

# Association of Histone Modification Patterns With Transcription Factor Binding Revealed by Systematic Analysis\*

LI Wen, LI Wen-Ting, CAO Shu-Juan, YU Wen-Xiang, LI Wen-Jie, DAI Jun-Biao\*\*, SUN Zhi-Rong\*\*

(MOE Key Laboratory of Bioinformatics, State Key Laboratory of Biomembrane and Membrane Technology,  
School of Life Sciences, Tsinghua University, Beijing 100084, China)

**Abstract** Selective binding of transcription factors (TFs) to *cis*-regulatory elements plays an important role in cell-type specific gene expression in mammalian cells. This process is potentially guided by epigenetic states of the chromatin. Recent studies provide large amounts of genome-wide ChIP-seq data for both TF binding and histone modification loci, enabling large-scale analysis of spatial and regulatory interplay between TFs and epigenetic marks. In this paper, the authors report an integrative analysis of multiple public ChIP-seq and RNA-seq data sets, concerning 85 TFs, 9 histone modifications and 5 cell lines, to investigate the genome-wide localization correlations between transcription factor binding sites (TFBSs), histone modification patterns and transcription in human. This study reveals that genome-wide co-localization with histone modifications follow the same pattern for different TFs, and active histone marks typically adjoin TFBSs at a distance around 500 bp. TF occupancy at conserved sequences is found positively correlated with levels and bimodal pattern of active histone marks, and the bimodal and co-localized patterns track with higher gene expression. The correlation among histone modification patterns, TF occupancy and gene transcription suggests the existence of a possible regulatory mechanism that cells may implement to regulate transcription.

**Key words** epigenomics, transcription factor, systems biology, integrative approach

**DOI:** 10.3724/SP.J.1206.2012.00165

Transcription factors (TFs) are key regulators of gene expression. In mammalian cells, binding of TFs to *cis*-regulatory elements in a cell-type-specific way is to some degree responsible for the selective expression of genes. TFs with specific binding consensus sequences are found occupying only a small fraction of those patterned sequences, which cannot be explained merely by their sequence context, suggesting that epigenetic states may help direct recruitment of TFs to chromatin<sup>[1]</sup>. Yet, such regulatory mechanisms remain to be confirmed. Thus mining data related to epigenetic states and TF recruitments is key to understanding cell-type specific gene expression<sup>[2]</sup>.

Recent studies take advantage of genome-wide ChIP-chip or ChIP-seq experiments to investigate spatial correlations of TFs and histone marks in mammalian cells, and the subsequent expression changes of target genes. Lupien *et al.*<sup>[3]</sup> discovered that H3K4 methylation level correlates with cell-type-

specific recruitment of FOXA1, which translates this epigenetic signal into differential gene expression. In a similar study, distal FOXA2 and STAT1 binding sites are found symmetrically flanked by H3K4me1<sup>[4]</sup>. Also, GATA1 occupancy is positively correlated with H3K4me1 and negatively with H3K27me3<sup>[5]</sup>. Such spatial correlation of histone modifications (HMs) with TFs has been applied to TF occupancy prediction models in several studies as a very informative parameter<sup>[6-10]</sup>.

\*This work was supported by grants from National Basic Research Program of China (2009CB918801), and The National Natural Science Foundation of China (31171274).

\*\*Corresponding author.

SUN Zhi-Rong. Tel: 86-10-62772237, E-mail: sunzhr@mail.tsinghua.edu.cn

DAI Jun-Biao. Tel: 86-10-62796190, E-mail: jbdai@biomed.tsinghua.edu.cn

Received: April 1, 2012 Accepted: April 25, 2012

Refined positioning patterns of histone marks and their role in TF recruitment have attracted a great deal of attention as well. Hoffman *et al.*<sup>[11]</sup> reported three categories of H3K4 methylations distribution patterns, namely bimodal, monomodal and low-signal, around FOXA2-, PDX1-, and HNF4- occupied loci in two mouse tissues and suggested a flanking pair of H3K4me-marked nucleosomes may be responsible for the bimodal pattern. Two other studies revealed similar symmetric distributions of histone marks around RNAP II - and TAF1-marked transcription start sites (TSSs)<sup>[12]</sup> and H3K4me1/2 around p300-marked enhancers and Oct4 binding sites<sup>[8]</sup>.

However, previous findings on spatial correlations and patterns of histone modifications are limited to a few cell lines, transcription factors or histone marks. Plenty of large-scale ChIP-seq data are recently available<sup>[13]</sup> and remained to be mined to make more general discoveries. The aim of this study is to systematically analyze global spatial correlations of TFs and different histone modifications and to discover refined positioning patterns of histone modifications at TFBS, in hope of providing new insights into the link between epigenetic and transcriptional regulation of gene expression.

## 1 Materials and methods

### 1.1 Data sources

All ChIP-seq data of TF binding and histone modifications is from the ENCODE project<sup>[13]</sup>. Data in five cell lines (Gm12878, HeLa-S3, HepG2, HUVEC, and K562) from two TFBS datasets (*HAIB TFBS* and *Yale TFBS*) and two histone modification datasets (*UW Histone* and *Broad Histone*) are used. Descriptions of cell lines can be found on the ENCODE website (<http://genome.ucsc.edu/ENCODE/cellTypes.html>). All data is mapped to Human Genome Build 36 (hg18).

ENCODE RNA-seq data for gene transcription from the same five cell lines are downloaded from NCBI GEO. Accession numbers of used data are listed in Table S1 (Supplementary online, [http://www.pibb.ac.cn/cn/ch/common/view\\_abstract.aspx?file\\_no=20120165&flag=1](http://www.pibb.ac.cn/cn/ch/common/view_abstract.aspx?file_no=20120165&flag=1)).

### 1.2 Peak calling from ChIP-seq tag data

The SISSRs tool (v1.4) developed by NIH<sup>[14]</sup> is used to identify peaks from tag-alignment data, a processed format of original ChIP-seq data. Parameters are: Background model=yes; Genome length-

3137161264 bp; *E*-value-10; *P*-value-0.005; Scanning window size-2.

### 1.3 Measuring co-localization of TF-HM pairs

Co-localization of a TF with a histone modification (HM) is measured by a Boolean indicator of whether centers of the TF peak and HM peak are less than *L* away from each other. *L* is 50 bp for results shown in Figure 1.

Let *N*(*i*) be HM peak number found within  $\pm L$  of peak center of TFBS no. *i*, *t* be the total peak number of TFBSs, *h* be the total peak number of HM, *l* be the length of the genome, then HM peak density in the  $\pm L$  bp flanking region of all TFBSs for a particular TF (denoted as *D<sub>f</sub>*) is:

$$D_f = \frac{\sum_{i=1}^t N(i)}{2tL}$$

And background HM peak density (denoted as *D<sub>b</sub>*) is:

$$D_b = \frac{h}{l}$$

Thus, HM enrichment fold change (*FC*) is:

$$FC = \frac{D_f}{D_b} = \frac{l \sum_{i=1}^t N(i)}{2htL}$$

For TF-HM pairs with multiple replicates, the arithmetic mean *FC* of all replicate combinations is calculated. Similarity between TFs in terms of their co-localization patterns with different HM is indicated by Pearson correlation coefficient of *FC* scores described above.

### 1.4 Identification of TF clusters and multiple transcription factor loci (MTLs)

Pairwise overlap ratio of TF binding loci is defined as the geometric mean of the two TFs' overlapped fraction in DNA length. For TFs with multiple ChIP-seq replicates, the maximum overlap ratio is taken. TFs are clustered according to overlap ratios using Euclidean distance.

Multiple transcription factor loci (MTLs) are identified by scanning 500 bp long windows along the genome by a step of 50 bp for windows that contain peak center of all TF cluster members. Consecutive windows are combined into one locus.

### 1.5 TF occupancy in conserved binding sites

Binding consensus information of TFs are from two sources, SwissProt annotations<sup>[15]</sup> and UCSC TFBS Conserved Sites<sup>[16]</sup>. Consensus sites that overlap with any ChIP-seq binding signal of corresponding TFs are

defined as occupied, and the remaining are defined as unoccupied. Occupancy ratio of a particular TF is measured by the percentage of occupied sites in all annotated consensus sites.

