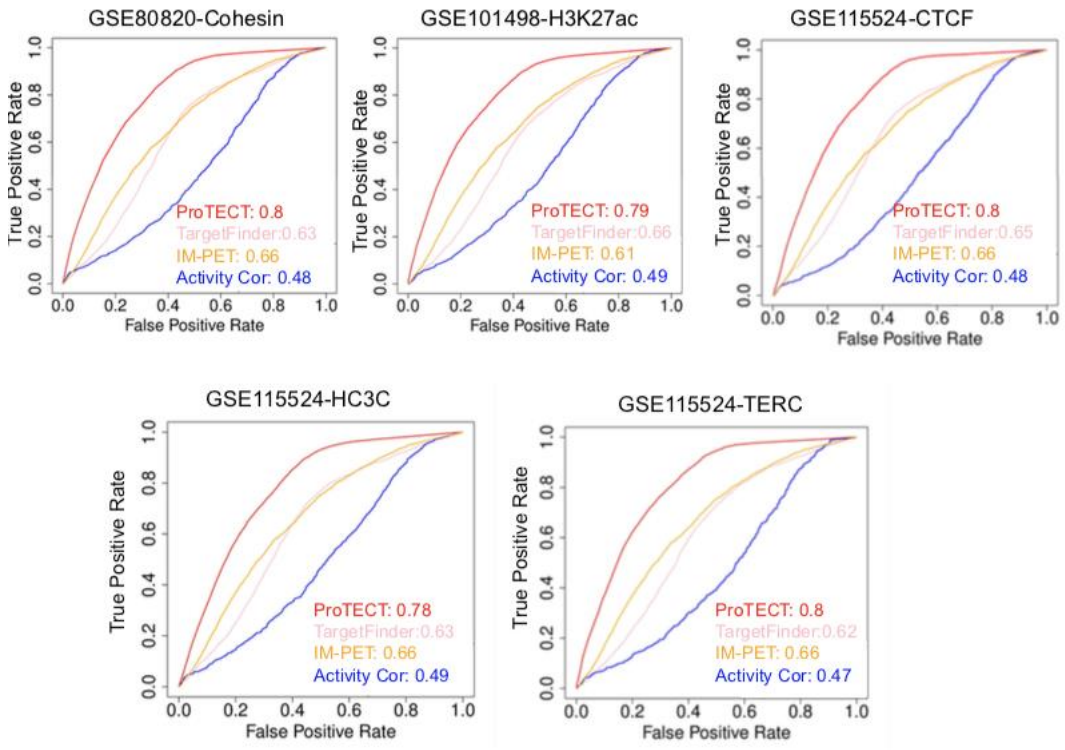


Figure B.10. Performance comparison using five different Hi-ChIP datasets as the gold-standards in GM12878. The model is trained on the balanced training data. ProTECT uses the DNase-seq signals to quantify enhancer activity. The threshold for the PPI confidence scores is set to 100.

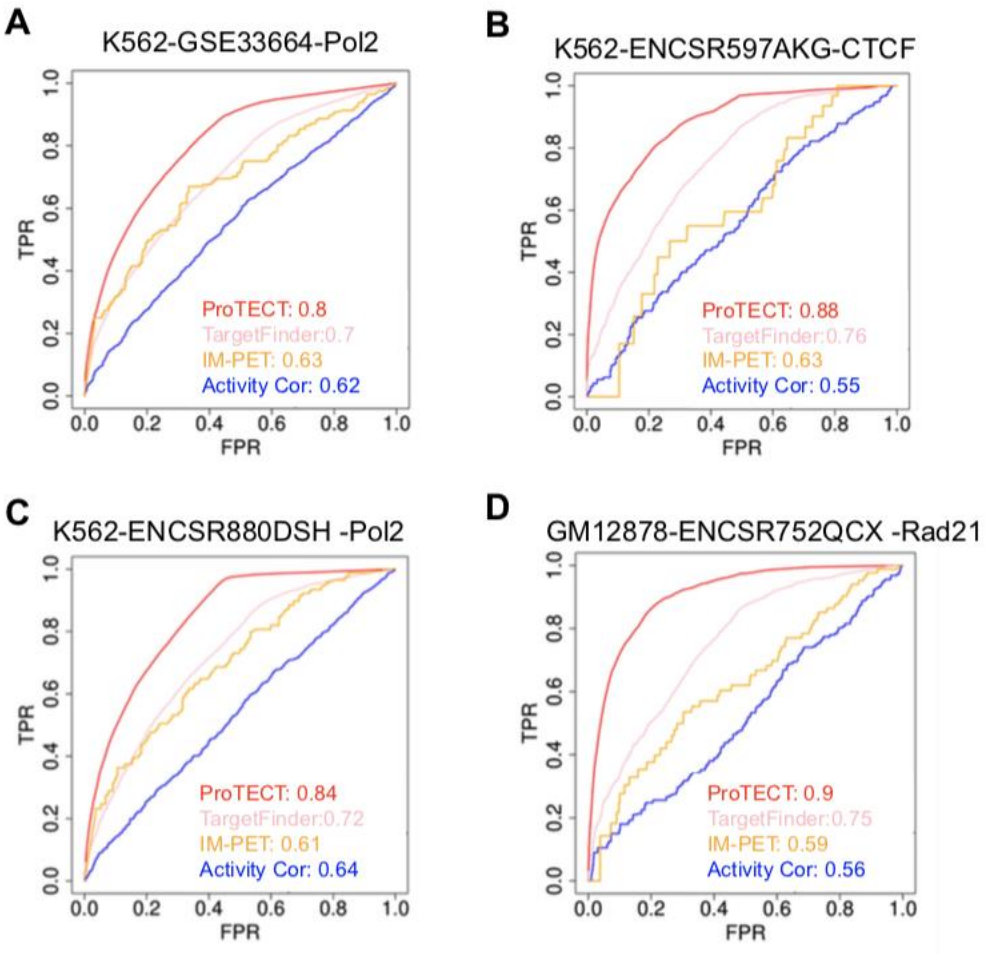


Figure B.11. Performance comparison using four different ChIA-PET datasets as the gold-standards in (A-C) K562 and (D) GM12878. ProTECT uses the DNase-seq signals to quantify enhancer activities. The threshold for the PPI confidence scores is set to 100.

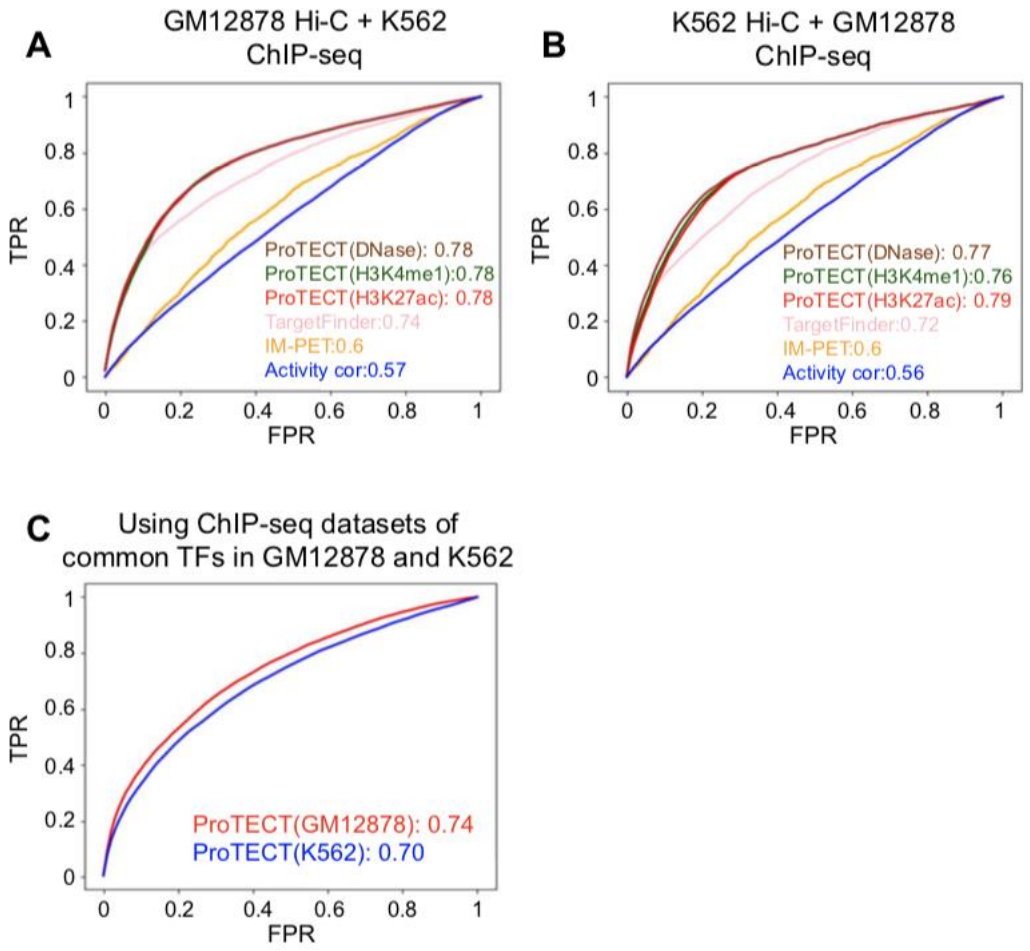


Figure B.12. Performance comparison based on different combinations of Hi-C data and TF ChIP-seq data. (A-B) Performance comparison by pairing Hi-C data and TF ChIP-seq data from different cell-types. (A) The training enhancer-gene pairs are generated using the Hi-C data in GM12878. The TF-related features are generated using the TF ChIP-seq datasets in K562. (B) The training enhancer-gene pairs are generated using the Hi-C data in K562. The TF-related features are generated using the TF ChIP-seq datasets in GM12878. (C) The training data are generated using the Hi-C in GM12878 (red) and K562 (blue) respectively. Only 83 TFs with ChIP-seq data available in both GM12878 and K562 (intersection subset) are used to generate the PPI features.

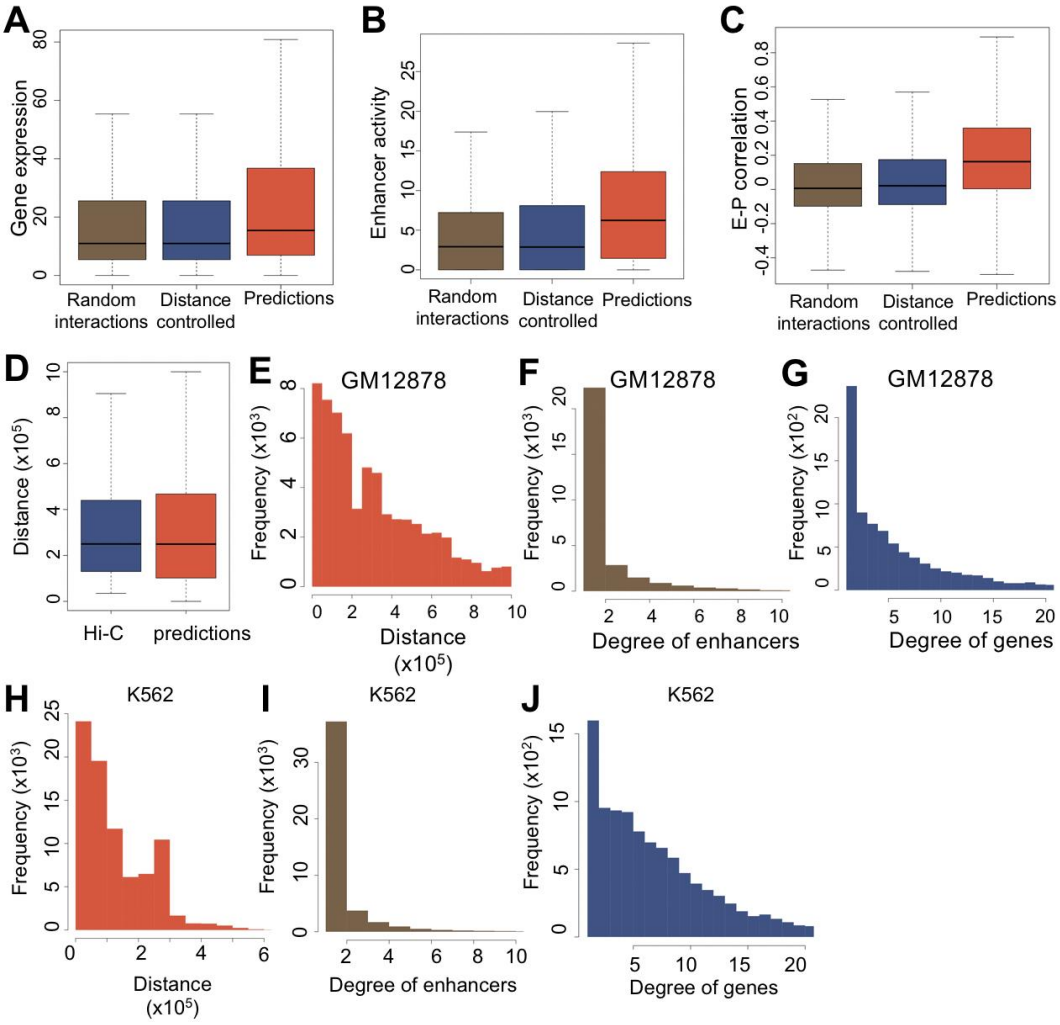


Figure B.13. Summary of genome-wide predictions by ProTECT in GM12878 and K562. (A-C) Compared with controls, the ProTECT predictions have (A) higher gene expressions, (B) higher enhancer activities, (C) higher activity-correlations between enhancers and genes. The controls are generated in the same way as described in Supplementary Figure 1. (D-E) Summary of distance distributions of ProTECT predictions in GM12878. (D) ProTECT predictions have similar distance distributions as real GM12878 Hi-C interactions. (E) The distribution of the genomic distances for predicted enhancer-promoter interactions by ProTECT. (F-G) The degree distributions for (F) enhancers (i.e. the number of target genes regulated per enhancer), and (G) genes (i.e. the number of enhancers co-regulating the same genes), based on ProTECT predictions in GM12878. (H-J) The distribution of genomic distances and the degree distributions for enhancers and genes in K562.

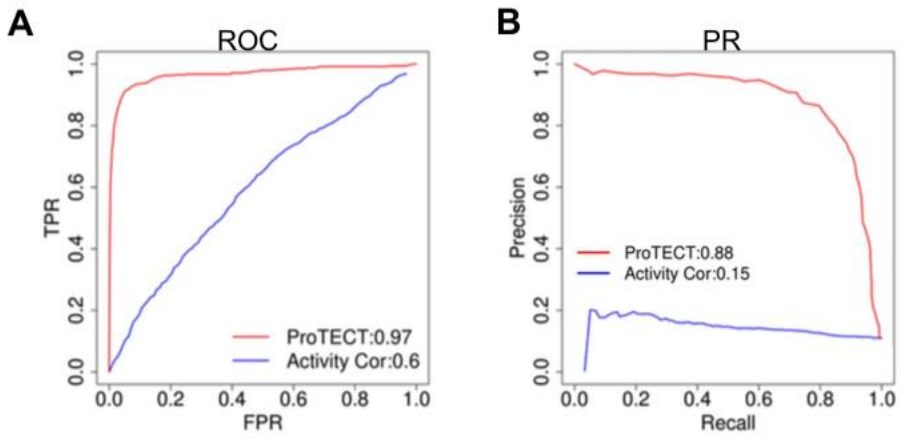


Figure B.14. Validation of predicted enhancer-gene links with enhancer degrees greater than one. ProTECT is applied to the whole genome to predict enhancer-gene links. The enhancer-gene links are considered as positive samples if the links are supported by the Hi-C interactions and the enhancer has degree greater than one across all positive samples. The negative samples are selected by pairing the same enhancer set as the positive samples with random genes. The positive to negative ratio is set to 0.1. The performance is evaluated by (A) ROC curves and (B) PR curves.

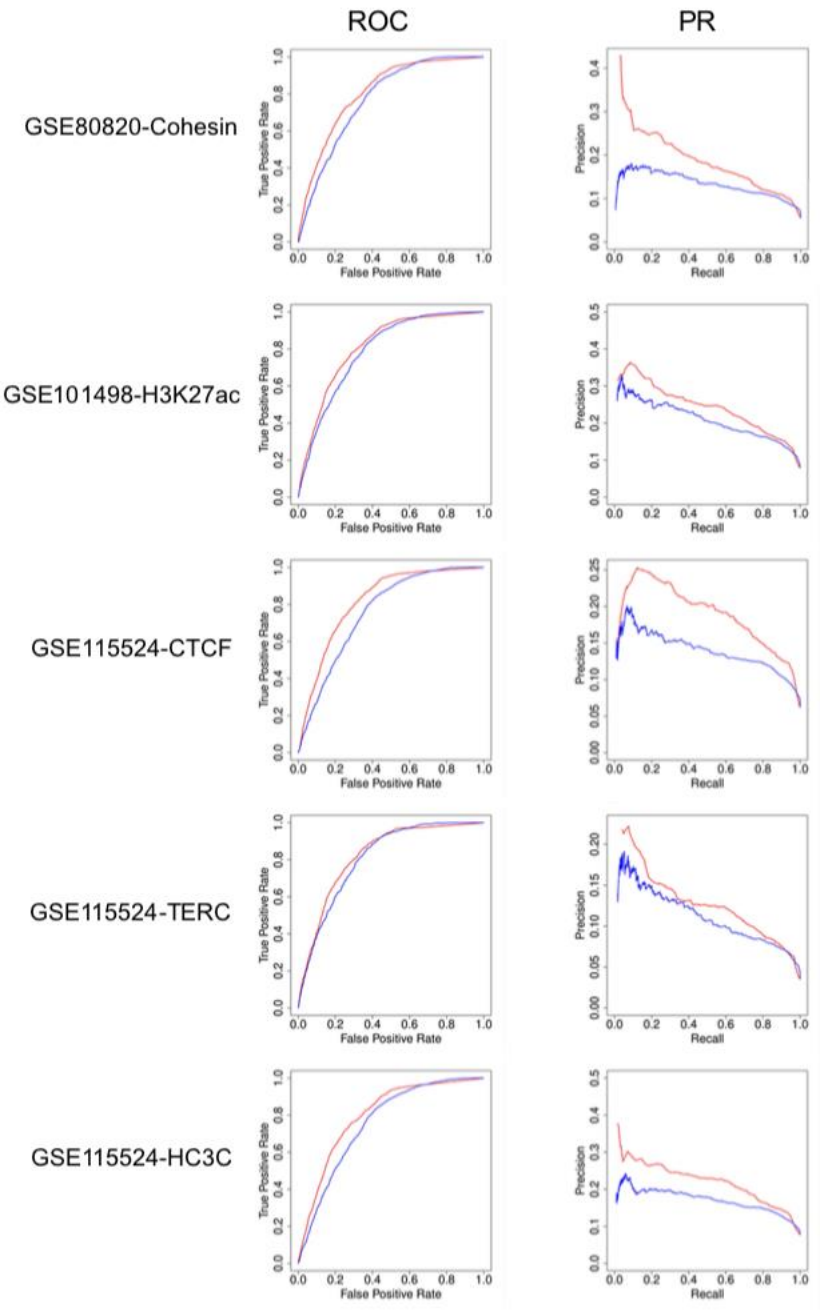


Figure B.15. Performance comparison with the ABC model in the whole genome. The same set of enhancer-gene pairs are ranked by ProTECT (red) and ABC model (blue) respectively. Five different Hi-ChIP datasets are used as the gold-standards. The performance is evaluated by ROC curves and PR curves.

