Whole-Genome Analysis of Histone Marks Reveal Bivalent Promoter Dynamics upon Differentiation

Qingyan Li[1, 4], Lin Li[2], Yong Yang[1], Xianhua Dai[1, 3*]

[1] School of Information Science and Technology, Sun Yat-Sen University, Guangzhou 510006, China.

[2] College of Computer and Engineering, Jimei University, Xiamen 361021, China.

[3] SYSU-CMU Shunde International Joint Research Institute, Shunde 528300, China.

[4] School of Science, Jimei University, Xiamen 361021, China.

* Corresponding author. Tel.: 8602039943331; email: issdxh@mail.sysu.edu.cn

Abstract: In ES cells, many promoters are co-modified by the activating H3K4me3 mark and the repressive H3K27me3 mark. These bivalent domains are thought to play a pivotal role related to pluripotency by maintaining lineage commitment programs in a poised state. Enhancers are long-range regulatory elements that affect development and differentiation, which are marked by the presence of H3K4me1. However, how enhancers functional interact with bivalent promoters has remained unknown. In this paper, we found bivalent promoters are significantly flanked by poised enhancers which are distinguished by the enrichment of H3K4me1 and the absence of the H3K27ac in hESC. After differentiation of ESC to NPC, a neuron-specific subset of bivalent promoters acquires a chromatin signature with active modification marks.

Key words: Bivalent promoter, enhancer, association, H3K27me3.

1. Introduction

Histone post-translational modifications have emerged as important players in the regulation of gene expression. The four core histones — H2A, H2B, H3, and H4 — have dozens of different modifications, including methylation, acetylation, phosphorylation and ubiquitination. Histone H3 methylations of lysine 4 (H3K4) and lysine 27 (H3K27) have been shown to relate with active and repressed states, respectively [1]. These methylations are catalyzed by Trithorax group (trxG) and Polycomb group (PcG) proteins and play key roles in lineage-specific developmental functions [2]. Trithorax-associated H3K4 trimethylation (H3K4me3) positively regulates transcription by recruiting nucleosome remodeling enzymes and histone acetylases [3]-[5], whereas Polycomb-associated H3K27 trimethylation (H3K27me3) negatively regulates transcription by promoting a compact chromatin structure [6], [7]. The colocalization of these H3K4me3 and H3K27me3 histone modifications, termed 'bivalent domains', was found highly enriched in ES cells by mapping mouse genome [8], [9]. There is a large cohort of developmental regulatory promoters that are concomitantly marked by this paradoxical histone modifications. By exhibiting both active and repressive features, bivalent genes maintain a poised transcriptional state for activation upon suitable developmental cues or environmental stimuli. Whole-genome mapping found that H3K4me3 peaks were enriched in the region within 2kb of the TSS of Refseq annotations, and H3K27me3 peaks were also enriched in a band

center around the TSS with a greater width; moreover, most H3K27me3 peaks localized on promoters that were already marked with H3K4me3 [10].

Enhancers are distal-regulating elements that orchestrate the regulation of genes that are critical for cell differentiation and development [11]. In recent years, genome wide studies have identified several key characteristics of enhancer elements. For example, enhancer elements correlate with the enrichment of the histone mark H3K4me1, which are located in DNase I hypersensitive sites, and often display cell type-specific localization patterns across the genome [12], [13]. Although many approaches have been developed to find out how enhancers interact with target promoters, the dynamic associations between enhancers and bivalent promoters remains unclear.

In this work, we analyzed genome-wide ChIP-seq data sets to map four chromatin marks in embryonic stem cells (ESC) and ESC-derived neuronal progenitor cells (NPC), and used combinations of histone modifications to define poised and active enhancers, repressive, bivalent and active promoters [14], [15]. We established H3K27ac as an important enhancer mark that distinguishes between active and poised enhancer elements. The resulting chromatin state maps allow us to connect bivalent promoters to putative regulated enhancers, and to predict cell type-specific developmental potency. Furthermore, it offers some insights into understanding the dynamic interactome of bivalent promoters and enhancers during differentiation of ESC. our paper.

2. Methods

2.1. Sequencing Data Alignment and Analysis

The following publically available ChIP-seq data sets were obtained from the Sequence Read Archive (http://www.ncbi.nih.gov/sra): hESC H3K4me1 (SRX012373), hESC H3K4me3 (SRX012501), hESC H3K27me3 (SRX012368), hESC H3K27ac (SRX012366), hNPC H3K4me1 (SRX101243), hNPC H3K4me3 (SRX101246), hNPC H3K27me3 (SRX101236), and hNPC H3K27ac (SRX101234). SRA file were converted to FASTQ format by using the fastq-dump utility of the SRA toolkit from NCBI, and the FASTQ files were mapped to the hg19 reference genome using Bowtie with the default settings and output to SAM format [16]. The peaks of ChIP-seq data were called using model-based analysis for ChIP-Seq (MACS) [17].