### 1.6 Categorizing histone modification patterns around TF binding consensus

To categorize the pattern of a particular histone modification at a particular TF consensus site, three scoring functions are derived: a co-localization score  $S_c$  indicating overlap situation, a bimodal score  $S_b$  indicating symmetry and closeness, and a low-signal score  $S_l$  indicating distance.

$$S_c = \begin{cases} \frac{1}{D+d}, & \text{when overlapped} \\ -1, & \text{when not overlapped} \end{cases}$$

Where  $D$  (bp) is the distance between peaks of TF binding consensus site and histone modification,  $d$  is a pseudo distance in case of zero-denominator,  $d=1$  bp. Sites with positive  $S_c$  are categorized into the co-localization class (class C). Higher  $S_c$  indicates closer localization of histone modification to TF consensus.

$$S_l = \frac{\min(D_l, D_r)}{1000}$$

Where  $D_l$  (bp) and  $D_r$  (bp) stand for distances of closest histone modification signal at left and right of the TF consensus site, respectively.  $S_l$  indicates the degree of histone mark depletion. Consensus sites that are not flanked by histone modification within  $\pm 1000$  bp range, *i.e.* with  $S_l > 1$ , are categorized into the low-signal class (class L).

$$S_b = \frac{k}{(D_l^2 + D_r^2)(|D_l - D_r| + d)}$$

Where  $D_l$  and  $D_r$  are as described above,  $d=50$  bp is pseudo distance for denominator adjustment, and  $k=10^7$  bp<sup>2</sup> is a parameter to adjust scale of the score.  $S_b$  increases as distances of histone modification centers decreases and symmetry increases. Sites that do not belong to Class C or L, and at the same time with bimodal scores over 1 are categorized into the bimodal class (Class B).

Sites not categorized in any of these three classes form the monomodal class (class M).

### 1.7 Gene expression analysis

A TF binding consensus site within  $\pm 1000$  bp of any UCSC known gene are considered proximal<sup>[3]</sup>. Modification patterns of histone acetylation and methylation at these sites are categorized into four classes, as described before. Genes are then grouped

according to histone modification patterns proximal TFBSs.

For each of the five cell lines, Gm12878, HeLa-S3, HepG2, HUVEC and K562, we averaged RNA-seq transcription levels in four replicates from two samples derived from the nucleus. Significance test for transcription levels of different gene groups is described in the Statistics section.

### 1.8 Statistics

**1.8.1** Occupancy at consensus sites with or without histone marks. The problem of whether occupancy of TFBSs is different between consensus sites with or without histone marks is a problem of test for homogeneity of binomial proportion. Fisher's exact test is performed.

**1.8.2** Transcription levels of genes associated with different proximal histone modification patterns at TFBSs. Under the assumption that transcription levels for both groups (pattern A-associated and pattern B-associated) of genes follow Gaussian distribution, test of average transcription level is a Behrens-Fisher problem. Thus, two-sample *t*-test is performed to determine the significance.

### 1.9 Chromatin immunoprecipitation

Chromatin immunoprecipitation (ChIP) assay was carried out using standard ChIP protocol. Briefly, K562 cells were fixed with 1% formaldehyde for 10 min, washed twice with ice-cold phosphate buffer saline (PBS). The cells were lysed with RIPA buffer (50 mmol/L Tris-HCl pH 7.4, 150 mmol/L NaCl, 1% NP40, 0.25% Na-deoxycholate, 1 mmol/L PMSF, 1× Roche complete protease inhibitor) and sonicated 30 times (10 s "on", 20 s "off" in each cycle) with the ultrasonicator (Sonic VC750) at 15% power. Sonicated chromatin solution was precleared by protein IgG magnetic beads (Invitrogen 100.01D) and immunoprecipitation was performed overnight with anti-acetyl histone H3K27 (Millipore 06-866) or normal rabbit IgG and magnetic beads. After incubation, the beads were washed with RIPA buffer, and reverse-crosslinked at 65°C for 6 h. DNA extracted from immunoprecipitates was subjected to PCR analysis, then quantified by real-time PCR using Sybr-green system (BioRAD IQ5). The sequences of primers were listed in Table S2 (Supplementary online, [http://www.pibb.ac.cn/cn/ch/common/view\\_abstract.aspx?file\\_no=20120165&flag=1](http://www.pibb.ac.cn/cn/ch/common/view_abstract.aspx?file_no=20120165&flag=1)). DNA directly purified from sonicated chromatin solution was used as a positive control (input).

## 2 Results

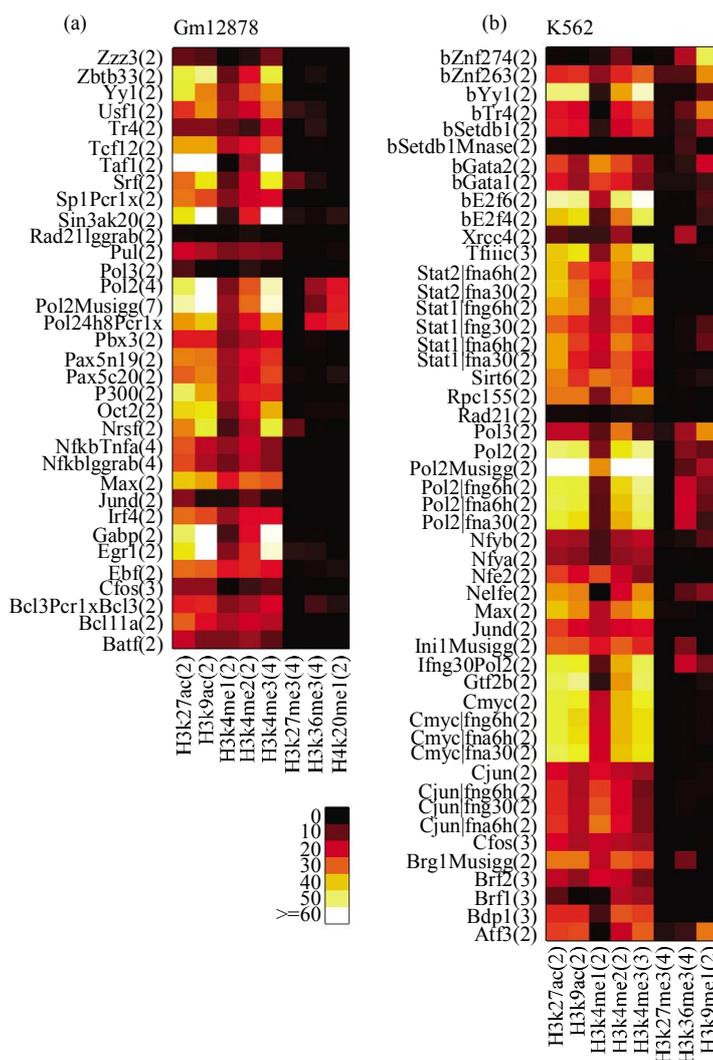
### 2.1 Levels of genome-wide co-localization with histone modifications follow the same pattern for different TFs

By exploring over 400 published genome-wide ChIP-seq profiles from the ENCODE project<sup>[13]</sup>, we investigated the spatial correlation between 9 kinds of HMs and binding loci of 105 TFs (summarized in

Table 1). For each of the 5 distinct human cell lines, namely Gm12878, HeLa-S3, HepG2, HUVEC and K562, we generated a co-localization heat map illustrating degree of co-localization between a specific HM and a given TF. The degree of co-localization is measured by fold change (*FC*) of TF and HM peak signals within 50 bp of each other relative to the background (see **Materials and methods**). As shown in Figure 1 and Figure S1(Supplementary online, <http://>

**Table 1 Summary of ChIP-seq data used in co-localization profiling**

Cell line	Gm12878	HeLa-S3	HepG2	HUVEC	K562
Number of TFs	34	28	25	3	50
Number of total TFBS replicates	80	64	55	7	105
Number of HMs	8	3	7	9	8
Number of total HM replicates	22	6	17	24	23



**Fig. 1 TF-HM co-localization heat maps of Gm12878 and K562 cell lines**

Peak centers of ChIP binding signals within 50 bp of each other are counted as co-occurring signal pairs; co-localization score of each TF-HM pair is the fold difference between observed co-occurring signal pairs divided by the number expected from a random distribution. Numbers of parallel replicates of TF/HM are labeled in the brackets. For TF-HM pairs with multiple replicates, average co-localization scores are mapped.

