COMPUTATIONAL ANNOTATIONS OF CELL TYPE SPECIFIC TRANSCRIPTION FACTORS BINDING AND LONG-RANGE ENHANCER-GENE INTERACTIONS

By

Wenjie Qi

A DISSERTATION

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

Biomedical Engineering – Doctor of Philosophy
Computational Mathematics, Science, and Engineering – Dual Major

2022

ABSTRACT

Precise execution of cell-type-specific gene transcription is critical for cell differentiation and development. The accurate lineage-specific gene regulation lies in the proper combinatorial binding of transcription factors (TFs) to the cis-regulatory elements. TFs bind to the proximal DNA sequences around the genes to exert control over gene transcription. Recently, experimental studies revealed that enhancers also recruit TFs to stimulate gene expression by forming long-range chromatin interactions, suggesting the interplay between gene, enhancer, and TFs in the 3D space in specifying cell fates. Identification of transcription factor binding sites (TFBSs) as well as pinpointing the long-range chromatin interactions is pivotal for understanding the transcriptional regulatory circuits. Experimental approaches have been developed to profile protein binding as well as 3D genome but have their limitations. Therefore, accurate and highly scalable computation methods are needed to comprehensively delineate the gene regulatory landscape. Accordingly, I have developed a supervised machine learning model, TF-wave, to predict TFBSs based on DNase-Seq data. By incorporating multi-resolutions features generated by applying Wavelet Transform to DNase-Seq data, TF-wave can accurately predict TFBSs at the genome-wide level in a tissue-specific way. I further designed a matrix factorization model, EP^{3}ICO, to jointly infer enhancer-promoter interactions based on protein-protein interactions (PPIs) between TFs with combined orders. Compared with existing algorithms, EP^{3}ICO not only identifies underlying mechanistic regulators that mediate the 3D chromatin interactions but also achieves superior performance in predicting long-range enhancer-promoter links. In conclusion, our models provide new computational approaches for profiling the cell-type specific TF bindings and high-resolution chromatin interactions.

ACKNOWLEDGEMENTS

I would like to acknowledge and give the warmest thanks to my advisor and committee chair, Professor Jianrong Wang. I have learned a lot from Dr. Wang about how to conduct research with specific goals and directions. His guidance carried me through all the stage of completing my projects as well as helped me to establish computational algorithm development.

I would also like to express the appreciation to my committee members, Professor Adam Alessio, Professor Yuehua Cui and Professor Jens Schmidt as well as Professor Sudin Bhattacharya, for their insightful feedback in my research and professional guidance in my career skill development.

Special thanks to Professor Adam Alessio. Dr. Alessio helped me go through the tough stages during my graduate school. Working with him, I have not only developed my technical skills, but also learned how to become a caring person who would always be there and help others. I feel blessed that I have Dr. Alessio on my committee.

I would like to thank my lab members, Jiaxin Yang, Dr. Binbin Huang and Dr. Hao Wang for their help in my research. I will remember the days we worked together.

Thanks to my friends, David Filipovic, Muneeza Amat for their listening, understanding and mental support. Graduate school could be hard, I am more than grateful that I have met them and have their accompany through this journey.

I want to express my deepest thanks to my parents, my brother and his family. They are always behind me and encourage me to go further. Without their love and support, I would never have the achievements today.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Understanding the gene regulatory network is critical to gain clear insights of cellular process as well as improve the understanding of human health. The formation of lineage-specific gene regulatory network achieved by cell-type specific regulatory relationship between transcription factors (TFs) and their target genes. Through the binding to cis-regulatory elements including promoters and enhancers in the non-coding genomic region, TFs exert control over target genes transcription activity such as enhance or inhibit target genes expression level. On the one hand, TFs bind on promoters near their target gene transcription start sites and directly recruit RNA polymerase or other accessary factors to regulate gene transcription activity. On the other hand, TFs also bind to distal enhancers that could be 1Mb away from target genes' core promoters to regulate their transcription activity. The regulatory input from enhancers is achieved by the proper folding of chromatins that brings regulatory elements close to promoters in a three-dimensional space. Therefore, identification of TFs binding sites (TFBSs) is the first step for constructing the cell-type specific gene regulatory networks. Moreover, pinpointing the long-range enhancer-promoter interactions is also critical for delineating the gene regulatory network in the 3D space.

Experimental technologies including ChIP-Seq, ChIP-exo, and ChIP-nexus have been developed to detect TFs bindings at genome-wide level in different cell types. However, all of the ChIP-based methods are limited by the quality of antibodies. In addition, given that there are approximate 1600 TFs in the human genome, profiling every TFs in every cell line at every stage is currently not experimentally feasible. Experimental approaches

have also been developed to profile the chromatin conformation across cell types. For example, Hi-C and Capture-C can profile long-range chromatin contacts but it has high false positive rates and lacks tissue variability. ChIA-PET can detect chromatin interactions at high resolution. However, ChIA-PET can only detect chromatin interactions facilitated by a specific protein and therefore has high false negative rates. Overall, experimental methods used for detecting TFBSs as well as chromatin structures are cost-expensive and not able to identify underlying mechanisms that determine transcription factor binding or facilitate long-range chromatin interactions. Therefore, accurate and highly scalable computational methods are needed to address the problems.

Diverse genomic datasets have been profiled by high throughput sequencing technologies and collected by large data consortia such as ENCODE, Roadmap. All of the previous efforts make it promising to develop computational methods that integrate multi-omics data to predict TFBSs and infer long-range enhancer-promoter interactions. Here, we developed two main computational models, TF-wave and EP^{3}ICO to predict cell-type specific TFBSs and long-range enhancer-promoter interactions respectively. TF-wave uses a Gradient Boosting Tree model to predict based on DNase-Seq data. By applying Wavelet Transform to DNase-Seq data to extract multi-resolution features as input, TF-wave predicts cell-type TFBSs at genome-wide level accurately. TF-wave can also be applied to distinguish different TFs binding accurately. Moreover, by using local chromatin accessibility information, TF-wave can predict TF bind probability at single-nucleotide level, which opens a new avenue in predicting TF footprinting at high-resolution. As a case study, we have developed an XGBoost model that takes multi-omics data to predict Aryl Hydrocarbon Receptor's (AHR) binding sites in multiple tissues.

Determinants of AHR binding are identified by extracting the important features from the XGBoost model. To detect long-range enhancer-promoter interactions, we have developed EP^{3}ICO, which applies matrix factorization models to predict long-range enhancer-promoter interactions based on protein-protein interactions (PPIs) between TFs. By considering second-order PPIs between TFs, EP^{3}ICO has boosted accuracy for enhancer-promoter interactions that cannot be reconstructed by first-order PPIs. EP^{3}ICO can be further applied to infer higher-order PPIs between TFs that regulate chromatin interactions. Dispute that predicting long-range enhancer-promoter interactions is challenging, EP^{3}ICO has achieved superior performance when compared with existing methods. EP^{3}ICO can identify PPIs between TFs as mechanistic regulators that mediate 3D chromatin interactions. The predicted long-range enhancer-promoter links are also enriched with cis-eQTLs, providing insights of unrevealing how genetic variants affect gene regulation, ultimately leading to different phenotypes.

CHAPTER 2

PREDICTING TRANSCRIPTION FACTOR FOOTPRINT USING WAVELET DECOMPOSED DNASE-SEQ DATA

This chapter is adapted from our in-preparation work: Qi W., Yang J., Wang J.. Predicting transcription factor footprint using Wavelet decomposed DNase-seq data.

2.1 INTRODUCTION

Identification of site-specific transcription factors (TF) binding at cis-regulatory elements is the key for elucidating regulatory mechanisms underlying transcriptional process and disease progress[1–4]. Characterizing TF binding sites (TFBS) across the entire genome is a monumental task[2]. Recently, high-throughput sequencing-based methods such as chromatin immunoprecipitation followed by DNA-sequencing (ChIP-Seq) are widely used to profile and detect TF binding sites at the genome-wide level[5]. However, ChIP-Seq has low resolution and is cost-expensive[6]. Newer experimental methods such as ChIP-exo and ChIP-nexus have been developed to detect DNA-binding TFs with higher resolution and cost efficiency[7,8]. However, all of the ChIP-based techniques have several shortcomings. Firstly, ChIP-based methods require high-quality antibodies to pull down TFs and therefore are limited to TFs that have high-quality antibodies. Secondly, ChIP-based methods cannot be applied to distinguish TF bindings between primary binding and secondary binding[2,6,9]. Lastly, ChIP-based experiments can only characterize one TF per experiment. Given that there are approximate 1600 TFs in the human genome, it is currently not experimentally feasible to profile every TF in every tissue/cell line at different stages.

With the advent of high-throughput sequencing technologies, another experimental assay that identifies DNase I Hypersensitive Sites (DHS) has been developed to detect the open regions of chromatin[10–12]. In vivo binding of TFs shields bound DNA elements from the DNase I's attack. After deep-sequencing, the chromatin accessibility data as well as the TF-specific DNase-I protection profiles at single-nucleotide resolution are collected in the open chromatin region. A majority of TFBSs can be identified at single-nucleotide resolution by detecting a footprint-like region with low DNase I cutting frequency[13].

By taking advantage of the booming DNase-Seq data, computational methods have been developed to predict TFBS by investigating the TF footprint patterns in open chromatin regions. The computational footprinting methods can be grouped into two main categories, including motif-centric models (BinDNase, CENTIPEDE, FLR, DeFCoM, PIQ) and de-novo models (Wellington, TRACE, DNase2TF)[2,13–19].

On the one hand, motif-centric methods usually scan the genome to generate candidate binding sites for TF of interest and predict TF-specific binding activity at the candidate sites. BinDNase takes up- and down-stream base pair resolution DNase-Seq data (100bp) around the motifs as features and select the discriminatory features by a backward greedy method. BinDNase then deploys supervised logistic regression model to predict TFBSs based on the selected features. CENTIPEDE used position weight matrices (PWM) to scan the genome and collect all positions with substantial sequence similarity with the TF motif as the candidate sites. Then CENTIPEDE applied an unsupervised Bayesian mixture model to integrate DNase-Seq data, DNA sequence data, and histone modification data to infer which candidate site is likely to be bound by the TF. FLR detects active TFBSs by a mixture of multinomial models. Firstly, FLR first learns the mixture of

two multinomial distributions, representing TF footprints and background by the expectation maximization algorithm. Secondly, footprints are scored by log-odds ration or the footprint-versus-background model. One main advantage of FLR is that it uses a small window size around the motif, i.e. ±25bp, which makes it a more targeted approach for detecting footprints for TF of interest. DeFCoM extracts both local and global DNase-Seq features by using different size of segments around the motifs. The DeFCoM applies kernel SVM model with either a linear kernel or a radical kernel to integrate DNase-Seq features in a non-linear way to predict TFBSs at their motif sites. PIQ uses Gaussian process to model and smooth the footprint signals around candidate motif sites. Then PIQ estimates the footprints with an expectation propagation algorithm. Finally, PIQ selects the set of motifs whose footprints are distinguishable from the noise as the final predictions.

On the other hand, the de-novo methods do not require the pre-generated candidate binding sites. De-novo methods detect TFBSs in the open chromatin region based on the DNase I digestion pattern. Wellington[19] is a sliding window approach that detects TF footprints based on the binomial test. For a given candidate footprint site, Wellington tests the hypothesis that there are less reads within the footprint than its flanking regions. The major novelty of Wellington is that it tests each strand independently as it considers that different strands have different effect in inhibiting DNase I activity. However, Wellington is unable to detect the binding sites for TFs of interest specifically. DNase2TF detects footprints by calculating the significance of DNase I cut depletion around the motif region based on a binomial Z score. Then DNase2TF interactively merge the candidate footprints sites to identify the regions that produces the most significant depletion regions.

TRACE uses a Hidden Markov Model (HMM)[20,21] to predict footprints and label binding sites for desired TFs by integrating PWMs and DNase-Seq Data in the open chromatin[2]. However, most de-novo methods were not designed to predict TF-specific footprinting and cannot label binding sites for TFs of interest.

Assigning TFs to their footprints on the basis of matching consensus sequences enables the analysis of TF-mediated gene regulatory networks[6,22]. Experimental studies have revealed that different TFs have different TF footprint shapes[13,17]. However, previous computational footprinting methods mainly focus on binary prediction, i.e. if TFs will bind on a given candidate motif or an accessible open region. Yet, existing methods haven't explored the different TFs footprint shape information and lack the ability to assign different TFs to their footprinted regions when there exist multiple candidate regions.

Feature extraction methods such as short-time Fourier transform (STFT) and Discrete Wavelet Transform have been widely applied in spectrum signal processing[23–27]. Wavelet Transform applies a low pass filter, i.e. scaling and a high pass filter to a signal, i.e. wavelet transforms and decomposes the original signals into signals with low resolution and high resolution respectively. The low-resolution signals, which are the approximate coefficients, represent a summary of the original signals. On the other hand, the high-resolution signals, which are the detailed coefficients represent the fine details in the original signals[23,28]. Both low-resolution coefficients and high-resolution coefficients are half sizes of the original signals. Multi-resolution signals are obtained by repeating the rounds of scaling and wavelet transform process to the low-resolution signals.

Inspired by the different TF footprint patterns and the successful feature extraction application of Wavelet Transform, we develop a supervised machine-learning model, TF-

wave, which applies Gradient Boosted trees to utilizes different frequency features decomposed by Discrete Wavelet Transform (DWT) from DNase-Seq data to predict specific TF bindings. By applying Wavelet decomposition to the DNase-Seq signal, TF-wave considers the low and high-frequency signals contained in the spectrum in different TF's footprints and can accurately infer specific TF binding sites as well as distinguish different TFs' binding sites. Furthermore, we introduce a convolutional neuron network (CNN) model that can predict TF footprint probability at single-nucleotide level based on the local chromatin accessibility. Our results provide a framework for both binary and nucleotide-level footprinting model of TFs could be applied as the first step in analyzing TF-focused gene regulatory networks.

2.2 MATERIALS AND METHODS

2.2.1 Datasets and data generation

TF ChIP-Seq data in BED format and DNase-Seq data in BAM and BED format in K562 and GM12878 were collected from ENCODE. FRiP score for the TF ChIP-Seq files was calculated to quantify the quality of ChIP-Seq experiments by R Libray ChIPQCsample[29]. The ChIP-Seq Bed file with the highest FRiP score is kept for TFs with multiple datasets. To filter out TFs ChIP-Seq data with low quality, TFs with ChIP-Seq FRiP score < 5 were removed. Overall, 46 TFs in K562 and 19 TFs in GM12878 with ChIP-Seq FRiP score => 5 were kept for the subsequent analysis in each cell line. TFs' motifs are downloaded from JASPAR[30]. For each TF, candidate motif sites were identified by MOODS[31] with threshold $p < 0.0001$ in open chromatin regions obtained from DNase-Seq BED files. Motif sites under TF ChIP-Seq peaks are labeled as positive samples, while motif sites under the DNase-Seq peak only are labeled as negative samples. Motifs located on both

forward and reverse strands are collected. Using the motif midpoint as the center, we extended each motif site to upstream 100bp and downstream 100bp to obtain the local chromatin accessibility from corresponding DNase-Seq BAM files around the TF motifs for the following analysis.

2.2.2 Data processing

The DNase-Seq read depth at each nucleotide for each motif site was obtained by SAMtools[32] based on the DNase-Seq BAM file. DNase-Seq read-depth at each base pair for 200 bp was extracted by 5'-3' orientation corresponding to the orientation of the TF motif on the forward strand naturally. However, for the motifs located in the complementary strand, we extracted the DNase-Seq read-depth at each nucleotide from 3'-5' with respect to the orientation of the motifs. Then we inverted the DNase-Seq read-depth vector for motifs on the complimentary strand to have the 5'-3' direction. DNase-Seq read-depth vectors on leading strands and inverted DNase-Seq read-depth vectors on complementary strands were used to construct the feature matrix at 5'-3' orientation.

2.2.3 Discrete Wavelet Transform and Gradient Boosted Trees model

To extract features with different frequencies contained in the TF footprints as well as motifs sites in open chromatin regions without being footprinted by TFs, the DNase read-depth data is first normalized by Z-Score normalization. Then we applied the Biorthogonal wavelet to decompose the 200bp normalized DNase-Seq vector into approximate coefficients and detailed coefficients with multiresolution. The approximate coefficients and detailed coefficients are then concatenated together as features to train the Gradient Boosting Trees (GBT). GBT is a fast and effective tree-boosting algorithm for

classification and regression[33]. GBT model learns the data by adding additional trees into the model to minimize the loss function of previous trees. Similar to the tree-based method, GBT can learn the non-linear interactions between different features and is proved to have high classification performance.

2.2.4 Model training and evaluation

In order to train the GBT model in TF-wave to obtain high accuracy, we tuned two kinds of parameters during the training process. Firstly, previous studies have observed that different TFs have different footprint shapes[13,22], a single Wavelet Transform model with one decomposition level is unable to capture the most informative features for all TFs. Therefore, all of the 14 Biorthogonal wavelets model in pywt packages[34] were used to decompose the DNase-Seq. The parameter of decomposition level is set up to six. The decomposed signals of a particular wavelet model with a specific decomposition level were used to construct a feature matrix and train the model. Secondly, the hyperparameters of GBT including the number of trees including [100, 500, 1000, 2000] and learning rate including [0.05, 0.1, 0.5] were tuned during the training process. 5-fold cross-validation with AUPR metric was used to evaluate the model performance. The best-performing model with the highest AUPR is selected and it consists of features constructed by a specific Biorthogonal wavelet model and the corresponding decomposition level. Then the best-performing model is applied to predict TFBSs at genome-wide level and in cell-specific way.

2.2.5 Models performance comparison

Due to the limited access to other model, we can only compare the performance of our model with Wellington (http://jpiper.github.io/pyDNase). We used the open chromatin regions containing TFs motifs as evaluation data. Specifically, candidate binding sites (motif sites) that overlap with DNase-seq peaks and ChIP-seq were used as the positive set, and candidate sites that are under DNase-seq peaks but not under ChIP-seq peaks are used as the negative set. We applied Wellington to DNase-seq peaks (with 100-bp flanking regions to each side) containing the same sets of motif sites that were included in our model and We compared our model with Wellington[19]. The fdrlimit in the output of Wellington was set to 0 to have predictions for all input DNase-seq peaks. The absolute values were used as scores in the evaluation. Only the predictions overlapping with motif sites of tested TFs were included in the evaluation on the same dataset.

2.2.6 Distinguish different TFs footprints

TFs have unique footprint shapes as it is shown in Figure 2.1. To investigate if different TFBSs is distinguishable by their DNase I footprinting shapes, we trained GBT models to predict different TFs' binding sites based on the Wavelet decomposed DNase-Seq signals. Specifically, given two TFs, TF A and TF B, the DNase-Seq signal of these two TFs' binding sites was collected and decomposed by Wavelet as input features for the GBT model using the methods described above. Then the binding sites of TF A are labeled to be 1 as the positive data and the binding sites of TF B are labeled to be 0 as the negative data for training the GBT model. GBT models were then trained and evaluated by 5-fold cross-validation. We also used original DNase-Seq signals to train the GBT model as

baseline models and compared it with the model trained by Wavelet decomposed DNase-Seq signals using AUCs on the same datasets.

2.2.7 Consensus footprint generation

For each TF, the trained GBT was used to predict TFBS for each candidate motif site. To obtain the consensus footprints for each TF, the 200bp DNse-Seq read-depth of the top 5% predicted binding sites were used to generate the TF's consensus footprints. The consensus footprints were presented by aggregating DNase-Seq read-depth at each nucleotide. The examples were obtained by the UCSC genome browser[35].

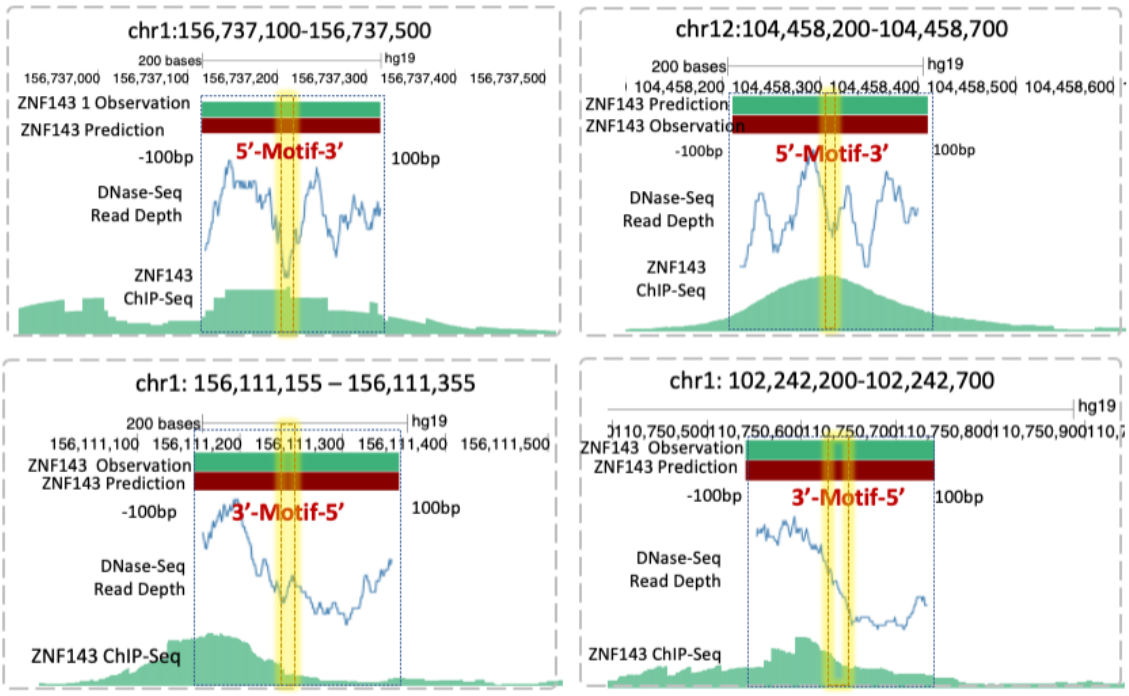

Figure 2.1. TFs leave footprints to DNase-Seq signal. Four examples of DNase-Seq footprints left by TF ZNF143. From the top to bottom for each panel: observed ZNF143 binding site (green) flanking the motif (yellow), predicted ZNF143 binding site by TF-wave (red), DNase-Seq read depth of 200 base pairs flanking the ZNF143 motif (blue). ZNF143 ChIP-Seq signal from bigwig files (green). The orientation of the DNase-Seq data is indicated by 5'-3' or 3'-5'.

2.2.8 Single-nucleotide level footprint probability prediction

Mutations in TFs motif results in the gain of function or loss of function binding, which will lead to the cis-regulatory evolution[22]. Therefore, it is critical to determine the binding probability at each nucleotide for a TF to understand how the variants influence the gene regulatory networks mediated by the TF binding. In order to investigate the TF binding probability, we trained a CNN model to predict the TF binding at each nucleotide centering the TF motifs based on the 2kb flanking regions DNase-Seq signals. CNN has proven highly effective in a number of diverse tasks including biological sequence analysis[36–38]. The input layer of CNN has concatenated Wavelet decomposed DNase-Seq signals. Following the input layer, there are three convolutional layers. In each convolutional layer, batch normalization was applied to center the data. The kernel width is 3, and stride size is 1. Then Relu activation function was used to the filter output. Finally, Maxpool was applied to pool adjacent values by taking the maximum in a small window to reduce the dimension of input for the next layer (Figure 2.7a).

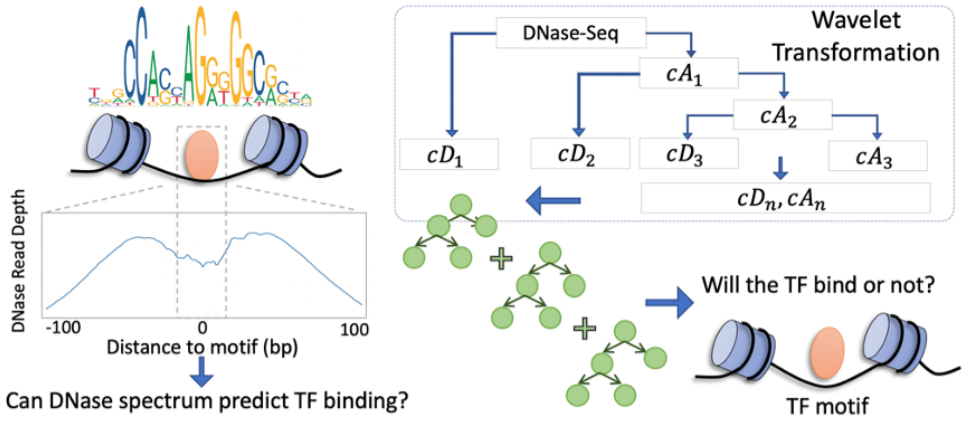

Figure 2.2. TF-wave predicts TF binding sites based on wavelet-transformed DNase-Seq signal. Left: schematic figure of binding of CTCF in open chromatin leaves

Figure 2.2 (cont'd)

footprint on DNase-Seq read-depth, centering around the CTCF motif, shown on the top. Right: schematic figure of applying Wavelet Transform to DNase-Seq to predict TF binding sites using gradient boosting tree model. Wavelet Transform is applied to decompose DNase-Seq signal. The extracted approximate coefficients ($\mathrm{cA_n}$) and detailed coefficients ($\mathrm{dA_n}$) are concatenated as input features to train the Gradient Boosting Tree models (GBT), the trained GBT is applied to predict TF binding status for a candidate motif in the open chromatin.

2.3 RESULTS

2.3.1 TF-wave predicts TFBS based on Wavelet decomposed DNase-Seq signals

The binding of TFs protects DNA from DNase I digestion and therefore leaves the TFs footprints on the DNase-Seq signal (Figure 2.1). The footprints of different TFs are distinguishable and contain unique features for specific TFs. We applied a Discrete Wavelet Transform to extract the multi-resolution features underlying the DNase-Seq signals centering the TF motifs. Gradient Boosting Tree model were applied to predict TF binding probability at candidate motif sites by using the multi-resolution features (Figure. 2.2).

By using 5-fold cross-validation with auROC and auPR to evaluate the model's performance, we demonstrated that Wavelet decomposed DNase-Seq signal can predict with-in TF binding sites at the genome-wide level accurately.