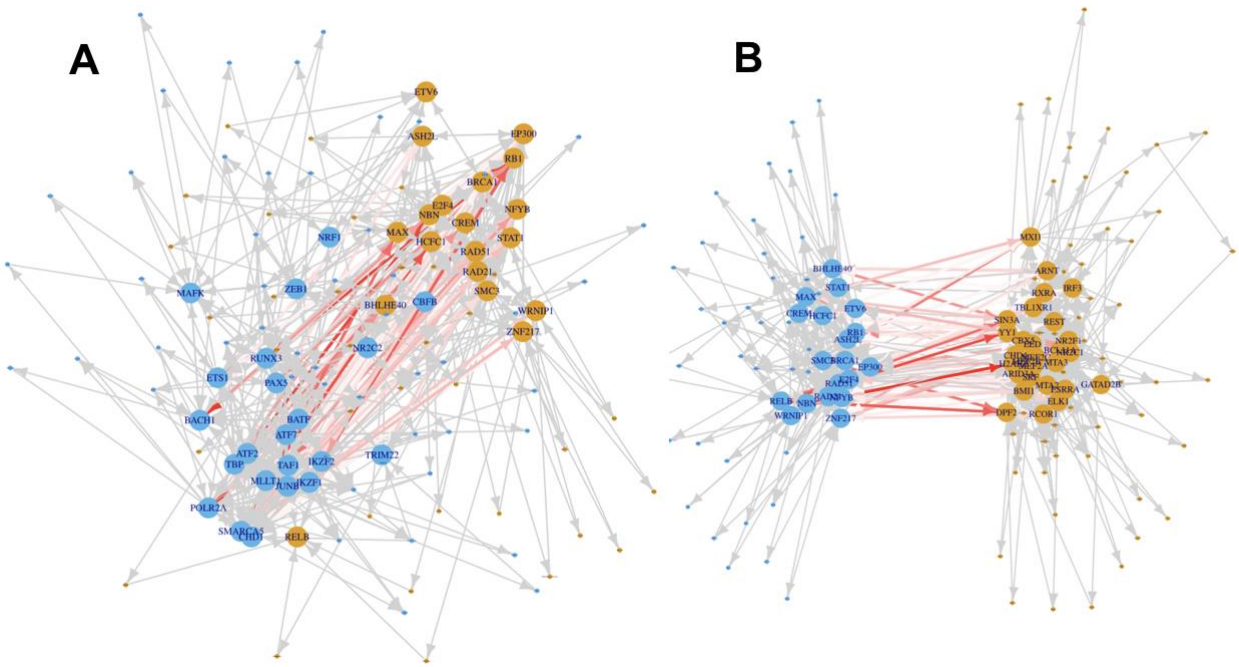


Figure B.16. Examples of prioritized module-level TF PPI features. The two module-level TF PPI features are selected based on the highest feature importance inferred by the random forest model. Two types of nodes are included in the network: TFs (large size) and non-TF proteins (small). Nodes from the first module are colored as blue and TFs from the second module are colored as orange. Edges represent the inter-module TF-level PPIs that connect the TF module pairs. Edges point from enhancer-binding TFs to promoter-binding TFs. The edges are colored by the occurrence frequencies of TF PPIs in ProTECT predictions. Important TF-level PPIs in each module-level feature are further prioritized based on the occurrence frequencies, e.g.RELB-YY1 and SMC3-POLR2A.

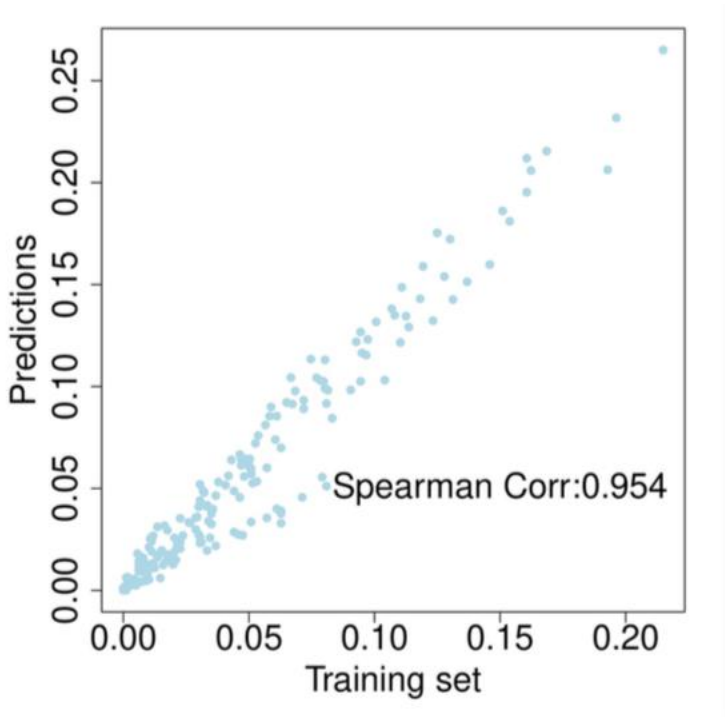
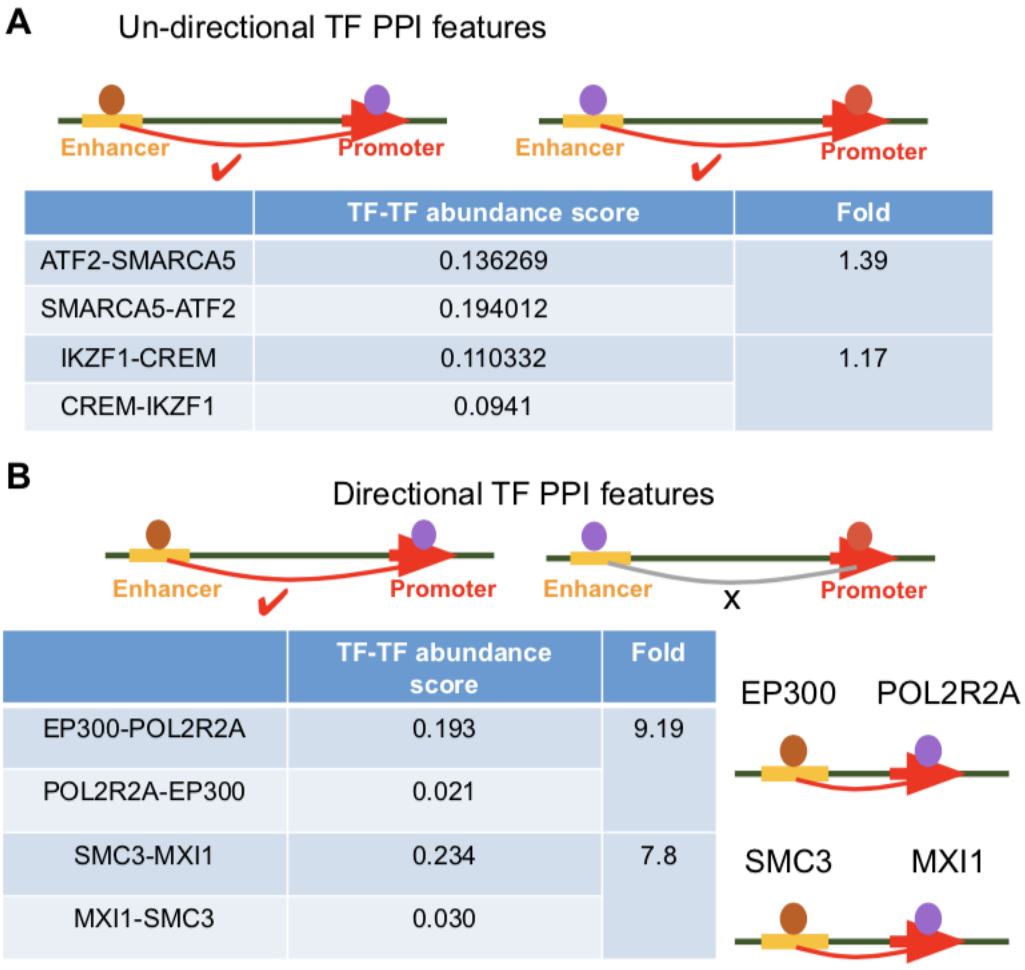


Figure B.17. Comparing the TF-level PPI abundance scores in the Hi-C supported enhancer-gene links (training set, x-axis) and the ProTECT predictions (y-axis). Dots represent the TF PPIs. The Spearman correlation is 0.954.



	TF-TF abundance score	Fold
ATF2-SMARCA5	0.136269	1.39
SMARCA5-ATF2	0.194012	
IKZF1-CREM	0.110332	1.17
CREM-IKZF1	0.0941	

	TF-TF abundance score	Fold
EP300-POL2R2A	0.193	9.19
POL2R2A-EP300	0.021	
SMC3-MXI1	0.234	7.8
MXI1-SMC3	0.030	

Figure B.18. Identification of the directions of TF PPI features. For each pair of TF PPI features with opposite directions, the fractions of predicted enhancer-promoter interactions containing the specific TF PPI features are used as the abundance scores. (A) Examples of un-directional TF PPI features, where the abundance scores of two directional features are similar to each other. (B) Examples of directional TF PPI features, where the abundance scores of the two directional features are substantially different.

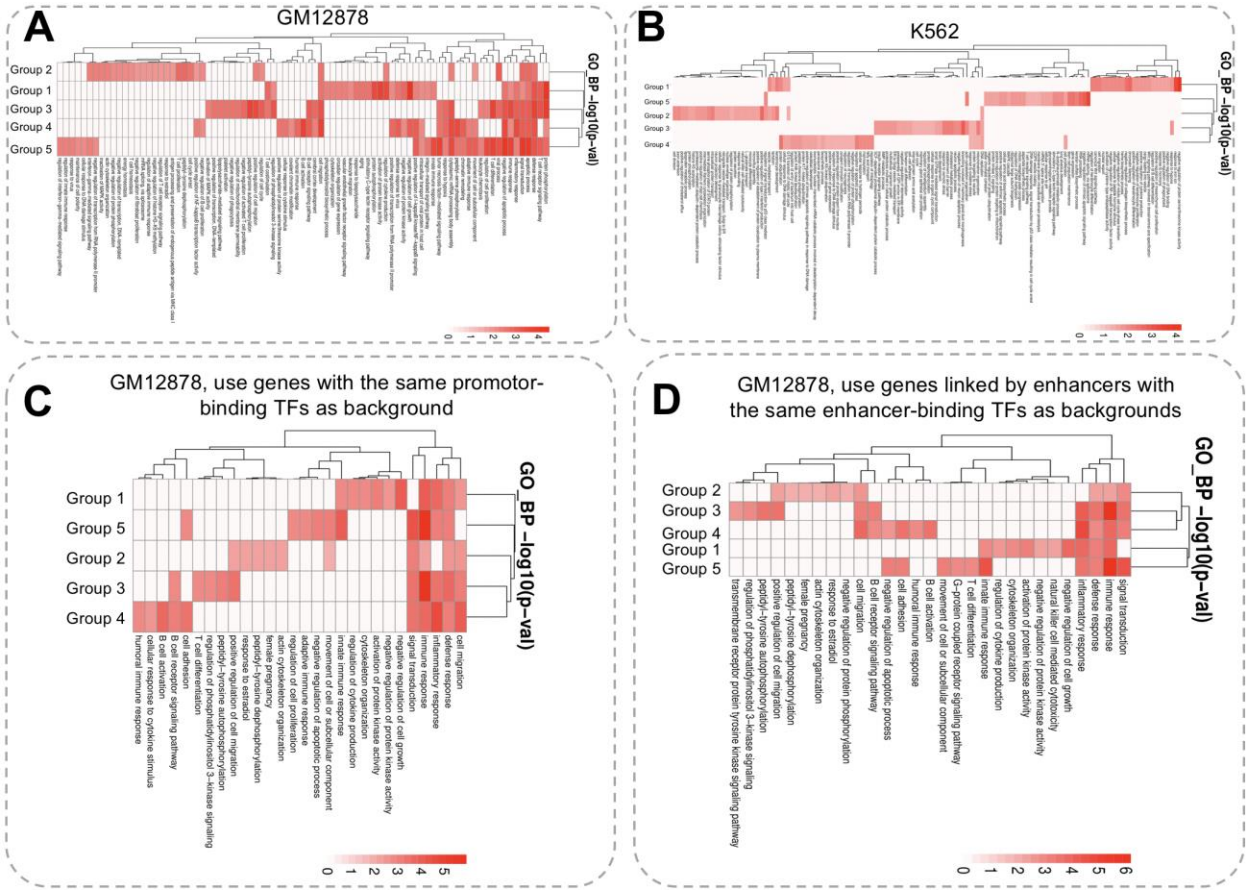


Figure B.19. Differential pathway enrichments of genes regulated by different module-level TF PPIs based on the ProTECT predictions. (A-B) Using all genes as the background gene sets in (A) GM12878 and (B) K562. (C) The genes with the same promoter-binding TFs are used as the background gene sets in GM12878 for the GO analysis. (D) The genes linked by the enhancers with the same enhancer-binding TFs are used as the background gene sets in GM12878 for the GO analysis.

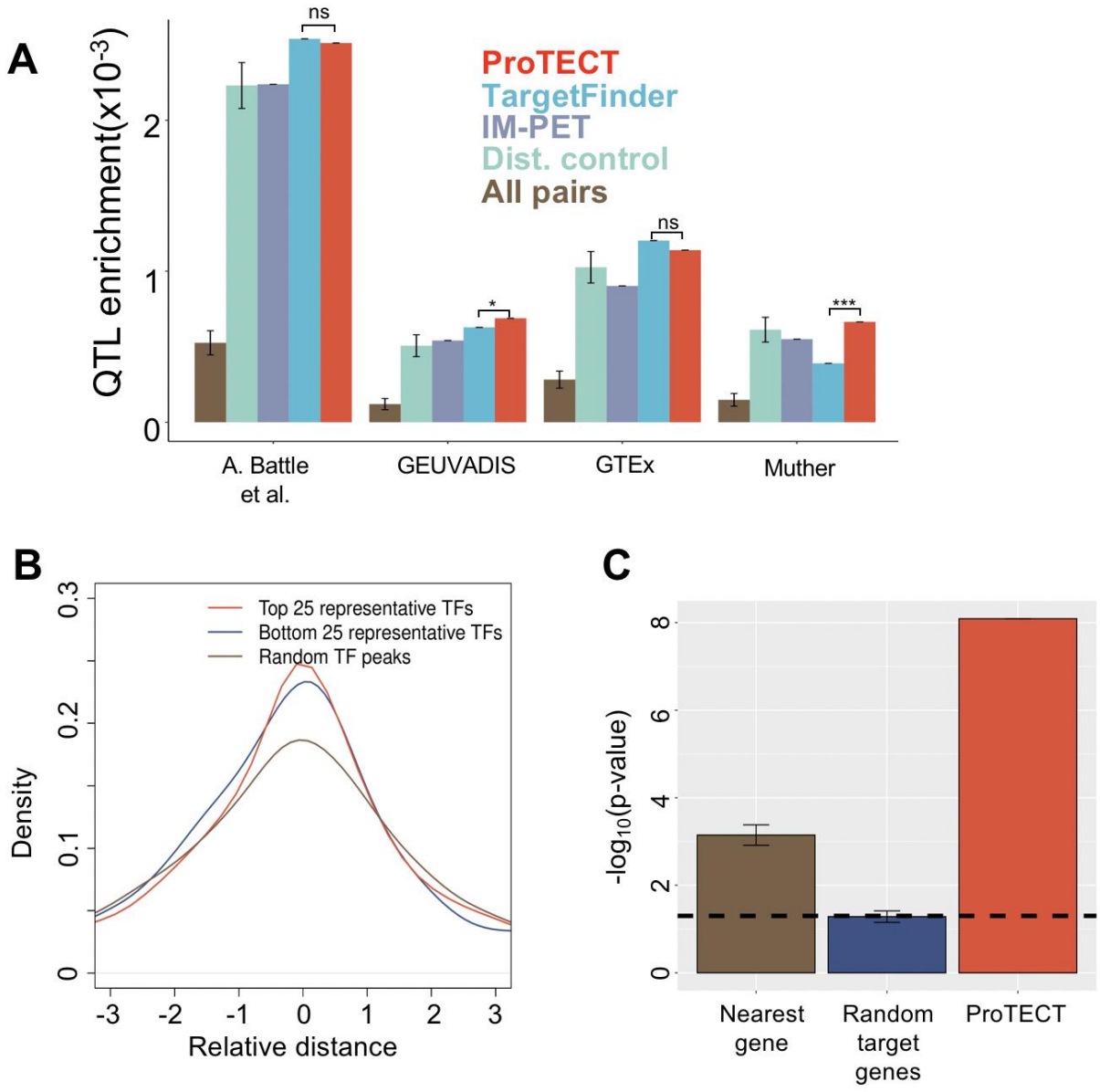


Figure B.20. QTL enrichment analysis in K562. (A) Enrichment of ProTECT predictions with cis-eQTLs (y-axis) from multiple datasets (x-axis). Two versions of controls are generated: 1) Random pairs (brown): randomly pairing enhancers and promoters within 2Mb distance windows. 2) Distance controlled (blue): randomly pairing enhancers and genes while the genomic distances follow the same distribution as ProTECT predictions (red). Controls are generated 1,000 times and standard deviations are used for error bars. Enrichments of TargetFinder and IM-PET are also included. Empirical p-values are calculated. (***: p-value < 10^{-3}, **: p-value < 10^{-2}) (B) Relative genomic distance distribution between cis-eQTL SNPs and the summits of ChIP-seq peaks of the prioritized TFs by the model (red). The relative distance is calculated as the distance between cis-eQTL SNPs and TF ChIP-seq peak summits, normalized by the size of TF peaks. The same number of peaks of bottom-ranked TFs (grey) and randomly selected enhancer-binding TFs (blue) are used as controls. The p-value equals to 0.0518 based on the Kolmogorov-Smirnov test. (C) Trans-eQTL analysis in K562. The hypergeometric test is used to test the enrichment of

overlaps between enhancer-mediated TF-gene pairs and trans-eQTLs, whose SNPs located in the TF's gene body and the eQTL target gene is the same as the TF's target gene (red). The $-\log_{10}$(p-value) of the hypergeometric test is shown. The p-values are highly significant in ProTECT predictions (p-value=8.12×10^{-9}) compared with two controls: 1) nearest gene to enhancers (p-value=0.052, brown), and 2) random target genes (p-value=7.13×10^{-4}, blue).

