HISTONE MODIFICATION OF STRESS RESPONSIVE REGULATORY REGIONS, A BIOINFORMATICS STUDY

By

GUANGYAO LI

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2013

To my dear wife Jingyi and our baby Sophia

ACKNOWLEDGMENTS

This remarkable journey of achieving the Ph.D. degree fulfills with my most valuable memories. It would not have been possible for me to complete this doctoral dissertation without the help of the following individuals.

First, I would like to express my appreciation to my advisor, Dr. Lei Zhou, for his continuous support of my Ph.D. study. His patience, inspiration and motivation made my achievements possible. He has shown me the curiosity, enthusiasm and persistence a real scientist should carry, and he always encourages me to overcome all obstacles and guides me to become independent in every aspect of my life.

I also want to express a special thanks to my dissertation committee: Dr. Thomas Yang, Dr. Suming Huang, Dr. Luciano Brocchieri, Dr. Samuel Wu, and my former committee Dr. Alberto Riva, for their valuable suggestions, inspiration and continuous support.

My sincere thanks also go to the formal and current members of Zhou laboratory, especially Dr. Yanping Zhang, Dr. Nianwei Lin, Dr. Can Zhang, Dr. Bo Liu, Michelle Chung, Jordan Reuter, Denis Titov for their kind and patient help. I also would like to thank my classmates and best friends, including Jianxing Zhang, Yihai Wang, Qingchun Shi, Tong Lin, Shaojun Tang, Yajie Yang, Ming Tang, Yuanqing Yan, Ruli Gao, Bing Yao, Weiyi Ni, Shuibin Lin, Chen Ling, Yi Guo, Wei Wang, Shanjun Helian, Liangjie Yin, Mei Zhang and so on. I am so grateful to have met them in Gainesville and become lifetime friends.

Last and most importantly, I want to thank my parents and my family, whose love and encouragement allowed me to finish this journey. A special thanks to my beloved

wife Jingyi He and our little daughter Sophia Li, who fulfills my life and makes everything meaningful to me.

TABLE OF CONTENTS

page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BEAF-32	Boundary element associated factor of 32kD
CATCH-IT	Covalent attachment of tags to capture histones and identify turnover
CHIP	Chromatin Immuno-precipitation
CP190	Centrosomal protein 190kD
CTCF	CCCTC-binding factor
CTR	Chromatin transitional region
DCTCF	*Drosophila* ortholog of mammalian CCCTC-binding factor
ENCODE	ENCyclopedia of DNA element
GAF	GAGA factor
H3K9ME3	Trimethylated histone 3 lysine 9
H3K27ME3	Trimethylated histone 3 lysine 27
IRER	Irradiation responsive enhancer region
ILB	IRER left boundary
MNASE	Micrococcal nuclease
MODENCODE	Model organism ENCyclopedia of DNA element
MOD(MDG4)	Modifier of mdg4
NURF	Nucleosome remodeling factor
PCG	Polycomb group
PRE	Polycomb response elements
PUMA	p53 upregulated modulator of apoptosis
QPCR	Quantitative polymerase chain reaction
SU(HW)	Suppressor of hairy wing
TRXG	Trithorax group
TSS	Transcription start site

UAS	Upstream activation sequence
USF1	Upstream stimulatory factor 1

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

HISTONE MODIFICATION OF STRESS RESPONSIVE REGULATORY REGIONS, A BIOINFORMATICS STUDY

By
Guangyao Li

December 2013

Chair: Lei Zhou
Major: Genetics and Genomics

The juxtaposed distribution of euchromatic and heterochromatic chromatin domains defines the expressivity of genes and thus the identity of individual cells. Due to the self-propagating nature of the heterochromatic modification H3K27me3 (tri-methylation of lysine 27 on histone H3 tail), chromatin barrier activities are required to demarcate the boundary and prevent it from encroaching into euchromatic regions. Studies in *Drosophila* and vertebrate systems have revealed several important chromatin barrier elements and their respective binding factors. However, epigenomic data indicate that the binding of these factors are not exclusive to chromatin boundaries. To gain a comprehensive understanding of facultative heterochromatin boundaries, we developed a two-tiered method to identify the Chromatin Transitional Region (CTR), i.e. the 200bp region that shows the greatest transition rate of the H3K27me3 modification as revealed by ChIP-Seq (chromatin immunoprecipitation followed by high-throughput sequencing). This approach was applied to identify CTRs in *Drosophila* S2 cells and human HeLa cells. Although many insulator proteins have been characterized in *Drosophila*, less than half of the CTRs in S2 cells are associated with known insulator proteins, indicating unknown mechanisms remain to be characterized. Our analysis also

revealed that the peak binding of insulator proteins are usually 200~600bp away from the CTR. Comparison of CTR-associated insulator protein binding sites vs. those in heterochromatic region revealed that boundary -associated binding sites are distinctively flanked by nucleosome destabilizing sequences, which correlates with significantly decreased nucleosome density and increased binding intensities of co-factors. A subgroup of facultative heterochromatin boundaries have enhanced H3.3 incorporation but reduced nucleosome turnover rate. Together, our genome-wide study reveals that diverse mechanisms are employed to define the boundaries of facultative heterochromatin.

CHAPTER 1
BACKGROUND AND INTRODUCTION

Polycomb Group Proteins and Polycomb Silencing

In eukaryotic genomes, DNA is compacted into nucleus in the high order structure called chromatin. As the structural unit of chromatin, nucleosome is composed of a histone octamer and 147bp of DNA sequence wrapped around it. The histone octamer consists of two copies of each of the histones H2A, H2B, H3 and H4. The nucleosomes are linked together by linker DNA to form the “beads on string” structure. Then nearby nucleosomes are condensed together by the linker histone H1 to form the 30nm chromatin fiber (Allan et al., 1980). With the function of scaffold proteins, chromatin further forms the higher order looping structure. The N-terminal tails of histones can protrude out of the nucleosome particle and are subject to different post-transcriptional modifications, such as methylation, acetylation, phosphorylation, ubiquitination and so on. The different combinations of histone modifications, known as histone code (Jenuwein and Allis, 2001), will result in different chromatin structure and subsequently, distinct expression pattern of associated genes.

PcG Proteins

Polycomb group (PcG) proteins are originally discovered in *Drosophila* as chromatin structure remodelers to prevent inappropriate expression of homeotic (Hox) genes during embryonic development (Lewis, 1978). PcG proteins begin to function in 3-hour-old fly embryo in which the expression pattern of Hox genes has been shaped by upstream transcription factors (Pirrotta, 1998; Zhang and Bienz, 1992). After early embryogenesis, PcG silencing takes place and Hox genes will maintain their state of expression throughout the rest of development.

PcG-mediated silencing in *Drosophila* involves at least three multiprotein complexes, which are PRC1, PRC2 and PhoRC complexes. The key components of PRC1 (Polycomb repressive complex 1) and PRC2 (Polycomb repressive complex 2) are the polycomb protein PC (Polycomb) and E(z) (Enhancer of zeste), respectively (Czermin et al., 2002; Shao et al., 1999). Polycomb-mediated silencing is usually associated with trimethylation of lysine 27 of histone 3 (H3K27), which is catalyzed by the PRC2 complex. This histone modifying activity requires a minimum of three components in PRC2 – E(z), Esc and Su(z)12 (Cao and Zhang, 2004; Nekrasov et al., 2005). The chromodomain of PC can recognize and specifically bind to trimethylated H3K27 (Fischle et al., 2003), then PRC2 is recruited and the neighboring nucleosome is trimethylated by histone methyltransferase (HMTase) activity of PRC2. The polycomb silencing will self-propagate until the chain reaction is broken. The core components of the other PcG complex PhoRC (Pleiohomeotic repressive complex) is PHO (Pleiohomeotic) and PHOL (Pleiohomeotic-like), which are the only known DNA-binding PcG proteins (Brown et al., 2003; Brown et al., 1998; Fritsch et al., 1999). The role of PhoRC on polycomb silencing is unclear, as the mutation of PHO and PHOL did not significantly abolish the binding of PRC1 and PRC2 on polytene chromosomes (Brown et al., 2003). But the direct binding of PhoRC to DNA may mediate the initiation of PcG repression (Mohd-Sarip et al., 2005).

Targeting of PcG Silencing

How the PcG-mediated silencing gets initiated still remains to be fully understood. In *Drosophila*, the specific regulatory region named Polycomb Response Element (PRE) which serves as docking platforms for PcG proteins and initiates PcG repression, have long been discovered. Although some sequence-specific DNA binding proteins have

been found at PREs, including PHO, PHOL, GAGA factor, Zeste and so on. The PREs share little similarity at the sequence level. Based on the clustered pairs of these transcription factors binding sites, an algorithm has been developed and more than a hundred of PREs have been predicted in *Drosophila* genome (Ringrose et al., 2003). However, the result was challenged by recent genome-wide ChIP-Chip analyses where only limited overlap were found between PcG protein binding sites and the predicted PREs (Negre et al., 2006; Schwartz et al., 2006; Tolhuis et al., 2006). Nowadays, more and more evidences indicate the polycomb silencing is a complex progress, its initiation may involve different mechanisms such as non-coding RNA and RNAi machinery (Petruk et al., 2006; Rinn et al., 2007; Sanchez-Elsner et al., 2006).

TrxG Proteins

The identification of Trithorax group (TrxG) proteins was also originated from the early study of Hox genes in *Drosophila*. Compared to PcG silencing, TrxG proteins set up the active state for Hox genes. For example, in the absence of Trithorax (TRX), multiple Hox genes become repressed in early stage embryos where they normally express, and flies show segmental transformations consequently (Breen and Harte, 1991; Orlando and Paro, 1995). The maintenance of active chromatin state by TrxG proteins was achieved through direct histone modification (Strahl and Allis, 2000) or ATP-dependent nucleosome remodeling (Vignali et al., 2000). The main *Drosophila* TrxG proteins, TRX, is itself a histone methyltransferase which catalyzes H3K4 trimethylation (Santos-Rosa et al., 2002). And the SWI-SNF nucleosome remodeling complex in TrxG family contains ATP-dependent nucleosome remodeling proteins that can change chromatin structure to facilitate transcription machinery (Smith and Peterson, 2005). The appropriate regulation between PcG and TrxG proteins, through

catalyzing either repressive or active histone modifications, respectively, is crucial for the temporal and spatial expression pattern for Hox genes as well as other target genes.

Chromatin Insulators

Chromatin insulators are regulatory DNA elements that are bound by insulator proteins and cofactors to prevent inappropriate communication between different chromatin domains.

Two Functions of Chromatin Insulators

Two types of functions have been attributed to insulators. The first is called enhancer-blocking activity, i.e. blocking the interaction between enhancer and promoter when locate in-between to prevent inappropriate gene expression. The other is chromatin barrier activity, which antagonizes the propagation of heterochromatin silencing (Gaszner and Felsenfeld, 2006).

Insulators, such as the *gypsy* insulator, were originally identified for their enhancer-blocking activity (Geyer et al., 1986). Later, it was revealed that most of them also have barrier activity. It was not clear whether the two activities are separable until the characterization of the cHS4 insulator in the chicken β-globin locus. The complete cHS4 has both enhancer-blocking and barrier activity. However, a series of mechanistic studies indicated that the two activities are separable and carried out by distinct DNA elements. The enhancer-blocking activity of cHS4 is mediated by CTCF, while its barrier activity against heterochromatin formation requires a binding site for USF1 (Upstream Stimulatory Factor 1). Binding of USF1 to cHS4 recruits chromatin-modifying enzymes that catalyze histone modifications incompatible with heterochromatin formation, thus

preventing the propagation of suppressive histone modification (Huang et al., 2007; West et al., 2004).

Recently, a novel chromatin barrier that lacks any detectable enhancer-blocking function has also been identified in *Drosophila* (Lin et al., 2011). This ~200bp element is located at the left boundary of IRER (Irradiation Responsive Enhancer Region), a 33 kb intergenic regulatory region controlling stress-induced expression of multiple pro-apoptotic genes (Zhang et al., 2008a). When tested in transgenic animals, ILB (IRER Left Boundary) is fully capable of blocking the propagation of H3K27me3 initiated by a strong Polycomb response element (PRE) (Lin et al., 2011). The chromatin barrier function of ILB is evolutionarily conserved. When tested in a vertebrate system, it blocked heterochromatin propagation as effectively as the cHS4 (Lin et al., 2011). Although many insulator/boundary-associated proteins have been characterized in *Drosophila*, including Su(Hw), dCTCF, BEAF-32, GAF, CP190 and Mod(mdg4) (Gurudatta and Corces, 2009), none of those was found associated with ILB.

Models for Chromatin Insulator Functions

Although the molecular mechanisms of insulators are not as well-studied as promoters or enhancers, several models have been proposed for the two functions of chromatin insulators based on recent investigations on yeast, *Drosophila* and vertebrates. The three models supporting the enhancer-blocking activity are called promoter decoy model, physical barrier model, and loop domain model (Bushey et al., 2008; Raab and Kamakaka, 2010).

Three models have also been proposed for the chromatin barrier activity. The first model, called nucleosome gap model, got inspiration from the studies in yeast (Bi and Broach, 1999, 2001). This model proposes that a nucleosome-free region forms at

chromatin insulator, and it disrupts the spread of repressive histone modifications since the modifiers need to interact with the adjacent nucleosome. The second model proposes that the recruited acetytransferases can compete with the spreading deacetylation and methylation activities to counteract the propagation of heterochromatin. The observation that “hot nucleosomes” are associated with heterochromatin boundaries in yeast leads to the third model. In this model, rapid and constant replacement of nucleosomes erases the repressive histone modification, before the repressive modifications can spread any further (Dion et al., 2007). Although different models have been proposed, it was observed *in vivo* that multiple mechanisms are involved in demarcating the boundaries between heterochromatin and euchromatin (Oki et al., 2004).

Epigenetic Regulation of Gene Expression

In contrast to genetics, which studies the phenotype variation due to the changes of primary DNA sequences, epigenetics is the study of heritable and reversible changes in accessibility and expression status of the underlying genetic information. Major epigenetic mechanisms include DNA methylation, histone modification, and noncoding RNAs (Henikoff, 2008).

DNA Methylation

DNA methylation in mammals is predominantly found on cytosine residues of CpG dinucleotides, between 60% and 90% of all CpGs are methylated (Ehrlich et al., 1982; Tucker, 2001). The genomic regions with high density of CpGs are referred to as CpG islands, and DNA methylation of these CpG islands correlates with transcriptional repression (Goll and Bestor, 2005). Unmethylated CpG islands often present in the 5’

regulatory regions of active or poised genes, whereby the underlying chromatins have H3K4 methylation.

DNA methylation may achieves the transcriptional repression state by two mechanisms. First, DNA methylation itself can prevent the binding of transcriptional factors to gene regulatory regions (Choy et al., 2010). And second, methylated DNA can be bound by the methyl-CpG-binding proteins (MBDs), which further recruit histone deacetylases and chromatin remodeling proteins, results in the compact and inactive chromatin structure called heterochromatin (Zhang et al., 1998a).

DNA methylation involves in many biological processes like centromere silencing, X-chromosome inactivation in female mammals, and mammalian imprinting (Yang and Kuroda, 2007). Abnormal DNA methylation like hypermethylation at promoter CpG islands of tumor suppressor genes are associated with tumorigenesis (Esteller et al., 2001).

Histone Modification

As afore-mentioned, nucleosome consists of an octamer of two copies of the histone proteins H2A, H2B, H3, H4, and 147bp of DNA sequences wrapped around the octamer complex. The histone tails are subject to a variety of post-translational modifications, among which methylation, acetylation and phosphorylation are the best studied. The histone modifications pattern can be extremely complex not only because the variety of modifications, but also because that lots of amino acids on the histone tails can be modified. For instance, lysine 4, 9, 14, 18, 27, 36, and arginine 2, 3, 8, 17, 26 on histone H3 tail can accept methyl or acetyl groups. Histone modifications have a direct impact on gene expression pattern, it is now understand that transcription activators will recruit histone acetyltransferases to acetylate histones, while transcription

repressors will recruit histone deacetyltransferases to deacetylate histones (Allis et al., 2007).

Recent advantages in ChIP-Chip and ChIP-Seq technologies (chromatin immunoprecipitation followed by microarray or high-throughput sequencing) have allowed the study of histone modifications on a genome-wide scale. And these studies have shown that certain histone modifications are consistently associated with certain expression patterns at certain regulatory regions. For example, H3K4me3 is associated with active promoters, H3K4me1 is associated with active and poised enhancers, H3K27ac is associated with active enhancers, and H3K36me3 is enriched in gene body specifically in exons. Whereas H3K27me3 is associated with repressed regions, and H3K9me3 is enriched at centromere and telomere regions (Guttman et al., 2009; Heintzman et al., 2009a; Heintzman et al., 2007; Ozsolak et al., 2008a; Won et al., 2008).

The Coordination between DNA Methylation and Histone Modifications in Gene Repression

Chromatin structure is well known to have a large impact on the pattern of gene expression during animal development. DNA methylation and histone modifications are both involved in the establishment of chromatin structure as well as gene expression regulation. Repressive histone methylations, like H3K27me3 and H3K9me3, will cause local formation of inaccessible heterochromatin structure, which is reversible due to developmental signals and environmental stresses, whereas DNA methylation will lead to stable repression pattern (Cedar and Bergman, 2009). The appropriate coordination between the transient and stable repression forces is crucial for animal development and responses to stimuli. For example, the early developmental genes like Hox genes

need to remain inactive after certain stages, whereas the tissue-specific genes and tumor suppressor genes need to be reactivated in certain cell types.

According to the recent model that the establishment of DNA methylation in early development is mediated by histone modification (Ooi et al., 2007), the H3K4 methylation at promoters and enhancers might be formed before *de novo* DNA methylation. It was proposed that H3K4 methylation is mediated by direct binding of RNA polymerase II, which recruits H3K4 methyltransferases (Guenther et al., 2007). As a result, *de novo* DNA methylation, which is carried out by the DNA methyltransferases enzymes, happens exclusively at CpG sites that do not have H3K4 methylation, since DNA methylation and H3K4 methylation are strongly anti-correlated (Meissner et al., 2008; Mohn et al., 2008; Weber et al., 2007).

Maintaining the repressed state of pluripotency genes during stem cell differentiation provides a good example for illustrating the coordination between histone modification and DNA methylation. The repression stability is achieved through three steps. In the first step, transcription is shut down by the repressor molecules which bind to promoter regions of pluripotency genes. This initial repression is reversible once the repressors are removed. In the next step, the histone deacetylation and histone methyltransferase lead to the formation of local heterochromatin regions marked by methylated H3K9. This new layer of repression provided by the change of chromatin structure is much more stable than repressor binding alone. However, it is not sufficient to maintain the repression stability against reprogramming, as the pluripotency genes which are silenced by H3K9 methylation alone can be reactivated during differentiation (Epsztejn-Litman et al., 2008; Feldman et al., 2006). The final step of pluripotency gene

inactivation involves DNA methylation to their promoters, after which the reactivation becomes almost impossible if the key factors are not artificially altered (Cedar and Bergman, 2009).

Noncoding RNAs

The mechanisms of noncoding RNAs in epigenetic regulation are less well understood than DNA methylation and histone modification. A generally accepted classification of noncoding RNAs is based on the length, which divides noncoding RNAs into small (less than 200 nucleotides) and long (more than 200 nucleotides) categories.

Small noncoding RNAs usually derive from large RNA precursors, and include microRNA (miRNA), short interfering RNA (siRNA), and so on. miRNAs are about 22 nucleotides in length, and are the products of imperfect hairpin structures in long noncoding RNA precursors or introns. They function by base-pairing with the complementary sequences within target mRNAs, and usually resulting in gene silencing through translational inhibition or mRNA degradation. It was recently reported that miRNA can regulate *de novo* DNA methylation in mouse embryonic stem cells (Benetti et al., 2008; Sinkkonen et al., 2008).

Long noncoding RNAs can sometimes reach more than 10kb in length, and they were initially considered as noisy transcription or artifacts from the contamination with genomic DNA or pre-mRNA (Mattick, 2005; Struhl, 2007). However it is now generally accepted that large amounts of such long noncoding RNAs exist in eukaryotes genomes. Long noncoding RNAs have an effect on transcription by regulating the activity or localization of transcription factors within cells, and by processing to various types of small regulatory RNAs. For example, as the first well characterized long noncoding RNA, HOTAIR was identified in human HOXC locus, it functions *in trans* on

HOXD locus by targeting PRC2 complex to HOXD and generating a transcriptionally repressed chromosomal domain (Rinn et al., 2007).

Epigenetic Regulation and Cancer

The concerted coordination of epigenetic mechanisms, including DNA methylation, histone modification, noncoding RNA and so on, forms an epigenetic regulatory network which dynamically adjusts gene expression pattern and cellular properties according to developmental signals and environmental stresses. Dys-regulation of this network will lead to various of diseases including cancer (Baylin and Ohm, 2006; Feinberg, 2007; Jirtle and Skinner, 2007; Rodriguez-Paredes and Esteller, 2011).

Dys-Regulation of Distinct Epigenetic Mechanisms Leads to Cancer

Cancer has been previously considered as a pure genetic disease. Due to the consistent study over the last decade, it becomes clear that epigenetic regulation is implicated in tumorigenesis. Aberrant DNA methylation, histone modification and nucleosome remodeling are the common epigenetic processes which take place during cancer development (Rodriguez-Paredes and Esteller, 2011). These abnormal disruptions of epigenome have been shown to lead to tumorigenesis by silencing tumor-suppressor genes and reactivating oncogenic retroviruses (Perry et al., 2010).

It has been widely observed that hypermethylation of site-specific CpG island promoters is responsible for the silencing of numerous tumor-suppressor genes (Esteller, 2007). Global DNA hypomethylation, which preferentially appears at repetitive sequences, can account for genomic instability (Esteller, 2008a) and the resuscitation of protooncogenes(Rodriguez-Paredes and Esteller, 2011).

Global mis-organization of histone modification is another hallmark of cancer. One of the most well-studied examples is that the global reduction of monoacetylation at

H4K16 and trimethylation at H4K20 appear early and accumulate in multiple primary tumors (Fraga et al., 2005). Abnormalities in global histone modification levels can also serve as a prediction for cancer recurrence and prognosis. For instance, lower global levels of H3K4me2 and H3K18ac are associated with higher risk of prostate cancer recurrence (Seligson et al., 2009), as well as lower survival probabilities in both lung and kidney cancer patients. And reduced global level of H3K9me2 is also prognostic of poorer outcome for prostate or kidney cancer patients (Seligson et al., 2009).

An increasing number of cases have shown that nucleosome remodeling caused by the loss of specific landmarks or proteins is another reason for the aberrations in epigenomic landscape of cancer. For example, recent evidences have demonstrated that the loss of the insulator protein CTCF binding will lead to the spreading of facultative heterochromatin into the genebody of tumor-suppressor genes p16 (Witcher and Emerson, 2009) and p53 (Soto-Reyes and Recillas-Targa, 2010), eventually cause the silencing of these tumor-suppressor genes. This kind of cancer progression through the loss of protection against heterochromatin propagation has no direct correlation with DNA methylation (Soto-Reyes and Recillas-Targa, 2010). And traditional epigenetic drugs such as DNA demethylating agent 5'-AZA-2'-deoxycytidine can neither permanently restore p16 expression (Esteller, 2007) nor reverse CTCF binding (Witcher and Emerson, 2009).

PcG Proteins and Cancer

Dysregulation of PcG components were also found in variety of cancers. For example, EZH2 was up-regulated in a wide range of hematopoietic and solid human malignancies, such as lymphoma, breast cancer, prostate cancer, colon cancer, and so on (Kleer et al., 2003; Mimori et al., 2005; van Kemenade et al., 2001; Varambally et al.,

2002; Visser et al., 2001). And EZH2 was also found to consistently over-express in metastatic prostate cancer compared to localized prostate cancer or normal tissues (Varambally et al., 2002). The elevated activity of EZH2 causes tumorigenesis by ectopic repression of tumor suppressor genes, for example, *DAB2IP* and *MSMB* were silenced by EZH2 in prostate cancer (Beke et al., 2007; Chen et al., 2005).

Several recent large scale transcriptome/exome studies, which aimed at identifying the tumor genetic abnormalities, have found that histone modifiers, such as the histone H3K27 demethylase UTX, were frequently mutated in a variety of cancers (Dalgliesh et al., 2010; Gui et al., 2011; van Haaften et al., 2009).

In addition, tumor cells may be mis-specified by PcG proteins to adopt stem cell properties, likely through the PcG-mediated silencing of genes responsible for differentiation and lineage specification (Bernstein et al., 2006a; Caretti et al., 2004; Ezhkova et al., 2009; Lee et al., 2006). The well known phenomena that tumor cells and stem cells share some common properties like extensive proliferation capacity and differentiation potential, lead to the “cancer stem cell” hypothesis (Pardal et al., 2003; Sparmann and van Lohuizen, 2006). Although this hypothesis is still debatable, it does not affect the fact that tumorigenesis can be initiated by PcG silencing of developmental genes.

P53 and Cancer

The origin of mammalian p53 family proteins (p53, p63 and p73) can probably go back as far as the divergence of animalia and fungi approximately 2 billion year ago (Fernandes and Atchley, 2008). p53 is unique than the other two mammalian p53 family proteins because of its prominent role as tumor suppressor. In addition, p53 also involves in regulating cellular responses upon stress and DNA damage as well as many

other physiological processes like stem cell maintenance, development, etc (Junttila and Evan, 2009). Although the origin of mammalian p53 is early, its role as tumor suppressor is likely a relative recent adaption, which is only needed for large, long-lived organisms. Several mysteries about its tumor suppression function are still need to be addressed, such as how p53 distinguish tumor cells from normal cells, when p53 is activated during tumorigenesis, and what signals cause the loss of p53 function in cancers. The understanding of these questions will improve the therapeutic efficacy in the cancer treatment.

Tumor Suppression Function of P53

In normal cells, the interaction between p53 and MDM2, MDM4 results in the steady and low activity level of p53, which is sufficient for most of the physiological functions of p53. In contrast, oncogenic and DNA damage signals will interrupt the interaction between p53 and MDM2, MDM4, which triggers the transcriptional activity of p53 and causes its accumulation and dramatically increased activity. The rapid induction of p53 activity by interrupting its interaction with MDM2 and MDM4 is vertebrate specific, as no counterparts of MDM2 or MDM4 exist in invertebrates (Brodsky et al., 2000; Nordstrom and Abrams, 2000).

It is still unclear whether the roles of p53 in tumor suppression, stress or DNA damage responses and other physiological processes are independent or not. The evidence that p53's role in DNA damage responses is evolutionarily conserved, has led to the concept that the tumor suppression role of p53 is achieved by antagonizing DNA damage and so preserving the genome integrity and preventing the accumulation of oncogenic mutations (Junttila and Evan, 2009). The idea was further sponsored by the evidences that dramatic genome instability shown in many cancer cells (Donehower et

al., 1995), and that oncogenic signals can induce DNA damage in some circumstances (Di Micco et al., 2006).

However, the concept that DNA damage functions as the principal trigger of p53-mediated tumor suppression has recently been challenged by the work in mouse shown that the tumor suppression function regulated by p53 is entirely $p19^{ARF}$ dependent (Christophorou et al., 2006). $p19^{ARF}$ is encoded by an alternative open-reading frame within the *Ink4a-Afr* locus (Quelle et al., 1995) and can activate p53 activity by inhibiting its interaction with MDM2 (Pomerantz et al., 1998; Stott et al., 1998; Zhang et al., 1998b). More importantly, only oncogenic signaling can specifically induce $p19^{ARF}$, but DNA damage could not (Christophorou et al., 2006; Kamijo et al., 1997; Zindy et al., 2003), which implies that oncogenic signaling is the mechanistic feature of tumor cells whereas DNA damage is dispensable.

p53 is a sequence-specific transcription factor that regulates many downstream genes, among which the induction of apoptotic genes is the predominant role p53 plays in tumor suppression and DNA damage responses. The mechanisms include p53-mediated induction of proteins (Marchenko et al., 2000), such as Bax, NOXA, PUMA. In addition, p53 may also induce apoptosis through relocation of death receptors, such as Fas and DR5 to the cell surface (Bennett et al., 1998); and regulation of translation by direct binding to the 5' UTR of certain genes (El-Deiry, 1998).

P53 and Cancer Therapy

The predominant role of p53 in tumor suppression makes it a potential target for cancer therapy. Most of the early efforts focused on gene therapy approach that transfers p53 to cancer patients that lack functional p53 by virus vectors (Peng, 2005; Roth et al., 1996; Senzer and Nemunaitis, 2009). A different strategy focusing on the

development of low-molecular-mass compounds that restore p53 activity is also under conduct. To this end, efforts have evolved to produce compounds that interact with mutant p53 proteins in tumor cells and restore their function by altering their conformation (Boeckler et al., 2008; Bykov et al., 2002); as well as to produce compounds that enhance p53 activity by disrupting its interaction with MDM2 (Vassilev, 2007; Vassilev et al., 2004). p53 mutations can also potentially predict patients prognosis (Aas et al., 1996; Young et al., 2008), however these expectations have not been fulfilled because of the genetic complexity and extensive diversity of individual tumors (Levine and Oren, 2009).

ChIP-Seq: Genome-Wide Monitoring of Epigenetic Regulation

ChIP-Seq combines chromatin immunoprecipitation with next-generation high-throughput parallel DNA sequencing to identify the binding sites of transcription factors and effective domains of histone modifications. Compared with ChIP-Chip (chromatin immunoprecipitation followed by microarray), ChIP-Seq has several advantages such as higher throughput, higher resolution, and lower systematic bias. Due to the recent development in next-generation sequencing, a lot of genome-wide epigenetic datasets became available for the model organisms as well as others. Among these, ENCODE (encyclopedia of DNA elements) (2004) and modENCODE (model organism ENCODE) (Roy et al., 2010) projects provide a huge amount of whole genome ChIP-Seq as well as other types of datasets for human, mouse, *Drosophila* and *C.elegans* in different cell lines, tissues and developmental stages (Bernstein et al., 2012; Djebali et al., 2012; Gerstein et al., 2012; Neph et al., 2012; Sanyal et al., 2012; Thurman et al., 2012).