2.3.2 TF-wave accurately predicts TFBS at the genome-wide level

To assess the performance of TF-wave and Wellington which also uses DNase-Seq data to predict TFBS, we evaluated the performance of TF-wave and Wellington on the same datasets. To ensure a fair comparison, for a TF of interest, we used the same DNase-

Seq datasets containing this TF's motif as input data to predict its binding sites. The area under the receiver operating characteristic curve and the area under the precision-recall

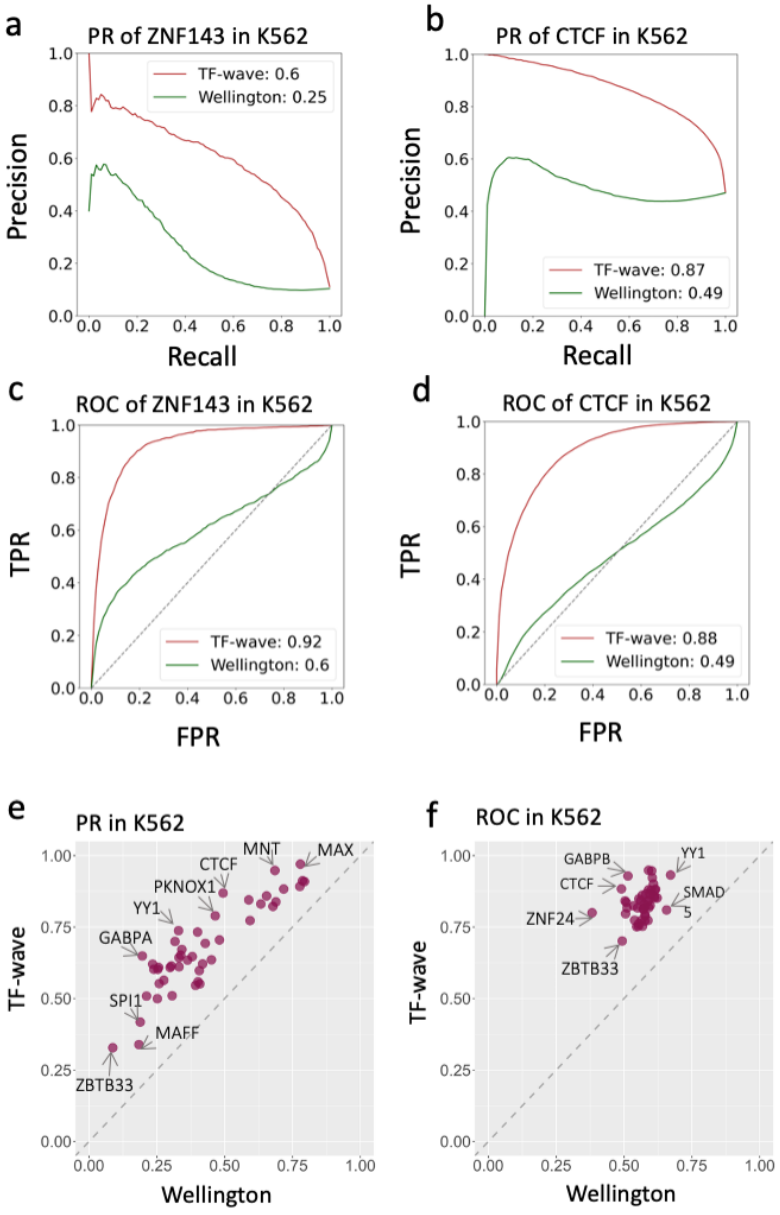

Figure 2.3. TF-wave predicts TF binding sites accurately and outperforms Wellington in K562. TF-wave and Wellington were evaluated on the same datasets using the averaged performance of 5-fold cross-validation. (a,b) auPRC of ZNF143 and CTCF in K562 for TF-Wave and Wellington. (c,d) auROC of ZNF143 and CTCF in K562 for TF-Wave and Wellington. (e) Performance comparison in K562 for TF-Wave and Wellington using auPR. The x-axis and y-axis are auPR of applying Wellington and TF-

Figure 2.3 (cont'd)

wave to predict TFBS on the same data. Each point represents a TF, points above the diagonal line indicate TF-wave performs better than Wellington. (f) Performance comparison in K562 for TF-Wave and Wellington using auROC. The x-axis and y-axis are auROC of applying Wellington and TF-wave to predict TFBS on the same data. Each point represents a TF, points above the diagonal line indicate TF-wave performs better than Wellington.

curve was calculated for TF-wave and Wellington based on the predicted binding probability and scores respectively. 5-fold cross-validation was used and the averaged auROC and auPRC were reported to evaluate the final prediction performance of TF-wave and Wellington. For all the TFs evaluated in the analysis, TF-wave achieves higher performances in identifying the TF binding sites from the candidate motif sites than Wellington in both K562 and GM12878 cell lines (Figure 2.3, Figure A.1). Examples of observed CTCF binding sites and TF-wave accurately predicted CTCF binding sites are shown in Figure 2.4.

2.3.3 Applying Wavelet Transform to DNase-Seq signal provides boosted prediction performance

To justify that multi-resolution features extracted by Discrete Wavelet Transform indeed contribute to the TFBSs prediction accuracy of TF-wave, we also trained the Gradient Booting Trees model on the Z-score normalized original DNase-Seq data as the baseline model and compared TF-wave with the baseline model on the same datasets. 5-fold cross-validation was used to evaluate TF-wave and the baseline model's performances. Averaged auROC and auPR were reported as the final results. We observed that TF-wave has higher auROC and auPR scores than the baseline model in both K562 and GM12878 cell lines (Figure A.2). The results demonstrated that TF-wave achieves

boosted performances for all TF binding sites predictions by using the multi-resolution features extracted from Wavelet decomposed DNase-Seq signal, suggesting the performance improvement of TF-wave is induced by applying Wavelet transformation DNase-Seq data.

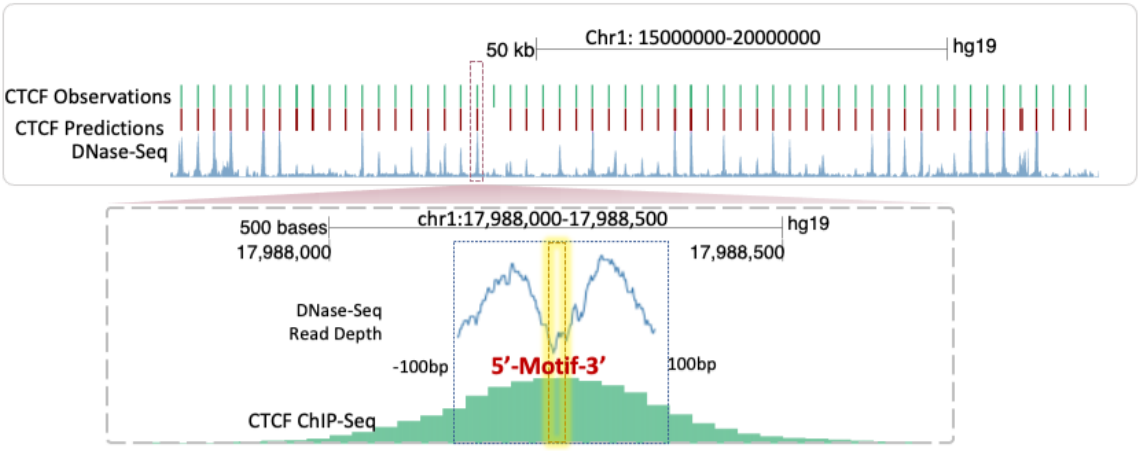

Figure 2.4. TF-wave identify TFBS accurately. Examples of observed CTCF binding sites at K562 chr1: 15000000-2000000. From top to bottom: observed CTCF binding sites (green), predicted CTCF binding sites by TF-wave (red), observed DNase-Seq indicating the open chromatin (blue). One correctly predicted CTCF binding site is highlighted by red rectangle and the CTCF footprinted DNase-Seq signal is shown below. Blue: CTCF footprinted DNase-Seq signal centering the CTCF motif(yellow). The orientation of the DNase-Seq read-depth data is 5'-3'. The CTCF ChIP-Seq track is shown at the bottom (green).

2.3.4 Predicted TF binding sites reveal the TF footprint shapes

In order to investigate if the predicted TF binding can reveal the different footprint shapes based on the DNase-Seq data, we ranked the predicted TF binding sites based on the GBT provided binding probability and selected the top 5% predicted binding sites to reconstruct the TF consensus footprints. Our results indicated that different TFs have different footprint shapes as shown in Figure 2.5 and Figure A.3. Yet, the TFs footprint

shapes are consistent with the pattern that there is one small DNase-Seq peak at each shoulder of the motif, indicating the high cleavage of DNase I at the accessible chromatin regions, and low cleavage of DNase I at the motif sites that are protected by TFs bindings. On the other hand, DNase-Seq read counts are relatively low at the centering motif region, suggesting the decreased DNase I cleavage rate at the motif sequences due to the protection of TF binding.

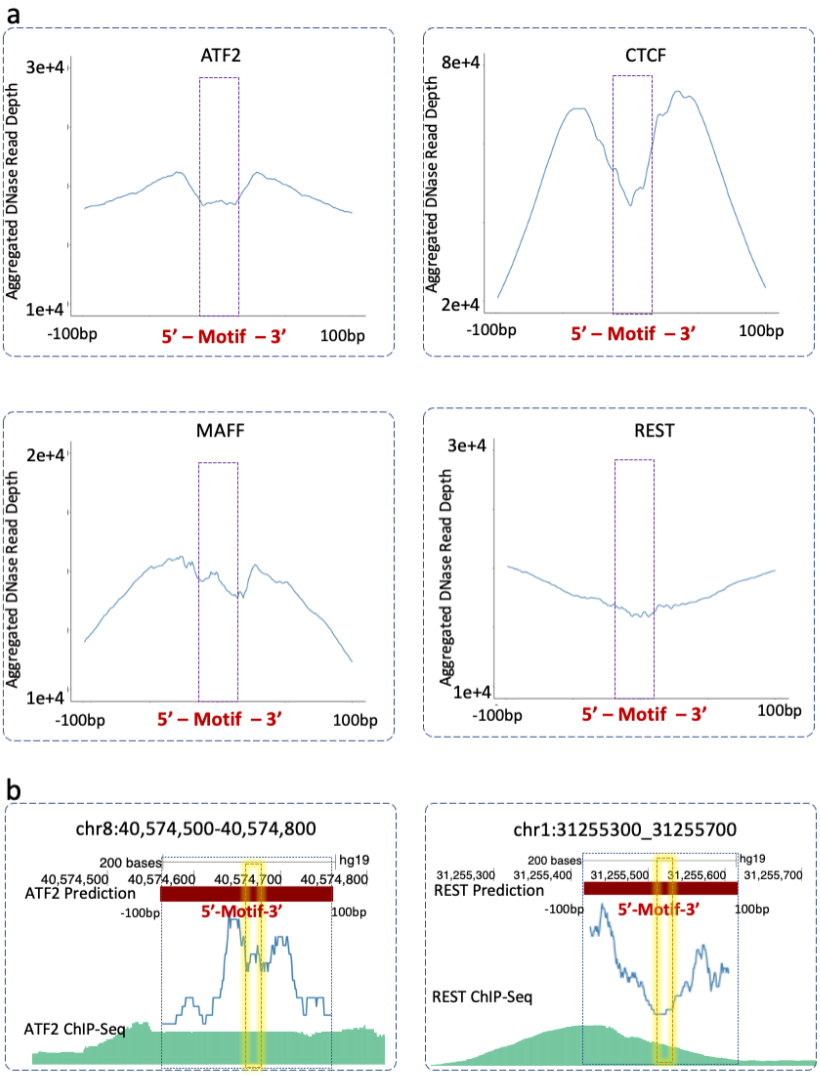

Figure 2.5. Different TFs have different consensus footprint shapes. (a) Aggregated

Figure 2.5 (cont'd)

DNase-Seq read-depth for top 5% predictions for ATF2, CTCF, MAFF, REST. Each TF has a unique footprint shape. The motif of ATF2 is palindromic, and the footprint of ATF2 is symmetric. (b) Examples of DNase-Seq read-depth for ATF2 and REST. The footprint shape of the specific ATF2 binding site and REST binding site is similar to the aggregated ATF2 DNase-Seq and REST DNase-Seq respectively. While the footprint shapes of ATF2 and REST are different.

2.3.5 Using Wavelet decomposed DNase-Seq data can distinguish different TFs binding

TFs bind at the accessible chromatin and different TFs leave different footprint shapes. Examples are shown in Figure 2.5. To test if different DNase-Seq signals centering the TFs footprinted region could distinguish different TFs binding, we applied Wavelet Transform to REST and ATF2 footprinted DNase-Seq data to extract features at different frequencies. The features at multi-resolution were used to train a Gradient Boosting Trees model to distinguish REST and ATF2 binding in the open chromatin (see details in the method section). The resulting AUCs are then compared to a baseline model that was trained on the original REST and ATF2 footprinted DNase-Seq data. Figure 2.6 shows that using Wavelet decomposed DNase-Seq signals can successfully distinguish the binding of REST and ATF2 and has higher accuracy (auROC 0.75 and auPR 0.65) than using DNase-Seq data only (auROC 0.72 and auPR 0.61). As it is shown in Figure. 2.6b, Figure. 2.6c, even though both REST and ATF2 are in the open chromatin region, the different footprinted DNase-Seq data decomposed by Wavelet have the capacity to distinguish their binding sites precisely.

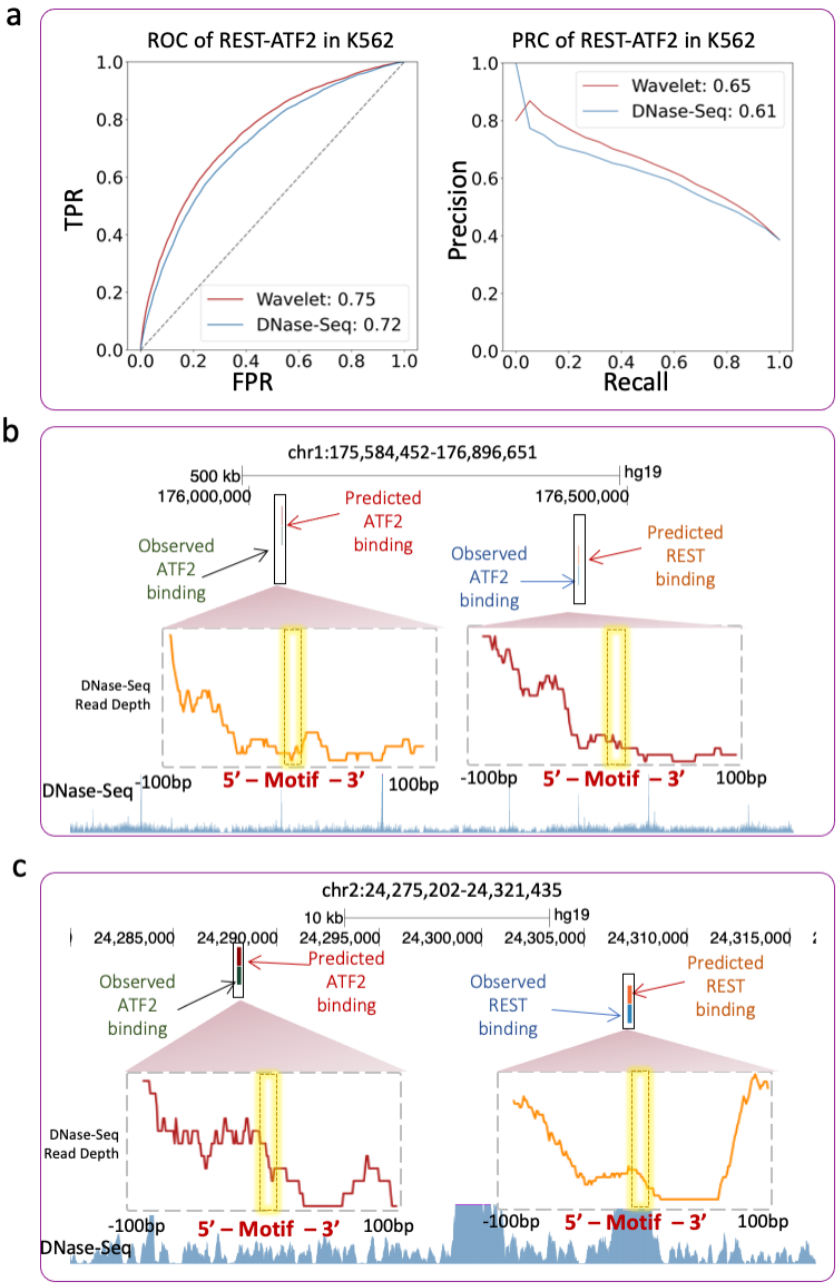

Figure 2.6. Applying wavelet transformation to DNase-Seq signal can distinguish TF binding sites accurately. (a) ROCs and PRC for predictions of REST-ATF2 binding sites. ROCs for predictions by applying Wavelet transform to DNase-Seq signals (Red). GBT is applied to DNase-Seq directly to predict TF binding sites as a baseline model (Green). (b, c) Examples of correctly predicted REST and ATF2 binding sites. From top to bottom: predicted ATF2 binding site, observed ATF2 binding site. Predicted REST binding sites, observed REST binding site. DNase-Seq read-depth at each motif site, DNase-Seq track in K562.

2.3.6 Local chromatin accessibility can predict TF binding at single-nucleotide level accurately

To investigate if local chromatin accessibility provides information for predicting TF binding probability at the single-nucleotide level, we obtained more DNase-Seq data including ± 1000bp from the summit of the TF ChIP-Seq peaks as features, then we trained a CNN model to learn the Wavelet decomposed DNase-Seq data and predict TF binding probability at each nucleotide for $\pm$25bp sequences flanking the summit of TF ChIP-Seq peaks containing the TF motif. Using CTCF ChIP-Seq data in K562 as an example data, we obtained multi-resolution features extracted by applying the Wavelet Transform to the K562 DNase-Seq data. The input layer for the CNN is the concatenated Wavelet decomposed DNase-Seq data. Subsequent convolutional layers use Maxpool and Relu activation functions (see methods). The final layer outputs 51 predictions for the probability of each nucleotide in the DNA sequence is being footprinted by the CTCF in K562.

To synthesize sensitivity and specificity, we assessed the CNN model using auROC. As the data is imbalanced, auPR is also used to measure the CNN model's performance. Using K562 CTCF ChIP-Seq as gold-standard, the CNN model achieves high accuracy with auROC 0.74 and auPR 0.51 (Figure. 2.7a). For the CTCF motifs within $\pm$25bp of the ChIP-Seq summit, the CNN model predicts high TF footprinted probability (> 0.5) for each nucleotide in the motifs, while for the flanking sequences around the motif, the predicted footprinted probability (< 0.5) is low (Figure 2.7b). These results suggested that local chromatin accessibilities can be applied to predict TF footprint probability at single-nucleotide level.

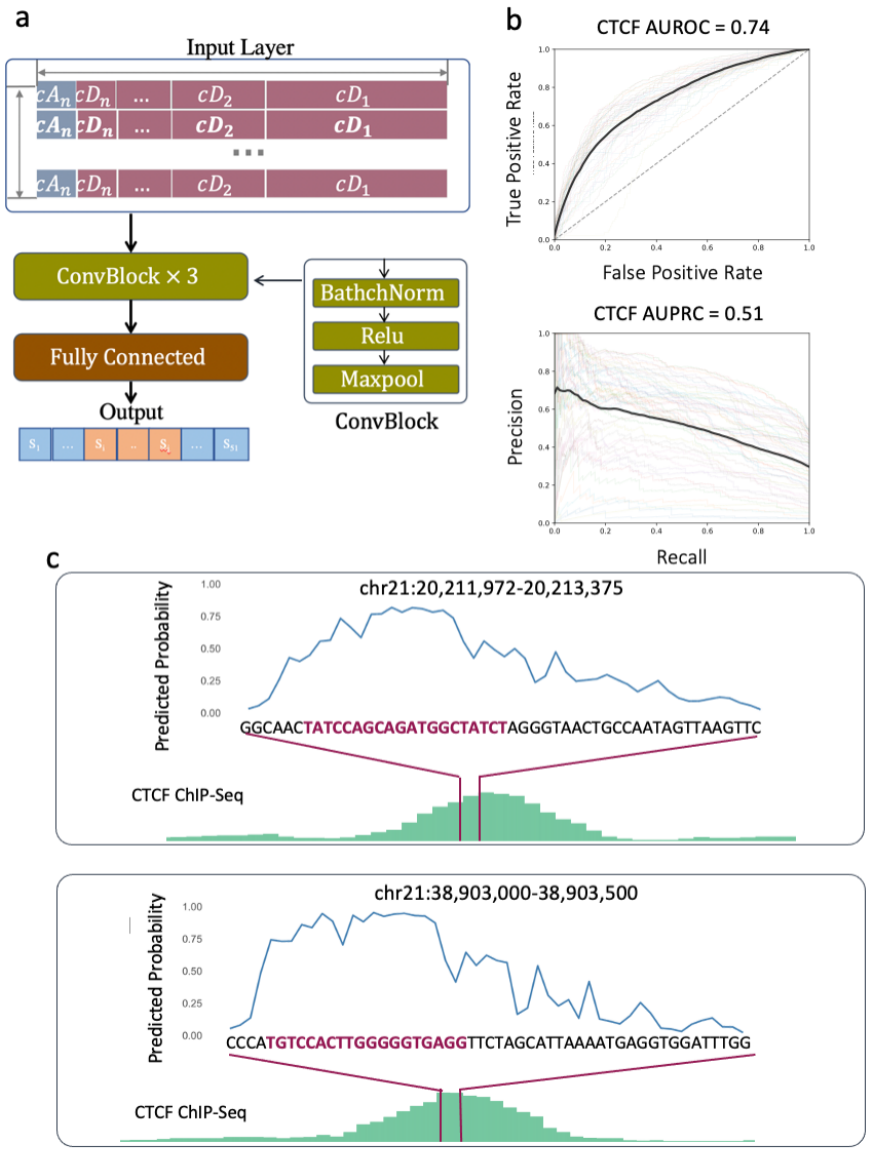

Figure 2.7. Schematic figure of applying wavelet transformation to DNase-Seq signals and using CNN model to predict TF binding at single-nucleotide levels. (a) The DNase-Seq signals are extracted and decomposed by wavelet transformation. The approximate coefficients and detailed coefficients are concatenated as the input layer for training the CNN model. The CNN model has three convolutional layers. In each convolutional layer, batch normalization was applied, followed by Relu activation, and Maxpool function. Finally, the output layer was the predicted footprinted probability for each nucleotide around the CTCF motif. (b) AUCs for CTCF prediction at the single nucleotide level. (c) Examples of predicted CTCF binding at every single nucleotide. From bottom to top: the observed CTCF track, the DNA sequences containing the CTCF motif (violet-red), and the predicted TF footprint probability for each nucleotide in the DNA sequence. The predicted footprint probability is high for the motif sequences.

2.4 DISCUSSION

Identifying TFBSs is the first-step for understanding the cell-specific gene regulatory network. In this study, we introduced TF-wave, a supervised Gradient Boosting Trees model that uses Wavelet decomposed DNase-Seq data to predict TF binding at candidate motif sites in a cell-specific way. We implement TF-wave in K562 and GM12878 cell lines. TF-wave demonstrates higher accuracy in TFBS prediction than other available computational footprinting method, Wellington, which also uses DNase-Seq data as the main feature. Moreover, we demonstrate that by applying Wavelet Transform to DNase-Seq data to extract multi-resolution features underlying the TF footprints as the input features, TF-wave achieved higher prediction performance than models trained on original DNase-Seq data.

Although the binding of TF in open chromatin protects DNA sequences from the cleavage of DNase I cleavage and therefore leaves footprints, different TFs yet have different footprint shapes. Existing computational footprinting methods focus on binary prediction of TFBS, i.e. if TFs bind or not for a given candidate motif or accessible chromatin region. They lack the capacity to distinguish different TFs binding in the open chromatin region. In this work, by using the Wavelet decomposed DNase-Seq signals, we demonstrate that TF-wave could accurately distinguish different TFs bindings when their motifs are in the open chromatin regions. This workflow provides the first step in understanding how TFs compete in the binding at the same accessible region as well as how TFs cooperate the binding.

DNase-Seq can profile TF footprint at high resolution by counting the read at 5’ end. However, previous TFBSs footprinting methods mainly focus predict binary binding

activity for a given DNA sequence. By taking advantage of the high-resolution DNase-Seq data, we developed the CNN model that can predict TF footprint probability at the single-nucleotide level. CNN model has been widely applied to biomedical filed and proved to be successful in multi-task prediction. The results show that trained CNN model can accurately predict TF binding at the motif site at single-nucleotide level. This single-nucleotide footprinting strategy enables the understanding of how genetic variants located in the TF footprints regions will result in the gain of function or loss of function TF binding as well as elucidate how variants influence the gene regulatory network. Finally, this workflow can open the revenue not only for identifying functional mutations but also for understanding the underlying mechanism for phenotypic variations.

CHAPTER 3

PREDICTIVE MODELS OF GENOME-WIDE ARYL HYDROCARBON RECEPTOR DNA BINDING REVEAL CELL SPECIFIC BINDING DETERMINANTS

This chapter is adapted from our in-preparation work: Filipovic D., Qi W., Kana O., Marri D., LeCluyse E., Andersen M., Cuddapah S., Bhattacharya S.. Predictive Models of Genome-Wide Aryl Hydrocarbon Receptor DNA Binding Reveal Tissue Specific Binding Determinants.

Section 3.3.2 in this chapter is adapted from previously published work (Desmet N. et al, 2021): Desmet N., Dhusia K., Qi W., Doseff A., Bhattacharya S., Gilad A. (2021) Bioengineering of Genetically Encoded Gene Promoter Repressed by the Flavonoid Apigenin for Constructing Intracellular Sensor for Molecular Events. Biosensors.

3.1 INTRODUCTION

The Aryl Hydrocarbon Receptor (AHR) is a ligand-activated transcription factor that belongs to the basic-helix-loop-helix (bHLH) PER-ARNT-SIM (PAS) family[39–42]. The AHR mediates toxic actions of environmental contaminants, such as 2,3,7,8-tetrachlorodibenzo-ρ-dioxin (TCDD)[43,44]. Before binding to ligands and being activated, the AHR is sequestered in the cytoplasm by its chaperone proteins including a dimer of the 90-kDa heat shock protein (HSP90), the AHR-interacting protein (AIP) and the cochaperone p23[45–48]. When activated by its ligands, the AHR translocates to the nucleus and forms a heterodimer with Aryl Hydrocarbon Nuclear Translocator (ARNT)[49–51]. The AHR-ARNT heterodimer binds to specific DNA sequences termed Aryl Hydrocarbon Response Elements (AHRE), Dioxin Response Elements (DREs), or xenobiotic response

elements (XREs) containing a consensus motif, 5'-GCGTG-3' to mediate genes transcription activity.

