UNIVERSITY OF CALIFORNIA

Los Angeles

Detection of Transcription Factor Co-Binding Patterns in Human Cells via Point Process Models

A thesis submitted in partial satisfaction

of the requirements for the degree Master of Science

in Statistics

by

Yuan Tian

2014

ABSTRACT OF THE THESIS

Detection of Transcription Factor Co-Binding Patterns

in Human Cells via Point Process Models

by

Yuan Tian

Master of Science in Statistics

University of California, Los Angeles, 2014

Professor Qing Zhou, Chair

Transcription factors usually work synergistically to regulate target gene expression. Characterizing their combinatorial binding patterns will provide key step towards elucidating the underlying gene regulatory mechanism. Accordingly, the goal of this thesis is to apply a newly designed test statistic based on inhomogeneous Poisson process and Ripley's K-function to investigate pairwise transcription factor binding patterns using chromatin immunoprecipitation sequencing (ChIP-seq) data. We applied the method to 21 selected transcription factors using ChIP-seq data from two different human cell types. Significant clustering patterns have been detected between most transcription factor pairs, and their optimal binding distances are reported. More interestingly, by comparing the two cell types, we identify tissue-specific co-binding patterns, which implicate tissue-specific transcriptional regulation. In summary, the presented work has demonstrated the development and utility of the designed test statistics on evaluating transcription factor binding patterns, which can help to infer transcription factor interactions on gene expression regulation.

The thesis of Yuan Tian is approved.

Rick Paik Schoenberg

Jingyi Jessica Li

Qing Zhou, Committee Chair

University of California, Los Angeles

2014

To my beloved family

for their unconditional love and endless support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SUPPLEMENTARY MATERIALS

1. The most significant distance t* in hESC for assessed TF pair

Chapter 1

Introduction

Genes to be functional needs first to be transcribed into RNA transcript, and this process is strictly controlled. Any dysregulation in the expression regulation process may lead to dysfunction of the downstream molecular pathways, which may further result in disease. Therefore, it is important to understand how gene expression is regulated and how they are related to different biological circumstances. Many factors have been involved in transcriptional regulation and one of the most important regulators are transcription factors (TFs), which are proteins that binds to specific DNA sequences to modulate transcriptional activity. Those specific DNA sequences are called TF binding sites. For a particular TF, its binding sites usually share similar sequence patterns, which

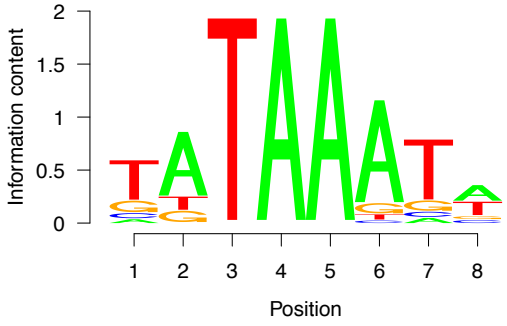

Figure 1.1: Example PWM for TATA box binding protein (TBP). PWM: a model for a fixed length sequence that specifies the probability of each nucleotide at each position. The height is correlated with the occurrence probability of the corresponding nucleotide in the transcription factor binding regions.

can be described by a position weight matrix (PWM)[1, 2](Figure 1.1). Moreover, TFs do not work along. They interact with each other remotely, or they can co-bind to DNA

sequence as a complex. Therefore, interacted TFs often show complex binding patterns among their binding sites in the target DNA sequences. Characterizing these combinatorial binding patterns will provide the key step towards elucidating the behind gene regulatory network (Figure 1.2).

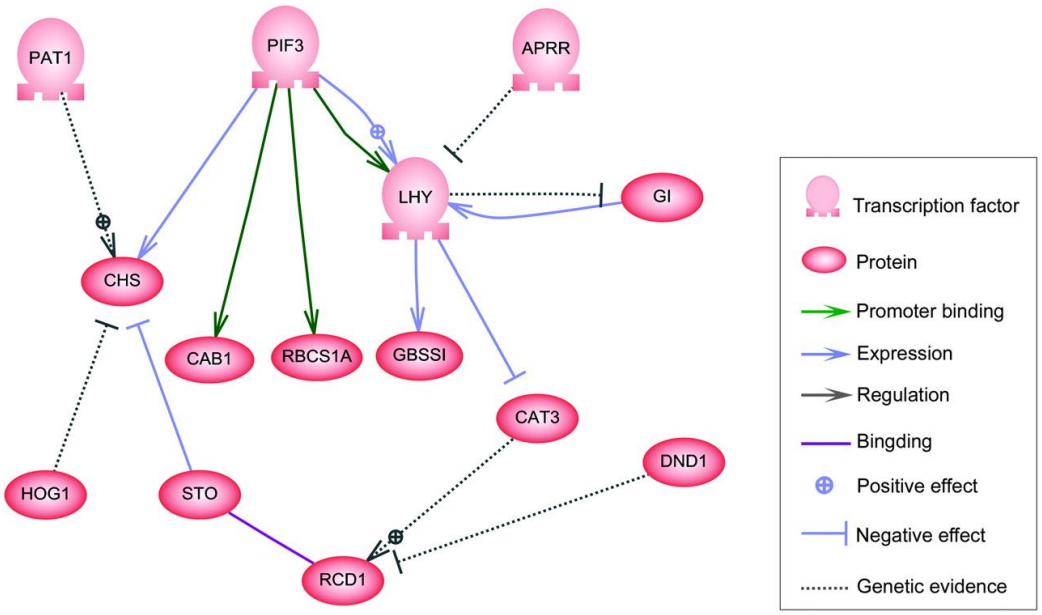

Figure 1.2: Example of a regulatory network built in Pathway Studio (Figure adapted from Song *et al*. 2010[3]).

Many methods have been proposed to identify binding motifs in DNA sequences, and therefore to elucidate the behind gene regulatory network. Those methods are essentially statistical testing on sequence matches by scanning the PWM against candidate gene promoter regions [4-6]. Furthermore, it is known that one TF usually regulate a set of target genes that are functionally related. Therefore, the methods have been further improved to evaluate the overrepresentation significance of a particular TF binding sites among a set of related gene promoter sequences [7-10]. One major advantage of these motif-based approaches is that only genomic DNA sequence and TF PWMs of interest are required for the analysis. However, as the TF binding sites are widespread in

genomes, occurrences of motifs along are not sufficient to predict real TF binding. In fact, most of motifs are not accessible to TFs by covering with various genomic and epigenomic modifications. Moreover, these motif search methods usually based on single TF, and are thereafter limited to infer the cooperative interactions among multiple TFs.

ChIP-seq is a molecular experimental method developed to analyze protein interactions with DNA. By combining chromatin immunoprecipitation (ChIP) with next generation DNA sequencing, it has been widely used to identify the binding sites of DNA-associated proteins, particularly TFs. While performing the experiments using specific tissues or cell types, ChIP-seq provides the unique opportunity to determine condition-based TF binding. More importantly, we can investigate the relationships among TFs by evaluating their ChIP-seq data in the same condition. Recent studies have used *ad hoc* methods to define and detect clusters of binding motifs using ChIP-seq data. With different statistical tests, including Poisson approximation and Fisher's exact test, people find several TFs with binding distance within 100bp [11-14]. However, current methods are usually performed on a small set of TFs, and the distance thresholds are arbitrarily chosen.

Choi and Zhou (2014) have recently developed a statistical method to detect pairwise TF binding patterns from ChIP-seq data [15]. They utilize the inhomogeneous Poisson process to estimate binding site distributions, and apply Ripley's K-function [16, 17] to summarize the binding patterns over multiple genes. Instead of using an arbitrary picked distance for statistical test, this method can evaluate various distance values and report the most significant one, which can be further used to infer the structures of the

corresponding TF complex. In their paper, they have applied this method on 14 TFs in mouse embryonic stem cells (ESC) with 91 different pairwise TF combinations. They have detected the clustering binding patterns between most of their analyzed TF pairs.

In this thesis, I aim to apply this method to human ChIP-seq data, and extend the analysis to 21 different TFs. In addition to analyzing the TF co-binding patterns within one context, I also compare the patterns between two different cell types, human ESCs and human lymphoblast cell lines, to investigate tissue-specific transcriptional regulation. The presented systematic analysis hare can provide the key step towards understanding the complex transcription regulation mechanism.

Chapter 2

Methods

2.1 ChIP-seq peak processing

The method is developed for identification of co-binding patterns using ChIP-seq data of selected pairwise TFs, as described in Choi *et al.* 2014[15]. Reported binding peaks in bed format are downloaded from Encyclopedia of DNA Elements (ENCODE) Consortium. Here, we defined the upstream (-8K, 2K] from the transcription start site (TSS) of every single gene as the candidate TF binding regions (promoter). For simplification, we regard the -8K end as the origin and choose the direction of gene transcription as the positive direction, so that an "promoter" region can be represented as (0, L] (L=10K). ChIP-seq peaks overlapped with those regions are assigned to the corresponding gene. In other words, genes are considered as targets of the analyzed TF if their "promoter" region overlaps with its ChIP-seq peaks. For every picked TF pairs, we only analyze the regions where both TF binding sites occur. Additionally, usually ChIP-seq peaks are around hundreds of base pairs long. To improve the resolution of the study, we further utilize the cisgneome (http://www.biostat.jhsph.edu/~hji/cisgenome/) tools [18] to scan the known TF binding motifs (PWM), against the peak region to find the exact match position. Then we summarize the detected location by the coordinates of its middle point to represent the TF binding site.