Bioinformatics Programs for ChIP-Seq Data Analysis

ChIP-Seq reads enrichment regions or ChIP-Seq peaks can generally be classified into three categories: sharp, broad and mixed. Sharp peaks of a few hundred base pairs width usually appear in DNA-binding protein profiles, which are usually used for the identification of transcription factor binding sites. While broad ChIP-Seq enrichment regions, which can span from several kilobases to even hundreds of kilobases covering tens of genes, are usually found in the histone modification profiles that mark transcription active or repressed regions.

Most of the current algorithms, such as MACS (Zhang et al., 2008b), PeakSeq (Rozowsky et al., 2009), etc, are specifically developed to search for peaks or transcription factor binding sites in the DNA-binding protein profiles. Several comprehensive comparisons have been conducted among these methods, but no consensus has been reached, partially due to the lack of well defined evaluation datasets (Laajala et al., 2009; Pepke et al., 2009; Qin et al., 2010; Wilbanks and Facciotti, 2010). Although it is possible to apply these peak-calling methods to histone modification profiles, there are obvious drawbacks. For example, the focus of the analysis of histone modification datasets is to identify a specific chromatin pattern that often spans a long range (Hon et al., 2008). However, peak calling for transcription factors is usually focused on much narrow peaks. This kind of problem can be clearly seen as the average width of the peaks identified by the peak-calling methods are in the range of 100bp to 3000bp when dealing with H3K27me3 profiles (Qin et al., 2010), while the real enrichment regions should be tens of kilobases wide.

Besides the methods specific for the detection of peaks in DNA-binding protein profiles, there are a few algorithms developed to identify the broad, low-intensity

enrichment regions in histone modification ChIP-Seq data, such as SICER (Zang et al., 2009), RSEG (Song and Smith, 2011), ChIPDiff (Xu et al., 2008), CCAT (Xu et al., 2010), ChromaSig (Hon et al., 2008) and Models 1-3 (Johannes et al., 2010). However, ChIPDiff and Models 1-3 are aimed to characterize the differential histone modification sites, and they both require ChIP-Seq data from multiple cell lines or developmental stages. CCAT is specified to identify the weak ChIP signals from background noise and it requires an input control profile.

SICER (spatial clustering approach for the identification of ChIP-enriched regions) is the mostly cited method to identify broad histone modification enrichment domains. It utilizes the clustering approach to identify candidate spatial islands of enriched signals and reports the islands with significant scores by comparing to a random background model. The other widely used program RSEG instead applies the two-state HMM (hidden Markov model) and also provides specific boundary calling function.

CHAPTER 2
HISTONE MODIFICATIONS, DNA ACCESSIBILITY, AND P53 BINDING PROFILES FOLLOWING DNA DAMAGE

Introduction

The origin of mammalian p53 family proteins (p53, p63 and p73) can probably go back as far as the divergence of animalia and fungi approximately 2 billion year ago (Fernandes and Atchley, 2008). p53 is unique than the other two mammalian p53 family members (p63 and p73) because of its prominent role as tumor suppressor. It is involved in the regulation of cellular responses upon oncogenic stresses such as DNA damage as well as many other physiological processes like stem cell maintenance, development, etc (Junttila and Evan, 2009). In normal cells, the interaction between p53 and MDM2/MDM4 results in the steady and low activity level of p53, which is sufficient for most of the physiological functions of p53. In contrast, oncogenic and DNA damage signals will interrupt the interaction between p53 and MDM2/MDM4, which triggers the transcriptional activity of p53 and causes its accumulation and dramatically increased activity. The rapid induction of p53 activity by interrupting its interaction with MDM2 and MDM4 could be vertebrate specific, as no counterparts of MDM2 or MDM4 has been identified in invertebrates (Brodsky et al., 2000; Nordstrom and Abrams, 2000). p53 is a sequence-specific transcription factor that regulates many downstream genes, among which the induction of apoptotic genes is the predominant role p53 plays in tumor suppression and DNA damage responses.

Nucleosome consists of a histone octamer which has two copies of the histone proteins H2A, H2B, H3, H4, and 147bp of DNA sequences wrapped around the complex. The histone tails are subject to a variety of post-translational modifications, whose pattern can be extremely complicated not only because the variety of

modifications there exist, but also because that lots of amino acids on the histone tails can be modified. Although single histone modification can change the chromatin property, the combination of histone modifications is believed to be the ultimate determinant of the chromatin structure. Recent advantages in ChIP-Chip and ChIP-Seq technologies (chromatin immunoprecipitation followed by microarray or sequencing) have allowed the study of histone modifications on a genome-wide scale. And these studies have shown that certain histone modifications are consistently associated with regulatory regions that controls or influences gene expression. For example, H3K4me3 is associated with active promoters, H3K4me1 is associated with active and poised enhancers, H3K27ac is associated with active enhancers, and H3K36me3 is enriched in gene body specifically in exons. Whereas H3K27me3 is mostly associated with repressed regions, and H3K9me3 is typically enriched at centromere and telomere regions (Guttman et al., 2009; Heintzman et al., 2009a; Heintzman et al., 2007; Ozsolak et al., 2008a; Won et al., 2008).

In our quest to understand what controls cellular sensitivity to p53-mediated pro-apoptotic response following DNA damage, we found that epigenetic regulation plays a pivotal role in controlling the responsiveness of pro-apoptotic genes. It has long been observed in *Drosophila* that while cells in early embryogenesis (stage 9-11) are extremely sensitive to irradiation induced apoptosis, cells in late stage embryos became very resistant, even though development cell death can still occur (Figure 2-1). Three pro-apoptotic genes, *reaper*, *hid*, and *sickle*, are rapidly induced within 15-20 minutes following DNA damage in early stage embryos (Lin et al., 2011; Zhang et al., 2008a). This induction is dependent on p53 and the regulatory region IRER (Irradiation

Responsive Enhancer Region) (Zhang et al., 2008a) (Figure 2-1). Remarkably, IRER is not only required for irradiation-induced expression of *reaper* and *sickle*, but also required for the responsiveness of *hid*, which is located ~210 kb away from *reaper*. In Df(IRER) mutant embryos, none of the three pro-apoptotic genes are responsive to DNA damage although the overall development expression patterns of the three genes are largely unchanged.

The IRER is in an “open” state in most cells during early embryogenesis, and consequently the three genes are very sensitive to irradiation and activated p53. However, during late embryogenesis, when most cells enter into post-mitotic differentiation, IRER becomes enriched for suppressive histone modifications H3K27me3 and H3K9me3. Correspondingly, this region is bound by Polycomb group proteins and HP1 (heterochromatin protein 1). As a consequence, DNA in IRER became inaccessible as measured by DNase I sensitivity assay. This epigenetic blocking of IRER renders all three pro-apoptotic genes unresponsive to irradiation, even though the responsiveness of other p53 target genes, such as Ku70 & Ku80, are unaffected at this stage (Zhang et al., 2008a). In embryos lacking the function of key Polycomb group proteins, this sensitive to resistant transition is significantly delayed.

Interestingly, the epigenetic blocking of the IRER is strictly limited to the IRER but never reaches the transcribed region of *reaper*, which is expressed in neural stem cells at later stage embryos (Lin et al., 2011; Zhang et al., 2008a). This is in sharp contrast with canonic PcG-mediated epigenetic regulation of homeotic genes, in which the whole transcribed region of the targeted gene is marked by repressive histone modifications. This restriction of heterochromatin formation is functionally significant. While *reaper*

becomes unresponsive to irradiation in later stage embryos, it is expressed in response to developmental cues and required for programmed neuroblast cell death in late embryogenesis. Work from Kristine White's group has shown that the later stage developmental expression of *reaper* is mediated by the intergenic Neuroblast Regulatory Region (NBRR) located downstream of *reaper* (Figure 2-1)(Tan et al., 2011).

Our data indicated that epigenetic regulation of the enhancer region mediating p53-indued pro-apoptotic gene expression could serve as an important mechanism to regulate cellular sensitivity to DNA damage during cellular differentiation. Given the fact that the role of p53 in mediating DNA damage-induced cell death is highly conserved, we are curious as to whether the enhancer-specific epigenetic regulation also plays a role in controlling cellular response to activated p53 in mammalian cells.

A recent study of p53-mediated DNA damage signaling provided a whole-genome map of p53 binding profiles in mouse ES cells (genotype 129X1/129S1) in both normal condition and DNA damaged condition induced by adriamycin (Adr) treatment (Li et al., 2012). Unfortunately, DNA accessibility and histone modification data was not available for mES cells of the same genotype. However, genome-wide DNA accessibility and histone modification profiles were available for the C57BL/6 genotype, generated by the ENCODE project (Shen et al., 2012). This provided us an opportunity to test whether histone modifications and DNA accessibility affect the binding profile of p53 in the mES cells, which has been rarely studied before. In this work, we tried to use these datasets to ask whether and how the histone modifications and DNA accessibility correlate with the binding profile of p53 upon DNA damage.

Materials and Methods

Dataset

The datasets used in this study are listed in Table 3-1.

Data Processing

For histone modification ChIP-Seq and MNase-Seq datasets, we first applied MACS (Zhang et al., 2008b) to count the number of tags that fell within the 10bp non-overlapping bins. Then we normalized the reads number in each bin to the total reads number, and then subtracted the normalized input signal from the normalized ChIP-Seq signal.

Mean Signals around P53 Binding Sites

Bx-python (Blankenberg et al., 2011) function (“aggregate scores in intervals”) was applied to calculate the average ChIP-Seq signal around p53 binding sites with window size as 200bp and step size as 20bp.

Grouping of P53 Binding Sites Based on DNA Accessibility

In order to robustly divide p53 binding sites based on DNA accessibility, we used two independent MNase-Seq datasets GSM769010 (Shen et al., 2012) and GSM849958 (Hu et al., 2013), which both measures the DNA accessibility in mouse ES cells with high resolution. And we used their average signals in the 800bp region surrounding the midpoint of a p53 binding site as a measure of its DNA accessibility. The high DNA accessibility group was defined as the p53 binding sites whose average MNase-Seq signals are both below their respective 10% quantile , while the low DNA accessibility group should have both signals in their respective top 10% percent. And the medium DNA accessibility group should satisfy that the first MNase-Seq

(GSM769010) signal is between 30% and 70% quantile, and the second MNase-Seq (GSM849958) signal is not below 10% or above 90% quantile.

Predict P53 Binding Sites using TransFac P53 Motif

The consensus p53 binding motif (V$P53_05) was extracted from TRANSFAC database (Wingender et al., 1996). And 67425 p53 binding sites were predicted in mouse genome (mm9) by MATCH (Kel et al., 2003) when setting both matrix and core cutoffs as 0.8.

Results

Constitutive and Conditional Intergenic P53 Binding Sites in Mouse ES Cells

A recent whole-genome study of p53-mediated DNA damage signaling provides a comprehensive map of p53 binding profiles in mouse ES cells in both normal condition and DNA damage condition induced by adriamycin (Adr) treatment (Li et al., 2012). With the peak-calling software MACS, Li et al. identified 7820 and 54564 p53 binding sites in normal and Adr-treated conditions, respectively (Li et al., 2012). We found that the majority (7778, 99.7%) of p53 binding sites in normal condition have a corresponding binding site in the Adr treated condition (Figure 2-2A). We thus named these as constitutive p53 binding sites. The remaining (46786) p53 binding sites were only detectable following Adr treatment. For the clarity of discussion we will henceforth refer those as conditional (Adr-induced) p53 binding sites.

Since p53 binding in promoter regions and transcribed regions can be affected by other transcription factors and the transcription process, we chose to focus on intergenic p53 binding sites that are at least 5 kb away from any TSS and not overlap with any annotated transcribed region. A total of 3640 constitutive and 19391 conditional p53 binding sites were identified as intergenic (Figure 2-2B). The number of intergenic

constitutive sites is about half of all p53 binding sites identified by the ChIP-Seq experiments. We believe that this set of binding sites provide a relatively clean context for us to probe how p53 binding may correlate with histone modifications and DNA accessibility in the genome scale.

Correlation between Histone Modification and DNA Accessibility in the C57BL/6 Mouse ES Cells

We first tried to assess whether we can verify the correlation between DNA accessibility and histone modification in the same mouse ES cell line. As show in table 2-1, the DNA accessibility data and histone modification data were generated by the Ren group as part of the ENCODE project (Shen et al., 2012). There is also MNase-Seq dataset in mouse ES cells from the Zhao group (Hu et al., 2013) (see Methods for detail). Based on these two data sets, we grouped the p53 binding sites that fall into high (top 10%), medium (30-70 percentile), and low (bottom 10%) DNA accessibility groups (Figure 2-3).

Instead of directly using histone modification ChIP-Seq signal, we applied the data processing method of ENCODE project (Shen et al., 2012), as we first normalized the IP and input signal to their total number of reads respectively, then subtracted the normalized input signal from the normalized IP signal. With this extra step, we eliminated the bias in histone modification signal introduced by the inconsistency in the underlying DNA accessibility.

When we plotted the histone modification ChIP-Seq value around the three groups of p53 binding sites, we found that DNA accessibility in those sites positively correlates with the active histone modifications like H3K27ac, H3K4me1/3, but negatively correlates with repressive marks like H3K27me3 and H3K9me3 (Figure 2-4). Overall,

the correlation between DNA accessibility and different histone modifications can be verified when using the datasets generated by the same laboratory with the mouse ES cells that have the same genetic background.

Lack of Correlation between DNA Accessibility and P53 Binding

We then tried to assess whether there is a correlation between DNA accessibility and p53 binding by extrapolating the DNA accessibility data to compare with the p53 binding data. We grouped the p53 binding sites into high (top 10%), medium (30-70 percentile), and low (bottom 10%) DNA accessibility groups as mentioned above (Figure 2-3). Interestingly, when DNA accessibility surrounding constitutive sites were compared against those surrounding the conditional sites, there was little difference between the two for any of the three groups of DNA accessibility (Figure 2-3). This lack of correlation is surprising since the level of p53 binding following Adr treatment was significantly higher at the constitutive sites (Figure 2-5).

We then proceeded to see whether DNA accessibility have an impact on DNA-damage induced p53 binding by plotting and compare the number of p53 ChIP-Seq reads following Adr treatment for these three different accessibility groups.

Although 19391 more p53 binding sites could be detected following DNA damage, their average binding intensity is much lower than that of the constitutive binding sites (Figure 2-5). More interestingly, the average intensity curves for the three DNA accessibility groups are almost overlapping in Adr treated condition (Figure 2-5B). This suggests that, for the dataset we analyzed, the number of p53 ChIP-Seq reads following DNA damage has no significant difference between those that are in high accessibility region vs. those that are in low accessibility region.

Correlation between Active Enhancer Marker H3K27ac and Constitutive P53 Binding Sites

The active histone modification H3K27ac has been found to distinguish active enhancers from inactive/poised enhancers containing H3K4me1 alone (Creyghton et al., 2010). We reasoned that the constitutive and conditional p53 binding sites contain active and poised enhancer elements respectively, although it requires intensive experiments to decide the exact identity. Indeed, the signal of active enhancer marker H3K27ac is higher around constitutive p53 binding sites than conditional ones, and the difference is most prominent in the high DNA accessibility group (Figure 2-4). Comparatively, the enhancer marker H3K4me1, promoter marker H3K4me3 and suppressive histone modifications H3K27me3 and H3K9me3 do not have significant difference between constitutive and conditional p53 binding sites (Figure 2-4).

P53 Binding Sites with Lower DNA Accessibility Tend to Contain Consensus P53 Binding Motif

We also interested in whether the underlying DNA sequence has difference for the different groups of p53 binding sites. We identified consensus p53 binding sites in the mouse genome using the p53 binding site matrices from TRANSFAC database (Wingender et al., 1996). In general greater portion of constitutive p53 binding sites contain consensus p53 motif than the conditional ones for all the three p53 groups based on DNA accessibility (Table 3-2). Interestingly, we found that the p53 binding sites within low DNA accessibility region are more likely to contain the consensus p53 motif. This is true for both conditional and constitutive binding sites (Table 3-2). These observations imply that conditional p53 binding sites, especially those with high DNA accessibility, may interact with DNA through different binding motifs or even through indirect interaction.

Discussions

The repressive histone modifications like H3K27me3 and H3K9me3 are supposed to mark the heterochromatin regions, which have compact chromatin structure and prevent the binding of transcription factors. However, with the data set that were available to us at the time of our work, we found that only the active enhancer mark H3K27ac shows significant difference between conditional and constitutive p53 binding sites in DNA accessibility high and medium groups, which is consistent with the previous finding that H3K27ac distinguishes active enhancers from poised ones (Creyghton et al., 2010). The intensities of other histone modifications, like the repressive H3K27me3, H3K9me3 and the active H3K4me1/3 have no detectable difference between conditional and constitutive p53 binding sites, no matter the DNA accessibility level. Also, DNA accessibility itself does not seem to have significant impact on p53 binding intensity.

The main limitation of our analysis is that the histone modification and DNA accessibility profiles were generated from the mouse strain that is genetically different with the one used to study p53 binding profiles. In addition, different experimental protocols and fragmentation methods were used. We note that histone modification profiles like H3K27me3 generated for the mES cells with different genotype and produced by different labs can have dramatic difference (Figure 2-6). So it would give us more confidence about any conclusion if the p53 and histone modification profiles generated from the same laboratory based on the same genetic background are available in the future.

With the limitation standing, our analysis failed to support the hypothesis that heterochromatin regions with low DNA accessibility prevent p53 binding upon DNA

damage. Since p53 binding sites are constrained to a relatively narrow space, not like other transcription factors, for example CTCF, whose binding occupies much longer DNA sequences. The chromatin structure may not have a significant impact on this kind of localized narrow binding.

Table 2-1. List of datasets used in this study

Dataset	GEO #	Cell line	Genetic background	Reference
p53 control	GSM647224	mES	129X1 x 129S1	(Li et al., 2012)
P53 Adr8h	GSM647225	mES	129X1 x 129S1	(Li et al., 2012)
Mnase-Seq	GSM769010	mES-Bruce4	C57BL/6	(Shen et al., 2012)
Mnase-Seq	GSM849958	mES	129S6/SvEvTac	(Hu et al., 2013)
H3K27me3	GSM1000089	mES-Bruce4	C57BL/6	(Shen et al., 2012)
H3K9me3	GSM1000147	mES-Bruce4	C57BL/6	(Shen et al., 2012)
H3K27ac	GSM1000099	mES-Bruce4	C57BL/6	(Shen et al., 2012)
H3K4me1	GSM769009	mES-Bruce4	C57BL/6	(Shen et al., 2012)
H3K4me3	GSM769008	mES-Bruce4	C57BL/6	(Shen et al., 2012)
H3K27me3	GSM970531	mES	129SvJae/C57BL/6	(Jia et al., 2012)

Table 2-2. Percentage of p53 binding sites that contain consensus p53 binding motif

p53 binding sites	DNA accessibility Group		
	Low	Medium	High
Conditional p53	239 (31.5%)	1250 (18.1%)	67 (10.6%)
Constitutive p53	69 (56.6%)	654 (50.0%)	37 (35.2%)

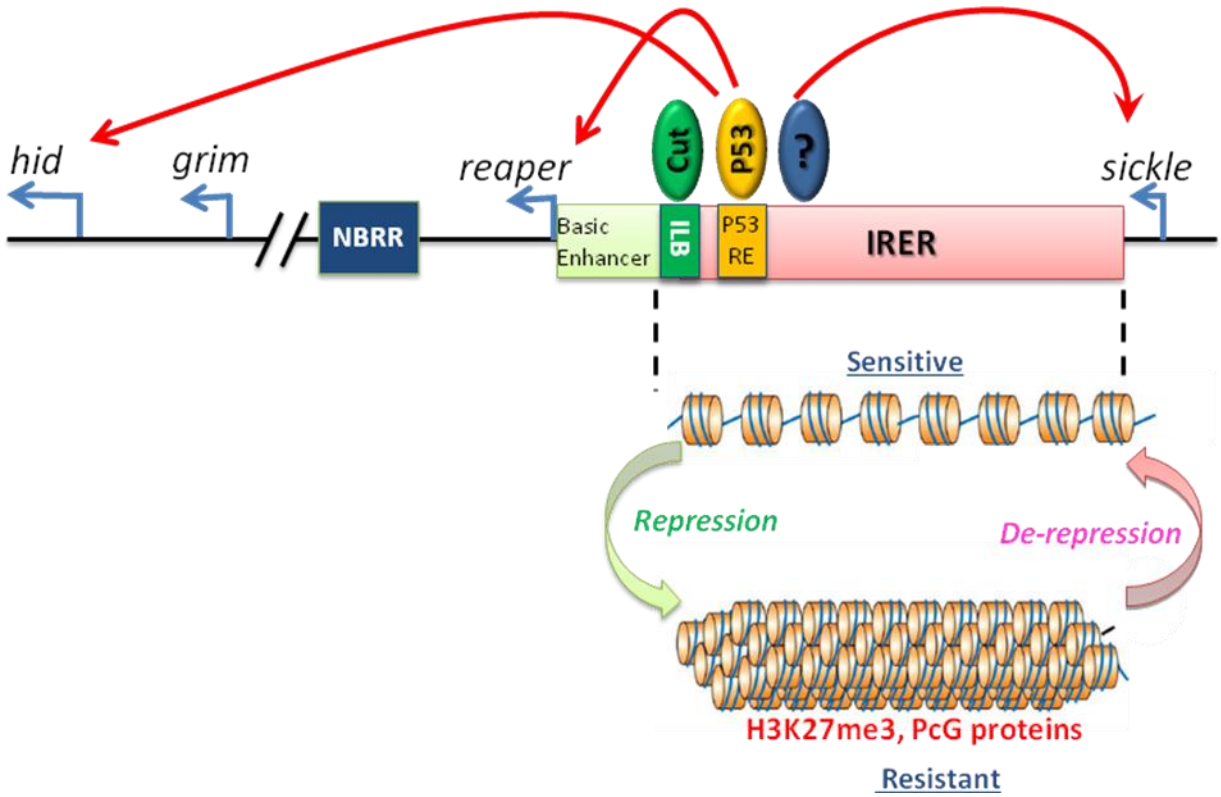


Figure 2-1. Schematic diagram summarizing previous findings regarding the IRER and ILB in our lab. Previous work from our lab has mapped the irradiation responsive enhancer region (IRER) to the 33kb intergenic sequence on the 3rd chromosome. It is located between two pro-apoptotic genes *reaper* and *sickle*, including the putative p53-response element (P53 RE). This enhancer region is subject to PcG-mediated epigenetic regulation and undergoes an open-to-closed transition during embryonic stage 11-12. The open chromatin structure in early stage embryos facilitates irradiation-induced transcription of *reaper* and *hid* and leads to apoptosis; whereas the condensed chromatin in late staged embryos precludes transcription and blocks apoptosis. The facultative heterochromatin formation is restricted to IRER by the IRER left barrier (ILB), which allows the *reaper* promoter to remain open throughout development and accessible to regulation. The barrier activity requires binding of the Cut protein, which may recruit chromatin modifying enzymes such as CBP; mechanistically, much remains to be elucidated. Modified from (Lin et al., 2011; Tan et al., 2011; Zhang et al., 2008a).

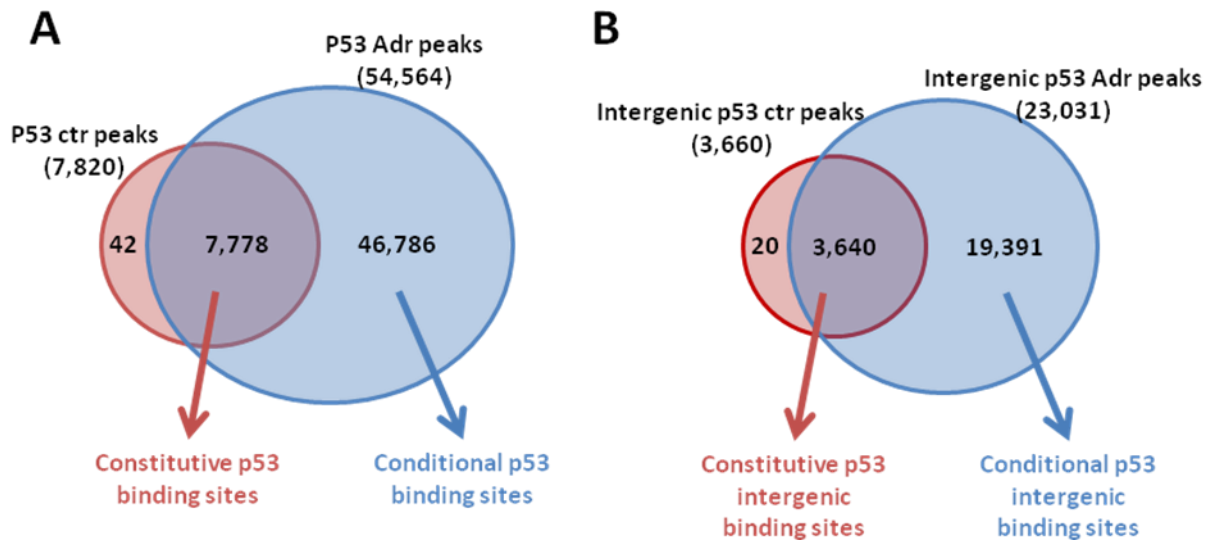


Figure 2-2. Constitutive and conditional p53 binding sites. Venn diagrams show the number of p53 binding sites in untreated (ctr) condition and Adr treated conditions, and the definition of constitutive and conditional p53 binding sites genome-wide (A), or in intergenic regions (B).

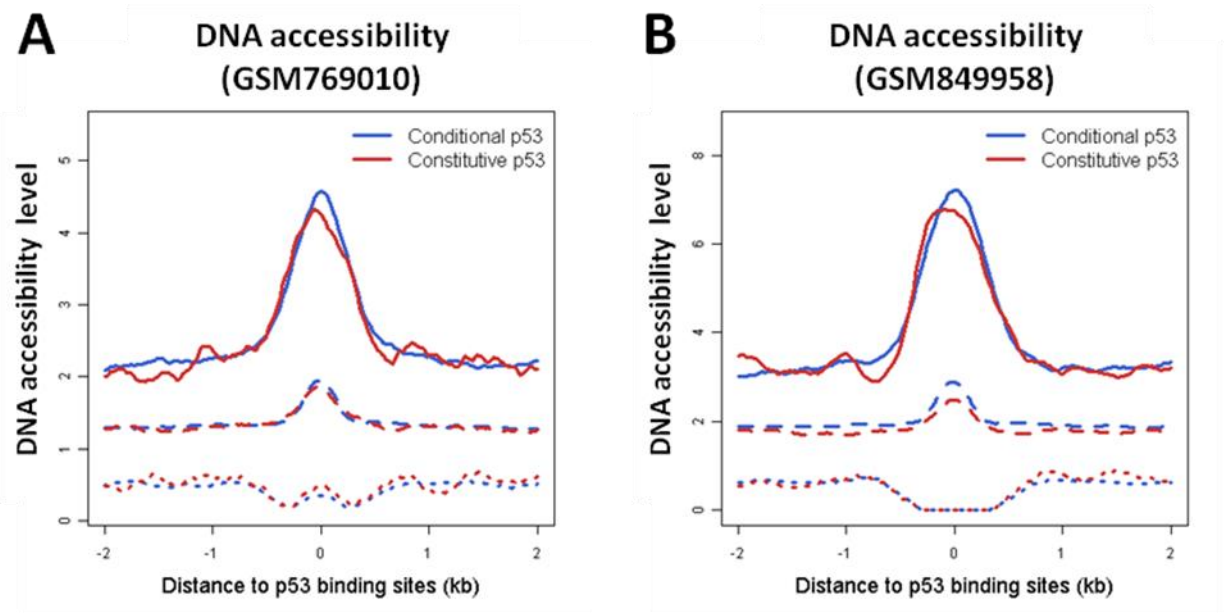


Figure 2-3. DNA accessibility around p53 binding sites. The signals from two independent MNase-Seq datasets (A) GSM769010, and (B) GSM849958, around conditional and constitutive p53 binding sites in intergenic regions. The solid, dashed and dotted lines indicate the DNA accessibility high, medium and low groups respectively.

