Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Kevin Thomas Van Bortle **Date**

The Combinatorial Role of Architectural Proteins in Insulator Function and Dynamics during the Ecdysone Response in *Drosophila melanogaster*

By

Kevin Thomas Van Bortle
Doctor of Philosophy

Graduate Division of Biological and Biomedical Science
Biochemistry, Cell, and Developmental Biology

_____

Victor G. Corces, Ph.D.
Advisor

_____

William G. Kelly, Ph.D.
Committee Member

_____

Kenneth Moberg, Ph.D.
Committee Member

_____

Maureen Powers, Ph.D.
Committee Member

_____

Daniel Reines, Ph.D.
Committee Member

Accepted:

_____

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

_____

Date

The Combinatorial Role of Architectural Proteins in Insulator Function and Dynamics during the Ecdysone Response in *Drosophila melanogaster*

By

Kevin Thomas Van Bortle
B. S., University of Rochester, 2009

Advisor: Victor G. Corces, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Graduate Division of Biological and Biomedical Science
Biochemistry, Cell, and Developmental Biology
2014

Abstract

The Combinatorial Role of Architectural Proteins in Insulator Function and Dynamics during the Ecdysone Response in *Drosophila melanogaster*

By Kevin Thomas Van Bortle

Recently developed genomic strategies for assaying chromosome architecture have significantly improved our ability to investigate the nature of genome organization and the players involved. In particular, recent high-throughput chromosome conformation capture studies provide evidence that eukaryotic genomes are organized into tissue-invariant, sub-megabase sized structures called Topologically Associating Domains (TADs). The conserved nature of TADs across diverse cell types suggests a pre-defined, bottom-up pattern of chromosome organization, yet how these structures are established and maintained during development remain important questions. The borders of TADs are highly enriched for architectural proteins, previously characterized for their role in insulator function. However, a majority of architectural protein binding sites (APBSs) localize within topological domains, suggesting sites associated with TAD borders represent a functionally different subclass of these regulatory elements. By mapping the genome-wide target sites for several *Drosophila* architectural proteins, including previously uncharacterized profiles for Pol III transcription factor TFIIIC and SMC-containing condensin complexes, I uncovered an extensive pattern of colocalization in which architectural proteins establish dense, high occupancy clusters at the borders of topological domains. Enhancer-blocking activity and TAD border strength scale with the occupancy level of APBSs, suggesting co-binding by multiple architectural proteins underlies the functional potential of these loci. Parallel analyses in mouse and human stem cells further suggest that clustering of architectural proteins is a general feature of genome organization, and that conserved APBSs may underlie the tissue-invariant nature of TADs. These results provide a novel, integrative model for understanding the role of architectural proteins in insulator function and topological domain organization. Profiling of architectural protein binding dynamics in response to signaling events, including 20-hydroxyecdysone (20HE) and TGF-β, further reveals a role for these proteins in defining the transcriptional response to extracellular stimuli. Finally, I present evidence that transcriptomic and proteomic codon usage is altered in response to 20HE in a manner that reflects the differentiation status of the cell. tRNA abundance predicts preferential codon incorporation in proteins, and increasing and decreasing tRNA isoacceptors lead to increased and decreased codon usage, respectively, in proteins, together suggesting tRNA levels play an important role in regulating the translational output of a cell.

The Combinatorial Role of Architectural Proteins in Insulator Function and Dynamics during the Ecdysone Response in *Drosophila melanogaster*

By

Kevin Thomas Van Bortle
B. S., University of Rochester, 2009

Advisor: Victor G. Corces, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Graduate Division of Biological and Biomedical Science
Biochemistry, Cell, and Developmental Biology
2014

ACKNOWLEDGEMENTS

The work presented in this dissertation was made possible by an endless amount of help from my incredibly talented labmates in the Corces lab, both past and present. In particular, I owe sincere thanks to my advisor, Victor Corces, for his patience and dedication to my development as a career scientist, for his continual mentorship and advise, and for his inspiration as a successful academic P.I. I would also like to thank Ashley Wood, Edward Ramos, Chintong Ong, Naomi Takenaka, Wendy Kellner, and Mike Nichols for the hundreds of hours they've spent working and discussing scientific ideas with me, and for their friendship. I thank my thesis committee, Bill Kelly, Ken Moberg, Maureen Powers, and Daniel Reines for their countless constructive ideas and suggestions these past five years.

I am thankful for the never ending support of my family and friends, and for the unconditional love and encouragement from Brenda and Scully, who've made my graduate years both fun and happy.

Table of Contents

CHAPTER 2: *DROSOPHILA* CTCF TANDEMLY ALIGNS WITH OTHER INSULATOR PROTEINS AT THE BORDERS OF H3K27ME3 DOMAINS

CHAPTER 3: INSULATOR FUNCTION AND TOPOLOGICAL DOMAIN BORDER STRENGTH SCALE WITH ARCHITECTURAL PROTEIN OCCUPANCY

List of Tables

CHAPTER 1: INTRODUCTION

List of Figures

CHAPTER 1: INTRODUCTION

CHAPTER 2: *DROSOPHILA* CTCF TANDEMLY ALIGNS WITH OTHER INSULATOR PROTEINS AT THE BORDERS OF H3K27ME3 DOMAINS

CHAPTER 3: INSULATOR FUNCTION AND TOPOLOGICAL DOMAIN BORDER STRENGTH SCALE WITH ARCHITECTURAL PROTEIN OCCUPANCY

CHAPTER 7: DISCUSSION

CHAPTER 1:

INTRODUCTION

Some introductory material and figures presented in this chapter are published in:

Van Bortle K, Corces VG. Annu Rev Cell Dev Biol. 2012. Nuclear organization and genome function. 28:163-87.

Eukaryotic genomes are packaged into higher-order chromatin structures and ultimately organized in a manner that functionally relates to gene expression. Understanding the mechanisms and molecular players involved in genome organization is therefore essential to fully comprehend the fundamental relationship between nuclear organization and genome function. Insulators are multiprotein DNA complexes proposed to underlie nuclear architecture on the basis of their ability to facilitate long-range physical interactions, to interact with nuclear substructures, and to cluster into nuclear foci termed insulator bodies. However, spatiotemporal expression and repression of genes pertinent to development and cell-type specification involve the function of additional regulatory elements, such as enhancers and Polycomb Response Elements (PREs), which suggests additional factors may also play a role in genome organization. The recent development of unbiased, high-throughput methods for mapping protein binding sites and genome-wide interactions has allowed an unprecedented look into the inner workings of genome biology, and new studies have provided valuable insights into the roles of insulators and chromatin structure in nuclear organization. In this introduction, I highlight the relationship between nuclear organization and genome function and emphasize the dynamic interplay among chromatin insulators, transcription activation, and Polycomb mediated repression in creating and/or maintaining a 3D arrangement of the chromatin that is conducive to the establishment of patterns of gene expression required for cell-type specification.

INTERPHASE CHROMOSOME ORGANIZATION

Eukaryotic cells are tasked with packaging the genome several thousandfold into the confines of the cell nucleus while maintaining gene accessibility and chromatin structure that accommodate highly dynamic processes, including gene transcription, replication, and DNA repair. Interphase chromosomes are organized into discrete territories that are distributed nonrandomly with respect to the nucleus, and whose placement can influence the potential for *trans* interactions and dictate whether a genomic locus is in an active or repressive nuclear environment [1, 2] The nucleus also harbors several discrete subnuclear foci, termed nuclear bodies, which are dynamically regulated structures that facilitate greater efficiency of many nuclear processes [3]. For example, active genes can relocate out of resident chromosome territories [4, 5], and cluster into subnuclear foci termed transcription factories for gene expression [6]. Gene silencing is also accomplished through recruitment to repressive nuclear structures, such as Polycomb bodies, and most biological processes are similarly compartmentalized into analogous nuclear bodies, which indicates an important relationship between nuclear organization and genome function.

Differentiation, Replication, and Genome Stability

The nonrandom order and significance of genome organization are perhaps best highlighted by its relationship to cellular differentiation, replication, and genome stability. The pathway from pluripotency to differentiated tissues is accompanied by changes in epigenomic landscapes, genome compaction, and some degree of chromosomal reorganization [7-10]. Developmental genes are differentially targeted to transcriptionally active or transcriptionally repressive nuclear substructures, and

differentiation is associated with restructuring of interactions between chromatin and the nuclear lamina [11]. An increase in genome compaction may accommodate the organization of nuclear foci associated with transcription, DNA repair, replication, and splicing while restricting the complexity of genome function by concealing irrelevant transcription factor binding sites [12]. Dynamic arrangement of interphase chromosomes also plays a critical role in organizing DNA replication into discrete subnuclear compartments and in maintaining genome integrity [13]. DNA damage gives rise to the accumulation of repair and DNA damage checkpoint proteins concomitant to increased chromatin accessibility [14, 15], and studies in both yeast and human cells demonstrate an important relationship between nuclear organization, DNA repeat stability, and telomere protection (for reviews see [16, 17]). Characterizing the mechanisms involved in 3D genome organization is therefore essential for understanding the apparent fundamental relationship between nuclear organization and cellular function.

Genomic Strategies

Microscopy studies have been invaluable in revealing insights into the distribution and organization of chromosomes in individual cells and their relationship with gene regulation. However, to break down the 3D organization of chromosomes and the relationship between nuclear organization and underlying chromatin proteins, new techniques were required that exceeded the resolution and throughput limits imposed by traditional light microscopy (Table 1). The advent of the chromosome conformation capture (3C) technique described by [18] marked the first approach to effectively map physical chromosomal interactions across the genome. Although 3C has been useful in

Method	Description
ChIP-PCR	Identify binding status of chromatin-associated protein at selected genomic loci; PCR against genomic region of interest in DNA fragments obtained by chromatin immunoprecipitation (ChIP) against protein of interest
ChIP-chip	Identify the genome-wide binding profile of a chromatin-associated protein; microarray hybridization of protein-associated DNA fragments enriched for by ChIP
ChIP-seq	Identify the genome-wide binding profile for chromatin-associated protein; high-throughput sequencing of protein-associated DNA regions enriched for by ChIP
3C	Measure the interaction frequency between two selected genomic loci; quantitative PCR against ligated restriction fragments of interest
4C	Map physical interactions between a selected locus and the entire genome; microarray hybridization or high-throughput sequencing of interacting sequences captured by inverse PCR
5C	Map physical interactions between genomic loci within a targeted locus; ligation-mediated amplification and quantitation of interacting DNA fragments
Hi-C	Measure genome-wide interaction frequencies between all genomic loci; high-throughput sequencing of interacting sequences obtained by purification of biotin-marked ligation junctions
ChIA-PET	Map genome-wide interactions bound by chromatin-associated protein; high-throughput sequencing of ligated fragments enriched for by ChIP against protein of interest

Table 1. **Genomic tools for assaying chromatin occupancy, structure, and organization**. ChIP: chromatin immunoprecipitation; ChIP-chip: chromatin immunoprecipitation followed by microarray analysis; ChIP-seq: chromatin immunoprecipitation followed by high-throughput sequencing; 3C: Chromosome conformation capture; 4C: Circular chromosome conformation capture; 5C: Carbon copy chromosome conformation capture; Hi-C: High-throughput chromosome conformation capture; ChIA-PET: chromatin interaction analysis by paired end tag sequencing.

identifying locus-specific interactions between regulatory elements and target genes[18, 19], derivations of the 3C technique have been introduced to extend the approach in an unbiased, high-throughput, genome-wide fashion. For example, the Hi-C method integrates an extended 3C protocol with massively parallel DNA sequencing, thereby capturing all genome-wide interactions at a resolution limited by the depth of sequencing [20]. Initial Hi-C analyses in a human lymphoblastoid cell line provided valuable insight into chromatin organization, supporting the fractal globule model in which chromosomes self-organize into a hierarchy of crumples, or series of globules governed by topological constraints [20-22]. Subsequent computational modeling supports the existence of chromosome territories and transcriptional foci and revealed new insights into the relationship between chromatin organization and CCCTC-binding factor (CTCF), a highly conserved architectural protein, as well as the nuclear lamina [23]. However, further conclusions about the chromosome topology and nuclear organization of chromatin in human cells will require higher resolution, which is likely to be obtained in the near future with greater sequencing depth.

Mediators of Nuclear Organization

Determination of how interphase chromosomes are anchored within the nuclear space and which proteins mediate structural arrangements conducive to gene regulation and locus plasticity remains a critical hurdle to understanding the mechanisms that regulate genome function. Fortunately, genome-wide mapping of chromatin-associated proteins has increased at an extraordinary pace during the past few years thanks to the ENCyclopedia of DNA Elements (ENCODE) Projects ([24, 25]) and the increasingly affordable option of high-throughput sequencing. Analyses combining the 3D

organization of interphase chromosomes - with genome-wide binding profiles of chromatin-associated factors - may ultimately establish which proteins functionally mediate nuclear organization. Information from microscopy and biochemical studies, combined with recent genome-wide mapping, has implicated multiple factors, including chromatin insulators, Polycomb repressive complexes, and noncoding RNAs as having roles in domain formation and chromatin organization..