Enhancer locations were defined by the peak of the H3K4me1 enriched regions. To avoid confounding active transcripts with gene bodies, we focused on intergenic enhancers, which were defined as those outside gene bodies and at least 1K from H3K4me3 enriched regions or a known transcript start site (TSS) of a gene (downloaded from the UCSC Tables Browser; http://genome.ucsc.edu/cgi-bin/hgTables?command =start Refseq). Enhancer peaks that overlap with H3K27AC peaks were defined as active enhancers, others were poised enhancers. A promoter is considered to be bivalent if it contains both an active H3K4me3 peak and a repressive H3K27me3 peak within ±2 kb of its TSS.

We do the above works in Ubuntu 12.04.5 LTS 64 system, example shell scripts were described as bellows:

```
fastq-dump SRR029609/SRR029609.sra;
fastq-dump SRR029620/SRR029620.sra;
bowtie -m 1 -S ../../hg19/hg19 SRR029609.fastq >H3K4me3a.sam;
bowtie -m 1 -S ../../hg19/hg19 SRR029620.fastq >H3K4me3b.sam;
samtools view -bSoH3K4me3a.bam H3K4me3a.sam;
samtools view –bSo H3K4me3b.bam H3K4me3b.sam;
samtools cat -o H3K4me3.bam H3K4me3a.bamH3K4me3b.bam
macs14 -t H3K4me3.bam -g hs -n H3K4me3 -S;
```

In total, we identified 111,938 enhancers in hESCs, of which 27,013 were active and 84,925 were poised. A total of 6163/4635 bivalent promoters were also recognized in hESCs. In hNPCs, 71,931 enhancers were identified, including 17,251 active enhancers and 54,680 poised ones (Table 1).

Table 1. Number of Bivalent Promoters and Enhancers in hESC and hNPC

Cell line	Active enhancer	Poised enhancer	Enhancer	Bivalent promoter
hESC	27,013	84,925	111,983	4635
hNPC	17,251	54,680	71,931	1205

2.2. Bivalent Promoters-Associated Enhancers

Each bivalent promoter was linked to its closest enhancer, based on the distance to Refseq TSS. The distance between TSS and their closest enhancers was calculated by using BEDTools [18]. Gene Ontology (GO) analysis and Pathway analysis were performed by using DAVID web services [19] and ConsensusPathDB[20], respectively.

2.3. Analysis Pipeline

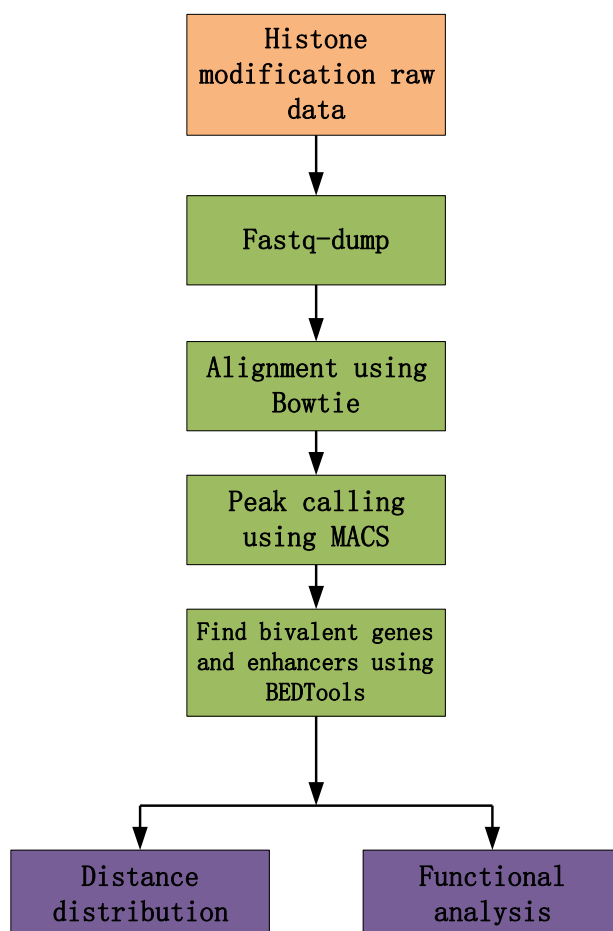The analysis pipe line was described in Fig. 1.