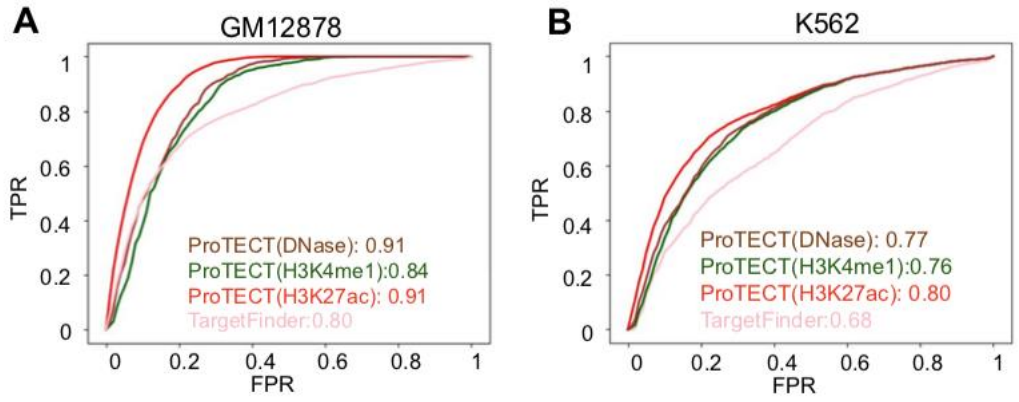


Figure B.21. ProTECT predicts enhancer-gene links based on the imputed TF binding sites. The TF binding sites are imputed by overlapping TF motifs with cell-type specific DNase-seq peaks. For each TF, strong motifs within DNase-seq peaks across the whole genome are considered as the imputed TF binding sites. ProTECT uses the imputed TF binding sites to generate the TF PPI features. (A) In GM12878, TF binding sites for 128 TFs are imputed and used. (B) In K562, TF binding sites for 270 TFs are imputed and used.

APPENDIX C

Supplementary materials for Chapter 4

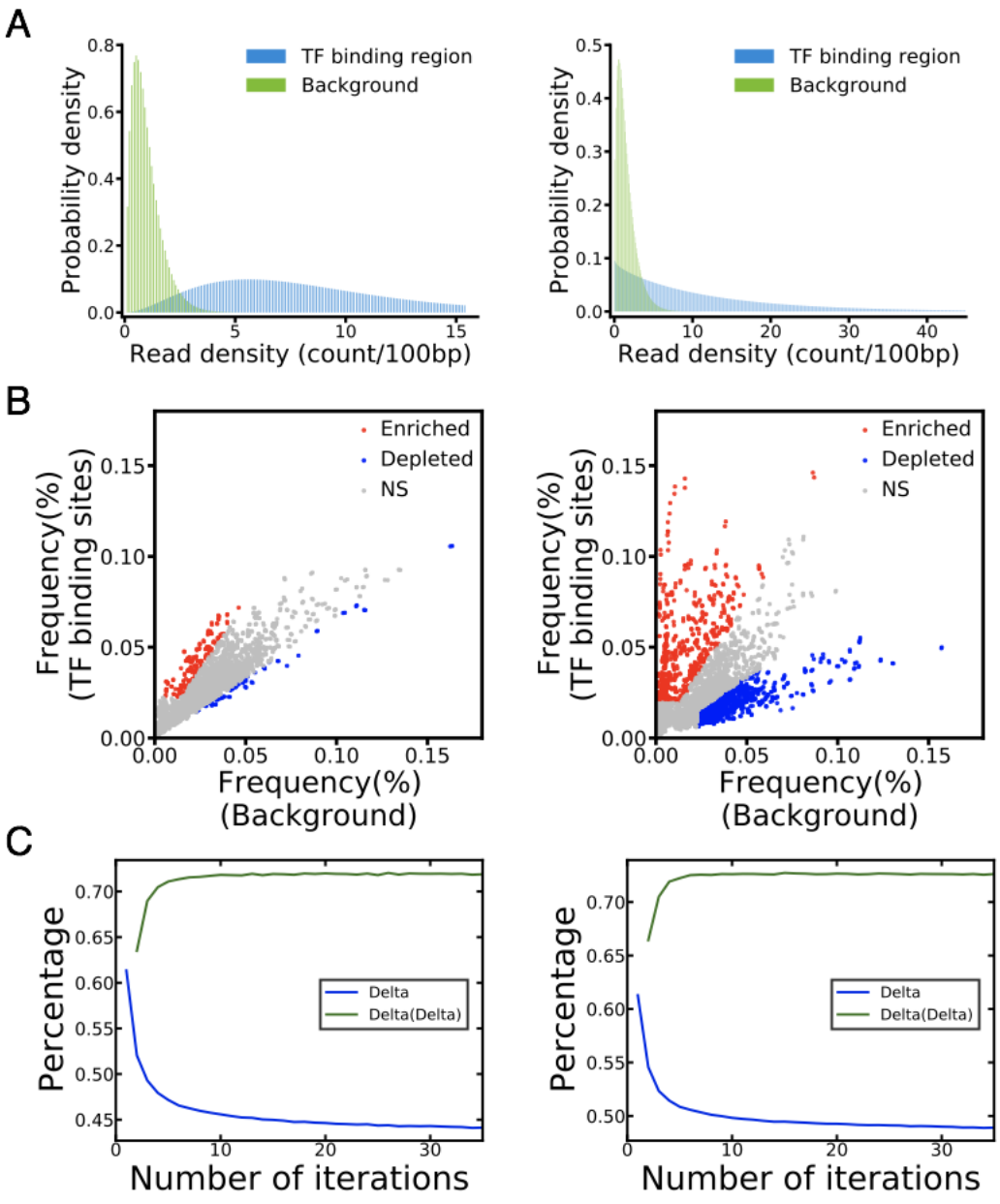


Figure C.1. The exploration of different parameters in the model. (A) The Ps, Pn distribution in a CTCF ChIP-seq dataset in GM12878 (left) and a NRF1 ChIP-seq dataset in K562 (right). (B) The kmer frequencies in the CTCF ChIP-seq dataset (left) and the NRF1 ChIP-seq dataset (right). (C) Model convergence with iterations in the CTCF ChIP-seq dataset (left) and the NRF1 ChIP-seq dataset (right).

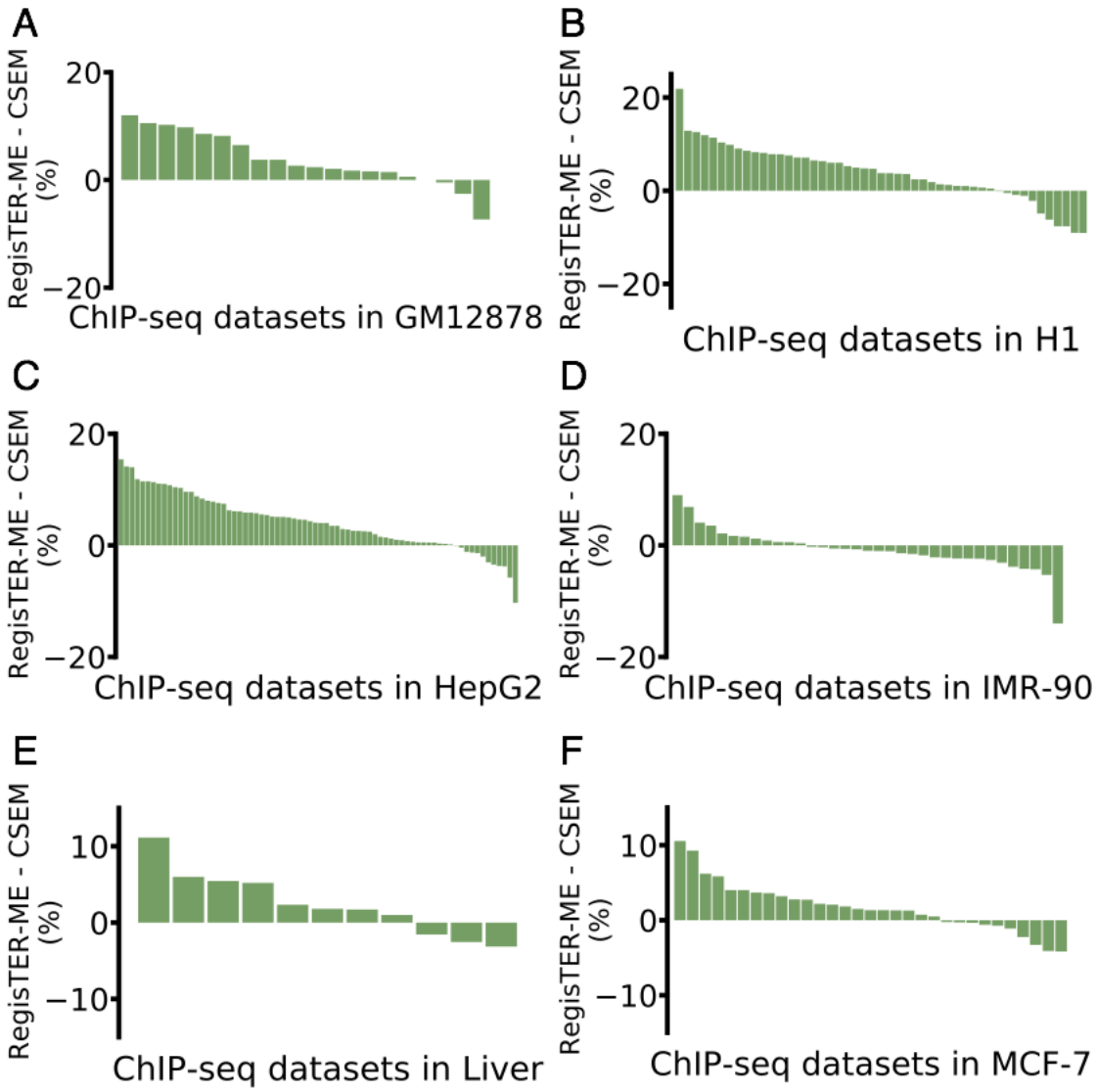


Figure C.2. Method comparison using fraction of peaks within promoter regions in different cell lines. (A) promoter enrichment difference across all the ChIP-seq datasets in GM12878. (B) promoter enrichment difference across all the ChIP-seq datasets in H1. (C) promoter enrichment difference across all the ChIP-seq datasets in HepG2. (D) promoter enrichment difference across all the ChIP-seq datasets in IMR-90. (E) promoter enrichment difference across all the ChIP-seq datasets in liver cell. (F) promoter enrichment difference across all the ChIP-seq datasets in MCF-7.

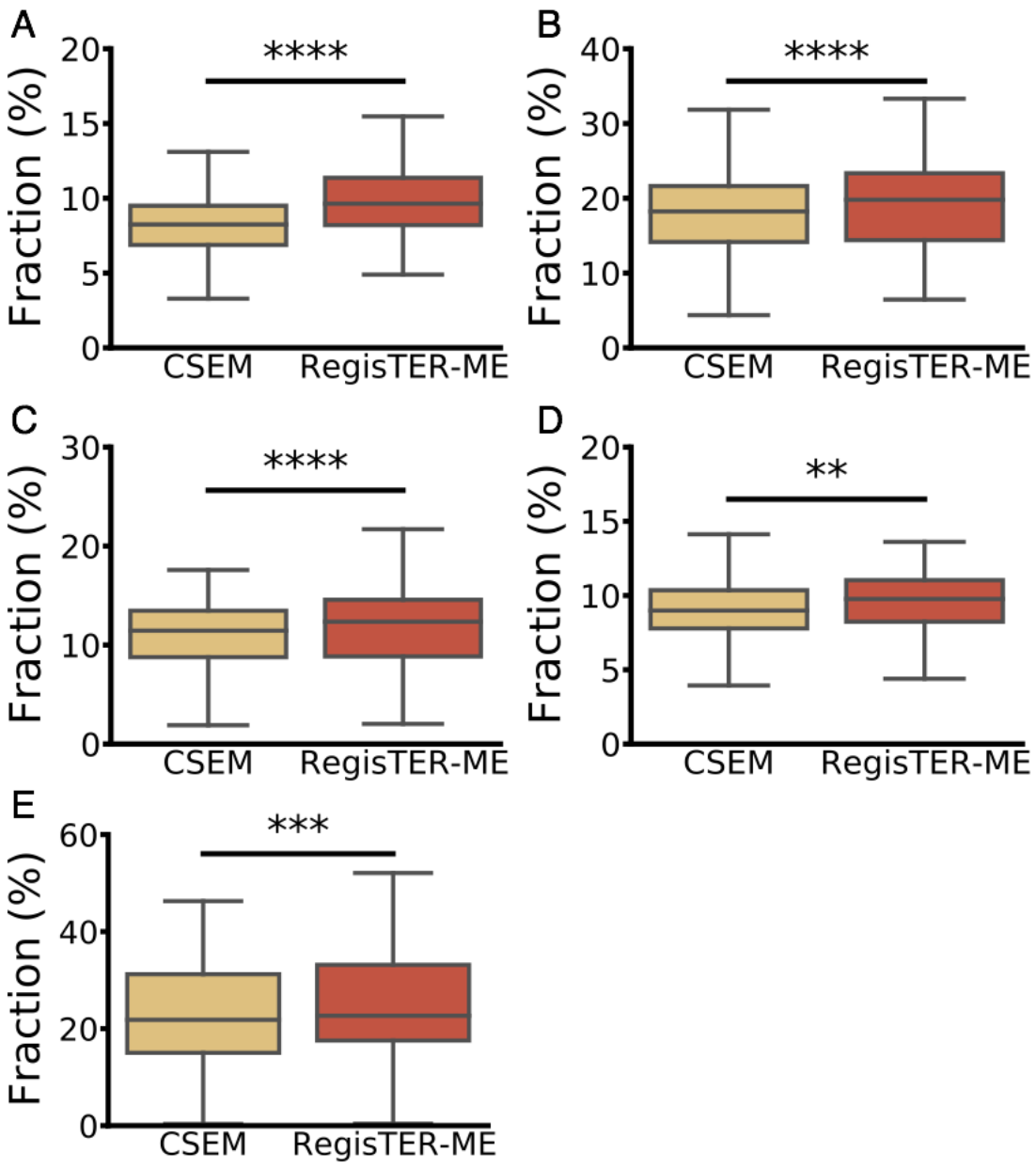


Figure C.3. Method comparison using fraction of peaks involved in enhancer-promoter interactions. (A) Fraction of RegisTER-ME peaks or CSEM peaks overlapping anchors in Hi-C data in K562. (B) Fraction of RegisTER-ME peaks or CSEM peaks overlapping anchors in Capture-C data in GM12878. (C) Fraction of RegisTER-ME peaks or CSEM peaks overlapping anchors in Hi-C data in GM12878. (D) Fraction of RegisTER-ME peaks or CSEM peaks overlapping anchors in Capture-C data in H1. (E) Fraction of RegisTER-ME peaks or CSEM peaks overlapping anchors in Hi-C data in H1.