[http://www.pibb.ac.cn/cn/ch/common/view\\_abstract.aspx?file\\_no=20120165&flag=1](http://www.pibb.ac.cn/cn/ch/common/view_abstract.aspx?file_no=20120165&flag=1)), most TFs behave alike in terms of their co-occurrence with histone marks.

One possible scenario is that TFs scored similarly on co-occurring fold change with HMs may themselves cluster together along the genome. A previous analysis in mESC revealed two co-binding TF clusters, each consisting of four TFs<sup>[17]</sup>. To answer the question of whether TF clusters lead to similar co-localization patterns, we first calculated pairwise ratios of overlapped binding length of TFs and found a few groups of TFs with noticeably high overlapping ratios. All nine cell-line-specific TF clusters with over 30% overlap are listed in Table S3 (Supplementary online,

[http://www.pibb.ac.cn/cn/ch/common/view\\_abstract.aspx?file\\_no=20120165&flag=1](http://www.pibb.ac.cn/cn/ch/common/view_abstract.aspx?file_no=20120165&flag=1)). Binding spots of those clusters, namely multiple transcription factor loci (MTLs), were scanned using a binning method (see **Materials and methods**). We measured similarity of co-localization fold change scores by Pearson correlation coefficient for each pair of TFs, and compared values of this indicator before and after removing members of co-occurring TF clusters (Table 2). Eliminating TF cluster members does not significantly lower the average pairwise linear correlation of FC scores, indicating that the presence of TF clusters is not the main cause of similar co-localization pattern of TFs with different HMs.

**Table 2 Linear correlation of co-localization fold change vectors between different TFs in three cell lines<sup>b</sup>**

Cell line	Pearson correlation coefficient ( $\bar{x} \pm s$ ) for all TFs	Pearson correlation coefficient ( $\bar{x} \pm s$ ) after eliminating members of TF clusters
Gm12878	0.8465 $\pm$ 0.1346	0.8190 $\pm$ 0.1462
HepG2	0.8864 $\pm$ 0.1121	0.8719 $\pm$ 0.1217
K562	0.6263 $\pm$ 0.3778	0.5846 $\pm$ 0.3880

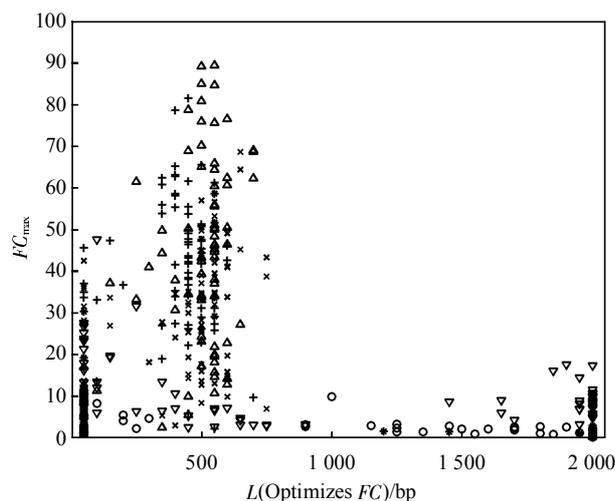
<sup>b</sup> Data sizes of the other two cell lines, HeLa-S3 and HUVEC, are too small to calculate correlation coefficients.

Active histone marks including acetylations on H3K27 and H3K9, and H3K4 methylations are generally enriched in binding loci of most TFs across different cell lines. Overall levels of fold co-localization separates active histone mark H3K4me2, 3, H3K9ac, and H3K27ac from repressive histone mark H3K27me3, H3K36me3, H4K20me1<sup>[1, 18-19]</sup>. The other two active histone marks, H3K4me1 and H3K9me1, which are mutually exclusive to H3K4me2/3 and H3K9ac respectively, are seen less enriched in TFBS. By using combinations of H3K4 methylations (H3K4me1/2/3) and H3K9ac/me1 (Figure S2, Supplementary online, [http://www.pibb.ac.cn/cn/ch/common/view\\_abstract.aspx?file\\_no=20120165&flag=1](http://www.pibb.ac.cn/cn/ch/common/view_abstract.aspx?file_no=20120165&flag=1)), we can see a clear separation between active and repressive histone marks in terms of co-localization with various TFs.

## 2.2 Active histone marks adjoin transcription factor binding loci at a distance around 500 bp

While measuring co-localization levels, an arbitrary distance cutoff  $L$  is used to distinguish peaks regarded as co-localized from those apart. In the above fold change heat maps (Figure 1 and S1 (Supplementary online, [http://www.pibb.ac.cn/cn/ch/common/view\\_abstract.aspx?file\\_no=20120165&flag=1](http://www.pibb.ac.cn/cn/ch/common/view_abstract.aspx?file_no=20120165&flag=1))),  $L$  equals 50 bp. To explore the positioning pattern

in a wider range, we scanned this parameter  $L$  from 0 to 2000 bp to optimize co-localization fold change (FC) for each pair of TF and HM, and found that  $L$  values at optimal FCs for active histone marks are typically around 500 bp long, as shown in Figure 2.



**Fig. 2 Co-localization fold change (FC) scores of TF-histone modification pairs are sensitive to distance cutoff  $L$  and reaches maximum at around 500 bp**

Optimized FC scores are plotted against corresponding  $L$  values for all 55 TFs in Gm12878 cell line. Typically maximum co-localization fold change scores occur when  $L$  falls in the 400~600 bp range, indicating that histone modification signal is abundant at this distance apart from TF binding loci. + : H3k27ac; o : H3k27me3; \* : H3k36me3; x : H3k4me1/2/3; Δ : H3k9ac/me1; ▽ : H4k20me1.

Most TFs plotted here reach optimal co-localization fold changes at around 400 to 600 bp (Table 3). We can conclude from Table 3 that active histone marks, H3K9ac/me1, H3K27ac, and H3K4me1/2/3, are

enriched at ~500 bp around various TFs, while repressive histone marks H3K27me3, H3K36me3, and H4K20me1 are not.

**Table 3 Co-localization distance cutoff  $L$  that maximizes co-localization fold change scores for different histone modifications**

Cell line		H3K27ac	H3K4me1/2/3	H3K9ac/me1	H3K27me3	H3K36me3	H4K20me1
Gm12878	$\bar{x} \pm s$	416 ± 142	472 ± 164	505 ± 116	588 ± 767	1032 ± 970	973 ± 862
	[400, 600]	78%	78%	84%	0%	0%	10%
HeLa-S3	$\bar{x} \pm s$	467 ± 592	N/A	N/A	1213 ± 903	290 ± 200	N/A
	[400, 600]	13%	N/A	N/A	3%	44%	N/A
HepG2	$\bar{x} \pm s$	482 ± 60	489 ± 134	608 ± 95	815 ± 848	1396 ± 909	1647 ± 665
	[400, 600]	91%	82%	60%	2%	0%	18%
HUVEC	$\bar{x} \pm s$	1443 ± 952	400 ± 65	464 ± 56	429 ± 39	1157 ± 1036	1621 ± 648
	[400, 600]	57%	100%	86%	0%	0%	14%
K562	$\bar{x} \pm s$	367 ± 148	397 ± 186	476 ± 197	635 ± 831	1062 ± 956	640 ± 725
	[400, 600]	66%	59%	68%	1%	1%	30%