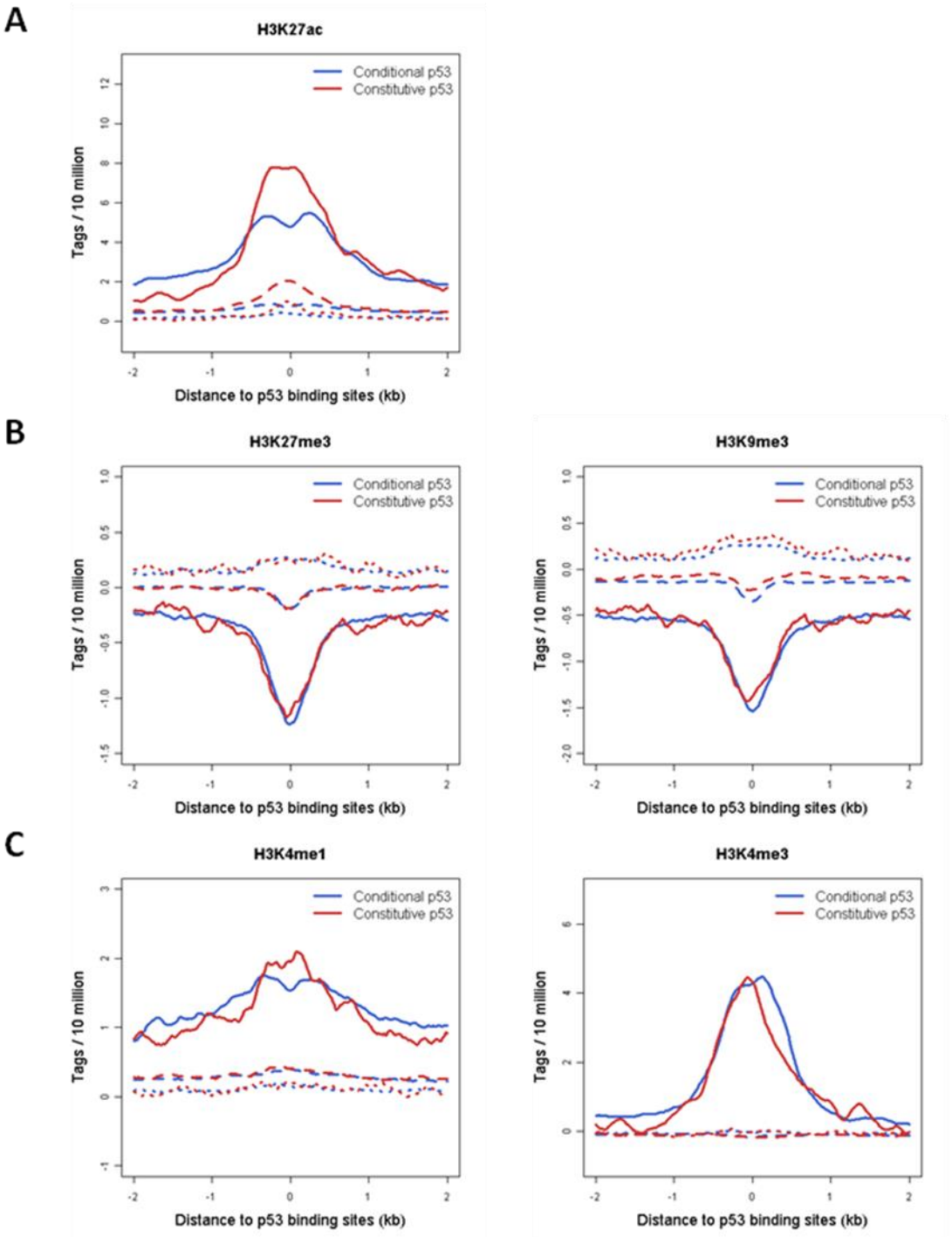


Figure 2-4. Histone modifications around p53 binding sites. Histone modifications, including the active enhancer mark H3K27ac (A), repressive marks H3K27me3, H3K9me3 (B), and active marks H3K4me1/3 (C) around conditional and constitutive p53 binding sites in intergenic regions. The solid, dashed and dotted lines indicate the DNA accessibility high, medium and low groups respectively.

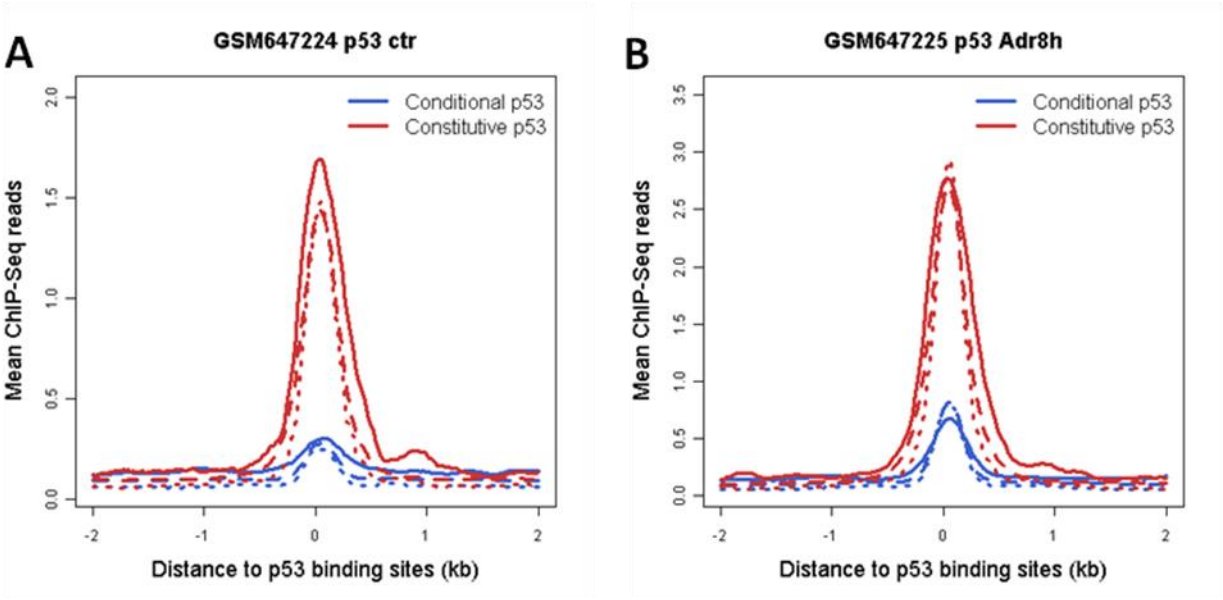


Figure 2-5. Binding intensities of conditional and constitutive p53 in untreated and Adr treated conditions. Average p53 ChIP-Seq reads number for conditional and constitutive p53 binding sites in control (A) and Adr treated (B) conditions. The solid, dashed and dotted lines indicate the DNA accessibility high, medium and low groups respectively.

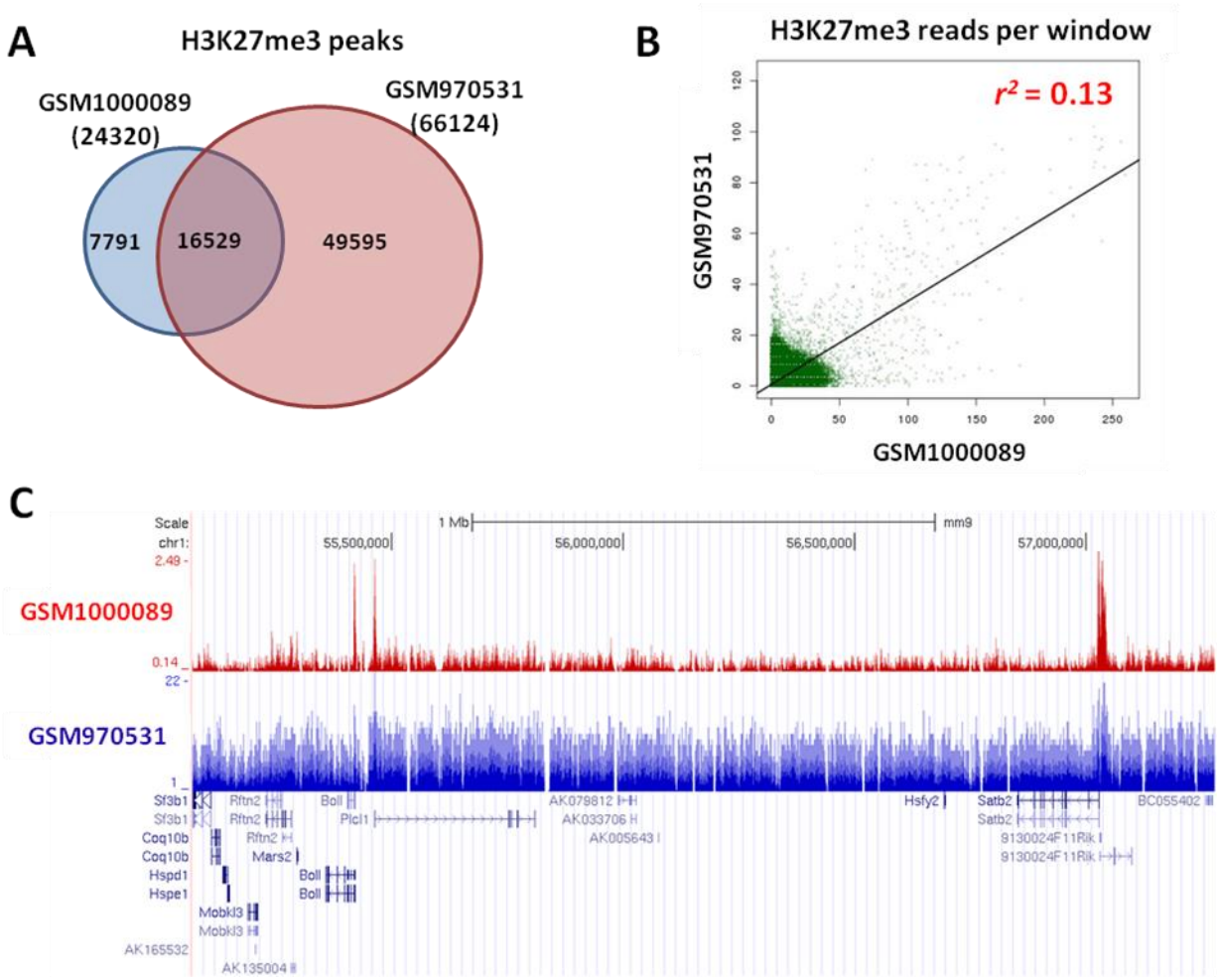


Figure 2-6. H3K27me3 ChIP-Seq profiles from different laboratories are not comparable. (A) Venn diagram shows how many H3K27me3 enriched domains are overlapping between the two datasets. (B) Scatter plot of ChIP-Seq signals from the two datasets in 200bp windows. Linear regression shows the coefficient of determination is only 0.13. (C) A representative region shows how different the two H3K27me3 datasets could be.

CHAPTER 3
GENOME-WIDE IDENTIFICATION OF CHROMATIN TRANSITIONAL REGIONS REVEALS DIVERSE MECHANISMS DEFINING THE BOUNDARY OF FACULTATIVE HETEROCHROMATIN

Introduction

Site-specific formation of facultative heterochromatin, mediated by PcG (Polycomb group) proteins, plays a fundamentally important role in controlling cellular differentiation and in defining the property of differentiated cells. The suppressive histone modification mark, H3K27me3, is catalyzed by Polycomb repressive complex 2 (PRC2). This suppressive modification has strong affinity to, and is usually bound by, Polycomb repressive complex 1 (PRC1). The interaction between PRC1 and PRC2 leads to the propensity to spread this suppressive histone modification until it is antagonized (reviewed in (Muller and Verrijzer, 2009; Schwartz and Pirrotta, 2007)). Although certain strong promoters of active genes can prevent the formation of facultative heterochromatin (Raab et al., 2012), under many circumstances, specialized DNA elements called chromatin barriers or barrier insulators are needed to demarcate the boundary of facultative heterochromatin (reviewed in (Gaszner and Felsenfeld, 2006)).

Insulators, such as the *gypsy* insulator, were originally identified for their enhancer-blocking activity, i.e. blocking the interaction between the enhancer and promoter when placed in-between (Geyer et al., 1986). Later, it was revealed that most of them also have barrier activity (Kahn et al., 2006; Roseman et al., 1993), i.e. blocking the propagation of repressive histone modifications. It was not clear whether the two activities are separable until the characterization of the cHS4 insulator in the chicken β-globin locus. The complete cHS4 has both enhancer-blocking and barrier activity. However, a series of mechanistic studies indicated that the two activities are separable

and carried out by distinct DNA elements. The enhancer-blocking activity of cHS4 is mediated by CTCF, while its barrier activity against heterochromatin formation requires a binding site for USF1 (Upstream Stimulatory Factor 1). Binding of USF1 to cHS4 recruits chromatin-modifying enzymes that catalyze histone modifications incompatible with heterochromatin formation, thus preventing the propagation of suppressive histone modification (Huang et al., 2007; West et al., 2004).

Recently, a novel chromatin barrier that lacks any detectable enhancer-blocking function has also been identified in *Drosophila* (Lin et al., 2011). This ~200bp element is located at the left boundary of IRER (Irradiation Responsive Enhancer Region), a 33 kb intergenic regulatory region controlling stress-induced expression of multiple pro-apoptotic genes (Zhang et al., 2008a). When tested in transgenic animals, ILB (IRER Left Boundary) is fully capable of blocking the propagation of H3K27me3 initiated by a strong Polycomb response element (PRE) (Lin et al., 2011). The chromatin barrier function of ILB is evolutionarily conserved. When tested in a vertebrate system, it blocked heterochromatin propagation as effectively as the cHS4 (Lin et al., 2011). Although many insulator/boundary-associated proteins have been characterized in *Drosophila*, including Su(Hw), dCTCF, BEAF-32, GAF, CP190 and Mod(mdg4) (reviewed in (Gurudatta and Corces, 2009)), none of those was found associated with ILB. The presence and prevalence of novel boundary-setting mechanisms were also implicated by epigenomic studies conducted in *Drosophila* and mammalian systems, which revealed that the majority of H3K27me3 boundaries are not associated with characterized insulator proteins.

Although lots of efforts have been directed towards partitioning the genome into large domains based on multiple histone modifications (Hon et al., 2008; Kharchenko et al., 2011) or protein binding profiles (Filion et al., 2010), there is much less focus on understanding how individual repressive histone modification is demarcated by chromatin barrier elements. To gain a comprehensive understanding of boundaries of facultative heterochromatin, we developed a novel bioinformatics approach to identify the chromatin transitional regions (CTRs). We reasoned that if the propagation of heterochromatin formation is stopped by a counter-acting mechanism as revealed by the models proposed by Felsenfeld and colleagues (Gaszner and Felsenfeld, 2006), then the boundary of the facultative heterochromatin should manifest as a rapid transitional region where the level of H3K27me3 shows dramatic changes. Using a two-tiered approach, we demonstrated that it is feasible to identify the CTRs based on H3K27me3 ChIP-Seq data from both *Drosophila* and mammalian cell lines. By locating CTRs to single nucleosome resolution, we found that CTRs are usually 200~600bp away from the binding sites of known insulator/boundary-associated factors. However, the majority of CTRs are not associated with any known insulator proteins. Conversely, only a small portion of insulator protein binding sites are associated with CTRs. Comparing insulator protein bindings associated with CTRs vs. those in H3K27me3-enriched regions revealed interesting distinctions in co-factor binding as well as in DNA sequences flanking the binding sites. Overall, our analysis suggests that diverse mechanisms can be employed to establish the boundaries of facultative heterochromatin (Li and Zhou, 2013).

Materials and Methods

CTRICS (Chromatin Transitional Regions Inference from ChIP-Seq) Algorithm

The program will take H3K27me3 ChIP-Seq data as input (and control dataset measures the input DNA level, if available). The datasets should be in BED (Browser Extensible Data) format. Redundant tags which map to the same genomic region will be kept as a single tag in order to minimize potential PCR bias. CTRICS then divides the whole genome into non-overlapping windows of size *w* (default is 200bp), counts ChIP-Seq tags in each window and generates a bedGRAPH file, which can be viewed with the UCSC Genome Browser (Kent et al., 2002).

We measure the rate of chromatin transition with T-score as,

$$\text{T-score} = \left(\sqrt{L(N)} - \sqrt{R(N)}\right) \Big/ \sqrt{\min(L(N), R(N))} \qquad (2\text{-}1)$$

where *L(N)* and *R(N)* denote the average of ChIP-Seq tags (when there is no input control file), or normalized ChIP-Seq tags in the *N* (default =20) windows upstream and downstream of the given window, respectively. Because H3K27me3 generally forms broad regions covering repressive genes and intergenic regions (Barski et al., 2007; Pauler et al., 2009), the long genomic region used in this initial evaluation should minimize the impact of enrichment level fluctuation observed for H3K27me3 enriched regions. Similar to other studies analyzing the change of chromatin modifications (He et al., 2010; Meyer et al., 2011), we took the square root transformation to minimize the variance introduced by higher counts. We will take the denominator as 1/*N* if the min(L(N), R(N)) was zero. A T-score greater than the threshold (see below), and has one side show significant enrichment of ChIP-Seq tags, will be taken as a candidate region where a transition exist.

In the next step, in seeking to pinpoint the CTR location, we calculate the T'-score for each window around the candidate CTRs, and the region that has the highest T'-score, i.e. the highest enrichment transition rate, will be reported as the predicted CTRs. T'-score is defined as,

$$\text{T'-score} = \text{T-score} \times \left|\sqrt{\text{L(n)}} - \sqrt{\text{R(n)}}\right| \Big/ \sqrt{\min(\text{L(n)},\text{R(n)})} \qquad (2\text{-}2)$$

which is the product of the T-score for the long genome region (*N* windows) and the absolute T-score for a short genome region (*n* windows; default=3).

To assess the statistical significance of each CTR (the probability that the observed T-score is by chance), we need to derive the distribution of T-score in the background model. In this program, we chose not to make any assumption about the background distribution of the ChIP-Seq tags because different datasets have variations and will not always follow a certain assumed distribution. Instead we applied a bootstrapping approach to get the background distribution of T-score. The bootstrapping was conducted by randomly choosing (with replacement) *N* windows from the whole genome as left windows and *N* windows as right windows, then calculating T-score with the randomly chosen windows. The T-score distribution in the random background model is obtained by repeating (with replacement) the above process for a large number of times (10^6). Based on the T-score distribution and the runtime input p-value, we will get the T-score threshold.

In the presence of the input control file, an extra step is needed before the estimation of T-score threshold and prediction of CTRs. We named this step as "background correction", which normalizes the tag number of each window in ChIP file to the tag number of the same window in control file by the following formula,

$$\frac{n_t / N_t}{n_c / N_c}$$

where n_t, n_c are the tag numbers of a given window in ChIP file and control file (again n_c will be set as one if it is zero), N_t, N_c are the total tag counts in ChIP and control files. After the background correction, the program will use the normalized tag counts to estimate T-score cutoff and to predict CTRs. Figure 3-17 shows the workflow of CTRICS.

CTRICS has been implemented in Perl, and it can be downloaded from http://159.178.28.30/CTRICS/home.htm.

Dataset

The datasets used in this study are listed in Table 3-1 and 3-2.

Parameters Used for Predicting CTRs and/or H3K27me3 Domains

CTRs were predicted in *Drosophila* S2 and human HeLa cells by CTRICS with default parameters (expect that the p-value was set to 0.005 for HeLa cells). We ran SICER without control file using default parameters suggested by the authors (window size = 200bp, gap size = 600bp, E-value = 100, p-value = 0.2), and we took the effective *Drosophila* genome size as 71.6%. RSEG was also run with default parameters defined by the program. The two boundaries of an H3K27me3 enrichment domain predicted by these programs were taken as two CTRs, and we discarded the boundaries which are less than 4kb to an unmappable region.

Statistical Analysis

The two-way hierarchical clustering (Ward method) and principal component analysis were carried out using JMP Genomics 5.0 (SAS Institute, Cary, NC). A binding

is considered positively associated with a CTR if the midpoint of the binding is within 1 kb of the CTR.

Wilcoxon rank sum test was performed in R programming environment (R version 2.9.2, R Development Core Team, 2009) to compare the gene expression levels on different sides of CTRs, as well as the binding intensity and width of binding sites of insulator proteins and co-factors.

Motif Discovery

The insulator protein binding motifs were identified using CisFinder (Sharov and Ko, 2009) with default setting. 400bp regions centered on the midpoint of the bindings were used as input. The predicted motifs were depicted as color logos using WebLogo (Crooks et al., 2004). The discriminative motifs were discovered using MEME (web-version 4.8.1)(Bailey et al., 2010).

Calculation of Nucleotides Content

Poly(dA:dT) (AAAA/TTTT) level was calculated using a sliding window approach with window size of 200bp and step of 25bp. The contents were further normalized to the genome average.

Results

Localize the Chromatin Transitional Regions (CTRs) Based on H3K27me3 ChIP-Seq Data

At the time of our study, several methodologies, such as SICER (Zang et al., 2009) and RSEG [(Song and Smith, 2011), have been developed to analyze genomic profiles of H3K27me3, the signature marker of facultative heterochromatin. Most of these methodologies focus on identifying broad domains enriched for a particular histone modification. Although these methodologies are very useful for identifying

H3K27me3-enriched regions, they were not designed for the purpose of specifying the boundary of facultative heterochromatin. The fact that there is a lack of experimentally verified data set of H3K27me3 boundaries also prevented objective comparison of these methodologies.

Drosophila melanogaster provides the best system for studying the boundaries of facultative heterochromatin. Several insulator proteins, such as Su(Hw) (Harrison et al., 1993), BEAF-32 (Gilbert et al., 2006), and dCTCF (Mohan et al., 2007), have been very well characterized in *Drosophila*. The genome-wide binding profiles for these proteins, as well as many other genomic and epigenomic information, are available for the *Drosophila* Schneider 2 (S2) cells due to the efforts of the modENCODE project (Roy et al., 2010) and many other individual labs. Taking the advantage of these data, we generated an empirical set of chromatin transitional regions for H3K27me3 (Figure 3-10). In essence, we selected regions where clear changes in H3K27me3 enrichment, as revealed by ChIP-Seq, were accompanied by experimentally verified binding of the insulator proteins and their respective co-factors (such as CP190).

Testing of the two popular H3K27me3 enrichment-calling algorithms with this empirical H3K27me3 boundary data set revealed inconsistency in precisely defining the transition region. We noticed that the enrichment-calling algorithms such as SICER is sensitive to the fluctuation of H3K27me3 enrichment levels in continuous facultative heterochromatin regions, and consequently predicts many “extra” boundaries in areas where the enrichment level of H3K27me3 fluctuated (Figure 3-1, Figure 3-11). On the other hand, methodology such as RSEG, which based on the two-state hidden Markov model and provided specific boundary calling function, seems to miss some putative

boundaries in our empirical data set (Figure 3-11). It is worth noting that RSEG also failed to predict a boundary at the ILB locus, which has been experimentally verified to function as chromatin barrier against Polycomb group (PcG)-mediated spreading of H3K27me3 (Lin et al., 2011).

To pinpoint the location of CTR, we developed a two-tiered analysis methodology called CTRICS (Chromatin Transitional Regions Inference from ChIP-Seq) (see Methods for detail). First, the existence of a transition was detected by comparing the enrichment of H3K27me3 in relatively large genomic intervals (4kb). The relatively large interval helps to minimize the false positives due to the fluctuation of H3K27me3 enrichment levels in facultative heterochromatin regions. After a transitional event has been identified, a secondary analysis is performed with short intervals to identify a 200bp region where the enrichment of H3K27me3 displays the most significant change. The number of CTRs identified by CTRICS is comparable to the boundaries identified by RSEG, and both are much less when compared with the boundaries predicted by SICER (Figure 3-11A). The majority of CTRs we identified overlap (i.e. within 2 kb) with the boundaries predicted by RSEG (Figure 3-11A). However, unlike RSEG, our method was able to identify more putative boundaries in our empirical data set (Figure 3-11B) as well as the ILB. Visual inspection indicated that some of the CTRs identified by CTRICS, but missed by RSEG can be corroborated with other evidences such as RNA-Seq or H3K4me3 data (Figure 3-11C). Thus we resorted to use CTRICS for genome-wide analysis of CTRs in S2 cells.

Genome-Wide Identification of CTRs in S2 Cells

Applying CTRICS to the H3K27me3 ChIP-Seq dataset derived from the *Drosophila* S2 cell line (Gan et al., 2010) identified a total of 2082 CTRs. From

sequencing depth analysis, we noticed that the H3K27me3 ChIP-Seq dataset, with a total of ~2.8 million uniquely mapped reads, had already reached saturation plateau for CTR detection (Figure 3-12).

Since CTRs define the boundaries between repressive facultative heterochromatin and accessible euchromatin, the active and repressive histone marks should have contrasting patterns around CTRs. Indeed, active histone marks, such as H3K4me1, H3K4me2, H3K4me3, H3K9ac, and H3K27ac, are enriched on the euchromatic side of CTRs, while depleted on the heterochromatic side (Figure 3-1A). We noticed that the enrichment levels of H3K9me3, which mostly associate with constitutive heterochromatin in centromeric and telomeric regions (Schones and Zhao, 2008), do not change significantly around the identified CTRs. This indicated that the CTRs we identified are specific to facultative heterochromatin. Although the two repressive histone marks overlap at some loci (Bilodeau et al., 2009; Hon et al., 2009; Lin et al., 2011), their global localizations are largely independent of each other. Our analysis also suggested that in most loci, the change of H3K27me3 level at the boundary was not associated with significant changes in H3K9me3.

In addition, we reasoned that genes locate on the heterochromatic sides of CTRs should in general be repressed compared to those on the euchromatic side. When the expression profile was evaluated using a companying RNA-Seq dataset from S2 cells (Gan et al., 2010), the difference was indeed obvious for genes on different sides of the CTRs. Compared with the global average, genes whose entire transcribed regions locate within the 4kb regions on the euchromatic side of CTRs had significantly higher level of expression, whereas genes on the heterochromatic sides were significantly

repressed (Figure 3-1B). Corresponding with the difference in gene expression levels, the binding of Pol II as well as active histone modification H3K4me3 show specific enrichment on the euchromatic side of CTRs (Figure 3-1C). These evidences all support that the CTRs identified by our method indeed are sharp boundaries interface H3K27me3-enriched and depleted regions.

The Spatial Relationships between CTRs and Known Boundary-Setting Proteins

The global binding profiles of the major insulator proteins Su(Hw), BEAF-32, dCTCF, GAF, and their important co-factors (such as CP190 and Mod(mdg4)) are available for the S2 cells. Comparison of H3K27me3 CTRs and the binding profiles indicated that less than 15% of the insulator proteins binding sites are within 1 kb of the identified CTRs (Figure 3-2A). The majority of the binding sites of these known insulator proteins are not in close association with CTRs. For instance, many (~49%) Su(Hw) binding sites are found in continuous H3K27me3 domains (Figure 3-2B).

Conversely, less than half (~42%) of the H3K27me3 CTRs are associated with any of the four DNA-binding insulator proteins, i.e. located within 1 kb (Figure 3-2C). However, for those that do associate with a binding site for the insulator proteins, the binding site is always preferentially located at the euchromatic side of the CTR (Figure 3-2D,E), which agrees very well with a recent genome-wide study of chromatin boundary elements conducted in human $CD4^+$ cells (Wang et al., 2012). When the intensity of these proteins were plotted, the peaks of insulator binding is located at about 200~600bp away from the CTR (Figure 3-2D). Very similar spatial relationship between CTRs and insulator protein binding was observed for BEAF-32, Su(Hw), and dCTCF. Compared with these insulator proteins, the spatial relationship between GAF binding and the correlated CTRs was somewhat different, with the enrichment region

more spread out and the peak of binding intensity about 1 more nucleosome space away from the CTR (Figure 3-2D). Figure 3-2E illustrates a CTR as an example, it is associated with BEAF-32 and CP190, and the peaks of both protein binding sites are located on the euchromatic side of this CTR, with about 400bp between the peaks of binding and the CTR.

The Diversity of Facultative Heterochromatin Boundaries

As mentioned above, the spatial relationship between CTR and the binding of known insulator proteins suggests that the CTRs observed for S2 cells are due to the barrier activity of insulator proteins. However, more than half of the CTRs are not co-localized with any of the known insulator proteins (Figure 3-2C). To gain a comprehensive understanding of the H3K27me3 CTRs identified in S2 cells, we expanded our analysis to include the binding profiles for other chromatin-associated proteins, all of which were generated by the modENCODE project (Roy et al., 2010) with the S2 cells. In this analysis we excluded proteins which have been shown to be directly involved in the establishment or maintenance of the facultative heterochromatin, such as the polycomb group proteins, the trithorax group proteins, and heterochromatin binding proteins. A total of 15 binding profiles were selected (Figure 3-3, Table 3-1), and the binding call was processed as described (Kharchenko et al., 2011). Similar to previous association studies (Cuddapah et al., 2009), we considered a binding within 1kb of a CTR as a positive association.

We then conducted unsupervised hierarchical clustering to classify CTRs based on the association with these chromatin-associated proteins. From the clustering analysis, the predicted CTRs can be clearly divided into eight groups (Figure 3-3A). We also performed principal component analysis on the 15 proteins and the first three

principal components turned out to account for 25.1%, 12.2% and 10.5% of the total variance respectively (Figure 3-13). After projecting the predicted CTRs on the first three principal components, the eight distinct CTR groups were also clearly separated (Figure 3-13), demonstrating that the grouping of CTRs was robust to different classification methods.

The protein occupancy in distinct CTR groups clearly suggested that the majority of CTRS in groups A, B and C are associated with the insulator protein CP190, whereas the other five groups are CP190 independent (Figure 3-3B). For the 3 CP190-associated groups, about 30% of CTRs in Group A are also associated with the insulator protein dCTCF, which requires CP190 as a co-factor (Mohan et al., 2007). The majority of CTRs in group B are also bound by insulator proteins Su(Hw), Mod(mdg4), which are the required trans factors for the *gypsy* insulator (Ghosh et al., 2001; Harrison et al., 1993; Pai et al., 2004). CTRs in Group C are enriched for insulator protein BEAF-32. Interestingly, CTRs in this group are also associated with the chromatin remodeling protein NURF, which has been shown to be required for establishing the chromatin barrier activity of cHS4 at the chicken β-globin locus (Li et al., 2011). Taken together, our unsupervised hierarchical clustering agrees very well with the model put forward based on genetic analysis of three insulator proteins, i.e. while the three insulator proteins Su(Hw), BEAF-32, and dCTCF barely overlap with each other, they all co-localize with CP190 (Bushey et al., 2009).