2.2 Inhomogeneous Poisson Process

Section 2.2 and 2.4 review relevant materials from Choi and Zhou (2014).

Let us consider a pair of TFs of X and Y, which share N promoter regions, denoted as $R_1, \dots, R_N$. Correspondingly, their binding sites in the region R_r are $x_r = (x_{r,i}, i = 1, \dots, n_r)$ and $y_r = (y_{r,i}, i = 1, \dots, m_r)$ for $r = 1, \dots, N$, respectively. Here, we assume the same model for every region R_r, thus the index r is suppressed in the follow-up descriptions.

Under the null model, TF binding sites are modeled by inhomogeneous Poisson point process. For subinterval $(a, b]$ in R_r, the count of binding sites, denoted as $N(a, b]$, follows a Poisson distribution with rate $\lambda_{ab} = \int_a^b \lambda(x)dx$, where $\lambda(x), x \in (0, L]$ is the intensity function, i.e.,

$$P\{N(a, b] = z\} = \frac{(\lambda_{ab})^z}{z!} e^{-\lambda ab}, z = 0,1,2, \dots$$

For the full region $N(0, L]$, we have

$$P\{N(0, L] = z\} = \frac{(\lambda_{ab})^z}{z!} e^{-\lambda 0L}, z = 0,1,2, \dots$$

Let $N(0, L] = n$. Furthermore, if $(a_1, b_1]$ and $(a_2, b_2]$ are not overlapped, $N(a_1, b_1]$ and $N(a_2, b_2]$ are independent. Therefore, the binding sites in one gene promoter are i.i.d. with density function

$$f(x) = \frac{\lambda(x)}{\lambda_{0L}}$$

for $x \in (0, L]$. By assuming $\lambda(x)$ is piecewise constant, we can have $\lambda(x_1) = \lambda(x_2)$, for $x_1, x_2 \in (k, k+1]$, where $k$is an integer. Accordingly, for two integers $a < b$, we have

$$\int_a^b \lambda(x)dx = \sum_{k=a+1}^{b} \lambda(k)$$

As described in Choi *et al*. 2014, inhomogeneous Poisson process is a decent model for approximating TF binding site location distribution for two reasons. First, observing a binding site in a small sub-interval $(x - \Delta, x]$ is a rare event with probability $\propto \lambda(x)\Delta$, and non-overlapping intervals are independent. Second, the inhomogeneity feature of the process has taken into account the common situation where TF binding sites are more likely to locate near TSS.

We next assume that $f(x)$ is identical across all regions. Then for transcription factor X, we will further have

$$\lambda_X^{(r)}(x) = \lambda_{X,0L}^{(r)}(x) f_X(x),$$

where $\lambda_{X,0L}^{(r)}(x)$ stands for the expected number of binding sites of TF X in R_r, and $\lambda_{X,0L}^{(r)}(x) = \int_0^L \lambda_X^{(r)}(x)dx$. We further assume that the binding sites densities are piecewise constant where the piece length is h bps. If every piece is denoted B_i for $i = 1, \ldots, L/h$, and total number of binding sits for TF X in B_i over all regions is $B_i(x)$. Then $f_X(x)$ can be estimate by

$$\widehat{f_X}(x) = \frac{B_i(x)\ over\ all\ regions}{h \cdot n_{\square}},$$

where $n_{\square} = \sum_{r=1}^{N} n_r$ represents the total number of binding sites for X. Accordingly,

$$\lambda_X^{(r)}(x) = n_r \widehat{f_X}(x)$$

Similarly, we can obtain the intensity function for TF Y is

$$\widehat{f_Y}(y) = \frac{B_i(y)\ over\ all\ regions}{h \cdot n_{\square}},$$

where $n_{\square} = \sum_{r=1}^{N} m_r$, and

$$\lambda_Y^{(r)}(x) = m_r \widehat{f_Y}(y)$$

2.3 Ripley's K-function to summarize patterns between two TFs

Next we utilize the Ripley's K-function [16, 17] to summarize the co-binding patterns between two TFs. For the two binding sites distribution that we analyzed above, the bivariate K-function is the expected number of paired binding sites between TF X and Y with distance $\leq t$, normalized by the product of the two intensity functions. Therefore, let the binding sites locations for X and Y are $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_m)$. Then estimated K-function for every analyzed promoter region is

$$\widehat{K_{XY}}(t) = \frac{1}{L} \sum_{i=1}^{n} \sum_{j=1}^{\square} \frac{\omega(x_i, y_j)^{-1} I(d_{x_i, y_j} \leq t)}{\lambda_X(x_i)\lambda_Y(y_i)},$$

where $I(\cdot)$is the indicator function, $\omega(x_i, y_j)$ is a weight function for edge correction, $d_{x_i,y_j} = |x_i - y_j|$ is the distance between x_i and y_j. The edge effects occurs when two X and Y binding sites are clustered, but one site in the area analyzed while the other is outside of the region. This can cause bias when estimate the K-function. Accordingly, we apply Ripley's method for edge correction, which is described by $\omega(x_i, y_j)$. $\omega(x_i, y_j) =$ 1 if $[x_i - d_{x_i,y_j}, x_i + d_{x_i,y_j}] \subseteq (0, L]$, and ½ otherwise.

The K-function, $\widehat{K_{XY}^{(r)}}(t)$, are calculated for each region with a defined $t \in (0, L]$. Under the null hypothesis, where the two TFs binds to DNA sequences independently, the expectation and variance of $\widehat{K_{XY}^{(r)}}(t)$ will be:

$$\mathbb{E}\left[\widehat{K_{XY}^{(r)}}(t)\middle| n_r, m_r, \mathcal{H}_c\right] = 2t,$$

$$Var\left[\widehat{K_{XY}^{(r)}}(t)\middle| n_r, m_r, \mathcal{H}_c\right] = \frac{V(t) + 4(m_r - 1)C_X t^2 + 4(n_r - 1)C_Y t^2}{L^2 n_r m_r},$$

$$\sigma_r(t) = \sqrt{Var\left[\widehat{K_{XY}^{(r)}}(t)\middle| n_r, m_r, \mathcal{H}_c\right]},$$

where,

$$V(t) = \int_0^L \int_0^L \frac{k^2(x, y; t)}{\widehat{f_X}(x)\widehat{f_Y}(y)} dydx - (2tL)^2,$$

$$C_X = \int_0^L \frac{1}{\widehat{f_X}(x)} dx - L^2,$$

$$C_Y = \int_0^L \frac{1}{\hat{f}_Y(y)} dy - L^2$$

As shown above, $V(t), C_X$, and C_Y do not depend on r and thereby identical across different upstream regions.

If the two TFs co-bind with a distance of t between their binding sites, then the calculated $\widehat{K_{XY}^{(r)}}(t)$ values will be significantly larger the expected value $2t$. Vice versa, $\widehat{K_{XY}^{(r)}}(t)$ will be substantially smaller than $2t$ if the actual distance is larger than t, which indicates a repulsive effect between the two TFs. An example is shown in Figure 2.2 (A). This is the $\widehat{K_{XY}^{(r)}}(t)$ values of Oct4 and Sox2 in mouse ESCs. As shown in the figure, the $\widehat{K_{XY}^{(r)}}(t)$ values curve against t is much higher than the expected $2t$ for distance t we evaluated. This strongly suggests a tight clustering pattern Oct4 and Sox2.

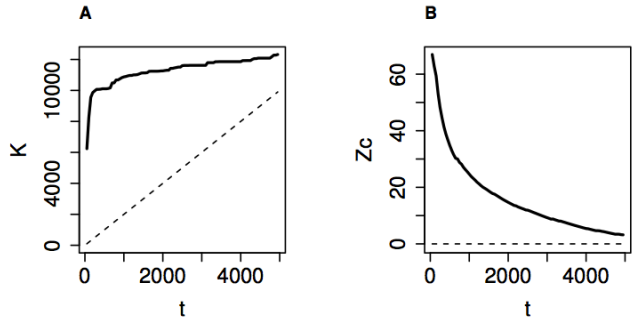

Figure 2.2: Example clustering binding patterns: (A) bivariate K-function of Oct4 and Sox2 (solid line), averaged over all regions, and the expected value 2t (dashed line); (B) Zc(t) of Oct4 and Sox2 ($t* = 50$). Figure adapted from Choi *et al.* 2014[15].