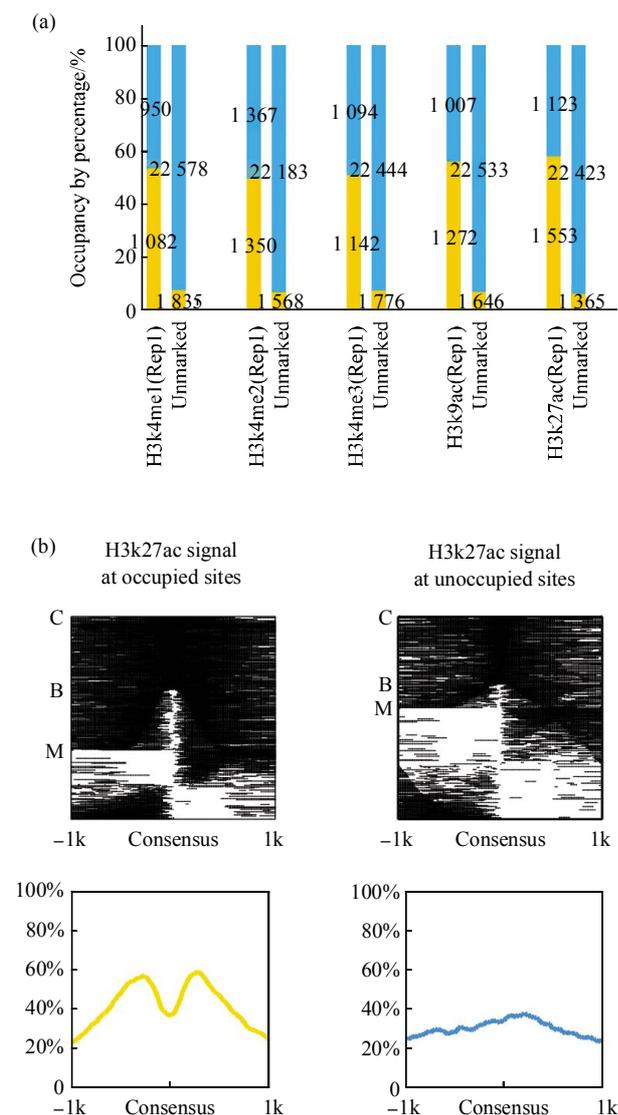
### 2.3 Binding preference of TFs to their consensus sites is positively correlated with active histone marks

In order to find out whether HMs have an impact on the selective binding of TFs to their consensus sites, we focus on the 16 TFs whose conserved binding sequence are annotated in SwissProt<sup>[15]</sup> or in UCSC TFBS Sites database<sup>[16]</sup> (listed in Table S4, online, [http://www.pibb.ac.cn/cn/ch/common/view\\_abstract.aspx?file\\_no=20120165&flag=1](http://www.pibb.ac.cn/cn/ch/common/view_abstract.aspx?file_no=20120165&flag=1)). Each consensus site is labeled either "occupied" or "unoccupied" according to whether a ChIP-seq binding signal is detected overlapping the site, and "marked" or "unmarked" by a certain HM according to whether a ChIP-seq modification signal is detected in a ±1 000 bp flanking region. The window length 1 000 bp is chosen in consideration that active histone marks are most abundant at ~500 bp around TF binding loci. Then we analyze overall occupancy (percentage of occupied sites) in marked and unmarked groups to see if there is a significant difference. Most of the 16 TFs analyzed are significantly more likely (Fisher's exact test  $P < 0.05$ ) to bind consensus sites with active histone marks (H3K27ac, H3K9ac/me1, and H3K4 methylation) than to unmarked sites, except for two TFs, SP1 and JunD (results not shown). Figure 3a shows occupancy

differences at c-Jun consensus sites with or without active histone marks as an example.

### 2.4 Bimodal patterned HMs are enriched in occupied TF binding consensus sites

It has been reported that transcription factor binding loci display different histone methylation patterns<sup>[11]</sup>. By scoring distance and symmetry of histone marks distribution around TF binding loci, we categorize positioning patterns of HMs into four classes: co-localization (C), monomodal (M), bimodal (B), and low-signal (L) (see **Materials and methods**). Class C are sites that overlaps signal of a particular histone mark; class B are sites with symmetric signals at both sides; class M are sites with signals at one side; class L are sites not flanked by HM signal ±1 000 bp. Under such classification, the proportion of bimodal modification patterns in TF-occupied motifs is much higher than (typically twice as much as) that of unoccupied ones for all 16 TFs examined (Table 4). As shown in Figure 3b, spatial distribution of H3K27ac differs between c-Jun occupied and unoccupied consensus sites. Higher proportions of bimodal-patterned H3K27ac signals are associated occupied c-Jun binding motifs (Figure 3B, upper panel), and overall distribution of this modification relative to TFBSs displays a two-peak shape as well.



**Fig. 3 Binding preference of c-Jun is associated with level and spatial distribution of histone modifications**

(a) Binding preference of c-Jun to its consensus sites is correlated with levels of active histone modifications in K562 cell line. Consensus sites show much higher occupancy when marked by H3K4me, H3K9ac, and/or H3K27ac. ■: Occupied; ■: Unoccupied. (b) Spatial distribution of histone acetylation differs between occupied (yellow) and unoccupied (blue) consensus sites. 'C', 'B', and 'M' labels on the left represent three of the four classes of histone modification distribution patterns, where "C" stands for "co-localization", "B" for "bimodal", and "M" for "monomodal". Higher proportions of bimodal-patterned histone acetylation are found in occupied c-Jun motifs. Individual histone marks  $\pm 1000$  bp relative to consensus sites are sorted according to spatial patterns and shown in a sparsity map (upper panel), and the averaged modification signal at each position is plotted (lower panel).

**Table 4 Fractions of active histone mark patterns vs. TF occupancy at annotated binding consensus sites**

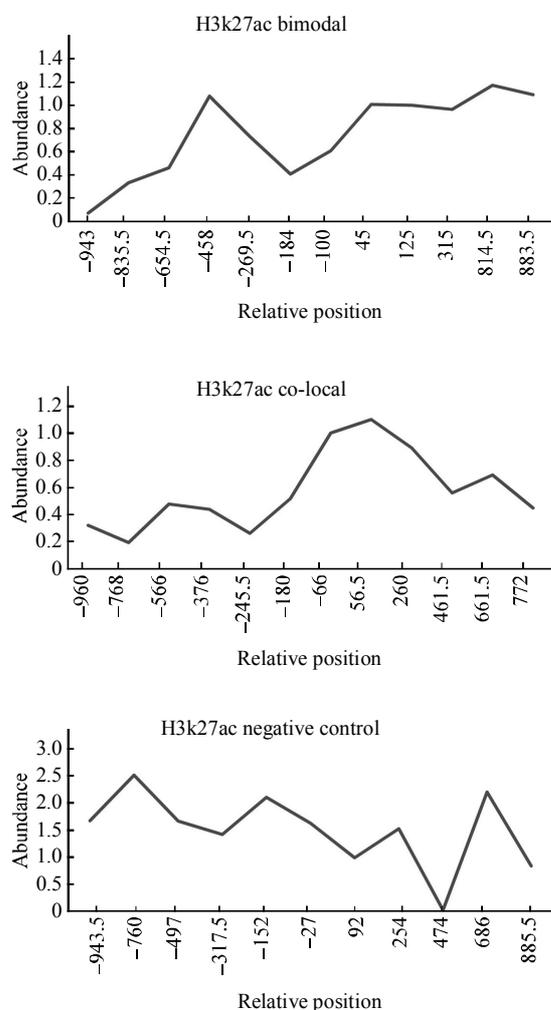
TF	Percentage of bimodal pattern in occupied loci/%	Percentage of bimodal pattern in unoccupied loci/%
ATF-3	39	27
E2F4	26	10
E2F6	20	9
c-Fos	28	13
c-Jun	26	14
c-Myc	23	8
Egr-1	25	17
JunD	26	15
Max	26	13
NF-E2	21	13
POU2F2 (Oct-2)	24	8
p300	37	17
Pax-5	24	13
SP1	31	12
STAT1	28	15
STAT2	27	11

Low-signal modifications are not considered in calculating fraction.

## 2.5 Experiment supports the histone modification patterns

To confirm the existence of bimodal distribution and co-localization in the genome as shown from the above analyses, we performed chromatin immunoprecipitation (ChIP) experiments using one of the cell lines (K562) as an example. We focused on c-Jun, which showed significant pattern with H3K27ac. Three sites representing bimodal, co-localization and low-signal respectively were analyzed using a serial of PCR primers spanning from  $-1000$  bp to  $+1000$  bp relative to the TF binding sites (Figure 4, Table S2, Supplementary online, [http://www.pibb.ac.cn/cn/ch/common/view\\_abstract.aspx?file\\_no=20120165&flag=1](http://www.pibb.ac.cn/cn/ch/common/view_abstract.aspx?file_no=20120165&flag=1)). As we expected, we didn't observe significant pattern across the low-signal locus. On the other hand, we detected strong signal at the co-localization patterned loci selected, agreeing with what we observed. At a loci classified as bimodal, we found this modification was specifically missing from the DNA region close to c-Jun binding site, with one peak around 400 bp upstream of the binding site and another peak about 200 bp downstream, confirming the bimodality at this region. Our experimental data

suggests that at least in the regions we tested in K562 cells, both bimodal and co-localization as categorized in our bioinformatics analysis could be verified.