Interestingly, the majority of CTRs in group E are associated with JIL1, which can maintain euchromatic state by terminating the constitutive heterochromatin spreading (Bao et al., 2007; Zhang et al., 2006). The colocalization suggests that JIL1 may also

antagonize facultative heterochromatin through a mechanism which is different from the other CTR groups. The separation of group D from group E is due to the presence of RNA polymerase II. However, CTRs in group D were not associated with annotated TSSs (Transcription Start Sites), while the majority of CTRs in groups A, C, and G are located close to TSS (Figure 3-14). In depth analysis indicated that the Pol II binding "peaks" associated with CTRs in group D are much smaller than those associated with bona fide TSS and they do not correlate with H3K4me2/3 enrichment. Close inspection suggested that the association of Pol II binding with this group was questionable and could be due to artifact of peak calling. Since all of the peak calling were generated by modENCODE with unifying standard (Kharchenko et al., 2011), we refrained from changing the calling specifically for the Pol II data. Group F is associated with the insulator protein GAF (Schweinsberg et al., 2004).

In groups G and H, most of the CTRs have no clear association with any of the investigated proteins, which suggests the existence of other proteins functioning at these CTRs. Interestingly, the novel chromatin barrier ILB we have recently identified (Lin et al., 2011) does not co-localize with any of the 15 proteins, and belongs to group H.

Strong Co-Factor Binding Distinguishes dCTCF and Su(Hw) Binding Associated with CTR vs. Those in H3K27me3-Enriched Regions

The strict spatial relationship between the binding of insulator proteins and the identified CTRs strongly suggests a cause-effect relationship between insulator protein binding and the formation of the boundary for the H3K27me3 modification. However, analysis of the global profiles indicated that only a small portion of binding sites for dCTCF and Su(Hw) are associated with CTRs. To reconcile the two seemingly

conflicting observations, we first asked whether there is any difference in terms of binding intensity by the respective insulator proteins. To address this question, we compared the binding profiles at sites associated with CTRs against those in regions enriched for H3K27me3, which are clearly not associated with any chromatin barrier activity.

We found that for dCTCF, Su(Hw), and GAF, there was no significant difference in terms of enrichment levels at the peaks of the binding (Figure 3-4A, 2-2B). There was only marginal difference for BEAF-32, where the peak intensity was about 50% higher in sites associated with CTR (Figure 3-4A). These findings suggested that insulator proteins such as dCTCF and Su(Hw) can bind with similar affinity to euchromatic regions associated with CTRs and facultative heterochromatic regions enriched for H3K27me3. Although the intensities at the peak were similar for both dCTCF and Su(Hw), we did notice that the binding for these two insulator proteins was more spread in heterochromatic regions and more constrained in binding sites associated with CTRs (Figure 3-4A,C, 2-2B). The functional significance of this difference is unclear.

It has been well documented that the binding of co-factors such as CP190 is required for the enhancer blocking function of Su(Hw) and dCTCF (Mohan et al., 2007; Pai et al., 2004). We found that the intensity of CP190 binding was much higher at sites associated with CTRs. This is true for both dCTCF and Su(Hw), where the CP190 binding intensities at sites associated with CTRs were significantly higher than those that are in H3K27me3-enriched regions (Figure 3-4B, 2-2B). Significant difference in binding intensity was also observed for another co-factor of Su(Hw), i.e. Mod(mdg4) (Ghosh et al., 2001), for which the binding intensity for sites associated with CTRs was

2.98 fold of those in heterochromatic regions (Figure 3-4B). These observations strongly suggested that co-factors such as CP190 and Mod(mdg4) are involved in establishing the chromatin barrier activity of dCTCF and Su(Hw).

Poly(dA:dT) Tracts and Decreased Nucleosome Density around the Insulator Binding Sites associated with CTR

We next asked whether there is any difference between the DNA sequences underlying the insulator protein binding sites associate with CTRs and those in H3K27me3-enriched regions. We inputted the 400bp regions around the CTR-associated binding sites to CisFinder (Sharov and Ko, 2009) to identify statistically overrepresented DNA motifs, which were then compared with the motifs obtained with the 400bp sequences surrounding the heterochromatic (H3K27me3-enriched) binding sites. There was no significant difference between the motifs identified from CTR-associated binding sites vs. those identified from the binding sites in heterochromatic region (Figure 3-5A). In fact, motifs identified from the aforementioned two subsets resembled the motifs identified using all binding sites identified in the S2 cells. This suggested that the DNA sequences interacting with the insulator proteins do not distinguish whether the association of the respective insulator protein can function as chromatin barrier or not.

We then asked whether sequences surrounding the CTR-associated insulator protein binding sites have discriminative patterns comparing to those surrounding the heterochromatic binding sites. Interestingly, when we supplied MEME (Bailey et al., 2010) with the CTR-associated binding sites as positive regions and the heterochromatic binding sites as negative regions to identify discriminative motifs, a motif with continues deoxyadenosine (multi-A) showed up for all of the four insulator

proteins (Figure 3-5B). Similar results were obtained when using the CisFinder program (Sharov and Ko, 2009). This indicated that a key distinction of insulator protein binding sites associated with CTRs was that they tend to be in close proximity to sequences with long stretch of dA/dTs (poly(dA:dT) tracts).

DNA sequences with poly(dA:dT) tracts where n(A/T)≥4 has been found to be rigid and discourage nucleosome binding (Mavrich et al., 2008a; Suter et al., 2000). When we compiled the levels of poly(dA:dT) (frequency of AAAA/TTTT) and the nucleosome density around the binding sites of the four known insulator proteins (Su(Hw), BEAF-32, GAF, dCTCF), we found that the binding sites associated with CTRs were strongly associated with increased poly(dA:dT) levels as well as dramatically decreased nucleosome occupation (increased sensitivity to MNase) (Figure 3-5C). In contrast, such an association was not observed for binding sites in H3K27me3-enriched regions. We concluded that the CTR-associated insulator protein binding sites tend to be surrounded by DNA sequences characterized with nucleosome-destabilizing poly(dA:dT) tracts and manifest as hypersensitive to MNase.

Poly(dA:dT) Tracts and Increased Sensitivity to MNase are Associated with CTRs that do not Bind with Known Insulator Proteins

To see how general are poly(dA:dT) tracts and increased MNase-sensitivity associated with CTRs of H3K27me3, we plotted the distribution for each of the groups identified by the hierarchical clustering (Figure 3-3A). We found that for CTRs in groups A, B, and C, which are all enriched for the binding of CP190, there is a clear trend of increased level of poly(dA:dT) (n≥4) and decreased nucleosome occupancy. The region of increased poly(dA:dT) levels roughly correlates with that of the increased sensitivity to MNase, and both peak at the euchromatic side of CTRs (Figure 3-6). CTRs in group

A, B, and C are enriched for the presence of binding of dCTCF, Su(Hw), and BEAF-32, respectively. However, not all of the CTRs in each group are associated with the corresponding insulator protein (Figure 3-3B).

We could not observe clear increase of poly(dA:dT) level, nor decreased nucleosome density, associated with the CTRs in groups D and E, which are associated with PolII and JIL1, respectively. While there is a slightly increased level of poly(dA:dT) and a decreased level of nucleosome density for CTRs in group F. However, the distribution is somewhat different from those observed for groups A, B, and C, in that there is no clear peak.

Interestingly, for CTRs in groups G and H, which are not enriched for the binding of any known insulator proteins or other chromatin-associated proteins investigated here, there is a clear trend of increased poly(dA:dT) level on the euchromatic side of CTRs. For group G, there is also a significant decrease of nucleosome density correlates with the increased level of poly(dA:dT). It is well known that nucleosome positioning sequences, including poly(dA:dT), are associated with promoters (Mavrich et al., 2008b). However, the majority of CTRs in group H are not close to TSS (Figure 3-14) or associated with Pol II binding, the increased multi-A level in the two groups is unlikely due to the nucleosome positioning sequences associated with promoters. This indicated that the presence of poly(dA:dT) tract and decreased nucleosome density is a general feature of CTRs beyond those that associate with the characterized insulator proteins.

Enrichment of H3.3 but Decreased Nucleosome Turnover at CTR-Associated dCTCF Binding Sites

It has been shown in mammalian systems that the binding of insulator proteins such as CTCF results in dynamic (unstable) nucleosomes and manifest as sites with increased enrichment of histone variants such as H3.3/H2A.Z at low salt isolation condition (Jin et al., 2009). The dynamics of nucleosomes in S2 cells has also been assayed with the rate of histone variant H3.3 replacement (Mito et al., 2005), and more recently, with the CATCH-IT technology (Deal et al., 2010). The latter is based on metabolic labeling of histones and is thus a direct measurement of nucleosome turnover rate independent of the composition of nucleosome. It has been shown that in general the profiles obtained with CATCH-IT correlate very well with the one based on H3.3 incorporation (Deal et al., 2010). In addition to the CATCH-IT profile, datasets for H3.3 enrichment at low vs. high salt isolation conditions (Henikoff et al., 2009), nucleosome density (ratio of nucleosomal/genomic) (Henikoff et al., 2009), and DNA accessibility evaluated with methylation footprinting (Bell et al., 2010) were also available for the same cell line.

When these profiles were evaluated around all of the H3K27me3 CTRs identified for S2 cells, we found that there was a conspicuous decrease of nucleosome density at the euchromatic side of CTRs (Figure 3-7A). The lowest point of nucleosome density is about 200~600bp away from the CTR, which corresponds well with the peak of the binding sites for known insulator proteins (Figure 3-2D), as well as the region enriched for poly(dA:dT) tracts (Figure 3-6). Correspondingly, consistent increase of DNA accessibility was also observed at the same relative position.

However, when the nucleosome dynamics data was evaluated, we noticed an apparent discrepancy between the H3.3 incorporation measurements and the CATCH-IT profiles. At the same relative location to CTRs, there is a significant increase of H3.3 incorporation (at low salt condition), which would have indicated an increased dynamics (turnover rate) at the insulator protein binding sites. However, this was contradicted by the CATCH-IT profile at these sites, which showed a sharp drop at the same relative position (Figure 3-7A).

To understand the cause of this discrepancy between the H3.3 incorporation and the turnover rate measured with CATCH-IT, we looked at these profiles associated with each individual insulator proteins (Figure 3-7B). We found that for the GAF binding sites, whether associated with CTRs or not, there is a consistent increase of both H3.3 and CATCH-IT. This agrees well with previous findings that GAF binding sites are marked by increased nucleosome dynamics (Deal et al., 2010). This also indicates that in terms of nucleosome dynamics, there is no difference between GAF binding sites associated with CTRs vs. those that are not associated.

However, a contrasting pattern was specifically observed between the H3.3 and CATCH-IT profiles around CTR-associated dCTCF binding sites. While there is a significant increase of H3.3 in these sites, the CATCH-IT data indicated that the turnover rate at these sites is not higher, but rather lower than the neighboring region (Figure 3-7B). This contrasting trend of H3.3 incorporation and nucleosome turnover rate suggested that, unlike GAF binding sites, the increased level of H3.3 incorporation is accompanied by decreased level of nucleosome turnover at the dCTCF binding sites close to CTRs. Interestingly, this contrasting trend was only obvious with dCTCF

binding sites associated with CTRs, but was not observed around dCTCF binding sites not associated with a CTR (more than 1kb away from the closest CTR) (Figure 3-7B).

As aforementioned, the contrasting pattern between the enrichment of H3.3 and decreased nucleosome turnover rate was obvious when the two profiles were evaluated for all H3K27me3 CTRs identified in S2 cells. For the groups of CTRs associated with known insulator factors, we found that this contrasting pattern is most prominent for group A (Figure 3-7C). About 30% of CTRs in this group has verified binding of dCTCF (Figure 3-3B). In addition, the enrichment of H3.3 was also prominent for CTRs in group G, which has no clear association with any of the known insulator proteins.

Chromatin Transitional Regions in the HeLa Cell Line

Applying the CTRICS program to H3K27me3 ChIP-Seq dataset derived from human HeLa cells (Cuddapah et al., 2009) identified a total of 10710 CTRs. The majority (8047) of which overlaps with the boundaries of H3K27me3 domains identified by Cuddapah *et al.* (Cuddapah et al., 2009), which identified a total of 32,704 H3k27me3 domains in HeLa cells (Figure 3-8A). The difference in the number of H3k27me3 boundaries identified by CTRICS and the H3K27me3 domain approach is likely due to the combined effect of 1.) the CTRICS methodology is less sensitive to fluctuation of H3K27me3 enrichment levels within H3K27me3-enriched domains (Figure 3-1C); and 2.) CTRICS is more stringent in that it will only identify boundaries with a significant drop of H3K27me3 level (Figure 3-15).

With this stringent set of CTRs in HeLa cells, there is a significant increase of DNA accessibility (DNase-Seq data set from (Thurman et al., 2012)) in at the immediate euchromatic side of the CTRs (Figure 3-8B). This is very similar to what we observed in the S2 cells. There is also a significant change of nucleosome density (MNase-Seq data

from (Tolstorukov et al., 2012)) around the predicted CTRs, which confirms that our method is identifying well defined facultative heterochromatin boundaries.

Similar to what was observed for S2 cells, CTCF was also enriched on the euchromatic side of CTRs with about 2-nucleosome space in between (Figure 3-8C). However, unlike dCTCF, there was a minor peak of the polled CTCF binding signal on the heterochromatic side of CTRs (Figure 3-8C). Interestingly, a similar major peak and minor peak pattern of CTCF binding was also observed independently for facultative heterochromatin boundaries in human $CD4^+$ cells identified with a consortium of histone modification profiles and a maximal segment algorithm (Wang et al., 2012). Overall the binding intensity for CTCF was moderately, but significantly, higher for binding sties associated with CTR than those in heterochromatic regions (Figure 3-8D). Since no co-factor such as CP190 was identified in mammalian systems, which prevented us to test whether similar distinction of co-factor binding also applies to human CTRs.

Discussions

In this work, we showed that it was possible to identify the boundaries of facultative heterochromatin based on H3K27me3 ChIP-Seq data. Our two-tiered method first identifies a heterochromatin to euchromatin transition event by considering the enrichment value for a relatively large region. Following that, the 200bp region that shows the greatest transition rate of enrichment values is designated as the CTR. The validity of this simple strategy was firstly verified by the dramatic difference in active/repressive histone modifications and gene expression levels on the heterochromatic vs. euchromatic side of the predicted CTRs (Figure 2-1). More importantly, the validity of this strategy was vindicated by the fact that, for CTRs

associated with the binding of known insulator proteins, there is a strict spatial relationship between the CTRs and the insulator protein binding sites.

Fixed vs. Variable Boundary for Facultative Heterochromatin

The method developed in this study is specifically suitable for the identification of fixed boundaries for facultative heterochromatin. Visual inspection of H3K27me3 profile has suggested that certain H3K27me3 domains do not have a fixed boundary (Schwartz et al., 2012). It is clear that for constitutive heterochromatin close to centromere, the boundary of heterochromatin marked by H3K9me2/3 can vary in different cells of the same tissue. This phenomenon was reflected as “variegated” expression of reporter/marker genes located close to centromere, i.e. position-effect variegation (PEV) (reviewed in (Girton and Johansen, 2008; Karpen, 1994)). It is possible that our method won’t be sufficient to identify boundaries that show variable locations in individual cells, for which the pooled ChIP-Seq data will lack a sharp transition region.

It is conceivable that due to its close association with euchromatic region, the boundaries of facultative heterochromatin need to be precisely defined to avoid the disruption of the transcriptional regulation of adjacent genes. In the case of the cHS4 chromatin barrier in the chicken β-globin locus, the binding of USF1 was responsible for recruiting histone modifying enzymes which in turn catalyze euchromatic histone modifications on adjacent nucleosomes (Huang et al., 2007; West et al., 2004). The USF1-directed euchromatic histone modifications effectively block the propagation of heterochromatic marks and results in a sharp transition of histone marks. Interestingly, a recent study revealed that NURF is recruited by USF1 to cHS4 and is required for establishing the chromatin barrier (Li et al., 2011). Our analysis indicated that the binding of NURF (NURF301, Figure 2-3) is associated with CTRs in groups A, B, C, and

G. It has been shown that *Drosophila* NURF is required for the enhancer-blocking activity of several insulators (Li et al., 2010). Our results suggest that its role in establishing chromatin barrier is also likely conserved over long evolutionary distance.

Our analysis of genome-wide H3K27me3 CTRs in S2 cells indicated that at least in this cell line, many boundaries of facultative heterochromatin, marked by the transition of H3K27me3 enrichment level, can be clearly identified. However, formation of facultative heterochromatin is, by definition, cell type specific. We found that clear boundaries cannot be reliably identified from H3K27me3 data obtained from homogenized animals (embryos or larvae). Since the binding profile of many insulator proteins as well as other epigenomic profile has been well studied in the S2 cells, the genome-wide identification of CTRs in this cell line allowed us to address several interesting questions in regards to chromatin barriers.

Binding of Insulator Protein Alone is not Sufficient for Establishing the H3K27me3 Boundary

Our analysis indicated that only a small portion of genome-wide binding sites for insulator proteins such as dCTCF and Su(Hw) are associated with the CTRs. This was not surprising, given that a genomic study conducted in mammalian cells also revealed that for CTCF binding sites observed for CD4+ T cells and HeLa cells, only a small percent (about 5.6% and 4.1%, respectively) are associated with the boundaries of H3K27me3-enriched domains (Cuddapah et al., 2009). Our results indicated that similar to what was observed for CTCF in mammalian cells, the majority binding sites for insulator proteins such as dCTCF and Su(Hw) do not co-reside with the boundaries of facultative heterochromatin. The same mammalian study also revealed that only a very small portion (less than 5%) of the H3K27me3 boundaries in those cells have a CTCF

binding site within 1 kb of distance. Although many more insulator proteins have been characterized in *Drosophila*, less than half of all H3K27me3 CTRs identified in S2 cells are associated with any of the known insulator proteins. This indicates that uncharacterized mechanisms, which do not involve any of those proteins known to play a role in this process, is responsible for establishing more than half of the facultative heterochromatin boundaries in S2 cells.

In this study, by narrowing down the transitional region to 200bp, we were able to reveal some very interesting relationships between the binding of insulator proteins and the CTRs. Central to these findings are the observation that there is a clear spatial relationship between the binding sites of insulator proteins and the CTRs. The binding of insulator proteins is at the euchromatic side of CTRs and the peak of binding is about 200~600bp away from the CTRs. This strict spatial relationship suggests that there is a functional relationship between the binding of these insulator proteins and the establishment of the sharp transition at the CTRs.

A prominent question in regards to insulator proteins binding and the formation of chromatin boundary is what distinguishes those sites associated with a chromatin boundary versus those do not. We found that compared with dCTCF and Su(Hw) binding sites in heterochromatic regions, the binding sites associated with CTRs were bound by higher levels of co-factors such as CP190 or/and Mod(mdg4). In contrast, such a distinction was not observed for CTRs associated with BEAF-32. A recent work revealed that, unlike dCTCF and Su(Hw), binding of CP190 at BEAF-32 binding sites was not affected when the insulator protein was knocked down (Schwartz et al., 2012). The same work also suggested that the inherited binding preferences of, not the

interaction between, the two proteins could be responsible for the observed colocalization of BEAF-32 and CP190. Our observations further support their argument. The increased binding intensities of co-factors at dCTCF and Su(Hw) sites associated with CTRs were not simply because those binding sites are located on the euchromatic side of the CTRs, since the intensity at CTR-associated sites was significantly higher when compared with that in euchromatic regions (Figure 2-16). These observations strongly suggested that there is a significant difference in co-factor binding between CTR-associated binding of dCTCF and Su(Hw) vs. those that are in heterochromatic regions.

Besides the difference in co-factor binding, the underlying DNA sequences surrounding CTR-associated binding sites are enriched for poly(dA:dT) tracts. Poly(dA:dT) tracts have been found to form rigid structures and discourage nucleosome formation (Mavrich et al., 2008b; Suter et al., 2000). The fact that poly(dA:dT) tracts distinguish CTR-associated insulator protein binding sites from those in heterochromatic region suggested that it plays a role in establishing/encouraging the barrier function of dCTCF and Su(Hw). One hypothetic model come out of our analysis is that the presence of nucleosome-destabilizing sequences flanking the insulator protein binding site associated with CTRs could change the dynamics of nucleosome formation as well as facilitate increased binding of co-factors. However, the enrichment of poly(dA:dT) tracts surrounding CTR-associated binding sites could simply be an indicator of nucleosome depletion, instead of playing a role in the formation of nucleosome depletion regions, as suggested by a recent study that these regions favor G/C to A/T mutations (Chen et al., 2012).

Nucleosome Dynamics, Histone Variants, and H3K27me3 Boundary

Increased nucleosome dynamics, often manifested as increased enrichment level of histone variants such as H3.3, has been linked with transcriptionally active genes in both *Drosophila* and mammalian systems. Our analyses indicated that distinctive patterns of nucleosome dynamics and histone variants incorporation are associated with different subgroups of CTRs.

For CTRs associated with GAF, there is an increased nucleosome turnover rate (measured by CATCH-IT) as well as an enrichment of H3.3 incorporation (Figure 2-7B). This agreement between turnover rate measured by CATCH-IT and H3.3 incorporation has been observed globally for TSSs (transcription start sites) and several important chromatin landmarks such as binding sites for ploycomb group proteins (Deal et al., 2010). It is conceivable that the dynamic nucleosome located at the binding site of GAF could serve to discourage the propagation of repressive histone modifications (Figure 2-9A), which is in consistent with the model proposed in yeast (Dion et al., 2007).

However, a surprising phenomenon was identified for those CTRs that are associated with dCTCF. Instead of increased turnover rate, the nucleosomes close to the binding sites actually showed decreased level of turnover as measured by CATCH-IT (Figure 2-7B). This reduced turnover rate at H3K27me3 CTRs was not limited to those that have binding of dCTCF. It was also prominent for CTRs in group G (Figure 2-7C). Intriguingly, the decreased level of turnover is accompanied by increased incorporation of H3.3 in those CTRs. This suggested that for certain subgroups of CTRs, the nucleosome at the boundary has reduced turnover rate but nonetheless has strong preference for the histone variant H3.3 (Figure 2-9B). The preference of H3.3 could potential serve as a deterrent for the spreading of H3K27me3. However, this

mechanism, if indeed contributes to the formation of H3K27me3 boundary, is likely redundant and dispensable. Since the deletion of H3.3 did not have significant impact on facultative heterochromatin formation and can be compensated by overexpression of H3 (Sakai et al., 2009).

Table 3-1. The list of ChIP-Chip profiles used in the clustering analysis

Protein	modENCODE Title	DCCid	Public Release Date	Antibody	Platform
Su(Hw)	Su(Hw)-VC.S2	modENCODE_331	10/11/2009	Su(Hw)-VC	Affymetrix Drosophila Tiling Arrays v2.0R
Mod(mdg4)	mod2.2-VC.S2	modENCODE_2674	02/15/2010	mod2.2-VC	Affymetrix Drosophila Tiling Arrays v2.0R
dCTCF	CTCF-VC.S2	modENCODE_283	10/11/2009	CTCF-VC	Affymetrix Drosophila Tiling Arrays v2.0R
GAF	GAF.S2	modENCODE_285	10/11/2009	GAF	Affymetrix Drosophila Tiling Arrays v2.0R
dMi-2	dMi-2_Q2626.S2	modENCODE_926	10/11/2009	dMi-2_Q2626	Affymetrix Drosophila Tiling Arrays v2.0R
SPT16	SPT16_Q2583.S2	modENCODE_3058	09/27/2010	SPT16_Q2583	Affymetrix Drosophila Tiling Arrays v2.0R
MRG15	MRG15_Q2481.S2	modENCODE_3047	09/27/2010	MRG15_Q2481	Affymetrix Drosophila Tiling Arrays v2.0R
JIL1	JIL1_Q3433.S2	modENCODE_945	10/12/2009	JIL1_Q3433	Affymetrix Drosophila Tiling Arrays v2.0R
RNAPolII	RNA pol II (ALG).S2	modENCODE_329	10/11/2009	RNA pol II (ALG)	Affymetrix Drosophila Tiling Arrays v2.0R
BRE1	BRE1_Q2539.S2	modENCODE_923	10/11/2009	BRE1_Q2539	Affymetrix Drosophila Tiling Arrays v2.0R
CP190	CP190-VC.S2	modENCODE_280	10/11/2009	CP190-VC	Affymetrix Drosophila Tiling Arrays v2.0R
NURF301	NURF301_Q2602.S2	modENCODE_947	10/12/2009	NURF301_Q2602	Affymetrix Drosophila Tiling Arrays v2.0R
Chriz	Chro(Chriz)BR.S2	modENCODE_278	10/11/2009	Chro(Chriz)BR	Affymetrix Drosophila Tiling Arrays v2.0R
BEAF	BEAF-HB.S2	modENCODE_274	10/11/2009	BEAF-HB	Affymetrix Drosophila Tiling Arrays v2.0R
WDS	WDS_Q2691.S2	modENCODE_953	10/12/2009	WDS_Q2691	Affymetrix Drosophila Tiling Arrays v2.0R

Table 3-2. List of datasets used in this study

Dataset	Cell line	Platform	GEO #
H3K27me3	S2	ChIP-Seq	GSM480157
Gene expression	S2	RNA-Seq	GSM480160
H3K4me1	S2	ChIP-Chip	GSE20786
H3K4me2	S2	ChIP-Chip	GSE20838
H3K4me3	S2	ChIP-Chip	GSE20787
H3K9ac	S2	ChIP-Chip	GSE20790
H3K9me3	S2	ChIP-Chip	GSE20794
H3K27ac	S2	ChIP-Chip	GSE20779
H3K27me3	S2	ChIP-Chip	GSE20781
RNA Polymerase II	S2	ChIP-Seq	GSM480159
H3.3 (low salt)	S2	ChIP-Chip	GSM333869
H3.3 (high salt)	S2	ChIP-Chip	GSM333871
Nucleosome density	S2	MNase-Chip	GSM333835
			GSM333840
			GSM333844
DNA accessibility	S2	Methylation footprinting	GSM441282
Nucleosome turnover	S2	CATCH-IT	GSM494308
H3K27me3	HeLa	ChIP-Seq	GSM325898
CTCF	HeLa	ChIP-Seq	GSM325897
DNA accessibility	HeLa	DNase-Seq	GSM816643
Nucleosome density	HeLa	MNase-Seq	GSM937970

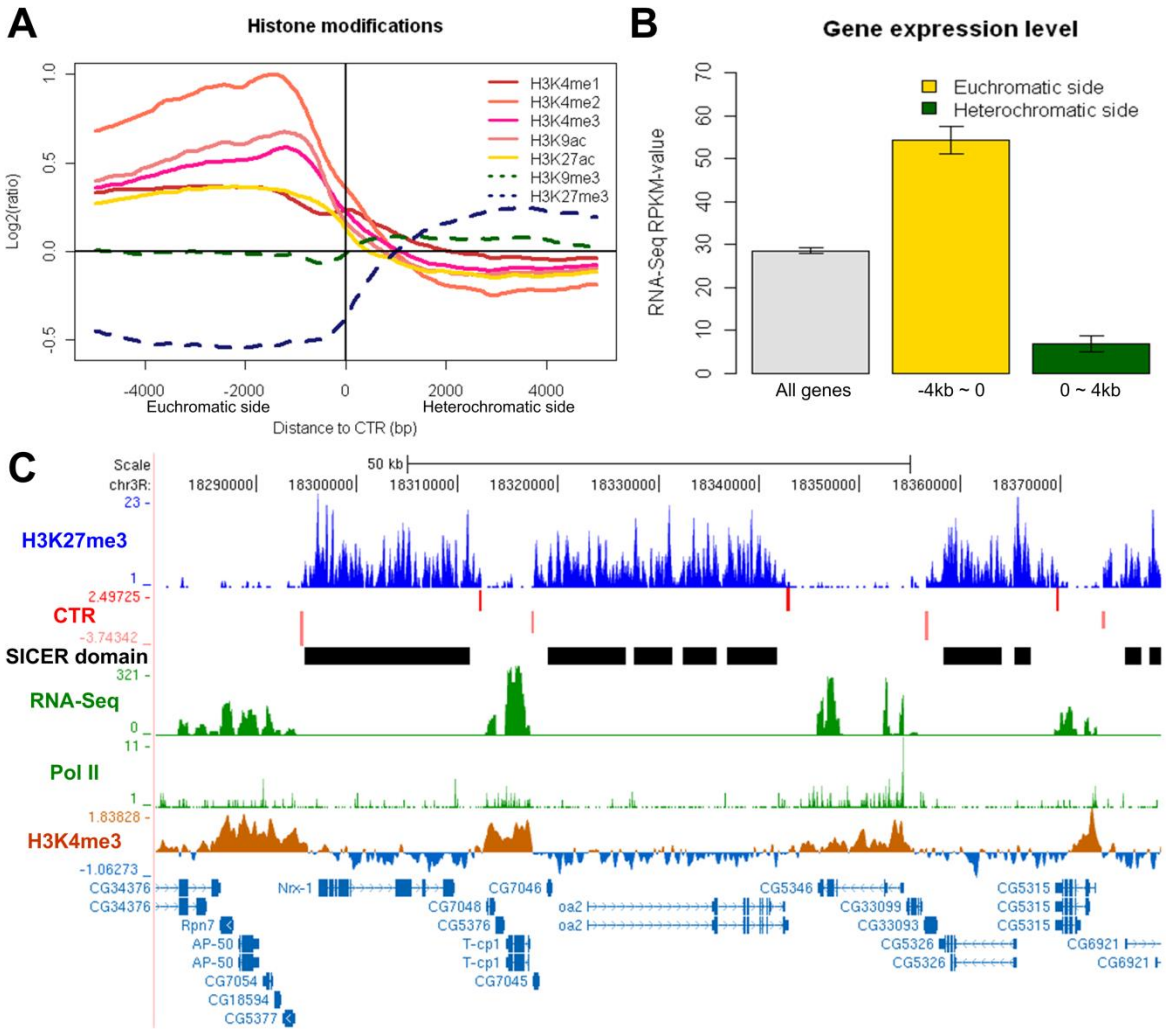


Figure 3-1. Histone modifications and gene expression levels on the euchromatic vs. heterochromatic side of the CTRs in *Drosophila* S2 cell line. (A) Enrichment levels of active (solid lines) and repressive (dashed lines) histone modifications around the H3K27me3 CTRs identified in S2 cells. Negative and positive distances indicate euchromatic and heterochromatic sides of the identified CTRs, respectively. (B) Expression levels of genes on the euchromatic or heterochromatic side of CTRs. Barplots represent Mean±SE for all genes (grey), genes within the 4kb region on the euchromatic side (yellow) or the heterochromatic side (green) of CTRs. The expression levels for genes on euchromatic side of CTRs are significantly greater than those of the genes on the heterochromatic side ($p<2.2E-16$, Wilcoxon rank sum test). (C) An example of 7 CTRs (red bars) predicted by CTRICS. Bar height reflects T-score, top and bottom rows denotes the orientation of the CTRs. The panel below CTR shows H3K27me3 domains called by SICER. RNA-Seq signal, RNA Pol II binding, as well as active histone modification (H3K4me3) are depleted in heterochromatic regions which have high H3K27me3, while they are enriched in euchromatic regions.