2.4 Z_c statistics to summarize genome-wide TF co-binding patterns

$\widehat{K_{XY}^{(r)}}(t)$ is computed for every promoter region. To understand the overall co-binding patterns of a TF pair over all of their common targets, we need a statistics to summarize them together. Here, used Z statistics as shown below:

$$Z_c(t) = \frac{1}{\sqrt{N}} \sum_{r=1}^{N} \frac{\widehat{K_{XY}^{(r)}}(t) - 2t}{\sigma_r(t)}.$$

It is easy to shown that $Z_c(t)$ has zero mean and unit variance under $\mathcal{H}_c$. Therefore, $Z_c(t)$ is a sum of N i.i.d. random variables, and it follows the standard normal distribution when N is large by Lindberg's central limit theorem.

$$Z_c(t)|\mathcal{H}_c \sim \mathcal{N}(0,1), as\ N \to \infty$$

In my thesis, I mainly focus on $t \in (0, 3000]$. According to previous studies, most TFs show significant clustering pattern when t is less than 100bp. Therefore, I picked t every 10bp when $t \in (0, 100]$, every 50bp when $t \in (100, 1000]$, and every 1000bp when $t \in (1000, 3000]$.

By plotting the $Z_c(t)$ curve (Figure 2.2B), this method can also report the optimal clustering distance t^* which maximizes $Z_c(t)$.

$$t^* = argmax_t Z_c(t)$$

Chapter 3.

General clustering patterns in hESC

3.1 Experimental settings

In this chapter, I investigate the TF binding pattern using the ChIP-Seq data generated by production groups in the ENCODE Consortium. ENCODE Consortium is an international collaboration of research groups, which aims at investigating the functional elements of human genome. Particularly, it provides 690 ChIP-seq datasets covering 161 unique regulatory factors across 91 human cell types under various treatment conditions, thus making it the most important human functional genomic database in the field [19-21]. Moreover, these ChIP-seq data are carefully analyzed using the standardized pipeline, and uniformed TF binding peaks are reported. Accordingly, given the size and quality of the data, ENCODE provides a substantial basis for our TF binding analysis.

I start my analysis with ChIP-seq experiments performed in human embryonic stem cells (hESC). hESCs are pluripotent stem cells derived from the inner cell mass of a blastocyst [22]. Many TFs with critical roles in various hESC activities haven been identified and well characterized individually. On one hand, transcription regulators, such as POU5F1, NANOG, and SOX2, are known to be crucial for maintaining the hESC capacity for self-renewal and pluripotency [23, 24]; On the other hand, introduction of other TFs, such as PAX6, MEF2, and FOXO3, leads to hESC differentiation in defined

trajectories [23, 25]. More importantly, those TFs usually function in a cooperative way. Therefore, it is very critical to investigate the TF binding patterns in hESC lines.

Table 3.1: List of the 21 TFs analyzed in the project with corresponding ENCODE ChIP-seq data and PWM from TRANSFAC database

TF	ChIP-Seq Data ID from ENCODE	Binding Site ID
ATF2	wgEncodeAwgTfbsHaibH1hescAtf2sc81188V0422111UniPk	M00040
ATF3	wgEncodeAwgTfbsHaibH1hescAtf3V0416102UniPk	M00513
CEBPB	wgEncodeAwgTfbsSydhH1hescCebpbIggrab	M00109,M00117
cMYC	wgEncodeAwgTfbsSydhH1hescCmycIggrabUniPk	M00322
cMYC	wgEncodeAwgTfbsUtaH1hescCmycUniPk	M00322
CREB1	wgEncodeHaibTfbsH1hescCreb1sc240V0422111PkRep1	M00039
CTCF	wgEncodeAwgTfbsBroadH1hescCtcfUniPk	N00015
CTCF	wgEncodeAwgTfbsHaibH1hescCtcfsc5916V0416102UniPk	N00015
CTCF	wgEncodeAwgTfbsUtaH1hescCtcfUniPk	N00015
GABPA	wgEncodeAwgTfbsHaibH1hescGabpPcr1x	M00341
JUND	wgEncodeAwgTfbsHaibH1hescJundV0416102	M00517,M00924,M00925,M00926
JUND	wgEncodeAwgTfbsSydhH1hescJundIggrab	M00517,M00924,M00925,M00926
JUN	wgEncodeAwgTfbsSydhH1hescCjunIggrab	M00517,M00924,M00925,M00926
MAX	wgEncodeAwgTfbsSydhH1hescMaxUcd	M00119
NANOG	wgEncodeAwgTfbsHaibH1hescNanogsc33759V0416102UniPk	N00002
NRF1	wgEncodeAwgTfbsSydhH1hescNrf1Iggrab	M00652
REST	wgEncodeAwgTfbsHaibH1hescNrsfV0416102	M00256,M01028
RFX5	wgEncodeAwgTfbsSydhH1hescRfx5200401194Iggrab	M00975
SP1	wgEncodeAwgTfbsHaibH1hescSp1Pcr1xUniPk	M00196
SP2	wgEncodeAwgTfbsHaibH1hescSp2V0422111	M00932,M00933
SP4	wgEncodeAwgTfbsHaibH1hescSp4v20V0422111	M00932,M00933
SRF	wgEncodeAwgTfbsHaibH1hescSrfPcr1x	M00152
TBP	wgEncodeAwgTfbsSydhH1hescTbpIggrab	M00471,M00980
USF1	wgEncodeAwgTfbsHaibH1hescUsf1Pcr1x	M00121,M00122,M00217,M00796
USF2	wgEncodeAwgTfbsSydhH1hescUsf2Iggrab	M00121,M00122,M00217,M00796
YY1	wgEncodeAwgTfbsHaibH1hescYy1sc281V0416102	M00059,M00069,M00793

Any binding site IDs starting in the format of M00XXX are PWM obtained from TRANSFAC database and M00XXX are index in the TRANSFAC database. Any IDs in the format of N000XX are PWM we generated from literatures.

Accordingly, here I apply the statistic method reviewed in Chapter 2 to 21 selected TFs (ATF2, ATF3, CEBPB, MYC, CREB1, CTCF, GABPA, JUND, JUN, MAX, NANOG, NRF1, REST, RFX5, SP1, SP2, SP4, TBP, USF1, USF2, and YY1, as

listed in Table 3.1), with available ChIP-seq data from ENCODE, and systematically investigate the co-binding patterns of all 210 resulting TF pairs. For TFs MYC, CTCF, and JUND, more than one ChIP-seq datasets are analyzed in this project to examine the reproducibility of our methods across different ChIP-seq experiments. Furthermore, to refine the exact TF binding location, I scan the known TF binding motifs (PWM) against the reported ChIP-seq peak regions to find the exact match position by utilizing cisgenome (http://www.biostat.jhsph.edu/~hji/cisgenome/) tools [18]. PWM are downloaded from the TRANScription FACtor database (TRANSFAC) (http://www.gene-regulation.com/pub/databases.html) [26, 27]. I should point out that for some TFs, multiple PWMs are included as listed in Table 3.1. As a result, a total of 1128 TF pairs are evaluated, and 837 of them shared at least 10 gene targets, which are thereafter used for follow-up analysis.

3.2 General clustering patterns

In this section, I will first use TF pair CTCF and cMYC as an example to illustrate the analysis results. In particular, three CTCF ChIP-seq data and one cMYC ChIP-seq data are utilized in this analysis, and one binding motif is evaluated per TF. First of all, the density function f_X for each TF is estimated, and the bivariate K-function for each upstream region is thereafter computed. After that, Z-scores are calculated afterwards for a sequence of distances $t \in [10,3000]$ (10 spaced in between [10,100]; 50 spaced in between [100, 1000]; 1000 spaced in between [1000,3000]). As shown in Figure 3.1 A, the maximum Zc values are obtained in the region of $t \in [100,150]$ for all

three CTCF ChIP-seq experiments with maximal Zc value of 11.82, 20.49, and 15.06, respectively, indicating that the clustering binding of CTCF and cMYC binding sites