CHROMATIN INSULATORS

Chromatin insulators originally were defined as regulatory elements that recruit proteins to establish boundaries between adjacent chromatin domains. Insulators were also shown to block the communication between enhancers and nearby promoters in an orientation-dependent manner, which led to intuitive models in which insulators limit the promiscuity of enhancers. However, further characterization of these sequences from yeast to humans has revealed that insulators colocalize to subnuclear foci called insulator bodies as well as transcription factories, and can mediate both long-range intra- and inter-chromosomal interactions, as revealed by Hi-C computational modeling in both yeast and humans. These findings suggest that insulators serve a greater purpose in chromatin organization.

Composition and Evolution

Studies of chromatin insulators in mammals have long been restricted to the highly conserved insulator protein CTCF, which until recently was the only characterized protein capable of insulator activity in humans. However, insulator activity in yeast

involves tRNA genes and the transcription factor TFIIIC, and this role now appears to be conserved in mammals [26, 27]. Concurrent research on insulators in yeast, *Drosophila*, and mammalian systems therefore suggests that these elements serve an evolutionarily conserved role in gene regulation and nuclear organization.

Insulator protein CTCF

CTCF contains a central domain composed of 11 zinc fingers and is ubiquitously expressed [28]. Early biochemical studies demonstrated >93% amino acid identity between human and chicken CTCF proteins [29], and studies in *Drosophila* later identified an orthologous CTCF factor with a similar domain structure and conserved insulator function [30], which suggests this zinc-finger protein plays a vital, highly conserved role in nuclear biology. Remarkably, CTCF primarily targets a highly similar core consensus sequence from *Drosophila* to humans, despite its ability to bind a variety of DNA sequences [31]. The proteins with which CTCF associates and the variant sequences it is able to bind have been suggested to underlie the numerous roles in which CTCF has been implicated [32, 33], such as X-chromosome inactivation, V(D)J rearrangement, and chromatin insulation. Many CTCF binding sites recruit the cohesin complex (Figure 1), which is required for functional CTCF insulator activity [34, 35]. The cohesin complex forms a ring-shaped structure and mediates cohesion between sister chromatids from S-phase until mitosis, which suggests that cohesin may specifically stabilize chromatin loops arranged by CTCF through a similar mechanism. CTCF also interacts with Yin and yang 1 (YY1), a transcription factor involved in X-chromosome inactivation [36] capable of recruiting Polycomb repressive complexes [37], and CTCF-

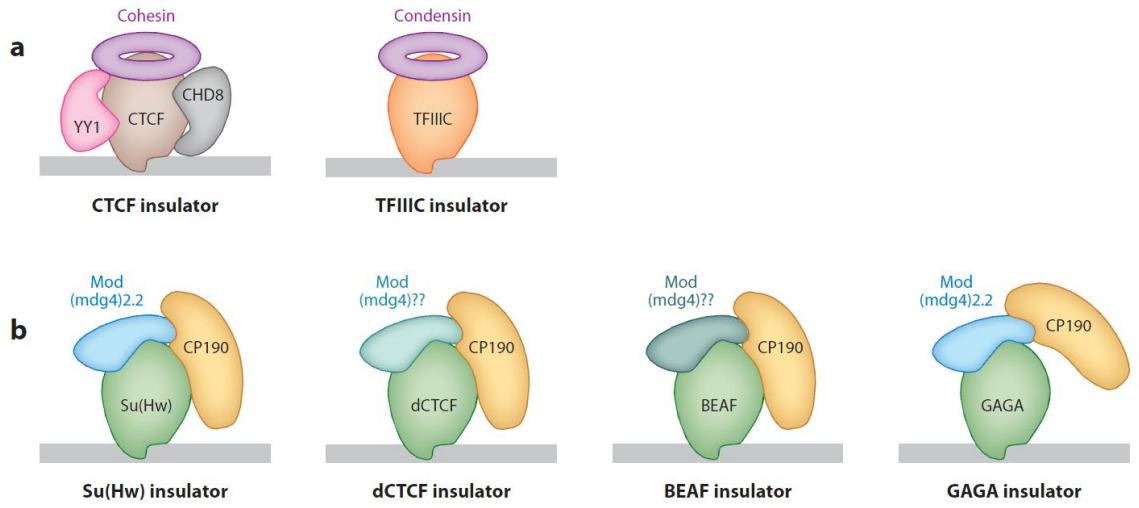


Figure 1. **Composition of vertebrate and *Drosophila* architectural protein complexes**. **(*a*)** Structure of the vertebrate CCCTC-binding factor (CTCF) and TFIIIC insulators. Indicated are factors associated with CTCF, such as cohesin, CHD8, and YY1, and with TFIIIC. **(*b*)** Each *Drosophila* insulator subclass contains a different binding protein that may define the specific function of the corresponding subclass. All insulators share the common protein Centrosomal Protein 190 kDa (CP190), although the role of this protein in the function of the GAGA insulator has not been demonstrated experimentally. In addition, all subclasses may also have one Modifier of Modg4 [Mod(mdg4)] isoform. The gypsy/ Suppressor of Hairy Wing [Su(Hw)] insulator contains Mod(mdg4)2.2. The dCTCF and Boundary Element Associated Factor (BEAF) insulators lack this isoform but contain a different variant of Mod(mdg4). GAGA has been shown to interact with Mod(mdg4)2.2

mediated insulator activity at the *H19/Igf2* requires the SNF2-like chromodomain helicase protein CHD8 (Figure 1) [38], the DEAD-box RNA helicase p68, and its associated RNA (SRA) [39].

Insulators in *D. melanogaster*

Drosophila insulator elements and their associated proteins have been particularly well characterized thanks to *in vivo* insulator activity reporter assays made easy by the genetic manipulations available in the fly model system. In addition to the *Drosophila* homolog of CTCF (dCTCF), which is highly conserved with respect to the DNA-binding zinc finger domains [30], several insulator proteins have been identified, including Suppressor of Hairy Wing [Su(Hw)], Boundary Element Associated Factor (BEAF-32), and GAGA factor (GAF) (Figure 1) [40]. Contrary to what happens in vertebrates, the *Drosophila* cohesin complex localizes to transcriptionally active genes independently of dCTCF [41]. Instead of cohesin, *Drosophila* insulator activity relies on fly-specific insulator proteins Centrosomal Protein 190 kDa (CP190) and Modifier of Modg4 [Mod(mdg4)], both of which contain BTB/POZ domains and are capable of forming stable multimers in vitro [42-44]. Genome-wide localization studies suggest that dCTCF –tandemly aligns with Su(Hw), BEAF-32, and CP190 at many sites throughout the genome, perhaps representing a unifying and perhaps synergistic role in facilitating chromosomal interactions and genome organization [45, 46]). Interestingly, insulator activity in *Drosophila* also relies on components of the RNA interference (RNAi) machinery [47, 48], though the mechanistic relationship remains poorly understood.

tDNA and TFIIIC

tRNA genes were first demonstrated to function as boundary elements flanking the repressed *HMR* locus in *Saccharomyces cerevisiae* and subsequently at the pericentromeric regions of *Schizosaccharomyces pombe* [49, 50]. The conservation of tDNA as an insulator element has been further extended by the demonstration that tRNA genes function as barrier and enhancer-blocking insulators in transgenic reporter assays in human cells [27]. Analysis of the *mat* locus in *S. pombe* revealed that a repeat of B-box elements, which are highly conserved intragenic promoter elements in tRNA genes that recruit RNA polymerase III (RNAPIII) transcription initiation factor TFIIIC, were responsible for barrier activity (Figure 1) [51]. Mutations in TFIIIC and TFIIIB, but not RNAPIII, affected insulator activity in *S. cerevisiae*, which suggests that insulator activity occurs independent of RNAPIII transcription [52]. Although TFIIIC is sufficient for gene insulation alone, insulator activity is strengthened by TFIIIB and utilizes chromatin remodelers, possibly to evict histones at tDNA insulators [53]. Interestingly, in yeast and humans, TFIIIC binds many regions devoid of RNAPIII, called Extra TFIIIC (ETC) loci [54], and these sites are associated with the cohesin complex and localize near CTCF-binding sites in mouse embryonic stem cells and humans [55, 56]. TFIIIC sites and tRNA genes also function as loading sites for the highly conserved condensin complex in yeast, which suggests insulators serve an important role in chromatin architecture during both interphase and mitosis [57].

Distribution and Chromatin Structure

Genome-wide localization studies by ChIP-chip and ChIP-seq have revealed that insulators are dispersed throughout eukaryotic genomes. CTCF is bound to thousands of independent sites in *Drosophila* [31, 45, 58] and tens of thousands of sites in human cell lines [59-61], consistent with a global role in genome function. ChIP-chip analysis further showed that CTCF colocalizes with cohesin at more than half of its sites [35, 62, 63], which suggests many CTCF sites are likely capable of functional insulator activity. Meanwhile, recent genome-wide localization of TFIIIC occupancy in mouse embryonic stem cells revealed that although all three TFIIIC subunits co-occupy <300 tRNA genes, as many as 2,200 independent TFIIIC-bound ETC loci are dispersed throughout the mouse genome [55]. Remarkably, as many as 85% of ETC loci lie within 20 kb of CTCF-binding sites, and cohesin subunits Smc1A and Smc3 were enriched at the ETC loci specifically, which suggests CTCF and TFIIIC distribution and insulator activity may be intimately associated. Evidence that suggests insulators may indeed collaborate comes from recent mapping of insulator proteins in *D. melanogaster* [46, 64]. CTCF tandemly aligns with other classes of *Drosophila* insulators, including Su(Hw) and BEAF-32, and these multi-insulator complexes then become enriched for CP190, Mod(mdg4), and additional co-factors [46]. Alignment of insulators suggest CTCF may cluster with other distinct insulator proteins to efficiently recruit essential co-factors important for establishing a robust insulator complex and perhaps capable of maintaining stable, long-range interactions. Future studies may uncover a similar relationship between CTCF and TFIIIC in humans.

The distribution of CTCF, tDNA, and aligned insulators correlate with recent mapping of physical domain borders in both *Drosophila* and mammals. For example, Cavalli and colleagues recently utilized an independent, high-throughput 3C derivative (3C-seq) to explore the 3D folding and functional organization principles of the *Drosophila* genome [65]. Their data suggest eukaryotic genomes are partitioned into physical domains that can be clustered on the basis of strong statistical association with linear epigenomic profiles. Physical domains were categorized as active, which correlate with active histone marks; null, which comprise large transcriptionally repressive regions lacking silent chromatin marks; Polycomb domains associated with histone H3 K27 trimethylation (H3K27me3) repression; or HP1/centromeric domains associated with classical heterochromatin. Physical domains are sharply demarcated, and contacts within domains abruptly decay at positions corresponding to physical domain edges. Remarkably, insulators are highly enriched at domain borders, and hierarchical clustering revealed recurrent combinations of insulators and active histone marks that are present at all combinations of physical boundaries (e.g., even between two similarly annotated physical domains, such as null--null). Analogous mapping of physical domains reveals similar partitioning of the human genome, as well as enrichment for CTCF and tRNA genes at chromatin domain borders [66, 67].

The enrichment of insulator proteins at all combinations of physical domain borders begs the question of what role insulators play in delimiting discrete chromatin domains. Earlier correlations for CTCF and tDNAs at the borders of repressed chromatin domains, typically in the form of H3K27me3[45, 60], have led many to believe insulators function

as chromatin boundaries, where they simply serve to prevent the spread of heterochromatin into flanking chromatin domains [68]. Recent compounding evidence strongly suggest that despite this correlation, endogenous insulator proteins are not important for restricting the spread of silencing chromatin. For example, RNAi depletion of CTCF in *D. melanogaster* does not lead to significant alterations in chromatin structure or gene activity [46, 64]. Similar findings are reported for depletion of other insulator proteins [46], and mutations in *Drosophila* insulator protein Su(Hw) have little consequence on several regions tested for gene activity [69]. Meanwhile, CTCF is not required for barrier activity at the well characterized β-*globin* locus[70-72], together suggesting endogenous insulator proteins are not involved in heterochromatin barrier formation. The discovery that insulators are present at all combinations of physical domain borders rather than just repressive H3K27me3 domains perhaps reinforces the notion that chromatin insulators are playing a paramount role, beyond the scope of barrier activities observed in transgenic assays that remove insulators from their natural genomic context. The ability to facilitate long-range interactions appears instead to be the defining feature of insulators, underlying their role in nuclear organization and genome function.

Long-Range Interactions

Insulators have also been defined by their ability to impede the interaction between promoters and distal enhancers in a direction-dependent manner. Meanwhile, observations at numerous genomic loci across species suggest that insulators influence chromatin structure by establishing chromatin loops through physical interactions. Concurrent models therefore proposed that insulators evolved to ensure the fidelity of enhancers and their target promoters *in vivo* by establishing chromatin loops and thereby

dictating the potential for enhancer-promoter interactions. However, accumulating data suggest that insulators facilitate long-range inter- and intrachromosomal interactions across the genome, including bridging connections between distant enhancers and target promoters, which suggests that local determination of enhancer-promoter interaction represents only part of a more significant role in chromosome organization.