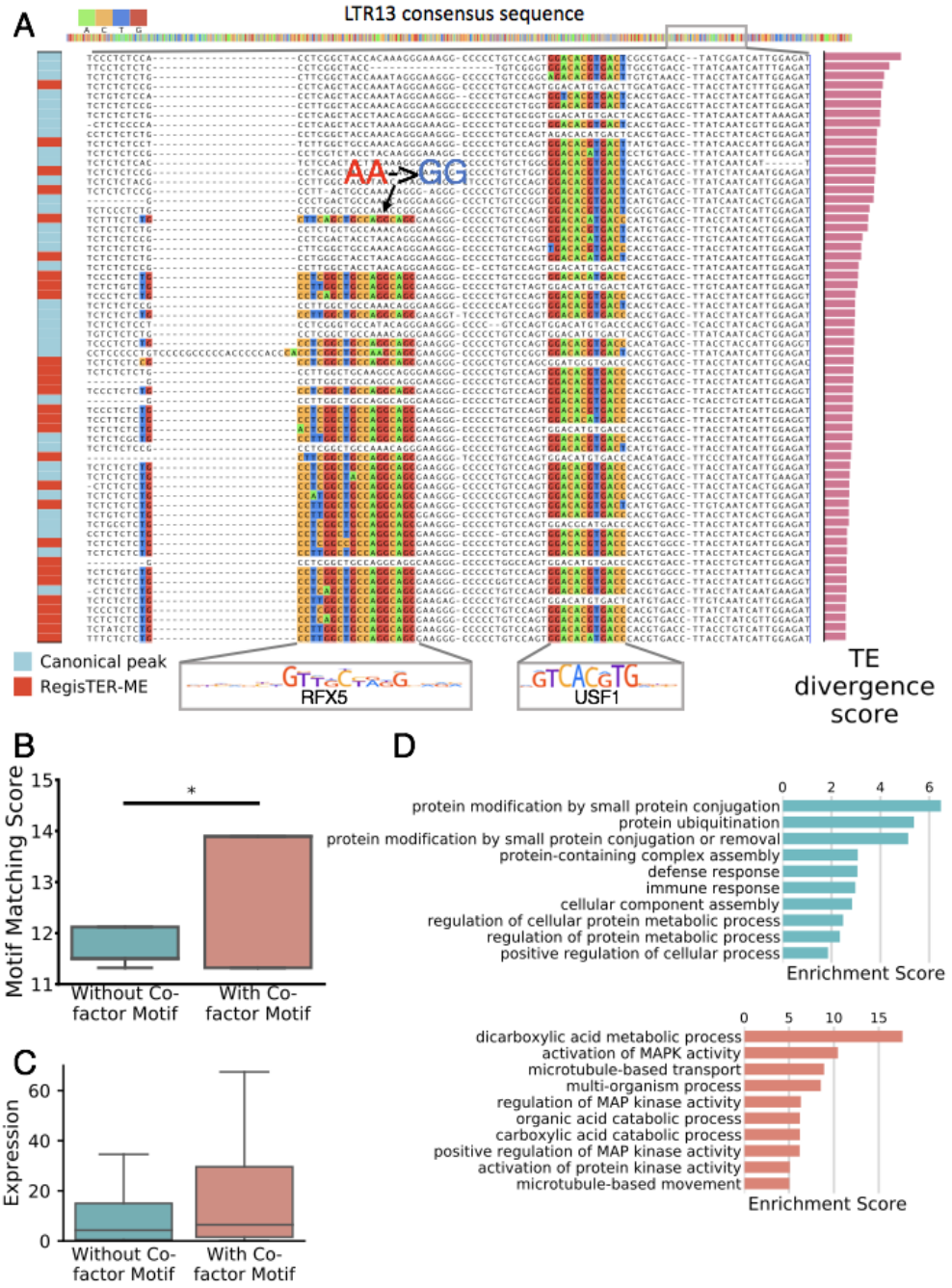


Figure C.4. Co-evolution of transcription factors and co-factors in TE-derived TF binding sites. (A) A sequence alignment for LTR13-derived USF1 binding regions. The left panel shows the identification of each binding site (blue: from canonical peaks, red: from RegisTER-ME peaks). The positions with motifs in the sequence alignment result are colored based on the nucleotides. The right panel shows the divergence score of each LTR13 sequence. (B) The motif matching scores for USF1 binding sites. The USF1 binding sites are divided into 2 groups: before having RFX5 motifs in the flanking regions and after having RFX5 motifs in the flanking regions. (C) The expression level of the nearest genes for the 2 groups of USF1 binding sites. (D) Enriched pathways for the 2 groups of nearest genes.

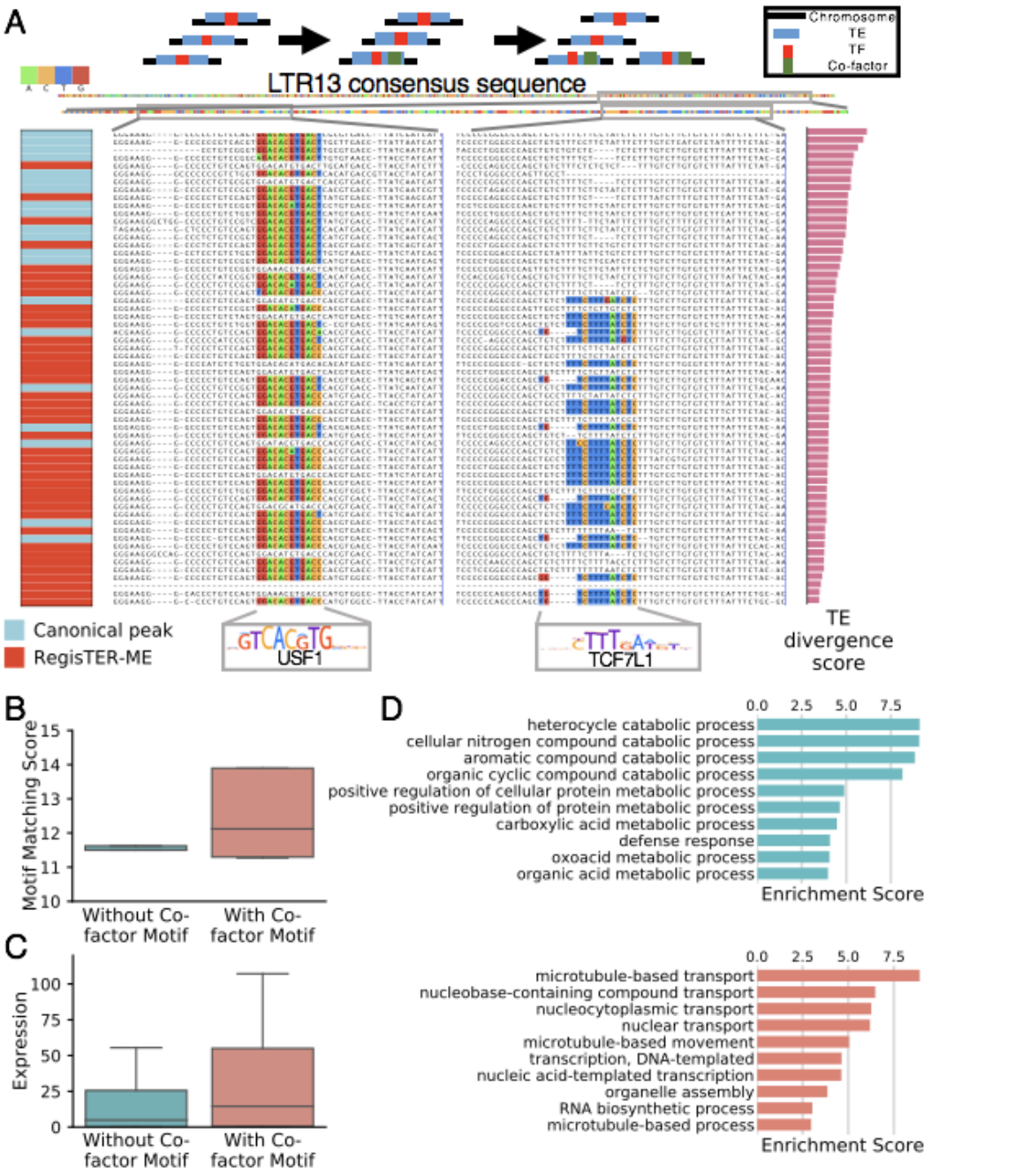


Figure C.5. Co-evolution of transcription factors and co-factors in TE-derived TF binding sites. (A) A sequence alignment for LTR13-derived USF1 binding regions. The left panel shows the identification of each binding site (blue: from canonical peaks, red: from RegisTER-ME peaks). The positions with motifs in the sequence alignment result are colored based on the nucleotides. The right panel shows the divergence score of each LTR13 sequence. (B) The motif matching scores for USF1 binding sites. The USF1 binding sites are divided into 2 groups: before having TCF7L1 motifs in the flanking regions and after having TCF7L1 motifs in the flanking regions. (C) The expression level of the nearest genes for the 2 groups of USF1 binding sites. (D) Enriched pathways for the 2 groups of nearest genes.

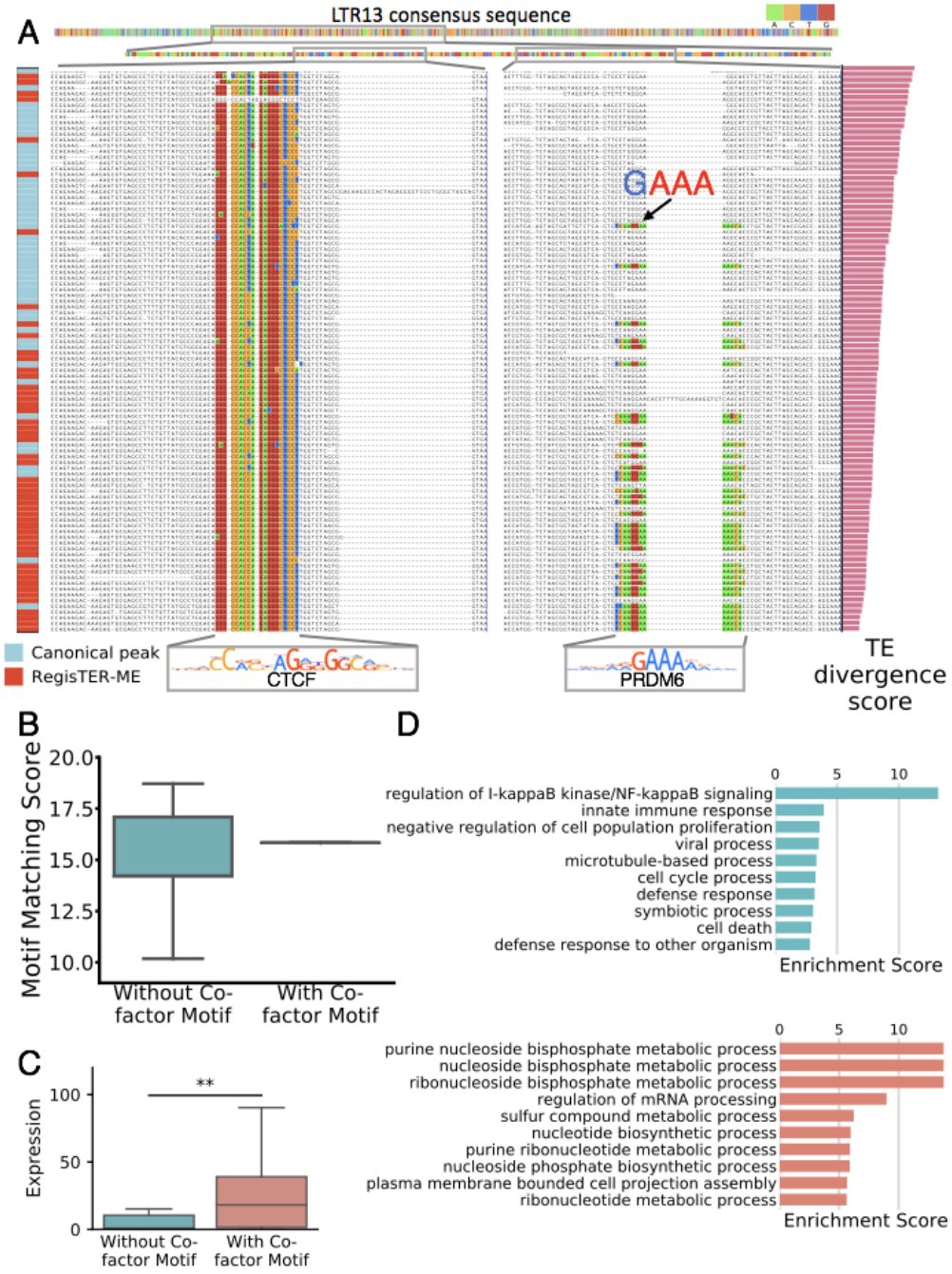


Figure C.6. Co-evolution of transcription factors and co-factors in TE-derived TF binding sites. (A) A sequence alignment for LTR13-derived CTCF binding regions. The left panel shows the identification of each binding site (blue: from canonical peaks, red: from RegisTER-ME peaks). The positions with motifs in the sequence alignment result are colored based on the nucleotides. The right panel shows the divergence score of each LTR13 sequence. (B) The motif matching scores for CTCF binding sites. The CTCF binding sites are divided into 2 groups: before having PRDM6 motifs in the flanking regions and after having PRDM6 motifs in the flanking regions. (C) The expression level of the nearest genes for the 2 groups of CTCF binding sites. (D) Enriched pathways for the 2 groups of nearest genes.

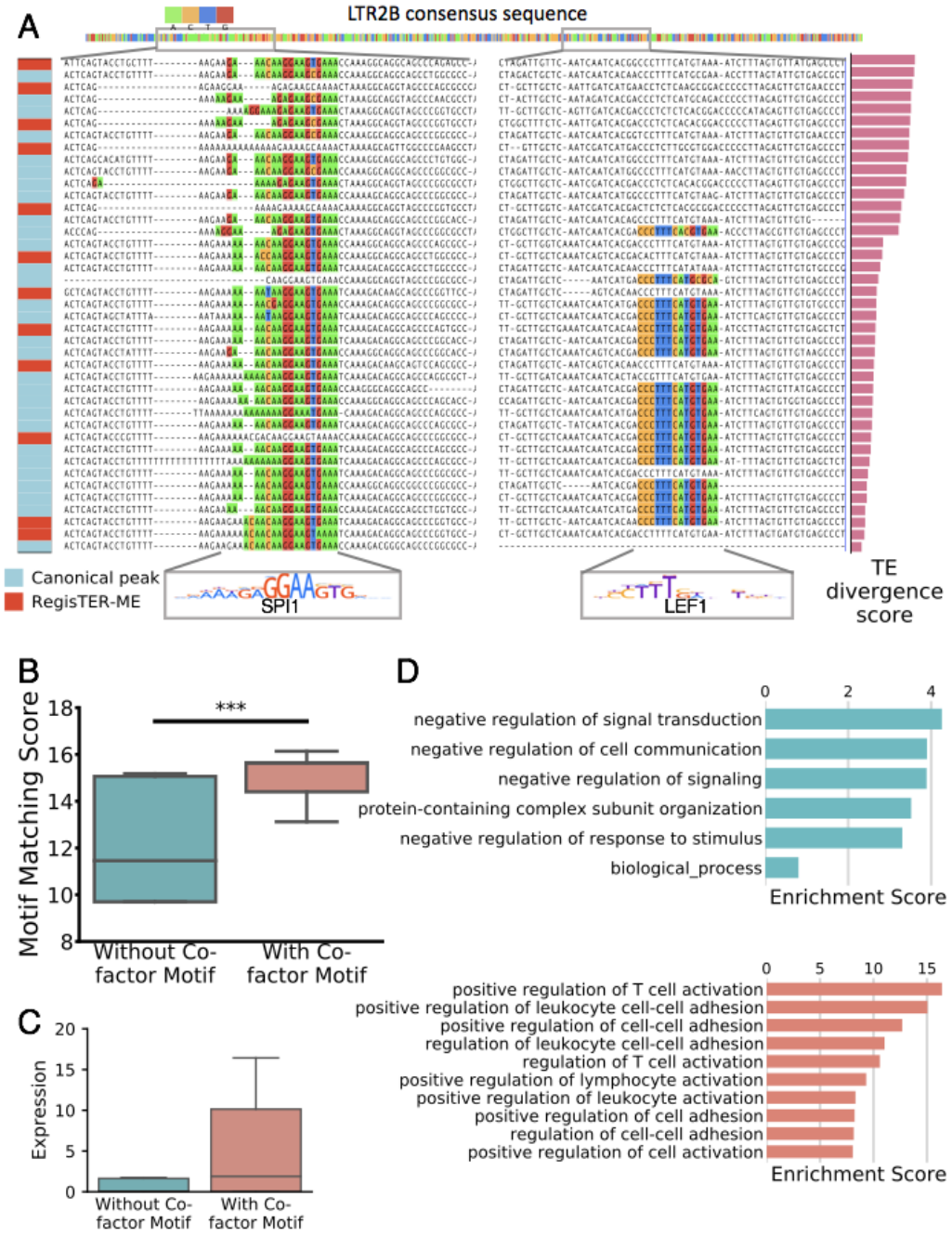


Figure C.7. Co-evolution of transcription factors and co-factors in TE-derived TF binding sites. (A) A sequence alignment for LTR2B-derived SPI1 binding regions. The left panel shows the identification of each binding site (blue: from canonical peaks, red: from RegisTER-ME peaks). The positions with motifs in the sequence alignment result are colored based on the nucleotides. The right panel shows the divergence score of each LTR2B sequence. (B) The motif matching scores for SPI1 binding sites. The SPI1 binding sites are divided into 2 groups: before having LEF1 motifs in the flanking regions and after having LEF1 motifs in the flanking regions. (C) The expression level of the nearest genes for the 2 groups of SPI1 binding sites. (D) Enriched pathways for the 2 groups of nearest genes.