**Fig. 4** Experimental validation of bimodal, co-localization and low-signal patterns of H3K27ac ChIP-seq data at c-Jun binding sites

For each modification pattern, we take one 2 000 bp long locus centering a c-Jun binding consensus site as an example. Note that the center of each site here is not valley or peak center in the original genome-wide ChIP-seq data, but the TF consensus site. Abundance of PCR products from multiple primers are plotted to represent signal of H3K27ac, and is not comparable across the three sites.

## 2.6 Bimodal and co-localized patterns of active histone marks track with higher gene expression

Under the assumption that differential TF recruitment translates HM patterns into gene expression differences, the bimodal pattern related to

high TF occupancy is predicted to affect transcription levels. Histone mark patterns of proximal(TSS  $\pm$  1 kb)<sup>[3]</sup> TF consensus sites were categorized into four classes as previously described. Transcription levels in the corresponding cell lines were acquired from RNA-seq experiments published also in the ENCODE project (Table S1, Supplementary online, [http://www.pibb.ac.cn/cn/ch/common/view\\_abstract.aspx?file\\_no=20120165&flag=1](http://www.pibb.ac.cn/cn/ch/common/view_abstract.aspx?file_no=20120165&flag=1)). Genes associated with bimodal or co-localized patterned active histone marks proximal TFBSs are found with significantly higher transcription than genes associated monomodal patterns. Table 5 summarizes significance levels (measured by *P*-value in two-sample *t*-test) of gene expression differences between different modification patterns.

According to Table 5, HM patterns at consensus sites of 12/16 TFs have significant( $P < 0.05$ ) correlation with transcription levels. Genes associated with bimodal pattern (class B) acetylation/methylation are transcribed significantly higher than those associated with monomodal pattern (class M). For most TFs, as shown in Figure S3 (Supplementary online, [http://www.pibb.ac.cn/cn/ch/common/view\\_abstract.aspx?file\\_no=20120165&flag=1](http://www.pibb.ac.cn/cn/ch/common/view_abstract.aspx?file_no=20120165&flag=1)), expression distributions of genes associated with class B or class C (co-localization) methylation at promoter consensus sites are almost entirely overlapping, while class M sites and class L sites associated genes have lower expressions.

## 3 Discussion

### 3.1 Systematic approaches help discover global associations between transcription factors and epigenetic marks

Integration of multiple one-dimensional ChIP-seq maps of diverse types is providing novel insights into the relations among histone modifications and other functional components of the genome. Although findings of such studies may be largely consistent with prior discoveries, developing approaches to integrate increasing types of data into coherent knowledge is crucial to fully exploit them<sup>[20]</sup>. There are several approaches to integrative analysis: data complexity reduction, unsupervised integration and supervised integration<sup>[21]</sup>. In this study, we first reduced data complexity by discretizing ChIP-seq data and intersecting histone marked regions with TF binding consensus. Then unsupervised analysis of histone mark positioning patterns flanking TFBSs generated the

**Table 5 Significance of transcription level difference for gene groups with different methylation or acetylation patterns**

TF	Cell line	Acetylation		Methylation	
		B vs. C	B vs. M	B vs. C	B vs. M
ATF-3	K562	NS	$1.09 \times 10^{-4}$	$3.32 \times 10^{-2}$	$5.35 \times 10^{-6}$
E2F4	K562	NS	$6.23 \times 10^{-4}$	$9.46 \times 10^{-1}$	$1.70 \times 10^{-5}$
E2F6	K562	NS	$4.05 \times 10^{-4}$	$4.61 \times 10^{-1}$	$2.42 \times 10^{-6}$
c-Fos	Gm12878	NS	NS	$3.34 \times 10^{-1}$	NS
	K562	$5.91 \times 10^{-3}$	NS	$1.17 \times 10^{-1}$	$2.09 \times 10^{-3}$
c-Jun	HUVEC	NS	$2.54 \times 10^{-2}$	$8.14 \times 10^{-1}$	$1.91 \times 10^{-2}$
	K562	NS	$4.26 \times 10^{-3}$	$2.88 \times 10^{-1}$	$1.77 \times 10^{-5}$
c-myc	K562	NS	NS	$3.24 \times 10^{-1}$	NS
Egr-1	Gm12878	NS	$8.67 \times 10^{-3}$	$4.64 \times 10^{-1}$	$1.62 \times 10^{-3}$
JunD	Gm12878	NS	NS	$2.42 \times 10^{-1}$	$4.67 \times 10^{-2}$
	K562	$3.77 \times 10^{-3}$	NS	$1.16 \times 10^{-2}$	$6.91 \times 10^{-3}$
Max	Gm12878	NS	$3.56 \times 10^{-2}$	$3.06 \times 10^{-1}$	$6.60 \times 10^{-3}$
	HUVEC	NS	NS	$2.45 \times 10^{-1}$	$1.36 \times 10^{-6}$
	K562	NS	$4.22 \times 10^{-5}$	$4.66 \times 10^{-2}$	$1.54 \times 10^{-6}$
NF-E2	K562	NS	$3.23 \times 10^{-1}$	$2.70 \times 10^{-1}$	$1.94 \times 10^{-2}$
POU2F2 (Oct-2)	Gm12878	NS	$2.87 \times 10^{-2}$	$7.76 \times 10^{-2}$	$3.29 \times 10^{-1}$
p300	Gm12878	$3.20 \times 10^{-1}$	$3.14 \times 10^{-1}$	$3.15 \times 10^{-1}$	$5.56 \times 10^{-5}$
	HepG2	$2.92 \times 10^{-1}$	$6.06 \times 10^{-1}$	$3.18 \times 10^{-1}$	$2.81 \times 10^{-1}$
Pax-5	Gm12878	$1.64 \times 10^{-1}$	$1.07 \times 10^{-3}$	$1.76 \times 10^{-1}$	$3.81 \times 10^{-5}$
SP1	Gm12878	$7.83 \times 10^{-1}$	$2.00 \times 10^{-4}$	$7.77 \times 10^{-1}$	$3.72 \times 10^{-5}$
STAT1	K562	$3.35 \times 10^{-3}$	$7.81 \times 10^{-3}$	$1.35 \times 10^{-1}$	$4.98 \times 10^{-4}$
STAT2	K562	$1.16 \times 10^{-1}$	$6.01 \times 10^{-3}$	$1.61 \times 10^{-1}$	$4.64 \times 10^{-2}$

B: Bimodal; C: Co-localization; M: Monomodal; NS: Not significant ( $P > 0.05$ ).

hypothesis that a bimodal feature of histone modification is associated with TF recruitment. And to test the hypothesis that bimodal-patterned histone marked may link gene expression and TF recruitment, a supervised integration approach is carried out. We examined the correlation between transcription levels and four pre-defined histone modification patterns, and the results support the hypothesis.

### 3.2 Transcription factor binding, histone modification patterns and gene expression

It has been known for years that binding of TFs to the promoter regions is one mechanism for gene activation. Recent study also revealed the importance of histone modification in transcriptional regulation<sup>[1]</sup>. For example, both H3K4me1 and H3K4me3 were associated with actively transcribed genes<sup>[18]</sup>. However, until very recently the relationship between the binding of TFs and specific histone modifications began to be appreciated<sup>[11, 22]</sup>. In this paper, after detailed analysis of all available profiles in multiple cell lines, we provided

evidence that the collaborative operation between TF binding and histone modifications profiles may be a general phenomenon which cells used to regulate gene transcription.

It is still unclear how cells establish the pattern between TF binding and histone modifications and how it leads to the regulation of gene expression. One interesting finding is that the distance from the TF binding site to the peak of histone modification is around 500 bp, which is about two to three nucleosomes away<sup>[23]</sup>. In addition, as shown previously<sup>[11]</sup>, in bimodal loci TFs bind to the nucleosome-free region, followed by two flanking nucleosomes. Therefore one possible hypothesis is that the binding of TF may set the position for nucleosome formation and recruit the histone-modifying enzymes to the flanking nucleosomes to form the bimodal pattern<sup>[24]</sup>. The transcriptional machinery may recognize both TFs and histone modification pattern and subsequently leads to the activation of gene expression.