The enhancer-blocking activities of CTCF have been best characterized at the chicken *β-globin* locus and the murine *H19/Igf2* imprinted locus, both of which have been extensively reviewed [73]. These loci provide ideal scenarios for studying the role of insulators in allele-specific and developmental cell-type-specific gene regulation and chromatin architecture, and 3C experiments suggest CTCF underlies chromatin contacts at both genomic loci [74, 75]. Recent application of 3C has revealed that CTCF also underlies developmental higher-order architecture at the conserved homeobox gene A (*HOXA*) locus in mouse and humans [76]. CTCF and cohesin contribute to reorganization and selective gene activation at *HOXA* by partitioning silenced genes through chromatin loop formation upon differentiation. Furthermore, pluripotency factor OCT4 antagonizes cohesin loading at the CTCF binding site, thereby demonstrating developmental regulation of insulator activity and gene expression by inhibiting chromosome loop formation. Studies in *D. melanogaster* suggest insulators have a conserved role in developmental coordination of gene expression and chromatin structure. For example, genome-wide mapping revealed that dCTCF and *Drosophila*-specific insulator proteins are regulated through DNA binding and recruitment of CP190 during the ecdysone hormone response in Kc cells [77]. In addition, 3C analysis at the ecdysone-induced

Eip75B gene further revealed a developmentally regulated dCTCF site, wherein recruitment of CP190 and enhanced chromatin looping upon ecdysone stimulation suggested alterations in *Eip57B* locus chromatin organization.

Numerous studies have focused on the role of insulators in locus-specific gene regulation and chromatin architecture, but the emerging picture of insulators in genome-wide nuclear organization requires global analyses of chromatin interactions. The first genome-wide map of CTCF-mediated functional interactions has been obtained by combining ChIP with high-throughput sequencing of enriched chromatin interactions [78]. The authors identified ~1,500 *cis* and ~330 *trans* interactions facilitated by CTCF in mouse embryonic stem cells and classified them into five categories on the basis of distinct epigenetic patterns. CTCF interactions harbor chromatin loops enriched for active or repressive chromatin signatures, which suggests that CTCF harnesses clusters of genes with coordinated expression (Figure 2). CTCF interactions also create chromatin hubs conducive to enhancer and promoter activities, in support of recent evidence suggesting that CTCF and cohesin underlie enhancer-promoter interactions at the *INFG* and *MHC-II* loci [79, 80]. The authors therefore speculate that insulators may instead facilitate cell-type specific enhancer-promoter interactions. Ultimately, Handoko et al. (2011) provide a genome-wide interrogation of CTCF interactions and reveal several modes through which CTCF functionally organizes the genome.

Genome-wide interrogations of intra- and interchromosomal interactions in yeast and humans independently demonstrate that tDNA insulators also underlie long-range

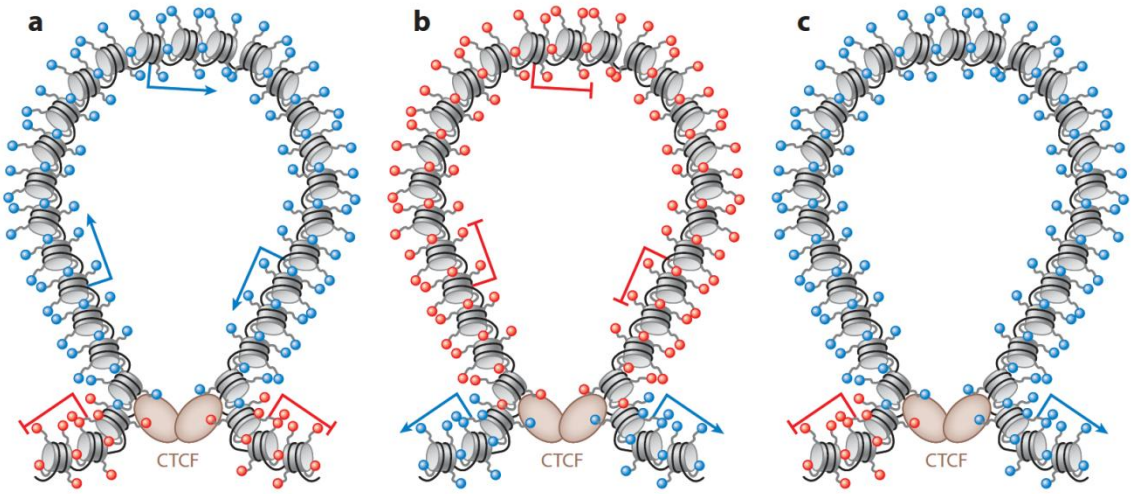


Figure 2. **Types of domains created by CCCTC-binding factor (CTCF)-mediated interactions in mouse embryonic stem cells**. Actively transcribed genes are represented by a blue arrow and silenced genes by a red inhibition line. Nucleosomes and the histone tails are represented in gray; active histone modifications are indicated as blue spheres and repressive modifications as red spheres. DNA is represented in black and CTCF as brown ovals. **(*a*)** CTCF forms a loop to separate a domain containing active histone modifications and transcribed genes from repressive marks and silenced genes. **(*b*)** CTCF forms a loop to separate a domain containing repressive histone modifications and silenced genes from active marks and transcribed genes. **(*c*)** CTCF forms a loop containing nucleosomes enriched in mono- and dimethylated H3K4, and trimethylated H3K4 at the boundaries of the loops, whereas the active transcription modification H3K36me3 and repressive H3K27me3 mark are observed outside the loops on opposite sides.

genomic interactions. Noble and colleagues recently employed a high-throughput 3C-based technique to query the 3D organization of the *S. cerevisiae* genome [81]. tRNA genes were significantly enriched for interactions with other tRNA genes, which suggests that insulator-to-insulator interactions are a conserved feature of eukaryotic genome organization. Hierarchical clustering revealed two clusters of colocalizing tRNA genes, one consistent with previously described nucleolar localization [82] and another with centromeres. Similar mapping of interactions at a tDNA insulator in humans revealed analogous long-range interactions between tDNAs as well as ETC loci, which suggests tRNA genes and TFIIIC play a conserved role in genome organization. Recent findings indicate that TFIIIC binding sites facilitate condensin binding in *S. cerevisiae* and *S. pombe* [57] and colocalize with cohesin in mammals [55], which suggests TFIIIC recruitment of condensin and cohesin complexes may underlie chromosomal interactions in yeast and humans analogous to those with CTCF.

Role in Nuclear Organization

Microscopy-based analyses of the physical and functional organization of eukaryotic nuclei have led to the identification of several discrete subnuclear organelles called nuclear bodies, which play host to a variety of nuclear processes, including transcription, splicing, processing, and epigenetic regulation [3]. Nuclear staining of insulator proteins has shown a clear propensity for insulators to concentrate into distinct nuclear foci, a feature that is conserved in yeast [51] *Drosophila* [83], and mammals [84]. Insulators also interact with and localize to nuclear substructures, including the nuclear and nucleolar peripheries, which suggests insulators tether associated chromatin to defined nuclear compartments [83, 85]. These findings, combined with knowledge about

insulator-mediated functional interactions, have led to models proposing that insulators ultimately interact to partition chromatin into structural and functional domains that are physically organized through insulator bodies (Figure 2). Although tantalizing, the functional importance and molecular underpinnings of insulator bodies remain poorly characterized.

Nuclear organization clearly involves spatial arrangement of chromosomes whose position with respect to the nuclear periphery correlates with chromatin structure and gene expression. Chromatin interactions at the nuclear periphery recently have been mapped in both *Drosophila* [86] and humans [87], revealing large, sharply defined lamina-associated domains (LADs) that correlate with low gene density and transcriptional repression. The borders of LADs in humans are enriched for CTCF [87, 88]), which suggests this protein may separate chromatin environments at the nuclear periphery. Mapping of insulators in *D. melanogaster* with respect to the nuclear lamina has also revealed a significant enrichment for the *Drosophila*-specific Su(Hw) insulator at the borders of LADs [78, 89]. In mouse embryonic stem cells, CTCF loops are depleted within LADs but enriched at LAD borders, which suggests that either CTCF orchestrates genome organization with respect to the nuclear lamina, or that interactions between DNA and lamins limit the accessibility of CTCF binding within LADs. An analogous role for TFIIIC in nuclear organization in yeast has been proposed on the basis of perinuclear staining of insulator bodies in *S. pombe* [51]. In support of this model, perinuclear localization and silencing of the *HMR* locus in *S. cerevisiae* was recently shown to rely on nuclear pore proteins that localize to a tDNA barrier insulator [90].

POLYCOMB

Genome plasticity and selective expression are essential features of multicellular development, and several early regulatory factors involved in body patterning and segmentation need to be strictly regulated to facilitate appropriate developmental decisions. Polycomb Group (PcG) proteins are evolutionarily conserved epigenetic transcriptional repressors that play an important role in establishing and maintaining cell fate by influencing the expression status of pertinent genes. PcG proteins specifically mediate the repression of numerous developmental genes through posttranslational modification of histone proteins, and recent studies demonstrate that PcG activity is involved in a broad scope of cellular processes, including differentiation, cell cycle regulation, X-inactivation, and cell signaling [91]. Microscopy studies have demonstrated that PcG proteins concentrate into nuclear foci, called Polycomb bodies, which suggests PcG proteins may also mediate the nuclear organization of their target genes. Here we review recent progress in determining the relationship between Polycomb and nuclear organization.

Composition and Evolution

The Polycomb genes were first discovered as chromatin repressors that maintain silencing of the homeotic regulatory genes throughout *Drosophila* development [92]. Further studies in the fruit fly demonstrated that PcG proteins are recruited through *cis*-regulatory elements called PREs [93], and a recently identified element at the *HoxD* locus in humans facilitating Polycomb-dependent transcriptional repression throughout cell differentiation suggests the mechanism of PcG recruitment may be conserved [94].

However, PREs are not easily broken into obvious DNA consensus sequences as described for insulator proteins, and the functional mechanism of PcG targeting remains unclear, though it appears to involve numerous players including DNA-binding proteins, histone posttranslational modification binding proteins, RNAi machinery proteins, and noncoding RNAs [95].

PcG proteins are present in two major complexes, Polycomb repressive complex 1 (PRC1) and 2 (PRC2), whose core components are largely conserved from flies to humans. The PcG family also includes several additional proteins that allow for the formation of diverse Polycomb chromatin-binding complexes with variable enzymatic activities [96]. PRC2 catalyzes H3K27 di- and trimethylation (H3K27me3, which is associated with transcriptional repression) through SET domain--containing subunit EZH1, as well as EZH2 depending on cellular context [97, 98]. PRC2 recruitment and activity is regulated by core components and ancillary subunits, such as PHF1, JARID2, and AEBP2, which stimulate PRC2 enzymatic activity [97, 99-101]. PRC1 subunits RING1B and BMI1 form a stable heterodimer capable of catalyzing H2AK119 ubiquitylation (H2AK119Ub1) [102], which likely underlies PRC1-mediated silencing [103]. There is also recent evidence that histone modification-binding proteins containing malignant brain tumor (MBT) modules contribute to Pc function. For example, L3MBTL2 interacts with and is required for PcG-mediated repression by a PRC1-like complex in human cells [104]. Interestingly, the *Drosophila* MBT protein L(3)mbt localizes specifically to chromatin insulators [105], and as we discuss below, recent evidence suggests insulator activity may play an important role in PcG repression.

Distribution and Chromatin Structure

The genome-wide localization of PcG proteins has been studied in several independent ChIP-chip and ChIP-seq experiments in both *Drosophila* and mammals. Polycomb complexes localize to putative PREs in *Drosophila* and correlate with broad repressive H3K27me3 domains that encompass genes involved in major developmental pathways [106, 107]. Most PREs are co-occupied by the major complexes PRC1 and PRC2, and the occupancy landscape of PcG proteins changes during development, consistent with its role in mediating cell fate restriction by differential gene silencing [108, 109]. Mapping of PRC1 and PRC2 complexes in mammals shows similar co-occupation of developmental pathway genes and displacement of PcG proteins during gene activation [110-112], which suggests Polycomb repression is a highly conserved feature of multicellular development. However, PcG proteins also localize to bivalent domains characterized by overlapping H3K27me3 and H3K4me3 encompassing genes poised for activation or repression upon cellular differentiation in mammals, a feature that is largely absent in fly embryos [113].

Genome-wide mapping of Trithorax group (TrxG) proteins, which catalyze H3K4 methylation and antagonize PcG repression through transcriptional activation, suggests a dynamic interplay between PcG repression and TrxG activation dependent on overlapping recruitment proteins and the relative levels of PcG- and TrxG- associated factors [113-115]. Interestingly, PRC2 also functionally associates with numerous noncoding RNAs (ncRNAs) [116]. For example, the 2.2-kb long ncRNA *HOTAIR* serves as a scaffold for both PRC2 and H3K4 demethylase LSD1 complexes, and thereby

coordinates targeting of PcG proteins to chromatin while coupling H3K27 trimethylation and H3K4 demethylation activities for epigenetic repression [117]. Mapping of the genome-wide occupancy of *HOTAIR* revealed >800 focal, transcription factor-like binding sites that co-occupied the genomic binding profiles for the PRC2 subunits EZH2, SUZ12, and H3K27me3. *HOTAIR* occupancy was maintained upon EZH2 depletion, which suggests *HOTAIR* actively binds chromatin and may underlie the nucleation of PRC2 domains [118].