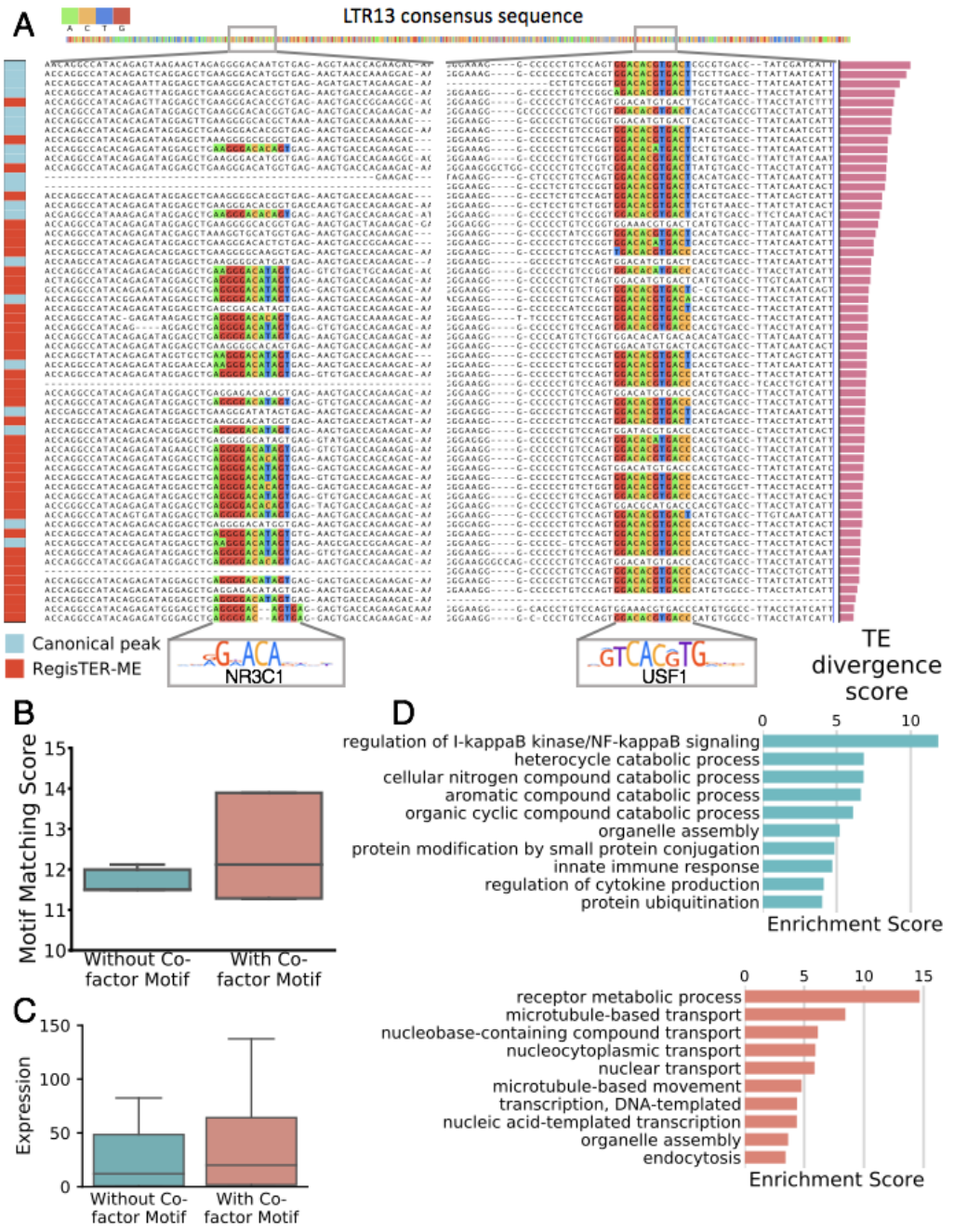


Figure C.8. Co-evolution of transcription factors and co-factors in TE-derived TF binding sites. (A) A sequence alignment for LTR13-derived USF1 binding regions. The left panel shows the identification of each binding site (blue: from canonical peaks, red: from RegisTER-ME peaks). The positions with motifs in the sequence alignment result are colored based on the nucleotides. The right panel shows the divergence score of each LTR13 sequence. (B) The motif matching scores for USF1 binding sites. The USF1 binding sites are divided into 2 groups: before having NR3C1 motifs in the flanking regions and after having NR3C1 motifs in the flanking regions. (C) The expression level of the nearest genes for the 2 groups of USF1 binding sites. (D) Enriched pathways for the 2 groups of nearest genes.

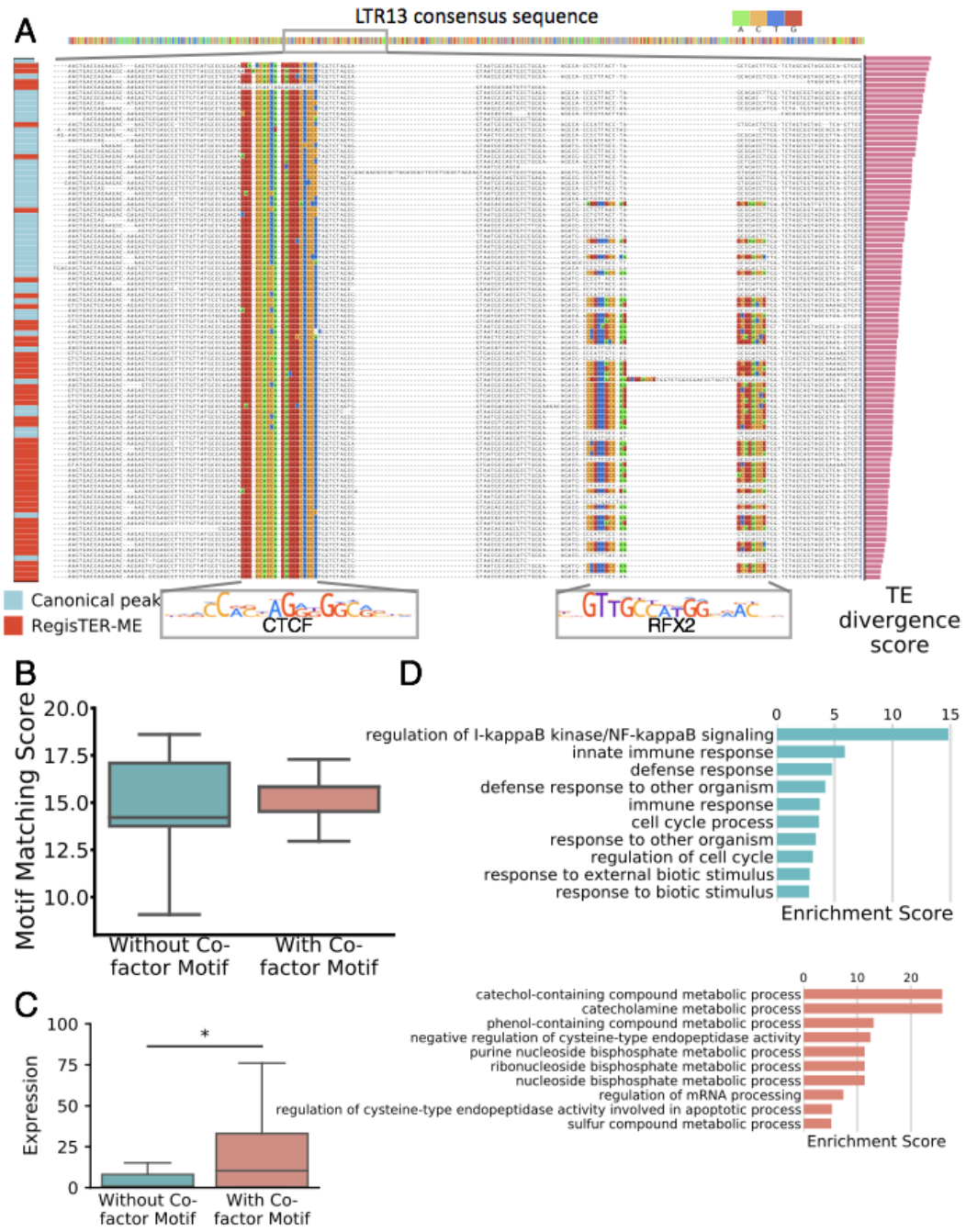


Figure C.9. Co-evolution of transcription factors and co-factors in TE-derived TF binding sites. (A) A sequence alignment for LTR13-derived CTCF binding regions. The left panel shows the identification of each binding site (blue: from canonical peaks, red: from RegisTER-ME peaks). The positions with motifs in the sequence alignment result are colored based on the nucleotides. The right panel shows the divergence score of each LTR13 sequence. (B) The motif matching scores for CTCF binding sites. The CTCF binding sites are divided into 2 groups: before having RFX2 motifs in the flanking regions and after having RFX2 motifs in the flanking regions. (C) The expression level of the nearest genes for the 2 groups of CTCF binding sites. (D) Enriched pathways for the 2 groups of nearest genes.

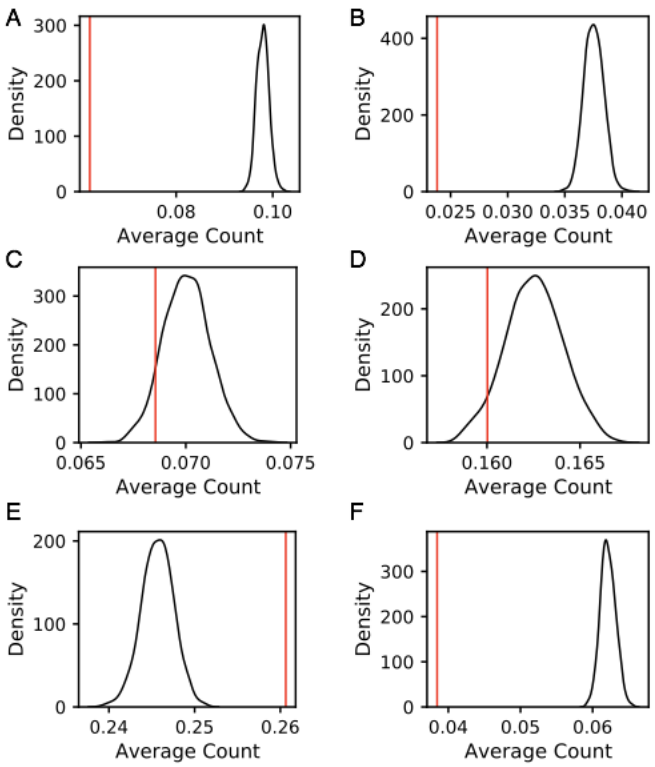


Figure C.10. Average number of TE-derived enhancers that regulate a gene among 1000 times of TE label shuffles. (A) Average number of ERV1-derived enhancers among shuffled ERV1 background. (B) Average number of ERVL-derived enhancers among shuffled ERVL background. (C) Average number of L1-derived enhancers among shuffled L1 background. (D) Average number of L2-derived enhancers among shuffled L2 background.(E) Average number of MIR-derived enhancers among shuffled MIR background. (F) Average number of ERVL-MaLR derived enhancers among shuffled ERVL-MaLR background.

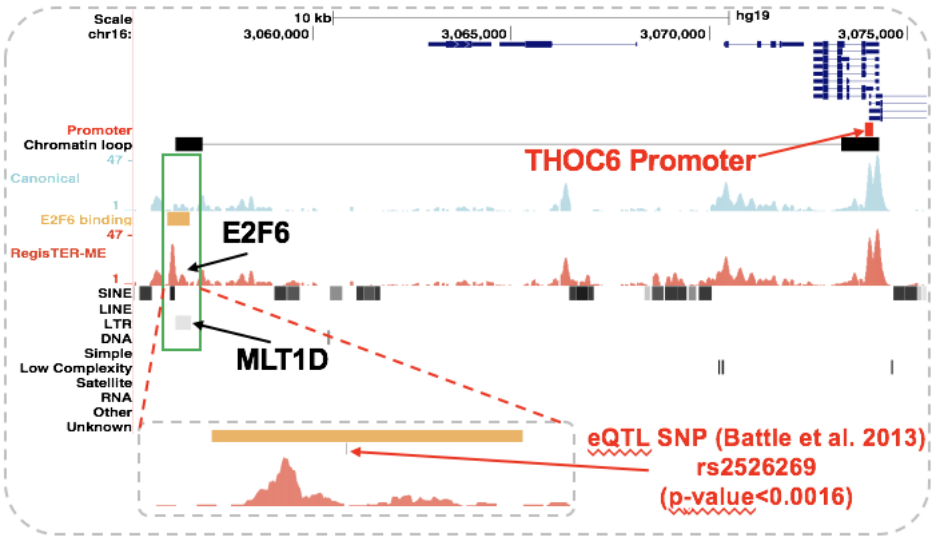


Figure C.11. A real example of eQTL interpretation in a E2F6 ChIP-seq dataset in K562.

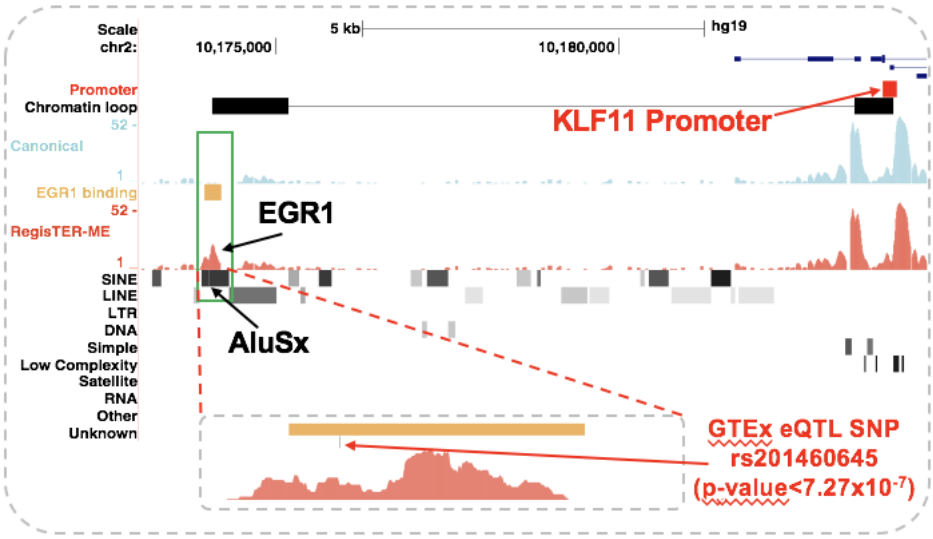


Figure C.12. A real example of eQTL interpretation in a EGR1 ChIP-seq dataset in K562.

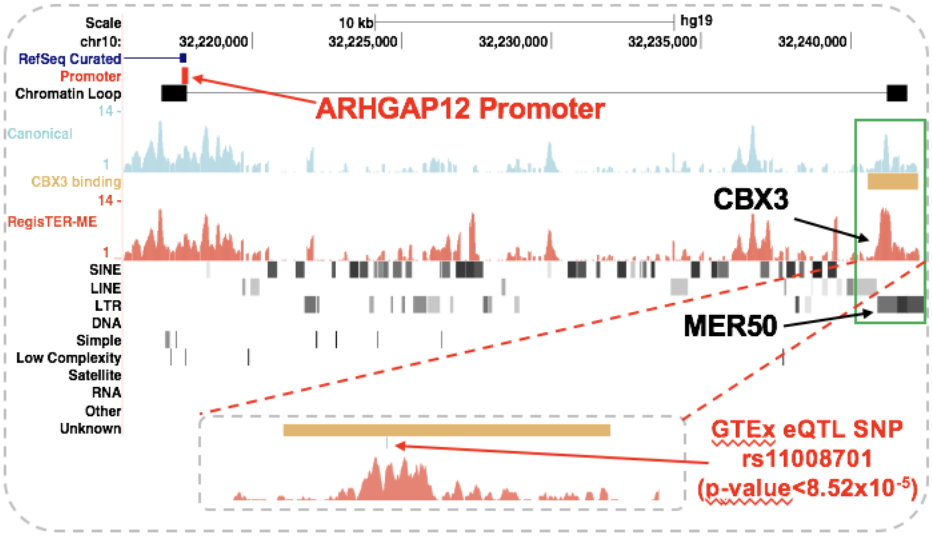


Figure C.13. A real example of eQTL interpretation in a CBX3 ChIP-seq dataset in K562.

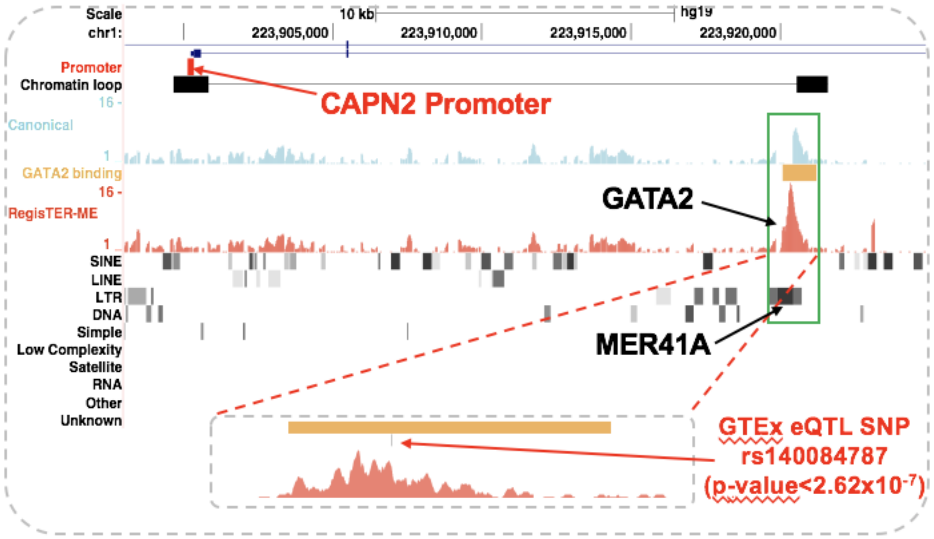


Figure C.14. A real example of eQTL interpretation in a GATA2 ChIP-seq dataset in K562.48

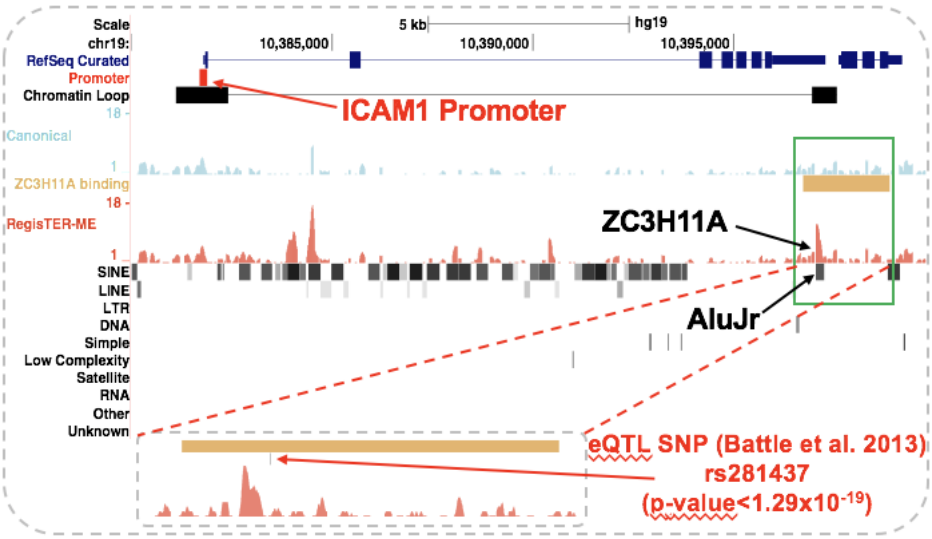


Figure C.15. A real example of eQTL interpretation in a ZC3H11A ChIP-seq dataset in K562.