**Acknowledgments** We thank the ENCODE Project, supported by NHGRI, for making data available (in particular, Michael Snyder, Mark Gerstein and Sherman Weissman at Yale University; Peggy Farnham at UC Davis; Kevin Struhl at Harvard; the Broad Institute; the Bradley E. Bernstein lab at the Massachusetts General Hospital/Harvard Medical School; UW ENCODE group; and the Myers Lab at the HudsonAlpha Institute for Biotechnology).

### References

- [1] Berger S L. The complex language of chromatin regulation during transcription. *Nature*, 2007, **447**(7143): 407–412
- [2] Eeckhoutte J, Metivier R, Salbert G. Defining specificity of transcription factor regulatory activities. *J Cell Science*, 2009, **122**(22): 4027–4034
- [3] Lupien M, Eeckhoutte J, Meyer C A, *et al.* FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell*, 2008, **132**(6): 958–970
- [4] Robertson A G, Bilenky M, Tam A, *et al.* Genome-wide relationship between histone H3 lysine 4 mono- and tri-methylation and transcription factor binding. *Genome Research*, 2008, **18** (12): 1906–1917
- [5] Cheng Y, Wu W S, Kumar S A, *et al.* Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications, and mRNA expression. *Genome Research*, 2009, **19**(12): 2172–2184
- [6] Joseph R, Orlov Y L, Huss M, *et al.* Integrative model of genomic factors for determining binding site selection by estrogen receptor-alpha. *Mol Syst Biol*, 2010, **6**: 456
- [7] Ramsey S A, Knijnenburg T A, Kennedy K A, *et al.* Genome-wide histone acetylation data improve prediction of mammalian transcription factor binding sites. *Bioinformatics*, 2010, **26** (17): 2071–2075
- [8] Won K J, Ren B, Wang W. Genome-wide prediction of transcription factor binding sites using an integrated model. *Genome Biol*, 2010, **11**(1): R7
- [9] Pique-Regi R, Degner J F, Pai A A, *et al.* Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research*, 2011, **21**(3): 447–455
- [10] Ernst J, Plasterer H L, Simon I, *et al.* Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Research*, 2010, **20**(4): 526–536
- [11] Hoffman B G, Robertson G, Zavaglia B, *et al.* Locus co-occupancy, nucleosome positioning, and H3K4me1 regulate the functionality of FOXA2-, HNF4A-, and PDX1-bound loci in islets and liver. *Genome Research*, 2010, **20**(8): 1037–1051
- [12] Heintzman N D, Stuart R K, Hon G, *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*, 2007, **39**(3): 311–318
- [13] Birney E, Stamatoyannopoulos J A, Dutta A, *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 2007, **447**(7146): 799–816
- [14] Jothi R, Cuddapah S, Barski A, *et al.* Genome-wide identification of *in vivo* protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res*, 2008, **36**(16): 5221–5231
- [15] Gasteiger E, Jain E, Bairoch A, *et al.* Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics*, 2009, **10**: 136
- [16] Fujita P A, Rhead B, Zweig A S, *et al.* The UCSC Genome Browser database: update 2011. *Nucleic Acids Res*, 2011, **39**: D876–D882
- [17] Chen X, Xu H, Yuan P, *et al.* Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 2008, **133**(6): 1106–1117
- [18] Li B, Carey M, Workman J L. The role of chromatin during transcription. *Cell*, 2007, **128**(4): 707–719
- [19] Zhao K J, Wang Z B, Zang C Z, *et al.* Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genetics*, 2008, **40**(7): 897–903
- [20] Zhou V W, Goren A, Bernstein B E. Charting histone modifications and the functional organization of mammalian genomes. *Nature Reviews Genetics*, 2011, **12**(1): 7–18
- [21] Hawkins R D, Hon G C, Ren B. Next-generation genomics: an integrative approach. *Nature Reviews Genetics*, 2010, **11**(7): 476–486
- [22] Barrera L O, Ren B. The transcriptional regulatory code of eukaryotic cells insights from genome-wide analysis of chromatin organization and transcription factor binding. *Curr Opin Cell Biol*, 2006, **18**(3): 291–298
- [23] Kato M, Onishi Y, Wada-Kiyama Y, *et al.* Dinucleosome DNA of human K562 cells: Experimental and computational characterizations. *J Mol Biol*, 2003, **332**(1): 111–125
- [24] Ren B, Barrera L O. The transcriptional regulatory code of eukaryotic cells insights from genome-wide analysis of chromatin organization and transcription factor binding. *Curr Opin Cell Biol*, 2006, **18**(3): 291–298

# 通过系统分析揭示组蛋白修饰模式 与转录因子结合之间的关联\*

李雯 李文婷 曹书娟 余文祥 李文杰 戴俊彪\*\* 孙之荣\*\*

(清华大学生命科学学院, 生物信息学教育部重点实验室, 生物膜与膜工程国家重点实验室, 北京 100084)

**摘要** 转录因子对顺势调控元件的选择性结合, 在哺乳动物细胞类型特异的基因表达中扮演重要的角色. 这个过程受到染色质表观遗传状态的潜在调控. 近期, 染色质免疫共沉淀结合测序的研究提供了大量泛基因组水平的数据, 阐述转录因子结合以及组蛋白修饰的位点, 这为系统地研究转录因子和表观遗传标记之间的空间及调控关系提供了基础. 该研究对公共数据库中的染色质免疫共沉淀结合测序数据进行整合分析, 涉及 5 个细胞系中的 85 种转录因子、9 种组蛋白修饰, 目的是研究转录因子结合位点与组蛋白修饰模式以及基因表达在泛基因组水平上的关联. 作者发现, 不同转录因子与组蛋白修饰的共定位模式高度一致, 并且组蛋白修饰在距离转录因子结合位点约 500 碱基对的位置富集. 作者还发现, 转录因子结合位点的占有率与活性组蛋白修饰的水平和双峰模式正相关, 并且启动子区域组蛋白修饰的双峰和共定位模式和基因的高转录水平相一致. 组蛋白修饰模式、转录因子结合位点的占有率与基因转录之间的关联暗示了细胞可能利用的基因表达调控机制.

**关键词** 表观遗传组, 转录因子, 系统生物学, 整合方法

**学科分类号** Q756, Q786

**DOI:** 10.3724/SP.J.1206.2012.00165

\* 国家重点基础研究发展计划(973)(2009CB918801), 国家自然科学基金(31171274)资助项目.