Long-Range Interactions

As discussed above, PcG proteins have long been shown to concentrate into nuclear foci called Polycomb bodies, whose number and size change upon cellular differentiation [119, 120], which suggests that PcG facilitates genome-wide interactions that are further compartmentalized within the nuclear space. Accumulating data suggest that PcG proteins are indeed involved in long-range interactions essential for Polycomb repression. Combinatorial use of 3C and fluorescence in situ hybridization (FISH) showed that PRE-PRE interaction occurs between the *bxd* and *Fab-7* elements separated by ~130 kb in the *Drosophila* bithorax complex (BX-C) and that colocalization occurs specifically in embryonic tissues and cell lines in which both the *AbdA* and *Ubx* genes are corepressed [121]. Long-range associations dependent on PRC2 core component EZH2 at the mammalian *GATA-4* locus, which is silenced in undifferentiated human TERA-2 cells, have also been observed [122, 123]. Using a 3C-based approach analogous to chromatin interaction analysis with paired-end tag sequencing (ChIA-PET), Tiwari et al. show that a limited number of intra- and interchromosomal interactions in human TERA-2 cells are

dependent on EZH2, suggesting that PRC2 is involved in mediating long-range interactions in mammals.

The formation of Polycomb bodies and identification of long-range PRE interactions support a global role in genome function, but understanding the role of PcG complexes in multigene regulation and nuclear organization requires genome-wide exploration of genomic interactions. To this end, van Steensel and colleagues recently adapted chromosome conformation capture on chip (4C), a 3C derivative that determines genome-wide interactions of a given locus [124], to explore where Polycomb domains are able to interact within the *Drosophila* genome [125]. The authors demonstrate long-range interactions between Polycomb target genes and independent Polycomb and/or H3K27me3 domains in larval brain tissue, and further show that PcG interactions are topologically constrained to a single chromosome arm. Accordingly, chromosome inversion dramatically altered the interaction profiles revealed by 4C, but global gene expression patterns were relatively unchanged. Although specific interaction partners were altered, Polycomb target genes nevertheless continue to interact with independent Polycomb domains, which suggests that PcG interactions are flexibly amenable to new partners for gene repression. Genome-wide mapping of chromosomal interactions by [65] independently identified 30 significant pairs of long-range interactions between Polycomb domains, thus supporting the role of PcG proteins in mediating specific long-range associations. However, the mechanisms by which these interactions are established and maintained and whether PcG proteins are directly responsible for mediating physical interactions are not clear.

Role in Nuclear Organization

The correlation of long-range chromatin interactions with Polycomb domains provides compelling evidence in support of the possibility that PcG complexes orchestrate the nuclear organization of repressed genes, but whether PcG proteins are required for long-range interactions has recently been called into question. Co-staining for insulator protein CTCF and PcG protein Pc2 shows clear colocalization of Polycomb and insulator bodies in HeLa cells [84] , and accumulating evidence suggests that insulators, rather than PcG proteins underlie the colocalization and nuclear organization of Polycomb domains [126].

The best-studied examples of physical interactions involving Polycomb targets occur between elements within *Drosophila Hox* gene clusters that have been shown to harbor insulator activity, and genome-wide localization studies have characterized numerous dCTCF insulator sites within the Antennapedia (Antp) complex and BX-C[31]. However, Pirrotta and colleagues have recently shown that interactions between copies of the BX-C *Fab-7* or *Mcp* elements are not dependent on PREs [126]. Importantly, both *Fab-7* and *Mcp* elements have been shown to harbor enhancer-blocking insulator activities, and other studies have demonstrated that interactions depend on insulators flanking the *Mcp* and *Fab-7* PREs [126-128]. A concurrent study independently revealed that the Su(Hw) insulator is capable of dictating PRE-target interactions through chromatin looping and topological constraint in *D. melanogaster* [129], consistent with earlier findings showing that the Su(Hw) insulator facilitates PRE-PRE contacts in *trans* [130]. Interestingly, the maintenance of long-range interactions between copies of the *Fab-7* element also require components of the RNAi machinery [120], including those required for insulator activity

[47], and recent mapping of RNAi component Argonaute2 (AGO2) shows specific colocalization with insulator proteins dCTCF and CP190 throughout the BX-C [48]. Although insulators likely play an important role in previously described Polycomb interactions, Pirrotta and colleagues now speculate that PcG complexes may contribute to the stability of physical interaction [126].

In support of this theory, mutations in PcG proteins significantly reduce the level of *Antp* and *Abd-B* colocalization at Polycomb bodies [131]. Meanwhile, current models suggest PcG proteins mechanistically facilitates repression through PRC1-mediated chromatin compaction [132], and mapping of genome accessibility in *D. melanogaster* reveals that H3K27me3 domains are indeed the most inaccessible [133]. Together, these findings suggest that insulators mediate the interactions between Polycomb targets and that PRCs likely strengthen the association and repression of Polycomb domains through histone modifications and chromatin compaction. In support of this model, RNAi depletion of CTCF and other *Drosophila* insulator proteins results in loss of H3K27me3 within associated Polycomb domains [46], suggesting insulators are important for mainting appropriate chromatin architecture within these domains.

The interrogation of physical interactions between distant *Hox* loci has provided new and important insight into the regulation of long-range interactions during development. In *D. melanogaster*, the homeotic genes *Antp* and *Abd-B*, which are corepressed and colocalize to Polycomb bodies in *Drosophila* embryo heads, are recruited to separate nuclear compartments when one gene becomes activated [131]. In mammals, the *HoxD* cluster

forms a single interaction domain in murine embryonic tissues in which all genes are inactive and transitions to a bimodal state in embryonic tissues in which *HoxD* genes are differentially expressed, with active genes segregated into an active domain [134]. This suggests that developmentally regulated genes are dynamically targeted to specific nuclear subcompartments, such as Polycomb bodies and transcription factories, for transcriptional repression or activation. The apparent role of insulators in Polycomb contacts suggests that insulators are likely involved in gene localization to both transcriptionally repressive and transcriptionally permissive environments.

TRANSCRIPTION FACTORIES

The organization of transcription within eukaryotic nuclei is far more complex than traditional textbook models of polymerase recruitment and gene tracking. Instead, transcription is spatially organized into discernable nuclear structures in which multiple RNA polymerases and active genes dynamically localize into nuclear bodies termed transcription factories. The formation of transcriptionally active subcompartments presumably allows for more efficient transcription by concentrating the molecular players, reactants, and DNA substrates within a confined nuclear volume. Recent evidence suggests that transcription factories are highly conserved features of nuclear organization, that long-range chromosomal interactions are a hallmark of gene expression, and that insulators likely play an important role at transcription factories.

Composition and Evolution

Transcription is a fundamental cellular process carried out by highly conserved multisubunit RNA polymerases that share a high degree of homology in bacteria,

archaea, and eukaryotes [135]. RNA polymerase I transcription of ribosomal genes is organized into a strongly conserved and highly organized nuclear substructure called the nucleolus [136], which represents the classical example of transcriptional clustering into a factory structure. Meanwhile, most protein-coding genes are transcribed by RNA polymerase II (RNAPII), and several findings now support models proposing that RNAPII transcription occurs at analogous factories. The localization of transcription into discrete sites was initially identified by detection of nascent transcripts and by RNAPII staining, which revealed a limited number of foci unable to account for the number of active genes in human nuclei [137]. Subsequent studies further revealed that transcription factories are large and relatively immobile proteinaceous structures whose numbers vary by cell type and nuclear morphology [6] and that active genes dynamically localize to factories for expression in a transcription-dependent manner [138]. Recent genome-wide interaction assays described in this review also provide supporting evidence for the existence of transcription factories and further demonstrate that clustering of active genes is a highly conserved phenomenon. Hi-C modeling of chromosomal contacts reveals preferential clustering and interactions among actively transcribed genes and active chromatin domains in *S. pombe*, *Drosophila*, and humans [23, 65, 139].

How transcription factories are physically organized and how genes are dynamically targeted to them remains poorly understood. To this end, Cook and colleagues recently isolated transcription complexes from human nuclei and identified the proteome of RNAPI, II, and III factories by mass spectrometry [140]. Each complex was shown to harbor a characteristic set of unique proteins, though several proteins involved in DNA or RNA metabolism were shared. Whereas most proteins isolated from RNAPI complexes

overlap those characteristic of nucleoli, RNAPII factories contain general transcription factors and CTCF, consistent with the finding that CTCF underlies organization of coregulated genes in mammals. Interestingly, RNAPII factories also contain repressive histone methyltransferases, including PRC2-core component EZH2, which suggests that the transition from epigenetic repression to gene activation may not require the ejection of PcG proteins. In support of this possibility, Pc2 was recently shown to relocate from Polycomb bodies to transcription factories dependent on demethylase KDM4C[141]. Targeting of Pc2 relies on ncRNAs, including *TUG1*, which interacts with PRC2 and acts as a scaffold at Polycomb bodies, and *NEAT2*, which associates with epigenetic regulators involved in gene activation. Taken together, genes likely are directed to repressive Polycomb bodies or active transcription factories through dynamic interplay of PcG and TrxG proteins, whose long-range interactions require the function of ncRNAs and chromatin insulators. This model, recently highlighted by [142], is consistent with early findings in *D. melanogaster*, wherein mutations in both PcG and TrxG genes were shown to modulate the activity and nuclear organization mediated by insulators [143].

Enhancer-Promoter Interactions

As the name suggests, enhancers are regulatory elements functionally defined by their ability to activate transcription and to do so regardless of their location, distance, or orientation with respect to gene promoters [144]. Enhancers underlie complex spatiotemporal regulation of tissue-specific gene expression and have been characterized by chromatin and transcription factor signatures in mammalian cells [59, 145, 146]. Enhancers are commonly separated by large genomic distances from their associated promoters, which makes accurate assignments of enhancer-promoter relationships

difficult and suggests that long-range interactions are a defining feature of gene regulation. Numerous 3C-based interaction studies have supported enhancer looping models wherein enhancers directly contact gene promoters for activation, and recent models suggest chromatin signatures at enhancers may act as epigenetic signals for transcriptional activation [147]. Interestingly, a distinct class of enhancer elements is also capable of recruiting RNAPII and is transcribed into enhancer-derived RNAs [148, 149]. The transcriptional output of neuronal activity-regulated enhancers positively correlates with the expression levels of associated genes [148], and alternative models of enhancer function have been proposed [150], including ones in which enhancers and their associated promoters colocalize by virtue of recruitment to transcription factories.

Enhancer-promoter interactions and transcriptional clustering of active genes consistent with transcription factory models is further supported by ChIA-PET analyses enriching for RNAPII-based chromosomal interactions. [151]. As many as 65% of RNAPII binding sites have been shown to be involved in a complex network of physical interactions in human cell lines. RNAPII interactions were intergenic (e.g., promoter-promoter), and extragenic (e.g., promoter-enhancer), with most contacts aggregated into ~1,500 interaction complexes. Multigene complexes typically consisted of related and coregulated genes, which suggests gene families functionally associate for cotranscription, whereas single-gene complexes tended to associate with tissue-specific or developmentally regulated genes and cell-type-specific enhancer-promoter interactions. The authors ultimately reveal a complex organization of transcription reflecting the importance of long-range chromatin interactions between coregulated promoters and between enhancers and promoters, possibly at transcription factories.

Nevertheless, the role of long-range enhancer-promoter interactions in eukaryotic gene activation and how these interactions are organized is not fully understood. Traditional models propose that enhancers underlie recruitment and assembly of the transcription machinery at core promoters [152]. However, in the case of the human *FOSL1* gene, cross talk between an enhancer and promoter triggers transcription elongation [153], which suggests some enhancers may function by releasing RNAPII from promoter-proximal pausing. Meanwhile, recent characterization of chromatin-associated proteins at the human *α-globin* genes and upstream MCS-R2 enhancer suggest distal enhancers stimulate gene expression by reversing PcG activities [154]. The MCS-R2 enhancer is required for recruitment of H3K27 demethylase JMJD3, although active chromatin marks indicative of Trx activity (H3K4me3) were present in the absence of the enhancer. This supports a model wherein PcG and TrxG complexes dynamically associate with target genes and enhancers promote gene induction by favoring Trx activity. This is consistent with ChIA-PET mapping of RNAPII contacts, which found high enrichment of active chromatin marks H3K4me1 and H3K4me3 coupled with a lack of repressive marks at RNAPII interaction sites [151].