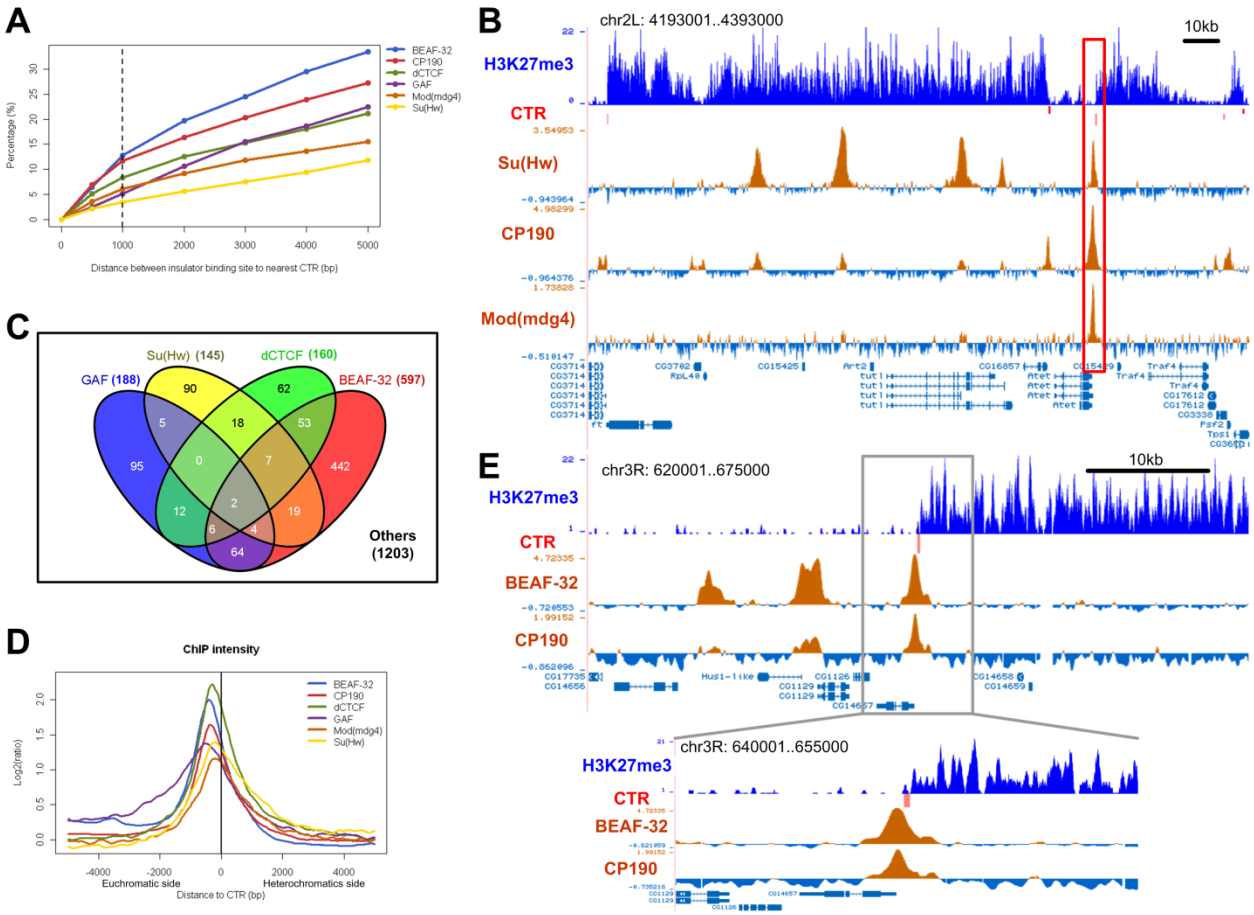


Figure 3-2. CTRs and the known insulator proteins in *Drosophila* S2 cell line. (A) Percentages of insulator protein binding sites are associated with a CTR. The x-axis shows the distance between insulator protein binding site and the nearest CTR, and y-axis shows the percentage of binding sites that are within a certain distance from the nearest CTR. The dashed line indicates the distance cutoff of 1kb, which is used for association analysis. (B) A 200kb region on chromosome 2L as an example. There are five Su(Hw) binding sites in this region, one is associated with a CTR (red bar, highlighted region), the others locate in regions enriched for H3K27me3. The intensities of co-factors (CP190, Mod(mdg4)) are relatively high at the CTR-associated binding site, and lower at the binding sites in the H3K27me3-enriched region. (C) Venn diagram shows the number of CTRs that are associated with four insulator proteins. Note that more than half (1203/2082) of the CTRs are not associated with any of the four insulator proteins. (D) Enrichment of insulator proteins in the ±5kb region around corresponding CTRs. The negative and positive distances also indicate the euchromatic and heterochromatic side of CTR, respectively. (E) An example illustrates the relative positions of a predicted CTR and the binding profiles of BEAF-32 and CP190. The peaks of the binding sites locate on the euchromatic side of the CTR, and the distance between the peaks of binding sites and the CTR midpoint is about 400bp.

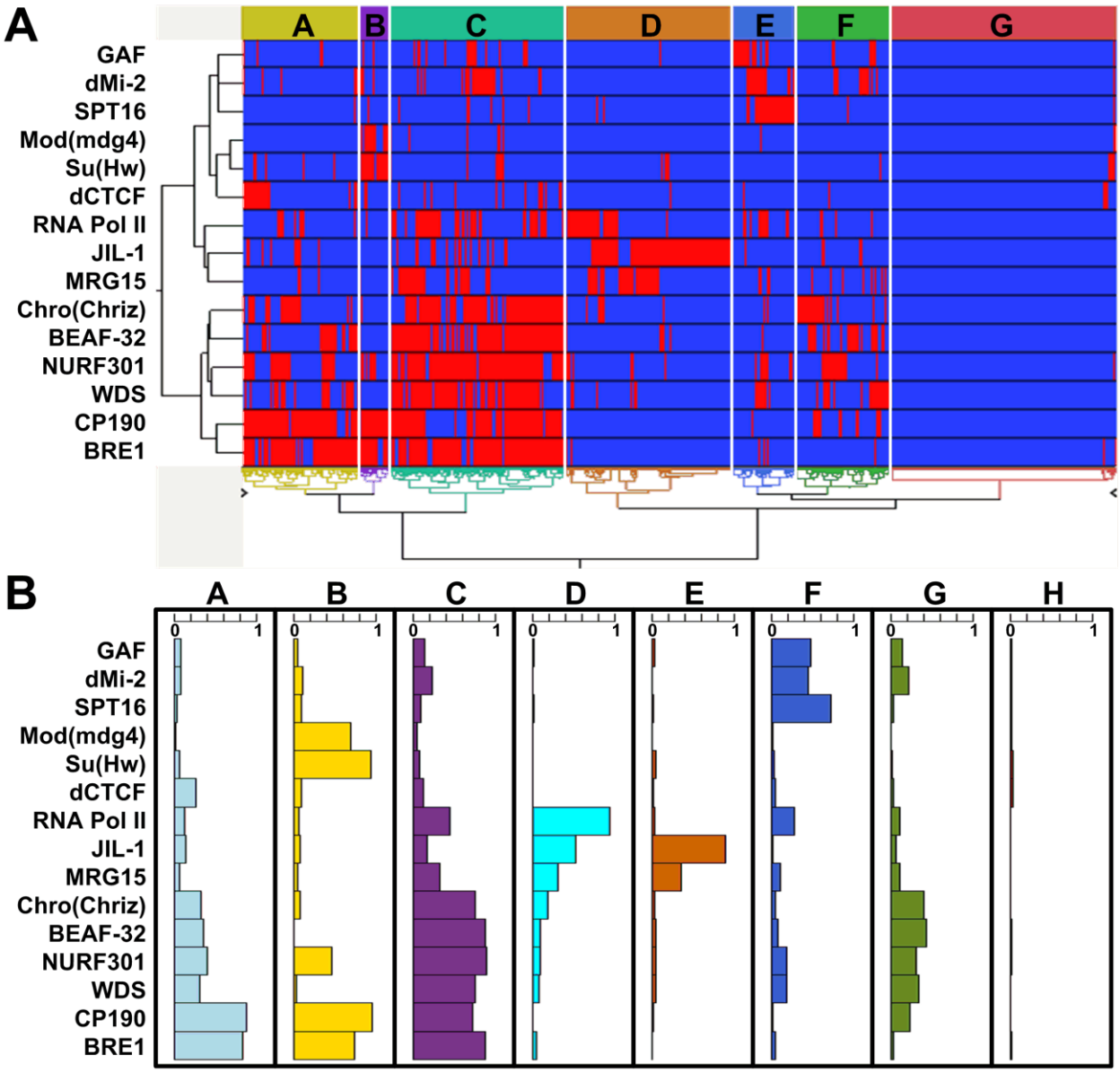


Figure 3-3. Subgroups of CTRs based on associated proteins in *Drosophila* S2 cell line. (A) Heat-map of the hierarchical clustering analysis result. Each column denotes a single CTR, and each row represents one protein included in the association analysis. The red and blue bars denote the presence or absence of an association with the corresponding CTR, respectively. Capital letters within colored boxes highlight the different subgroups of CTRs. (B) Proportions of CTRs in each subgroup (identified in (A)) that are associated with individual protein. The width of the bar indicates the percentage of CTRs in each group that are bound by the respective protein.

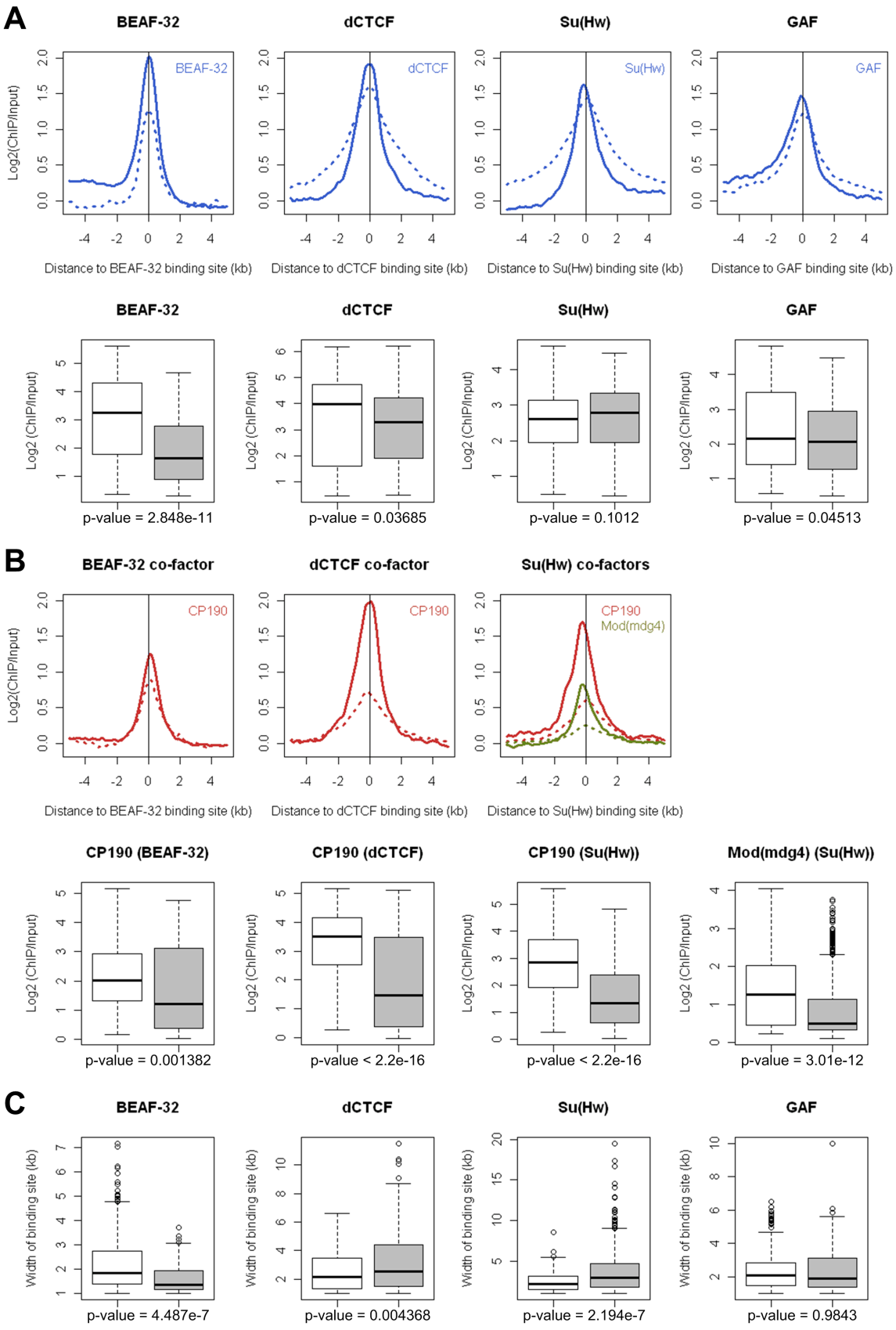

A
BEAF-32
dCTCF
Su(Hw)
GAF
Log2(ChIP/Input)
Distance to BEAF-32 binding site (kb)
Distance to dCTCF binding site (kb)
Distance to Su(Hw) binding site (kb)
Distance to GAF binding site (kb)
Log2 (ChIP/Input)
p-value = 2.848e-11
p-value = 0.03685
p-value = 0.1012
p-value = 0.04513
B
BEAF-32 co-factor
dCTCF co-factor
Su(Hw) co-factors
CP190
Mod(mdg4)
CP190 (BEAF-32)
CP190 (dCTCF)
CP190 (Su(Hw))
Mod(mdg4) (Su(Hw))
p-value = 0.001382
p-value < 2.2e-16
p-value < 2.2e-16
p-value = 3.01e-12
C
Width of binding site (kb)
p-value = 4.487e-7
p-value = 0.004368
p-value = 2.194e-7
p-value = 0.9843

Figure 3-4. Binding intensity and patterns of insulator proteins and co-factors associated with CTRs in *Drosophila* S2 cell line. The enrichment levels of respective insulator proteins (A) and co-factors (B) around binding sites associated with CTR (solid lines) or located in H3K27me3-enriched region (dashed lines). For CTR-associated binding sites, negative and positive distances denote euchromatic and heterochromatic side. Box plots show the peak values for individual insulator proteins (A) and co-factors (B) at binding sites associated with CTR (open box) or in heterochromatic regions (shaded box). (C) Box plots of the width of insulator proteins binding patterns at binding sites associated with CTR (open box) or in heterochromatic regions (shaded box). P-values were all calculated by Wilcoxon rank sum test.

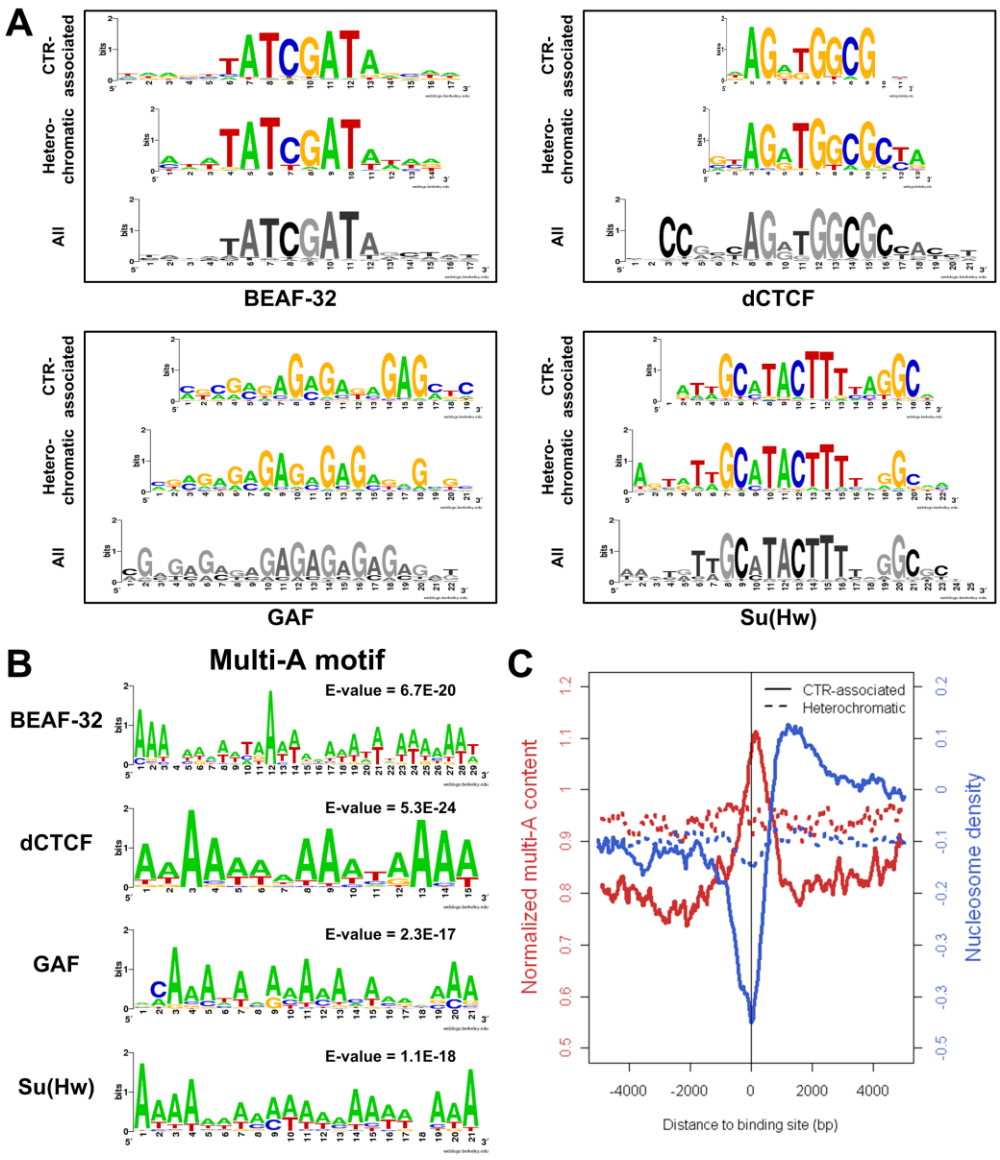


Figure 3-5. Cis-elements associated with CTRs in *Drosophila* S2 cell line. (A) Logos representation of motifs identified from DNA sequences underlying insulator protein binding sites associated with CTRs (CTR-associated) or in H3K27me3-enriched (Heterochromatic) regions. Motifs obtained with all binding sites are represented at the bottom. (B) Multi-A motifs are the discriminative motif identified by MEME for CTR-associated binding sites vs. heterochromatic binding sites. (C) Multi-A (AAAA/TTTT) content (normalized to genome average, red curve) and nucleosome density (blue curve) around CTR-associated insulator protein binding sites (solid line) and heterochromatic binding sites (dashed line). Data presents combined value for all the insulator proteins, dCTCF, Su(Hw), GAF, and BEAF-32. For CTR-associated binding sites, negative and positive distances denote euchromatic and heterochromatic side.

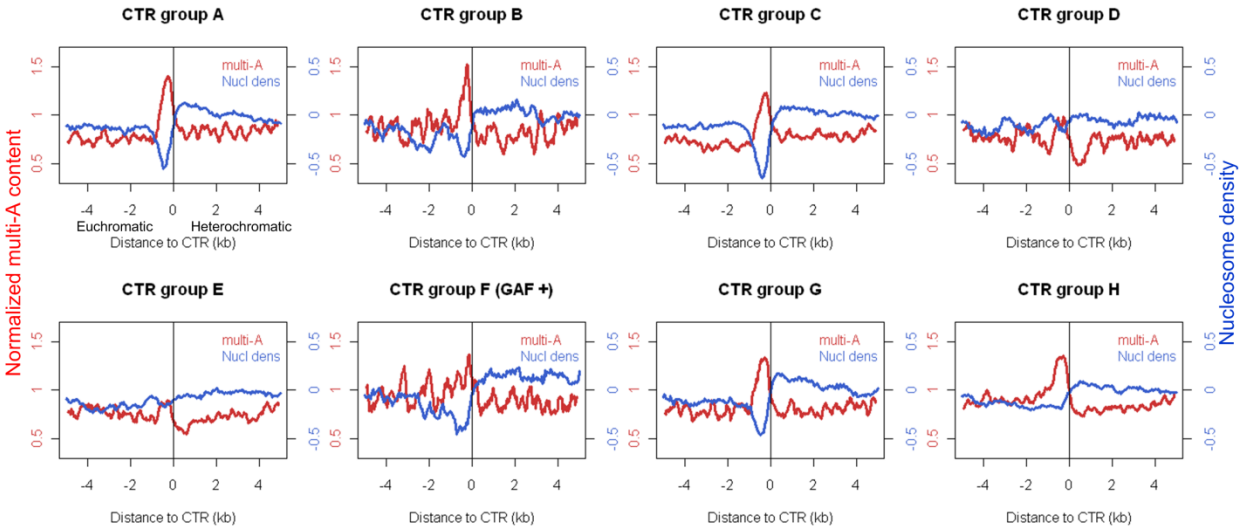
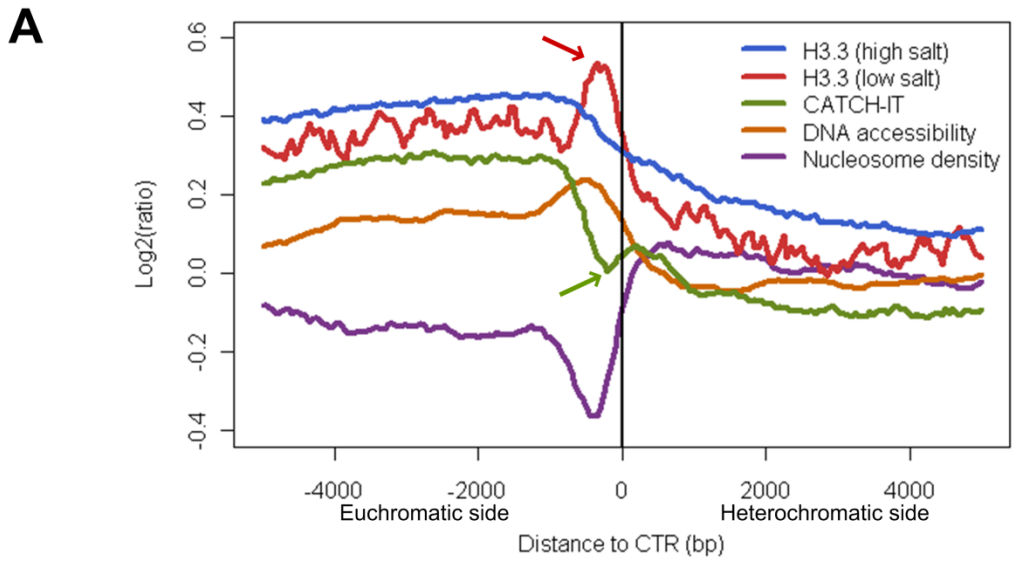
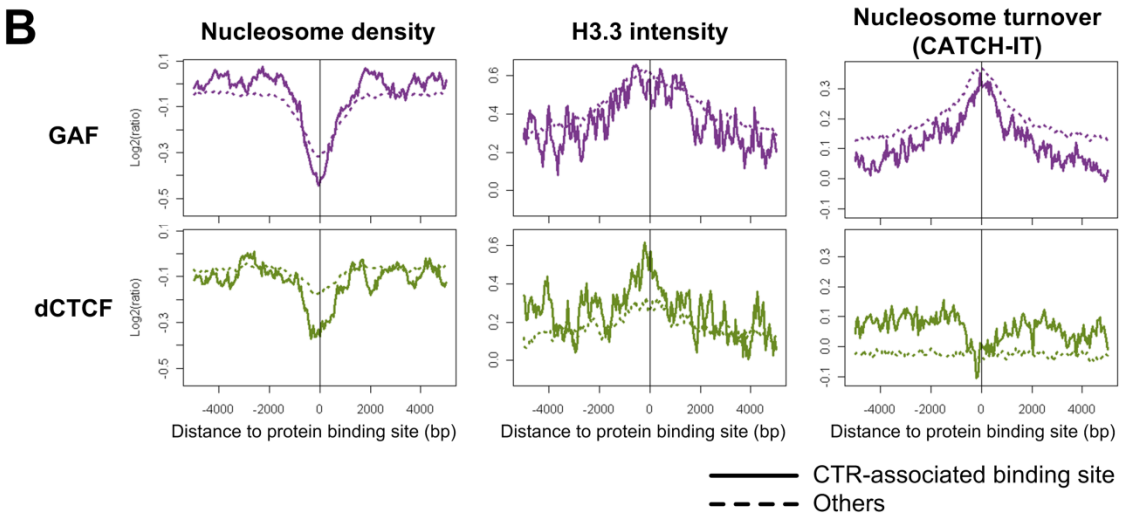


Figure 3-6. Multi-A (AAAA/TTTT) content (normalized to genome average, red curve) and nucleosome density (blue curve) around individual subgroup of CTRs in *Drosophila* S2 cell line (For group F only those co-localized with GAF were included). The negative and positive distances denote the euchromatic and heterochromatic sides of CTR, respectively.


A
H3.3 (high salt)
H3.3 (low salt)
CATCH-IT
DNA accessibility
Nucleosome density
Log2(ratio)
0.6
0.4
0.2
0.0
-0.2
-0.4
-4000
-2000
0
2000
4000
Euchromatic side
Heterochromatic side
Distance to CTR (bp)


B
Nucleosome density
H3.3 intensity
Nucleosome turnover (CATCH-IT)
GAF
dCTCF
Log2(ratio)
Distance to protein binding site (bp)
CTR-associated binding site
Others

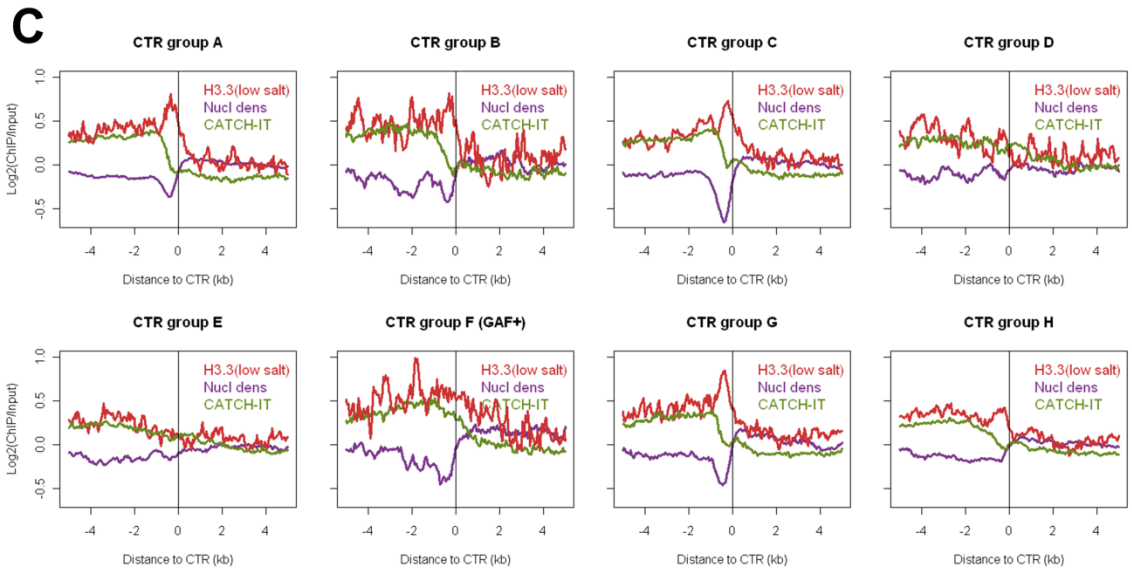

C
CTR group A
CTR group B
CTR group C
CTR group D
CTR group E
CTR group F (GAF+)
CTR group G
CTR group H
H3.3(low salt)
Nucl dens
CATCH-IT
Log2(ChIP/Input)
Distance to CTR (kb)

Figure 3-7. Contrasting patterns of H3.3 enrichment and nucleosome turnover rate associated with subgroups of CTRs in *Drosophila* S2 cell line. (A) Composite plot for all CTRs. H3.3 (low salt) incorporation is enriched on the euchromatic side of CTRs (red arrow), while nucleosome turnover rate (CATCH-IT) is drops down sharply at the same region (green arrow). (B) H3.3 enrichment and CATCH-IT measurements of nucleosome turnover rate moves to the same direction for GAF (both CTR-associated and others). In contrast, for CTR-associated dCTCF binding sites, the enrichment of H3.3 is accompanied by decreased turnover rate. (C) Plots of H3.3 enrichment (red), nucleosome turnover rate (green, measured with CATCH-IT), and nucleosome density (purple) for each subgroup of the CTRs (for group F only those co-localized with GAF were included). Note the contrasting pattern between H3.3 enrichment and CATCH-IT profile in subgroups A, B, C, G, but not in subgroups D and E.