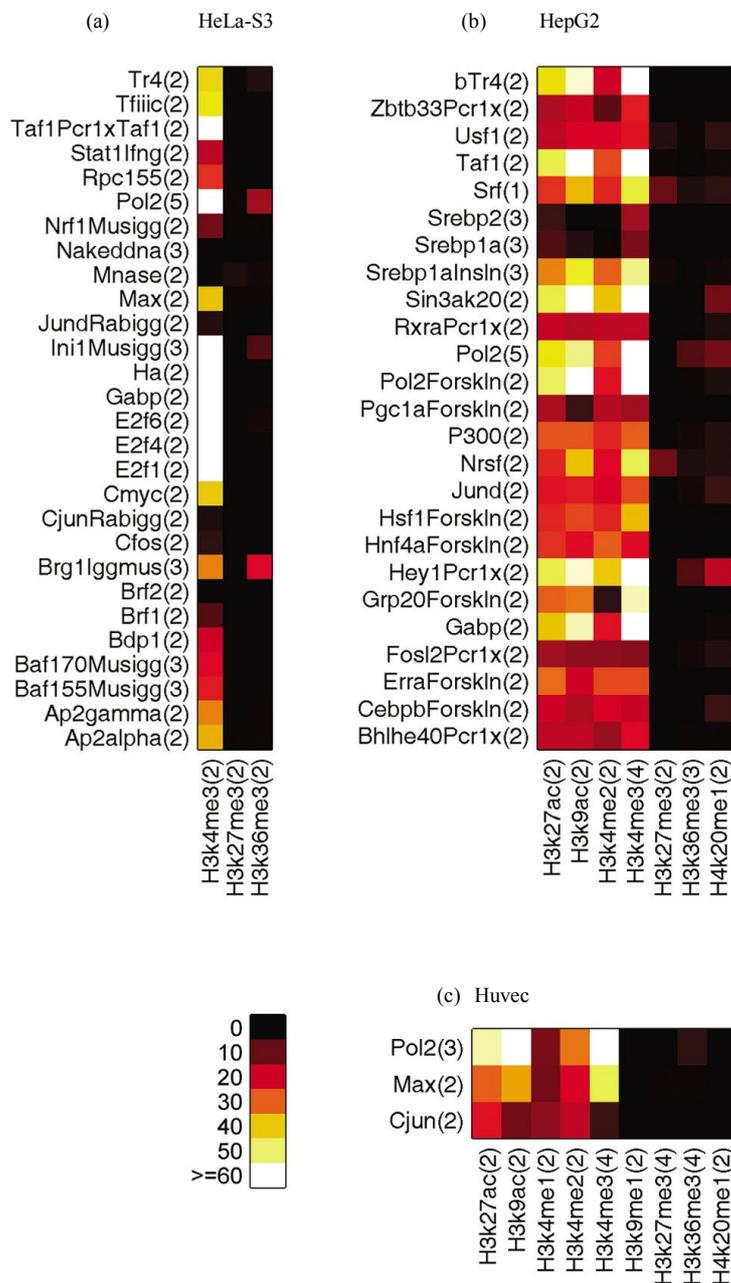
\*\* 通讯联系人.

孙之荣. Tel: 010-62772237, E-mail: sunzhr@mail.tsinghua.edu.cn

戴俊彪. Tel: 010-62796190, E-mail: jbdai@biomed.tsinghua.edu.cn

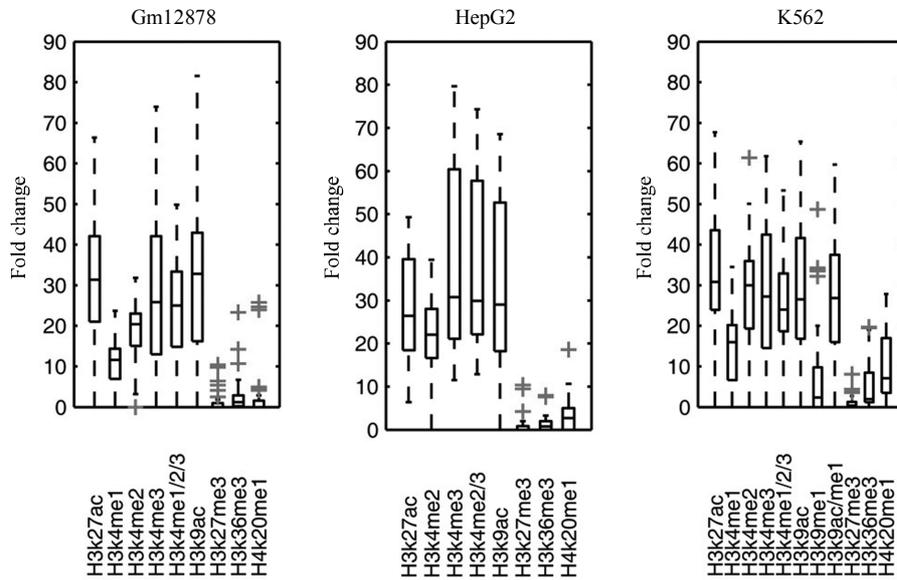
收稿日期: 2012-04-01, 接受日期: 2012-04-25

# 附录



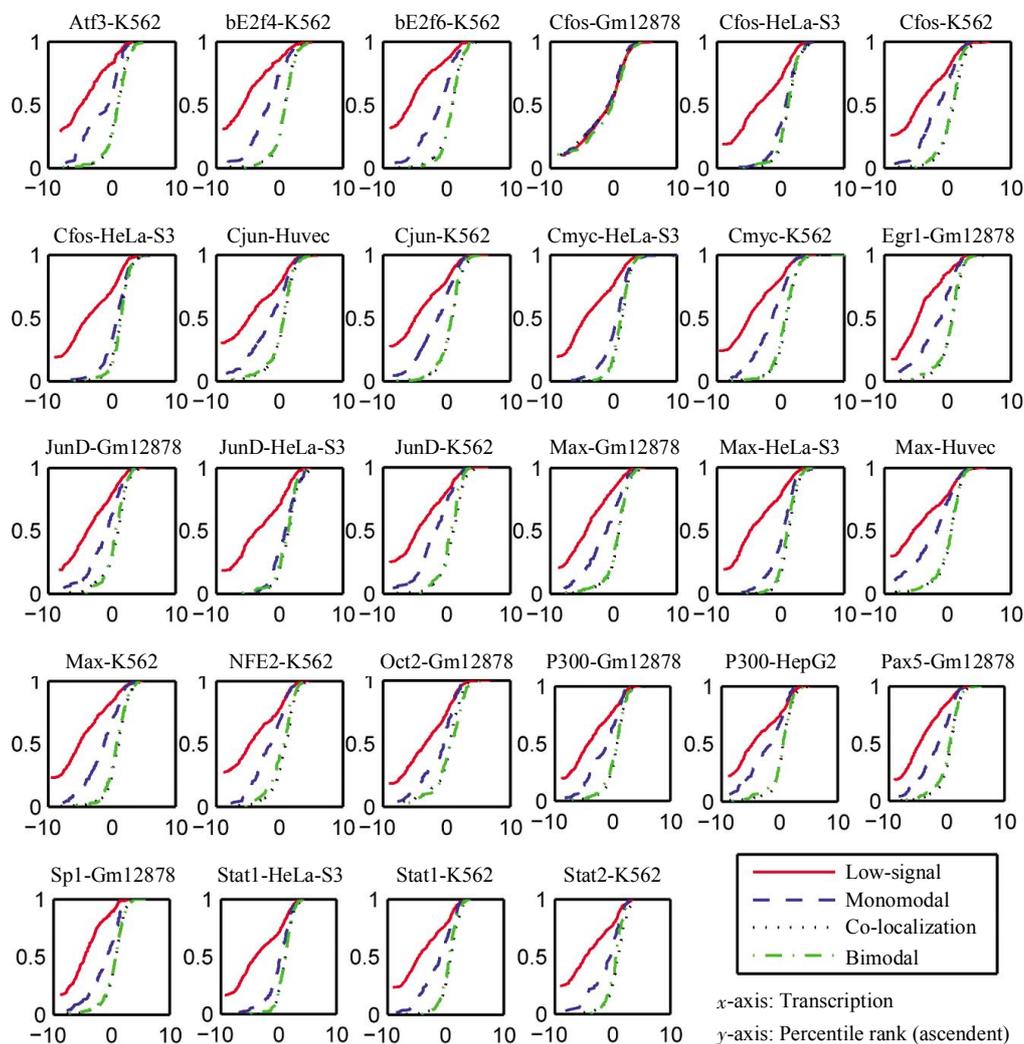
**Fig. S1 Heat maps of TF-histone modification co-localization profiles in (a) HeLa-S3, (b) HepG2, and (c) HUVEC**

Peak centers of ChIP binding signals within 50 bp of each other are counted as co-occurring signal pairs; co-localization score of each TF-HM pair is the fold difference between observed co-occurring signal pairs divided by the number expected from a random distribution. Parallel replicates of TF/HM are labeled in the brackets. For TF-HM pairs with multiple replicates, average co-localization scores are mapped.



**Fig. S2 Box-whisker plots of co-localization FC scores of mutually active and repressive histone marks or combinations with all TFs analyzed, in Gm12878, HepG2 and K562**

Active histone marks such as H3K4me3, H3K9ac, and H3K27ac show higher average FC, and their mutually exclusive counterparts H3K4me1/2, and H3K9me1 are less co-localized with most TFs. Combinations of mutually exclusive active histone marks (H3K4me1/2/3 and H3K9ac/me1) can be represented by the members with higher FC scores (H3K4me3 and H3K9ac) in terms of FC distribution. Repressive histone marks such as H3K27me3, H3K36me3 and H4K20me1 are much less co-localized with TFs.