Fig. 1. Analysis pipeline.

3. Result and Discussion

3.1. Bivalent Promoters Flanked by Poised Enhancers Significantly

From the distance distribution of all interaction between bivalent promoters and enhancers, we found that there are some characteristics on the interactome. Firstly, although there are examples of long-range interactions in Human and Drosophila [21], [22], the vast majority of bivalent promoters we found are within 20 kb of their nearest enhancer, with few known to exceed over 50 kb (Fig. 2(a)). The median distance between bivalent promoters and enhancers was 11 kb, and only 19 (0.4%) was over 200 kb (Fig. 2(b)).
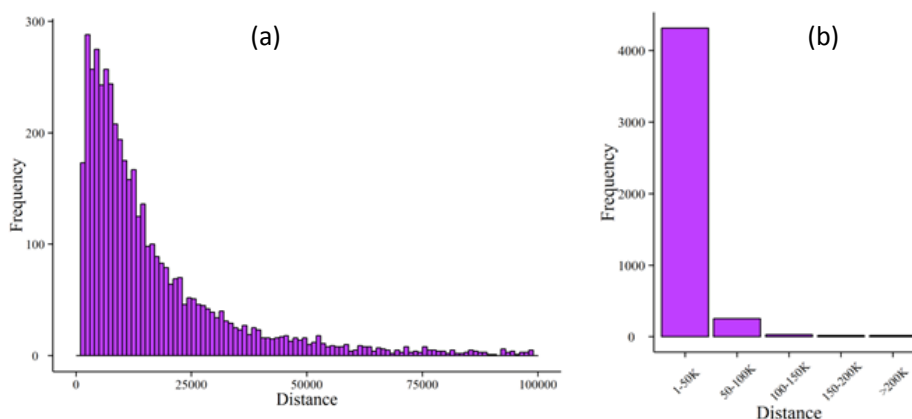

Fig. 2. Distribution of the distance between bivalent promoters and its nearest enhancer.

Secondly, of the 4635 bivalent promoters we found here in hESC, 3678 promoters were flanked by poised enhancers, while 957 promoters were flanked by active enhancers. Using all 27,013 active enhancers and 84,925 poised enhancers we identified as background, we found that bivalent promoters are more significantly flanked by poised enhancers ($P = 3 \times 10^8$, calculated by Fisher's exact test).

Thirdly, to test broader applicability of this significance, we extended our analysis to ES-derived neural progenitor cells. Although only 1,205 promoters were marked by bivalent modifications of H3K4me3 and H3K27me3, the nearest enhancers of 997 promoters were poised, while the other 208 promoters were more closed to active enhancers. We found similar distribution results as hESC, the bivalent promoters were preferentially flanked by poised enhancers in hNPC ($P = 6 \times 10^{-8}$, calculated by Fisher's exact test).

3.2. Most Bivalent Promoters Loose Repressive Mark upon Differentiation

We next examined the H3K4me3 and H3K27me3 modification patterns in the neuronal progenitor cell (NPC) which was differentiated from ESC. Contrast to 4635 found in the ES cell data, we only found 1205 bivalent promoters in this cell type, suggesting the same amount of bivalent promoters lost repressive mark during differentiation.

Among these 1205 bivalent promoters in NPC, there are 999 promoters also found in ESC and 206 newly formed bivalent promoters. Thus, the majority of TSSs (3637/4635) that show bivalent modifications in ES cells do not show bivalent domains in the NP cells. Moreover, the vast majority of these (3277/3637) TSSs that lost bivalent marks show H3K4me3 state.

Intriguingly, there are 206 do novo established bivalent promoters found in NPC. When using the human genes as background, Gene Ontology analysis of these newly formed bivalent promoters suggests that they are significantly enriched in neuron development and differentiation (Table 2). Pathway analysis of these de novo formed bivalent promoters discovered that the predominantly upregulated pathway was Ectoderm differentiation and Ectoderm commitment pathway (Fig. 3). Moreover, we also conducted the pathway analysis for the bivalent promoters that found both in hESC and hNPC, the most enriched pathway is

Neuroactive ligand-receptor interaction when using KEGG as source (Table 3), the result is consistent with the function of the hNPC.