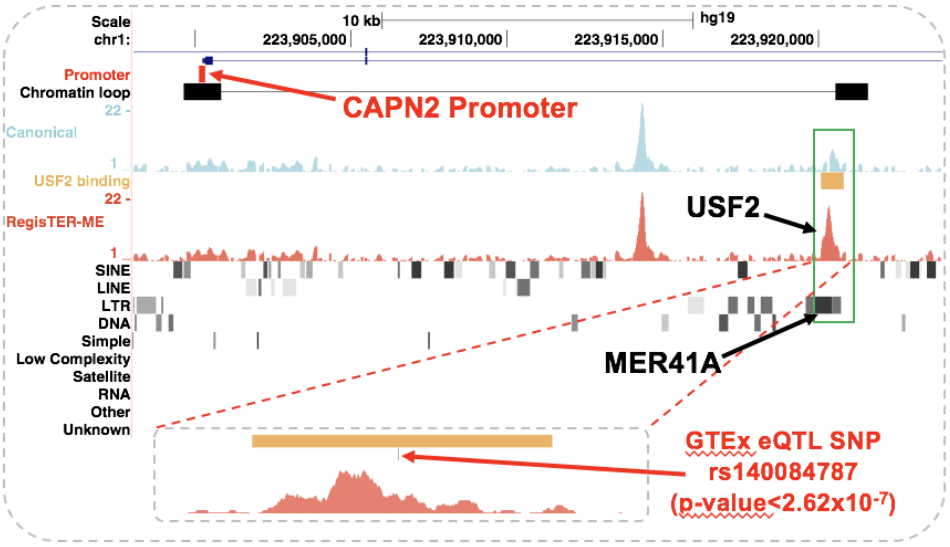


Figure C.16. A real example of eQTL interpretation in a USF2 ChIP-seq dataset in K562.

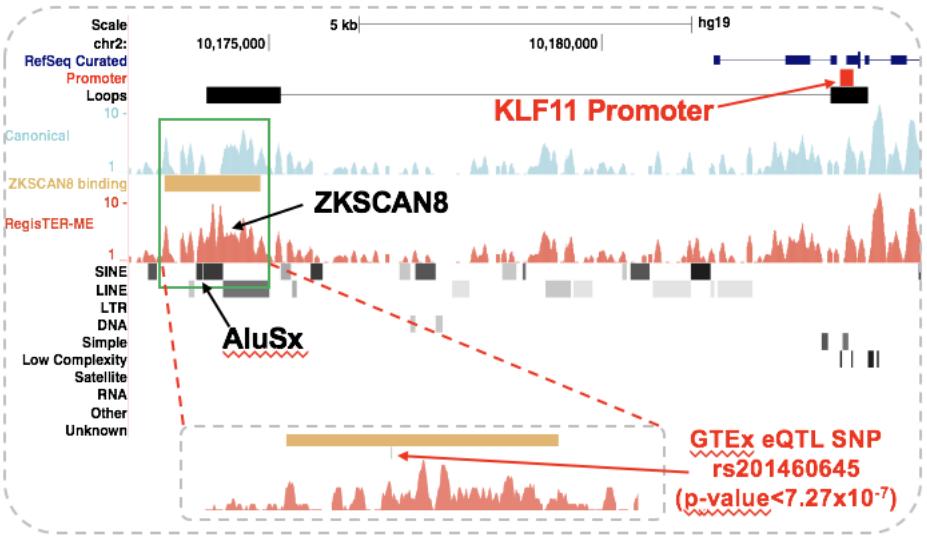


Figure C.17. A real example of eQTL interpretation in a ZKSCAN8 ChIP-seq dataset in K562.

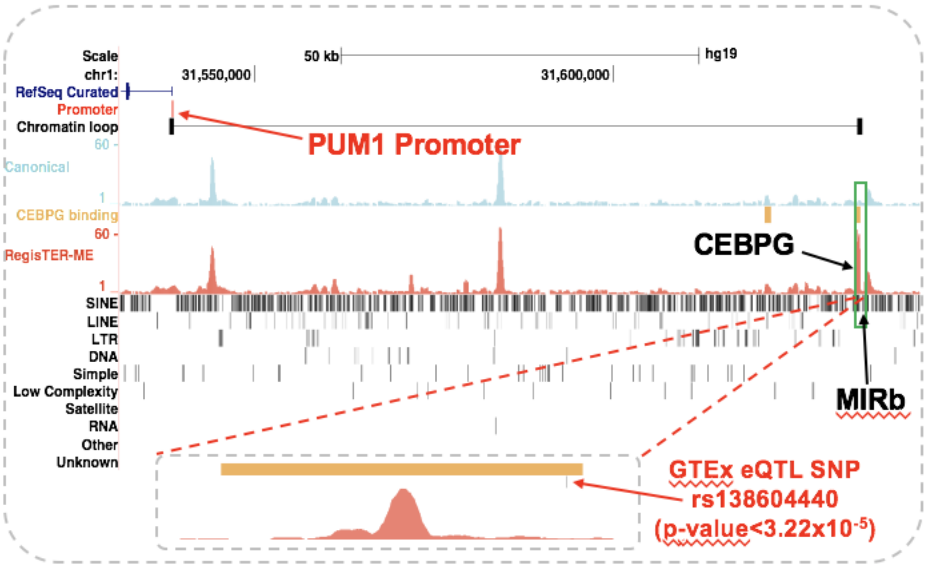


Figure C.18. A real example of eQTL interpretation in a CEBPG ChIP-seq dataset in K562.

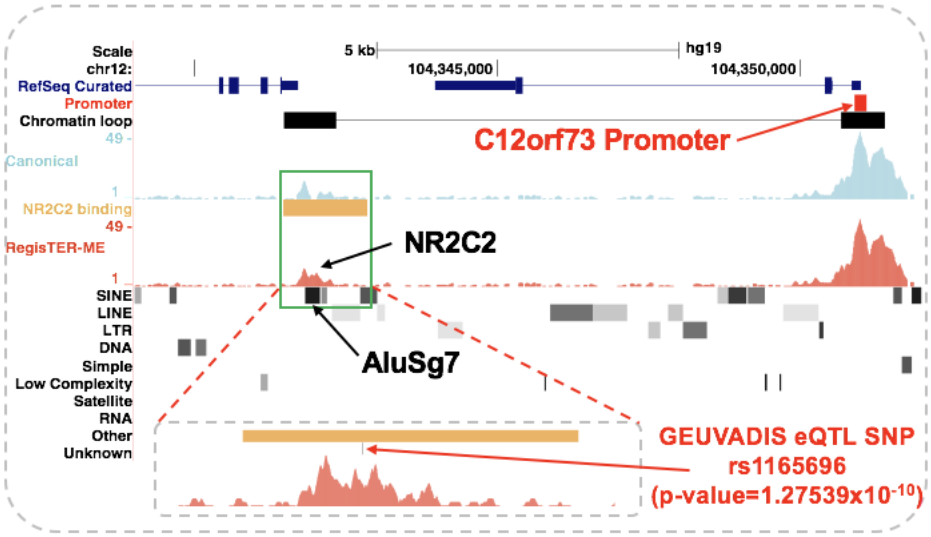


Figure C.19. A real example of eQTL interpretation in a NR2C2 ChIP-seq dataset in K562.

BIBLIOGRAPHY

BIBLIOGRAPHY

1. Walters KA (2015) Role of androgens in normal and pathological ovarian function. Reproduction 149(4):R193-218.

2. Walters KA, Simanainen U, & Gibson DA (2016) Androgen action in female reproductive physiology. Curr Opin Endocrinol Diabetes Obes 23(3):291-296.

3. Walters KA, Simanainen U, & Handelsman DJ (2010) Molecular insights into androgen actions in male and female reproductive function from androgen receptor knockout models. Hum Reprod Update 16(5):543-558.

4. Prizant H, Gleicher N, & Sen A (2014) Androgen actions in the ovary: balance is key. J Endocrinol 222(3):R141-151.

5. Sen A, et al. (2014) Androgens regulate ovarian follicular development by increasing follicle stimulating hormone receptor and microRNA-125b expression. Proc Natl Acad Sci U S A 111(8):3008-3013.

6. Franks S & Hardy K (2018) Androgen Action in the Ovary. Front Endocrinol (Lausanne) 9:452.

7. Stener-Victorin E, et al. (2020) Animal Models to Understand the Etiology and Pathophysiology of Polycystic Ovary Syndrome. Endocr Rev 41(4).

8. Walters KA (2016) Androgens in polycystic ovary syndrome: lessons from experimental models. Curr Opin Endocrinol Diabetes Obes 23(3):257-263.

9. Walters KA, Bertoldo MJ, & Handelsman DJ (2018) Evidence from animal models on the pathogenesis of PCOS. Best Pract Res Clin Endocrinol Metab 32(3):271-281.

10. Walters KA, Rodriguez Paris V, Aflatounian A, & Handelsman DJ (2019) Androgens and ovarian function: translation from basic discovery research to clinical impact. J Endocrinol 242(2):R23-R50.

11. Hu YC, et al. (2004) Subfertility and defective folliculogenesis in female mice lacking androgen receptor. Proc Natl Acad Sci U S A 101(31):11209-11214.

12. Shiina H, et al. (2006) Premature ovarian failure in androgen receptor-deficient mice. Proc Natl Acad Sci U S A 103(1):224-229.

13. Walters KA, et al. (2007) Female mice haploinsufficient for an inactivated androgen receptor (AR) exhibit age-dependent defects that resemble the AR null phenotype of dysfunctional late follicle development, ovulation, and fertility. Endocrinology 148(8):3674-3684.

14. Walters KA, et al. (2012) Targeted loss of androgen receptor signaling in murine granulosa cells of preantral and antral follicles causes female subfertility. Biol Reprod 87(6):151.

15. Sen A & Hammes SR (2010) Granulosa cell-specific androgen receptors are critical regulators of ovarian development and function. Mol Endocrinol 24(7):1393-1403.

16. Ma Y, et al. (2017) Androgen Receptor in the Ovary Theca Cells Plays a Critical Role in Androgen-Induced Reproductive Dysfunction. Endocrinology 158(1):98-108.

17. Wu S, et al. (2014) Conditional knockout of the androgen receptor in gonadotropes reveals crucial roles for androgen in gonadotropin synthesis and surge in female mice. Mol Endocrinol 28(10):1670-1681.

18. Caldwell ASL, et al. (2017) Neuroendocrine androgen action is a key extraovarian mediator in the development of polycystic ovary syndrome. Proc Natl Acad Sci U S A 114(16):E3334-E3343.

19. Laird M, et al. (2017) Androgen Stimulates Growth of Mouse Preantral Follicles In Vitro: Interaction With Follicle-Stimulating Hormone and With Growth Factors of the TGFbeta Superfamily. Endocrinology 158(4):920-935.

20. HUA WANG KA, HARUO HAGIWARA, LIU XIAOWEI, NOBUMASA KIKUCHI, YUMIKO ABE, KIYOHIKO YAMADA, RIASAT FATIMA, AND HIDEKI MIZUNUMA (2001) Effect of Adrenal and Ovarian Androgens on Type 4 Follicles Unresponsive to FSH in Immature Mice. Endocrinology 142(11):4930–4936.

21. Hickey TE, et al. (2005) Androgens augment the mitogenic effects of oocyte-secreted factors and growth differentiation factor 9 on porcine granulosa cells. Biol Reprod 73(4):825-832.

22. Hickey TE, Marrocco DL, Gilchrist RB, Norman RJ, & Armstrong DT (2004) Interactions between androgen and growth factors in granulosa cell subtypes of porcine antral follicles. Biol Reprod 71(1):45-52.

23. A. A. Murray RGG, V. Allison and N. Spears (1998) Effect of androgens on the development of mouse follicles growing in vitro. Journal of Reproduction and Ferlility(113):27-33.

24. P.R.Casson1 MSL, M.D.Pisarska, S.A.Carson and J.E.Buster (2000) Dehydroepiandrosterone supplementation augments ovarian stimulation in poor responders: a case series. Human Reproduction 15(10):2129–2132.

25. Barad D & Gleicher N (2006) Effect of dehydroepiandrosterone on oocyte and embryo yields, embryo grade and cell number in IVF. Hum Reprod 21(11):2845-2849.

26. Barad DH GN (2005) Increased oocyte production after treatment with dehydroepiandrosterone. Fertil Steril 84(3):756.

27. Mamas L & Mamas E (2009) Premature ovarian failure and dehydroepiandrosterone. Fertil Steril 91(2):644-646.

28. Hyman JH ME, Rabinowitz R, Tsafrir A, Gal M, Alerhand S, et al (2013) DHEA supplementation may improve IVF outcome in poor responders: a proposed mechanism. Eur J Obstet Gynecol Reprod Biol 168(1):49-53.

29. So¨nmezer M OB, Cil AP, Ozkavukc¸u S, Taşc¸i T, Olmuş H, et al (2009) Dehydroepiandrosterone sup- plementation improves ovarian response and cycle outcome in poor responders. Reprod Biomed Online 19(4):508–513.

30. Sunkara SK CA (2011) Androgen pretreatment in poor responders undergoing controlled ovarian stimulation and in vitro fertilization treatment. Fertil Steril 95(8):e73–74.

31. nzalez-Comadran M DnM, Solà I, Fa´bregues F, Carreras R, Checa MA, et al (2012) Effects of trans- dermal testosterone in poor responders undergoing IVF: systematic review and meta-analysis. Reprod Biomed Online 25(5):450–459.

32. Hammes SR LE (2011) Minireview: Recent advances in extranuclear steroid receptor actions. Endocrinology 152(12):4489–4495.

33. Hammes SR LE (2007) Extranuclear steroid receptors: nature and actions. Endocr Rev 28(7):726–741.

34. Levin ER HS (2016) Nuclear receptors outside the nucleus: extranuclear signalling by steroid receptors. Nat Rev Mol Cell Biol 17(12):783–797.

35. Sen A DCI, Defranco DB, Deng FM, Melamed J, Kapur P, et al (2012) Paxillin mediates extranuclear and intranuclear signaling in prostate cancer proliferation. J Clin Invest 122(7):2469–2481.

36. Sen A OMK, Wang Z, Raj GV, Defranco DB, Hammes SR, et al (2010) Paxillin regulates androgen and epidermal growth factor-induced MAPK signaling and cell proliferation in prostate cancer cells. J Biol Chem 285(37):28787–28795.

37. Sen A PH, Hammes SR (2011) Understanding extranuclear (nongenomic) androgen signaling: what a frog oocyte can tell us about human biology. Steroids 76(9):822–828.

38. Ma X HE, Biswas A, Seger C, Prizant H, Hammes SR, et al (2017) Androgens Regulate Ovarian Gene Expression Through Modulation of Ezh2 Expression and Activity. Steroids 158(9):2944–2954.

39. Vire´ E BC, Deplus R, Blanchon L, Fraga M, Didelot C, et al (2006) The Polycomb group protein EZH2 directly controls DNA methylation. Nature 439(7078):871–874.

40. Anthony M. Bolger ML, and Bjoern Usadel (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120.

41. Kevin L Howe PA, James Allen, et al (2021) Ensembl 2021. Nucleic Acids Res 49(1):884–891.

42. Alexander Dobin CAD, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, Thomas R Gingeras (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29(1):15-21.

43. Liao Y SGaSW (2013) The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Research 41(10):e108.

44. Love MI HW, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15(550).

45. Andrews S (2010) FastQC: A Quality Control Tool for High Throughput Sequence Data.

46. Langmead B & Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9(4):357-359.

47. Zhang Y LT, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biology 9(9):137.

48. Cairns J F-PP, Wingett SW, Várnai C et al (2016) CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. Genome Biology 17(1):127.

49. Vian L PA, Rao SSP, Kieffer-Kwon KR et al (2018) The Energetics and Physiological Impact of Cohesin Extrusion. Cell 173(5):1165-1178.

50. Guoqiang Li YL, Yanxiao Zhang, Rongxin Fang, Manolis Kellis, Bing Ren (2018) Simultaneous profiling of DNA methylation and chromatin architecture in mixed populations and in single cells. bioRxiv.

51. Chaudhri VK D-SK, Wu Z, Shrestha M et al (2020) Charting the cis-regulome of activated B cells by coupling structural and functional genomics. Nat Immunol 21(2):210-220.

52. Rao SS HM, Durand NC, Stamenova EK et al (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159(7):1665-1680.

53. Timothy L. Bailey JJ, Charles E. Grant, William S. Noble (2015) The MEME Suite. Nucleic Acids Research 43(1):39-49.

54. Shobhit Gupta JS, Timothy Bailey and William Stafford Noble (2007) Quantifying similarity between motifs. Genome Biology 8(2):24.