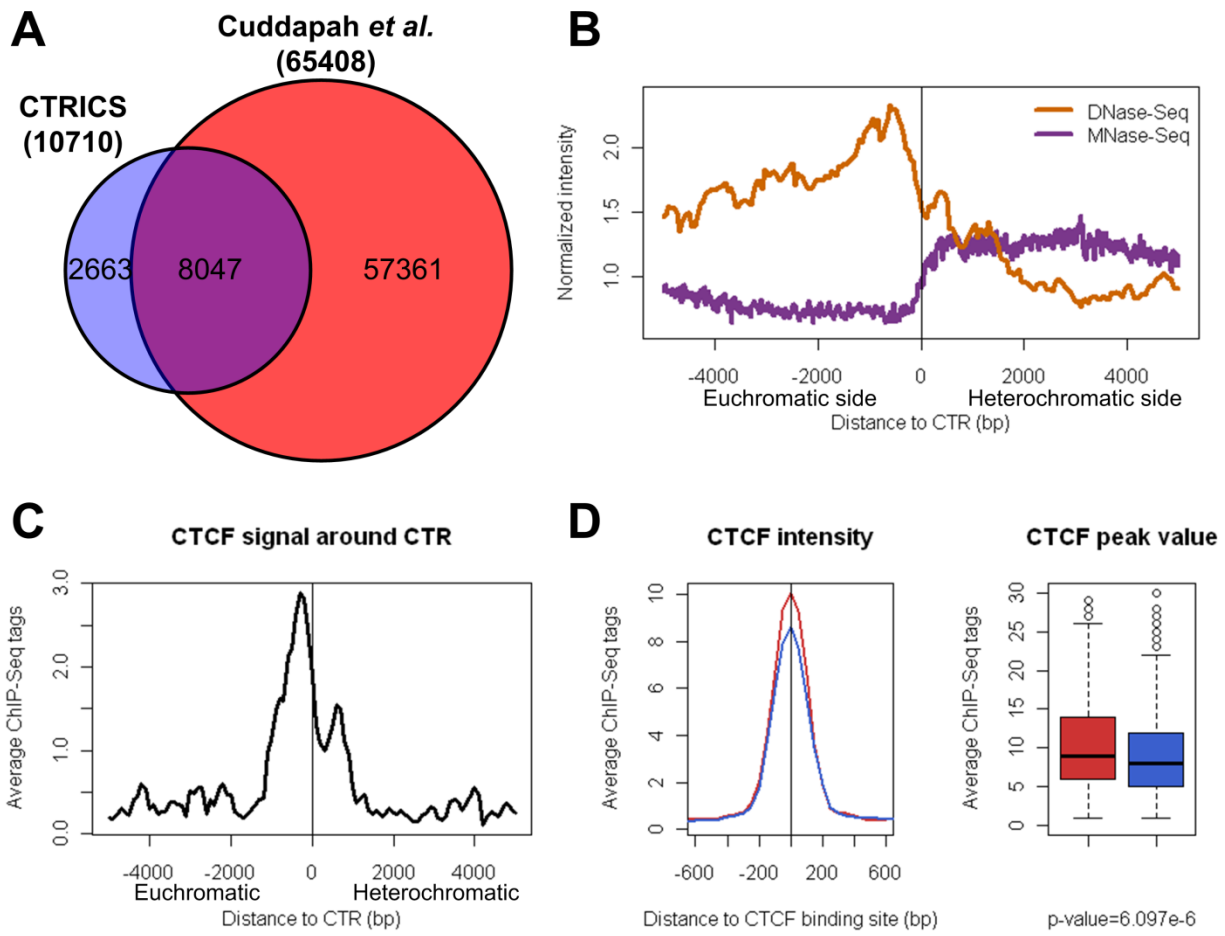


Figure 3-8. Chromatin transitional regions in human HeLa cell line. (A) 6852 predicted CTRs in HeLa cells are overlapping (within 1kb) with the chromatin barrier regions in the previous study. (B) DNA accessibility (measured by DNase-Seq) and nucleosome density (measured by MNase-Seq) around all the predicted CTRs in HeLa cell line (normalized to genome average). The negative and positive distances denote the euchromatic and heterochromatic sides of CTR, respectively. (C) Binding pattern of insulator protein CTCF around the CTRs which co-localize with CTCF. The negative and positive distances also denote the euchromatic and heterochromatic sides of CTR, respectively. (D) The enrichment level of CTCF around CTR-associated (red) and heterochromatic (blue) binding sites. Box-plots show the peak values of CTCF at the CTR-associated (red) and heterochromatic (blue) binding sites. The peak values of CTCF at CTR-associated binding sites were significantly greater than that at the heterochromatic binding sites (p-value = 6.097e-6, Wilcoxon rank sum test).

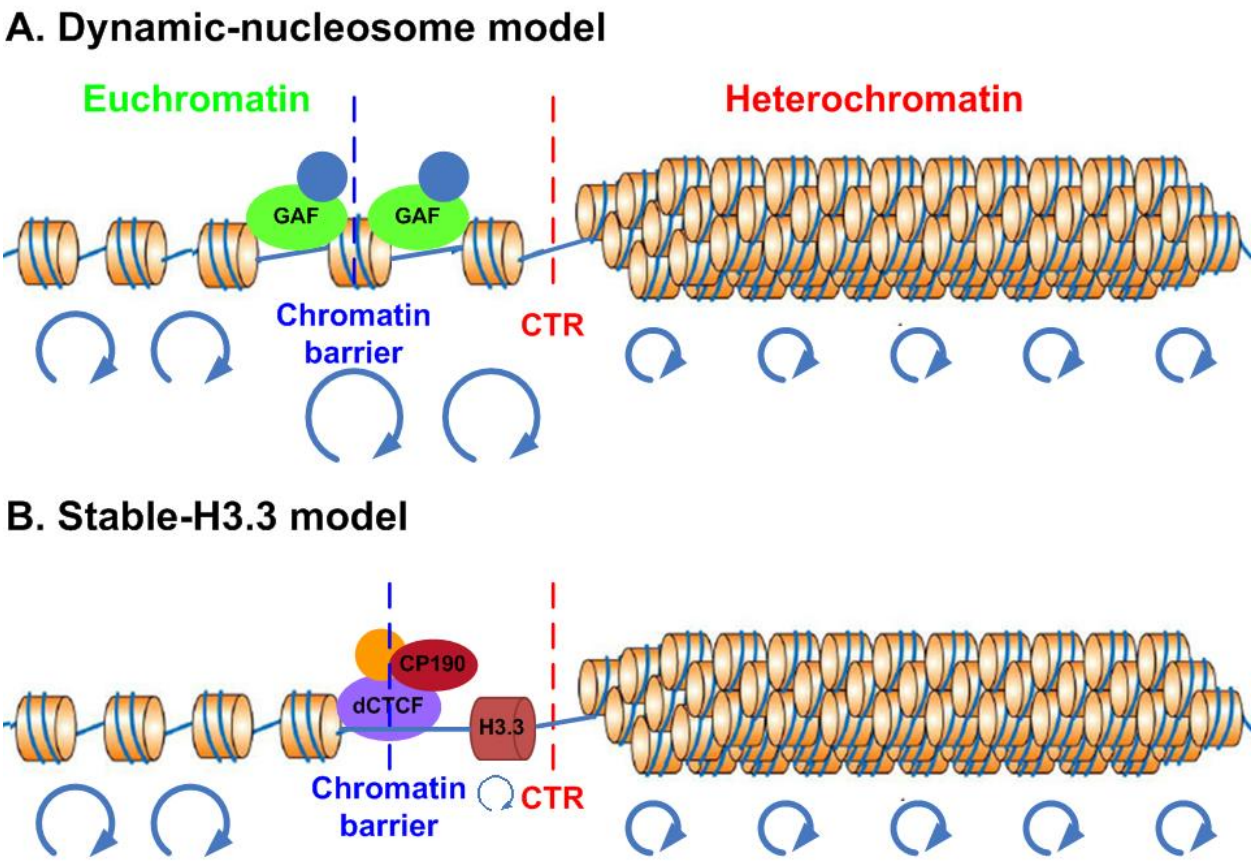


Figure 3-9. Proposed models for facultative heterochromatin boundary. Models represent distinct features of GAF-associated (A) vs. dCTCF-associated (B) CTRs. The red and blue dashed lines denote the position of CTR and chromatin barrier, respectively. The blue circles at the bottom of each model indicate the nucleosome turnover rate, the bigger the circles, the faster the nucleosomes turnover. For dCTCF-associated CTRs, the increased enrichment of H3.3 is coupled with decreased turnover rate.

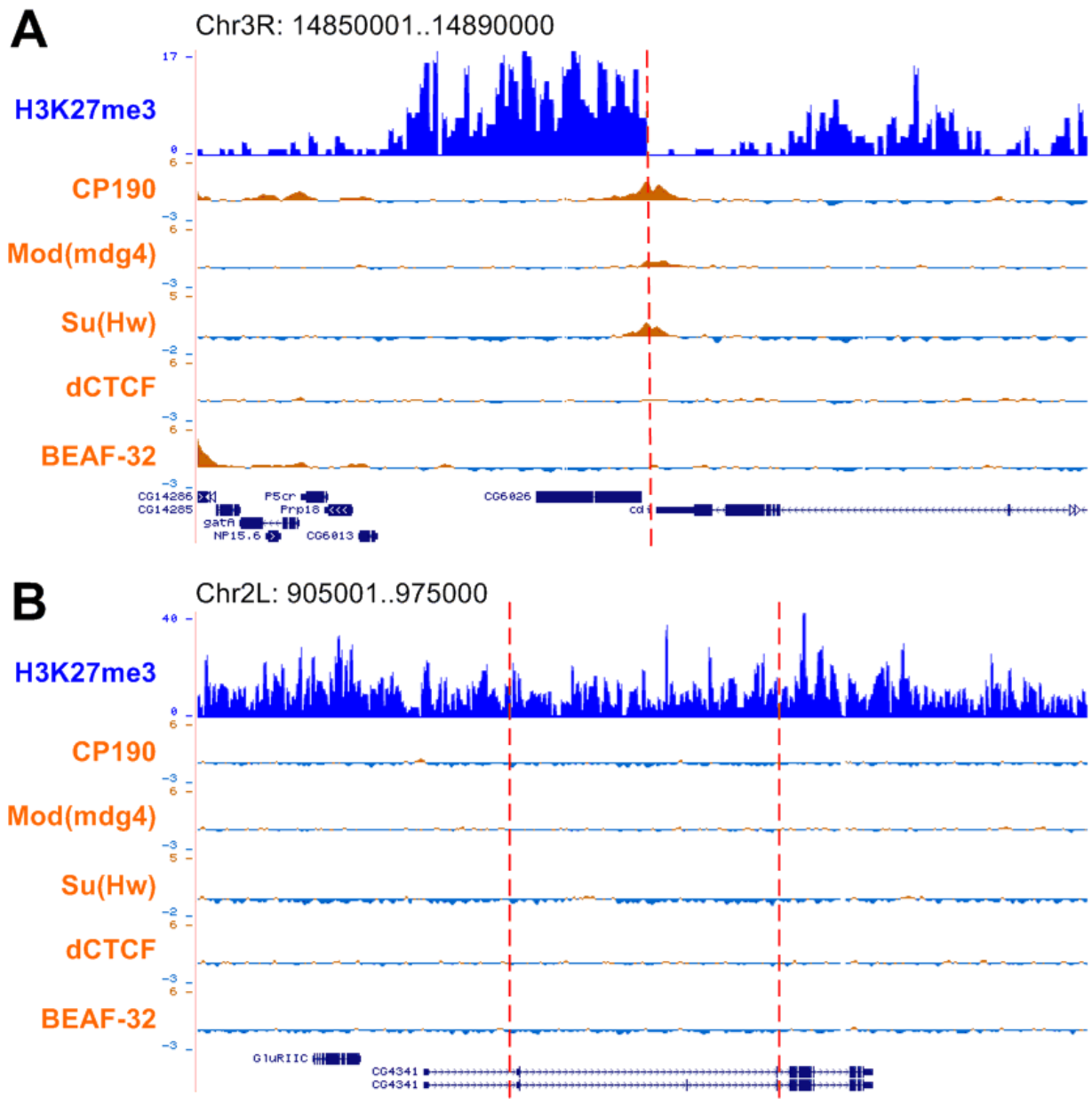


Figure 3-10. Construction of empirical positive and negative evaluation datasets.

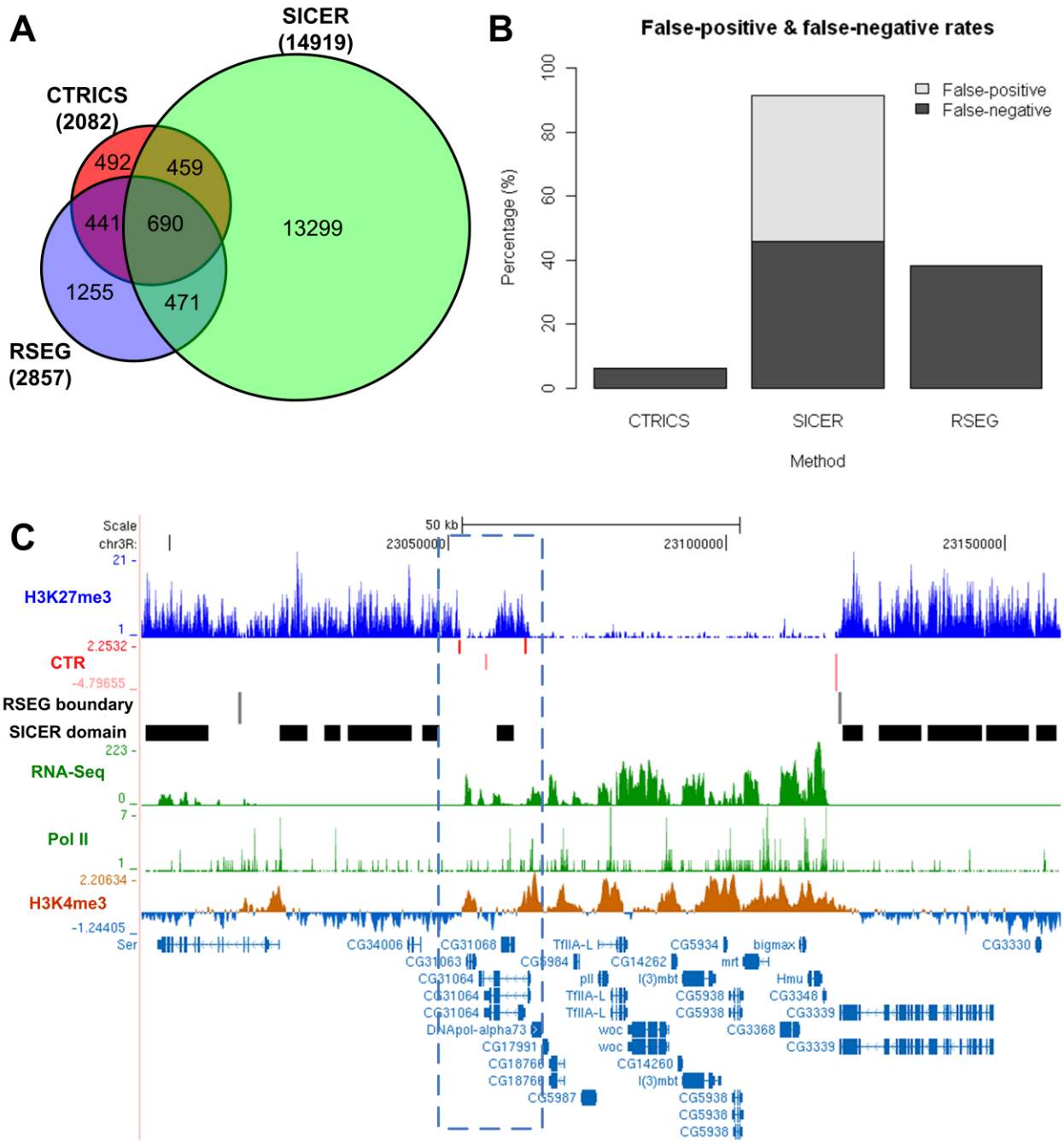


Figure 3-11. Comparison of CTRICS with SICER and RSEG. (A) The Venn diagram shows the number of CTRs (predicted by CTRICS) that are overlapping with the chromatin boundaries predicted by the other two methods in S2 cells. (B) False-positive and false-negative rates for different methods based on the empirical evaluation datasets. (C) A region shows several examples of CTRs predicted by CTRICS (red bar), H3K27me3 boundaries predicted by RSEG (grey bar), and H3K27me3 domains predicted by SICER. RSEG missed the three boundaries shown in the blue dashed block, which can be corroborated with RNA-Seq and H3K4me3 ChIP-chip data.

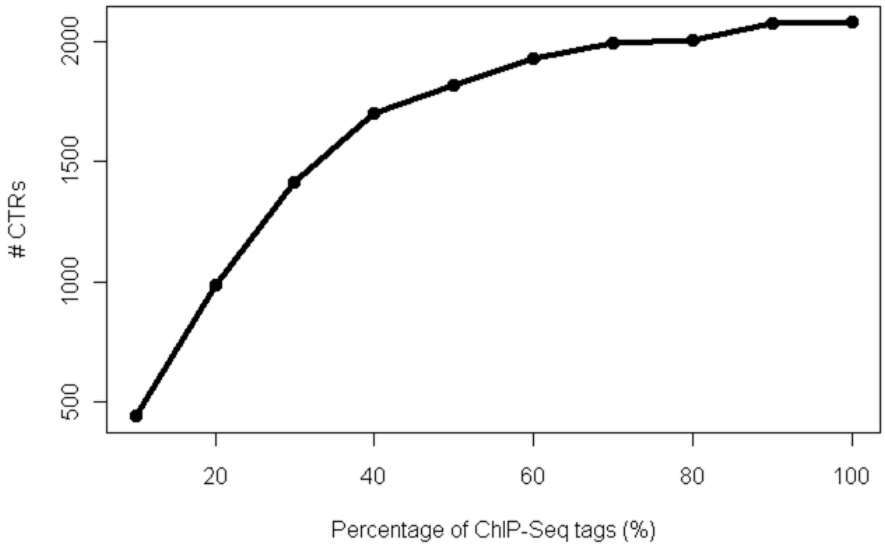


Figure 3-12. Sequencing depth analysis. In order to test if the H3K27me3 ChIP-Seq dataset has reached saturation status and if sequencing depth has any influence on the CTR prediction, we conducted the sequencing depth analysis. We first randomly extracted a series of subsamples (10%, 20%, 30%, and so on until 90% of the original tags) from the H3K27me3 ChIP-Seq dataset without replacement. We then identified chromatin transitional regions in each subsample using CTRICS with default parameters. The x-axis of the plot represents the percentage of subsample tags compared to the total tags (~2.8 x10^6), and y-axis indicates the number of CTRs identified.

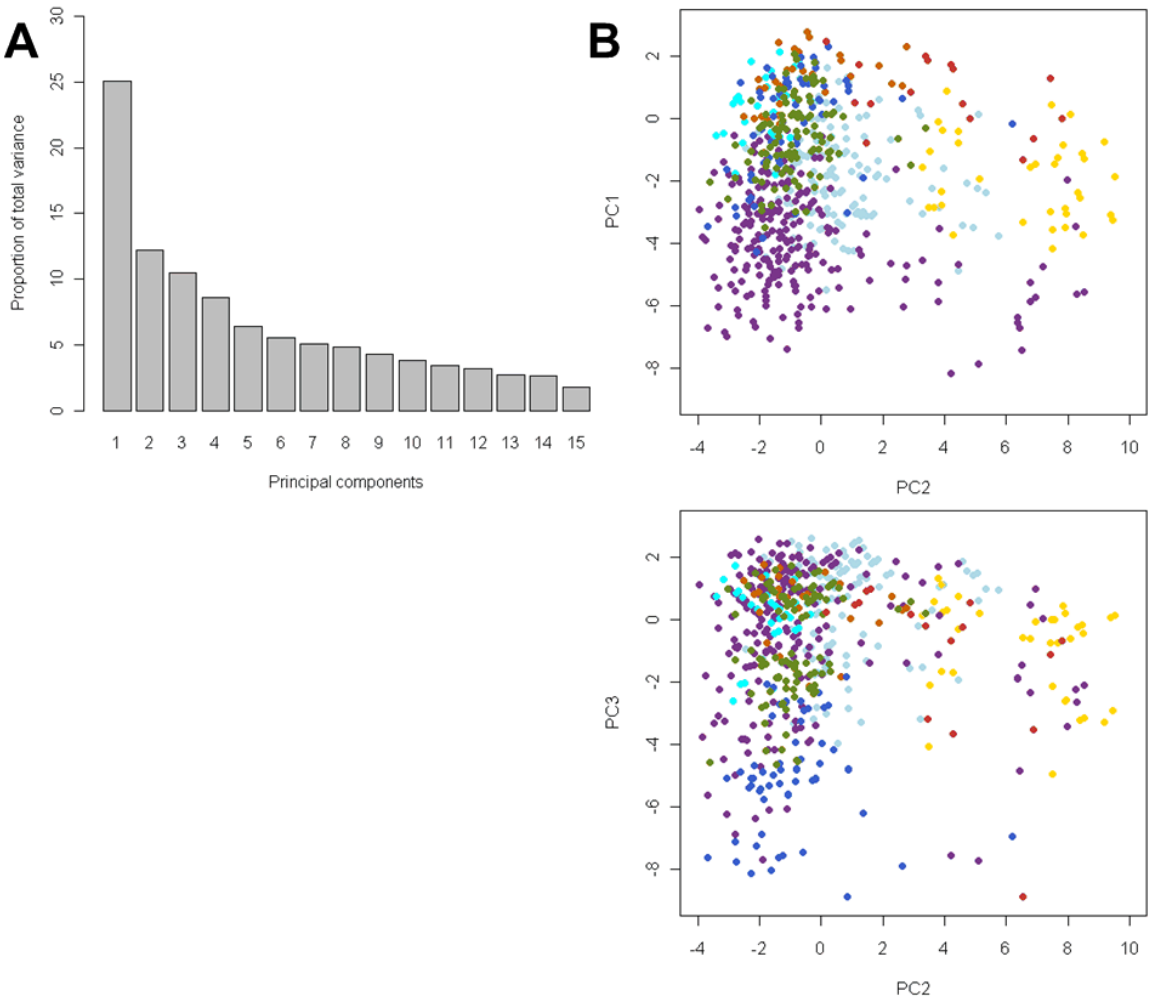
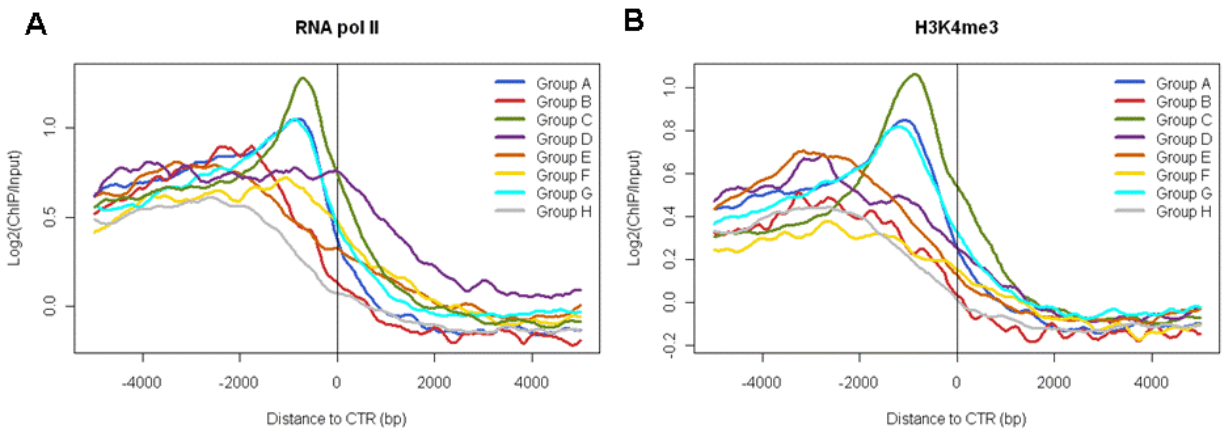


Figure 3-13. Principal component analysis of CTRs based on association with the 15 proteins. (A) Percentage of total variance accounted for by individual principal components. (B) Two-dimensional projections onto the first three principal components. Different colors of the dots represent different groups of CTRs corresponding to the groups shown in the hierarchical clustering result (Figure 3-3).



C

CTR group	TSS	TES	Intergenic	Exon	Intron
A (279)	135 (48.4%)	49 (17.6%)	63 (22.6%)	12 (4.3%)	20 (7.2%)
B (74)	14 (18.9%)	22 (29.7%)	20 (27.0%)	6 (8.1%)	12 (16.2%)
C (418)	265 (63.3%)	46 (11.0%)	50 (12.0%)	31 (7.4%)	26 (6.2%)
D (125)	34 (27.2%)	29 (23.2%)	4 (3.2%)	42 (33.6%)	16 (12.8%)
E (272)	48 (17.6%)	79 (29.0%)	19 (7.0%)	96 (35.3%)	30 (11.0%)
F (151)	21 (13.9%)	11 (7.3%)	41 (27.2%)	20 (13.2%)	58 (38.4%)
G (228)	100 (43.9%)	41 (18.0%)	52 (22.8%)	11 (4.8%)	24 (10.5%)
H (534)	72 (13.5%)	166 (31.1%)	135 (25.3%)	83 (15.5%)	78 (14.6%)

Figure 3-14. Genomic distribution of CTRs. The average intensities of RNA polymerase II (A) and H3K4me3 (B) around individual groups of CTR. (C) The distribution of CTRs in each group.

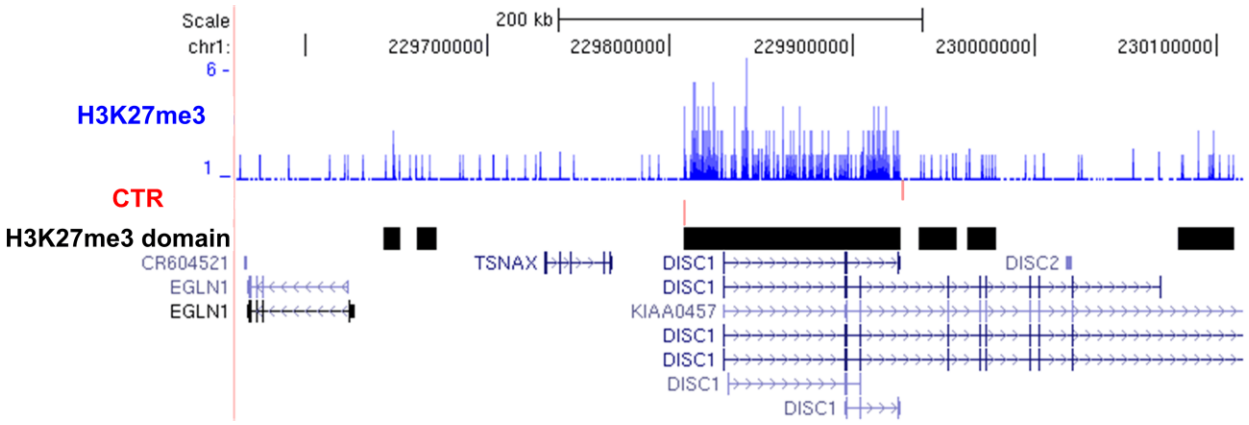


Figure 3-15. An example of 2 CTRs (red bars) predicted by CTRICS in human HeLa cells. The panel below CTR shows H3K27me3 domains predicted in Cuddapah et al. 2009. CTRICS identifies the boundaries with a significant drop of H3K27me3 level, but ignores the boundaries with minor changes in H3K27me3 signal.

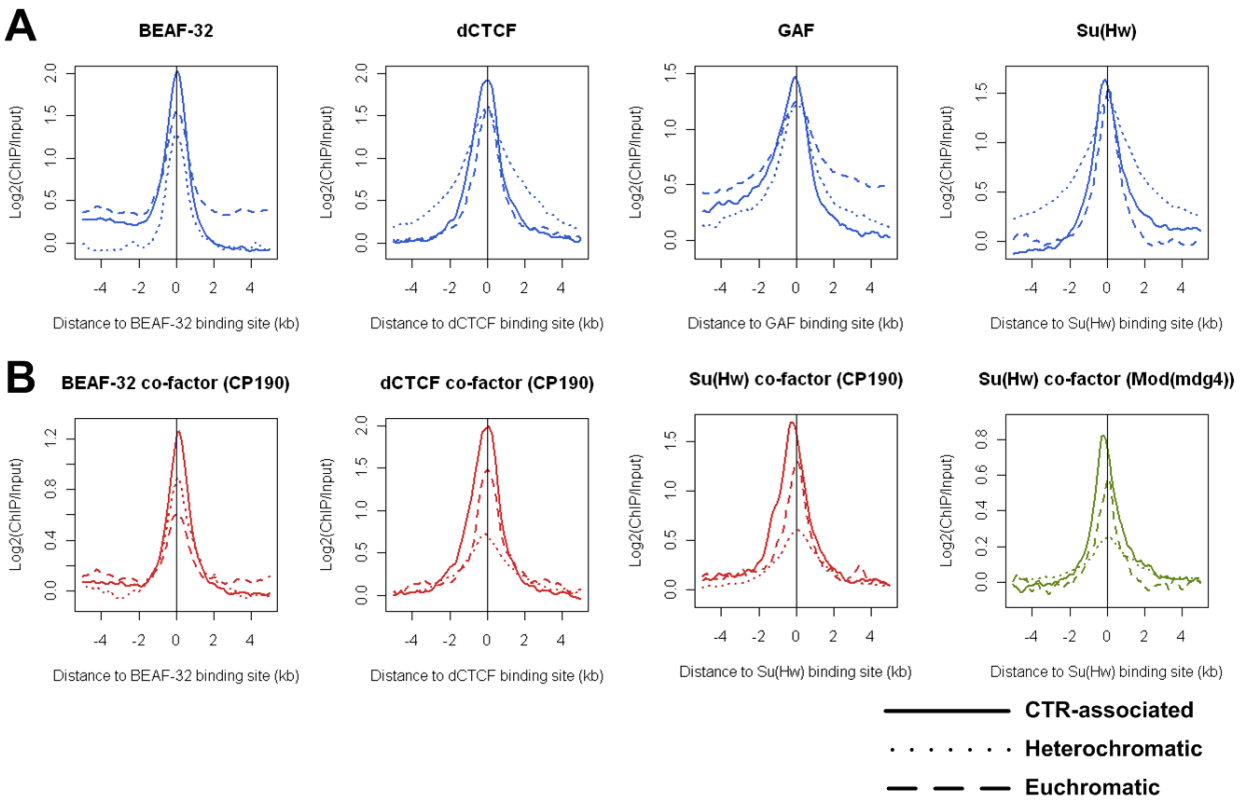


Figure 3-16. Binding patterns of co-factors are different for CTR-associated and euchromatic binding sites. Binding patterns of insulator proteins (A) and their co-factors (B) around CTR-associated (solid curve), heterochromatic (dotted curve) and euchromatic (break curve) binding sites in *Drosophila* S2 cells. For CTR-associated binding sites, negative and positive distances denote euchromatic and heterochromatic side.