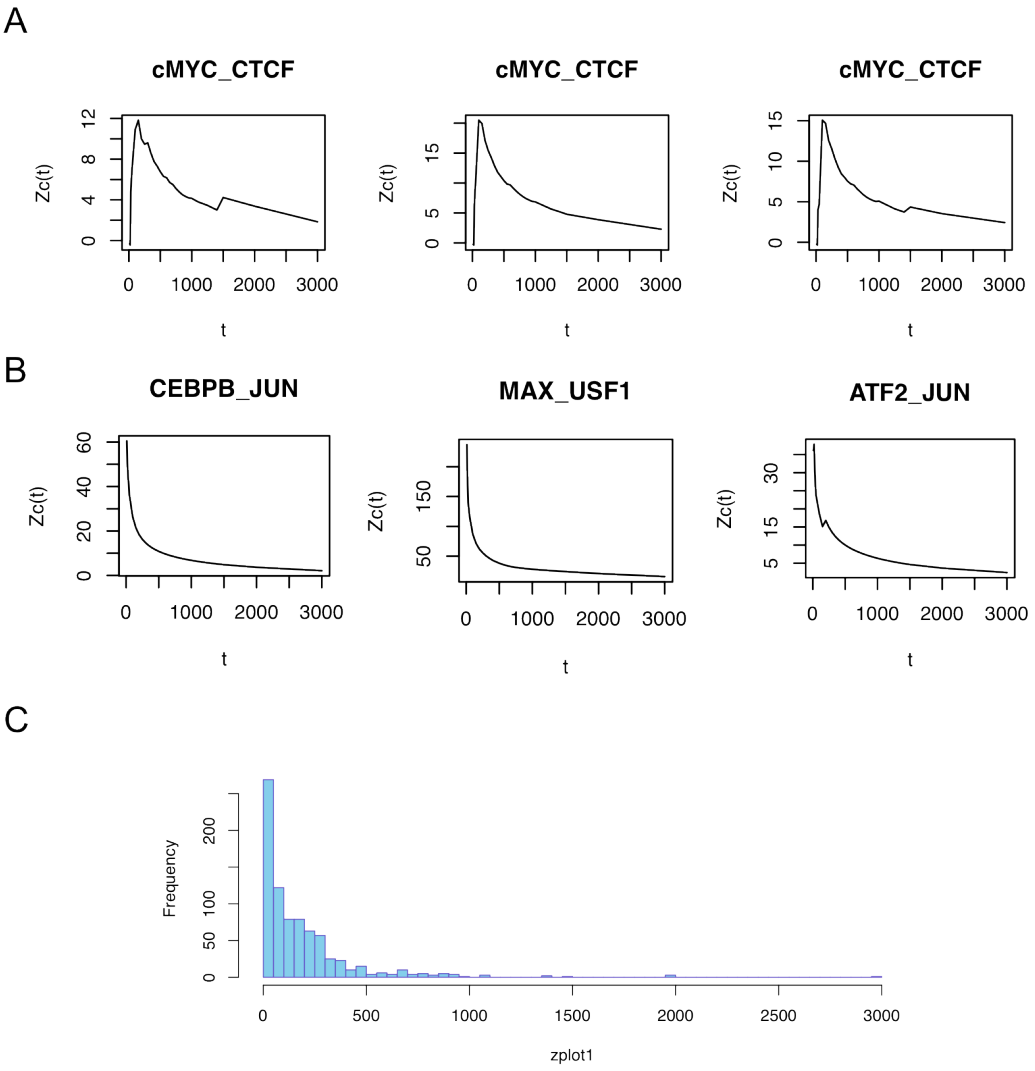

Figure 3.1: General clustering pattern in hESC. (A) Zc distribution against t for transcription factors cMYC and CTCF using three replicated CTCF ChIP-seq datasets from ENCODE. (B) Example TF pairs with cluster distance smaller than 50bp. Shown are the Zc statistic against t for transcription factors CEBPB and JUN, MAX and USF1, ATF2 and JUN. (C) Histogram of the most significant clustering distance t^* in hESC. Only significant TF pairs are shown in the plot.

become most prominent when examining their binding sites distance around 100bp. Furthermore, we can see that the Zc score decreases as t increased, and their clustering pattern becomes less significant as the distance threshold increases. Given the fact that CTCF and cMYC do interact with each other to cooperatively regulate cell activity, the clustering pattern observed here can thus be well-explained.

For every TF pairs, we also compute the most significant clustering distance t* based on obtained Zc(t). 177 out of 244 analyzed TF pairs show significant clustering patterns with FDR<0.05, which is consistent with the previous notion that most TFs function coherently in hESCs [11, 28]. Specifically, the optimal t* for selected TFs pairs are listed in supplementary Table 1.

Next, I evaluate the t* distribution among the analyzed TF pairs. As shown in Figure 3.1C, the most maximal distance thresholds t* are within 100bp, while only a few exceed 500bp. The median and mean values of t* are 150bp and 199bp, respectively. This is in line with previous observation of short-range co-occurrence among TF binding sites in regulatory regions [11-14]. In particular, 269 TF pairs have optimal cluster distance smaller than 50bp, suggesting that they may bind to composite sites as protein complexes. As the illustration, Zc(t) of protein pairs of ATF2-JUN, MAX-USF1, and CEBPB-JUN are shown in Figure 3.1B. Among them, ATF2 and JUN, as well as MAX and USF1, are known as the co-binding partners with composite binding sites which have already been identified [29, 30]. In contrast, the TFs whose binding sites are separated by a few hundred base pairs may provide line of evidence for previously suggested cis-regulatory modules

(CRM) [31]. Examples are CEBPB-SP2 (t*=950bp), CREB1-MAX (t*=900bp), NANOG-TBP (t*=700bp), etc..

3.3 Clustering pattern of different binding motifs (PWM) for the same TF

To systematically evaluate the clustering pattern among multiple TFs discussed in Section 3.1, I will next elaborate the identified co-binding patterns using the heatmap plots. Figure 3.2 shows the four heatmaps based on log_{10} transformed Zc values for all analyzed TF pairs, which represent four different clustering distances: 50, 100, 500, and 3000. It can be clearly seen from each plot that several red squares along the diagonal indicate extremely significant clustering pattern among different binding motifs of the same TF. For example, I investigate four binding motifs for JUND provided by the TRANSFAC database (M00517, M00924, M00925, and M00926). As shown in the heatmap plot, these four binding motifs always give the highest co-binding scores regardless of the t value. In fact, this is expected since most of those PWMs are duplicated reports on the same binding motif with minor differences. Similar pattern can also be observed for TFs JUN, SP2, SP4, TBP, USF1, and USF2.

Nevertheless, it is noted that there are some exceptions. As shown in Figure 3.3, different PWMs give different Zc distributions. Figure 3.3 (A-C) display the Zc values against t for two PWMs of CEBPB with ATF3, SP1, and CTCF, respectively. The sequence logo plots of the two corresponding CEBPB PWMs are shown below the distribution plots in Figure 3.3. It can be observed that even though the two CEBPB PWMs are nearly the same except two middle nucleotides, they lead to very different t^*

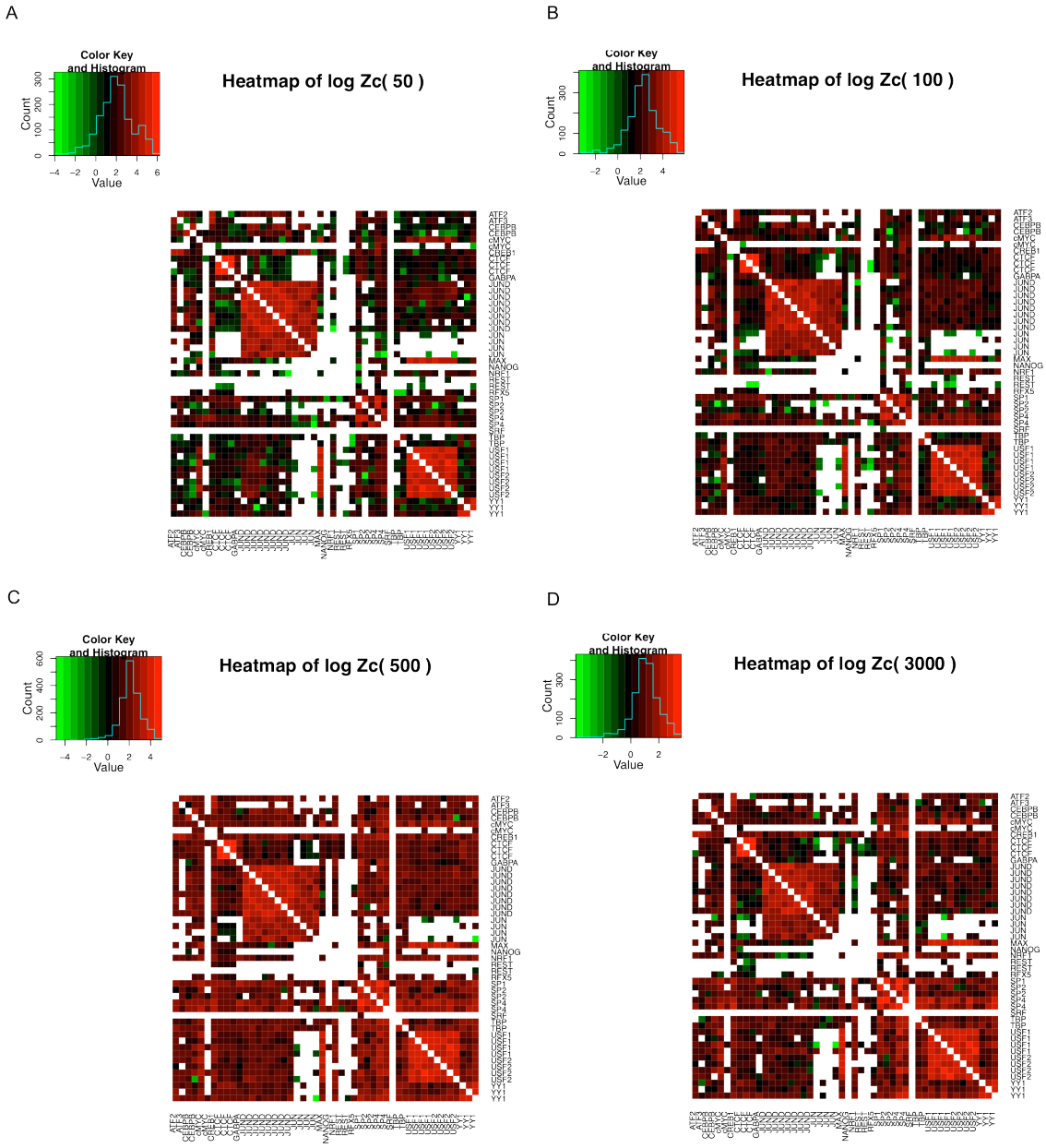

Figure 3.2: Heatmap using $\log_{10}$ transformed Zc statistics among the 21 TFs. (A) t=50. (B) t=100. (C) t=500. (D) t=3000. Only TF pairs with more than 10 shared targets are analyzed and displayed in the heatmap. TF pairs with less than 10 targets are colored by white.