**Fig. S3 Transcription level distribution of genes associated with different histone modification patterns at promoter TF binding consensus**

Percentile rank of consensus sites ( $y$  axis) is plotted against  $\log(e)$  transformed transcription value ( $x$  axis). Four classes of active histone marks are in different colors: Bimodal (green), Co-localization (black), Monomodal (blue), and Low-signal (red).

**Table S1 RNA-seq transcription data used in this paper**

GEO accession	Description
GSM767853	LICR_RnaSeq_GM12878_nucleus_longNonPolyA
GSM765386	LICR_RnaSeq_GM12878_nucleus_longPolyA
GSM767848	LICR_RnaSeq_HeLa-S3_nucleus_longNonPolyA
GSM765403	LICR_RnaSeq_HeLa-S3_nucleus_longPolyA
GSM767850	LICR_RnaSeq_HepG2_nucleus_longNonPolyA
GSM758568	LICR_RnaSeq_HepG2_nucleus_longPolyA
GSM767857	LICR_RnaSeq_HUVEC_nucleus_longNonPolyA
GSM758565	LICR_RnaSeq_HUVEC_nucleus_longPolyA
GSM767844	LICR_RnaSeq_K562_nucleus_longNonPolyA
GSM765387	LICR_RnaSeq_K562_nucleus_longPolyA

**Table S2 Locus information PCR primers used in ChIP validation of H3k27ac patterns in K562 cell line**

A bimodal pattern, c-Jun occupied; CHR:21 START:17904644 END:17906657

Primer pair	Relative position	Sequence (5'~3')
1	-943	CGACATAGATGTTACTGTGTGTGTC TTTTAAAGTTCAAATCATTCACAGAA
2	-835.5	AAAATTCGTGAAATGATTGAACTTT TGTGTTAACCTTTCTAGGAACTGC
3	-654.5	GGTGAACCTTGAATACATCACAC GGATATTTCAATCAATGGAATCG
4	-458	CAGGGTACTTTTTGAGGTGAT CCATGCAATTCACCCATTTA
5	-269.5	TAAATGGGTGAATTGCATGG TTTCGAAAAGTCAATGCTGA
6	-184	CACAAAACTAGTGGATCTATCAGGA TCCTTGAGGCAATCCAAAA
7	-100	TTTGATTGCCTCAAAGGAC CAATCTTAGGACAAGGGTTTGC
8	45	CCCTGTCCTAAGATTGTTTGA CCCAAAAAGGTGATGTTGCT
9	125	TTTGTGTTCCACATTTCCA TCCTTGTTGAATTTGCACTTG
10	315	CACACAAATCTCCTCCAGA GGTGGGCTGGGGTTATGTAT
11	814.5	CACCACGATACGATGGAAAA TTGTTGCTTTGCTGGAGTG
12	883.5	CAAGGAACCACACTAAAGAAATG GCAAACAGAAACAAATCCAGAA

## B colocalization pattern, c-Jun occupied; CHR:1 START:165783409 END:165785422

Primer pair	Relative position	Sequence (5'~3')
1	-960	CCTGCATGACAGAGCAAGAC CCTTAAGGTTAAGATCAAGATTTACAA
2	-768	AAAGGAAATGTAAGAATGCAGAGA TTACAGTGGTTTCCCACTTCG
3	-566	GCATATTTATCCATAACATGGA AAAAAGATGGAGGAGCAGCA
4	-376	TGACACGGTCAAAAAGCAAA ATCACCTGCTGAAGGGAAGA
5	-245.5	GCCCAGACATGAAAAACAT TTCTTACTAAAGAGCATCTGGTTGA
6	-180	CATCTGCTTGGCTTCAGAAAA AATTCCATGAAAACCAAACG
7	-66	ATGTTCTGACTGGGCATGG TACAGGCGTGAGCCACTATG
8	56.5	CATAGTGGCTCACGCCTGTA GGCTGGTCTTGAACCTCTGA
9	260	AATTAGCTGGGTGTGGTGGT TACTGCAACCTCTGCCTCCT
10	461.5	TGAAACCATCTCCCTCCTG CAAAGTCCCCTTTGTGTTCA
11	661.5	TGAAATATGGTTTCTGGAGAGG CCCTGAGTGAAGGTGATGT
12	772	ATTCTTTTCTGAAACCCTGAA ACAGCCTTGAATTTGGGTGT

## C low signal pattern, c-Jun occupied; CHR:15 START:69330650 END:69332661

Primer pair	Relative position	Sequence (5'~3')
1	-943.5	TGGTACCGGAAATAGCCACT CTCCAACCTGGACAACACAG
2	-760	CCTACAGGCATGCACTACCA GGAGGCTAAATTGGGAGCA
3	-497	GGCAGTCATAAGTAAGCACTCTTG CACTAAGCCCAGCCTTTTCCAC
4	-317.5	TGAACTTCCTTGGATGAAAAA GGGAAAGGCTTTGGATGTTT
5	-152	AGGTGGCAAAGGTTTTTGTG GCCAAAATTCGCTCCAGT
6	-27	GAAAAACAAGCCCATTTTCCA TTCCAGAAGCCGTAATTGCT
7	92	TTGGTGTCTCCACTTCAGCA TACAGGCGTGAGCCACTATG
8	56.5	CATAGTGGCTCACGCCTGTA GGCTGGTCTTGAACCTCTGA
9	260	AATTAGCTGGGTGTGGTGGT TACTGCAACCTCTGCCTCCT
10	461.5	TGAAACCATCTCCCTCCTG CAAAGTCCCCTTTGTGTTCA
11	661.5	TGAAATATGGTTTCTGGAGAGG CCCTGAGTGAAGGTGATGT

**Table S3 Co-occurring transcription factor clusters and multiple transcription factor loci (MTLs)**

No.	Cell line	Transcription factors	Minimum pairwise overlap	MTL count
1	HeLa-S3	E2F1, HA	0.59	4901
2	HeLa-S3	BDP1, BRF1, BRF2	0.49	360
3	Gm12878	BATF, BCL11A, IRF4	0.46	20793
4	K562	Brg1, Ini1	0.44	4918
5	HeLa-S3	c-Myc, Max, Pol2, Ini1	0.42	48558
6	Gm12878	POU2F2 (Oct-2), PAX5, SP1, TCF12	0.40	42858
7	HepG2	HEY1, Sin3A, TAF1, Pol2	0.39	27994
8	K562	c-Myc, Max	0.39	13528
9	HeLa-S3	c-Fos, c-Jun, JunD	0.37	17763
10	HepG2	JunD, p300, RXRA	0.36	15030

**Table S4 Annotated conserved binding sequences of TFs**

TF	Consensus symbol/sequence	No. of sites	Reference
ATF-3	CRE (5' GTGACGT[AC][AG] 3')	9654	SwissProt
E2F4	V\$E2F_02	7241	UCSC tfbsConsSites
E2F6	5' TTTC[CG]CGC 3'	17912	SwissProt
c-Fos	V\$AP1FJ_Q2; V\$AP1_01	23041	UCSC tfbsConsSites
c-Jun	V\$CREBP1CJUN_01; V\$AP1FJ_Q2; V\$AP1_01	26812	UCSC tfbsConsSites
c-Myc	V\$MYC_MAX_01; V\$MYC_MAX_02; V\$MYC_MAX_03	25886	UCSC tfbsConsSites
Egr-1	V\$EGR1_01	5234	UCSC tfbsConsSites
JunD	V\$AP1_01	15233	UCSC tfbsConsSites
Max	V\$MYC_MAX_03	8410	UCSC tfbsConsSites
NF-E2	V\$NFE2_01; V\$NFE2_01	13114	UCSC tfbsConsSites
POU2F2 (Oct-2)	5' ATTGTCAT 3'	28396	SwissProt
p300	V\$P300_01	7988	UCSC tfbsConsSites
Pax-5	V\$PAX5_01; V\$PAX5_02	15404	UCSC tfbsConsSites
SP1	V\$SP1_01; V\$SP1_Q6	3657	UCSC tfbsConsSites
STAT1	V\$STAT_01; V\$STAT1_01; V\$STAT1_03	19570	UCSC tfbsConsSites
STAT2	V\$STAT_01	11685	UCSC tfbsConsSites