Recent studies have also provided new insight into how enhancers interact with distant target promoters to induce gene transcription and have suggested roles for transcription factors, chromatin insulators, and a unique cohesin complex in enhancer-promoter organization. In the case of the well characterized *β-globin* locus, whose expression levels are developmentally regulated by a distal upstream locus control region (LCR),

long-range enhancer-promoter interactions require transcription factors EKLF and GATA-1 [155, 156], and genes coregulated by EKLF preferentially cluster into shared transcription factories [157]. Integration of an ectopic human LCR on a separate chromosome, for example, does not prevent the LCR from interacting in *trans* with EKLF- and GATA-1-regulated genes [158]. These results suggest that transcription factors coordinate the specificity and organization of enhancer-promoter intra- and inter-chromosomal interactions. Through the powerful combination of 4C and FISH, [158] also demonstrate that interchromosomal interactions between the LCR and *β-globin* genes are limited to specific "jackpot" cells actively transcribing *β-globin* genes, which suggests interactions are cell specific and reflect genome conformations that are conducive to enhancer-promoter association. In other words, enhancers preferentially interact with genes through shared transcription factors, but they do so stochastically in a restricted nuclear space that varies from cell to cell.

Insulators also play an important role in facilitating cell-type-specific chromatin organization conducive to enhancer-promoter interactions, and recent mapping of CTCF interactions in pluripotent cells supports this possibility [78]. CTCF-mediated tissue-specific chromatin architecture has been characterized at the apolipoprotein gene cluster; the imprinted *IGF2-H19* locus; and the developmentally regulated *IFNG*, *β-globin*, *MHC-II*, and *CFTR* loci [147, 159]. However, recent findings suggest that cohesin complexes, which are recruited by CTCF, are also capable of stabilizing enhancer-promoter chromatin looping in the absence of CTCF. For example, cell-type-specific enhancer-promoter interactions in murine embryonic stem cells require a unique cohesin

complex, including the transcriptional coactivator Mediator and cohesin loading factor nipped B-like protein (NIPBL) [160]. Cohesin and NIPBL were also recently shown to mediate interactions between the *β-globin* genes and upstream LCR, and whereas CTCF-coordinated organization at the locus does not directly influence gene expression [75], the cohesin-mediated LCR interaction regulates globin gene expression *in vivo* and *in vitro* [161].

PERSPECTIVES

The power of chromatin profiling and 3C-based genomic strategies for exploring genome-wide interactions has led to substantial progress in our understanding of nuclear organization over the past few years. Meanwhile, the recent identification of tDNA insulator activity in humans [27], mapping of CTCF-mediated chromosomal interactions in mammals [78], and identification of genome-folding principles in human cells [20, 23, 66, 67] and *Drosophila* embryos [65] have significantly expanded our knowledge of insulator function and the highly conserved role of insulators in genome organization. The requirement for insulators in mediating long-range interactions essential for Polycomb repression [126] and their localization to transcription factories [140] suggest insulators underlie the dynamic interplay between epigenetic gene repression and gene induction associated with developmental gene regulation. These observations can be used to derive a comprehensive model of the role of insulators and chromatin structure in nuclear organization (Figure 3), with emphasis on the evolutionarily conserved role of insulators in yeast, *Drosophila*, and mammals; the molecular players involved in each; and their role in gene localization to Polycomb bodies and transcription factories. Despite

substantial progress, several important questions remain, including how chromosomal associations and underlying insulator activities are regulated. The discovery of tDNA insulator activity in mammals raises the question of whether the highly conserved tDNA and CTCF insulators functionally cooperate. Finally, studies have only begun to characterize the new roles for ncRNAs in gene regulation and nuclear organization.

Interactions mediated by insulators, PcG proteins, and enhancer/promoter-bound factors result in the creation of a 3D arrangement of the DNA that must represent a fingerprint of the functional status of the nucleus. Therefore, a detailed understanding of all inter- and intrachromosomal interactions in the nucleus together with information on the nature and function of the interacting loci can lead to the establishment of structure-based functional maps of nuclear output that are a representation of cell identity. Some of these interactions may be a consequence of genome function, whereas others may be established during cell differentiation to elicit specific patterns of gene expression. As a consequence, the 3D architecture of the genetic material may carry epigenetic information in addition to that written into the 10-nm chromatin fiber. Understanding how this information is maintained during the cell cycle and how the 3D arrangement of chromosomes during interphase relates to their structure during mitosis remains a major challenge for the near future.

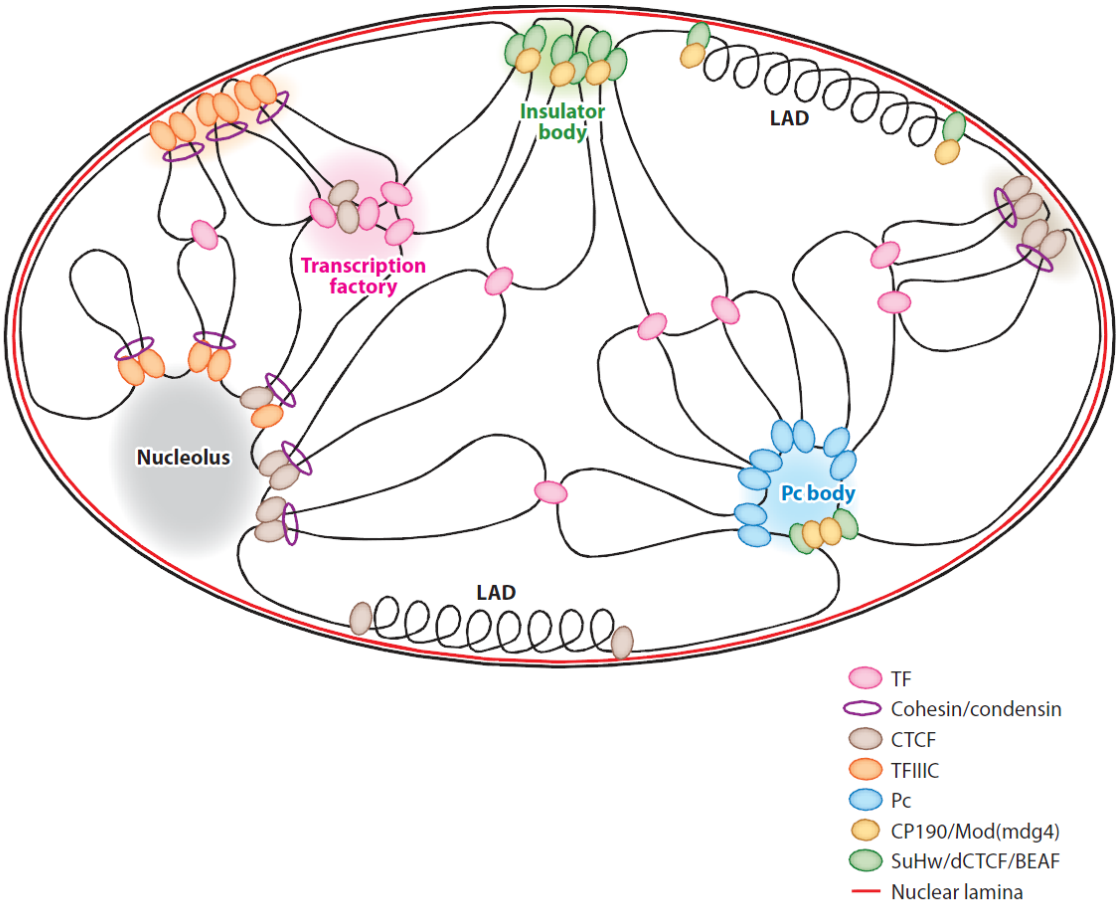


Figure 3. **Comprehensive model for the highly conserved role of insulators in nuclear organization**. Insulators in yeast (TFIIIC, *orange*), *Drosophila* [Su(Hw), dCTCF, BEAF, *green*; CP190, Mod(mdg4), *yellow*], and mammals (CTCF, brown; TFIIIC, *orange*) mediate long-range inter- and intrachromosomal interactions important for gene regulation and cluster into subnuclear foci called insulator bodies. Insulators underlie interactions necessary for Polycomb (Pc) body repression (*blue*) and localize with general transcription factors (TF, *pink)* to transcription factories. Insulators localize to subnuclear structures, including the nuclear lamina (*red*), where they are enriched at the borders of lamina-associated domains (LADs) and the nucleolus (*gray*). CTCF insulator activity in

mammals requires cohesin (*purple*), and TFIIIC insulator sites are associated with both cohesin and condensin (*purple*). Insulator activity in *Drosophila* relies on recruitment of fly-specific proteins CP190 and Mod(mdg4)

Remaining questions concerning the role of insulators in chromosome organization

The following chapters presented in this dissertation address two fundamental questions concerning the organization and function of insulator elements in genome biology. For one, a majority of sites bound by insulator proteins are not capable of classical insulator activities, such as enhancer-blocking, in transgenic reporter assays. Second, a majority of insulator binding sites localize within chromosome domains, rather than at the boundaries for which they are proposed to function as barriers. Thus, what differentiates boundary from non-boundary insulators, and sites capable versus incapable of enhancer-blocking, remained unresolved when I began my graduate research. By integrating genomic strategies aimed at mapping the binding profile of insulator proteins and the interactome of eukaryotic chromosomes, we discover a spectrum of insulator protein occupancy that scales with the heterogeneous partitioning of topological domains, the ability of these regulatory elements to function as insulators, and the role these factors play in chromosome organization and transcriptional regulation. In addition, I present evidence that insulator protein binding is dynamically regulated during steroid-hormone signaling, and that transforming-growth factor beta (TGF-β) effector proteins co-localize with insulator proteins in a dCTCF-dependent manner. These results introduce an important connection between insulator elements and extracellular signaling responses, suggesting the genomic response to signaling takes place at specific architectural protein binding sites. Finally, I present evidence from small RNA-seq analysis that tRNA levels are dynamically altered in response to signaling events in a manner that reflects changes in codon usage associated with the differentiation status of the cell.

CHAPTER 2:

DROSOPHILA CTCF TANDEMLY ALIGNS WITH OTHER INSULATOR PROTEINS AT THE BORDERS OF H3K27ME3 DOMAINS

Work presented in this chapter is published in:

Van Bortle K, Ramos E, Takenaka N, Yang J, Wahi JE, Corces VG. Genome Res. 2012. Drosophila CTCF tandemly aligns with other insulator proteins at the borders of H3K27me3 domains. Nov;22(11):2176-87.

My contributions include conceiving the project, performing the experiments, computational analysis of the data, interpretation of the results, and writing the paper.

ABSTRACT

Several multi-protein DNA complexes capable of insulator activity have been identified in *Drosophila melanogaster*, yet only CTCF, a highly conserved zinc-finger protein, and the transcription factor TFIIIC have been shown to function in mammals. CTCF is involved in diverse nuclear activities, and recent studies suggest that the proteins with which it associates and the DNA sequences it targets may underlie these various roles. Here we show that the *Drosophila* homolog of CTCF (dCTCF) aligns in the genome with other *Drosophila* insulator proteins such as Suppressor of Hairy wing (SU(HW)) and Boundary Element Associated Factor (BEAF-32) at the borders of H3K27me3 domains, which are also enriched for associated insulator proteins and additional co-factors. Disruption of insulators genome-wide by knockdown of dCTCF and combinatorial knockdown of other *Drosophila* insulator proteins leads to a reduction of H3K27me3 levels within repressed domains, suggesting chromatin insulators are important for the maintenance of appropriate repressive chromatin structure in Polycomb (Pc) domains. These results shed new insights into the roles of insulators in chromatin domain organization, and support recent models wherein insulators underlie interactions important for Pc-mediated repression. We reveal an important relationship between dCTCF and other *Drosophila* insulator proteins, and speculate that vertebrate CTCF may also cluster with other nuclear proteins to accomplish similar functions.

INTRODUCTION

Insulators were first characterized as regulatory elements that play an important role in establishing proper gene expression in eukaryotic cells. Early studies demonstrated the ability of insulators to act as barriers, preventing the spread of heterochromatin and thereby demarcating chromatin boundaries, as well as enhancer-blockers, preventing enhancers from activating nearby genes in a direction-dependent manner [162, 163]. Insulators have since been shown to be multiprotein-DNA complexes that can mediate inter- and intra-chromosomal interactions important for facilitating proper gene expression at specific loci, and more recently in genome-wide chromatin organization [73]. Insulator activity in vertebrates requires the essential, highly conserved, CCCTC-binding factor CTCF. Recent genome-wide studies have effectively mapped both the mammalian CTCF binding sites and the chromatin interactions they facilitate [61, 78]. However, how CTCF mediates these interactions and the proteins associated with functional insulator activity remains poorly understood.

The CTCF insulator protein contains a highly conserved central domain encoding 11 zinc fingers, and is ubiquitously expressed [28]. Interestingly, CTCF has been implicated in numerous, unique nuclear functions in addition to the classical enhancer-blocking and barrier activities that define insulators. These include X-chromosome inactivation [164], nucleolar stability [165], V(D)J recombination [166], and global chromatin organization [61, 78]. The combinatorial use of its 11 zinc fingers in binding to discrete DNA target sequences, as well as the diverse, context-dependent protein-interaction networks of CTCF, have been proposed to underlie these numerous roles [29, 32, 33]. Meanwhile,

recent studies in both *D. melanogaster* and humans have demonstrated that CTCF appears to demarcate physical chromatin domains [65], including a subset of repressive H3K27me3 domains [60, 68]. The proteins with which CTCF associates and the purpose for which CTCF demarcates chromatin boundaries requires further exploration.