values. Take ATF3-CEBPB pair as an example: the t^* value for the left PWM is 150bp, while the t^* for the right PWM is less than 10bp. Such difference suggests that there

might be two different ways for CEBPB to cooperate with ATF3. Similar patterns can also be observed for TF pair of cMYC and YY1. Consequently, our results imply that different mechanisms are involved in TF interactions, which might be determined by binding motifs switches. These different regulatory mechanisms usually lead to different sets of target genes, which eventually result in very different biological consequences. Further, motif-stratified functions may also partially explain how limited number of transcription factors control diverse cellular activities.

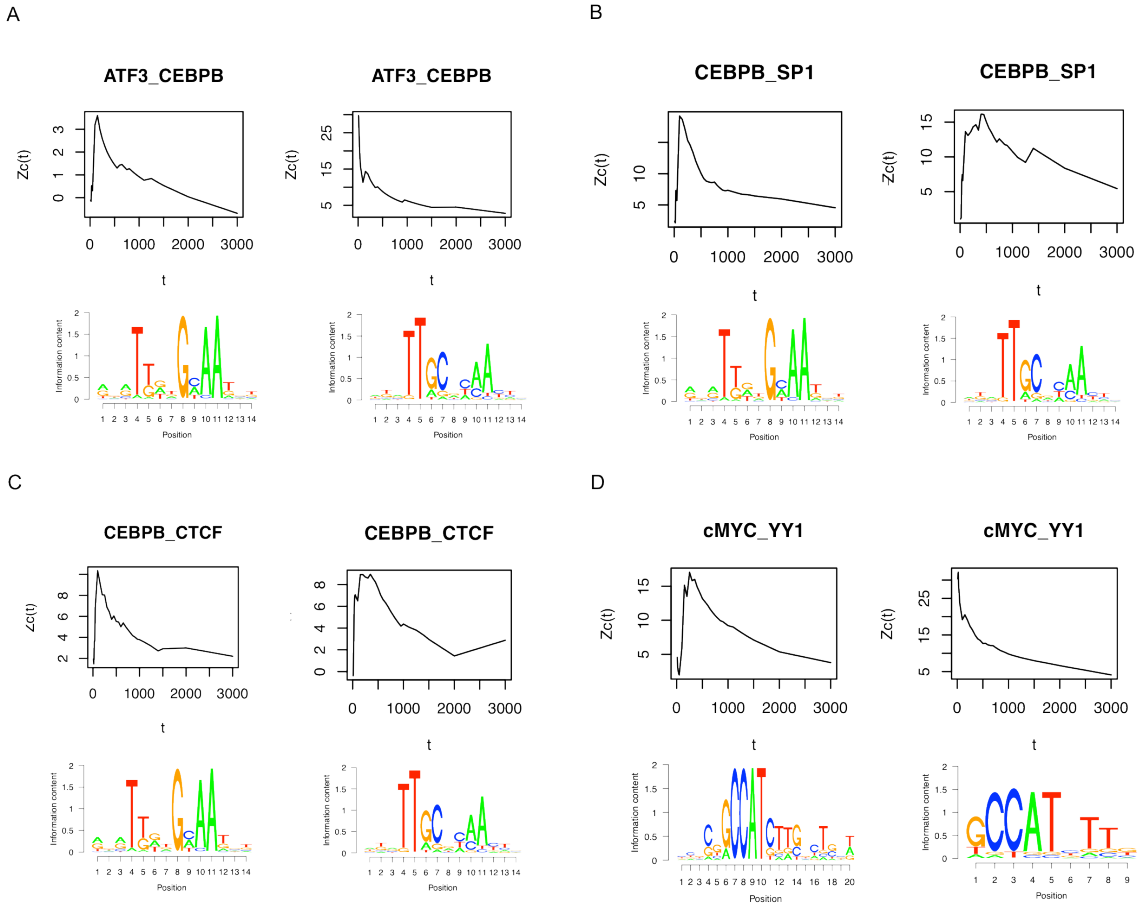

Figure 3.3: Clustering pattern of different binding motifs for the same transcription factor. Shown are Zc distribution against t for TF pairs: (A) ATF3 and CEBPB; (B) CEBPB and SP1; (C) CTCF and CEBPB; and (D) cMYC and YY1. In (A-C), the left and right plot represent two different PWM of CEBPB with corresponding PWM displayed below, while in D are two different PWM for YY1.

3.4 Clustering pattern of TFs in the same protein family

Figure 3.4 shows an example where two TFs, ATF2 and CREB1, from the same protein family—leucine zipper family of DNA binding proteins—display consistent co-binding patterns. Both of these two TFs stimulate transcription upon binding to the DNA cAMP response element (CRE). They are known to have similar biological functions, and usually form dimers to activate downstream transcriptions [32]. In fact, ATF2 and CREB1 binding sites are tightly clustered in our data with optimal distance t^* less than 10bp, demonstrating that these two proteins work as dimers (Figure 3.4). Furthermore, these two TFs share common co-binding patterns with other TFs, as illustrated in Figure 3.2. This implicates shared biological functions among them. In addition, other examples, such as USF1 and USF2, as well as JUND and JUN, are also displaying coherent Zc patterns.

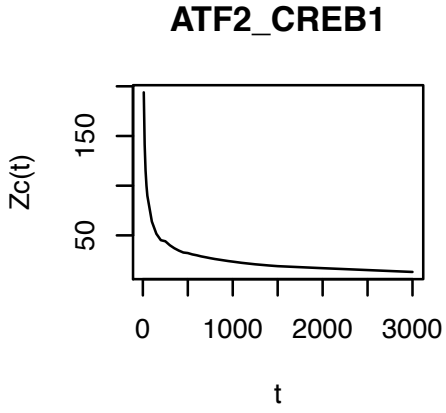

Figure 3.4: Zc statistics against t for ATF2 and CREB1.

3.5 TF co-binding pattern

As shown in Figure 3, it can be seen that the Zc patterns are highly dependent on distance t. Therefore, to compare the co-binding patterns among all analyzed TFs, a summary statistics over t is required. According to Choi et. al.[15], here I use average Zc(t) with t from 50 to 1000, denoted as $\overline{Z_{1000}}$, to summarize the Zc scores over various t for each protein pairs so as to avoid selection bias. As illustrated before [15], $\overline{Z_{1000}}$ provides detailed information about the co-binding pattern among various TFs. The heatmap based on $\overline{Z_{1000}}$ is shown in Figure 3.5.

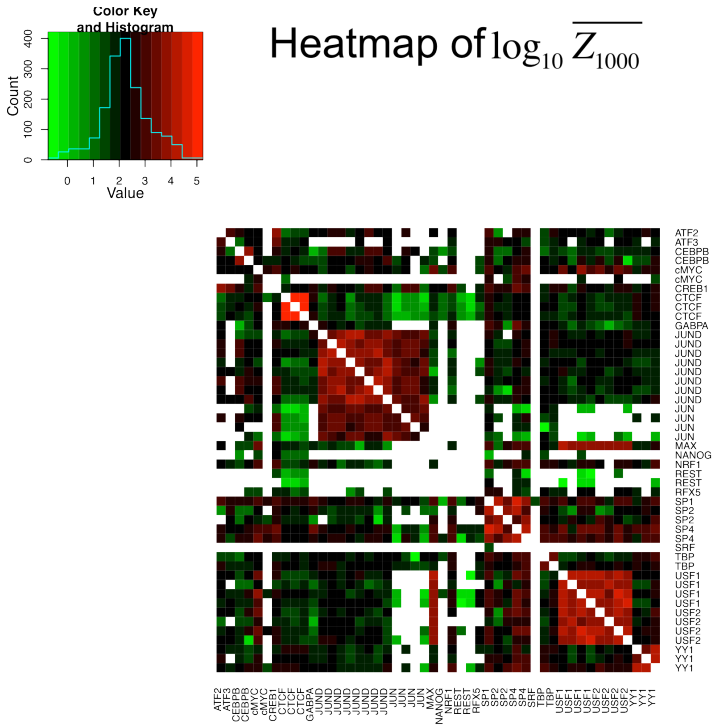

Figure 3.5: Heatmap using the $\log_{10}$ transformed average Zc(t) over t from 50 to 1000.