Through the binding to cis-regulatory elements including gene promoters and distal enhancers, the AHR regulates a variety of target genes including cytochrome p450 1A1 (CYP1A1), CYP1B1, and AHR Repressor. Identifying AHR binding sites is the first step for constructing its gene regulatory network and understanding its gene regulatory circuits, which is crucial for understanding the role of AHR in toxicity and disease, as well as its role in physiological functions such as immune response[52,53], circadian rhythm[54,55] cell cycle progression[54,56], and embryonic development[57,58]. High throughput sequencing technologies such as Chromatin Immunoprecipitation and Sequencing (ChIP-Seq), ChIP-exo, and ChIP-nexus have enabled the profiling of transcription factors binding sites at a genome-wide level[7,8,59]. However, these experimental approaches are limited by several constraints including cost, time, or biological materials such as high-quality antibodies[60–62]. AHR binding has been profiled in MCF-7 by ChIP-Seq experiment[53]. However, the determinants for the tissue specificity of AHR binding remain poorly understood.

In recent years, a lot of pioneer efforts have been made to develop computational approaches to predict the TF binding sites. One of the most commonly used models to infer TF binding is the position weight matrix (PWM). PWM is derived from experimentally validated DNA sequences bound by a particular TF and quantitively describes the binding sites of the TF. The representation of a particular TF PWM is its motif, which can be readily obtained from databases such as JASPAR, HOCOMOCO, TRABSFAC, or estimated de novo[30,63–66]. For a potential binding site, the PWM generates a quantitative score by adding up individual scores of each nucleotide making up the PWM and

overlapping the potential binding site. PWM is commonly used to scan the genome for candidate TF binding sites by a predefined threshold score[67–70]. The PWMs are often derived from in vitro experiments, including high throughput systematic evolution of ligands by exponential enrichment (HT-SELEX)[69] , and in vivo experiments such as ChIP-seq. However, many TFs exhibit high levels of in vivo binding to DNA sequences that do not possess the in-vitro or even the in-vivo derived binding motif.

On the other hand, TFs in eukaryotes generally do not bind DNA in isolation but rather in dense with other co-binding factors, which contributes to the tissue-specific binding activity. TF clusters with binding sites of multiple TFs co-occurring in close proximity[71,72]. Consequently, PWMs of co-bound TFs could potentially be used to predict the binding of a TF of interest. However, models incorporating PWMs of co-binding TFs have shown limited utility in improving model performance[70]. Nevertheless, given that TFs bind in clusters and that PWMs are not necessarily representative of actual TF binding, ChIP-seq signals of co-bound TFs, as a measure of their actual binding, are likely to provide information that cannot be obtained from PWM. In addition, interpretable machine learning models incorporating measures of co-bound TFs could provide mechanistic insights into the determinants of tissue-specific binding for a TF of interest, such as AHR.

Computational models have been developed to predict TF binding sites by integrating multi-omics data. Pique-Regi R et al applied Bayesian mixture models that integrate PWM, chromatin accessibility, and histone modifications to predict TF binding sites in open chromatin region[73]. TFBSImpute imputes missing TF binding directly from ChIP-Seq data by using a three-mode tensor[74], where the three dimensions are the TFs, cell lines and genome locations respectively. Virtual ChIP-Seq makes the TF binding prediction by

leveraging the associations between gene expression and TF bindings[70]. Deep learning models are also applied to predict TF binding in specific tissue based on DNA sequence and chromatin accessibility[60,74,75]. However, even though some of these models achieve high prediction performance for some TFs in both intra- and inter-cell lines, they generally lack interpretability and provide little mechanistic insight into what drives tissue-specific or cell-type specific TF binding. This limitation is especially acute for TFs exhibiting highly variable binding across tissues or cell lines. In addition, most computational models of TF binding to date have been developed for and tested on constitutively active TFs. The binding of inducible TFs, such as nuclear receptors or other ligand-activated TFs like the AHR, remains largely unexplored.

In this study, we applied a Gradient Boosting Tree model, XGBoost[76], to develop supervised machine learning models predicting the AHR binding status of DREs in open chromatin, i.e., DRE is bound or unbound, MCF-7 cell. Using AHR ChIP-Seq data derived from TCDD-treated MCF-7 cell line and corresponding chromatin accessibility experiments (DNase-Seq) in MCF-7 downloaded from ENCODE, we first detected tissue-specific AHR-bound and AHR-unbound DREs in open chromatin of each tissue. Then, we applied XGBoost to integrate multi-omics data to predict the binding status of DREs in open chromatin in MCF-7. Our results demonstrate highly accurate and robust models of within-tissue binding prediction. We identify several TFs as predictive features of AHR binding in individual tissues, such as GATA3 as well as histone modifications (HMs) – H3K4me1 and H3K4me3 in MCF-7 cells. Our tissue-specific models generalize well to the prediction of AHR binding sites without DREs, demonstrating the robustness of the models. In conclusion, we demonstrate that the patterns of TFs and HMs most predictive

of AHR binding. Moreover, we show that AHR binding is driven by a complex interplay of tissue-agnostic DNA sequences flanking the DRE and tissue-specific local chromatin context. The approach used here can be further adapted to other inducible TFs, such as steroid hormones and nuclear receptors.

3.2 MATERIALS AND METHODS

3.2.1 Identification of AHR bound and unbound DREs

To obtain the AHR bound or unbound DREs, all DREs in the human genome according to the reference sequence of the hg19 human assembly were collected based on the AHR motif firstly. Secondly, DNase-Seq experiment data for MCF-7 cell line was downloaded from ENCODE https://encodeproject.org/. The BroadPeak DNase-seq files for the hg19 genome assembly were used to mark the open chromatin. If there were multiple replicates, the intersection of all replicates was used for downstream analysis. Any DRE found under the peaks of DNase-seq intersection was considered to be in the open chromatin of the corresponding cell line and was used in the determination of bound and unbound DREs for the purposes of model training. DREs occurring in blacklisted regions were ignored in subsequent analyses. Thirdly, the AHR ChIP-seq file was downloaded from GEO (GEO: GSE90550)[77], where the original sequencing files have been processed uniformly following a standard processing pipeline. DREs in open chromatin as well as under the AHR ChIP-Seq peaks are assigned to be the AHR bound DREs and used as the positive data in the machine learning model. DREs in open chromatin but outside of the AHR ChIP-Seq peaks are assigned to be the AHR unbound DREs.

3.2.2 Genomic and epigenetic features generation

For each DRE in the human genome, the genomic sequence of seven nucleotides 5' and 3' flanking the DRE (5'-GCGTG-3') were obtained. These nucleotides were then one-hot encoded and used as genomic features for the machine learning models. There were around 1.6 million DREs located in the entire the human genome. While only a small fraction of DREs have fulfilled the criteria for bound and unbound DREs used in the model training and testing.

The epigenetic signals profiled by DNase-Seq, ChIP-Seq bigwig files were downloaded from ENCODE. The epigenetic signals extracted from bigwig files were then used as features in the models training and testing. Specifically, for each bound and unbound DRE, the epigenetic signal of 740 base pairs up and downstream from the DRE were extracted from the bigwig files, for a total of 1495 base pairs of signals (including the 5-base pairs DREs). The extracted signals were split into 15 genomic bins and each bin contains 99 base pairs. The epigenetic signals within each bin were averaged to generate 15 features corresponding to the particular DRE-genomic signal combination. During the averaging, any areas of missing signal are replaced with zeros.

3.2.3 XGBoost model training

For all the bound and unbound DREs appearing in open chromatin of that particular cell line, we have constructed the features matrix including genomic sequence, epigenetic features for all available DNase-Seq, histone mark as well as the co-binding activity of transcription factors. In order to tune the parameters of the XGBoost model, grid search was performed with the following of the hyperparameter space: max_depth = {3, 4, 5, 6, 7}, min_child_weight = {3, 4, 5, 6, 7}, subsample = {1.0, 0.9, 0.8, 0.7}, colsample_by_tree

= {1.0, 0.9, 0.8, 0.7} and eta = {0.05, 0.075, 0.1, 0.125, 0.15, 0.2, 0.3}. The average performances over all five folds were used to select best performing models in terms of hyperparameter selection.

3.2.4 XGBoost model evaluation

Area under Receiver Operating Characteristic (auROC) as well as area under Precision Recall curve (auPRC) was used to evaluate the performance of the model. The scores are the output of the XGBoost algorithm in the form of probabilities of each particular observation (DRE) belonging to a particular output class (bound or unbound). By using different thresholds for these probabilities above which the model predicts a DRE as bound, we obtain the number of true and false positives for each threshold, as well as true and false negatives relative to the ground truth of DRE binding obtained from the corresponding AHR ChIP-Seq experiment. Each threshold produces a point on the ROC and PRC curves; the area under the curve was calculated using a line interpolated through all the points.

3.3 RESULTS

3.3.1 Machine learning models accurately predict AHR binding

In order to detect AHR binding sites, we applied XGBoost to integrate multi-omics data to predict the binding status of DREs in open chromatin. We trained the models on DREs occurring under singleton, i.e., 1-DRE, peaks only. These DREs represent about one third of all AHR peaks, across all binding experiments. To avoid noise in the training data, multi-DRE peaks were not used for training since it is impossible to determine which specific DRE among the cluster of DREs under the ChIP-Seq peak were indeed bound

by AHR and can be used as positive labels for training the model. In the validation, multi-DRE peaks were used for validating the performance of the model. We developed all our machine learning models using the gradient boosted tree algorithm of the XGBoost family of algorithms, which has been shown to handle non-linear data well[78,79]. These algorithms also supply metrics of feature importance, i.e., the contribution of individual input features to improving the model performance.

Local chromatin features were used as input features for the models and are trained and validated to predict the binding status of singleton bound and isolated unbound DREs (see Methods), limited to DREs found in open chromatin. Model evaluation is performed using a 5-fold cross validation procedure (see Methods for details) (Figure 3.1). The local

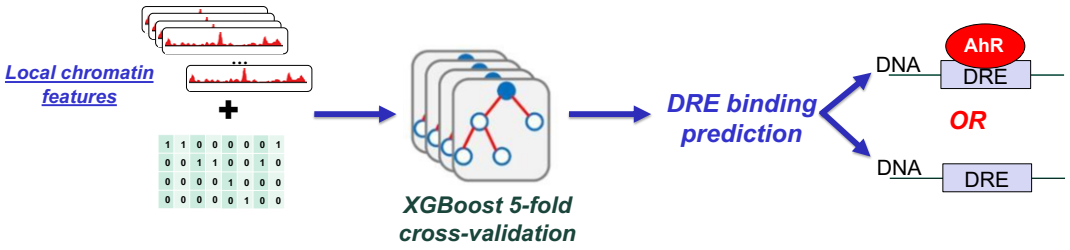

Figure 3.1. Machine learning models predicting AhR binding learn tissue specific and agnostic rules. Schematic figure of XGBoost architecture. The input features include epigenetic features and DNA sequence (method). XGBoost uses the local chromatin features to predict the binding status of DRE in open chromatin.

chromatin input features that were used include 1) The DNA sequence immediately flanking the DRE. We included the flanking sequence of up to 7 nucleotides directly up- and down- stream from the DRE–previously proposed to be involved in AHR binding. These sequences were one-hot encoded prior to being used as model inputs; 2) Binned mean values of bigWig signals of the MCF-7 cell line. We used the bigWig files of i)

DNase-seq (as representative of chromatin accessibility), ii) histone modification, and iii) transcription factor ChIP-seq signals from ENCODE – see methods for details. For each bigWig signal and each DRE, we created 15 bins of width 99 base pairs. Each bin was assigned a value representing the average bigWig signal across the width of that bin. The mid-point of the central bin was positioned at the middle nucleotide of the 5-bp DRE; 3) Indicator variables of whether the DRE is found in a strict (+/- 200 bp away from a transcription start site - TSS) or loose (+/- 1500 bp away from the TSS) definition of a promoter.

In order to achieve accurate prediction performance, an extensive hyperparameter search were conducted for each binding experiment and input feature set (see Methods) and the model with the highest performance were selected. In all instances, unless otherwise stated, model performance was reported as the area under the Receiver Operating Characteristic (ROC) and Precision Recall (PR), averaged over five folds using the 5-fold cross validation procedure. For class imbalanced datasets such as the ones used here, where unbound DREs far outnumber the bound (Supplementary Table 1), the area under the PR curve (auPRC) is considered a more appropriate metric of model performance. Therefore, the model producing the highest auPRC was selected as the best performing model. However, the area under the ROC curve (auROC) was still a useful metric to distinguish between poorly and well performing models when comparing between binding experiments (see Methods).

To better understand the tissue-specific determinants of AHR binding beyond the core DRE motif and chromatin accessibility, we developed a series of interpretable classifiers that use different combination of features to determine the binary binding status of DREs

in open chromatin. The model performance was investigated as a function of the input feature set. Different feature sets consisting of the following features were - 1) DNase-seq only, 2) flanking sequence only, 3) flanking sequence and DNase-seq, 4) flanking sequence, DNase-seq and histone modifications, 5) flanking sequence, DNase-seq, histone modifications and transcription factor binding (referred to as the full model). (Figure 3.2)

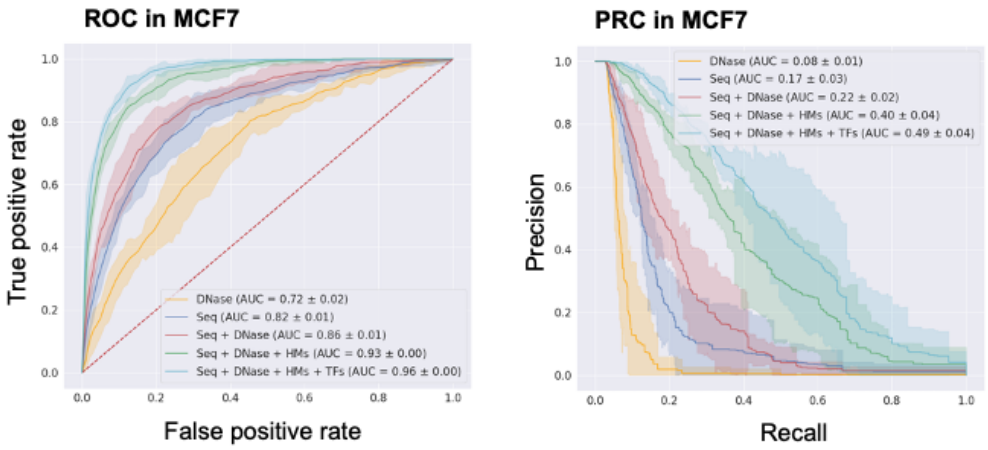

Figure 3.2. Performance of models predicting the binding status of DREs in open chromatin of the MCF-7 cells, with five different sets of input features. Performance of each set of features is represented as a mean line with a 95% confidence interval shaded around the line resulting from 5-fold cross-validation. The legend shows the features used, as well as area under the curve. Both receiver operating characteristic (ROC) - left panel, and precision-recall curves (PRC) – right panel, are shown.

XGBoost was extracted to determine, for each binding experiment, the total feature importance of non-sequence (i.e., TF binding and epigenomic) features (Figure 3.3A), relative feature importance of sequence features per flanking sequence nucleotide position (Figure 3.3B), and relative importance of individual bins of non-sequence features (Figure 3.3C). Further examination of non-sequence feature importance scores

revealed that the specific models are predominantly learning and making AHR-DRE binding predictions by relying on different features across different binding experiments. For example, within each binding experiment we observed three to six bigWig signals with feature importance 2-5 times higher than that of any other signal. These were in H3K4me1, H3K4me3, GATA3, CTCF, ZNF217 and FOXA1.

Figure 3.3. Feature importance of all local chromatin context features excluding DNA sequence flanking the DRE, measured as feature importance gain in XGBoost classifier model trained on a particular cell type. A Feature importance for each chromatin context feature is calculated as the average feature importance of all bins for that particular chromatin context feature. B Feature importance of DNA sequence flanking the DRE, measured as feature importance gain in XGBoost classifier model trained on a particular cell line or type. Feature importance of each nucleotide type at a particular position relative to the DRE is normalized to the nucleotide type with the highest feature importance at that nucleotide position. C Feature importance of all local chromatin context features excluding DNA sequence flanking the DRE in models predicting the DRE binding status of all DREs in open chromatin. Feature importance measured as feature importance gain in XGBoost classifier model trained on a particular cell line or type. Feature importance for each bin of a particular chromatin context feature is normalized to the bin with the highest feature importance for that chromatin context feature.

The importance of DNA sequence immediately flanking the DRE was evaluated via examining the importance scores of nucleotides flanking the DRE produced by two ways including 1) flanking sequence-only models, and 2) full models (inclusive of flanking

sequence, DNase-seq, histone mark and transcription factor features). And the importance of every nucleotide is shown in Figure 3.3B

3.3.2 Sequence-only AHR model can identify the Optimal Promoter Sequence for designing gene circuits

As a ligand-inducible transcription factor, the AHR could be activated by its ligand including flavonoids, and bind to DNA to turn on its target genes expression. Therefore, the AHR system could be used to design reporter genes or gene circuits controlled by AHR ligands.

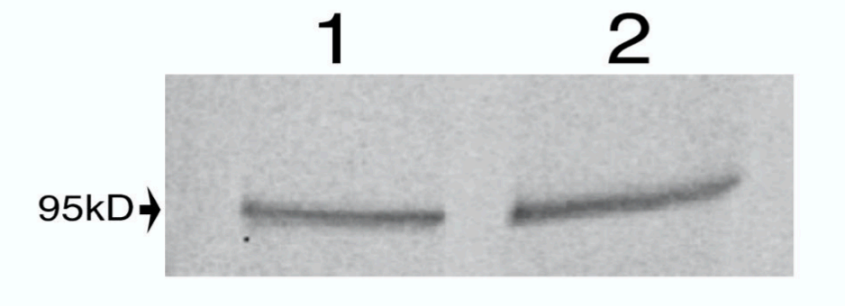

Figure 3.4. Western blot analysis of AHR expression in HEK293FT (1) and Hela (2).

In order to find the promoter with valid AHR binding sites, the XGBoost model trained based DNA sequence only was applied to predict the optimal promoter sequence for designing a reporter gene in response to flavonoids since the gene circuits don't contain any epigenetic features. Specifically, the XGBoost model was trained on the AHR binding sites found in human breast cancer MCF-7 cells. The trained model was then applied to predict the binding probabilities for the 10 base pair segments containing the AHR core motif (5'-GCGTG-3') are displayed in Table 1. Screening of a 2Kb promoter region upstream of the CYP1A1 transcription start site revealed 5 locations with putative binding probability of the AHR/ARNT complex greater than 0.5, suggesting that the complex

bound to these locations would initiate transcription of the reporter gene. Western blot confirmed that the reporter gene successfully expressed the AHR protein (Figure 3.4)

3.4 DISCUSSION

The binding of transcription factors (TFs) to DNA has been extensively studied by experimental studies, computational prediction models have been developed for TF binding imputation. However, the determinants and mechanisms underlying tissue-specific TF binding are still not well understood. In addition, unlike many constitutively active TFs, tissue-specific binding of ligand-inducible TFs such as AHR cannot be fully determined through chromatin accessibility, the extended binding motif, the motifs of other co-bound TFs, or any combination of these features.

PWM has been applied to scan the genome to identify candidate AHR binding sites, however, this method has high-false positive rate and cannot give us insights of determinants of AHR binding. In order to accurately identify AHR binding sites as well as as well as gain insights of mechanisms that mediate AHR binding, we developed, to the best of our knowledge, the first machine learning models that integrate multi-omics data to predict cell-specific and tissue-specific AHR binding sites. The machine learning model, XGBoost has demonstrated robust prediction performance across different tissue. Moreover, XGBoost can integrate predictive features in a non-linear way, and the important features can be extracted based on the Gradient Boosting tree structures, which provides us the probability of deciphering factors that determine AHR bindings. Furthermore, XGBoost trained on DNA sequence only has high scalability, and can be applied to predict AHR binding for building synthetic genes.

One of the limitations of predicting ligand-activated TFs is inconsistency in the data. In this study, the AHR ChIP-Seq is derived from TCDD treated MCF-7 cell line, while the epigenetics data used for training the machine-learning model collected from untreated cell lines. The treatment of TCDD might change the epigenetic environment in the genome, which results in the inconsistency of data used to generate features and labels for the model. Therefore, for future studies, comprehensive profiling of ligand-activated transcription factors is required not only to provide more biological information but also improve the predictive model's performance designed specifically for such transcription factors.

CHAPTER 4

JOINT INFERENCE OF PROTEIN-PROTEIN INTERACTIONS AND ENHANCER-GENE LINKS BY A MATRIX DECOMPOSITION MODEL

This chapter is adapted from our in-preparation work: Qi W., Yang J., Wang H., Wang J.. Joint Inference of Protein-Protein Interactions and Enhancer-Gene Links by A Matrix Decomposition Model.

4.1 INTRODUCTION

One of the fundamental questions in human biology is how one genome sequence can give rise to so many different cell fates. The answer to this question lies in the accurate execution of cell-type-specific gene transcription during cell differentiation and development[80–84]. Such lineage-specific regulation of gene transcription requires core promoters and proximal elements that locate around the genes' transcription start sites. Recent experimental studies have revealed that, in addition to the proximal regulation elements, distal cis-regulatory elements, e.g. enhancers, also appear to be as major contributors in regulation of gene transcription activity[85–87]. Far away from their target genes as the enhancers could be in the genome, they can stimulate the gene transcription through the formation of chromatin loops to that bring the enhancers close to their target gene promoters in the 3D space[88–91].

Proper folding of chromatin is critical for gene regulation. The alterations in chromatin structure could lead to developmental abnormalities or human disease[88–91]. For example, disruptions including deletions, inversions, and duplications cause the TAD boundaries spanning EPHA4 brachydactyly, F syndrome, and polysyndactyly in humans. Further 4C analyses in a CRISPR-Cas9 edited mouse model revealed that a cluster of limb-specific

enhancers associated with EPHA4 was misplaced and abnormally activated neighboring genes including Pax3, Wnt6, and Ihh[92]. In addition to developmental disorders, mutations in chromatin structures such as enhancer adoption or enhancer hijacking by oncogenes could also lead to tumorigenesis[91,93–96]. For instance, variations are frequently found at cohesin and CTCF binding sites and, perturbing the TAD boundaries in non-malignant cells, which upregulates the proto-oncogenes[97]. A study using 7,416 cancer genomes across 26 tumor types also revealed that even one single duplication of the genome could result in a formation of a new chromatin domain and lead to the overexpression of the IGF2 gene in colorectal cancer[98]. Therefore, delineating the genome in the 3D space is critical to expanding our understanding of cell development and disease progression.

Experimental methods have been developed to profile long-range chromatin interactions. For instance, chromatin conformation capture (3C) and 3C-derived techniques (4C and 5C) perform high-throughput sequencing to examine spatial topology at a regional scale in the genome[99–103]. Moreover, Hi-C was later developed to profiling chromatin interactions at a genome-wide level[103]. A similar method, Capture Hi-C[104] was developed to further improve the resolution of C based method. Chromatin interaction analysis with paired-end-tag sequencing (ChIA-PET) can detect cell-type specific long-range chromatin interactions at high resolution by targeting one protein out of interest[105]. Furthermore, CRISPR-dCase9-based techniques were developed to unbiasedly capture long-range DNA interactions as well as identify locus-specific chromatin-regulating protein complexes[106,107]. All of these technologies have provided large-scale chromatin contact maps in a diversity of cell types and tissues in the human genome as well as model species[108]. Imagining-based methods such as Fluorescence in situ hybridization

of DNA (DNA-FISH)[109] use fluorescently labeled probes to hybridize to their complementary target loci within the nucleus. The chromatin contacts are therefore inferred by setting an arbitrary threshold, which is usually 50 nm to 1 μm[110] , to the spatial distance based on the scale of genomic distances between the regions of interest and the resolution of the microscope.

However, even though the experimental methods have generated comprehensive references of the long-range chromatin interactions maps for a number of cell types and tissue in both humans and model species[108,111–114], there are limitations on in-depth profiling of enhancer-promoter interaction in a cell-type specific way. First, the resolution of existing C-methods including Hi-C and Capture Hi-C is usually 5kb-40kb, which is still relatively low and wherefore hard to precisely pinpoint specific enhancers that regulate the promoters[104,108]. Second, the chromatin interactions detected by Hi-C lack tissue variability and therefore have low specificity and too many false positive discoveries[115,116]. Third, mapping resolution is directly related to sequencing depth for Hi-C assays, hence, the cost of laboratories seriously hinders the popularity of Hi-C. Fourth, although ChIA-PET and CRISPR-dCas9 have higher resolutions and can detect cell-type specific interactions, ChIA-PET relies on the antibodies to specific proteins such as CTCF, RAD21[114], CRISPR-dCas9 can only generate locus-specific long-range interactions. Therefore, these two methods can only provide a subset of long-range chromatin interaction, which lead to high false negative discoveries. Imagining-based methods can detect chromatin contacts at all scales of chromosome folding, including contacts between chromosomes, while DNA-FISH remains to be limited to pre-selected genomic

regions and often used to validate findings, herein, not widely applied at the genome-wide level[110].

Due to the limitations of experimental technologies, computational methods have been developed to predict cell-type and tissue-type specific long-range enhancer-promoter interactions by integrating multi-omics signatures, including genomics, transcriptomics, and epigenomics[117–120]. There are in general two categories of computational methods of predicting chromatin interactions, e.g. unsupervised and supervised machine learning models. Unsupervised machine learning methods usually predict chromatin interactions by assigning each enhancer-promoter pair a score, ranking the pair based on the scores, and selecting the top-ranked pairs as the predicted interacting enhancer-promoter pairs[121,122]. Supervised models predict enhancer-promoter interactions by incorporating multi-omics signature features. The features include i) genomic sequences and genomic distance between enhancer-promoter pair, ii) gene expression activity profiled by RNA-Seq, iii) enhancer activity profiled by epigenetics signals such as H3K4me1, DNase-Seq. iv) epigenetic features at the genomic window region between enhancer and promoter. By integrating these or some of these features, supervised machine-learning models are trained on labels generated by experimental technologies and predict enhancer-promoter interactions on unknown data. Two top-performing methods IM-PET[123] and TargetFinder[124] were developed to infer long-range enhancer gene links. IM-PET predicts EP pairs by training a Random Forest model that takes four kinds of features 1) correlation between enhancer and promoter activity. 2) The correlation between the expression of transcription factor binding on enhancers and the promoters' activity. 3) the coevolution between the enhancer and its target promoter. 4) genomic distance between

enhancer and promoter. TargetFinder employs a gradient boosting tree model and integrates hundreds of genomic features including 1) chromatin accessibility, 2) methylation status of DNA, 3) gene expression, 4)ChIP-Seq signal of TFs, architectural proteins, modified histones, 5) quantified signal as well as the genomic distance between enhancer and promoter. 6) conserved synteny of the enhancer and promoter, 7) similarity of TF and target gene annotations. Despite the supervised machine learning models have better performance than unsupervised models [125], these models suffer from overfitting problems due to the large feature dimension [126]. In addition, these models didn't provide insights into the mechanisms that mediate long-range enhancer-promoter interactions[127].