In *Drosophila*, several insulator binding proteins have been identified and characterized, including the *Drosophila* homolog of CTCF (dCTCF), Boundary Element Associated Factor of 32 kDa (BEAF-32), and Suppressor of Hairy Wing (SU(HW)) [40]. These DNA-binding proteins require additional proteins for functional insulator activity, including Centrosomal Protein 190 (CP190) and Modifier of mdg4 (MOD(MDG4)) [43, 44, 167]. We have recently identified the genome-wide binding sites of dCTCF, BEAF-32, SU(HW), and CP190 with high resolution ChIP-seq, and demonstrated that recruitment of these proteins is regulated during the ecdysone response in *D. melanogaster* [77]. However, the functional relationship between these different insulator proteins remains poorly characterized.

Here we present a comprehensive map of direct insulator-binding sites throughout the *Drosophila* genome, and show that as many as 40% of dCTCF sites align tightly with the *Drosophila* specific insulators SU(HW) and BEAF-32. dCTCF sites are enriched for three similar, but distinct, DNA motifs, potentially representing discrete binding modes throughout the *Drosophila* genome. Aligned insulators are enriched for additional co-factors and commonly occur at the borders of H3K27me3 domains, where they are essential for maintaining appropriate chromatin structure. Surprisingly, we find that

disruption of insulators genome-wide by knockdown of insulator components does not significantly affect the expression of genes flanking H3K27me3 domains, nor does H3K27me3 spread beyond domain borders as one might expect based on classical barrier models for insulator function at chromatin boundaries. Instead, H3K27me3 is lost within domains, suggesting chromatin insulators serve an important role in the maintenance of silenced chromatin in Polycomb (Pc) domains. Our findings support recently proposed models, wherein chromatin insulators are involved in mediating long-range interactions important for Pc-mediated repression [142].

RESULTS

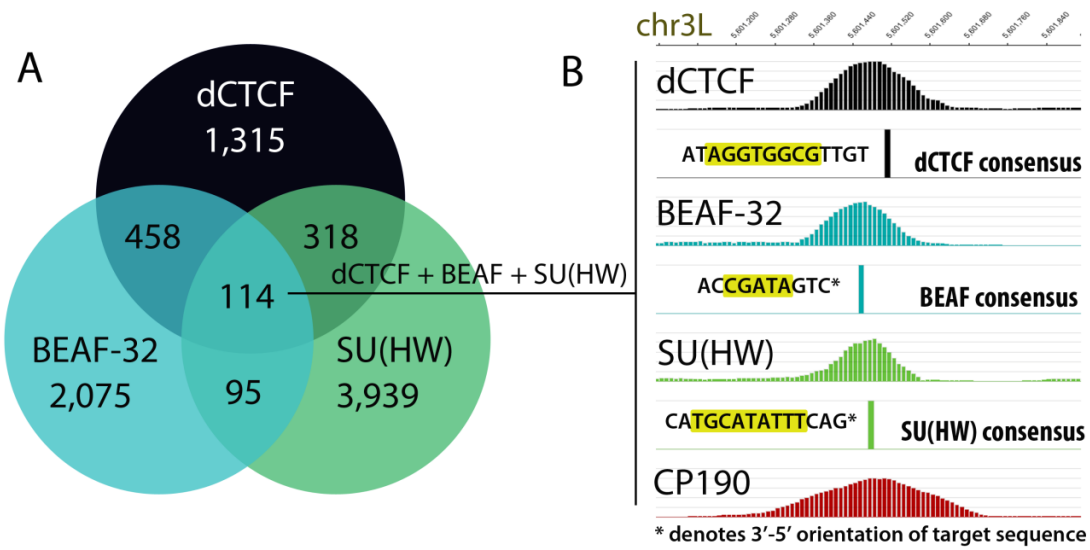
dCTCF sites align with *Drosophila* specific insulator proteins Su(Hw) and BEAF-32

Two recent studies independently identified the genome wide binding sites of insulator proteins in *Drosophila melanogaster* by combining chromatin immunoprecipitation with microarray hybridization (ChIP-chip). Whereas one study demonstrated unique genome-wide distributions and gene ontologies between dCTCF, SU(HW), and BEAF-32 [58], the other observed an enrichment of dCTCF and BEAF-32 as co-localizing, and therefore split insulators into two main classes: dCTCF/BEAF-32/CP190 (Class I), and SU(HW) (Class II) [45]. However, previous studies have shown that CP190 is an essential component of both the dCTCF and SU(HW) insulators [43, 167]. The functional implications of insulator classes and why dCTCF might co-localize with other insulator proteins requires further exploration.

To better determine the genome wide binding sites of *Drosophila* insulators with greater accuracy and resolution, we recently re-mapped dCTCF, SU(HW), BEAF-32, and CP190 sites by combining chromatin immunoprecipitation with high-throughput sequencing (ChIP-seq). Here we analyze peaks repeatedly called in three independent ChIP-seq experiments during the ecdysone response in *Drosophila* Kc cells, which are therefore most likely to represent real, stable insulator binding sites [77]. We then determined enriched consensus sequence motifs by MEME-ChIP [168]. Results confirm previously identified position weight matrices for each respective insulator protein [31, 45, 169] (Supplemental Fig. S1). Given the ability of distant insulator proteins to interact with

each other, it is possible that different insulator proteins bound to these sites may co-precipitate during the ChIP procedure, thus appearing to co-localize when, in fact, they are located in different genomic locations. As a consequence, we speculate that genome-wide binding profiles for each insulator protein likely contain many indirect binding sites. Therefore, we excluded sites that do not contain appropriate target sequences for each respective insulator protein (see Methods), thereby providing a stringent list of insulator sites that are highly likely to represent real, direct binding sites for each protein.

Results from this analysis suggest that insulators indeed cluster together in the genome often as previously reported [45], and do so while binding their own discrete target sequence. As many as 40% of all dCTCF sites align with SU(HW) or BEAF-32, and as many as 5% of all dCTCF sites tightly align with both SU(HW) and BEAF-32 (Figure 4). Though previous studies broke insulators into two or three classes, we find that dCTCF aligns with SU(HW) (432 (49% of aligned sites)) and/or BEAF-32 (572 (64% of aligned sites)) at many sites. Given the number of SU(HW) binding sites throughout the genome (4,466 sites with SU(HW) consensus, Table 1), earlier correlation analyses of co-localization were likely biased by thousands of independent SU(HW) sites. The high-resolution obtained by ChIP-seq demonstrates that these insulators align tightly, within only 200-300 bp (Figure 4), and sequential ChIP for insulator proteins dCTCF, BEAF-32, and SU(HW) suggests these proteins co-localize at these sites in individual cells (Supplemental Fig. S2). In addition, by removing insulator sites lacking



C

	# of ChIP-seq sites (Wood et al., 2011)	# of ChIP-seq sites with consensus	# of sites aligned with different insulator(s)	# of independent (non-aligned) sites
dCTCF	2,871	2,205	890 (40%)	1,315 (60%)
BEAF-32	3,617	2,742	667 (24%)	2,075 (76%)
SU(HW)	6,617	4,466	527 (12%)	3,939 (88%)

Figure 4. **dCTCF aligns with *Drosophila* DNA-binding insulator proteins Su(Hw) and BEAF-32**. (A) Venn diagram depicting overlap of dCTCF binding sites with those of BEAF-32 and SU(HW). Overlap represented as number of sites (summits +/- 150 bp) in which dCTCF intersects BEAF-32 and/or SU(HW), wherein target sequences are present for each insulator protein, suggesting close alignment (within 150 bp). (B) Example ChIP-seq profile for dCTCF, BEAF-32, SU(HW), and CP190 on chromosome 3L, in which case dCTCF aligns with both BEAF-32 and SU(HW), where each cognate target is present. (C) Number of sites in which dCTCF, BEAF-32, and SU(HW) contain appropriate target sequences. Percentages of sites in which dCTCF, BEAF-32, and

SU(HW) align closely with other DNA-binding insulator proteins. Most (90%) alignments include dCTCF, and as many as 40% of dCTCF sites align with either BEAF-32 and/or SU(HW).

known target sequences, we demonstrate that each insulator protein binds to its own target sequence (notice DNA sequence, Figure 4), and that overlap is not a consequence of indirect binding. The alignment of dCTCF with both SU(HW) and BEAF-32 suggests the possibility of synergistic cooperation in insulator function.

dCTCF sites are enriched for multiple DNA motifs

In addition to its ability to interact with numerous nuclear proteins, the versatility of CTCF in genome biology may also be attributable to its wide range of potential target sequences. However, genome-wide analyses of CTCF binding sites have revealed a primarily enriched core target sequence that is strikingly similar between invertebrates and vertebrates despite millions of years of evolution [31] (Supplemental Fig. S1). This is not entirely surprising given that CTCF encodes 11 highly conserved zinc fingers that confer target specificity. However, early characterization of CTCF identified its ability to bind to a wide range of sequences dependent on its combinatorial use of zinc fingers [29, 170], and recent work has identified similar regulatory elements bound by CTCF in the human genome [171]. These data suggest that CTCF may bind to unique DNA target sequences, not represented in the conserved target sequence.

Motif analysis of dCTCF ChIP-seq data by MEME-ChIP [168] indeed identifies the primary consensus sequence of dCTCF as previously reported (Figure 5, Supplemental Fig. S1). However, results also indicate strong enrichments for a strikingly similar, but novel secondary consensus sequence (Figure 5), also independently obtained using Weeder 1.3 [172], suggesting the variability in target sequence specificity holds true for

Drosophila. There is also enrichment for an additional motif accounting for less than 10% of dCTCF sites (Figure 5). Comparison of dCTCF read intensities at these three motifs suggests that the highly conserved core consensus (motif 1) recruits higher occupancy levels of dCTCF, whereas motifs 2 and 3 recruit lower but similar occupancy levels (Supplemental Fig. S3). This finding is similar to previous reports suggesting CTCF targets different occupancy-based motif classes in vertebrates [173], and suggests that these unique target sites may underlie distinct roles.

Studies in *CP190* mutants demonstrated a dependence of dCTCF on CP190 for binding to a subset of its DNA binding sites on polytene chromosomes [43, 174]. Earlier studies have shown that although CP190 physically associates with insulator proteins SU(HW) and MOD(MDG4)2.2 and is essential for functional insulator activity, it does not directly bind to insulator sequences present on the *gypsy* retrotransposon [167], and therefore likely relies on dCTCF and SU(HW) to associate with insulator sites. In support of this notion, recent biochemical studies demonstrated that CP190 function at dCTCF, SU(HW) and BEAF-32 sites requires its BTB/POZ (protein interaction) domains, whereas its zinc fingers were dispensable [175]. Interestingly, we find significant enrichments for CP190 at dCTCF sites containing motifs 2 or 3 when compared to the primary conserved consensus (Figure 5B). Given the dependence of dCTCF on CP190 for binding to a subset of its sites, we speculate that interactions between dCTCF and CP190 facilitate its interaction with these low occupancy, and presumably lower affinity, target sequences. Furthermore, we find enrichments for the novel secondary sequence at sites where

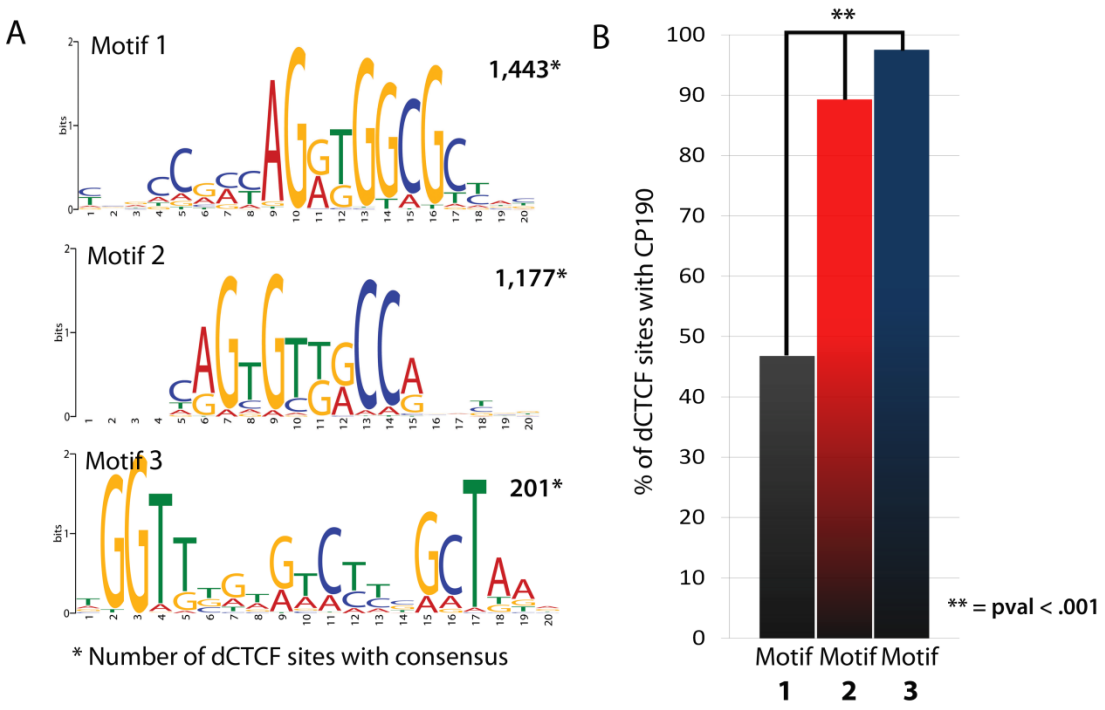


Figure 5. **dCTCF sites are enriched for three distinct DNA motifs, including a similar but novel secondary motif enriched for insulator protein CP190**. (A) Position weight matrices for primary target sequence and secondary target sequences obtained by MEME-ChIP, and confirmed with Weeder 1.3. Number of sites provided represents sites in which dCTCF summits are within +/- 150 bp of the DNA motif. (B) Percentage of dCTCF sites in which CP190 is present when containing each DNA motif.