55. Kulakovskiy IV, et al. (2018) HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res 46(D1):D252-D259.

56. Yue F CY, Breschi A, Vierstra J, Wu W, Ryba T, et al (2014) A comparative encyclopedia of DNA ele- ments in the mouse genome. Nature 515(7527):355–364.

57. Nilsson EE SM (2003) Bone morphogenetic protein-4 acts as an ovarian follicle survival factor and promotes primordial follicle development. Biol Reprod 69(4):1265–1272.

58. P N (2015) Genetic Models for the Study of Luteinizing Hormone Receptor Function. Front Endocrinol (Lausanne) 6:152.

59. Richards JS H-GI, Gonzalez-Robayna I, Teuling E, Lo Y, Boerboom D, et al (2005) Regulated expression of ADAMTS family members in follicles and cumulus oocyte complexes: evidence for specific and redundant patterns during ovulation. Biol Reprod 72(5):1241–1255.

60. Light A HS (2015) LH-Induced Steroidogenesis in the Mouse Ovary, but Not Testis, Requires Matrix Metalloproteinase 2- and 9-Mediated Cleavage of Upregulated EGF Receptor Ligands. Biol Reprod 93(3):65.

61. Carbajal L BA, Niswander LM, Prizant H, Hammes SR (2011) GPCR/EGFR cross talk is conserved in gonadal and adrenal steroidogenesis but is uniquely regulated by matrix metalloproteinases 2 and 9 in the ovary. Mol Endocrinol 25(6):1055–1065.

62. Jamnongjit M GA, Hammes SR (2005) Epidermal growth factor receptor signaling is required for normal ovarian steroidogenesis and oocyte maturation. Proc Natl Acad Sci USA 102(45):16257– 16262.

63. Fitzpatrick SL RJ (1991) Regulation of cytochrome P450 aromatase messenger ribonucleic acid and activity by steroids and gonadotropins in rat granulosa cells. Endocrinology 129(3):1452–1462.

64. Tetsuka M HS (1997) Differential regulation of aromatase and androgen receptor in granulosa cells. J Steroid Biochem Mol Biol 61(3-6):233–239.

65. Vendola KA ZJ, Adesanya OO, Weil SJ, Bondy CA (1998) Androgens stimulate early stages of follicular growth in the primate ovary. J Clin Invest 101(12):2622–2629.

66. Weil S VK, Zhou J, Bondy CA (1999) Androgen and follicle-stimulating hormone interactions in primate ovarian follicle development. J Clin Endocrinol Metab 84(8):2951–2956.

67. Weil SJ VK, Zhou J, Adesanya OO, Wang J, Okafor J, et al (1998) Androgen receptor gene expres- sion in the primate ovary: cellular localization, regulation, and functional correlations. J Clin Endocrinol Metab 83(7):2479–2485.

68. Vendola K ZJ, Wang J, Famuyiwa OA, Bievre M, Bondy CA (1999) Androgens promote oocyte insulin- like growth factor I expression and initiation of follicle development in the primate ovary. Biol Reprod 61(2):353–357.

69. Owens LA KS, Lerner A, Christopoulos G, Lavery S, Hanyaloglu AC, et al (2019) Gene Expres- sion in Granulosa Cells From Small Antral Follicles From Women With or Without Polycystic Ovaries. J Clin Endocrinol Metab 104(12):6182–6192.

70. Ambekar AS KD, Pinto SM, Sharma R, Hinduja I, Zaveri K, et al (2015) Proteomics of follicular fluid from women with polycystic ovary syndrome suggests molecular defects in follicular development. Clin Endocrinol Metab 100(2):744–753.

71. Salehi E AR, Moeini A, Yamini N, Asadi E, Khosravizadeh Z, et al (2017) Apoptotic biomarkers in cumulus cells in relation to embryo quality in polycystic ovary syndrome. Arch Gynecol Obstet 296(6):1219–1227.

72. Peng Y ZW, Yang P, Tian Y, Su S, Zhang C, et al (2017) ERBB4 Confers Risk for Polycystic Ovary Syndrome in Han Chinese. Sci Rep 7:42000.

73. Li L LK, Choi BC, Baek KH (2013) Relationship between leptin receptor and polycystic ovary syndrome. Gene 527(1):71–74.

74. Nord,A.S., Blow,M.J., Attanasio,C., Akiyama,J.A., Holt,A., Hosseini,R., Phouanenavong,S., Plajzer-Frick,I., Shoukry,M., Afzal,V. et al. (2013) Rapid and pervasive changes in genome-wide

enhancer usage during mammalian development. Cell, 155, 1521–1531.

75. Schoenfelder,S. and Fraser,P. (2019) Long-range enhancer–promoter contacts in gene expression control. Nat. Rev. Genet., 20, 437–455.

76. Vicente,C.T., Edwards,S.L., Hillman,K.M., Kaufmann,S., Mitchell,H., Bain,L., Glubb,D.M., Lee,J.S., French,J.D. and Ferreira,M.A. (2015) Long-range modulation of PAG1 expression by 8q21 allergy risk variants. Am. J. Hum. Genet., 97, 329–336.

77. Martin,P., McGovern,A., Orozco,G., Duffus,K., Yarwood,A., Schoenfelder,S., Cooper,N.J., Barton,A., Wallace,C., Fraser,P. et al. (2015) Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. Nat. Commun., 6, 10069.

78. Deng,W., Lee,J., Wang,H., Miller,J., Reik,A., Gregory,P.D., Dean,A. and Blobel,G.A. (2012) Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. Cell, 149, 1233–1244.

79. Ragoczy,T., Bender,M.A., Telling,A., Byron,R. and Groudine,M. (2006) The locus control region is required for association of the murine beta-globin locus with engaged transcription factories during erythroid maturation. Genes Dev., 20, 1447–1457.

80. Lettice,L.A., Heaney,S.J., Purdie,L.A., Li,L., de Beer,P., Oostra,B.A., Goode,D., Elgar,G., Hill,R.E. and de Graaff,E. (2003) A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum. Mol. Genet., 12, 1725–1735.

81. Jeong,Y., El-Jaick,K., Roessler,E., Muenke,M. and Epstein,D.J. (2006) A functional screen for sonic hedgehog regulatory elements across a 1 Mb interval identifies long-range ventral forebrain enhancers. Development, 133, 761–772.

82. Sagai,T., Amano,T., Tamura,M., Mizushina,Y., Sumiyama,K. and Shiroishi,T. (2009) A cluster of three long-range enhancers directs regional Shh expression in the epithelial linings. Development, 136, 1665–1674.

83. Smemo,S., Tena,J.J., Kim,K.H., Gamazon,E.R., Sakabe,N.J., Gomez-Marin,C., Aneas,I., Credidio,F.L., Sobreira,D.R., Wasserman,N.F. et al. (2014) Obesity-associated variants within FTO form long-range functional connections with IRX3. Nature, 507, 371–375.

84. Dryden,N.H., Broome,L.R., Dudbridge,F., Johnson,N., Orr,N., Schoenfelder,S., Nagano,T., Andrews,S., Wingett,S., Kozarewa,I. et al. (2014) Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. Genome Res., 24, 1854–1868.

85. McGovern,A., Schoenfelder,S., Martin,P., Massey,J., Duffus,K., Plant,D., Yarwood,A., Pratt,A.G., Anderson,A.E., Isaacs,J.D. et al. (2016) Capture Hi-C identifies a novel causal gene, IL20RA, in the pan-autoimmune genetic susceptibility region 6q23. Genome Biol., 17, 212.

86. Jager,R., Migliorini,G., Henrion,M., Kandaswamy,R., Speedy,H.E., Heindl,A., Whiffin,N., Carnicer,M.J., Broome,L., Dryden,N. et al. (2015) Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. Nat. Commun., 6, 6178.

87. Shlyueva,D., Stampfel,G. and Stark,A. (2014) Transcriptional enhancers: from properties to genome-wide predictions. Nat. Rev. Genet., 15, 272–286.

88. Buecker,C. and Wysocka,J. (2012) Enhancers as information integration hubs in development: lessons from genomics. Trends Genet., 28, 276–284.

89. Consortium,E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. Nature, 489, 57–74.

90. Roadmap Epigenomics,C., Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J. et al. (2015) Integrative analysis of 111 reference human epigenomes. Nature, 518, 317–330.

91. Hoffman,M.M., Buske,O.J., Wang,J., Weng,Z., Bilmes,J.A. and Noble,W.S. (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat. Methods, 9, 473–476.

92. Ernst,J. and Kellis,M. (2017) Chromatin-state discovery and genome annotation with ChromHMM. Nat. Protoc., 12, 2478–2492.

93. Pennacchio,L.A., Bickmore,W., Dean,A., Nobrega,M.A. and Bejerano,G. (2013) Enhancers: five essential questions. Nat. Rev. Genet., 14, 288–295.

94. Mumbach,M.R., Satpathy,A.T., Boyle,E.A., Dai,C., Gowen,B.G., Cho,S.W., Nguyen,M.L., Rubin,A.J., Granja,J.M., Kazane,K.R. et al. (2017) Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. Nat. Genet., 49, 1602–1612. 95.Gondor,A. and Ohlsson,R. (2009) Chromosome crosstalk in three dimensions. Nature, 461, 212–217.

96. Kvon,E.Z., Kamneva,O.K., Melo,U.S., Barozzi,I., Osterwalder,M., Mannion,B.J., Tissieres,V., Pickle,C.S., Plajzer-Frick,I., Lee,E.A. et al. (2016) Progressive loss of function in a limb enhancer during snake evolution. Cell, 167, 633–642.

97. Claussnitzer,M., Dankel,S.N., Kim,K.H., Quon,G., Meuleman,W., Haugen,C., Glunk,V., Sousa,I.S., Beaudry,J.L., Puviindran,V. et al. (2015) FTO obesity variant circuitry and adipocyte browning in humans. N. Engl. J. Med., 373, 895–907.

98. Dekker,J., Rippe,K., Dekker,M. and Kleckner,N. (2002) Capturing chromosome conformation. Science, 295, 1306–1311.

99. Zhao,Z., Tavoosidana,G., Sjolinder,M., Gondor,A., Mariano,P., Wang,S., Kanduri,C., Lezcano,M., Sandhu,K.S., Singh,U. et al. (2006) Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat. Genet., 38, 1341–1347.

100. Dostie,J., Richmond,T.A., Arnaout,R.A., Selzer,R.R., Lee,W.L., Honan,T.A., Rubio,E.D., Krumm,A., Lamb,J., Nusbaum,C. et al. (2006) Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res., 16, 1299–1309.

101. Lieberman-Aiden,E., van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O. et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science, 326, 289–293.

102. Rao,S.S., Huntley,M.H., Durand,N.C., Stamenova,E.K., Bochkov,I.D., Robinson,J.T., Sanborn,A.L., Machol,I., Omer,A.D., Lander,E.S. et al. (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell, 159, 1665–1680.

103. Jung,I., Schmitt,A., Diao,Y., Lee,A.J., Liu,T., Yang,D., Tan,C., Eom,J., Chan,M., Chee,S. et al. (2019) A compendium of promoter-centered long-range chromatin interactions in the human genome. Nat. Genet., 51, 1442–1449.

104. Mifsud,B., Tavares-Cadete,F., Young,A.N., Sugar,R., Schoenfelder,S., Ferreira,L., Wingett,S.W., Andrews,S., Grey,W., Ewels,P.A. et al. (2015) Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat. Genet., 47, 598–606.

105. Schoenfelder,S., Javierre,B.M., Furlan-Magaril,M., Wingett,S.W. and Fraser,P. (2018) Promoter capture Hi-C: high-resolution, genome-wide profiling of promoter interactions. J. Vis. Exp., 136, 57320.

106. Fullwood,M.J. and Ruan,Y. (2009) ChIP-based methods for the identification of long-range chromatin interactions. J. Cell. Biochem., 107, 30–39.

107. Li,X., Luo,O.J., Wang,P., Zheng,M., Wang,D., Piecuch,E., Zhu,J.J., Tian,S.Z., Tang,Z., Li,G. et al. (2017) Long-read ChIA-PET for base-pair-resolution mapping of haplotype-specific chromatin interactions. Nat. Protoc., 12, 899–915.

108. Smith,E.M., Lajoie,B.R., Jain,G. and Dekker,J. (2016) Invariant TAD boundaries constrain cell-type-specific looping interactions between promoters and distal elements around the CFTR locus. Am. J. Hum. Genet., 98, 185–201.

109. Yardimci,G.G., Ozadam,H., Sauria,M.E.G., Ursu,O., Yan,K.K., Yang,T., Chakraborty,A., Kaul,A., Lajoie,B.R., Song,F. et al. (2019) Measuring the reproducibility and quality of Hi-C data. Genome Biol., 20, 57.

110. Li,G., Fullwood,M.J., Xu,H., Mulawadi,F.H., Velkov,S., Vega,V., Ariyaratne,P.N., Mohamed,Y.B., Ooi,H.S., Tennakoon,C. et al. (2010) ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. Genome Biol., 11, R22.

111. Meuleman,W., Muratov,A., Rynes,E., Halow,J., Lee,K., Bates,D., Diegel,M., Dunn,D., Neri,F., Teodosiadis,A. et al. (2020) Index and biological spectrum of human DNase I hypersensitive sites. Nature, 584, 244–251.

112. Consortium,E.P., Moore,J.E., Purcaro,M.J., Pratt,H.E., Epstein,C.B., Shoresh,N., Adrian,J., Kawli,T., Davis,C.A., Dobin,A. et al. (2020) Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature, 583, 699–710.

113. Yen,A. and Kellis,M. (2015) Systematic chromatin state comparison of epigenomes associated with diverse properties including sex and tissue type. Nat. Commun., 6, 7973.

114. Roy,S., Siahpirani,A.F., Chasman,D., Knaack,S., Ay,F., Stewart,R., Wilson,M. and

Sridharan,R. (2015) A predictive modeling approach for cell line-specific long-range regulatory interactions. Nucleic Acids Res., 43, 8694–8712.

115. Hait,T.A., Amar,D., Shamir,R. and Elkon,R. (2018) FOCS: a novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer–promoter map. Genome Biol., 19, 56.

116. Gao,T. and Qian,J. (2019) EAGLE: an algorithm that utilizes a small number of genomic features to predict tissue/cell type-specific enhancer-gene interactions. PLoS Comput. Biol., 15, e1007436.

117. Cao,Q., Anyansi,C., Hu,X., Xu,L., Xiong,L., Tang,W., Mok,M.T.S., Cheng,C., Fan,X., Gerstein,M. et al. (2017) Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. Nat. Genet., 49, 1428–1436.

118. He,B., Chen,C., Teng,L. and Tan,K. (2014) Global view of enhancer–promoter interactome in human cells. Proc. Natl. Acad. Sci. USA, 111, E2191–E2199.

119. Whalen,S., Truty,R.M. and Pollard,K.S. (2016) Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. Nat. Genet., 48, 488–496.

120. Fishilevich,S., Nudel,R., Rappaport,N., Hadar,R., Plaschkes,I., Iny Stein,T., Rosen,N., Kohn,A., Twik,M., Safran,M. et al. (2017) GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database (Oxford), 2017, bax028.

121. Thurman,R.E., Rynes,E., Humbert,R., Vierstra,J., Maurano,M.T., Haugen,E., Sheffield,N.C., Stergachis,A.B., Wang,H., Vernot,B. et al. (2012) The accessible chromatin landscape of the human genome. Nature, 489, 75–82.

122. Corradin,O., Saiakhova,A., Akhtar-Zaidi,B., Myeroff,L., Willis,J., Cowper-Sal lari,R., Lupien,M., Markowitz,S. and Scacheri,P.C. (2014) Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. Genome Res., 24, 1–13.

123. Moore,J.E., Pratt,H.E., Purcaro,M.J. and Weng,Z. (2020) A curated benchmark of enhancer-gene interactions for evaluating enhancer-target gene prediction methods. Genome Biol., 21, 17.

124. Cao,F. and Fullwood,M.J. (2019) Inflated performance measures in enhancer–promoter interaction-prediction methods. Nat. Genet., 51, 1196–1198.

125. Whitaker,J.W., Nguyen,T.T., Zhu,Y., Wildberg,A. and Wang,W. (2015) Computational schemes for the prediction and annotation of enhancers from epigenomic assays. Methods, 72, 86–94.

126. Nolis,I.K., McKay,D.J., Mantouvalou,E., Lomvardas,S., Merika,M. and Thanos,D. (2009) Transcription factors mediate long-range enhancer–promoter interactions. Proc. Natl. Acad. Sci. USA, 106, 20222–20227.

127. Hnisz,D., Shrinivas,K., Young,R.A., Chakraborty,A.K. and Sharp,P.A. (2017) A phase

separation model for transcriptional control. Cell, 169, 13–23.

128. Quevedo,M., Meert,L., Dekker,M.R., Dekkers,D.H.W., Brandsma,J.H., van den Berg,D.L.C., Ozgur,Z., van,I.W.F.J., Demmers,J., Fornerod,M. et al. (2019) Mediator complex interaction partners organize the transcriptional network that defines neural stem cells. Nat. Commun., 10, 2669.

129. Maksimenko,O. and Georgiev,P. (2014) Mechanisms and proteins involved in long-distance interactions. Front Genet, 5, 28.

130. Li,Y., Haarhuis,J.H.I., Sedeno Cacciatore,A., Oldenkamp,R., van Ruiten,M.S., Willems,L., Teunissen,H., Muir,K.W., de Wit,E., Rowland,B.D. et al. (2020) The structural basis for cohesin-CTCF-anchored loops. Nature, 578, 472–476.

131. Beagan,J.A., Duong,M.T., Titus,K.R., Zhou,L., Cao,Z., Ma,J., Lachanski,C.V., Gillis,D.R. and Phillips-Cremins,J.E. (2017) YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. Genome Res., 27, 1139–1152.