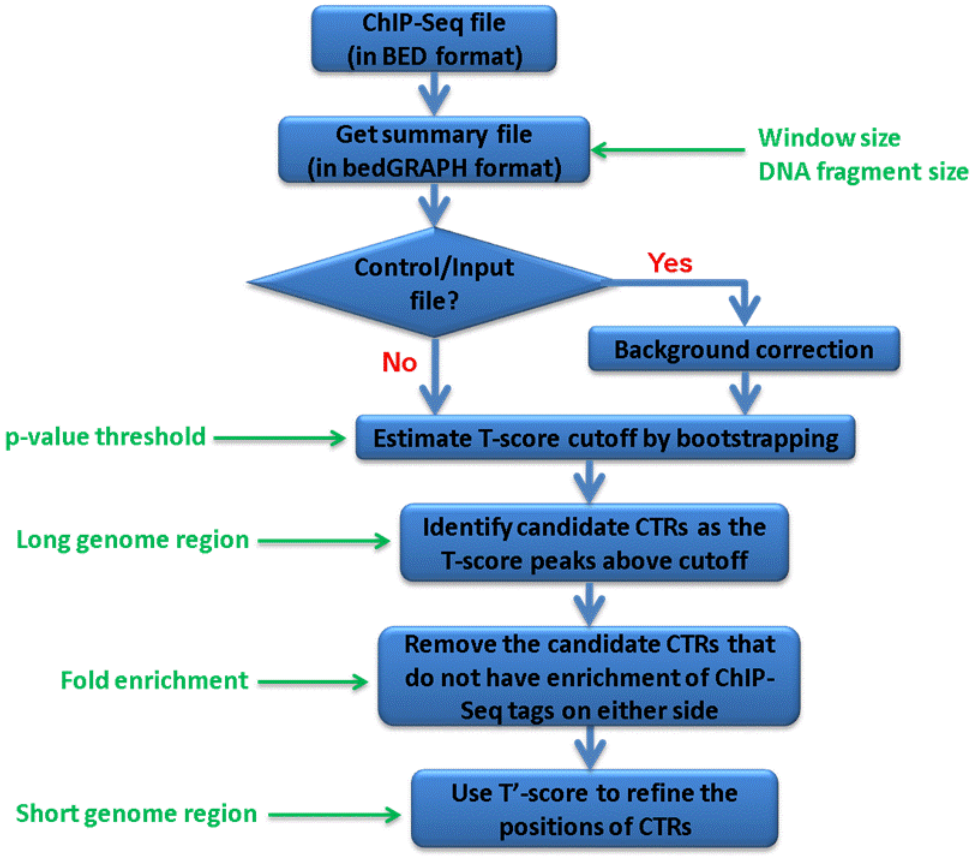


Figure 3-17. Flowchart of CTRICS. The green characters are the parameters needed in each step.

CHAPTER 4
DISCUSSIONS, EXPLORATIVE WORKS, AND PERSPECTIVES

More and more evidences have demonstrated the crucial role of epigenetic regulation in multiple biological processes such as apoptosis, development, cell differentiation, and tumorigenesis. The high-order chromatin structure is not only an efficient way to compact DNA into nucleus, but also an elaborate approach to store the heritable information of epigenomic landscapes. In specific, the close coordination between appropriate chromatin modifications and systematic changes of chromatin state will help to maintain the stable gene expression profile in particular developmental stages, as well as to dynamically adjust the gene expression pattern in response to developmental and environmental stimuli.

The series of studies in our lab have shown *in vitro* and *in vivo* that chromatin barriers and enhancer elements play a fundamentally crucial role in the regulation of chromatin landscapes. My *in silico* work takes advantage of the current available genome-wide datasets to expand and verify our observations and hypothesis in the genome scale.

Epigenomics Era: New Opportunities and New Challenges

Since the development of new techniques, especially next-generation high-throughput sequencing, we entered into the era of epigenomics, which studies DNA methylation and histone modifications layer on top of the genome. One representative feature of epigenomics research is the generation of tons of genome-scale, high-resolution datasets. Several national and international collaborative projects such as modENCODE and ENCODE projects, and lots of individual laboratory's effort have

greatly advanced epigenomics study by producing high quality datasets and state-of-the-art analysis tools and algorithms.

Potential Opportunities from Large-Scale Epigenomics

The understanding of genetic information of many organisms has been well established, but the epigenomic landscapes have not been systematically studied in the genome-wide scale. The major hurdle is the large number of epigenomes there may exist even within a single individual. Each individual has essentially one genome, whereas the individual is believed to have a distinct epigenome in each cell type and tissue.

Given the extreme complexity involved in cellular regulation, the emerging epigenome maps will help us reveal new principles in the process. For example, the genome-wide epigenomics datasets will provide us a comprehensive chromatin map, which will facilitate the identification of transcription factors, regulatory molecules and pathways, target genes of epigenetic features (Ernst and Kellis, 2010). Mapping of different epigenetic marks, such as histone modifications, DNA methylation and non-coding RNAs simultaneously in the same cell type will help us to better understand the coordination among these regulatory mechanisms. Epigenomic datasets also have the valuable power to identify *cis*-regulatory elements like enhancers, insulators, and loci poised for activation (Bernstein et al., 2006b; Dindot et al., 2009; Heintzman et al., 2009b; Ozsolak et al., 2008b).

With respect to diseases, epigenomic maps will also provide a unique resource allowing us to identify responsible factors that contribute to the disease as well as downstream genes affected by the disease. In addition, epigenetic states can serve as

biomarkers for certain diseases and maybe useful for disease diagnosis and prognosis (Esteller, 2008b).

Challenges with Large-Scale Epigenomics

The possible combinations of cell types, disease states, individual variations and environmental stresses make the number of possible distinct epigenomes seems astronomical. Thus the effective application and interpretation of these datasets generated from diverse laboratories requires the development of standards, which will help in improving comparability between datasets, avoiding duplication of effort, and so on. The epigenomics community is going to benefit most from a systematic and organized collaboration in which similar epigenetic features are investigated in a defined set of cell types and tissues, using a standardized protocols and quality controls, eventually generating high-quality datasets that are comparable with one another (Satterlee et al., 2010). The standardized approach would guarantee the reliable identification of epigenomic features in particular cell states.

Another challenge is the development and standardization of computational methods to process and display the large epigenomics datasets. As the scale of epigenomic datasets continues to increase, more sophisticated methods, such as statistical modeling and machine learning techniques, are needed to uncover the underlying patterns behind the massive and complicated data. In addition, user-friendly data visualizing tools are also extremely welcomed by not only experimental biologists, since they will provide the valuable direct visualization and interpretation of the datasets.

Application of Machine Learning to the Prediction of Chromatin Boundaries

As mentioned above, more sophisticated methods, like machine learning are needed to elucidate the mysteries underlying large biological datasets. In Chapter 3, it has been shown that the predicted CTRs in *Drosophila* S2 cells can be divided into 8 subgroups based on the binding patterns of 15 proteins (Figure 3-3). I also tried to apply the machine learning method called Support Vector Machine (SVM) to study the relationship between the 15 proteins and CTRs, and to predict chromatin boundaries solely based on the 15 proteins or even less information (Figure 4-1).

Support vector machines are supervised learning models that can be applied to classification (binary output) and regression (continuous output). The essential of SVM is to use the kernel functions to map the inputs (of dimension n) into a higher dimensional space (dimension>n) so that different classes can be classified with a linear hyperplane. It has been proved that this hyperplane always exists when using the appropriate kernel functions. So in practice, the most important thing is to choose the right kernel function with the most suitable parameters, in order to get the best prediction accuracy as well as not to over-fit the model. Over-fitting, which means the model fits perfectly with training data but has little prediction power with new data, is a common drawback should be carefully examined and avoid.

In order to apply SVM to predict chromatin boundaries, the independent and dependent variables need to be specified. Here the binary CTR status was defined as dependent variable, where the 2082 CTRs in *Drosophila* S2 cells were taken as positive set (chromatin boundaries), and the binding sites of the 4 insulator-binding proteins (including BEAF-32, dCTCF, GAF, Su(Hw)) in heterochromatic regions were taken as negative set (not chromatin boundaries). The peak or mean signals of the 15 proteins

(Figure 3-3) or the 6 insulator proteins (including BEAF-32, dCTCF, GAF, Su(Hw), CP190, Mod(mdg2)) in the 2kb region centered on CTR were taken as independent variables (Figure 4-1A). We are aiming to differentiate the chromatin boundaries from the non-boundaries binding sites based on the signals of the proteins using support vector machine. In addition, the 10-fold cross-validation was used to evaluate the model, where positive and negative sets were divided into 10 subgroups respectively, and 9 positive and 9 negative subgroups were taken as training set, whereas the other subgroups were taken as testing set. Then repeat this process 10 times and each time using a different subgroup as testing set.

As a result, we find that the models generated by SVM can successfully differentiate chromatin boundaries from the non-boundary regions. The area under the ROC (receiver operating characteristic) curves for the models using peak or mean signals of the 15 proteins can reach as high as 0.971, and model with only 6 insulator proteins also has great prediction power (AUC=0.936) (Figure 4-1B). In the future studies, the SVM models may be applied to predict chromatin boundaries in other *Drosophila* cell types and developmental stages. And it may even be applied to mammalian systems, but of course other independent variables should be utilized since some of the 15 proteins do not have homologies in mammals.

Experimental Verification of the Predicted Chromatin Boundaries

In Chapter 3, it has been observed that for some groups of CTR, the signal of nucleosome density has a dip while DNA accessibility signal has a peak on the euchromatic side of CTRs (Figure 3-6, 3-7). This means for these CTRs, the regions on their euchromatic side are open and accessible to DNase I digestion. In order to experimentally verify this observation, I conducted DNase I sensitivity assay on

Drosophila S2 cells, followed by qPCR assay on 5 selected CTRs. As a result, 4 of the selected CTRs are DNase I hypersensitivity sites (Figure 4-2).

It is of great benefit for computational biologists to be directly involved in molecular biology experiments, since such a practice can help the understanding of biological processes, promote the generation of biological hypothesis, and facilitate the interpretation of computational outcomes. During my Ph.D. training, I tried to be actively involved in the bench work and to establish an approach to experimentally verify the findings from my bioinformatics analysis. These experiences taught me that reasoning in a biologically meaningful way is crucial for the success of computational biologists.

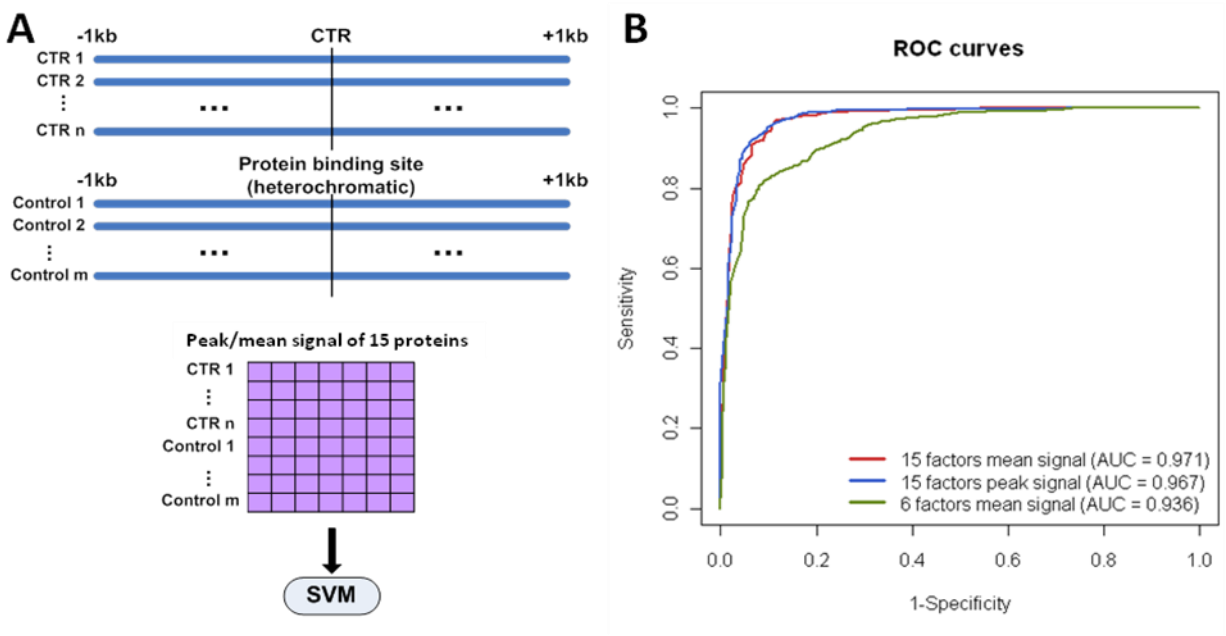


Figure 4-1. Application of support vector machine (SVM) to predict chromatin boundaries. (A) Schematic figure shows how the independent and dependent variables were chosen. (B) Receiver operating characteristic (ROC) curves for different models. The AUC (area under curve) for the three models are 0.971, 0.967, 0.936 respectively.

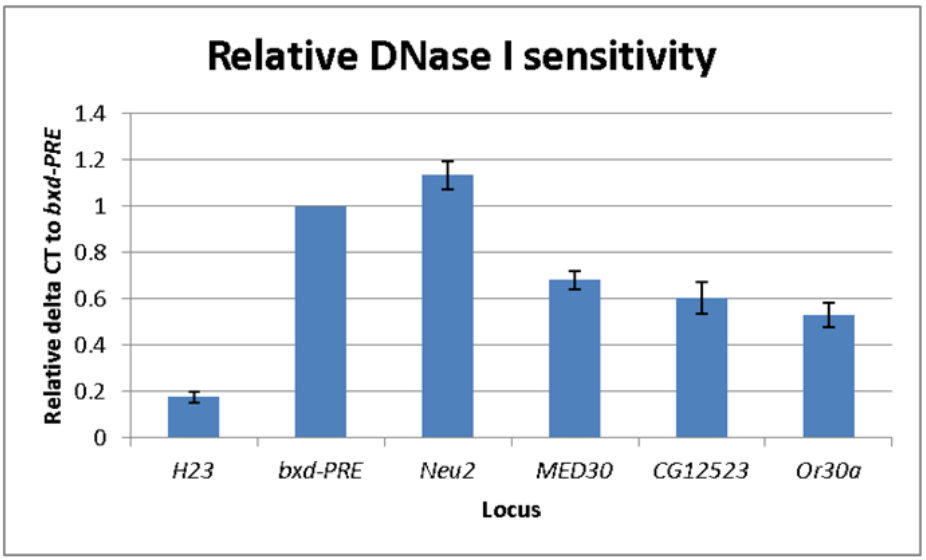


Figure 4-2. Experimental verification of chromatin boundaries. I performed DNase I sensitivity assay on *Drosophila* S2 cells, and five intergenic CTRs predicted by CTRICS were selected for qPCR verification. Among them, four CTRs near genes *Neu2*, *MED30*, *CG12523* and *Or30a* were proved to be DNase I hypersensitive sites. H23 and bxd-PRE were used as negative and positive controls, respectively. Barplot shows Mean±SEM.

LIST OF REFERENCES

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

Aas, T., Borresen, A.L., Geisler, S., SmithSorensen, B., Johnsen, H., Varhaug, J.E., Akslen, L.A., and Lonning, P.E. (1996). Specific P53 mutations are associated with de novo resistance to doxorubicin in breast cancer patients. Nature Medicine *2*, 811-814.

Allan, J., Hartman, P.G., Crane-Robinson, C., and Aviles, F.X. (1980). The structure of histone H1 and its location in chromatin. Nature *288*, 675-679.

Allis, C.D., Berger, S.L., Cote, J., Dent, S., Jenuwien, T., Kouzarides, T., Pillus, L., Reinberg, D., Shi, Y., Shiekhattar, R.*, et al.* (2007). New nomenclature for chromatin-modifying enzymes. Cell *131*, 633-636.

Bailey, T.L., Boden, M., Whitington, T., and Machanick, P. (2010). The value of position-specific priors in motif discovery using MEME. BMC Bioinformatics *11*, 179.

Bao, X., Deng, H., Johansen, J., Girton, J., and Johansen, K.M. (2007). Loss-of-function alleles of the JIL-1 histone H3S10 kinase enhance position-effect variegation at pericentric sites in Drosophila heterochromatin. Genetics *176*, 1355-1358.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823-837.

Baylin, S.B., and Ohm, J.E. (2006). Epigenetic gene silencing in cancer - a mechanism for early oncogenic pathway addiction? Nat Rev Cancer *6*, 107-116.

Beke, L., Nuytten, M., Van Eynde, A., Beullens, M., and Bollen, M. (2007). The gene encoding the prostatic tumor suppressor PSP94 is a target for repression by the Polycomb group protein EZH2. Oncogene *26*, 4590-4595.

Bell, O., Schwaiger, M., Oakeley, E.J., Lienert, F., Beisel, C., Stadler, M.B., and Schubeler, D. (2010). Accessibility of the Drosophila genome discriminates PcG repression, H4K16 acetylation and replication timing. Nat Struct Mol Biol *17*, 894-900.

Benetti, R., Gonzalo, S., Jaco, I., Munoz, P., Gonzalez, S., Schoeftner, S., Murchison, E., Andl, T., Chen, T., Klatt, P.*, et al.* (2008). A mammalian microRNA cluster controls DNA methylation and telomere recombination via Rbl2-dependent regulation of DNA methyltransferases. Nat Struct Mol Biol *15*, 268-279.

Bennett, M., Macdonald, K., Chan, S.W., Luzio, J.P., Simari, R., and Weissberg, P. (1998). Cell surface trafficking of Fas: A rapid mechanism of p53-mediated apoptosis. Science *282*, 290-293.

Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., and Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K.*, et al.* (2006a). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315-326.

Bernstein, B.E., Mikkelsen, T.S., Xie, X.H., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K.*, et al.* (2006b). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315-326.

Bi, X., and Broach, J.R. (1999). UASrpg can function as a heterochromatin boundary element in yeast. Genes Dev *13*, 1089-1101.

Bi, X., and Broach, J.R. (2001). Chromosomal boundaries in S. cerevisiae. Curr Opin Genet Dev *11*, 199-204.

Bilodeau, S., Kagey, M.H., Frampton, G.M., Rahl, P.B., and Young, R.A. (2009). SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state. Genes Dev *23*, 2484-2489.

Blankenberg, D., Taylor, J., and Nekrutenko, A. (2011). Making whole genome multiple alignments usable for biologists. Bioinformatics *27*, 2426-2428.

Boeckler, F.M., Joerger, A.C., Jaggi, G., Rutherford, T.J., Veprintsev, D.B., and Fersht, A.R. (2008). Targeted rescue of a destabilized mutant of p53 by an in silico screened drug. P Natl Acad Sci USA *105*, 10360-10365.

Breen, T.R., and Harte, P.J. (1991). Molecular characterization of the trithorax gene, a positive regulator of homeotic gene expression in Drosophila. Mech Dev *35*, 113-127.

Brodsky, M.H., Sekelsky, J.J., Tsang, G., Hawley, R.S., and Rubin, G.M. (2000). mus304 encodes a novel DMA damage checkpoint protein required during Drosophila development. Gene Dev *14*, 666-678.

Brown, J.L., Fritsch, C., Mueller, J., and Kassis, J.A. (2003). The Drosophila pho-like gene encodes a YY1-related DNA binding protein that is redundant with pleiohomeotic in homeotic gene silencing. Development *130*, 285-294.

Brown, J.L., Mucci, D., Whiteley, M., Dirksen, M.L., and Kassis, J.A. (1998). The Drosophila Polycomb group gene pleiohomeotic encodes a DNA binding protein with homology to the transcription factor YY1. Mol Cell *1*, 1057-1064.

Bushey, A.M., Dorman, E.R., and Corces, V.G. (2008). Chromatin insulators: regulatory mechanisms and epigenetic inheritance. Mol Cell *32*, 1-9.

Bushey, A.M., Ramos, E., and Corces, V.G. (2009). Three subclasses of a Drosophila insulator show distinct and cell type-specific genomic distributions. Genes Dev *23*, 1338-1350.

Bykov, V.J.N., Issaeva, N., Shilov, A., Hultcrantz, M., Pugacheva, E., Chumakov, P., Bergman, J., Wiman, K.G., and Selivanova, G. (2002). Restoration of the tumor suppressor function to mutant p53 by a low-molecular-weight compound. Nature Medicine *8*, 282-288.

Cao, R., and Zhang, Y. (2004). SUZ12 is required for both the histone methyltransferase activity and the silencing function of the EED-EZH2 complex. Mol Cell *15*, 57-67.

Caretti, G., Di Padova, M., Micales, B., Lyons, G.E., and Sartorelli, V. (2004). The Polycomb Ezh2 methyltransferase regulates muscle gene expression and skeletal muscle differentiation. Genes Dev *18*, 2627-2638.

Cedar, H., and Bergman, Y. (2009). Linking DNA methylation and histone modification: patterns and paradigms. Nature Reviews Genetics *10*, 295-304.

Chen, H., Tu, S.W., and Hsieh, J.T. (2005). Down-regulation of human DAB2IP gene expression mediated by polycomb Ezh2 complex and histone deacetylase in prostate cancer. J Biol Chem *280*, 22437-22444.

Chen, X., Chen, Z., Chen, H., Su, Z., Yang, J., Lin, F., Shi, S., and He, X. (2012). Nucleosomes suppress spontaneous mutations base-specifically in eukaryotes. Science *335*, 1235-1238.

Choy, M.K., Movassagh, M., Goh, H.G., Bennett, M.R., Down, T.A., and Foo, R.S. (2010). Genome-wide conserved consensus transcription factor binding motifs are hyper-methylated. BMC Genomics *11*, 519.

Christophorou, M.A., Ringshausen, I., Finch, A.J., Swigart, L.B., and Evan, G.I. (2006). The pathological response to DNA damage does not contribute to p53-mediated tumour suppression. Nature *443*, 214-217.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A.*, et al.* (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A *107*, 21931-21936.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res *14*, 1188-1190.

Cuddapah, S., Jothi, R., Schones, D.E., Roh, T.Y., Cui, K., and Zhao, K. (2009). Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. Genome Res *19*, 24-32.

Czermin, B., Melfi, R., McCabe, D., Seitz, V., Imhof, A., and Pirrotta, V. (2002). Drosophila enhancer of Zeste/ESC complexes have a histone H3 methyltransferase activity that marks chromosomal Polycomb sites. Cell *111*, 185-196.

Dalgliesh, G.L., Furge, K., Greenman, C., Chen, L., Bignell, G., Butler, A., Davies, H., Edkins, S., Hardy, C., Latimer, C.*, et al.* (2010). Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. Nature *463*, 360-363.

Deal, R.B., Henikoff, J.G., and Henikoff, S. (2010). Genome-wide kinetics of nucleosome turnover determined by metabolic labeling of histones. Science *328*, 1161-1164.

Di Micco, R., Fumagalli, M., Cicalese, A., Piccinin, S., Gasparini, P., Luise, C., Schurra, C., Garre, M., Nuciforo, P.G., Bensimon, A.*, et al.* (2006). Oncogene-induced senescence is a DNA damage response triggered by DNA hyper-replication. Nature *444*, 638-642.

Dindot, S.V., Person, R., Strivens, M., Garcia, R., and Beaudet, A.L. (2009). Epigenetic profiling at mouse imprinted gene clusters reveals novel epigenetic and genetic features at differentially methylated regions. Genome Res *19*, 1374-1383.

Dion, M.F., Kaplan, T., Kim, M., Buratowski, S., Friedman, N., and Rando, O.J. (2007). Dynamics of replication-independent histone turnover in budding yeast. Science *315*, 1405-1408.

Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F.*, et al.* (2012). Landscape of transcription in human cells. Nature *489*, 101-108.

Donehower, L.A., Godley, L.A., Aldaz, C.M., Pyle, R., Shi, Y.P., Pinkel, D., Gray, T., Bradley, A., Medina, D., and Varmus, H.E. (1995). Deficiency of P53 Accelerates Mammary Tumorigenesis in Wnt-1 Transgenic Mice and Promotes Chromosomal Instability. Gene Dev *9*, 882-895.

Ehrlich, M., Gama-Sosa, M.A., Huang, L.H., Midgett, R.M., Kuo, K.C., McCune, R.A., and Gehrke, C. (1982). Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. Nucleic Acids Res *10*, 2709-2721.

El-Deiry, W.S. (1998). Regulation of p53 downstream genes. Seminars in Cancer Biology *8*, 345-357.

Epsztejn-Litman, S., Feldman, N., Abu-Remaileh, M., Shufaro, Y., Gerson, A., Ueda, J., Deplus, R., Fuks, F., Shinkai, Y., Cedar, H.*, et al.* (2008). De novo DNA methylation promoted by G9a prevents reprogramming of embryonically silenced genes. Nature Structural & Molecular Biology *15*, 1176-1183.

Ernst, J., and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol *28*, 817-U894.

Esteller, M. (2007). Cancer epigenomics: DNA methylomes and histone-modification maps. Nat Rev Genet *8*, 286-298.

Esteller, M. (2008a). Epigenetics in cancer. N Engl J Med *358*, 1148-1159.

Esteller, M. (2008b). Molecular origins of cancer: Epigenetics in cancer. New Engl J Med *358*, 1148-1159.

Esteller, M., Corn, P.G., Baylin, S.B., and Herman, J.G. (2001). A gene hypermethylation profile of human cancer. Cancer Res *61*, 3225-3229.

Ezhkova, E., Pasolli, H.A., Parker, J.S., Stokes, N., Su, I.H., Hannon, G., Tarakhovsky, A., and Fuchs, E. (2009). Ezh2 orchestrates gene expression for the stepwise differentiation of tissue-specific stem cells. Cell *136*, 1122-1135.

Feinberg, A.P. (2007). Phenotypic plasticity and the epigenetics of human disease. Nature *447*, 433-440.

Feldman, N., Gerson, A., Fang, J., Li, E., Zhang, Y., Shinkai, Y., Cedar, H., and Bergman, Y. (2006). G9a-mediated irreversible epigenetic inactivation of Oct-3/4 during early embryogenesis. Nat Cell Biol *8*, 188-U155.

Fernandes, A.D., and Atchley, W.R. (2008). Biochemical and functional evidence of p53 homology is inconsistent with molecular phylogenetics for distant sequences. Journal of Molecular Evolution *67*, 51-67.

Filion, G.J., van Bemmel, J.G., Braunschweig, U., Talhout, W., Kind, J., Ward, L.D., Brugman, W., de Castro, I.J., Kerkhoven, R.M., Bussemaker, H.J.*, et al.* (2010). Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. Cell *143*, 212-224.

Fischle, W., Wang, Y., Jacobs, S.A., Kim, Y., Allis, C.D., and Khorasanizadeh, S. (2003). Molecular basis for the discrimination of repressive methyl-lysine marks in histone H3 by Polycomb and HP1 chromodomains. Genes Dev *17*, 1870-1881.

Fraga, M.F., Ballestar, E., Villar-Garea, A., Boix-Chornet, M., Espada, J., Schotta, G., Bonaldi, T., Haydon, C., Ropero, S., Petrie, K.*, et al.* (2005). Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer. Nat Genet *37*, 391-400.

Fritsch, C., Brown, J.L., Kassis, J.A., and Muller, J. (1999). The DNA-binding polycomb group protein pleiohomeotic mediates silencing of a Drosophila homeotic gene. Development *126*, 3905-3913.

Gan, Q., Schones, D.E., Ho Eun, S., Wei, G., Cui, K., Zhao, K., and Chen, X. (2010). Monovalent and unpoised status of most genes in undifferentiated cell-enriched Drosophila testis. Genome Biol *11*, R42.

Gaszner, M., and Felsenfeld, G. (2006). Insulators: exploiting transcriptional and epigenetic mechanisms. Nat Rev Genet *7*, 703-713.

Gerstein, M.B., Kundaje, A., Hariharan, M., Landt, S.G., Yan, K.K., Cheng, C., Mu, X.J., Khurana, E., Rozowsky, J., Alexander, R.*, et al.* (2012). Architecture of the human regulatory network derived from ENCODE data. Nature *489*, 91-100.