dCTCF aligns with BEAF-32, and where dCTCF aligns with both BEAF-32 and SU(HW) (Supplemental Fig. S4), suggesting these DNA target sequences exhibit distinct features with respect to insulator recruitment and alignment.

dCTCF recruits unique MOD(MDG4) isoform(s)

The recruitment of CP190 is of particular interest given its ability to form stable homodimers and homotetramers *in vitro,* supporting the notion that active insulators function through loop formation via interactions with other insulators. CP190 contains a unique BTB domain that excludes it from the ttk (tramtrack) group of BTB/POZ proteins, and inhibits it from interacting with ttk members [42]. However, SU(HW) was first characterized as recruiting an additional BTB/POZ protein, MOD(MDG4), also essential for insulator activity [44, 176]. MOD(MDG4) in fact, belongs to the ttk group of BTB/POZ containing proteins, and has been shown to form higher-order homo-oligomers [42]. Meanwhile, the observation that in diploid cell nuclei insulators form large structures called insulator bodies suggests many insulators associate together within the nucleus, and the presence of CP190 and MOD(MDG4) supports this possibility [43, 44].

Whereas CP190 has been shown to associate with dCTCF, SU(HW), and BEAF-32 [58], currently only the SU(HW) insulator has been shown to recruit MOD(MDG4). Although the *mod(mdg4)* gene encodes for at least 26 alternatively spliced variants, each containing a common N-terminal region encoding the ttk-family BTB/POZ domain [177], SU(HW) insulator activity requires one specific isoform, MOD(MDG4)2.2 [176].Staining of

Drosophila polytene chromosomes reveals MOD(MDG4) specific bands, unaccounted for with MOD(MDG4)2.2 staining alone (Figure 6A), suggesting additional isoforms are recruited to DNA. Whether dCTCF and BEAF-32 recruit unique MOD(MDG4) isoforms is currently unknown. We therefore carried out ChIP-seq analyses in *Drosophila* Kc cells using two different antibodies that recognize either all MOD(MDG4) isoforms or MOD(MDG4)2.2.

The binding profile for MOD(MDG4)2.2 is significantly accounted for at SU(HW) sites (Figure 6B). Given that MOD(MDG4)2.2 is required for SU(HW) insulator activity, the MOD(MDG4)2.2 map may reveal a subset of active SU(HW) sites throughout the *Drosophila* genome. Here, we find that the ChIP-seq profile of MOD(MDG4), which includes significantly more binding sites than MOD(MDG4)2.2 alone (Figure 6B), reveals unique peaks at discrete dCTCF and BEAF-32 sites, suggesting additional isoforms of MOD(MDG4) recruited by dCTCF and BEAF-32 must exist (Figure 6C). Whereas average read intensities for MOD(MDG4)2.2 are strongest at SU(HW) sites, dCTCF and BEAF-32 sites show an opposite trend, with stronger read intensities for MOD(MDG4) (Supplemental Fig. S5). Finally, as many as 64% of dCTCF sites and 38% of BEAF-32 sites co-localize with an isoform of MOD(MDG4). These data suggest that dCTCF and BEAF-32 indeed recruit unique isoforms of Mod(md4), and that all three *Drosophila* insulators function similarly through the recruitment of BTB domain-containing proteins CP190 and MOD(MDG4).

Aligned dCTCF sites are enriched for CP190, MOD(MDG4) isoforms, and additional co-factors

The tight alignment of dCTCF with BEAF-32 and SU(HW), combined with their common insulator function, suggests that insulators synergize, perhaps to create a more active insulator complex. Given that CP190 and MOD(MDG4) form homo-oligomers, it is also intuitive to imagine closely aligned insulators cooperatively recruiting CP190 and MOD(MDG4), increasing the likelihood that each member of the insulator cluster is functionally active. In support of this hypothesis, we find enrichment for aligned dCTCF sites containing CP190 and MOD(MDG4) when compared to independent dCTCF insulators (Figure 6D-F). This suggests that by associating with BEAF-32 and/or SU(HW), dCTCF might ensure that it will become a functionally active insulator complex by recruiting essential co-factors.

Many additional proteins have been functionally associated with insulators in *D. melanogaster*, suggesting these insulator clusters may represent hubs for recruiting other co-factors. For example, Lethal (3) malignant brain tumor (L(3)MBT) has been recently shown to co-localize with the *Drosophila* chromatin insulators dCTCF, BEAF-32, SU(HW), and CP190 [105], and other studies found direct interactions between L(3)MBT and the SU(HW) insulator protein [178]. L(3)MBT imparts transcriptional regulation of the Salvador-Warts-Hippo (SWH) pathway, likely repressing SWH target genes important for cell proliferation and organ size control. Whereas recently published L(3)MBT sites [105] are enriched primarily at independent dCTCF sites over the BEAF-32 and SU(HW) insulators, we find enrichment of L(3)MBT sites, identified

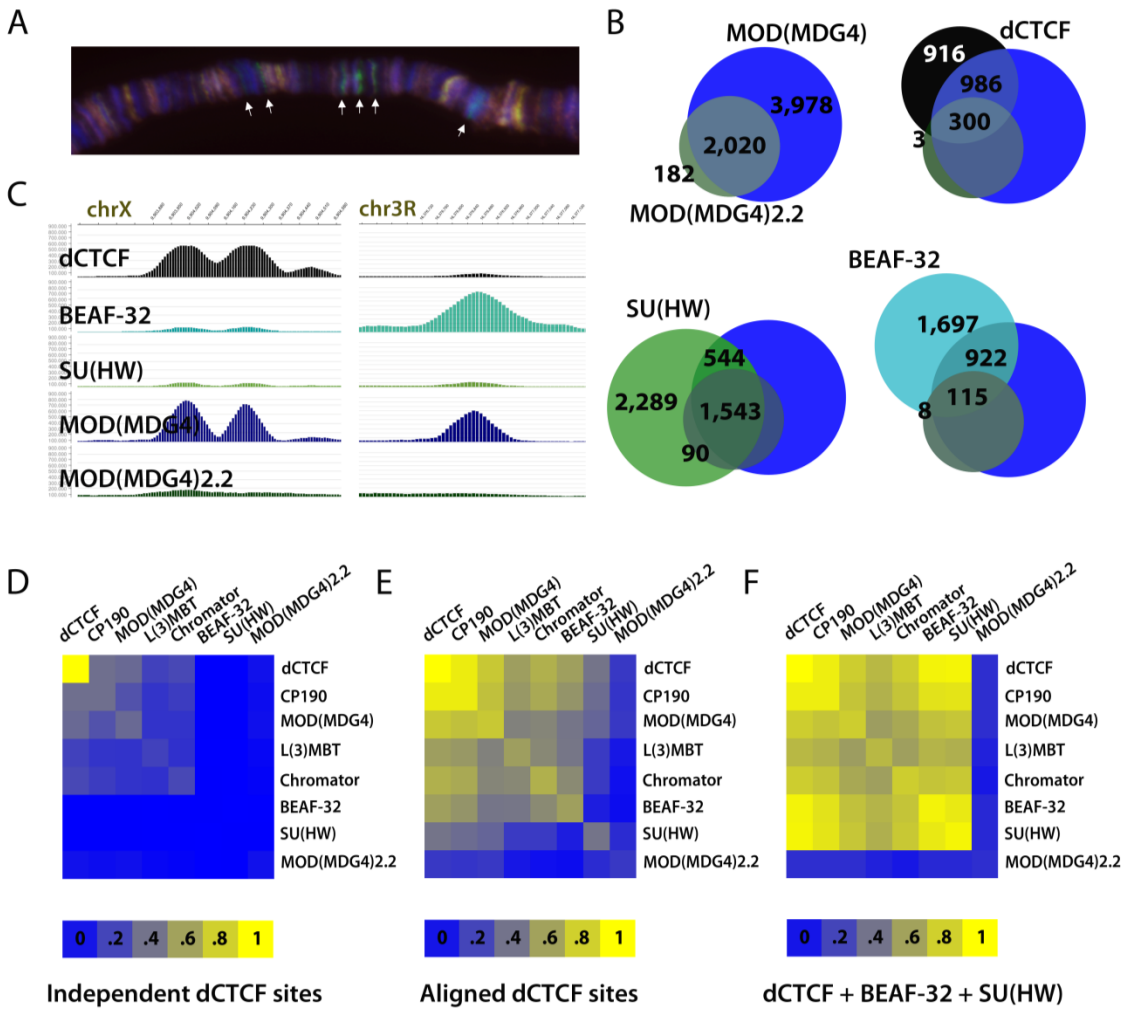


Figure 6. **dCTCF and BEAF-32 recruit isoform(s) of Mod(mdg4) different from Mod(mdg4)2.2**. Aligned dCTCF sites are enriched for Mod(mdg4) and additional co-factors (A) Immunofluorescence microscopy of Mod(mdg4) (green) and Mod(mdg4)2.2 (red) on *Drosophila* polytene chromosomes. Mod(mdg4) staining includes many discrete bands not accounted for by Mod(mdg4)2.2 specific antibodies, depicted with white arrows. (B) Genome wide overlap of dCTCF, BEAF-32, and Su(Hw) with Mod(mdg4) and Mod(mdg4)2.2. Many dCTCF (45%) and BEAF-32 (34%) sites overlap Mod(mdg4) isoform(s) not represented by Mod(mdg4)2.2. Meanwhile, many Su(Hw) (37%) sites

overlap Mod(mdg4) 2.2 sites, as expected. (C) ChIP-seq profile for Mod(mdg4) and Mod(mdg4) 2.2 reveals many unique peaks specifically in the Mod(mdg4) profile, accounted for at BEAF-32 and dCTCF sites. (D-F) Heatmap representation of percentages of dCTCF sites in which CP190, Mod(mdg4), Mod(mdg4) 2.2, BEAF-32, Su(Hw), L(3)MBT, and/or Chromator co-occur at independent dCTCF sites, aligned dCTCF sites, and sites where dCTCF aligns with both BEAF-32 and Su(Hw).

independently by ChIP-seq in *Drosophila* Kc cells, at aligned dCTCF sites when compared to independent dCTCF sites (Figure 6D-F). In addition to SU(HW), L(3)MBT interacts with a chromodomain protein, Chromator, that has also recently been shown to co-localize and cooperate with the BEAF-32 insulator [179, 180]. Indeed, using publicly available ChIP-chip data for Chromator in *Drosophila* Kc cells [25], aligned dCTCF sites also show an apparent enrichment for Chromator (Figure 6D-F). Interestingly, Chromator and zinc finger protein Z4 are important for maintaining polytene chromosome structure [181], suggesting a functional relationship between insulators and Chromator in chromatin domain organization.

Together, these data suggest dCTCF may team up with *Drosophila* specific insulator proteins in order to more efficiently recruit co-factors essential for insulator activity. These insulator sites are enriched for additional insulator-related proteins L(3)MBT and Chromator, suggesting these sites are different from independent insulator sites, and appear to represent large complexes of proteins associated with insulator activity.

Aligned dCTCF sites commonly flank the borders of H3K27me3 domains

The correlation of dCTCF with SU(HW) and BEAF-32 is striking, but why dCTCF function requires clustering with other insulator proteins requires further exploration. Recent interrogation of chromosome architecture in *D. melanogaster* revealed recurrent combinations of insulators and active histone marks at the borders of physical domains, including enrichments for Chromator [65]. Comparison with physical domains analyzed by Sexton et al reveals enrichment for aligned dCTCF sites within 5 Kb of domain

borders, suggesting these tandemly aligned insulators are involved in demarcating chromatin domains (Figure 7A-B). Recent studies in both *Drosophila* and humans have also demonstrated an enrichment of CTCF and other insulators at the borders of H3K27me3 domains [45, 60, 68].