In addition to the predictive features used in previous studies as mentioned above, recent experimental studies revealed that protein-protein interactions (PPIs) between TFs also have been identified to participate in the formation of long-range chromatin interactions and herein facilitate interactions between distal enhancers with their target genes[128–131]. The most well-known example is the DNA loop extrusion formed by the cohesin complex, where cohesin loads onto chromatin, leading to the formation and enlargement of DNA loops that are eventually arrested at boundary elements such as CTCF[132–134]. In addition, instead of being an insulator to maintain the TAD boundary, intra-TAD CTCF binding, together with the cohesin complex, is also reported to stabilize the enhancer-promoter interactions and maintain robust gene expression. Deleting the CTCF binding sites compromises the interactions between enhancers and promoters[135,136]. Moreover, The zinc-finger transcription factor Yin Yang 1 (YY1) has also been identified as a structural regulator of enhancer-promoter loops. YY1 could form dimers to promote DNA

interactions by binding to active enhancers and promoters identified by chromatin immunoprecipitation with mass spectrometry (ChIP-MS)[131,137].

Furthermore, experimental studies have reported that TFs bind together to secure long-range enhancer-promoter interactions. For example, the zinc-finger protein, ZNF143, is also demonstrated to be involved in CTCF-mediated chromatin interaction loops by overlapping with cohesin binding sites and CTCF binding sites. Deletion of the ZNF143 binding sites is reported to result in the loss of CTCF-mediated chromatin loops. Together, these observations implicate that ZNF143 functions as a partner of CTCF to establish and stabilize the chromatin structures by cooperating with the cohesin complex[138–142].

These observations from previous experimental work establish a mechanistic hypothesis that PPIs between specific TFs may mediate long-range enhancer regulation. Therefore, integrating PPIs between TFs as a new set of features into the computational models is expected to improve the prediction accuracy of long-range enhancer-promoter links. By interpreting the computational model to extract the most predicted PPIs, we can gain insights of cell/tissue-type specific mechanisms that regulate long-range enhancer-gene links. Furthermore, by considering TF PPIs with combined orders, the computational models are expected to achieve boosted accuracy in predicting long-range enhancer-gene links.

In this study, we developed a matrix factorization model to infer enhancer-promoter interactions based on TF PPIs with combined order (EP^{3}ICO). EP^{3}ICO jointly predicts long-range enhancer-promoter interactions and optimizes first-order TF-TF interactions that regulate enhancer-promoter links. Prioritized first-order TF-TF interactions can be applied to accurately to predict enhancer-promoter links. For the enhancer-promoter pairs

that cannot be reconstructed by the first-order TF-TF interactions, EP^{3}ICO further decomposes these enhancer-promoter pairs and identify second-order PPIs between TFs that facilitate the long-range interaction between enhancer-promoter pairs that cannot be explained by the first-order TF PPIs only. By using TAD-split cross-validation and controlling confounding factors including genomic distance between enhancer and promoter, we demonstrated the superior performance of EP^{3}ICO compared to existing models. EP^{3}ICO provides a workflow to prioritize novel multi-order PPIs between TFs that regulate long-range enhancer-promoter interactions, making it possible to understand the underlying mechanisms that organize the 3D chromosome.

4.2 MATERIALS AND METHODS

4.2.1 Datasets

EP^{3}ICO takes the continuous chromatin contact frequency matrices derived from Hi-C data as input. High-resolution Hi-C data from K562 was downloaded from GEO (GEO: GSE63525)[103,104,108]. Knight-Ruiz normalization (KR) was first applied to remove the biases such as library size, fragment length, GC content, sequence mappability, copy number variations, and other unknown factors[143,144]. ChIA-PET in K562 was used as ground truth to evaluate the model performance[111]. Enhancer-promoter pairs were labeled as positive samples if they overlap with ChIA-PET interactions. Otherwise, they will be labeled as negative samples.

Gene promoters are defined as up-stream and down-stream 1kb regions of the gene transcriptional start sites (TSS). The TSS data is obtained based on the annotation from GENCODE v17[145]. Gene expressions in K562 were measured by RPKM values of the RNA-seq dataset from the Roadmap Epigenomics project[70]. Enhancer coordinates were

obtained based on ENCODE and Roadmap enhancer annotations. Enhancer activities in K562 were quantified by the DNase-Seq signals[146,147]. In order to have high-activity enhancers and promoters, enhancers or promoters that have activity < 0.05 in K562 are removed. Enhancers that locate on the same Hi-C anchors with promoters are also filtered out. Correlation coefficients between enhancer-gene pairs were calculated based on the enhancer activity and gene expression across 111 cell types.

The ChIP-Seq IDR narrow peaks datasets for TF binding in K562 were downloaded from ENCODE[148]. The ChIP-Seq experiments with treatments were removed to maintain the consistency with the cell-specific Hi-C and ChIA-PET data used in this work. If multiple datasets exit for one TF, ChIP-Seq file with the highest FRiP score[149] is selected first, then ChIP-Seq datasets with FRiP score < 5 are further filtered out to maintain the high quality of ChIP-Seq data. The significant peaks identified by MACS2 were used to label the TF binding sites in the genome. Overall, 114 TFs in K562 were used for constructing the TF-TF interaction matrix for the following analysis.

Protein-protein interactions are collected from STRING dataset v11[150]. To obtain the high-quality PPIs, only PPIs with the confidence score greater than 100 in the 'Experiments' were used to for the following analysis.

4.2.2 Generation of data matrices

EP^{3}ICO applies matrix factorization models to optimize first-order PPIs and second-order PPIs between TFs that regulate long-range enhancer-promoter interactions. Five kinds of matrices containing genomic information are required for the matrix factorization model: 1) intra-TAD enhancer-promoter interaction frequency matrix, 2) TF binding on enhancer matrix and TF binding on promoter matrix, 3) first-order TF-TF interaction matrix and

second-order TF-TF interaction matrix, 4) TF binding correlation matrices, 5) genomic distance between enhancer-promoter pairs.

Hi-C data is used to construct the enhancer-promoter contact frequency matrix[108]. For each enhancer-promoter pair, the enhancer is overlapped with one Hi-C anchor, the promoter is overlapped with the other Hi-C anchor, KR normalized Hi-C contacts between these two anchors were used as the contact frequency between the enhancer-promoter pair. Previous studies have reported that inter-TAD Hi-C interactions' data quality is substantially reduced compared with intra-TAD interactions[151,152]. Therefore, only intra-TAD enhancer-promoter pairs were used to construct the enhancer-promoter pair contact matrix.

ChIA-PET interactions were used as golden-standard to evaluate the model[153]. Specifically, for each intra-TAD enhancer-promoter pair where the enhancer overlaps with one ChIA-PET fragment and the promoter overlaps with the other ChIA-PET fragment, this enhancer-promoter pair will be assigned with 1 if there exists validated ChIA-PET interaction between these two fragments. Otherwise, the enhancer-promoter pair will be assigned 0. The binary matrix was then used as a label to evaluate model performance.

To construct the binary TF binding on the enhancer matrix as well as TF binding on the promoter matrix, we overlapped the TF ChIP-Seq peak coordinates with enhancers' or promoters' coordinates, respectively. If the TF peak overlaps with enhancer coordinates, the TF-enhancer entry value is labeled to be 1, otherwise, 0. Similarly, if the TF peak overlaps with promoter coordinates, the TF-promoter entry value is labeled to be 1, otherwise, 0.

The first-order TF PPI matrix was constructed by labeling the TF-TF links according to the PPIs data, i.e. for each TF-TF pair, if there exist high-quality PPIs between them, then the value between these two TFs is labeled as 1, otherwise 0. All TFs with available ChIP-Seq data from ENCODE that meet the criteria described above were used to construct the first-order TF PPI matrix.

To obtain the second-order TF PPI matrix, the correlation coefficient between two TFs was firstly calculated based on their binding activities on enhancer and promoter respectively. Only correlation coefficients that are greater than 0.6 were used to make the E, and P matrices. Correspondingly, experimental validated PPIs between TF-TF pairs with a correlation greater than 0.6 on both enhancer binding and promoter binding sites were used in the second-order PPIs matrix.

4.2.3 Cluster topological associated domains

It is observed that TFs have high binding to enhancers or promoters in certain domains while low bindings in others. To reduce the variances caused by the TF binding, hierarchical clustering with the walds method were applied to cluster the TAD based on the TFs binding in promoters and enhancers. Specifically, for each TF, the proportion of enhancers or promoters this TF bind to was calculated, and a vector is used to represent the TFs binding abundance on promoters and enhancers for each domain. Ward minimum variance method is applied for the clustering[154]. Four clusters were identified so that there are enough training data (> 100 TADs) and similarity is maintained within each cluster.

4.2.4 Remove genomic distance variance from Hi-C data

To remove the distance variance of Hi-C data, we first fit a linear model between enhancer-promoter contacts and their genomic distance. $\boldsymbol{Y} = \boldsymbol{\gamma D} + \boldsymbol{\gamma_0}$, where $\boldsymbol{Y}$ is the enhancer-promoter contacts matrix, $\boldsymbol{D}$ is the genomic distance between enhancers and promoters. $\boldsymbol{\gamma}$ is solved by fitting a linear model. Residuals $\boldsymbol{R_1}$ between the reconstructed enhancer-promoter interaction frequency using genomic distance and observed enhancer-promoter interaction frequency is calculated: $\boldsymbol{R_1} = \boldsymbol{Y} - (\boldsymbol{\gamma D} + \boldsymbol{\gamma_0})$. $\boldsymbol{R_1}$ is used as input for EP^{3}ICO to optimize the TF-TF interactions that regulate enhancer-promoter interactions and predict enhancer-promoter links (Figure 4.1D).

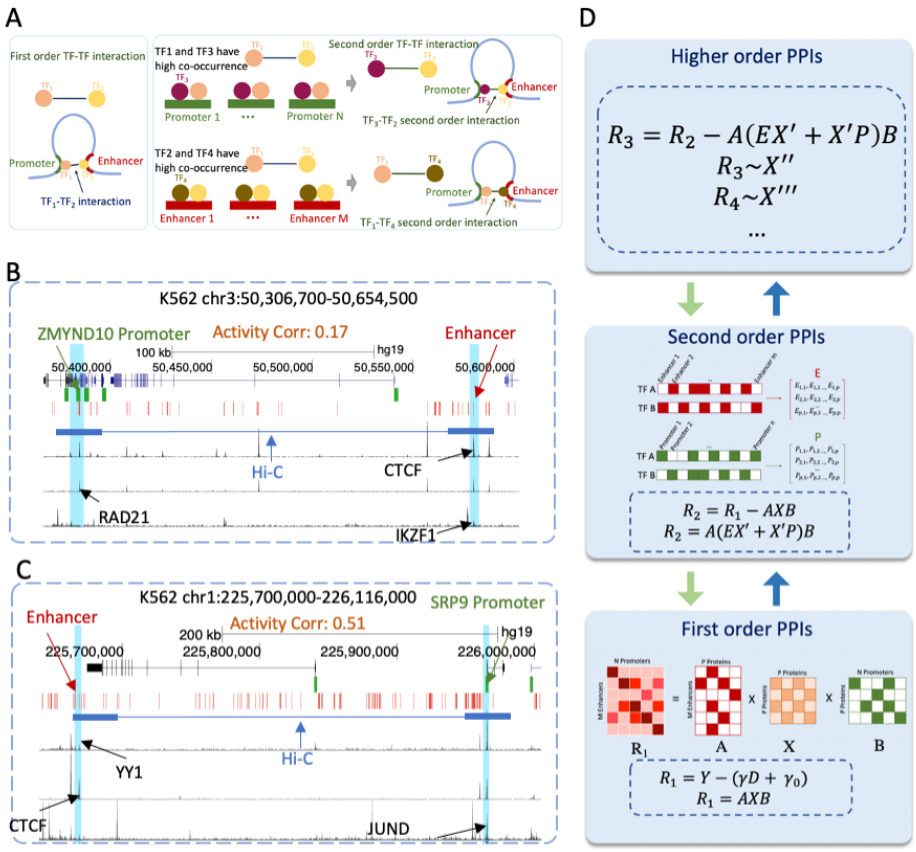

Figure 4.1. EP^{3}ICO infers long-range enhancer-promoter interaction based on TF-TF PPI features. A The Enhancer-promoter links are mediated by multi-order PPIs. Left,

Figure 4.1 (cont'd)

enhancer-promoter interactions are regulated by first order PPIs between enhancer-binding TFs (TF1, yellow) and promoter-binding TFs (TF2, beige). Right, enhancer-promoter interactions are regulated by second-order PPIs. The co-binding of TFs on promoters (TF1 and TF3, purple) or enhancers (TF2 and TF3, dark olive-green) facilitate the second-order PPIs between TF2 and TF3 and TF1 and TF4 respectively. The second-order PPIs regulate long-range enhancer-promoter links. B Examples of second-order PPI between IKZF1-CTCF-RAD21 mediated Hi-C interactions. CTCF and RAD21 have first-order PPIs, the co-binding of IKZF1 and CTCF on promoters provides second-order PPIs between IKZF1 and RAD21. The second-order PPI between IKZF1-CTCF-RAD21 regulates interactions between ZMYND10 and enhancer located 100kb away. C Examples of second-order PPI between CTCF-YY1-JUND mediated Hi-C interactions. YY1 and JUND has first-order PPIs, the co-binding of YY1 and CTCF on enhancer provides second-order PPIs between CTCF and JUND. The second-order PPI between CTCF-YY1-JUND regulates interactions between SRP90 and enhancer located 200kb away. D The flexible framework of EP^{3}ICO for jointly inferring multiple orders of PPIs that mediate chromatin loops and predicting long-range enhancer-promoter links using prioritized TF-TF PPIs. Bottom: a matrix optimization model that prioritizes first-order TF-TF PPIs that regulate enhancer-promoter interactions, X is the first-order PPI matrix. Middle: a matrix optimization model that prioritizes second-order TF-TF PPIs that regulate enhancer-promoter interactions, X' is the second-order PPI matrix. Top: EP^{3}ICO can be further applied to identify higher-order PPIs that mediate chromatin interactions.

4.2.5 Matrix decomposition model is applied to optimize first-order TF-TF interactions

EP^{3}ICO applies the model decomposition model $\boldsymbol{R_1 = AXB}$ to jointly infer the first-order TF-TF interactions that regulate long-range enhancer-promoter links as well as predict long-range enhancer-promoter interactions (Figure 4.1D), where $\boldsymbol{A}$ is the TF-enhancer matrix, $\boldsymbol{B}$ is the TF-promoter matrix, and $\boldsymbol{X}$ is the TF-TF interaction matrix that need to be optimized. Biologically, only a subset of TF PPIs mediates the long-range enhancer-promoter interactions. Therefore, the optimized TF-TF interaction matrix is expected to be a sparse matrix. To make the optimized matrix be sparse, L1-regularization is added to the loss function to increase the sparsity in the optimized PPI matrix as well as avoid over-fitting of the models $\boldsymbol{L(X) = ||R_1 - AXB||_F^2 + \lambda|X|}$. The gradient is derived for the

loss function and is used in the gradient descent method to minimize the loss function. We aim to identify TF-TF interactions from experimentally validated PPIs, thus, only entries with experimentally validated TF-TF interactions will be updated during the optimization progress, and remaining entries will be kept zero. In order to quickly achieve the global optimum, the Barzilai-Borwein method is used to calculate the step size at each iteration in the optimization process[155].

4.2.6 Matrix decomposition model to optimize second-order TF-TF interactions

The first-order TF-TF interaction matrix is optimized in each cluster. Then the prioritized TF-TF interactions are applied to reconstruct long-range enhancer-promoter contact frequency together the genomic distance model in each TAD within the cluster. The residuals between observed and reconstructed enhancer-promoter interaction frequency is calculated: $\boldsymbol{R_2} = \boldsymbol{Y} - (\boldsymbol{\gamma D} + \boldsymbol{\gamma_0}) - \boldsymbol{AXB}$. $\boldsymbol{R_2}$ is used to optimize the second-order TF-TF interactions $\boldsymbol{X'}$: $\boldsymbol{R_2} = \boldsymbol{A(EX' + X'P)B}$, where $\boldsymbol{X'}$ is the second-order TF-TF interaction matrix. Specifically, enhancer-promoter pairs with the top as well as bottom 5% residuals are considered the ones that cannot be solely explained by the first-order TF-TF interactions and are therefore used to prioritize the second-order PPIs. Similarly, the L1-regularization is applied in the matrix factorization model to increase the sparsity and avoid over-fitting $\boldsymbol{L(X')} = \left|\left|\boldsymbol{R_2} - (\boldsymbol{A(EX' + X'P)B})\right|\right|_{\boldsymbol{F}}^{\boldsymbol{2}} + \boldsymbol{\beta}|\boldsymbol{X'}|$.

4.2.7 Model evaluation and performance comparison

Area under the Receiver Operating Characteristic curve (AUROC) and area under the precision-recall curve (AUPR) are used to evaluate both first-order TF-TF interactions and second-order TF-TF interactions model's performance by 5-fold cross-validation. In each

cluster, TADs were randomly split into training and testing sets. EP^{3}ICO was trained on the training TADs and was applied to predict EP pairs in the testing TADs.

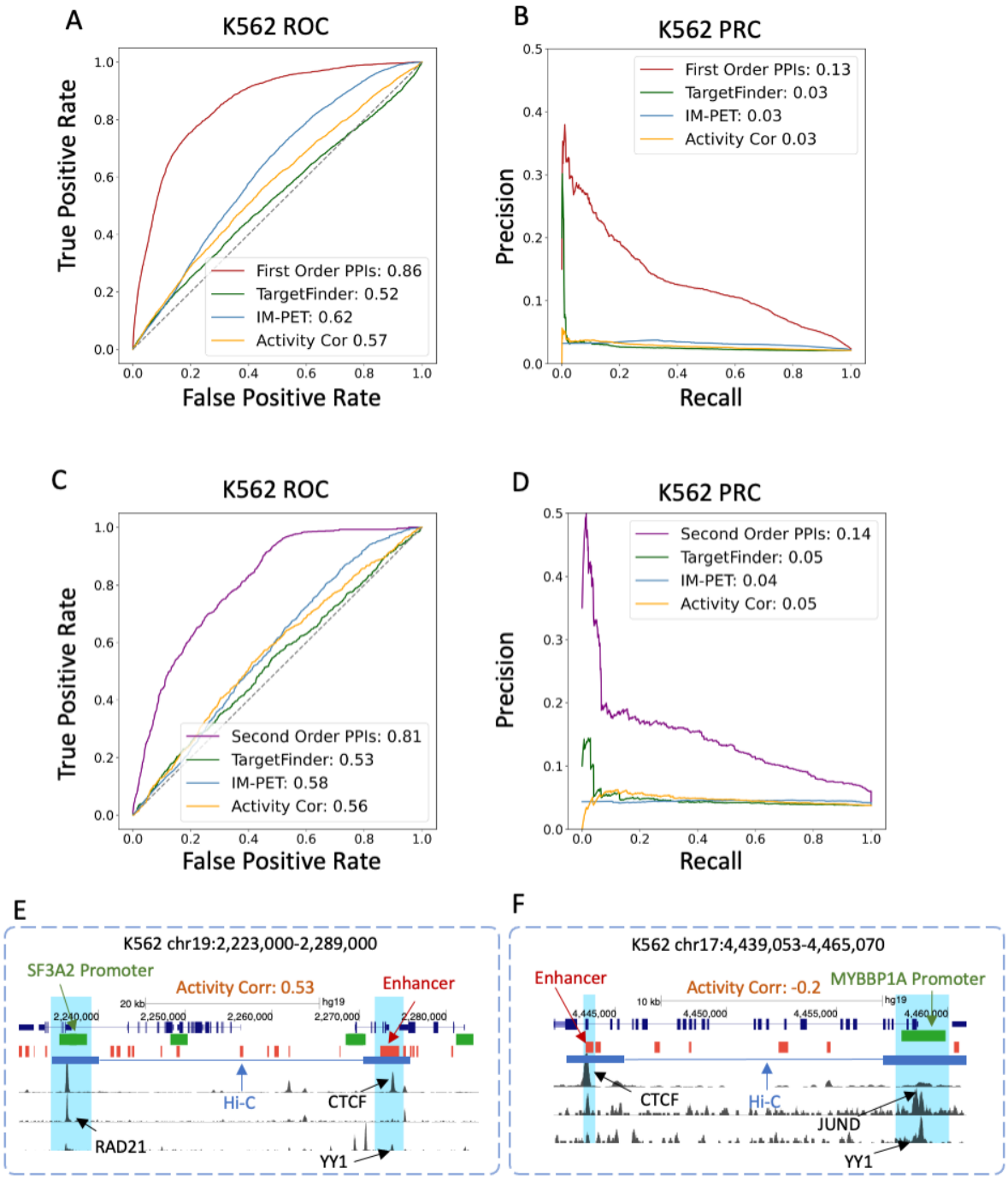

Figure 4.2. Performance comparison in K562 cells. EP^{3}ICO, ProTECT, TargetFinder,

Figure 4.2 (cont'd)

and IM-PET are applied on the same input datasets and are evaluated based on the averaged performance of 5-fold cross-validation. As a baseline comparison, enhancer-gene activity correlations are also included in the analysis. A ROC curves and B PR curves of first-order PPIs performance in K562 cells. C ROC curves and D PR curves of second-order PPIs performance in K562 cells. E, F Examples of enhancer-promoter interactions predicted by EP^{3}ICO second-order PPIs. In each example, the highlighted enhancer (red) is predicted to interact with the highlighted promoter (green) by EP^{3}ICO. Both predictions are supported by cell-type specific Hi-C interactions (blue paired lines). The prioritized TF PPIs mediating the interactions are second-order PPIs RAD21-CTCF-YY1 (E) and YY1-JUND-CTCF (F), respectively.

The training and testing data have high imbalance and the imbalanced dataset might inflate the model's performance. To justify the robustness of EP^{3}ICO, a balanced dataset with the same number of positive and negative samples is also generated by downsampling for cross-validation. Furthermore, the genomic distance between the enhancer and promoter pair might be a confounding factor that dominates the model performance. Thus, a balanced dataset with genomic distance control is further generated to evaluate the model performance.

The performance of our EP^{3}ICO were also evaluated and compared with two state of art models, TargetFinder and IM-PET, which also leverage TF binding features in their model[124,156]. Both TargetFinder and IM-PET integrate activity-based features, genomic distance, as well as TF binding information in enhancers and promoters. Additionally, TargetFinder also used the TF binding information in the windows region between enhancers and promoters. Using the same set of TF ChIP-Seq peaks to generate the TF features in the window between enhancer-promoter pair, TargetFinder is trained on the Hi-C data according to the data processing procedure in the paper. The trained model is then applied to the same testing dataset to make the predictions. IM-PET is implemented

to the same dataset. As IM-PET automatically predicts enhancer-promoter pairs with genomic distance < 2Mb, only common enhancer-gene pairs are used to evaluate performances, resulting in a fair comparison among the three models. By comparing our first-order PPIs model TargetFinder and IM-PET, we demonstrate that PPIs information can greatly improve the accuracy in the predicting of enhancer-promoter links.

The prioritized second-order PPIs are used to predict enhancer-promoter interactions on the subset of enhancer-promoter pairs (see methods). By comparing the second-order PPIs with the first-order PPI's performance on the same dataset, we demonstrate that second-order PPIs can provide additional information for predicting enhancer-promoter links. ProTECT applies Random Forest model to integrate TF-TF interactions modules to predict enhancer-promoter interactions[157]. In order to demonstrate the additional information provided by the second-order PPIs, we further compare our second-order PPI model with ProTECT. Only the common enhancer gene pairs among all models were used to evaluate the model performance. 5-fold cross-validation is applied and the average AUCs are used to report the final performances.

4.2.8 TF shuffling

In order to quantitively check if TFs binding, as well as TF PPIs, indeed contribute to the high accuracy of EP^{3}ICO, we randomly permute the TFs binding on the enhancers and genes as well as permute the PPIs between TFs with the degrees of each TFs unchanged and the numbers of TFs that bind on enhancer or promoter maintained respectively. The First-order PPI model was trained and tested on the same datasets using permuted TF bindings and TF PPIs. The AUCs were used to compare the first-order PPI model trained with the original dataset. For the second-order PPIs model, the TF binding on the

enhancer correlation matrix, i.e, E matrix, as well as the TF binding on the promoter, i.e P matrix is calculated based on the permuted TF binding on the enhancers and promoters, respectively. Then, the second-order PPI model was trained using permuted E, P, and second-order PPI matrix on the same training set used for training the original second-order PPI matrix, and the AUCs were compared to the original second-order PPI model on the same testing dataset. Balanced dataset as well as distance-controlled balanced data were also generated to evaluate the model performance based on shuffled PPIs.

4.2.9 EP^{3}ICO predict enhancer-promoter interactions based on imputed cell-specific TF bindings

In order to assess if EP^{3}ICO can be transferred to cell lines or tissues when TF ChIP-Seq data is not accessible, we impute cell-line specific TF bindings in K562 and implement EP^{3}ICO based on the imputed TF bindings.

To impute the TFs binding, the TF motifs coordinates discovered from ENCODE ChIP-Seq data were collected[158]. Firstly, TF motif coordinates located in close chromatin were filtered out and motifs in accessible chromatin were used to impute the TF binding. Secondly, TFs that have lower gene expressions than 0.6 were further removed to avoid the high false positives. Finally, imputed TF binding sites at enhancers and promoters are generated with TFs pass the criteria by overlapping these TFs' imputed binding sites with enhancers and promoters through BEDtools[159] (see methods). These TFs are then used to construct first-order and second-order TF-TF interactions matrices and applied in optimization models.