Geyer, P.K., Spana, C., and Corces, V.G. (1986). On the molecular mechanism of gypsy-induced mutations at the yellow locus of Drosophila melanogaster. EMBO J *5*, 2657-2662.

Ghosh, D., Gerasimova, T.I., and Corces, V.G. (2001). Interactions between the Su(Hw) and Mod(mdg4) proteins required for gypsy insulator function. EMBO J *20*, 2518-2527.

Gilbert, M.K., Tan, Y.Y., and Hart, C.M. (2006). The Drosophila boundary element-associated factors BEAF-32A and BEAF-32B affect chromatin structure. Genetics *173*, 1365-1375.

Girton, J.R., and Johansen, K.M. (2008). Chromatin structure and the regulation of gene expression: the lessons of PEV in Drosophila. Adv Genet *61*, 1-43.

Goll, M.G., and Bestor, T.H. (2005). Eukaryotic cytosine methyltransferases. Annu Rev Biochem *74*, 481-514.

Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R., and Young, R.A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. Cell *130*, 77-88.

Gui, Y., Guo, G., Huang, Y., Hu, X., Tang, A., Gao, S., Wu, R., Chen, C., Li, X., Zhou, L.*, et al.* (2011). Frequent mutations of chromatin remodeling genes in transitional cell carcinoma of the bladder. Nat Genet *43*, 875-878.

Gurudatta, B.V., and Corces, V.G. (2009). Chromatin insulators: lessons from the fly. Brief Funct Genomic Proteomic *8*, 276-282.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P.*, et al.* (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223-227.

Harrison, D.A., Gdula, D.A., Coyne, R.S., and Corces, V.G. (1993). A leucine zipper domain of the suppressor of Hairy-wing protein mediates its repressive effect on enhancer function. Genes Dev *7*, 1966-1978.

He, H.H., Meyer, C.A., Shin, H., Bailey, S.T., Wei, G., Wang, Q., Zhang, Y., Xu, K., Ni, M., Lupien, M.*, et al.* (2010). Nucleosome dynamics define transcriptional enhancers. Nat Genet *42*, 343-347.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W.*, et al.* (2009a). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108-112.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W.*, et al.* (2009b). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108-112.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A.*, et al.* (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet *39*, 311-318.

Henikoff, S. (2008). Nucleosome destabilization in the epigenetic regulation of gene expression. Nat Rev Genet *9*, 15-26.

Henikoff, S., Henikoff, J.G., Sakai, A., Loeb, G.B., and Ahmad, K. (2009). Genome-wide profiling of salt fractions maps physical properties of chromatin. Genome Res *19*, 460-469.

Hon, G., Ren, B., and Wang, W. (2008). ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome. PLoS Comput Biol *4*, e1000201.

Hon, G., Wang, W., and Ren, B. (2009). Discovery and annotation of functional chromatin signatures in the human genome. PLoS Comput Biol *5*, e1000566.

Hu, G., Cui, K., Northrup, D., Liu, C., Wang, C., Tang, Q., Ge, K., Levens, D., Crane-Robinson, C., and Zhao, K. (2013). H2A.Z facilitates access of active and repressive complexes to chromatin in embryonic stem cell self-renewal and differentiation. Cell Stem Cell *12*, 180-192.

Huang, S., Li, X., Yusufzai, T.M., Qiu, Y., and Felsenfeld, G. (2007). USF1 recruits histone modification complexes and is critical for maintenance of a chromatin barrier. Mol Cell Biol *27*, 7991-8002.

Jenuwein, T., and Allis, C.D. (2001). Translating the histone code. Science *293*, 1074-1080.

Jia, J., Zheng, X., Hu, G., Cui, K., Zhang, J., Zhang, A., Jiang, H., Lu, B., Yates, J., 3rd, Liu, C.*, et al.* (2012). Regulation of pluripotency and self- renewal of ESCs through epigenetic-threshold modulation and mRNA pruning. Cell *151*, 576-589.

Jin, C., Zang, C., Wei, G., Cui, K., Peng, W., Zhao, K., and Felsenfeld, G. (2009). H3.3/H2A.Z double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions. Nat Genet *41*, 941-945.

Jirtle, R.L., and Skinner, M.K. (2007). Environmental epigenomics and disease susceptibility. Nat Rev Genet *8*, 253-262.

Johannes, F., Wardenaar, R., Colome-Tatche, M., Mousson, F., de Graaf, P., Mokry, M., Guryev, V., Timmers, H.T., Cuppen, E., and Jansen, R.C. (2010). Comparing genome-wide chromatin profiles using ChIP-chip or ChIP-seq. Bioinformatics *26*, 1000-1006.

Junttila, M.R., and Evan, G.I. (2009). p53-a Jack of all trades but master of none. Nature Reviews Cancer *9*, 821-829.

Kahn, T.G., Schwartz, Y.B., Dellino, G.I., and Pirrotta, V. (2006). Polycomb complexes and the propagation of the methylation mark at the Drosophila ubx gene. J Biol Chem *281*, 29064-29075.

Kamijo, T., Zindy, F., Roussel, M.F., Quelle, D.E., Downing, J.R., Ashmun, R.A., Grosveld, G., and Sherr, C.J. (1997). Tumor suppression at the mouse INK4a locus mediated by the alternative reading frame product p19(ARF). Cell *91*, 649-659.

Karpen, G.H. (1994). Position-effect variegation and the new biology of heterochromatin. Curr Opin Genet Dev *4*, 281-291.

Kel, A.E., Gossling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O.V., and Wingender, E. (2003). MATCH: A tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res *31*, 3576-3579.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res *12*, 996-1006.

Kharchenko, P.V., Alekseyenko, A.A., Schwartz, Y.B., Minoda, A., Riddle, N.C., Ernst, J., Sabo, P.J., Larschan, E., Gorchakov, A.A., Gu, T.*, et al.* (2011). Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. Nature *471*, 480-485.

Kleer, C.G., Cao, Q., Varambally, S., Shen, R., Ota, I., Tomlins, S.A., Ghosh, D., Sewalt, R.G., Otte, A.P., Hayes, D.F.*, et al.* (2003). EZH2 is a marker of aggressive breast cancer and promotes neoplastic transformation of breast epithelial cells. Proc Natl Acad Sci U S A *100*, 11606-11611.

Laajala, T.D., Raghav, S., Tuomela, S., Lahesmaa, R., Aittokallio, T., and Elo, L.L. (2009). A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments. BMC Genomics *10*, 618.

Lee, T.I., Jenner, R.G., Boyer, L.A., Guenther, M.G., Levine, S.S., Kumar, R.M., Chevalier, B., Johnstone, S.E., Cole, M.F., Isono, K.*, et al.* (2006). Control of developmental regulators by Polycomb in human embryonic stem cells. Cell *125*, 301-313.

Levine, A.J., and Oren, M. (2009). The first 30 years of p53: growing ever more complex. Nature Reviews Cancer *9*, 749-758.

Lewis, E.B. (1978). A gene complex controlling segmentation in Drosophila. Nature *276*, 565-570.

Li, G.Y., and Zhou, L. (2013). Genome-Wide Identification of Chromatin Transitional Regions Reveals Diverse Mechanisms Defining the Boundary of Facultative Heterochromatin. Plos One *8*.

Li, M., Belozerov, V.E., and Cai, H.N. (2010). Modulation of chromatin boundary activities by nucleosome-remodeling activities in Drosophila melanogaster. Mol Cell Biol *30*, 1067-1076.

Li, M., He, Y., Dubois, W., Wu, X., Shi, J., and Huang, J. (2012). Distinct regulatory mechanisms and functions for p53-activated and p53-repressed DNA damage response genes in embryonic stem cells. Mol Cell *46*, 30-42.

Li, X., Wang, S., Li, Y., Deng, C., Steiner, L.A., Xiao, H., Wu, C., Bungert, J., Gallagher, P.G., Felsenfeld, G.*, et al.* (2011). Chromatin boundaries require functional collaboration between the hSET1 and NURF complexes. Blood *118*, 1386-1394.

Lin, N., Li, X., Cui, K., Chepelev, I., Tie, F., Li, G., Liu, B., Harte, P., Zhao, K., Huang, S.*, et al.* (2011). A Barrier-only Boundary Element Delimits the Formation of Facultative Heterochromatin in Drosophila and Vertebrates. Mol Cell Biol.

Marchenko, N.D., Zaika, A., and Moll, U.M. (2000). Death signal-induced localization of p53 protein to mitochondria - A potential role in apoptotic signaling. Journal of Biological Chemistry *275*, 16202-16212.

Mattick, J.S. (2005). The functional genomics of noncoding RNA. Science *309*, 1527-1528.

Mavrich, T.N., Ioshikhes, I.P., Venters, B.J., Jiang, C., Tomsho, L.P., Qi, J., Schuster, S.C., Albert, I., and Pugh, B.F. (2008a). A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. Genome Res *18*, 1073-1083.

Mavrich, T.N., Jiang, C., Ioshikhes, I.P., Li, X., Venters, B.J., Zanton, S.J., Tomsho, L.P., Qi, J., Glaser, R.L., Schuster, S.C.*, et al.* (2008b). Nucleosome organization in the Drosophila genome. Nature *453*, 358-362.

Meissner, A., Mikkelsen, T.S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B.E., Nusbaum, C., Jaffe, D.B.*, et al.* (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature *454*, 766-770.

Meyer, C.A., He, H.H., Brown, M., and Liu, X.S. (2011). BINOCh: binding inference from nucleosome occupancy changes. Bioinformatics *27*, 1867-1868.

Mimori, K., Ogawa, K., Okamoto, M., Sudo, T., Inoue, H., and Mori, M. (2005). Clinical significance of enhancer of zeste homolog 2 expression in colorectal cancer cases. Eur J Surg Oncol *31*, 376-380.

Mito, Y., Henikoff, J.G., and Henikoff, S. (2005). Genome-scale profiling of histone H3.3 replacement patterns. Nat Genet *37*, 1090-1097.

Mohan, M., Bartkuhn, M., Herold, M., Philippen, A., Heinl, N., Bardenhagen, I., Leers, J., White, R.A., Renkawitz-Pohl, R., Saumweber, H.*, et al.* (2007). The Drosophila insulator proteins CTCF and CP190 link enhancer blocking to body patterning. EMBO J *26*, 4203-4214.

Mohd-Sarip, A., Cleard, F., Mishra, R.K., Karch, F., and Verrijzer, C.P. (2005). Synergistic recognition of an epigenetic DNA element by Pleiohomeotic and a Polycomb core complex. Genes Dev *19*, 1755-1760.

Mohn, F., Weber, M., Rebhan, M., Roloff, T.C., Richter, J., Stadler, M.B., Bibel, M., and Schubeler, D. (2008). Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. Mol Cell *30*, 755-766.

Muller, J., and Verrijzer, P. (2009). Biochemical mechanisms of gene regulation by polycomb group protein complexes. Curr Opin Genet Dev *19*, 150-158.

Negre, N., Hennetin, J., Sun, L.V., Lavrov, S., Bellis, M., White, K.P., and Cavalli, G. (2006). Chromosomal distribution of PcG proteins during Drosophila development. PLoS Biol *4*, e170.

Nekrasov, M., Wild, B., and Muller, J. (2005). Nucleosome binding and histone methyltransferase activity of Drosophila PRC2. EMBO Rep *6*, 348-353.

Neph, S., Vierstra, J., Stergachis, A.B., Reynolds, A.P., Haugen, E., Vernot, B., Thurman, R.E., John, S., Sandstrom, R., Johnson, A.K.*, et al.* (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. Nature *489*, 83-90.

Nordstrom, W., and Abrams, J.M. (2000). Guardian ancestry: fly p53 and damage-inducible apoptosis. Cell Death and Differentiation *7*, 1035-1038.

Oki, M., Valenzuela, L., Chiba, T., Ito, T., and Kamakaka, R.T. (2004). Barrier proteins remodel and modify chromatin to restrict silenced domains. Mol Cell Biol *24*, 1956-1967.

Ooi, S.K.T., Qiu, C., Bernstein, E., Li, K.Q., Jia, D., Yang, Z., Erdjument-Bromage, H., Tempst, P., Lin, S.P., Allis, C.D.*, et al.* (2007). DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. Nature *448*, 714-U713.

Orlando, V., and Paro, R. (1995). Chromatin multiprotein complexes involved in the maintenance of transcription patterns. Curr Opin Genet Dev *5*, 174-179.

Ozsolak, F., Poling, L.L., Wang, Z., Liu, H., Liu, X.S., Roeder, R.G., Zhang, X., Song, J.S., and Fisher, D.E. (2008a). Chromatin structure analyses identify miRNA promoters. Genes Dev *22*, 3172-3183.

Ozsolak, F., Poling, L.L., Wang, Z.X., Liu, H., Liu, X.S., Roeder, R.G., Zhang, X.M., Song, J.S., and Fisher, D.E. (2008b). Chromatin structure analyses identify miRNA promoters. Gene Dev *22*, 3172-3183.

Pai, C.Y., Lei, E.P., Ghosh, D., and Corces, V.G. (2004). The centrosomal protein CP190 is a component of the gypsy chromatin insulator. Mol Cell *16*, 737-748.

Pardal, R., Clarke, M.F., and Morrison, S.J. (2003). Applying the principles of stem-cell biology to cancer. Nat Rev Cancer *3*, 895-902.

Pauler, F.M., Sloane, M.A., Huang, R., Regha, K., Koerner, M.V., Tamir, I., Sommer, A., Aszodi, A., Jenuwein, T., and Barlow, D.P. (2009). H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome. Genome Res *19*, 221-233.

Peng, Z.H. (2005). Current status of gendicine in China: Recombinant human Ad-p53 agent for treatment of cancers. Human Gene Therapy *16*, 1016-1027.

Pepke, S., Wold, B., and Mortazavi, A. (2009). Computation for ChIP-seq and RNA-seq studies. Nat Methods *6*, S22-32.

Perry, A.S., Watson, R.W., Lawler, M., and Hollywood, D. (2010). The epigenome as a therapeutic target in prostate cancer. Nat Rev Urol *7*, 668-680.

Petruk, S., Sedkov, Y., Riley, K.M., Hodgson, J., Schweisguth, F., Hirose, S., Jaynes, J.B., Brock, H.W., and Mazo, A. (2006). Transcription of bxd noncoding RNAs promoted by trithorax represses Ubx in cis by transcriptional interference. Cell *127*, 1209-1221.

Pirrotta, V. (1998). Polycombing the genome: PcG, trxG, and chromatin silencing. Cell *93*, 333-336.

Pomerantz, J., Schreiber-Agus, N., Liegeois, N.J., Silverman, A., Alland, L., Chin, L., Potes, J., Chen, K., Orlow, I., Lee, H.W*., et al.* (1998). The Ink4a tumor suppressor gene product, p19(Arf), interacts with MDM2 and neutralizes MDM2's inhibition of p53. Cell *92*, 713-723.

Qin, Z.S., Yu, J., Shen, J., Maher, C.A., Hu, M., Kalyana-Sundaram, S., and Chinnaiyan, A.M. (2010). HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. BMC Bioinformatics *11*, 369.

Quelle, D.E., Zindy, F., Ashmun, R.A., and Sherr, C.J. (1995). Alternative Reading Frames of the Ink4a Tumor-Suppressor Gene Encode 2 Unrelated Proteins Capable of Inducing Cell-Cycle Arrest. Cell *83*, 993-1000.

Raab, J.R., Chiu, J., Zhu, J., Katzman, S., Kurukuti, S., Wade, P.A., Haussler, D., and Kamakaka, R.T. (2012). Human tRNA genes function as chromatin insulators. EMBO J *31*, 330-350.

Raab, J.R., and Kamakaka, R.T. (2010). Insulators and promoters: closer than we think. Nat Rev Genet *11*, 439-446.

Ringrose, L., Rehmsmeier, M., Dura, J.M., and Paro, R. (2003). Genome-wide prediction of Polycomb/Trithorax response elements in Drosophila melanogaster. Dev Cell *5*, 759-771.

Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E*., et al.* (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell *129*, 1311-1323.

Rodriguez-Paredes, M., and Esteller, M. (2011). Cancer epigenetics reaches mainstream oncology. Nat Med *17*, 330-339.

Roseman, R.R., Pirrotta, V., and Geyer, P.K. (1993). The su(Hw) protein insulates expression of the Drosophila melanogaster white gene from chromosomal position-effects. Embo J *12*, 435-442.

Roth, J.A., Nguyen, D., Lawrence, D.D., Kemp, B.L., Carrasco, C.H., Ferson, D.Z., Hong, W.K., Komaki, R., Lee, J.J., Nesbitt, J.C*., et al.* (1996). Retrovirus-mediated wild-type p53 gene transfer to tumors of patients with lung cancer. Nature Medicine *2*, 985-991.

Roy, S., Ernst, J., Kharchenko, P.V., Kheradpour, P., Negre, N., Eaton, M.L., Landolin, J.M., Bristow, C.A., Ma, L., Lin, M.F*., et al.* (2010). Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science *330*, 1787-1797.

Rozowsky, J., Euskirchen, G., Auerbach, R.K., Zhang, Z.D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M., and Gerstein, M.B. (2009). PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol *27*, 66-75.

Sakai, A., Schwartz, B.E., Goldstein, S., and Ahmad, K. (2009). Transcriptional and developmental functions of the H3.3 histone variant in Drosophila. Curr Biol *19*, 1816-1820.

Sanchez-Elsner, T., Gou, D., Kremmer, E., and Sauer, F. (2006). Noncoding RNAs of trithorax response elements recruit Drosophila Ash1 to Ultrabithorax. Science *311*, 1118-1123.

Santos-Rosa, H., Schneider, R., Bannister, A.J., Sherriff, J., Bernstein, B.E., Emre, N.C., Schreiber, S.L., Mellor, J., and Kouzarides, T. (2002). Active genes are tri-methylated at K4 of histone H3. Nature *419*, 407-411.

Sanyal, A., Lajoie, B.R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. Nature *489*, 109-113.

Satterlee, J.S., Schubeler, D., and Ng, H.H. (2010). Tackling the epigenome: challenges and opportunities for collaboration. Nat Biotechnol *28*, 1039-1044.

Schones, D.E., and Zhao, K. (2008). Genome-wide approaches to studying chromatin modifications. Nat Rev Genet *9*, 179-191.

Schwartz, Y.B., Kahn, T.G., Nix, D.A., Li, X.Y., Bourgon, R., Biggin, M., and Pirrotta, V. (2006). Genome-wide analysis of Polycomb targets in Drosophila melanogaster. Nat Genet *38*, 700-705.

Schwartz, Y.B., Linder-Basso, D., Kharchenko, P.V., Tolstorukov, M.Y., Kim, M., Li, H.B., Gorchakov, A.A., Minoda, A., Shanower, G., Alekseyenko, A.A.*, et al.* (2012). Nature and function of insulator protein binding sites in the Drosophila genome. Genome Res.

Schwartz, Y.B., and Pirrotta, V. (2007). Polycomb silencing mechanisms and the management of genomic programmes. Nat Rev Genet *8*, 9-22.

Schweinsberg, S., Hagstrom, K., Gohl, D., Schedl, P., Kumar, R.P., Mishra, R., and Karch, F. (2004). The enhancer-blocking activity of the Fab-7 boundary from the Drosophila bithorax complex requires GAGA-factor-binding sites. Genetics *168*, 1371-1384.

Seligson, D.B., Horvath, S., McBrian, M.A., Mah, V., Yu, H., Tze, S., Wang, Q., Chia, D., Goodglick, L., and Kurdistani, S.K. (2009). Global Levels of Histone Modifications Predict Prognosis in Different Cancers. Am J Pathol *174*, 1619-1628.

Senzer, N., and Nemunaitis, J. (2009). A review of contusugene ladenovec (Advexin) p53 therapy. Current Opinion in Molecular Therapeutics *11*, 54-61.

Shao, Z., Raible, F., Mollaaghababa, R., Guyon, J.R., Wu, C.T., Bender, W., and Kingston, R.E. (1999). Stabilization of chromatin structure by PRC1, a Polycomb complex. Cell *98*, 37-46.

Sharov, A.A., and Ko, M.S. (2009). Exhaustive search for over-represented DNA sequence motifs with CisFinder. DNA Res *16*, 261-273.

Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V.V.*, et al.* (2012). A map of the cis-regulatory sequences in the mouse genome. Nature *488*, 116-120.

Sinkkonen, L., Hugenschmidt, T., Berninger, P., Gaidatzis, D., Mohn, F., Artus-Revel, C.G., Zavolan, M., Svoboda, P., and Filipowicz, W. (2008). MicroRNAs control de novo DNA methylation through regulation of transcriptional repressors in mouse embryonic stem cells. Nat Struct Mol Biol *15*, 259-267.

Smith, C.L., and Peterson, C.L. (2005). ATP-dependent chromatin remodeling. Curr Top Dev Biol *65*, 115-148.

Song, Q., and Smith, A.D. (2011). Identifying dispersed epigenomic domains from ChIP-Seq data. Bioinformatics *27*, 870-871.

Soto-Reyes, E., and Recillas-Targa, F. (2010). Epigenetic regulation of the human p53 gene promoter by the CTCF transcription factor in transformed cell lines. Oncogene *29*, 2217-2227.

Sparmann, A., and van Lohuizen, M. (2006). Polycomb silencers control cell fate, development and cancer. Nat Rev Cancer *6*, 846-856.

Stott, F.J., Bates, S., James, M.C., McConnell, B.B., Starborg, M., Brookes, S., Palmero, I., Ryan, K., Hara, E., Vousden, K.H.*, et al.* (1998). The alternative product from the human CDKN2A locus, p14(ARF), participates in a regulatory feedback loop with p53 and MDM2. Embo Journal *17*, 5001-5014.

Strahl, B.D., and Allis, C.D. (2000). The language of covalent histone modifications. Nature *403*, 41-45.

Struhl, K. (2007). Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nat Struct Mol Biol *14*, 103-105.

Suter, B., Schnappauf, G., and Thoma, F. (2000). Poly(dA.dT) sequences exist as rigid DNA structures in nucleosome-free yeast promoters in vivo. Nucleic Acids Res *28*, 4083-4089.

Tan, Y., Yamada-Mabuchi, M., Arya, R., St Pierre, S., Tang, W., Tosa, M., Brachmann, C., and White, K. (2011). Coordinated expression of cell death genes regulates neuroblast apoptosis. Development *138*, 2197-2206.

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

Tolhuis, B., de Wit, E., Muijrers, I., Teunissen, H., Talhout, W., van Steensel, B., and van Lohuizen, M. (2006). Genome-wide profiling of PRC1 and PRC2 Polycomb chromatin binding in Drosophila melanogaster. Nat Genet *38*, 694-699.

Tolstorukov, M.Y., Goldman, J.A., Gilbert, C., Ogryzko, V., Kingston, R.E., and Park, P.J. (2012). Histone Variant H2A.Bbd Is Associated with Active Transcription and mRNA Processing in Human Cells. Mol Cell *47*, 596-607.

Tucker, K.L. (2001). Methylated cytosine and the brain: a new base for neuroscience. Neuron *30*, 649-652.

van Haaften, G., Dalgliesh, G.L., Davies, H., Chen, L., Bignell, G., Greenman, C., Edkins, S., Hardy, C., O'Meara, S., Teague, J.*, et al.* (2009). Somatic mutations of the histone H3K27 demethylase gene UTX in human cancer. Nat Genet *41*, 521-523.

van Kemenade, F.J., Raaphorst, F.M., Blokzijl, T., Fieret, E., Hamer, K.M., Satijn, D.P., Otte, A.P., and Meijer, C.J. (2001). Coexpression of BMI-1 and EZH2 polycomb-group proteins is associated with cycling cells and degree of malignancy in B-cell non-Hodgkin lymphoma. Blood *97*, 3896-3901.

Varambally, S., Dhanasekaran, S.M., Zhou, M., Barrette, T.R., Kumar-Sinha, C., Sanda, M.G., Ghosh, D., Pienta, K.J., Sewalt, R.G., Otte, A.P.*, et al.* (2002). The polycomb group protein EZH2 is involved in progression of prostate cancer. Nature *419*, 624-629.

Vassilev, L.T. (2007). MDM2 inhibitors for cancer therapy. Trends in Molecular Medicine *13*, 23-31.

Vassilev, L.T., Vu, B.T., Graves, B., Carvajal, D., Podlaski, F., Filipovic, Z., Kong, N., Kammlott, U., Lukacs, C., Klein, C.*, et al.* (2004). In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. Science *303*, 844-848.

Vignali, M., Hassan, A.H., Neely, K.E., and Workman, J.L. (2000). ATP-dependent chromatin-remodeling complexes. Mol Cell Biol *20*, 1899-1910.

Visser, H.P., Gunster, M.J., Kluin-Nelemans, H.C., Manders, E.M., Raaphorst, F.M., Meijer, C.J., Willemze, R., and Otte, A.P. (2001). The Polycomb group protein EZH2 is upregulated in proliferating, cultured human mantle cell lymphoma. Br J Haematol *112*, 950-958.

Wang, J., Lunyak, V.V., and Jordan, I.K. (2012). Genome-wide prediction and analysis of human chromatin boundary elements. Nucleic Acids Res *40*, 511-529.

Weber, M., Hellmann, I., Stadler, M.B., Ramos, L., Paabo, S., Rebhan, M., and Schubeler, D. (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat Genet *39*, 457-466.

West, A.G., Huang, S., Gaszner, M., Litt, M.D., and Felsenfeld, G. (2004). Recruitment of histone modifications by USF proteins at a vertebrate barrier element. Mol Cell *16*, 453-463.

Wilbanks, E.G., and Facciotti, M.T. (2010). Evaluation of algorithm performance in ChIP-seq peak detection. Plos One *5*, e11471.

Wingender, E., Dietze, P., Karas, H., and Knuppel, R. (1996). TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res *24*, 238-241.

Witcher, M., and Emerson, B.M. (2009). Epigenetic silencing of the p16(INK4a) tumor suppressor is associated with loss of CTCF binding and a chromatin boundary. Mol Cell *34*, 271-284.

Won, K.J., Chepelev, I., Ren, B., and Wang, W. (2008). Prediction of regulatory elements in mammalian genomes using chromatin signatures. BMC Bioinformatics *9*, 547.

Xu, H., Handoko, L., Wei, X., Ye, C., Sheng, J., Wei, C.L., Lin, F., and Sung, W.K. (2010). A signal-noise model for significance analysis of ChIP-seq with negative control. Bioinformatics *26*, 1199-1204.

Xu, H., Wei, C.L., Lin, F., and Sung, W.K. (2008). An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. Bioinformatics *24*, 2344-2349.

Yang, P.K., and Kuroda, M.I. (2007). Noncoding RNAs and intranuclear positioning in monoallelic gene expression. Cell *128*, 777-786.

Young, K.H., Leroy, K., Moller, M.B., Colleoni, G.W.B., Sanchez-Beato, M., Kerbauy, F.R., Haioun, C., Eickhoff, J.C., Young, A.H., Gaulard, P.*, et al.* (2008). Structural profiles of TP53 gene mutations predict clinical outcome in diffuse large B-cell lymphoma: an international collaborative study. Blood *112*, 3088-3098.

Zang, C., Schones, D.E., Zeng, C., Cui, K., Zhao, K., and Peng, W. (2009). A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics *25*, 1952-1958.

Zhang, C.C., and Bienz, M. (1992). Segmental determination in Drosophila conferred by hunchback (hb), a repressor of the homeotic gene Ultrabithorax (Ubx). Proc Natl Acad Sci U S A *89*, 7511-7515.

Zhang, W., Deng, H., Bao, X., Lerach, S., Girton, J., Johansen, J., and Johansen, K.M. (2006). The JIL-1 histone H3S10 kinase regulates dimethyl H3K9 modifications and heterochromatic spreading in Drosophila. Development *133*, 229-235.

Zhang, Y., LeRoy, G., Seelig, H.P., Lane, W.S., and Reinberg, D. (1998a). The dermatomyositis-specific autoantigen Mi2 is a component of a complex containing histone deacetylase and nucleosome remodeling activities. Cell *95*, 279-289.

Zhang, Y., Lin, N., Carroll, P.M., Chan, G., Guan, B., Xiao, H., Yao, B., Wu, S.S., and Zhou, L. (2008a). Epigenetic blocking of an enhancer region controls irradiation-induced proapoptotic gene expression in Drosophila embryos. Dev Cell *14*, 481-493.

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

Zhang, Y.P., Xiong, Y., and Yarbrough, W.G. (1998b). ARF promotes MDM2 degradation and stabilizes p53: ARF-INK4a locus deletion impairs both the Rb and p53 tumor suppression pathways. Cell *92*, 725-734.

Zindy, F., Williams, R.T., Baudino, T.A., Rehg, J.E., Skapek, S.X., Cleveland, J.L., Roussel, M.F., and Sherr, C.J. (2003). Arf tumor suppressor promoter monitors latent oncogenic signals in vivo. P Natl Acad Sci USA *100*, 15930-15935.

BIOGRAPHICAL SKETCH

Guangyao Li was born in Zhengzhou, Henan province, P.R. China. He is currently a Ph.D. candidate at the University of Florida, majoring in Genetics and Genomics. He received the Bachelor of Science degree in Mathematics and Applied Mathematics from Sun Yat-sen University in Guangzhou in 2008, and immediately started the graduate study in Bioinformatics at the same university following his interest in applying mathematics and statistics to biological sciences. In 2009, he was admitted to the Ph.D. program in Genetics and Genomics at the University of Florida. In 2010, he joined the laboratory of Dr. Lei Zhou in the Genetics Institute and Department of Molecular Genetics and Microbiology, with the dissertation topic focusing on revealing chromatin structure and epigenetic regulation using bioinformatics approaches. He passed the qualifying exam and became a Ph.D. candidate in September 2011. In May 2013, he received the Master of Statistics degree from Department of Statistics at the University of Florida. He finished his Ph.D. dissertation and received the Ph.D. degree in Genetics and Genomics in December 2013.