Table 2. The 10 Top Most Enriched GO Terms of Biological Processes about Neuron Development in Newly Formed Bivalent Promoters in NP Cells

No.	GO term	Bivalent promoter n (%)[a, b]	E-ratio[c]	*P*-value
1	GO:0007417~central nervous system development	21 (10.9)	5.0	5.5E-9
2	GO:0030182~neuron differentiation	18 (9.4)	4.2	1.3E-6
3	GO:0030900~forebrain development	11 (5.7)	7.3	2.4E-6
4	GO:0007420~brain development	14 (7.3)	4.9	5.1E-6
5	GO:0048666~neuron development	15 (7.8)	4.5	5.8E-6
6	GO:0022008~neurogenesis	20 (10.4)	3.4	6.1E-6
7	GO:0048699~generation of neurons	19 (9.9)	3.5	8.5E-6
8	GO:0007399~nervous system development	27 (14.1)	2.5	1.7E-5
9	GO:0007409~axonogenesis	11 (5.7)	5.8	2.0E-5
10	GO:0048667~cell morphogenesis involved in neuron differentiation	11 (5.7)	5.3	3.9E-5

[a]Num., number of proteins annotated; [b]Per., percentiles of proteins annotated; [c]E-ratio, enrichment ratio of bivalent genes
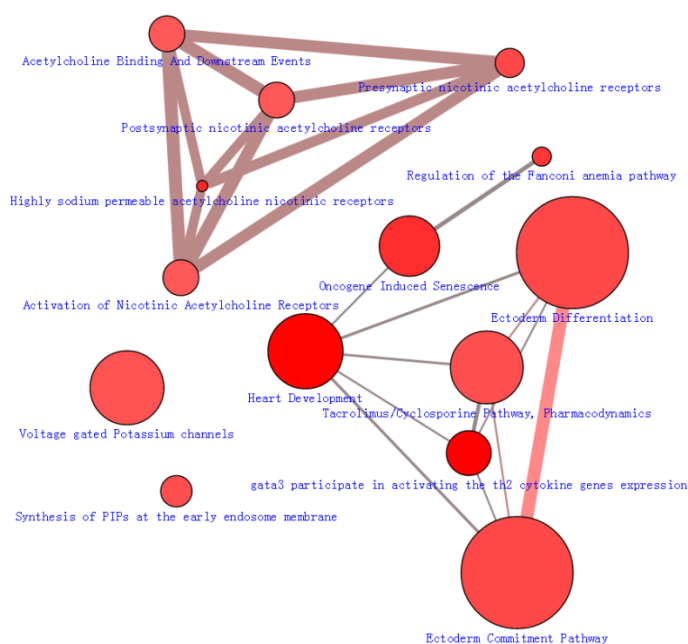

Fig. 3. Enriched pathways of newly formed bivalent promoters.

Table 3. The Most Enriched Pathway of the Bivalent Promoters Which was Found Both in hESC and hNPC

pathway	p-value
Neuroactive ligand-receptor interaction	6.80E-14
Calcium signaling pathway	5.53E-08
cAMP signaling pathway	1.90E-07
Inflammatory mediator regulation of TRP channels	8.34E-06
Circadian entrainment	7.60E-05
Pathways in cancer	9.66E-05
Vascular smooth muscle contraction	0.00011366
Insulin secretion	0.000144948
Serotonergic synapse	0.000580285
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	0.001781832
Thyroid cancer	0.001906382
Dilated cardiomyopathy	0.002539808
Oxytocin signaling pathway	0.002708684
MAPK signaling pathway	0.003042026
HTLV-I infection	0.003870178
PI3K-Akt signaling pathway	0.004005475
Cholinergic synapse	0.004441927
Type II diabetes mellitus	0.004501796
Axon guidance	0.004816924
ECM-receptor interaction	0.005579308
Rap1 signaling pathway	0.007083635
Melanogenesis	0.007337503
Salivary secretion	0.008359651

4. Conclusions

Here, we examined genome-wide relationship between bivalent promoters and enhancers during early neural differentiation of hESCs. Our major findings are that: (i) aconsiderable number of bivalent promoters show spatial proximity to poised enhancers; (ii) the epigenetic dynamics of bivalent promoters and close enhancers upon differentiation. Our results shed light on the association between bivalent promoters and enhancers during development in ES cells and NP cells. We proposed that the co-localization of bivalent promoters and poised enhancers may provide a robust epigenetic state, which will keep developmental genes poised in ES cells.

Further studies are needed to define the association between bivalent promoters and their distal enhancers using alternate techniques, such as 4C, 5C and Hi-C etc.

Acknowledgements

References

[1] Bracken, A. P., Dietrich, N., Pasini, D., Hansen, K. H., & Helin, K. (2006). Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions. *Genes Dev.*, *20*, 1123-1136.