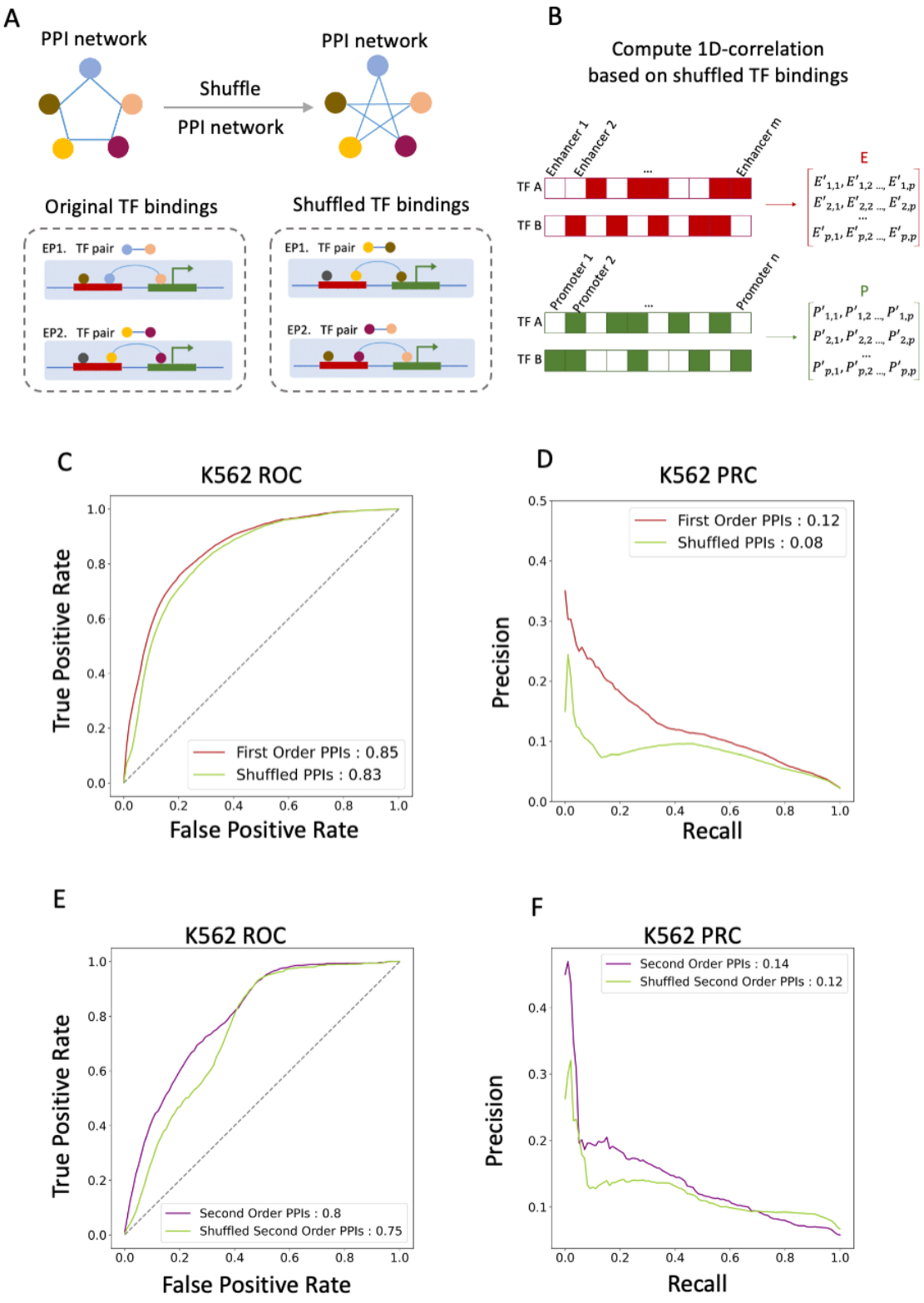

Figure 4.3. TF PPI features provide additional information beyond TF bindings. A Schematic figure of the permutation test on TF PPI features. The shuffled PPIs are

Figure 4.3 (cont'd)

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. B Correspondingly, the co-occurrence matrices of TFs, E, and P are generated based on the shuffle TF bindings. Based on the shuffled PPIs, new matrix factorization models were trained and then evaluated by the same cross-validation procedure for the first-order and second-order PPIs respectively. C, D ROC, and PR plots for the models based on the original first-order TF PPI features (red), the models based on the shuffled first-order TF PPI features (green). E, F ROC and PR plots for the models based on the original second-order TF PPI features (purple), the models based on the shuffled second-order TF PPI features (green).

4.2.10 cis-eQTL enrichment analysis for predicted long-range enhancer-gene interactions

To predict the enhancer-promoter interactions, the prioritized first-order PPIs were applied to reconstruct the score of the enhancer-promoter pairs. The top 10% predicted enhancer-promoter pairs were selected to be the final predictions. Furthermore, optimized second-order PPIs were applied to predict enhancer-promoter links on the subset of data we used to learn the second-order PPIs. And the top 10% predictions in the testing data from the subset were used as the final predictions.

cis-eQTL from matched cell lines also provides orthogonal information in validating the accuracy of genome-wide predictions generated by our model. As we made the genome-wide enhancer-promoter interactions in the K562 cell line, we collected three eQTL datasets that are profiled from either lymphoblastoid cells or blood tissue[156,160,161]. A predicted enhancer-promoter interaction is considered to be supported by the cis-eQTL if the SNP overlaps with the enhancer and the promoter matches with the gene. The fraction of predicted enhancer-promoter links was calculated for each eQTL dataset as the enrichment score. The overlapping fraction of SNP-gene pairs of the top 10% predictions of IM-PET were also used as a comparison for the first-order PPIs predictions.

The top 5% predictions of IM-PET were used to calculate the overlapping fractions of eQTLs as a comparison to our second-order PPIs predicted enhancer-gene pairs. Wilcoxon signed-rank test was used to check the statistical significance.

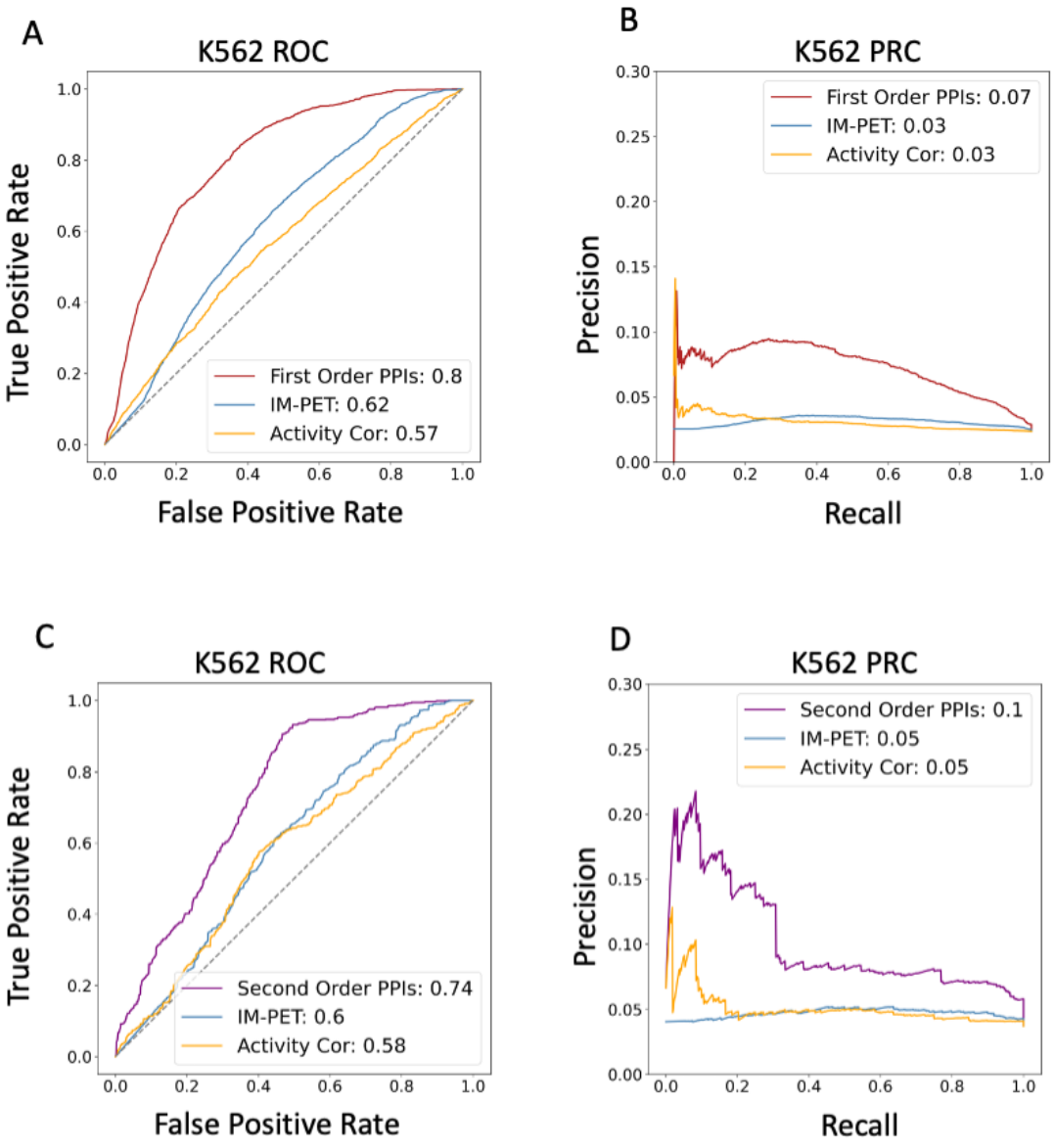

Figure 4.4. Imputed TF bindings can accurately predict enhancer-promoter interactions. EP^{3}ICO uses imputed TF binding to predict long-range enhancer-promoter interactions and is compared with IM-PET on the same input datasets. Performance is

Figure 4.4 (cont'd)

evaluated based on the averaged performance of 5-fold cross-validation. As a baseline comparison, enhancer-gene activity correlations are also included in the analysis. A ROC curves and B PR curves of first-order PPIs performance in K562 cells based on imputed TF binding. C ROC curves and D PR curves of second-order PPIs performance in K562 cells based on imputed TF binding.

4.3 RESULTS

4.3.1 Predict long-range enhancer-promoter interactions based on multi-order PPIs between TFs

Previous experimental studies have identified that PPIs between specific TFs mediate long-range enhancer-promoter interactions through the binding to promoters and enhancers respectively. Two examples of first-order PPI-mediated Hi-C enhancer-gene interactions are shown in Figure C.1. In the first example, PPI between RAD21 and CTCF regulates gene TTC31's promoter with a distal enhancer located 100kb away. In the second example, PPI between CTCF and YY1 mediates the long-range interaction between gene E2F2's promoter and an enhancer located 20kb away.

Moreover, higher-order PPIs between TFs are also found to participate in the regulation of long-range chromatin loops [138–142]. Two second-order PPI facilitated Hi-C enhancer-promoter interaction examples are shown in (Figure 4.1B, 4.1C). The first example shows that IKZF1 maintains a second-order PPI with RAD21 through the co-binding with CTCF on enhancer. The second-order PPI between IZKF1 and RAD21 facilitates long-range links between gene ZMYND10 with an enhancer located 100kb away. Without the co-binding of IKZF1 on enhancer, PPI between CTCF and RAD21 alone is not able to facilitate the long-range chromatin loops for certain enhancer-promoter pairs. The second example shows second-order PPI between CTCF-YY1-JUND mediated Hi-C interactions.

YY1 and JUND has first-order PPIs, the co-binding of YY1 and CTCF on enhancer provides second-order PPIs between CTCF and JUND. The second-order PPI between CTCF-YY1-JUND regulates interactions between SRP90 and enhancer located 200kb away.

A predictive model, especially a linear model can easily suffer from overfitting problems due to the large feature dimensions. To effectively avoid the overfitting problems, we applied two steps to increase the feature sparsity. First, before the training of the model, we removed PPIs with a confidence score < 100 to manually reduce the feature dimension as well as control the input feature quality. Secondly, we added L1-regularization to the matrix factorization model we developed. L1-regularization shrinks the less important features' weights to zero, resulting in a sparse optimized matrix. The sparse features will avoid over-fitting issues effectively as well as preserve the robustness of the model[162].

The first-order PPI models are trained on a high-resolution Hi-C dataset from human K562 cell line. The Hi-C data was normalized by KR normalization method (see methods). The enhancer-gene contact frequency matrix in each TAD was constructed from the KR normalized matrix, and the variance caused by genomic distance was removed by fitting a linear model between enhancer-promoter distance and contacts (see methods). The residual matrix was computed by subtracting the prediction from the observation. The residual matrix was decomposed to into three matrices including the TF-enhancer matrix, the first-order PPIs interaction matrix, and the TF-promoter matrix. The learned first-order PPIs were optimized through an optimization process. Prioritized first-order TF-TF interactions were applied to predict genome-wide enhancer gene interaction contacts.

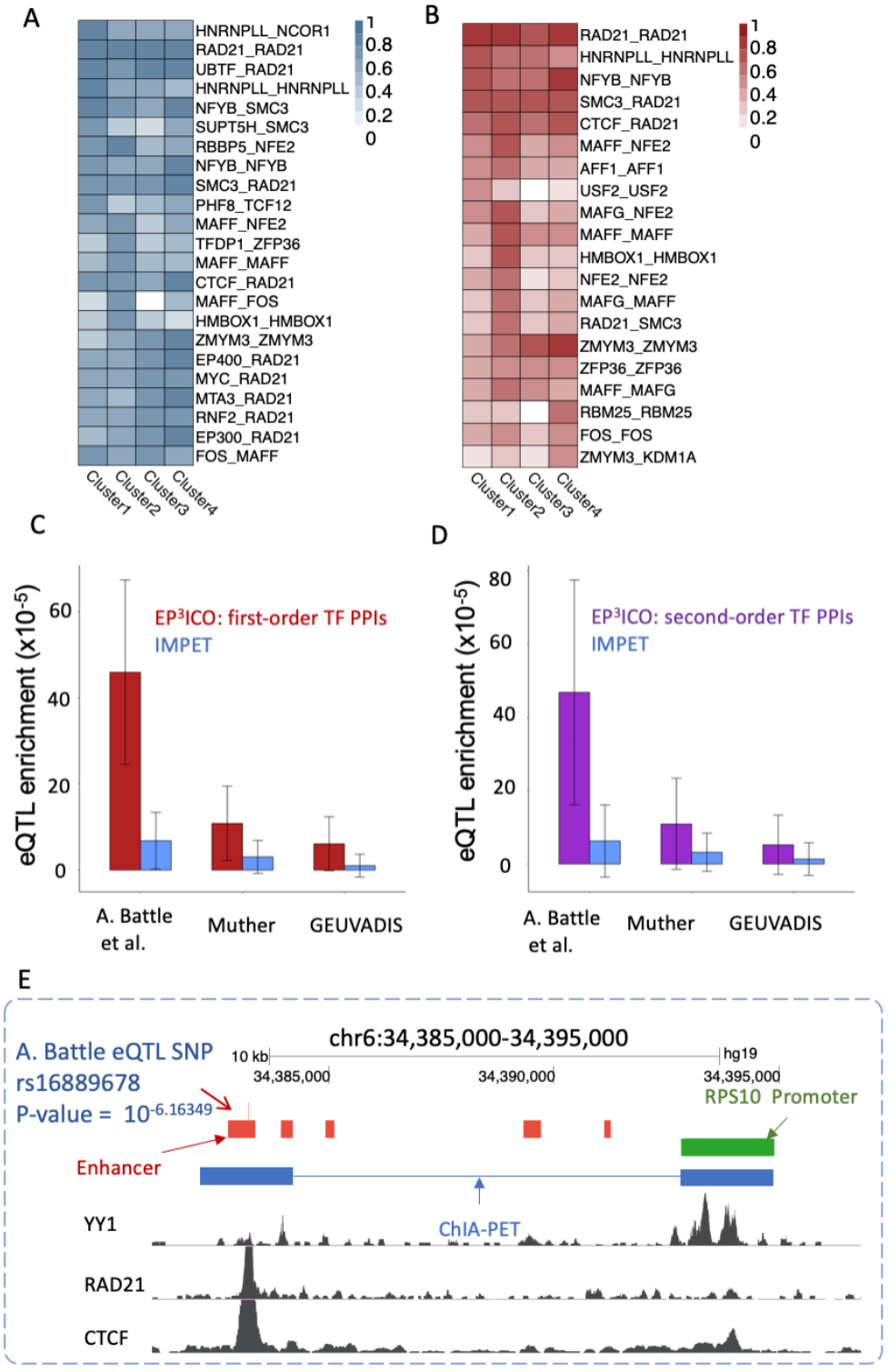

Figure 4.5. EP^{3}ICO identifies functional TF PPIs that regulate enhancer-promoter interactions. A,B Prioritized first-order (A) and second-order (B) TF PPIs in four clusters.

Figure 4.5 (cont'd)

Min-Max normalization is applied to the TF PPIs optimized weights in each cluster, top 20 optimized TF PPIs are selected based on the normalized weights in each cluster. The union of PPIs is used to plot the heatmap. The higher the number, the more important TF PPI is in regulating long-range enhancer-promoter interactions. C D cis-eQTLs from multiple datasets (x- axis) are significantly enriched in first-order TF PPIs (C) predicted enhancer-promoter interactions and second-order PPIs (D) predicted enhancer-promoter interactions in K562 (red). The fractions of enhancer-promoter interactions overlapping with cis-QTLs (y-axis) are compared with IM-PET. Error bars represent sd. E Example of a cis-eQTL, i.e. the rs16889678-RPS10 pair, overlapping with a second-order PPI predicted enhancer-promoter interaction. The predicted interaction is supported by ChIA-PET (blue paired lines). The prioritized PPI feature is YY1-CTCF-RAD21, consistent with the ChIP-seq signal tracks (black signals).

Differences between the observed enhancer-promoter pair contacts and predicted enhancer-promoter contacts were calculated. Enhancers-promoter pairs with top 5% and bottom 5% residuals were then selected to be further decomposed to learn the second-order PPIs information. The prioritized second-order PPIs were then applied to boost the prediction for enhancer-promoter pairs in the subset (Figure 4.1D). The prioritized first-order PPIs as well as second-order PPIs between TFs provide new insights for understanding the complex interplay between TFs, enhancers, and genes. Further cis-eQTL analysis based on the predicted first-order TF PPIs as well as second-order TF PPI predicted enhancer-promoter interactions provide a new platform for understanding the human genetic analyses.

4.3.2 TF PPIs features provide boosted performance

Using the genome TAD-split cross-validation strategy, we tested the prediction accuracy of the first-order PPI model as well as compared our model with two other supervised machine learning models, TagetFinder and IM-PET in K562 cell line. Our first-order PPI model achieved the highest performance, with AUROCs are 0.86. IM-PET ranked second

with AUROC being 0.57. TargetFinder has the lowest AUROC value 0.52. Since the testing data has a high imbalance, the AUPR is also used to assess the performance of all of the models. Our model has the highest performance again, with an AUPR are 0.13. As a baseline comparison, the activity correlation between the enhancer and gene pair is also calculated as a prediction score and used in the comparison. The AUROC and AUPR of activity correlation are 0.57 and 0.03 respectively, similar to IM-PET, which uses activity correlation as the main feature in their model (Figure 4.2A, 4.2B). As the data set is highly imbalanced, we also downsampled balanced data in two ways and evaluate the model performance on the balanced dataset. First, we randomly select a negative dataset that have the same number of samples as the positive dataset. By assessing the models' performance on the balanced data, EP^{3}ICO still shows the highest accuracy (Figure C.4A). Secondly, we performed the downsampling with a more stringent strategy by controlling the genomic distance. Specifically, instead of randomly selecting negative pairs, we selected the negative pairs that have the same genomic distance distribution with the positive pairs. By evaluating the models' performance on the genomic-distance controlled data, EP^{3}ICO again shows the highest accuracy (Figure C.6A).

Two examples of predicted enhancer-promoter interactions are shown in Figure C.2. The first enhancer-gene pair example, where the enhancer locates 50kb away from the enhancer, has Hi-C validated interactions. The activity correlation between enhancer and gene is 0, which indicates a false negative prediction based on the activity correlation. However, our prioritized PPI between CTCF and RAD21 reconstructed the long-range enhancer-promoter interaction with CTCF binding on the enhancer and RAD21 binding on the promoter. The second enhancer-gene pair example have a low activity correlation,

i.e. 0.4 while with validated Hi-C interaction. Again, the prioritized PPIs between EP300 and RAD21 have reconstructed this long-range enhancer-promoter pair with EP300 binding on the enhancer and RAD21 binding on the promoter.

A subset of enhancer-promoter pairs can't be reconstructed by our first-order PPI model, we further developed a matrix factorization model and trained the model on the subset enhancer-gene pairs to identify second-order PPIs required for enhancer-promoter regulation (Figure 4.1D). By comparing the second-order PPI model with the first-order PPI model, we further demonstrated that second-order PPIs boosted the accuracy in predicting long-range enhancer-promoter links that could not be predicted by first-order PPIs only (Figure C.3). By generating the balanced dataset and evaluate the models on balanced data, we further proved the robust performance of second-order PPI model (Figure C. 4B, Figure C. 6B). ProTECT[157] is another supervised random forest model that uses PPI modules as main features to predict long-range enhancer-promoter interactions. However, ProTECT doesn't use higher order PPIs as their features. To justify that higher-order PPIs can provide additional information, we compared our second-order PPI model with ProTECT, TargetFinder, IM-PET, and the baseline model on the common data. The results show that our second-order PPI model overperforms ProTECT as well as other three models, suggesting there are more information provided by the second-order PPIs (Figure C. 8).

Two predicted examples using second-order PPIs are presented in Figure 4.2E, 4.2F. In the first example, YY1 maintains a second-order PPI with RAD21 through the co-binding with CTCF on the enhancer and leading to the regulation of this enhancer's interaction with a gene located 20kb away. In the second example, JUND has second-order PPI in

CTCF through the co-binding with YY1. The second-order PPIs regulate the long-range enhancer-promoter interaction where the enhancer is 10kb away from the promoter.

As we used a subset to learn the second-order PPIs, our second-order PPI model was compared with TargetFinder and IM-PET on the same testing dataset. Still, our second-order PPI model has the highest performance with AUROC and AUPR are 0.81 and 0.14 respectively (Figure 4.2C, 4.2D). The results not only demonstrate that our second-order PPIs model has superior performance, but also our matrix factorization model is robust on different datasets.

4.3.3 TF PPI features provide additional information beyond TF bindings and activity-based features

To further justify the superior performance of indeed result from the information of first-order TF PPIs features, we permuted the TF bindings on the enhancer and promoters with the numbers of TFs on the enhancers or promoters being strictly maintained. Moreover, the TF-TF interactions were shuffled randomly with the degree of PPI partner remained (Figure 4.3A). Correspondingly, the E and P matrices were calculated based on the shuffled TF bindings to have the shuffled second-order PPIs (Figure 4.3B). We trained EP^{3}ICO on the data y using the permuted TF bindings and the shuffled first-order as well as shuffled second-order TF-TF interactions, and compared the model performance with using the original data. We observed that using the shuffle TF-TF features decreases the accuracy of both first-order's and second-order's predictions in the K562 cell line (Figure 4.3 C-F). Balanced dataset was also generated by downsample same number of enhancer-promoter pairs that are not supported by called ChIP-PET interactions. Balanced dataset was generated with distance controlled as well as in a

random way, i.e distance is not controlled. The models' performance was further compared on these two kinds of balanced dataset. The results further demonstrated that original TF bindings and TF-TF interactions can reconstruct the enhancer-promoter interactions more accurately (Figure C.5, Figure C.7). The difference in prediction accuracy suggests the boosted performance of EP^{3}ICO is contributed by the TFs bindings and the TF PPI features.

4.3.4 cis-eQTL are enriched in the first-order PPIs and second-order PPIs predicted long-range enhancer-promoter interactions.

The prioritized first-order PPIs are applied to predict enhancer-promoter interactions on the testing dataset. We selected the top 10% predictions of enhancer-promoter links based on the reconstructed interaction scores as the final predictions. As we trained the model on the Hi-C data, cis-eQTLs were used as orthogonal evidence to evaluate the accuracy of predicted enhancer-promoter interactions. By calculating the fractions of predicted enhancer-promoter pairs supported with SNP-gene pairs of significant eQTLs, we compared the overlapping enrichment scores of EP^{3}ICO with IM-PET. Compared with IM-PET, the enhancer-promoter links predicted by first-order PPIs have significantly higher fractions overlapping with eQTLs in K562 for all three eQTL datasets (p-values < 5.50e-08, 6.03e-04, and 7.45e-04 respectively, Figure 4.5C). Moreover, the prioritized second-order PPIs were also applied to predicted enhancer-promoter interactions on the testing set from the subset data. The top 10% reconstructed enhancer-promoter links were selected as the predictions. The same cis-eQTL datasets were applied to check the enrichment scores of EP^{3}ICO with IM-PET. EP^{3}ICO has significantly higher overlapping

fractions in K562, the p-values < 2.54e-06, 0.0171, 0.0187 respectively (Figure 4.5D). The results suggest high accuracy of EP^{3}ICO predictions.

4.4 DISCUSSION

In this study, we develop a matrix factorization model, EP^{3}ICO to infer multi-order PPIs between TFs that regulate long-range enhancer-promoter interactions. Using the optimized multi-order TF-TF interactions, EP^{3}ICO accurately predicts long-range enhancer-promoter interactions and achieves superior performance compared with existing methods. By incorporating higher-order PPIs between TFs, EP^{3}ICO further improved the prediction performance on a subset of enhancer-promoter interactions that can't be reconstructed accurately by first-order TF-TF interactions.

cis-eQTLs are also applied to validate the predictions of EP^{3}ICO as the orthogonal evidence. Enrichment scores of cis-eQTLs in EP^{3}ICO predicted enhancer-promoter links are compared with enrichment scores of IM-PET predicted enhancer-promoter interactions. The results show that prioritized first-order PPIs and second-order PPIs reconstructed enhancer-promoter links have significantly higher fractions than IM-PET. The promising enrichment analysis further indicates the predictions of EP^{3}ICO can be used as a platform to characterize the non-coding SNP's effects propagated through 3D chromatin interactions.