First of all, according to the summarized heatmap plot, our analysis verifies known TF co-binding patterns. For example, for all ATF2 involved protein pairs, JUND and JUN give highest $\overline{Z_{1000}}$ values. This is consistent with the known fact that ATF2 usually forms a homodimer or a heterodimer with c-Jun to stimulates CRE-dependent transcription [33]. Furthermore, tight clusters are also observed among known transcription co-regulators of CREB1-cJUN, as well as USF-cMyc-MAX complex [34]. On one hand, the consistency with previous findings strongly support the validity of this designed statistic method. On the other hand, this analysis provides additional line of evidence to support synergistic interaction among those TFs.

Moreover, TBP is the TATA Box Binding Protein, which binds to the core promoter to position the polymerase properly. I find that the TBP binding sites are tightly clustered with SP1/2/4, which also binds to the promoter regions. This is consistent with the fact that SP TFs usually interact with TBP to form the transcription initiation complex. Additionally, since TBP binding sites represent the core promoter region, evaluation the clustering pattern of other TFs with respect to TBP can provide distance information about their closeness to promoters. As shown in the Figure 3.6, most TFs binding sites show a clustering distance within 300bp to the promoters. The mean and median optimal distance t^* with regard to TBP binding sites are 269bp and 200bp, respectively, as compared to 199bp and 150bp for the overall t^* distribution. More remarkably, the first quartile for TBP clustering distance is 150bp, while that of the overall distribution is only 30bp. This suggests that TFs usually act in clusters on the DNA sequence around 200bp away from the promoter region.

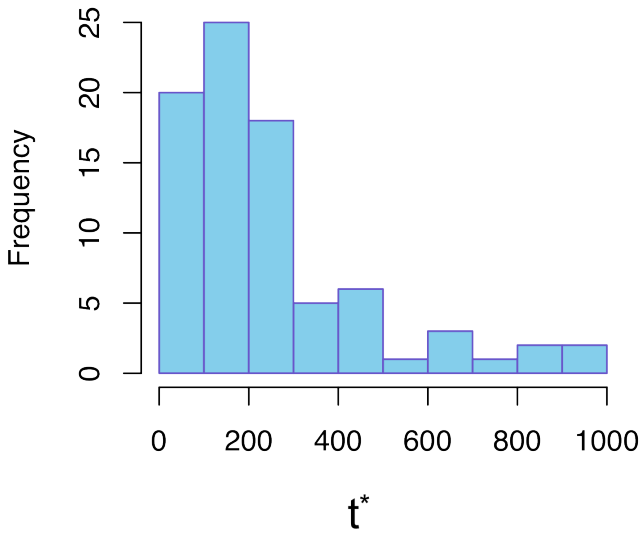

Figure 3.6: Histogram of the most significant clustering distance t^* with TBP in hESC.

Chapter 4.

Tissue-specific TF co-binding patterns

4.1 General clustering pattern in GM12878 cell lines.

Table 4.1: List of the 11 TFs analyzed in GM12878 cell line.

TF	ChIP-seq Data ID from ENCODE	PWM ID
ATF2	wgEncodeAwgTfbsHaibGm12878Atf2sc81188V0422111	M00040
ATF3	wgEncodeAwgTfbsHaibGm12878Atf3Pcr1x	M00513
CEBPB	wgEncodeAwgTfbsHaibGm12878Cebpbsc150V0422111	M00109
cMYC	wgEncodeAwgTfbsUtaGm12878Cmyc	M00322
CREB1	wgEncodeHaibTfbsGm12878Creb1sc240V0422111PkRep1(broadPeak)	M00039
CTCF	wgEncodeAwgTfbsSydhGm12878Ctcfsc15914c20	N00015
GABPA	wgEncodeAwgTfbsHaibGm12878GabpPcr2x	M00341
JUND	wgEncodeAwgTfbsSydhGm12878Jund	M00517
MAX	wgEncodeAwgTfbsSydhGm12878MaxIggmus	M00119
USF1	wgEncodeAwgTfbsHaibGm12878Usf1Pcr2x	M00217
USF2	wgEncodeAwgTfbsSydhGm12878Usf2Iggmus	M00217

Any binding site IDs starting in the format of M00XXX are PWM obtained from TRANSFAC database and M00XXX are index in the TRANSFAC database. Any IDs in the format of N000XX are PWM we generated from literatures.

It is known that gene expression varies across tissues, which implicates tissue-specific transcriptional regulation. Accordingly, I hypothesized that the binding sites of transcription factors, which are involved in tissue-specific regulation, are clustered with different distances across tissues. To verify this hypothesis, I next evaluate the TF co-binding pattern in GM12878 cell lines. GM12878 is the lymphoblastoid cell lines derived from the peripheral blood mononuclear cells via Epstein-Barr Virus transformation. This cell line represents the mesoderm cell lineage, and is widely used in biological researches. In this study, among the 21 TFs assessed in hESC, 11 are included in

GM12878 analysis. They are ATF2, ATF3, CEBPB, cMYC, CREB1, CTCF, GABPA, JUND, MAX, USF1, and USF2 (Table 4.1), and 50 resulting TF pairs are evaluated. According to the hESC analysis, different binding motifs of the same TF usually give consistent clustering patterns. Thus, only one binding motif is selected for each TF in GM12878 analysis, as listed in Table 4.1.

Table 4.2: The most significant distance t^* in GM12878 cell lines for all assessed TF pairs.

TF1	TF2	GM12878			hESC			
		Zcmax	TmaxZc (bp)	PmaxZc	Zcmax	TmaxZc (bp)	PmaxZc	Δt^* (bp)
ATF2	ATF3	70.7	t_10	0.0E+00	87.2	t_10	0.0E+00	0
ATF2	CEBPB	16.3	t_90	0.0E+00	47.5	t_10	0.0E+00	80
ATF2	cMYC	18.4	t_350	0.0E+00	12.3	t_350	0.0E+00	0
ATF2	CREB1	316.7	t_10	0.0E+00	193.9	t_10	0.0E+00	0
ATF2	CTCF	3.0	t_300	2.4E-03	4.0	t_300	6.7E-05	0
ATF2	GABPA	9.3	t_350	0.0E+00	12.3	t_250	0.0E+00	100
ATF2	MAX	11.7	t_70	0.0E+00	15.6	t_100	0.0E+00	-30
ATF2	USF1	13.2	t_200	0.0E+00	10.8	t_250	0.0E+00	-50
ATF2	USF2	16.0	t_300	0.0E+00	7.6	t_300	2.5E-14	0
ATF3	cMYC	8.9	t_250	0.0E+00	18.2	t_150	0.0E+00	100
ATF3	CREB1	93.9	t_10	0.0E+00	124.3	t_10	0.0E+00	0
ATF3	CTCF	11.0	t_30	0.0E+00	9.0	t_200	0.0E+00	-170
ATF3	USF2	6.8	t_70	7.8E-12	9.5	t_300	0.0E+00	-230
CEBPB	cMYC	9.2	t_30	0.0E+00	14.2	t_150	0.0E+00	-120
CEBPB	CREB1	34.0	t_10	0.0E+00	38.4	t_10	0.0E+00	0
CEBPB	CTCF	6.1	t_250	8.6E-10	10.4	t_100	0.0E+00	150
CEBPB	GABPA	0.9	t_200	3.7E-01	3.4	t_1500	6.3E-04	-1300
CEBPB	MAX	14.8	t_70	0.0E+00	7.9	t_150	2.9E-15	-80
CEBPB	USF1	3.9	t_70	9.8E-05	7.5	t_200	5.4E-14	-130
CEBPB	USF2	7.0	t_1000	2.8E-12	7.2	t_400	8.3E-13	600
cMYC	CREB1	20.2	t_300	0.0E+00	21.2	t_400	0.0E+00	-100
cMYC	CTCF	14.6	t_350	0.0E+00	11.8	t_150	0.0E+00	200
cMYC	GABPA	25.3	t_70	0.0E+00	14.6	t_200	0.0E+00	-130
cMYC	MAX	75.9	t_10	0.0E+00	92.0	t_10	0.0E+00	0
cMYC	USF1	225.3	t_10	0.0E+00	195.1	t_10	0.0E+00	0
cMYC	USF2	203.6	t_10	0.0E+00	153.7	t_10	0.0E+00	0