[2] Ringrose, L., & Paro, R. (2004). Epigenetic regulation of cellular memory by the Polycomb and Trithorax group proteins. *Annu. Rev. Genet.*, *38*, 413-443.

[3]   Santos-Rosa, H., *et al*. (2003). Methylation of histone H3 K4 mediates association of the Isw1p ATPase with chromatin. *Mol. Cell*, *12*, 1325-1332.

[4]   Pray-Grant, M. G., Daniel, J. A., Schieltz, D., Yates, J. R., & Grant, P. A. (2005). Chd1 chromodomain links histone H3 methylation with SAGA-and SLIK-dependent acetylation. *Nature*, *433*, 434-438.

[5]   Sims III, R. J., *et al*. (2005). Human but not yeast CHD1 binds directly and selectively to histone H3 methylated at lysine 4 via its tandem chromodomains. *J. Biol. Chem*., *280*, 41789-41792.

[6]   Francis, N. J., Kingston, R. E., & Woodcock, C. L. (2004). Chromatin compaction by a polycomb group protein complex. *Science*, *306*, 1574-1577.

[7]   Ringrose, L., Ehret, H., & Paro, R. (2004). Distinct contributions of histone H3 lysine 9 and 27 methylation to locus-specific stability of polycomb complexes. *Mol. Cell*, *16*, 641-653.

[8]   Bernstein, B. E., *et al*. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, *125*, 315-326.

[9]   Azuara, V., *et al*. (2006). Chromatin signatures of pluripotent cell lines. *Nat. Cell Biol*., *8*, 532-538.

[10] Pan, G., *et al*. (2007). Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. *Cell Stem Cell*, *1*, 299-312.

[11] Bulger, M., & Groudine, M. (2010). Enhancers: The abundance and function of regulatory sequences beyond promoters. *Dev. Biol.*, 339, 250-257.

[12] Heintzman, N. D., *et al*. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, *459*, 108-112.

[13] Xi, H., *et al*. (2007). Identification and characterization of cell type–specific and ubiquitous chromatin regulatory structures in the human genome. *PLoS Genet*., *3*, e136.

[14] Ernst, J. & Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol*., *28*, 817-825.

[15] Ernst, J., *et al*. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, *473*, 43-49.

[16] Langmead, B., Trapnell, C., Pop, M., Salzberg, S. L., *et al*. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*., *10*, R25.

[17] Feng, J., Liu, T., Qin, B., Zhang, Y., & Liu, X. S. (2012). Identifying ChIP-seq enrichment using MACS. *Nat. Protoc*., *7*, 1728-1740.

[18] Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*, 841-842.

[19] Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, *Nat. Protoc*., *4(1)*, 44-57.

[20] Kamburov, A., Stelzl, U., Lehrach, H., & Herwig, R. (2013). The consensus path DB interaction database: 2013 update, *Nucleic Acids Res*., 41*(1)*, 793-800.

[21] Sexton, T., *et al*. (2012). Three-dimensional folding and functional organization principles of the drosophila genome. *Cell*, *148*, 458-472.

[22] Zhang, Y., *et al*. (2013). Chromatin connectivity maps reveal dynamic promoter–enhancer long-range associations. *Nature*, *504*, 306-310.

Qingyan Li was born on April 13, 1979 in Xiamen, China. He is a Ph.D. candidate in bioinformatics at the Sun Yat-Sen University. He received the MSc from Fuzhou University, with specific training in computer science. For more than seven years, he has been working as a teacher and researcher at the School of Science at the University of Jimei. His research has focused mainly at genomics and epigenetics.

Lin Li was born on January 10, 1980 in Zhangzhou, China. She obtained her M.S. degree in computer sciences from Fuzhou University in 2005. Her current research topic is focused on the bioinformatics. Now, she is a teacher of the Computer Engineering College at the University of Jimei.

Yong Yang was born on March 20, 1980 in Lingshui, China. He received his M.S. degree in computational mathematics from South China Normal University. Actually he is a Ph.D. candidate in bioinformatics at the Sun Yat-Sen University. His Ph.D. topic was focused on the eukaryotes genomic regulation element, such as enhancer and promoter.

Xianhua Dai obtained his M.S. degree and Ph.D. degree in signal and information processing from Southeast University on 1988 and 1992 respectively. Throughout his practice, he has developed research projects in the areas of signal processing and bioinformatics. He has been a teacher at the Sun Yat-Sen University for more than thirteen years.