The K562 cell line has comprehensive profiling of TF bindings sites, which provides the data for EP^{3}ICO. However, there remain cell lines or tissues that have limited data sets. To address the data limitation issue, we impute TF bindings based on the TF motifs, TF gene expression level, and chromatin accessibility. By using the imputed TF bindings, EP^{3}ICO still achieves superior performance for both first-order PPIs and second-order

PPIs predicted enhancer-promoter links. The results further demonstrate the accuracy of EP^{3}ICO and its generalizability.

The major novelty of EP^{3}ICO is the inclusion of multi-order TF PPIs as features. The first-order TF PPI can reconstruct and explain the majority of enhancer-promoter links. However, a subset of enhancer-promoter interactions cannot be predicted by first-order TF PPIs only, second-order TF PPIs can be applied to improve the predictions. Moreover, EP^{3}ICO is a flexible workflow and can be further applied to identify higher-order PPIs that regulate the 3D genome.

CHAPTER 5

FUTURE DIRECTIONS

Identification of transcription factors binding sites (TFBSs) and charactering chromatin interactions in the 3D space are critical for understanding the gene regulatory networks. In this dissertation, we presented the studies that were conducted in both directions to decode the gene regulatory rhythms by integrating the multi-omics data. We developed a supervised machine learning model, TF-wave, that uses DNase-Seq data to predict TFBSs within cell lines. Next, we designed a matrix factorization model, EP^{3}ICO, that jointly infers long-range enhancer-promoter interactions and multi-order PPIs that mediate the 3D genome.

Although TF-wave has achieved superior performance in predicting intra-cell transcription factors footprints, inter-cell line prediction is not investigated yet and will be one of the future directions we will focus on. Another future direction is applying ATAC-Seq data as features to predict TFBSs. Moreover, single-cell ATAC-Seq (scATAC-Seq) data is available in many cell lines, we will future take use of the scATAC-Seq to predict TFBS at single-cell level.

EP^{3}ICO achieves high accuracy in predicting long-range enhancer-promoter links as well as provides mechanistic insights of 3D genome. We will continue working on exploring more on the roles of epigenetic features including histone modifications in shaping the 3D -genome. By integrating new epigenetics features, we will further improve the accuracy of mathematical models in predicting long-range enhancer-promoter links.

BIBLIOGRAPHY

1. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* (2018) doi:10.1016/j.cell.2018.01.029.

2. Ouyang, N. & Boyle, A. P. TRACE: Transcription factor footprinting using chromatin accessibility data and DNA sequence. *Genome Res.* (2020) doi:10.1101/gr.258228.119.

3. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science (80-.).* (2007) doi:10.1126/science.1141319.

4. Levy, S. & Hannenhalli, S. Identification of transcription factor binding sites in the human genome sequence. *Mamm. Genome* (2002) doi:10.1007/s00335-002-2175-6.

5. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* (2007) doi:10.1038/nature06008.

6. Li, H., Quang, D. & Guan, Y. Anchor: Trans-cell type prediction of transcription factor binding sites. *Genome Res.* (2019) doi:10.1101/gr.237156.118.

7. Rhee, H. S. & Pugh, B. F. ChIP-exo: A Method to Identify Genomic Location of DNA-binding proteins at Near Single Nucleotide Accuracy. *Curr. Protoc. Mol. Biol.* (2008).

8. He, Q., Johnston, J. & Zeitlinger, J. ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nat. Biotechnol.* (2015) doi:10.1038/nbt.3121.

9. Ernst, J., Plasterer, H. L., Simon, I. & Bar-Joseph, Z. Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Res.* (2010) doi:10.1101/gr.096305.109.

10. Boyle, A. P. *et al.* High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell* (2008) doi:10.1016/j.cell.2007.12.014.

11. Zhong, J. *et al.* Mapping nucleosome positions using DNase-seq. *Genome Res.* (2016) doi:10.1101/gr.195602.115.

12. Funk, C. C. *et al.* Atlas of Transcription Factor Binding Sites from ENCODE DNase Hypersensitivity Data across 27 Tissue Types. *Cell Rep.* (2020) doi:10.1016/j.celrep.2020.108029.

13. Yardimci, G. G., Frank, C. L., Crawford, G. E. & Ohler, U. Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. *Nucleic Acids Res.* (2014) doi:10.1093/nar/gku810.

14. Kähärä, J. & Lähdesmäki, H. BinDNase: A discriminatory approach for transcription factor binding prediction using DNase I hypersensitivity data. *Bioinformatics* (2015) doi:10.1093/bioinformatics/btv294.

15. Pique-Regi, R. *et al.* Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* (2011) doi:10.1101/gr.112623.110.

16. Quach, B. & Furey, T. S. DeFCoM: Analysis and modeling of transcription factor binding sites using a motif-centric genomic footprinter. *Bioinformatics* (2017) doi:10.1093/bioinformatics/btw740.

17. Sherwood, R. I. *et al.* Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat. Biotechnol.* (2014) doi:10.1038/nbt.2798.

18. Gusmao, E. G., Allhoff, M., Zenke, M. & Costa, I. G. Analysis of computational footprinting methods for DNase sequencing experiments. *Nat. Methods* (2016) doi:10.1038/nmeth.3772.

19. Piper, J. *et al.* Wellington: A novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Res.* (2013) doi:10.1093/nar/gkt850.

20. Rabiner, L. R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. IEEE* (1989) doi:10.1109/5.18626.

21. Speed, T. P. Biological sequence analysis. in *Selected Works of Terry Speed* (2012). doi:10.1007/978-1-4614-1347-9_14.

22. Vierstra, J. *et al.* Global reference mapping of human transcription factor footprints. *Nature* (2020) doi:10.1038/s41586-020-2528-x.

23. Wickerhauser, M. V. The Discrete Wavelet Transform. in *Adapted Wavelet Analysis* (2020). doi:10.1201/9781439863619-11.

24. Kim, J. *et al.* Stacked auto-encoder based CNC tool diagnosis using discrete wavelet transform feature extraction. *Processes* (2020) doi:10.3390/PR8040456.

25. Bailey, D. H. & Swarztrauber, P. N. A Fast Method for the Numerical Evaluation of Continuous Fourier and Laplace Transforms. *SIAM J. Sci. Comput.* (1994) doi:10.1137/0915067.

26. Mateo, C. & Talavera, J. A. Short-time Fourier transform with the window size fixed in the frequency domain. *Digit. Signal Process. A Rev. J.* (2018) doi:10.1016/j.dsp.2017.11.003.

27. Mateo, C. & Talavera, J. A. Short-Time Fourier Transform with the Window Size

Fixed in the Frequency Domain (STFT-FD): Implementation. *SoftwareX* (2018) doi:10.1016/j.softx.2017.11.005.

28. Liu, C.-L. A Tutorial of the Wavelet Transform. *History* (2010).

29. de Santiago, I. & Carroll, T. Analysis of ChIP-seq data in R/Bioconductor. in *Methods in Molecular Biology* (2018). doi:10.1007/978-1-4939-7380-4_17.

30. Castro-Mondragon, J. A. *et al.* JASPAR 2022: The 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* (2022) doi:10.1093/nar/gkab1113.

31. Korhonen, J., Martinmäki, P., Pizzi, C., Rastas, P. & Ukkonen, E. MOODS: Fast search for position weight matrix matches in DNA sequences. *Bioinformatics* (2009) doi:10.1093/bioinformatics/btp554.

32. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* (2009) doi:10.1093/bioinformatics/btp352.

33. Natekin, A. & Knoll, A. Gradient boosting machines, a tutorial. *Front. Neurorobot.* (2013) doi:10.3389/fnbot.2013.00021.

34. Lee, G., Gommers, R., Waselewski, F., Wohlfahrt, K. & O'Leary, A. PyWavelets: A Python package for wavelet analysis. *J. Open Source Softw.* (2019) doi:10.21105/joss.01237.

35. Navarro Gonzalez, J. *et al.* The UCSC genome browser database: 2021 update. *Nucleic Acids Res.* (2021) doi:10.1093/nar/gkaa1070.

36. Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* (2016) doi:10.1101/gr.200535.115.

37. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* (2015) doi:10.1038/nmeth.3547.

38. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* (2015) doi:10.1038/nbt.3300.

39. Dere, E., Lo, R., Celius, T., Matthews, J. & Zacharewski, T. R. Integration of Genome-Wide Computation DRE Search, AhR ChIP-chip and Gene Expression Analyses of TCDD-Elicited Responses in the Mouse Liver. *BMC Genomics* (2011) doi:10.1186/1471-2164-12-365.

40. Mimura, J. & Fujii-Kuriyama, Y. Functional role of AhR in the expression of toxic effects by TCDD. in *Biochimica et Biophysica Acta - General Subjects* (2003). doi:10.1016/S0304-4165(02)00485-3.

41. Sogawa, K. & Fujii-Kuriyama, Y. Ah receptor, a novel ligand-activated transcription factor. *Journal of Biochemistry* (1997) doi:10.1093/oxfordjournals.jbchem.a021864.

42. Beischlag, T. V., Morales, J. L., Hollingshead, B. D. & Perdew, G. H. The aryl hydrocarbon receptor complex and the control of gene expression. *Critical Reviews in Eukaryotic Gene Expression* (2008) doi:10.1615/CritRevEukarGeneExpr.v18.i3.20.

43. Poland, A. & Glover, E. Comparison of 2,3,7,8 tetrachlorodibenzo p dioxin, a potent inducer of aryl hydrocarbon hydroxylase, with 3 methylcholanthrene. *Mol. Pharmacol.* (1974).

44. Nebert, D. W. & Gelboin, H. V. The in vivo and in vitro induction of aryl hydrocarbon hydroxylase in mammalian cells of different species, tissues, strains, and developmental and hormonal states. *Arch. Biochem. Biophys.* (1969) doi:10.1016/0003-9861(69)90253-7.

45. Chen, H. S. & Perdew, G. H. Subunit composition of the heteromeric cytosolic aryl hydrocarbon receptor complex. *J. Biol. Chem.* (1994) doi:10.1016/s0021-9258(18)47020-2.

46. Petrulis, J. R. & Perdew, G. H. The role of chaperone proteins in the aryl hydrocarbon receptor core complex. *Chem. Biol. Interact.* (2002) doi:10.1016/S0009-2797(02)00064-9.

47. Larigot, L., Juricek, L., Dairou, J. & Coumoul, X. AhR signaling pathways and regulatory functions. *Biochimie Open* (2018) doi:10.1016/j.biopen.2018.05.001.

48. Guyot, E., Chevallier, A., Barouki, R. & Coumoul, X. The AhR twist: Ligand-dependent AhR signaling and pharmaco-toxicological implications. *Drug Discovery Today* (2013) doi:10.1016/j.drudis.2012.11.014.

49. Murray, I. A., Patterson, A. D. & Perdew, G. H. Aryl hydrocarbon receptor ligands in cancer: Friend and foe. *Nature Reviews Cancer* (2014) doi:10.1038/nrc3846.

50. Denison, M. S., Fisher, J. M. & Whitlock, J. P. The DNA recognition site for the dioxin-Ah receptor complex. Nucleotide sequence and functional analysis. *J. Biol. Chem.* (1988).

51. Korkalainen, M., Lindén, J., Tuomisto, J. & Pohjanvirta, R. Effect of TCDD on mRNA expression of genes encoding bHLH/PAS proteins in rat hypothalamus. *Toxicology* (2005) doi:10.1016/j.tox.2004.11.003.

52. Esser, C., Rannug, A. & Stockinger, B. The aryl hydrocarbon receptor in immunity. *Trends in Immunology* (2009) doi:10.1016/j.it.2009.06.005.

53. Yang, S. Y., Ahmed, S., Satheesh, S. V. & Matthews, J. Genome-wide mapping

and analysis of aryl hydrocarbon receptor (AHR)- and aryl hydrocarbon receptor repressor (AHRR)-binding sites in human breast cancer cells. *Arch. Toxicol.* (2018) doi:10.1007/s00204-017-2022-x.

54. Jaeger, C. & Tischkau, S. A. Role of Aryl Hydrocarbon Receptor in Circadian Clock Disruption and Metabolic Dysfunction. *Environmental Health Insights* (2016) doi:10.4137/EHI.S38343.

55. Shimba, S. & Watabe, Y. Crosstalk between the AHR signaling pathway and circadian rhythm. *Biochemical Pharmacology* (2009) doi:10.1016/j.bcp.2008.09.040.

56. Kalmes, M., Hennen, J., Clemens, J. & Blömeke, B. Impact of aryl hydrocarbon receptor (AhR) knockdown on cell cycle progression in human HaCaT keratinocytes. *Biol. Chem.* (2011) doi:10.1515/BC.2011.067.

57. Nacarino-Palma, A. *et al.* The aryl hydrocarbon receptor promotes differentiation during mouse preimplantational embryo development. *Stem Cell Reports* (2021) doi:10.1016/j.stemcr.2021.08.002.

58. Gialitakis, M. *et al.* Activation of the Aryl Hydrocarbon Receptor Interferes with Early Embryonic Development. *Stem Cell Reports* (2017) doi:10.1016/j.stemcr.2017.09.025.

59. Kim, T. H. & Dekker, J. ChIP-seq. *Cold Spring Harb. Protoc.* (2018) doi:10.1101/pdb.prot082644.

60. Quang, D. & Xie, X. FactorNet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. *Methods* (2019) doi:10.1016/j.ymeth.2019.03.020.

61. Gilfillan, G. D. *et al.* Limitations and possibilities of low cell number ChIP-seq. *BMC Genomics* (2012) doi:10.1186/1471-2164-13-645.

62. Carroll, T. S., Liang, Z., Salama, R., Stark, R. & de Santiago, I. Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. *Front. Genet.* (2014) doi:10.3389/fgene.2014.00075.

63. Matys, V. *et al.* TRANSFAC®: Transcriptional regulation, from patterns to profiles. *Nucleic Acids Research* (2003) doi:10.1093/nar/gkg108.

64. Kulakovskiy, I. V. *et al.* HOCOMOCO: A comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Res.* (2013) doi:10.1093/nar/gks1089.

65. Wasserman, W. W. & Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics* (2004) doi:10.1038/nrg1315.

66. Kiesel, A. *et al.* The BaMM web server for de-novo motif discovery and regulatory sequence analysis. *Nucleic Acids Res.* (2018) doi:10.1093/nar/gky431.

67. GuhaThakurta, D. & Stormo, G. D. Identifying target sites for cooperatively binding factors. *Bioinformatics* (2001) doi:10.1093/bioinformatics/17.7.608.

68. Kulakovskiy, I. *et al.* From binding motifs in chip-seq data to improved models of transcription factor binding sites. in *Journal of Bioinformatics and Computational Biology* (2013). doi:10.1142/S0219720013400040.

69. Atherton, J. *et al.* A model for sequential evolution of ligands by exponential enrichment (SELEX) data. *Ann. Appl. Stat.* (2012) doi:10.1214/12-AOAS537.

70. Karimzadeh, M. & Hoffman, M. M. Virtual ChIP-seq: predicting transcription factor binding by learning from the transcriptome. *bioRxiv* (2019) doi:10.1101/168419.

71. Gotea, V. *et al.* Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Res.* (2010) doi:10.1101/gr.104471.109.

72. Yan, J. *et al.* XTranscription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell* (2013) doi:10.1016/j.cell.2013.07.034.

73. Pique-Regi, R. *et al.* Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* (2011) doi:10.1101/gr.112623.110.

74. Guo, W. L. & Huang, D. S. An efficient method to transcription factor binding sites imputation: Via simultaneous completion of multiple matrices with positional consistency. *Mol. Biosyst.* (2017) doi:10.1039/c7mb00155j.

75. Keilwagen, J., Posch, S. & Grau, J. Accurate prediction of cell type-specific transcription factor binding. *Genome Biol.* (2019) doi:10.1186/s13059-018-1614-y.

76. Chen, T. & Guestrin, C. XGBoost: A scalable tree boosting system. in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016). doi:10.1145/2939672.2939785.

77. Yang, S. Y., Ahmed, S., Satheesh, S. V. & Matthews, J. Genome-wide mapping and analysis of aryl hydrocarbon receptor (AHR)- and aryl hydrocarbon receptor repressor (AHRR)-binding sites in human breast cancer cells. *Arch. Toxicol.* (2018) doi:10.1007/s00204-017-2022-x.

78. Elith, J., Leathwick, J. R. & Hastie, T. A working guide to boosted regression trees. *Journal of Animal Ecology* (2008) doi:10.1111/j.1365-2656.2008.01390.x.

79. Gregorutti, B., Michel, B. & Saint-Pierre, P. Correlation and variable importance in random forests. *Stat. Comput.* (2017) doi:10.1007/s11222-016-9646-1.

80. Gorkin, D. U., Leung, D. & Ren, B. The 3D genome in transcriptional regulation and pluripotency. *Cell Stem Cell* (2014) doi:10.1016/j.stem.2014.05.017.

81. Denholtz, M. & Plath, K. Pluripotency in 3D: Genome organization in pluripotent cells. *Current Opinion in Cell Biology* (2012) doi:10.1016/j.ceb.2012.11.001.

82. Salomoni, P. & Pandolfi, P. P. Transcriptional regulation of cellular transformation. *Nature Medicine* (2000) doi:10.1038/77459.

83. Atlasi, Y. & Stunnenberg, H. G. The interplay of epigenetic marks during stem cell differentiation and development. *Nature Reviews Genetics* (2017) doi:10.1038/nrg.2017.57.

84. Li, Y., Hu, M. & Shen, Y. Gene regulation in the 3D genome. *Human Molecular Genetics* (2018) doi:10.1093/hmg/ddy164.

85. Zabidi, M. A. *et al.* Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* (2015) doi:10.1038/nature13994.

86. Ay, F. & Noble, W. S. Analysis methods for studying the 3D architecture of the genome. *Genome Biol.* **16**, 1–15 (2015).

87. Bonev, B. & Cavalli, G. Organization and function of the 3D genome. *Nature Reviews Genetics* (2016) doi:10.1038/nrg.2016.112.

88. Zheng, H. & Xie, W. The role of 3D genome organization in development and cell differentiation. *Nature Reviews Molecular Cell Biology* (2019) doi:10.1038/s41580-019-0132-4.

89. Mirabella, A. C., Foster, B. M. & Bartke, T. Chromatin deregulation in disease. *Chromosoma* (2016) doi:10.1007/s00412-015-0530-0.

90. Boltsis, I., Grosveld, F., Giraud, G. & Kolovos, P. Chromatin Conformation in Development and Disease. *Frontiers in Cell and Developmental Biology* (2021) doi:10.3389/fcell.2021.723859.

91. Kaiser, V. B. & Semple, C. A. When TADs go bad: Chromatin structure and nuclear organisation in human disease. *F1000Research* (2017) doi:10.12688/f1000research.10792.1.

92. Lupiáñez, D. G. *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* (2015) doi:10.1016/j.cell.2015.04.004.

93. Valton, A. L. & Dekker, J. TAD disruption as oncogenic driver. *Current Opinion in Genetics and Development* (2016) doi:10.1016/j.gde.2016.03.008.

94. Gröschel, S. *et al.* A single oncogenic enhancer rearrangement causes

concomitant EVI1 and GATA2 deregulation in Leukemia. *Cell* (2014) doi:10.1016/j.cell.2014.02.019.

95. Northcott, P. A. *et al.* Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* (2014) doi:10.1038/nature13379.

96. Lotta, L. A. *et al.* Integrative genomic analysis implicates limited peripheral adipose storage capacity in the pathogenesis of human insulin resistance. *Nat. Genet.* (2017) doi:10.1038/ng.3714.

97. Hnisz, D. *et al.* Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science (80-.).* (2016) doi:10.1126/science.aad9024.

98. Weischenfeldt, J. *et al.* Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. *Nat. Genet.* (2017) doi:10.1038/ng.3722.

99. Pucéat, M. Capturing chromosome conformation. in *Methods in Molecular Biology* (2021). doi:10.1007/978-1-0716-0664-3_1.

100. Zhao, Z. *et al.* Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat. Genet.* (2006) doi:10.1038/ng1891.

101. Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.* (2006) doi:10.1038/ng1896.

102. Belton, J. M. & Dekker, J. Chromosome conformation capture carbon copy (5C) in budding yeast. *Cold Spring Harb. Protoc.* (2015) doi:10.1101/pdb.prot085191.

103. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (80-.).* (2009) doi:10.1126/science.1181369.

104. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* (2015) doi:10.1038/ng.3286.

105. Wei, C. L. *et al.* A global map of p53 transcription-factor binding sites in the human genome. *Cell* (2006) doi:10.1016/j.cell.2005.10.043.

106. Liu, X. *et al.* In Situ Capture of Chromatin Interactions by Biotinylated dCas9. *Cell* (2017) doi:10.1016/j.cell.2017.08.003.

107. Morgan, S. L. *et al.* Manipulation of nuclear architecture through CRISPR-mediated chromosomal looping. *Nat. Commun.* (2017) doi:10.1038/ncomms15993.

108. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* (2014) doi:10.1016/j.cell.2014.11.021.

109. Boyle, S., Rodesch, M. J., Halvensleben, H. A., Jeddeloh, J. A. & Bickmore, W. A. Fluorescence in situ hybridization with high-complexity repeat-free oligonucleotide probes generated by massively parallel synthesis. *Chromosom. Res.* (2011) doi:10.1007/s10577-011-9245-0.

110. Kempfer, R. & Pombo, A. Methods for mapping 3D chromosome architecture. *Nature Reviews Genetics* (2020) doi:10.1038/s41576-019-0195-2.

111. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* (2012) doi:10.1038/nature11247.

112. Jung, I. *et al.* A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nat. Genet.* (2019) doi:10.1038/s41588-019-0494-8.

113. Li, X. *et al.* Long-read ChIA-PET for base-pair-resolution mapping of haplotype-specific chromatin interactions. *Nat. Protoc.* (2017) doi:10.1038/nprot.2017.012.

114. Li, G. *et al.* ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome Biol.* (2010) doi:10.1186/gb-2010-11-2-r22.

115. Yardımcı, G. G. *et al.* Measuring the reproducibility and quality of Hi-C data. *Genome Biol.* (2019) doi:10.1186/s13059-019-1658-7.

116. Smith, E. M., Lajoie, B. R., Jain, G. & Dekker, J. Invariant TAD Boundaries Constrain Cell-Type-Specific Looping Interactions between Promoters and Distal Elements around the CFTR Locus. *Am. J. Hum. Genet.* (2016) doi:10.1016/j.ajhg.2015.12.002.

117. Roy, S. *et al.* A predictive modeling approach for cell line-specific long-range regulatory interactions. *Nucleic Acids Res.* (2015) doi:10.1093/nar/gkv865.

118. Hait, T. A., Amar, D., Shamir, R. & Elkon, R. FOCS: A novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer-promoter map. *Genome Biol.* (2018) doi:10.1186/s13059-018-1432-2.

119. Gao, T. & Qian, J. Eagle: An algorithm that utilizes a small number of genomic features to predict tissue/ cell type-specific enhancer-gene interactions. *PLoS Comput. Biol.* (2019) doi:10.1371/journal.pcbi.1007436.

120. Cao, Q. *et al.* Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. *Nat. Genet.* (2017) doi:10.1038/ng.3950.

121. Fishilevich, S. *et al.* GeneHancer: Genome-wide integration of enhancers and

target genes in GeneCards. *Database* (2017) doi:10.1093/database/bax028.

122. Corradin, O. *et al.* Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res.* (2014) doi:10.1101/gr.164079.113.

123. He, B., Chen, C., Teng, L. & Tan, K. Global view of enhancer-promoter interactome in human cells. *Proc. Natl. Acad. Sci. U. S. A.* (2014) doi:10.1073/pnas.1320308111.

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

125. Moore, J. E., Pratt, H. E., Purcaro, M. J. & Weng, Z. A curated benchmark of enhancer-gene interactions for evaluating enhancer-target gene prediction methods. *Genome Biol.* (2020) doi:10.1186/s13059-019-1924-8.

126. Cao, F. & Fullwood, M. J. Inflated performance measures in enhancer–promoter interaction-prediction methods. *Nature Genetics* (2019) doi:10.1038/s41588-019-0434-7.

127. Whitaker, J. W., Nguyen, T. T., Zhu, Y., Wildberg, A. & Wang, W. Computational schemes for the prediction and annotation of enhancers from epigenomic assays. *Methods* (2015) doi:10.1016/j.ymeth.2014.10.008.

128. Nolis, I. K. *et al.* Transcription factors mediate long-range enhancer-promoter interactions. *Proc. Natl. Acad. Sci. U. S. A.* (2009) doi:10.1073/pnas.0902454106.

129. Maksimenko, O. & Georgiev, P. Mechanisms and proteins involved in long-distance interactions. *Frontiers in Genetics* (2014) doi:10.3389/fgene.2014.00028.

130. Dall'Agnese, A. *et al.* Transcription Factor-Directed Re-wiring of Chromatin Architecture for Somatic Cell Nuclear Reprogramming toward trans-Differentiation. *Mol. Cell* (2019) doi:10.1016/j.molcel.2019.07.036.

131. Stadhouders, R., Filion, G. J. & Graf, T. Transcription factors and 3D genome conformation in cell-fate decisions. *Nature* (2019) doi:10.1038/s41586-019-1182-7.

132. Arnould, C. *et al.* Loop extrusion as a mechanism for formation of DNA damage repair foci. *Nature* (2021) doi:10.1038/s41586-021-03193-z.

133. Fudenberg, G. *et al.* Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep.* (2016) doi:10.1016/j.celrep.2016.04.085.

134. Kim, Y., Shi, Z., Zhang, H., Finkelstein, I. J. & Yu, H. Human cohesin compacts DNA by loop extrusion. *Science (80-.).* (2019) doi:10.1126/science.aaz4475.

135. Ren, G. *et al.* CTCF-Mediated Enhancer-Promoter Interaction Is a Critical Regulator of Cell-to-Cell Variation of Gene Expression. *Mol. Cell* (2017) doi:10.1016/j.molcel.2017.08.026.

136. Guo, Y. *et al.* CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell* (2015) doi:10.1016/j.cell.2015.07.038.

137. Weintraub, A. S. *et al.* YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell* (2017) doi:10.1016/j.cell.2017.11.008.

138. Bailey, S. D. *et al.* ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. *Nat. Commun.* (2015) doi:10.1038/ncomms7186.

139. Wen, Z., Huang, Z. T., Zhang, R. & Peng, C. ZNF143 is a regulator of chromatin loop. *Cell Biol. Toxicol.* (2018) doi:10.1007/s10565-018-9443-z.

140. Ye, B. Y. *et al.* ZNF143 is involved in CTCF-mediated chromatin interactions by cooperation with cohesin and other partners. *Mol. Biol.* (2016) doi:10.1134/S0026893316030031.