CREB1	CTCF	10.1	t_90	0.0E+00	7.3	t_400	2.5E-13	-310
CREB1	GABPA	20.7	t_40	0.0E+00	20.0	t_400	0.0E+00	-360
CREB1	MAX	15.0	t_550	0.0E+00	11.7	t_900	0.0E+00	-350
CREB1	USF1	13.0	t_900	0.0E+00	15.6	t_300	0.0E+00	600
CREB1	USF2	14.2	t_300	0.0E+00	18.8	t_300	0.0E+00	0
CTCF	GABPA	10.4	t_100	0.0E+00	21.1	t_100	0.0E+00	0
CTCF	MAX	7.1	t_50	9.8E-13	6.1	t_100	8.8E-10	-50
CTCF	USF1	10.8	t_400	0.0E+00	11.4	t_400	0.0E+00	0
CTCF	USF2	9.1	t_350	0.0E+00	7.5	t_150	9.0E-14	200
GABPA	MAX	13.0	t_70	0.0E+00	9.5	t_50	0.0E+00	20
GABPA	USF1	26.4	t_50	0.0E+00	11.5	t_50	0.0E+00	0
GABPA	USF2	14.7	t_70	0.0E+00	2.8	t_200	4.6E-03	-130
MAX	USF1	157.1	t_10	0.0E+00	257.1	t_10	0.0E+00	0
MAX	USF2	246.8	t_10	0.0E+00	236.3	t_10	0.0E+00	0
USF1	USF2	616.4	t_10	0.0E+00	560.0	t_10	0.0E+00	0

Zcmax: the maximum Zc scores over t.
TmaxZc: the distance t^* that maximize Zc scores. It is also named as "the most significant distance".
PmaxZc: the p-value at the most significant distance.
Δt^*: the difference in the most significance distances between hESC and GM12878 ($\Delta t^* = t^*_{GM12878} - t^*_{hESC}$)

The most significant distance t* in GM12878 cell lines for all assessed TF pairs are reported in table 4.2. The distribution of the most significant distance t* is comparable to what I obtained in hESC analysis (Figure 4.1A), with average t* of 181bp and 199bp in GM12878 and hESC, respectively. Similar with previous results in hESCs, almost all of the TF pairs have optimal distance t* within 500bps. In addition, as shown in Figure 4.1B, TFs belong to the same protein family tend to have more significant clustering Zc statistics over t, such as ATF2 and CREB pair, and USF1 and USF2 pair, manifesting their constitute functional relatedness across cell types. Furthermore, tight clustering is also found for known TF binding partners of cMYC-USF1/2-MAX, which is consistent with the observation in hESC. The most significant distances among their bindings sites are less than 10bp in both hESC and GM12878 cell lines. This implicates

that these three TFs form a tight complex to activate downstream transcriptions in both hESCs and GM12878 cell lines.

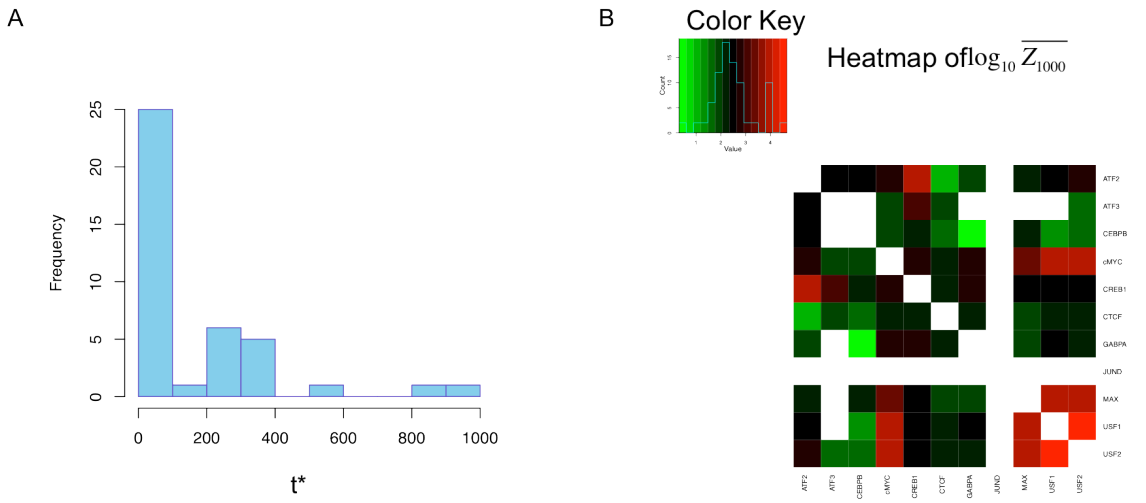

Figure 4.1: General clustering pattern in GM12878. (A) Histogram of the most significant clustering distance t^* for all TF pairs in GM12878. Only significant TF pairs are shown in the plot. (B) Heatmap using the $\log_{10}$ transformed average of Zc(t) over t from 50 to 1000 in GM12878.

4.2 Comparison of the TF co-binding pattern between hESC and GM12878

The most significant distance t^*s in hESC cells are also attached with corresponding TF pairs in Table 4.2, and the difference in t* between the two cell types, denoted as Δt* (t* in GM12878 – t* in hESC), is calculated. The distribution of Δt* is shown in Figure 4.2. The quantiles of Δt* are -1300(0%), -100(25%), 0(50%), 0(75%), and 600(100%), respectively. As expected, for most of the TF pairs, their Δt^*'s are in the range of [-100, 100]. The small change suggests that most TFs preserve their co-binding patters across hESCs and GM12878 cell lines.

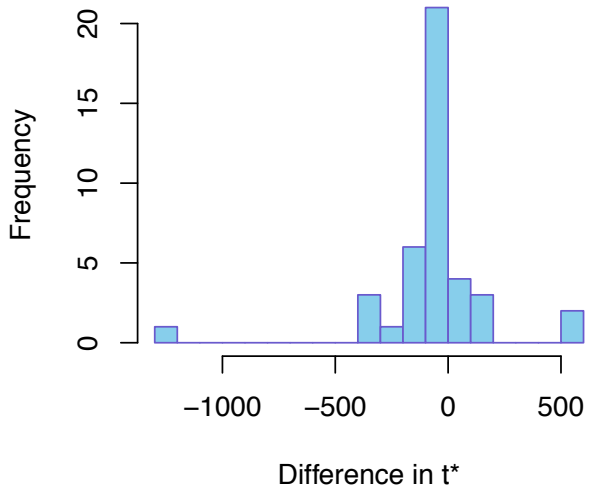

Figure 4.2: Histogram of difference in t* between hESC and GM12878 cell lines.

In addition to the similarities discussed above, my analysis also identifies tissue-specific protein co-binding patterns. It can be seen that there are 15 TF pairs with |Δt*| larger than 100bp. Particularly, three of them are above 500bp: CEBPB-GABPA (Δt*=-1300), CREB1-USF1(Δt*=600), and CEBPB-USF2(Δt*=600). The Zc(t) curves of some example TFs are illustrated in Figure 4.3. For example, ATF2 and CEBPB tends to have an overlapping binding sites in hESC cells (t*<10), while the distances between their binding sites in GM12878 are much larger and less significant(Figure 4.3A). Additionally, for the other three TF pairs, they show significant clustering pattern in both hESC and GM12878, however, the most significant clustering distance t* are different in the two cell types. CEBPB and CTCF tend to have a closer interaction in hESCs than GM12878 cell lines, while ATF3-CTCF and CREB1-GABPA show opposite patterns.

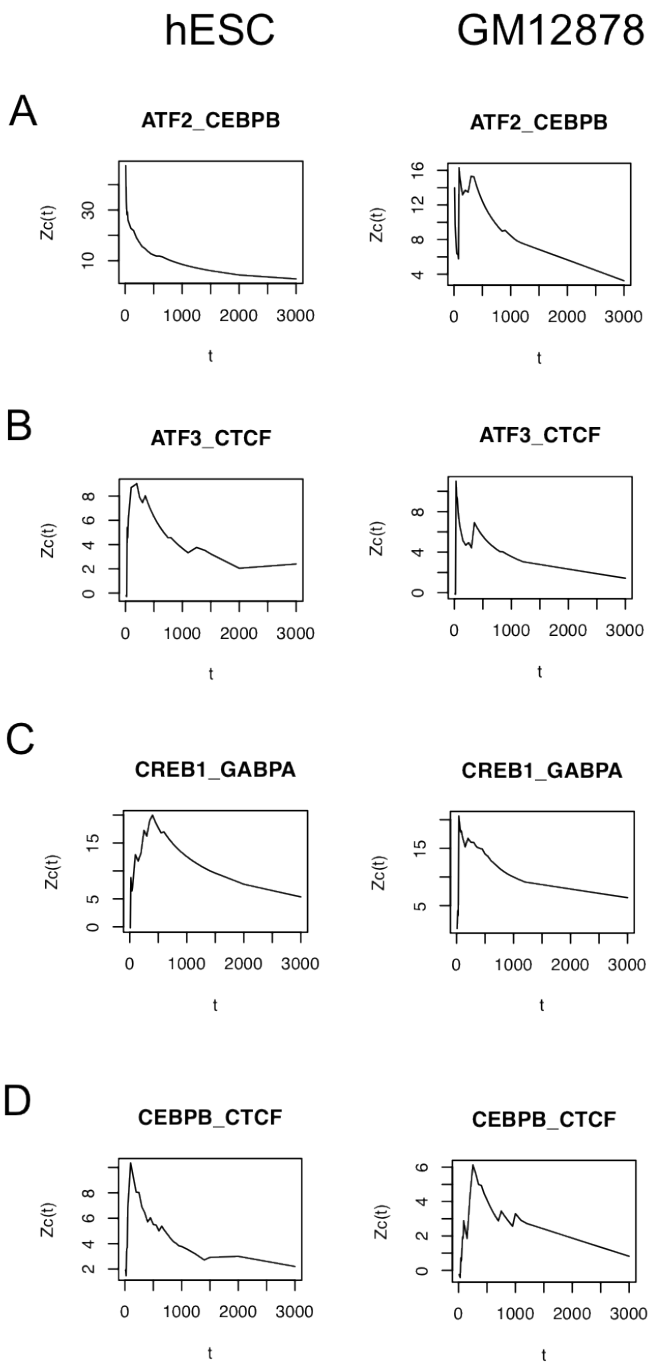

Figure 4.3: Examples illustrating different co-binding patterns in hESC and GM12878 cell line. Shown are the Zc statistic against t for TF pairs of (A) ATF2_CEBPB, (B) ATF3_CTCF, (C) CREB1_GABPA, and (D) CEBPB_CTCF.