132. Weintraub,A.S., Li,C.H., Zamudio,A.V., Sigova,A.A., Hannett,N.M., Day,D.S., Abraham,B.J., Cohen,M.A., Nabet,B., Buckley,D.L. et al. (2017) YY1 is a structural regulator of enhancer-promoter loops. Cell, 171, 1573–1588.

133. Morgan,S.L., Mariano,N.C., Bermudez,A., Arruda,N.L., Wu,F., Luo,Y., Shankar,G., Jia,L., Chen,H., Hu,J.F. et al. (2017) Manipulation of nuclear architecture through CRISPR-mediated

chromosomal looping. Nat. Commun., 8, 15993.

134. Zhang,K., Li,N., Ainsworth,R.I. and Wang,W. (2016) Systematic identification of protein combinations mediating chromatin looping. Nat. Commun., 7, 12249.

135. Wang,R., Wang,Y., Zhang,X., Zhang,Y., Du,X., Fang,Y. and Li,G. (2019) Hierarchical cooperation of transcription factors from integration analysis of DNA sequences, ChIP-Seq and ChIA-PET data. BMC Genomics, 20, 296.

136. Kato,M., Hata,N., Banerjee,N., Futcher,B. and Zhang,M.Q. (2004) Identifying combinatorial regulation of transcription factors and binding motifs. Genome Biol., 5, R56.

137. Michaelis,C., Ciosk,R. and Nasmyth,K. (1997) Cohesins: chromosomal proteins that prevent premature separation of sister chromatids. Cell, 91, 35–45.

138. Sanborn,A.L., Rao,S.S., Huang,S.C., Durand,N.C., Huntley,M.H., Jewett,A.I., Bochkov,I.D., Chinnappan,D., Cutkosky,A., Li,J. et al. (2015) Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc. Natl. Acad. Sci. U.S.A., 112, E6456–E6465.

139. Tan,K., Shlomi,T., Feizi,H., Ideker,T. and Sharan,R. (2007) Transcriptional regulation of protein complexes within and across species. Proc. Natl. Acad. Sci. U.S.A., 104, 1283–1288.

140. Szklarczyk,D., Gable,A.L., Lyon,D., Junge,A., Wyder,S., Huerta-Cepas,J., Simonovic,M., Doncheva,N.T., Morris,J.H., Bork,P. et al. (2019) STRING v11: protein–protein association

networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res., 47, D607–D613.

141. Harrow,J., Denoeud,F., Frankish,A., Reymond,A., Chen,C.K., Chrast,J., Lagarde,J., Gilbert,J.G., Storey,R., Swarbreck,D. et al. (2006) GENCODE: producing a reference annotation for ENCODE. Genome Biol., 7(Suppl.1), S4.

142. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. et al. (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol., 9, R137.

143. Amoutzias,G.D., Robertson,D.L., Van de Peer,Y. and Oliver,S.G. (2008) Choose your partners: dimerization in eukaryotic transcription factors. Trends Biochem. Sci., 33, 220–229.

144. Dixon,J.R., Selvaraj,S., Yue,F., Kim,A., Li,Y., Shen,Y., Hu,M., Liu,J.S. and Ren,B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature, 485, 376–380.

145. Akdemir,K.C., Le,V.T., Chandran,S., Li,Y., Verhaak,R.G., Beroukhim,R., Campbell,P.J., Chin,L., Dixon,J.R., Futreal,P.A. et al. (2020) Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. Nat. Genet., 52, 294–305.

146. Chesi,A., Wagley,Y., Johnson,M.E., Manduchi,E., Su,C., Lu,S., Leonard,M.E., Hodge,K.M., Pippin,J.A., Hankenson,K.D. et al. (2019) Genome-scale Capture C promoter interactions implicate effector genes at GWAS loci for bone mineral density. Nat. Commun., 10, 1260.

147. Pugacheva,E.M., Kubo,N., Loukinov,D., Tajmul,M., Kang,S., Kovalchuk,A.L., Strunnikov,A.V., Zentner,G.E., Ren,B. and Lobanenkov,V.V. (2020) CTCF mediates chromatin looping via N-terminal domain-dependent cohesin retention. Proc. Natl. Acad. Sci. U.S.A., 117, 2020–2031.

148. Vishwanathan,S.V.N., Borgwardt,K.M., Risi Kondor,I. and Schraudolph,N.N. (2008) In: Graph Kernels.

149. Pons,P. and Latapy,M. (2005) In: Computing Communities in Large Networks Using Random Walks (long version).

150. Newman,M.E. (2006) Modularity and community structure in networks. Proc. Natl. Acad. Sci. U.S.A., 103, 8577–8582.

151. Hauenstein,S., Dormann,C.F. and Wood,S.N. (2016) Computing AIC for black-box models using Generalised Degrees of Freedom: a comparison with cross-validation. arXiv doi: https://arxiv.org/abs/1603.02743, 09 March 2016, preprint: not peer reviewed.

152. Storey,J.D. (2002) A direct approach to false discovery rates. J. R. Stat. Soc. B (Stat. Methodol.), 64, 479–498.

153. Huang da,W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc., 4, 44–57.

154. Consortium,G.T., Laboratory,D.A., Coordinating Center -Analysis Working, G., Statistical

Methods groups-Analysis Working, G. and Fund, N.I.H.C., Nih/Nci, Nih/Nhgri, Nih/Nimh, Nih/NidaCoordinating Center -Analysis Working, G., Statistical Methods groups-Analysis Working, G. and Enhancing,G.G.2017) Genetic effects on gene expression across human tissues. Nature, 550, 204–213.

155. Lappalainen,T., Sammeth,M., Friedlander,M.R., Hoen,P.A., Monlong,J., Rivas,M.A., Gonzalez-Porta,M., Kurbatova,N., Griebel,T., Ferreira,P.G. et al. (2013) Transcriptome and genome sequencing uncovers functional variation in humans. Nature, 501, 506–511.

156. Grundberg,E., Small,K.S., Hedman,A.K., Nica,A.C., Buil,A., Keildson,S., Bell,J.T., Yang,T.P., Meduri,E., Barrett,A. et al. (2012) Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat. Genet., 44, 1084–1089.

157. Battle,A., Mostafavi,S., Zhu,X., Potash,J.B., Weissman,M.M., McCormick,C., Haudenschild,C.D., Beckman,K.B., Shi,J., Mei,R. et al. (2014) Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res., 24, 14–24.

158. Grubert,F., Zaugg,J.B., Kasowski,M., Ursu,O., Spacek,D.V., Martin,A.R., Greenside,P., Srivas,R., Phanstiel,D.H., Pekowska,A. et al. (2015) Genetic control of chromatin states in humans involves local and distal chromosomal interactions. Cell, 162, 1051–1065.

159. Gong,J., Mei,S., Liu,C., Xiang,Y., Ye,Y., Zhang,Z., Feng,J., Liu,R., Diao,L., Guo,A.Y. et al. (2018) PancanQTL: systematic identification of cis-eQTLs and trans-eQTLs in 33 cancer types. Nucleic Acids Res., 46, D971–D976.

160. Mumbach,M.R., Granja,J.M., Flynn,R.A., Roake,C.M., Satpathy,A.T., Rubin,A.J., Qi,Y., Jiang,Z., Shams,S., Louie,B.H. et al. (2019) HiChIRP reveals RNA-associated chromosome conformation. Nat. Methods, 16, 489–492.

161. Mumbach,M.R., Rubin,A.J., Flynn,R.A., Dai,C., Khavari,P.A., Greenleaf,W.J. and Chang,H.Y. (2016) HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nat. Methods, 13, 919–922.

162. Fulco,C.P., Nasser,J., Jones,T.R., Munson,G., Bergman,D.T., Subramanian,V., Grossman,S.R., Anyoha,R., Doughty,B.R., Patwardhan,T.A. et al. (2019) Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. Nat. Genet., 51, 1664–1669.

163. Jiang,Y., Huang,J., Lun,K., Li,B., Zheng,H., Li,Y., Zhou,R., Duan,W., Wang,C., Feng,Y. et al. (2020) Genome-wide analyses of chromatin interactions after the loss of Pol I, Pol II, and Pol III. Genome Biol., 21, 158.

164. Dyson,N.J. (2016) RB1: a prototype tumor suppressor and an enigma. Genes Dev., 30, 1492–1502.

165. Marke,R., van Leeuwen,F.N. and Scheijen,B. (2018) The many faces of IKZF1 in B-cell precursor acute lymphoblastic leukemia. Haematologica, 103, 565–574.

166. Sarvagalla,S., Kolapalli,S.P. and Vallabhapurapu,S. (2019) The two sides of YY1 in cancer:

a friend and a foe. Front. Oncol., 9, 1230.

167. Stengel,K.R. and Hiebert,S.W. (2015) Class I HDACs affect DNA replication, repair, and chromatin structure: implications for cancer therapy. Antioxid. Redox. Signal., 23, 51–65.

168. Losada,A., Hirano,M. and Hirano,T. (1998) Identification of Xenopus SMC protein complexes required for sister chromatid cohesion. Genes Dev., 12, 1986–1997.

169. Lee,T.C. and Ziff,E.B. (1999) Mxi1 is a repressor of the c-Myc promoter and reverses activation by USF. J. Biol. Chem., 274, 595–606.

170. Visel,A., Blow,M.J., Li,Z., Zhang,T., Akiyama,J.A., Holt,A., Plajzer-Frick,I., Shoukry,M., Wright,C., Chen,F. et al. (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature, 457, 854–858.

171. Lynch,C.J., Bernad,R., Calvo,I., Nobrega-Pereira,S., Ruiz,S., Ibarz,N., Martinez-Val,A., Grana-Castro,O., Gomez-Lopez,G., Andres-Leon,E. et al. (2018) The RNA polymerase II factor RPAP1 is critical for mediator-driven transcription and cell identity. Cell Rep., 22, 396–410.

172. Hu,Z., Killion,P.J. and Iyer,V.R. (2007) Genetic reconstruction of a functional transcriptional regulatory network. Nat. Genet., 39, 683–687.

173. Albert,F.W., Bloom,J.S., Siegel,J., Day,L. and Kruglyak,L. (2018) Genetics of trans-regulatory variation in gene expression. Elife, 7, e35471.

174. Brynedal,B., Choi,J., Raj,T., Bjornson,R., Stranger,B.E., Neale,B.M., Voight,B.F. and Cotsapas,C. (2017) Large-scale trans-eQTLs affect hundreds of transcripts and mediate patterns of transcriptional co-regulation. Am. J. Hum. Genet., 100, 581–591.

175. Johanson,T.M., Lun,A.T.L., Coughlan,H.D., Tan,T., Smyth,G.K., Nutt,S.L. and Allan,R.S. (2018) Transcription-factor-mediated supervision of global genome architecture maintains B cell identity. Nat. Immunol., 19, 1257–1264.

176. Ebert,A., McManus,S., Tagoh,H., Medvedovic,J., Salvagiotto,G., Novatchkova,M., Tamir,I., Sommer,A., Jaritz,M. and Busslinger,M. (2011) The distal V(H) gene cluster of the Igh locus contains distinct regulatory elements with Pax5 transcription factor-dependent activity in pro-B cells. Immunity, 34:175–187.

177. Arvey,A., Tempera,I., Tsai,K., Chen,H.S., Tikhmyanova,N., Klichinsky,M., Leslie,C. and Lieberman,P.M. (2012) An atlas of the Epstein-Barr virus transcriptome and epigenome reveals host-virus regulatory interactions. Cell Host Microbe, 12:233–245.

178. Bult,C.J., Blake,J.A., Smith,C.L., Kadin,J.A., Richardson,J.E. and Mouse Genome Database,G. (2019) Mouse Genome Database (MGD) 2019. Nucleic Acids Res., 47:D801–D806.

179. Li,H., Quang,D. and Guan,Y. (2019) Anchor: trans-cell type prediction of transcription factor binding sites. Genome Res., 29:281–292.

180. Keilwagen,J., Posch,S. and Grau,J. (2019) Accurate prediction of cell type-specific transcription factor binding. Genome Biol., 20:9.

181. Frankish A, et al. (2019) GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res 47(D1):D766-D773.

182. Visel A, et al. (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature 457(7231):854-858.

183. Park SJ, Kim JH, Yoon BH, & Kim SY (2017) A ChIP-Seq Data Analysis Pipeline Based on Bioconductor Packages. Genomics Inform 15(1):11-18.

184. Wang J, Huda A, Lunyak VV, & Jordan IK (2010) A Gibbs sampling strategy applied to the mapping of ambiguous short-sequence tags. Bioinformatics 26(20):2501-2508.

185. Chung D, et al. (2011) Discovering transcription factor binding sites in highly repetitive regions of genomes with multi-read analysis of ChIP-Seq data. PLoS Comput Biol 7(7):e1002111.

186. Zeng X, et al. (2015) Perm-seq: Mapping Protein-DNA Interactions in Segmental Duplication and Highly Repetitive Regions of Genomes with Prior-Enhanced Read Mapping. PLoS Comput Biol 11(10):e1004491.

187. Zhang Y, et al. (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol 9(9):R137.

188. Consortium EP (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414):57-74.

189. Langmead B & Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9(4):357-359.

190. Landt SG, et al. (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res 22(9):1813-1831.

191. Chung D, et al. (2011) Discovering transcription factor binding sites in highly repetitive regions of genomes with multi-read analysis of ChIP-Seq data. PLoS Comput Biol 7(7):e1002111.

192. Kharchenko PV, Tolstorukov MY, & Park PJ (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol 26(12):1351-1359.

193. Li Q, Brown JB, Huang H, & Bickel PJ (2011) Measuring reproducibility of high-throughput experiments. The Annals of Applied Statistics 5(3).

194. O'Leary NA, et al. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional

195. Jung I, et al. (2019) A compendium of promoter-centered long-range chromatin interactions in the human genome. Nat Genet 51(10):1442-1449.

196. Li G, et al. (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell 148(1-2):84-98.

197. Rao SS, et al. (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159(7):1665-1680.

198. Klopfenstein DV, et al. (2018) GOATOOLS: A Python library for Gene Ontology analyses. Sci Rep 8(1):10872.

199. Grant CE, Bailey TL, & Noble WS (2011) FIMO: scanning for occurrences of a given motif. Bioinformatics 27(7):1017-1018.

200. Kulakovskiy IV, et al. (2018) HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res 46(D1):D252-D259.

201. Szklarczyk D, et al. (2019) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 47(D1):D607-D613.

202. Karolchik D, et al. (2004) The UCSC Table Browser data retrieval tool. Nucleic Acids Res 32(Database issue):D493-496.

203. Storer J, Hubley R, Rosen J, Wheeler TJ, & Smit AF (2021) The Dfam community resource of transposable element families, sequence models, and genome annotations. Mob DNA 12(1):2.

204. Madeira F, et al. (2019) The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res 47(W1):W636-W641.

205. Waterhouse AM, Procter JB, Martin DM, Clamp M, & Barton GJ (2009) Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics 25(9):1189-1191.

206. Buniello A, et al. (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res 47(D1):D1005-D1012.

207. Battle A, et al. (2014) Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res 24(1):14-24.

208. Lappalainen T, et al. (2013) Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501(7468):506-511.

209. Grundberg E, et al. (2012) Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat Genet 44(10):1084-1089.

210. Quinlan AR & Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26(6):841-842.

211. Navarro Gonzalez J, et al. (2021) The UCSC Genome Browser database: 2021 update. Nucleic Acids Res 49(D1):D1046-D1057.

212. Lin Yang TZ, Iris Dror (2014) TFBSshape: a motif database for DNA shape features of transcription factor binding sites. Nucleic Acids Research 42(1):148-155.

213. Pouya Kheradpour MK (2013) Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments Nucleic Acids Research 42(5):2976–2987.

214. Arisa Sakamoto RY, Reona Yamaguchi, Shinji Narahara, Hiroyuki Sugiuchi, Yasuo Yamaguchi (2018) Cross-talk between the transcription factor Sp1 and C/EBPb modulates TGFb1 production to negatively regulate the expression of chemokine RANTES. Heliyon 4:e00679.

215. Sophie Lanciano, Gael Cristofari (2020) Measuring and interpreting transposable element expression. Nature Reviews Genetics 21:721-736.

216. James F Collins ZH (2007) Promoter analysis of intestinal genes induced during iron-deprivation reveals enrichment of conserved SP1-like binding sites. BMC Genomics 420.

217. Suresh Cuddapah RJ, Dustin E. Schones, Tae-Young Roh, Kairong Cui and Keji Zhao (2009) Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. Genome Research 19:24-32.

218. Yong Peng DX, Lun Zhao, Weizhi Ouyang, et al (2019) Chromatin interaction maps reveal genetic regulation for quantitative traits in maize. Nature Communications 10(2632).

219. Qi Zhang XZ, Sam Younkin, Trupti Kawli, Michael P. Snyder & Sündüz Keleş (2009) Systematic evaluation of the impact of ChIP-seq read designs on genome coverage, peak identification, and allele-specific binding detection. BMC Bioinformatics 17(96).

220. Guillaume Bourque BL, Vinsensius B. Vega, Xi Chen, Yen Ling Lee, et al (2008) Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Res 18(11):1752–1762.

221. Feschotte C (2008) Transposable elements and the evolution of regulatory networks. Nature Reviews Genetics 9:397–405.

222. Jieqiong Qu GYHZ (2019) p63 cooperates with CTCF to modulate chromatin architecture in skin keratinocytes. Epigenetics & Chromatin 12(31).

223. Eleonora Candi et al(2019) Metabolic pathways regulated by p63. Biochemical and Biophysical Research Communications 482(3):440-444.

224. Babu D, Fullwood MJ (2015) 3D genome organization in health and disease: emerging opportunities in cancer translational medicine. Nucleus. 6:382–93.

225. Melissa B. McPeak DY, Danielle A. Williams,Christopher L. Pritchett,Zhi Q. Yao,Charles E. McCall,Mohamed El Gazzar (2017) Frontline Science: Myeloid cell-specific deletion of Cebpb decreases sepsis-induced immunosuppression in mice. Journal of Leukocyte Biology 102:191-200.