141. Ye, B. *et al.* ZNF143 in Chromatin Looping and Gene Regulation. *Frontiers in Genetics* (2020) doi:10.3389/fgene.2020.00338.

142. Zhou, Q. *et al.* ZNF143 mediates CTCF-bound promoter–enhancer loops required for murine hematopoietic stem and progenitor cell function. *Nat. Commun.* (2021) doi:10.1038/s41467-020-20282-1.

143. Knight, P. A. & Ruiz, D. A fast algorithm for matrix balancing. *IMA J. Numer. Anal.* (2013) doi:10.1093/imanum/drs019.

144. Forcato, M. *et al.* Comparison of computational methods for Hi-C data analysis. *Nat. Methods* **14**, 679–685 (2017).

145. Harrow, J. *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome Biol.* (2006).

146. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* (2015) doi:10.1038/nature14248.

147. Ernst, J. & Kellis, M. ChromHMM: Automating chromatin-state discovery and characterization. *Nature Methods* (2012) doi:10.1038/nmeth.1906.

148. Feingold, E. A. *et al.* The ENCODE (ENCyclopedia of DNA Elements) Project. *Science* (2004) doi:10.1126/science.1105136.

149. Ji, H. *et al.* An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.* (2008) doi:10.1038/nbt.1505.

150. Szklarczyk, D. *et al.* STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* (2019) doi:10.1093/nar/gky1131.

151. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* (2012) doi:10.1038/nature11082.

152. Akdemir, K. C. *et al.* Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. *Nat. Genet.* (2020) doi:10.1038/s41588-019-0564-y.

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

154. Koehler, A. B. Journal of the American Statistical Association. *Int. J. Forecast.* (1995) doi:10.1016/0169-2070(95)90067-5.

155. Barzilai, J. & Borwein, J. M. Two-point step size gradient methods. *IMA J. Numer. Anal.* (1988) doi:10.1093/imanum/8.1.141.

156. Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* (2017) doi:10.1038/nature24277.

157. Wang, H., Huang, B. & Wang, J. Predict long-range enhancer regulation based on protein-protein interactions between transcription factors. *Nucleic Acids Res.* (2021) doi:10.1093/nar/gkab841.

158. Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.* (2014) doi:10.1093/nar/gkt1249.

159. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* (2010) doi:10.1093/bioinformatics/btq033.

160. Battle, A. *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* (2014) doi:10.1101/gr.155192.113.

161. Nica, A. C. *et al.* The architecture of gene regulatory variation across multiple human tissues: The muTHER study. *PLoS Genet.* (2011) doi:10.1371/journal.pgen.1002003.

162. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B* (1996) doi:10.1111/j.2517-6161.1996.tb02080.x.

APPENDIX A

SUPPLEMENTARY FIGURES FOR CHAPTER 2

Table A.1. List of TFs in K562 cell line

ID	TF	FRiP	Motif
ENCFF002DBD	CTCF	35.5666667	MA0139.1
ENCFF926FUM	YY1	11.395	MA0095.2
ENCFF664XPS	SPI1	8.045	MA0080.5
ENCFF495MHZ	NFE2	7.095	MA0841.1
ENCFF465JKF	MYC	9.53	MA0147.3
ENCFF002CVW	FOS	6.25333333	MA0476.1
ENCFF497XOD	MITF	8.69	MA0620.3
ENCFF178MOP	SMAD5	7.55	MA1557.1
ENCFF182QDI	EGR1	13.28	MA0162.4
ENCFF958KNK	ATF3	14.1	MA0605.2
ENCFF010UHD	RFX1	12.235	MA0509.2
ENCFF948TXN	CREM	12.2	MA0609.2
ENCFF973LDQ	ZNF24	6.835	MA1124.1
ENCFF544XKC	PKNOX1	23.15	MA0782.2
ENCFF334FMW	USF1	10.75	MA0093.3
ENCFF895QLA	REST	11.385	MA0138.2
ENCFF710IEF	ATF4	12.3	MA0833.2

Table A.1 (cont'd)

ENCFF917COW	E2F6	15.925	MA0471.2
ENCFF833FCO	GABPA	15.25	MA0062.3
ENCFF422NGZ	MAX	24.15	MA0058.3
ENCFF564WEB	ELF1	16.3	MA0473.3
ENCFF492GXZ	FOXK2	5.31	MA1103.2
ENCFF706ISJ	ZBTB7A	10.21	MA0750.2
ENCFF059ONJ	MNT	9.9	MA0825.1
ENCFF968JVX	IKZF1	15.85	MA1508.1
ENCFF213EPU	ESRRA	6.62	MA0592.3
ENCFF493ABN	NRF1	38.9	MA0506.1
ENCFF106DAY	E2F1	10.375	MA0024.3
ENCFF502KHR	ELF4	5.43	MA0641.1
ENCFF670ZCR	MGA	6.805	MA0801.1
ENCFF664ZGR	NR2C1	5.91	MA1535.1
ENCFF558DSF	HMBOX1	7.31333333	MA0895.1
ENCFF209MQX	TEAD4	9.16	MA0809.2
ENCFF370ENX	NFIC	10.585	MA0161.2
ENCFF255EOB	NR2F2	5.02	MA1111.1
ENCFF613RNG	MEIS2	10.8	MA0774.1
ENCFF968KBN	ATF2	12.265	MA1632.1
ENCFF408FQC	ZBTB33	12.35	MA0527.1
ENCFF175IIE	NR2F1	10.36	MA0017.2

Table A.1 (cont'd)

ENCFF868QLL	ATF7	10.555	MA0834.1
ENCFF308IXJ	MAFF	5.19	MA0495.3
ENCFF337DKJ	JUND	11.4	MA0491.2
ENCFF114IWY	ZNF143	6.595	MA0088.2
ENCFF113PMT	NFYB	13.65	MA0502.2
ENCFF179NDS	BHLHE40	5.9	MA0464.2
ENCFF429XKT	CEBPB	10.855	MA0466.2

Table A.2. List of TFs in GM12878 cell line

ID	TF	FRiP	Motif
ENCFF002DAJ	CTCF	28.9333333	MA0139.1
ENCFF967ACD	YY1	14.6	MA0095.2
ENCFF911BYP	MEF2B	9.845	MA0660.1
ENCFF405NFV	TBX21	11.21	MA0690.1
ENCFF810CEL	NR2F1	8.62	MA0017.2
ENCFF141SAU	ZNF143	12.4	MA0088.2
ENCFF095GMM	BHLHE40	19.7	MA0464.2
ENCFF726VEK	ATF7	11.85	MA0834.1
ENCFF628QJU	NFIC	10.41	MA0161.2
ENCFF339KUO	PAX5	6.28	MA0014.3

Table A.2 (cont'd)

ENCFF456FQB	RELB	10.385	MA1117.1
ENCFF609KOY	ATF2	7.045	MA1632.1
ENCFF593FBF	BATF	11.54	MA1634.1
ENCFF394DLH	EBF1	14.1	MA0154.4
ENCFF807AKG	ELF1	8.64	MA0473.3
ENCFF939TZS	JUNB	7.935	MA0490.2
ENCFF328QLX	MEF2A	7.16	MA0052.4
ENCFF467NRS	NFYB	8.26	MA0502.2
ENCFF248QFF	RUNX3	21.1	MA0684.2

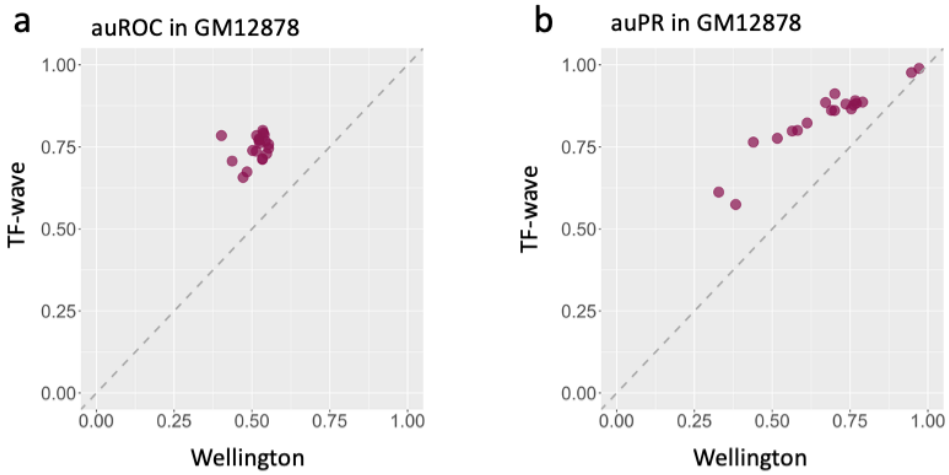

Figure A.1. TF-wave predicts TF binding sites accurately and outperforms Wellington in GM12878. TF-wave and Wellington were evaluated on the same datasets using the averaged performance of 5-fold cross-validation. (a) Performance comparison in GM12878 for TF-Wave and Wellington using auPR. The x-axis and y-axis are auPR of applying Wellington and TF-wave to predict TFBS on the same data. Each point represents a TF, points above the diagonal line indicate TF-wave performs better than Wellington. (f) Performance comparison in GM12878 for TF-Wave and Wellington using auROC. The x-axis and y-axis are auROC of applying Wellington and TF-wave to predict TFBS on the same data. Each point represents a TF, points above the diagonal line indicate TF-wave performs better than Wellington.

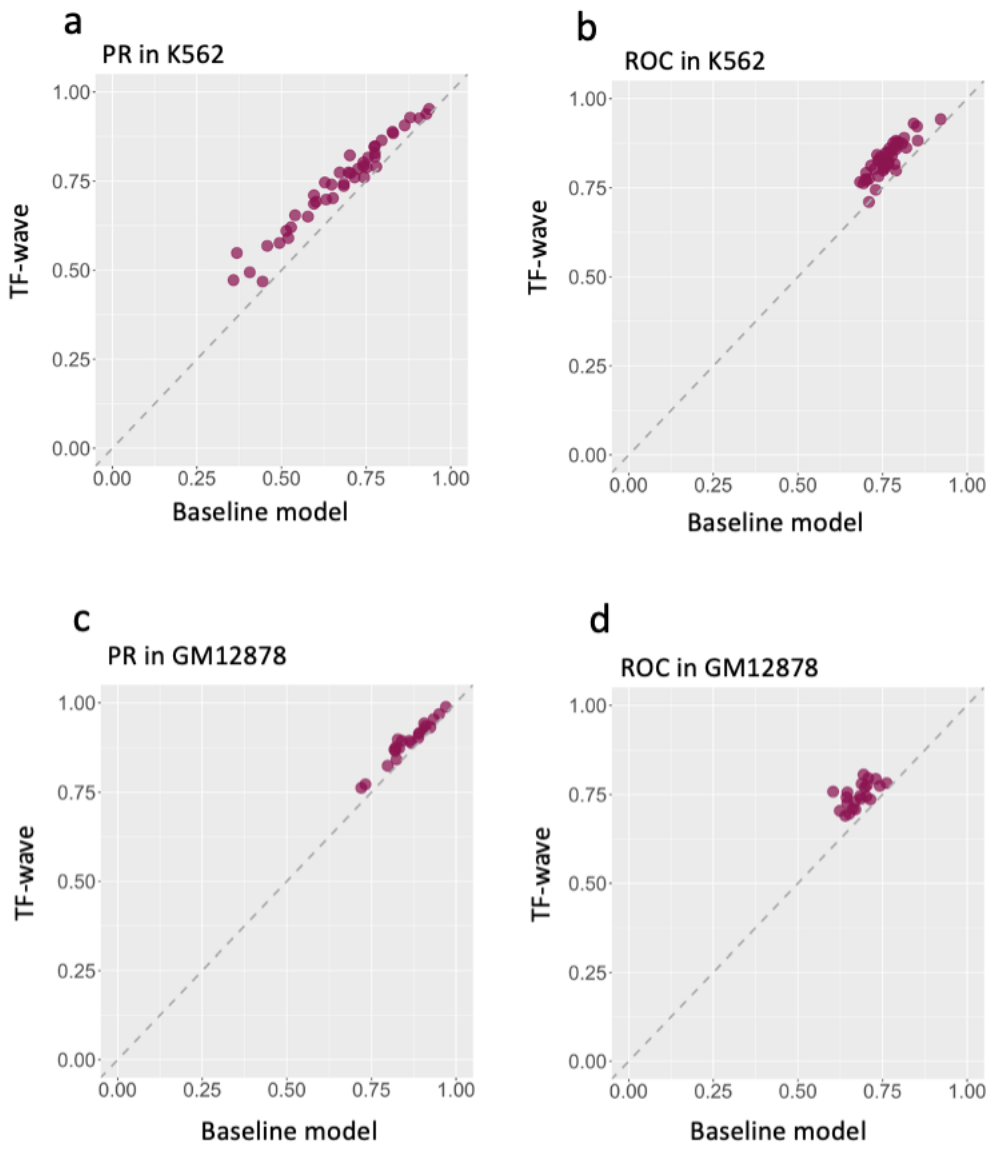

Figure A.2. Wavelet transform can boot the TFBS prediction performance in K562 and GM12878. TF-wave and GBT trained on the original DNase-Seq (baseline model) signal were evaluated on the same datasets using the averaged performance of 5-fold cross-validation. (a, c) Performance comparison in K562 GM12878 for TF-Wave and baseline model using auPR. The x-axis and y-axis are auPR of applying baseline model and TF-wave to predict TFBSs on the same data. Each point represents a TF, points above the diagonal line indicate TF-wave performs better than baseline model. (b, d) Performance comparison in K562 and GM12878 for TF-Wave and baseline model using auROC. The x-axis and y-axis are auROC of applying Wellington and TF-wave to predict TFBS on the same data. Each point represents a TF, points above the diagonal line indicate TF-wave performs better than baseline model.

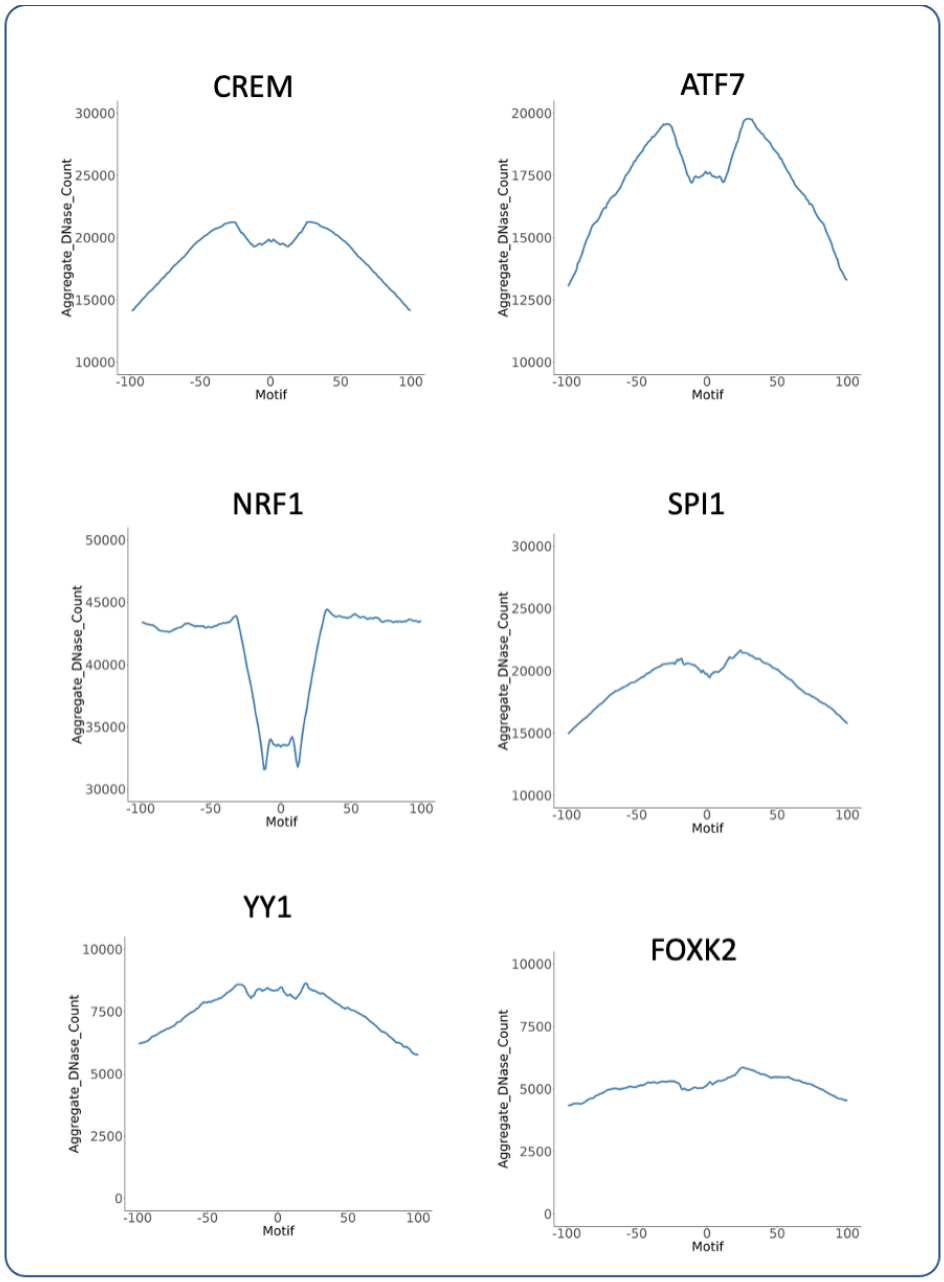

Figure A.3. Different TFs have different consensus footprint shapes. Aggregated DNase-Seq read-depth for top 5% predictions for NRF2, ATF7, YY1, SPI1, CREM, and FOXK2. Each TF has a unique footprint shape. Palindrome motifs such as CREM and ATF7 leave a symmetric footprint. NRF1 is similar to palindrome motifs, and also leaves a symmetric footprint. Motifs that are not palindrome such as SPI1, YY1, FOXK2 leaves asymmetric footprints.

APPENDIX B

SUPPLEMENTARY FIGURES FOR CHAPTER 3

Motif	Binding Probability
ACGCTGGGCGTGCAGATGC	0.17791858
CCGGCTCGCGTGCGCCGGC	0.6637833
CTAGCTTGCGTGCGCCGGC	0.5893966
AGGCGTTGCGTGAGAAGGA	0.82652086
GCGCGCGGCGTGGGGTTGG	0.15912299
TAGGTCTGCGTGTGGCTTC	0.6745489
TGTATTTGCGTGCCTAGCT	0.87979823
CCCCCTCGCGTGACTGCGA	0.4820609
GCCACAGGCGTGGACCGAA	0.15255722
ATTACAGGCGTGGGCCACC	0.2313509

Figure B.1. Ten putative AHR binding sites and their predicted binding probabilities

APPENDIX C

SUPPLEMENTARY MATERIAL FOR CHAPTER 4

Table C. 1. List of TFs used in K562 cell line

ID	TF
ENCFF002DBD	CTCF
ENCFF926FUM	YY1
ENCFF254YOX	NR2C2
ENCFF002CXU	RAD21
ENCFF664XPS	SPI1
ENCFF911VSD	NFE2
ENCFF465JKF	MYC
ENCFF002CVW	FOS
ENCFF947KPB	POLR2A
ENCFF497XOD	MITF
ENCFF178MOP	SMAD5
ENCFF076IFG	ZEB2
ENCFF701MXF	TFDP1
ENCFF116DIO	ZNF148
ENCFF182QDI	EGR1
ENCFF958KNK	ATF3
ENCFF490VVG	HDAC2
ENCFF644WLI	RNF2

Table C.1 (cont'd)

ENCFF948TXN	CREM
ENCFF232KAH	ATF1
ENCFF973LDQ	ZNF24
ENCFF511QHY	ZMYM3
ENCFF284LRP	TAL1
ENCFF440JJW	MLLT1
ENCFF822FWM	TAF1
ENCFF122QSN	HNRNPLL
ENCFF544XKC	PKNOX1
ENCFF294HEI	PRDM10
ENCFF334FMW	USF1
ENCFF895QLA	REST
ENCFF517KRT	GABPB1
ENCFF710IEF	ATF4
ENCFF622RBW	SMARCE1
ENCFF417LEJ	ZBTB40
ENCFF917COW	E2F6
ENCFF788EBS	ZNF316
ENCFF320THV	NR2C1
ENCFF685KAG	GABPA
ENCFF493TZM	TCF12
ENCFF186QUP	VEZF1

Table C.1 (cont'd)

ENCFF248AOD	ZNF766
ENCFF152VMJ	CTBP1
ENCFF422NGZ	MAX
ENCFF823SYE	DPF2
ENCFF692PVV	FOSL1
ENCFF076MSV	AFF1
ENCFF273TYA	ZNF592
ENCFF168HAG	PBX2
ENCFF273EYJ	POLR2G
ENCFF564WEB	ELF1
ENCFF366UBB	NCOR1
ENCFF492GXZ	FOXK2
ENCFF872JJJ	POLR2B
ENCFF706ISJ	ZBTB7A
ENCFF462JFW	ZBTB11
ENCFF996ZGL	C11orf30
ENCFF417TXD	USF2
ENCFF738WCE	KDM1A
ENCFF019ALY	HDAC1
ENCFF059ONJ	MNT
ENCFF968JVX	IKZF1
ENCFF801YZV	POLR2H

Table C.1 (cont'd)

ENCFF429XKT	CEBPB
ENCFF388YOB	ZNF395
ENCFF213EPU	ESRRA
ENCFF606CCB	CEBPG
ENCFF493ABN	NRF1
ENCFF120WOF	ZNF639
ENCFF096XMD	RAD51
ENCFF314ULQ	L3MBTL2
ENCFF106DAY	E2F1
ENCFF883TOD	SMARCA4
ENCFF602AXP	ZNF589
ENCFF502KHR	ELF4
ENCFF157UUF	NFRKB
ENCFF993GXU	CBFA2T3
ENCFF642BNC	CBFA2T2
ENCFF356ASJ	E2F5
ENCFF670ZCR	MGA
ENCFF083YCQ	E4F1
ENCFF881QBT	PML
ENCFF558DSF	HMBOX1
ENCFF209MQX	TEAD4
ENCFF521CRG	ZFP36

Table C.1 (cont'd)

ENCFF241TBP	SOX6
ENCFF680WBN	RBM25
ENCFF370ENX	NFIC
ENCFF057ZUY	BCOR
ENCFF925ANU	EP400
ENCFF255EOB	NR2F2
ENCFF927JBT	MAFG
ENCFF225MPC	ARID1B
ENCFF613RNG	MEIS2
ENCFF968KBN	ATF2
ENCFF408FQC	ZBTB33
ENCFF484HCG	SUPT5H
ENCFF646VQW	TRIM24
ENCFF350YXB	MTA3
ENCFF433RKB	ZFX
ENCFF175IIE	NR2F1
ENCFF868QLL	ATF7
ENCFF002CVU	CCNT2
ENCFF626KTJ	PHF8
ENCFF002CXW	POLR3A
ENCFF379MPS	RBBP5
ENCFF474NLG	HCFC1

Table C.1 (cont'd)

ENCFF671DOL	UBTF
ENCFF549TYR	EP300
ENCFF308IXJ	MAFF
ENCFF337DKJ	JUND
ENCFF114IWY	ZNF143
ENCFF113PMT	NFYB
ENCFF179NDS	BHLHE40
ENCFF041YQC	SMC3
ENCFF380FJL	TBP

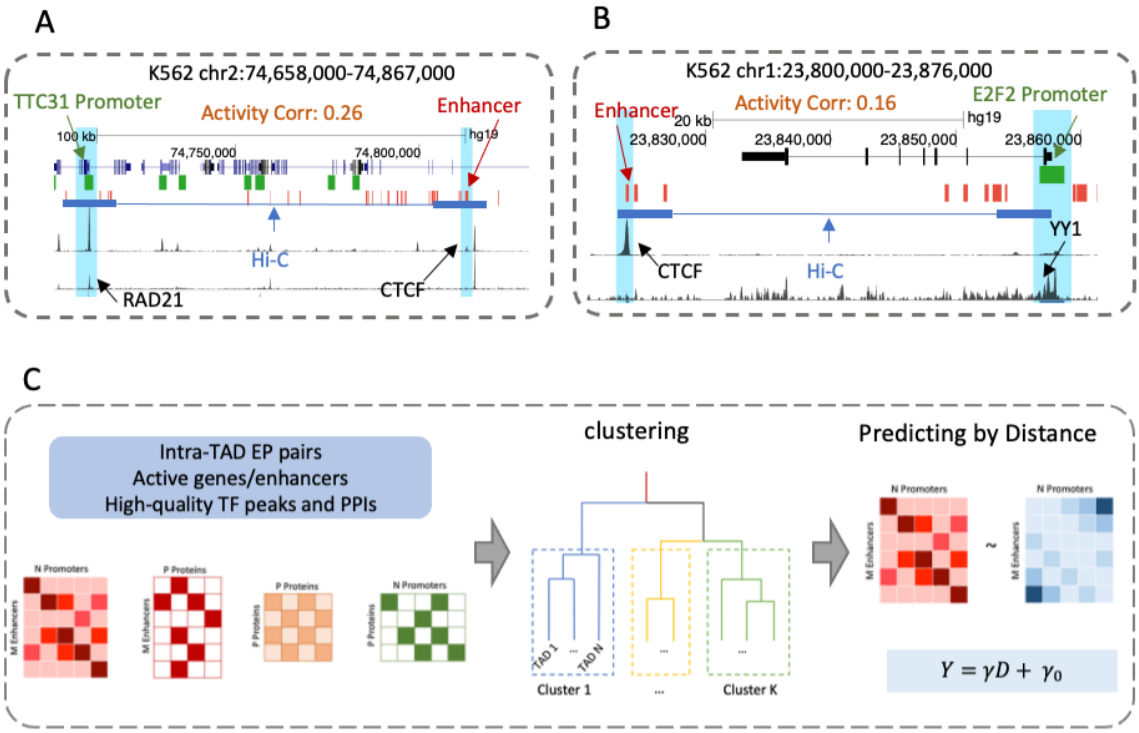

Figure C.1. A,B Examples of first-order PPI mediated long-range enhancer promoter interactions in K562 cell line. A, RAD21 binds on TTC31 promoter, CTCF binds on enhancer, the PPI between RAD21 and CTCF regulates the long-range enhancer-promoter interactions. B, YY1 binds on the E2F2 promoter, CTCF binds on the

Figure C.1 (cont'd)

enhancer, the PPI between YY1 and CTCF regulates this long-range enhancer-promoter interactions. C, matrices generation within each TAD, clustering of TAD based on the TF binding on enhancers and promoters within each TAD. A linear model was fitted between the enhancer-promoter interaction frequency and genomic distance.