Chapter 5

Conclusions and future directions

5.1 Conclusions

We have applied a novel test statistics to investigate the combinatorial binding patterns among TFs using ChIP-seq data. Specifically, we utilize the inhomogeneous Poisson process to estimate TF binding site distributions, and then apply the Ripley's K-function to summarize the binding patterns between two selected TFs. By approximated by the standard normal distribution, the clustering significance is consequently estimated using Z-statistics of the K-functions. Further evaluating the distribution of the Z scores, our method can prioritize the most significant clustering distance t* for every assessed TF pairs. More importantly, as previously shown, these procedures are easy to scale up from pairwise analysis to multiple TF combination using heatmap graphs.

To the best of my knowledge, this is the first work that comprehensively evaluates TF co-binding patterns in human genomes, which will greatly advance the future researches in the field of transcription regulation. By applying the method to two different human cell lines that are widely used in biological researches, we show the co-occurrence binding site patterns of selected TFs. The observed patterns provide direct evidences for the suggested interactions among TFs on expression regulation.

In addition, I also compare the identified co-binding patterns between human ESCs and human lymphoblastoid cell lines (GM12878). Several TFs are found to be

bound differently between the two cell lines. This implicates that different transcriptional machineries/mechanisms are involved, which warrants further experimental investigation.

5.2 Future directions

In current work, we mainly focus on evaluating pairwise TF co-binding patterns. Since usually more than two TFs are involved in a transcription regulation complex, the main goal for future researches is to scale up the study to multiple TF analysis. A simple analysis we could do is to analyze the change in pairwise patterns with respect to a few other TFs, discussing the pattern alterations with and without the existence of other TFs. This will not only be helpful to show the group binding patterns, but also be useful to prioritize the core factors in a transcription regulation complex. Furthermore, we may improve the test statistics so as to be able to investigate the clustering patterns of multiple TFs.

Since epigenetic modifications, including histone methylation and acetylation, have great impact on transcription binding, the next goal is to examine how the clustering patterns would be affected by different histone modification. We have done a preliminary analysis with regard to the H3K9 acetylation marks. H3K9 acetylation usually represents transcription activation, and interacts with many well-known TFs, such as CREB1. Here, we assessed the Zc(t) distribution stratified by H3K9 acetylation status. Several TF pairs are identified by showing different clustering patterns. The most prominent examples are TFs pairs where CTCF is involved. For example, a significant clustering pattern is observed for TF pairs, MAX-CTCF and USF2-CTCF, when their binding sites falls into

the H3K9 acetylation region, while no significant pattern are detected when their binding sites locate out of the H3K9 acetylation region. This suggests the involvement of H3K9 acetylation in CTCF participated protein complex organization. To extend this analysis, more histone modifications need to be investigated.

5.3 Summary

In summary, the presented work has demonstrated unequivocally the development and utility of the designed test statistics on evaluating TF binding site clustering patterns. The identified patterns can help to infer the TF interactions and their effects on gene expression regulation.

REFERENCES

1. Stormo, G.D. DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16-23 (2000).
2. Stormo, G.D., Schneider, T.D., Gold, L. & Ehrenfeucht, A. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res* **10**, 2997-3011 (1982).
3. Song, G.S. et al. Comparative transcriptional profiling and preliminary study on heterosis mechanism of super-hybrid rice. *Mol Plant* **3**, 1012-1025.
4. Claverie, J.M. & Audic, S. The statistical significance of nucleotide position-weight matrix matches. *Comput Appl Biosci* **12**, 431-439 (1996).
5. Schones, D.E., Smith, A.D. & Zhang, M.Q. Statistical significance of cis-regulatory modules. *BMC Bioinformatics* **8**, 19 (2007).
6. Staden, R. Methods for calculating the probabilities of finding patterns in sequences. *Comput Appl Biosci* **5**, 89-96 (1989).
7. Bailey, T.L. et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**, W202-208 (2009).
8. Bhar, A. et al. Coexpression and coregulation analysis of time-series gene expression data in estrogen-induced breast cancer cell. *Algorithms Mol Biol* **8**, 9.
9. Sarkar, C. & Maitra, A. Deciphering the cis-regulatory elements of co-expressed genes in PCOS by in silico analysis. *Gene* **408**, 72-84 (2008).
10. Veerla, S. & Hoglund, M. Analysis of promoter regions of co-expressed genes identified by microarray analysis. *BMC Bioinformatics* **7**, 384 (2006).

11. Chen, X. et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106-1117 (2008).

12. Kazemian, M., Pham, H., Wolfe, S.A., Brodsky, M.H. & Sinha, S. Widespread evidence of cooperative DNA binding by transcription factors in Drosophila development. *Nucleic Acids Res* **41**, 8237-8252.

13. Lee, Y. & Zhou, Q. Co-regulation in embryonic stem cells via context-dependent binding of transcription factors. *Bioinformatics* **29**, 2162-2168.

14. Oh, Y.M., Kim, J.K., Choi, S. & Yoo, J.Y. Identification of co-occurring transcription factor binding sites from DNA sequence using clustered position weight matrices. *Nucleic Acids Res* **40**, e38.

15. Choi, M. & Zhou, Q. Detecting Clustering and Ordering Binding Patterns Among Transcription Factors via Point Process Models. *Bioinformatics*.

16. Dixon, P.M. Ripley's K function, Vol. 3. (Encyclopedia of Environmetrics, 2002).

17. Ripley, B.D. 2nd-Order Analysis of Stationary Point Processes. *Journal of Applied Probability* **13**, 255-266 (1976).

18. Ji, H. et al. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat Biotechnol* **26**, 1293-1300 (2008).

19. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* **9**, e1001046.

20. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636-640 (2004).

21. Qu, H. & Fang, X. A brief review on the Human Encyclopedia of DNA Elements (ENCODE) project. *Genomics Proteomics Bioinformatics* **11**, 135-141.

22. Thomson, J.A. et al. Embryonic stem cell lines derived from human blastocysts. *Science* **282**, 1145-1147 (1998).

23. Tam, W.L. & Lim, B. Genome-wide transcription factor localization and function in stem cells. (2008).

24. Yeo, J.C. & Ng, H.H. The transcriptional regulation of pluripotency. *Cell Res* **23**, 20-32.

25. Yamamizu, K. et al. Identification of Transcription Factors for Lineage-Specific ESC Differentiation. *Stem Cell Reports* **1**, 545-559.

26. Wingender, E. et al. The TRANSFAC system on gene expression regulation. *Nucleic Acids Res* **29**, 281-283 (2001).

27. Wingender, E. et al. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res* **28**, 316-319 (2000).

28. Marson, A. et al. Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* **134**, 521-533 (2008).

29. Ramirez-Carrozzi, V. & Kerppola, T. Asymmetric recognition of nonconsensus AP-1 sites by Fos-Jun and Jun-Jun influences transcriptional cooperativity with NFAT1. *Mol Cell Biol* **23**, 1737-1749 (2003).

30. Terragni, J. et al. The E-box binding factors Max/Mnt, MITF, and USF1 act coordinately with FoxO to regulate expression of proapoptotic and cell cycle control genes by phosphatidylinositol 3-kinase/Akt/glycogen synthase kinase 3 signaling. *J Biol Chem* **286**, 36215-36227.

31. Zhou, Q. & Wong, W.H. CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc Natl Acad Sci U S A* **101**, 12114-12119 (2004).

32. Tacke, F., Liedtke, C., Bocklage, S., Manns, M.P. & Trautwein, C. CREB/PKA sensitive signalling pathways activate and maintain expression levels of the hepatitis B virus pre-S2/S promoter. *Gut* **54**, 1309-1317 (2005).

33. Hayakawa, J. et al. Identification of promoters bound by c-Jun/ATF2 during rapid large-scale gene activation following genotoxic stress. *Mol Cell* **16**, 521-535 (2004).

34. Walhout, A.J., Gubbels, J.M., Bernards, R., van der Vliet, P.C. & Timmers, H.T. c-Myc/Max heterodimers bind cooperatively to the E-box sequences located in the first intron of the rat ornithine decarboxylase (ODC) gene. *Nucleic Acids Res* **25**, 1493-1501 (1997).