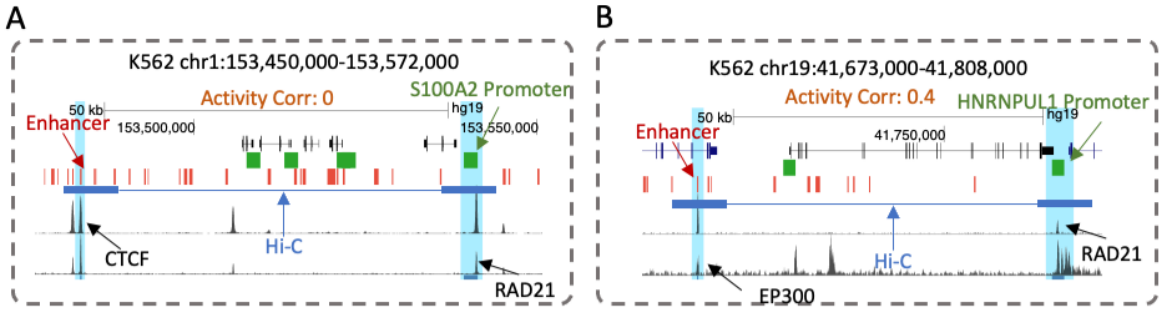

Figure C.2. Prioritized first-order PPIs can predict long-range enhancer-promoters accurately. A, RAD21 binds on S100A2 promoter, CTCF binds on enhancer, the prioritized PPI between RAD21 and CTCF regulate this long-range enhancer-promoter interactions. B, RAD21 binds on HNRNPUL1 promoter, RAD21 binds on enhancer, the prioritized PPI between EP300 and RAD21 regulate this long-range enhancer-promoter interactions.

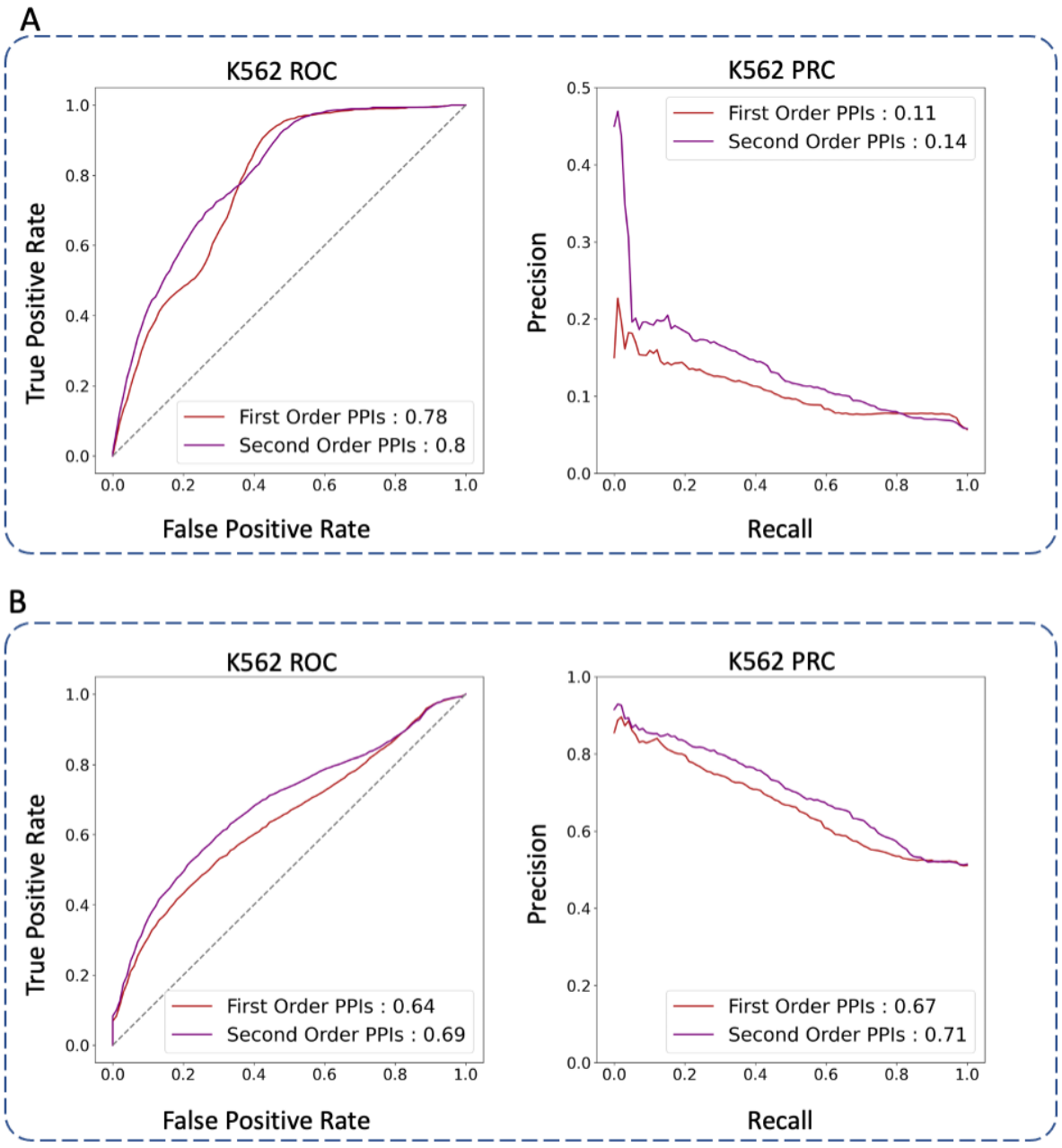

Figure C.3. Performance comparison between first-order PPI model and second-order PPI model. A. model comparison on the dataset used for second-order PPI model. B. model comparison on balanced-distance controlled dataset used for second-order PPI model

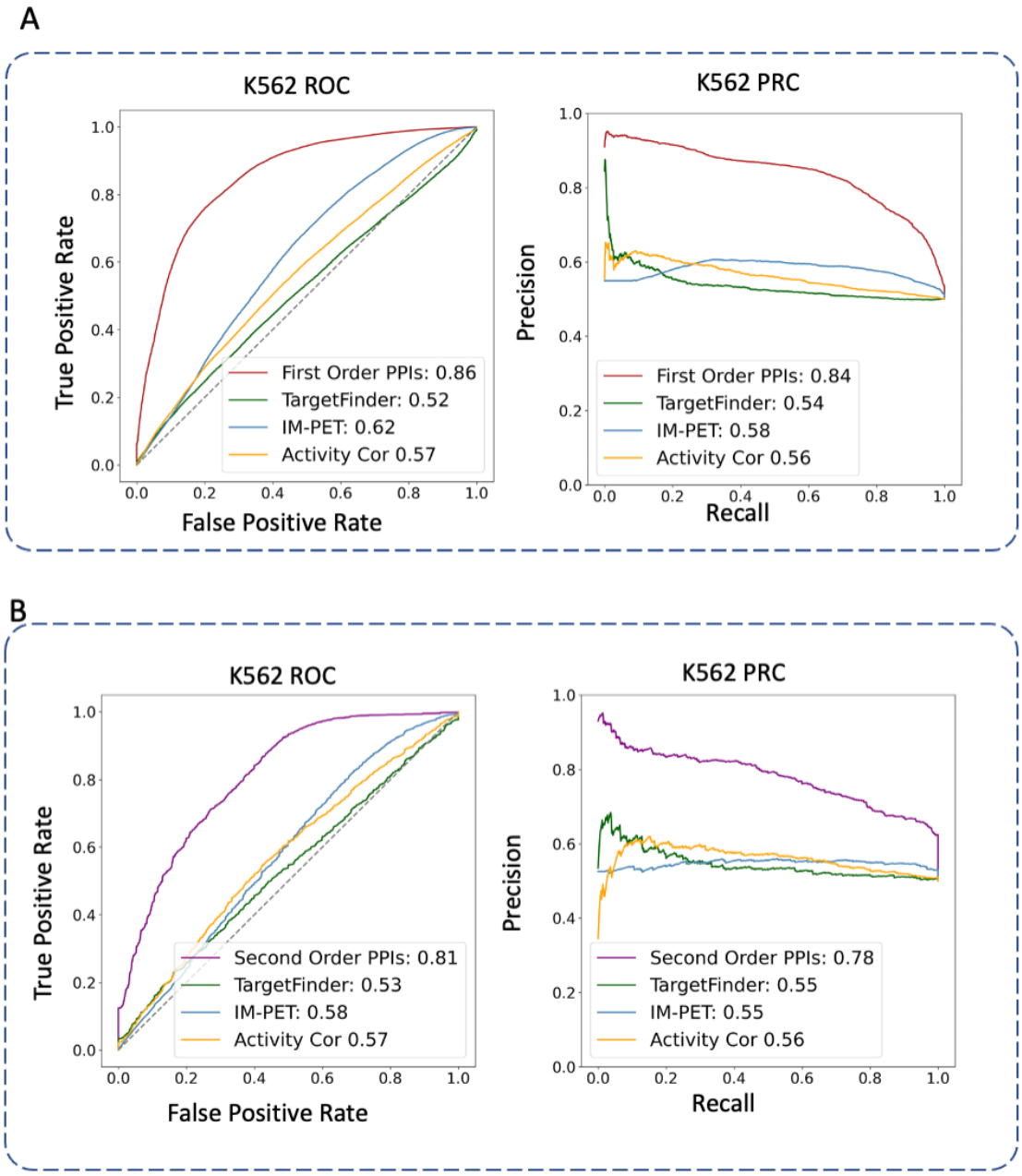

Figure C.4. Performance comparison in K562 cells on balanced data. EP^{3}ICO, ProTECT, TargetFinder, and IM-PET are applied on the same input datasets and are evaluated based on the averaged performance of 5-fold cross-validation on a balanced data. As a baseline comparison, enhancer-gene activity correlations are also included in the analysis. A ROC curves and PR curves of first-order PPIs performance in K562 cells. B ROC curves PR curves of second-order PPIs performance in K562 cells.

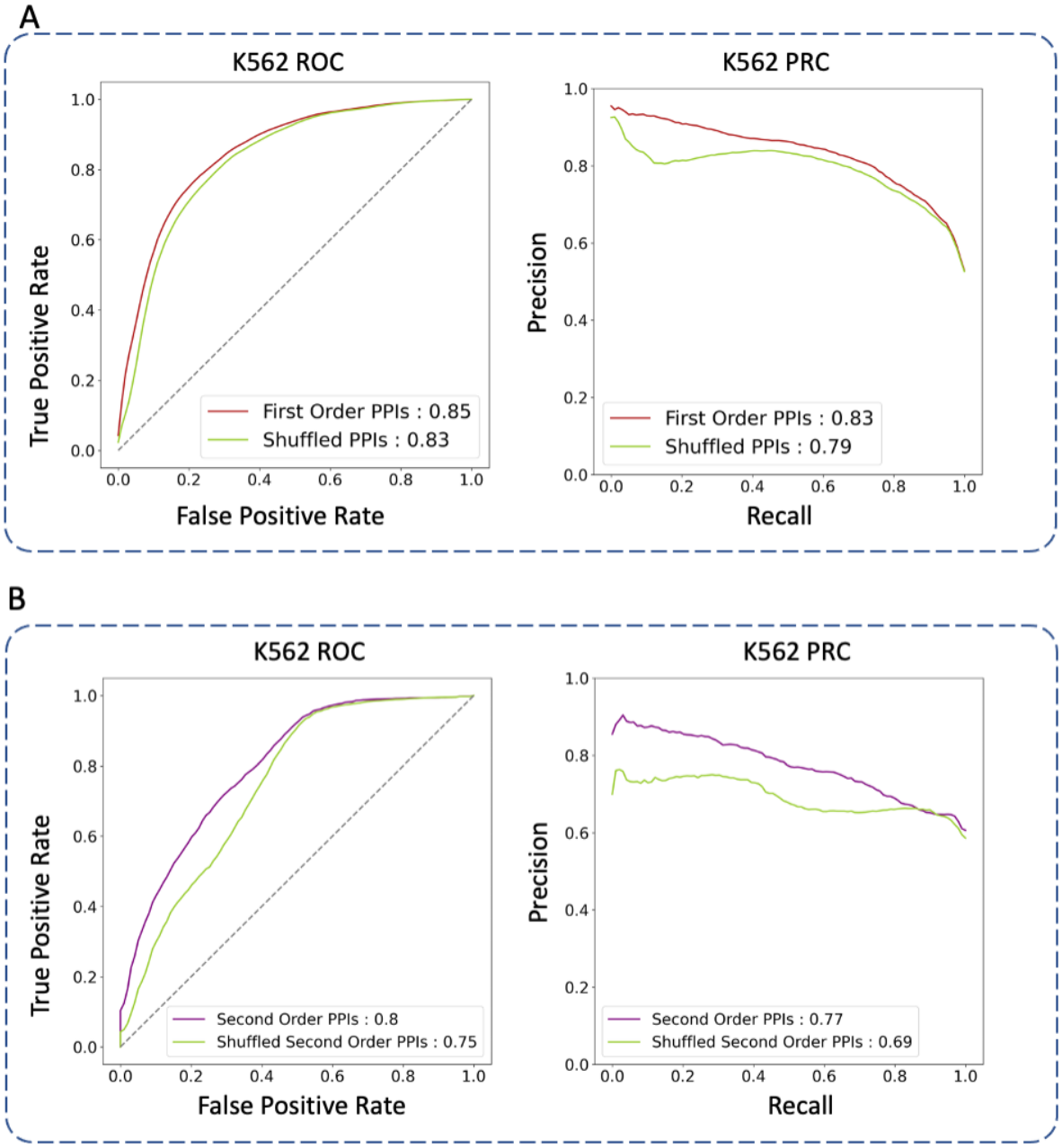

Figure C.5. Comparison with shuffled PPIs on balanced datasets. A. ROC, and PR plots for the models based on the original first-order TF PPI features (red), the models based on the shuffled first-order TF PPI features (green). B. ROC and PR plots for the models based on the original second-order TF PPI features (purple), the models based on the shuffled second-order TF PPI features (green).

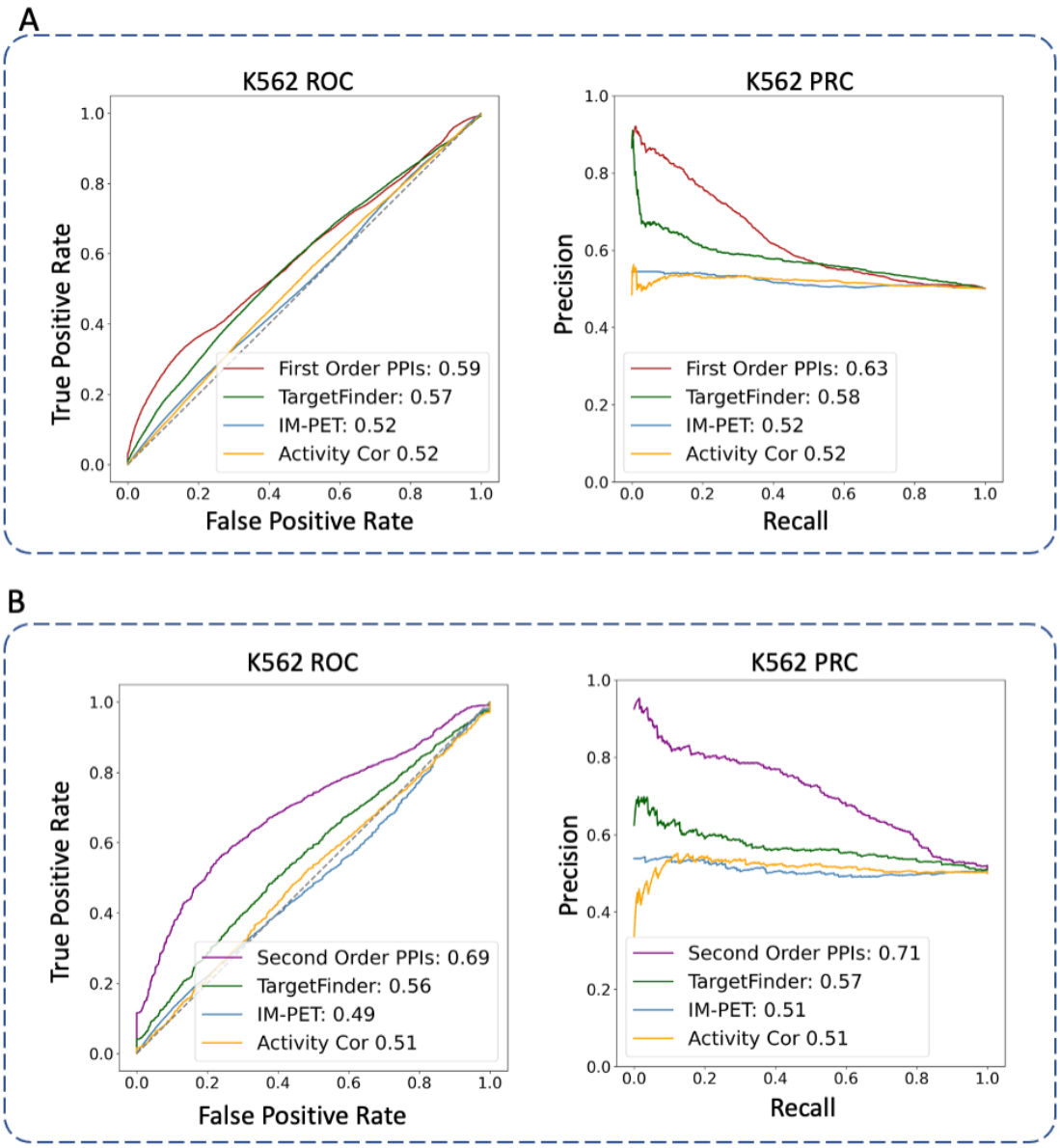

Figure C.6. Performance comparison in K562 cells on balanced data with distance control. EP^{3}ICO, ProTECT, TargetFinder, and IM-PET are applied on the same input datasets and are evaluated based on the averaged performance of 5-fold cross-validation on a balanced data with distance control. As a baseline comparison, enhancer-gene activity correlations are also included in the analysis. A ROC curves and PR curves of first-order PPIs performance in K562 cells. B ROC curves PR curves of second-order PPIs performance in K562 cells.

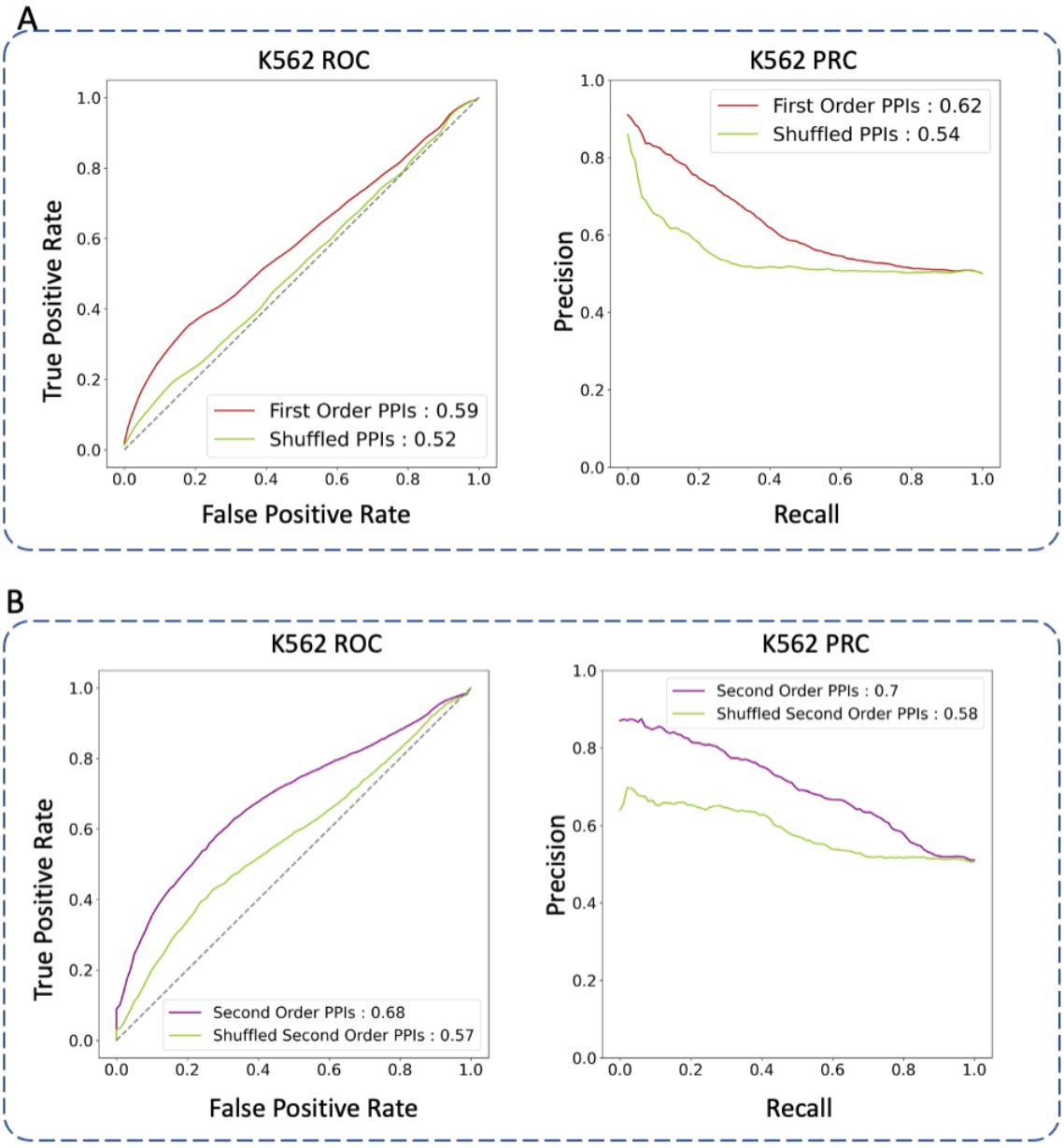

Figure C.7. Comparison with shuffled PPIs on balanced datasets with distance control. A. ROC, and PR plots for the models based on the original first-order TF PPI features (red), the models based on the shuffled first-order TF PPI features (green). B. ROC and PR plots for the models based on the original second-order TF PPI features (purple), the models based on the shuffled second-order TF PPI features (green).

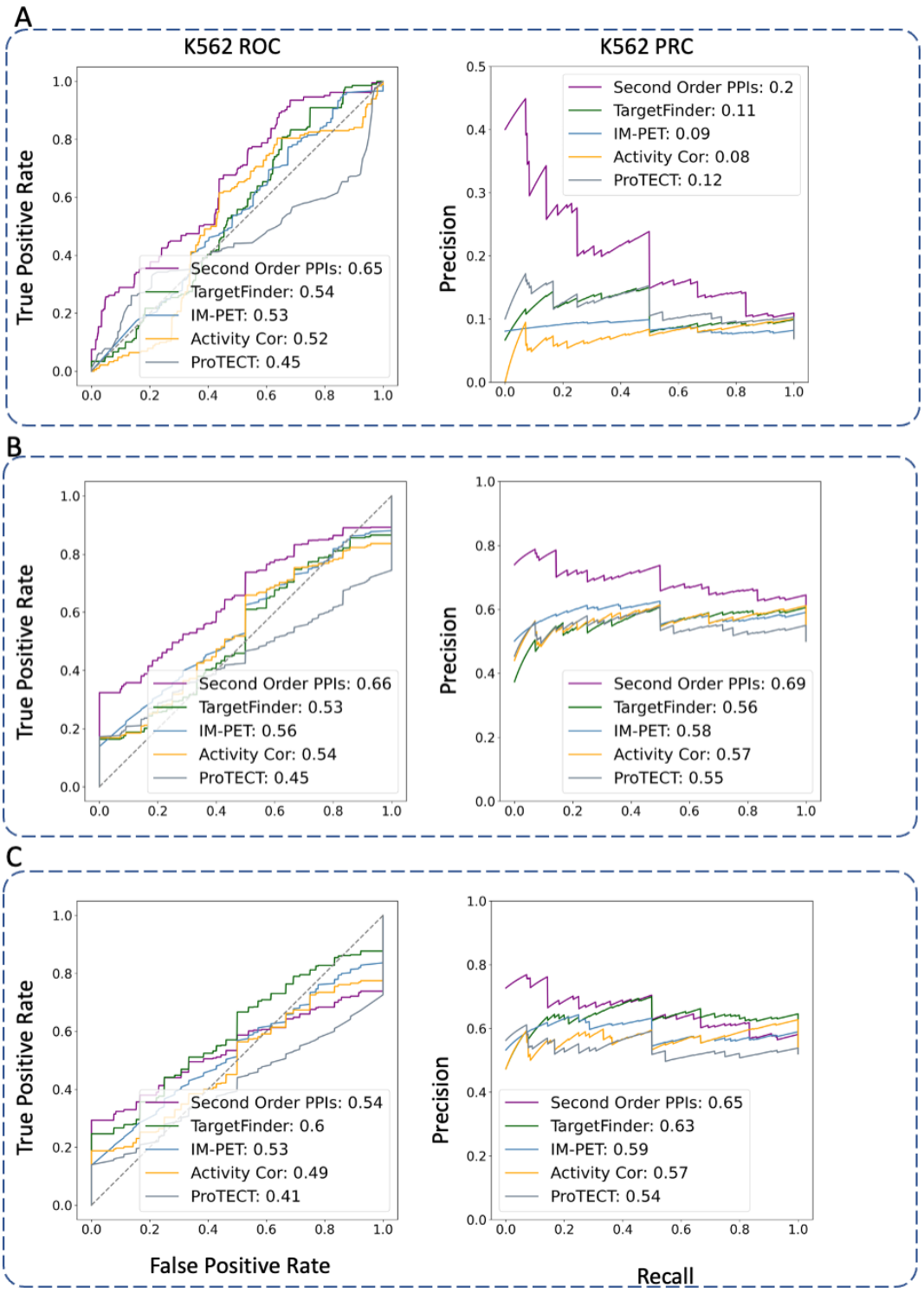

Figure C.8. Comparison with ProTECT, TargetFinder, IM-PET, and baseline model based on a common dataset. A. ROC, and PR plots for the models on the common dataset. B. ROC and PR plots for the models on the balanced dataset. C. ROC and PR plots for the models on the balanced dataset with distance control.