In order to determine whether aligned dCTCF insulator sites occur at H3K27me3 domain borders in *Drosophila* Kc cells, we independently mapped repressive chromatin domains by ChIP-seq against H3K27me3. We find an enrichment of insulator proteins immediately outside of H3K27me3 domain borders, and significant enrichment of aligned dCTCF sites within 5 kb of H3K27me3 domains (Figure 7C-D). Interestingly, read intensities for each insulator protein flanking domain borders (Figure 7D) suggest a periodicity of insulator presence beginning with dCTCF, consistent with the observation that insulators tandemly align. There is no significant enrichment for dCTCF aligned with BEAF-32 vs. dCTCF aligned with SU(HW) at domain borders (Supplemental Fig. S6), suggesting dCTCF aligns with either BEAF-32 and/or SU(HW) at the borders of repressed chromatin domains. However, the functional significance of insulators at chromatin domain borders and dCTCF alignment remains poorly characterized

Insulator knockdown results in H3K27me3 loss within repressed domains

Previous analyses of H3K27me3 levels immediately flanking domain borders in *dCTCF* and *CP190* mutants suggest these sites functionally maintain chromatin architecture at these domains by preventing the spread of heterochromatin [68]. Given recent data that CTCF associates with various nuclear proteins in a context dependent fashion, it is

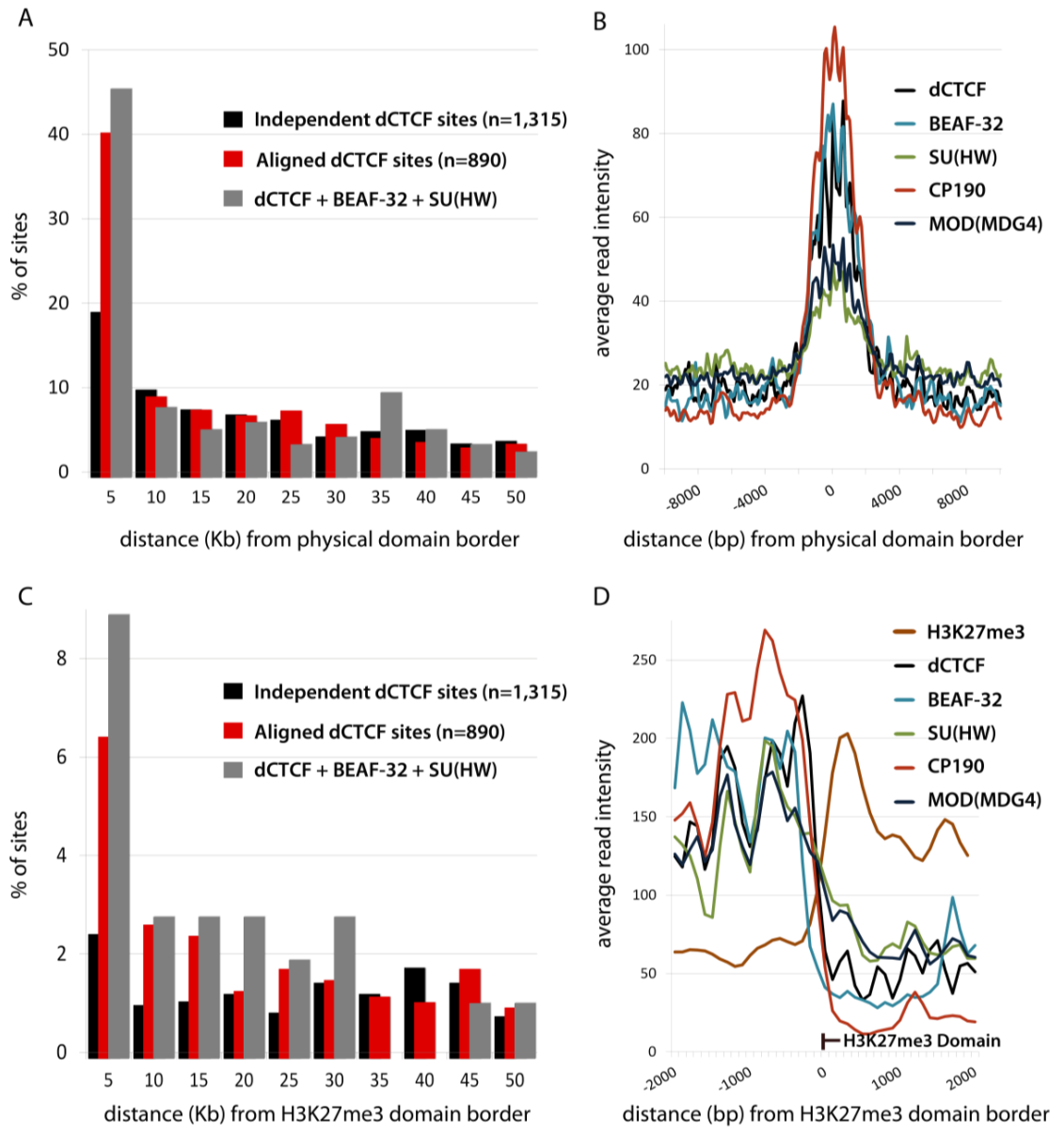


Figure 7. **Aligned dCTCF sites are enriched at the borders of H3K27me3 and physical domains**. (A) Percentage of independent and aligned dCTCF sites within 5 Kb from recently mapped physical domain boundaries [65]. (B) Average read intensity for insulator proteins at physical domain boundaries, +/- 10 Kb. Comparison of insulator profiles normalized by total read numbers. (C) Percentage of independent and aligned dCTCF sites within 5 kb from H3K27me3 domain borders. (D) Average read intensity

for H3K27me3 and insulator proteins at H3K27me3 domain borders, +/- 2 kb. Comparison of insulator profiles normalized by total read numbers.

conceivable that dCTCF aligns with other insulators to form a robust barrier at the borders of repressive domains to effectively prevent the spread of heterochromatin. Therefore, we next sought to determine whether insulators are important for maintaining appropriate chromatin architecture and gene expression at these H3K27me3 domains by combinatorial knockdown of insulator proteins in *Drosophila* Kc cells (Supplemental Fig. S7).

Surprisingly, we find no evidence for downregulation of domain-flanking genes compared to genome wide averages when insulators are disrupted genome-wide (Figure 8A), as one might expect if heterochromatin spreads beyond domain boundaries. We therefore carried out ChIP-seq for H3K27me3 in *Drosophila* Kc cells in which insulator protein dCTCF was knockdown. Results revealed decreased levels of H3K27me3 immediately within domain borders as well as throughout H3K27me3 domains, but not an increase outside of domain boundaries (Figure 8B). H3K27me3 levels were specifically reduced in Polycomb (Pc) domains containing dCTCF, indicating loss of H3K27me3 is a direct effect of dCTCF knockdown rather than general a consequence of disrupted chromatin architecture (Supplemental Fig. S8). ChIP-PCR against H3K27me3 levels at several loci confirms significant loss of H3K27me3 in dCTCF knockdown, as well as under various combinations of insulator knockdown, including MOD(MDG4), suggesting the recruitment of MOD(MDG4) to dCTCF sites and enrichment at aligned insulators is functionally significant (Figure 9). Importantly, expression of the *Enhancer of zeste* (*E(z)*) gene, the Pc methyltransferase responsible for H3K27me3, is unaffected by any combination of insulator knockdown, and nuclear levels of H3K27me3 remain

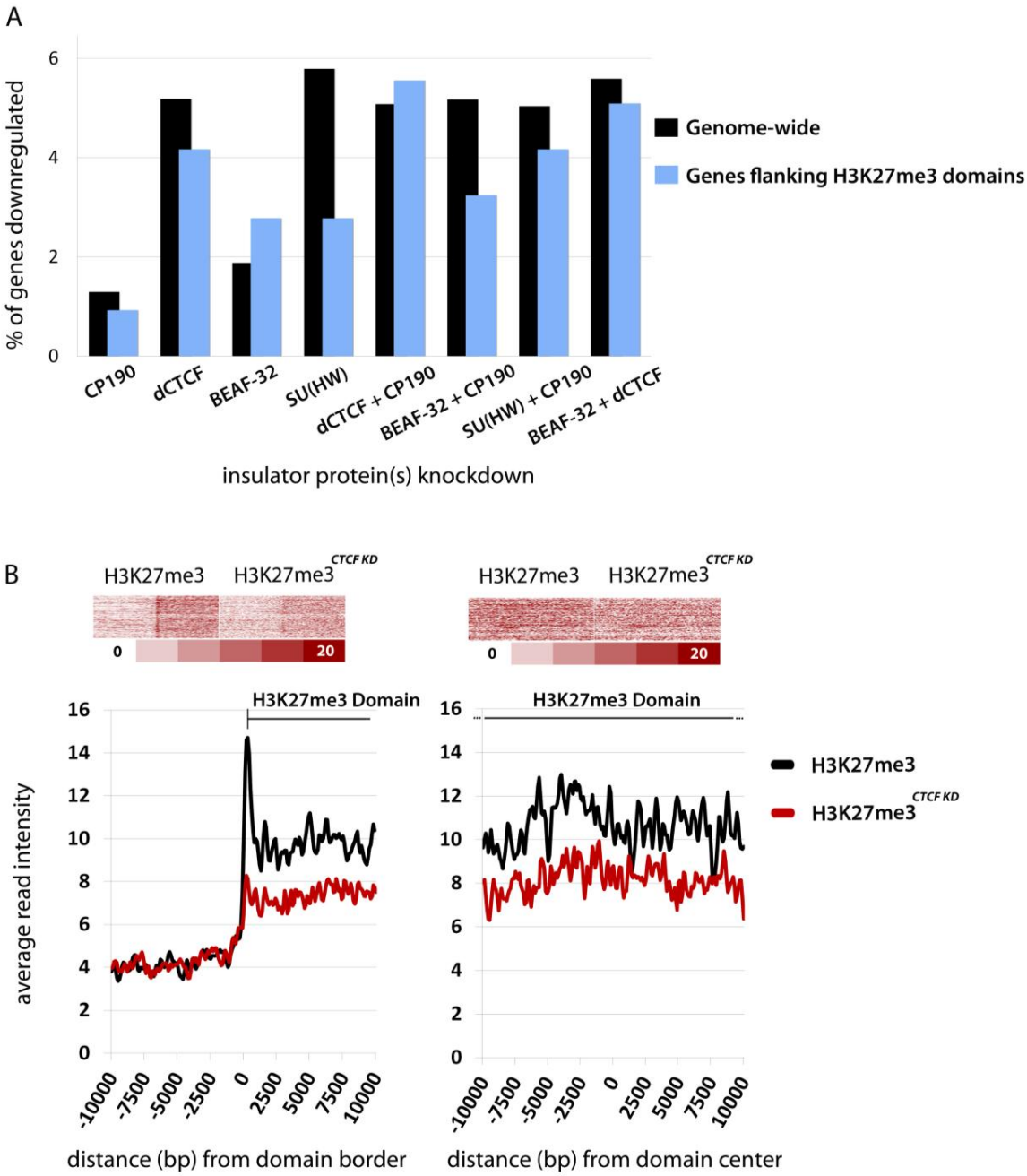


Figure 8. **RNAi depletion of insulator proteins causes reduction in H3K27me3 levels within domains but has no effect on H3K27me3 domain flanking genes**. (A) Percentage of genes downregulated (>2 fold) in *Drosophila* Kc cells after knockdown of insulator protein expression. Comparison of genes flanking H3K27me3 domains with

genome-wide averages. (B) Average read intensity for H3K27me3 in *Drosophila* Kc cells surrounding H3K27me3 domain borders and domain centers, +/- 10 Kb. H3K27me3 levels were determined by ChIP-seq before and after RNAi depletion of insulator protein dCTCF. For appropriate comparison, ChIP-seq data for H3K27me3 was rank normalized as previously described [182], and represented as average read intensity.

unchanged (Supplemental Fig. S7), indicating this is not an indirect consequence of disruption in methyltransferase activity. The reduction in H3K27me3 levels suggests insulators actively play a role important for the maintenance of H3K27me3 levels within Pc domains. Despite reduced H3K27me3 levels, gene expression is relatively unaffected for genes within these domains when insulators are knockdown (Supplemental Fig. S8), meaning the mechanisms underlying gene repression in Pc domains have not been entirely compromised, or that additional steps are necessary for the activation of Pc domain containing genes.

The *even-skipped* gene provides a model for dCTCF alignment at H3K27me3 domain borders

In mammals, broad domains of repressive H3K27me3 characterized by Polycomb have been shown to silence clusters of developmentally important genes [111, 183]. Recent findings have demonstrated similar repression of developmental genes in stable, cell-stage independent H3K27me3 domains in *D. melanogaster* [184]. Genes within H3K27me3 domains are highly enriched for developmental genes in Kc cells, including the *even-skipped* (*eve*) gene (Supplemental Table S1), consistent with previous results. The *eve* gene thus provides an excellent model to analyze the role of tandemly aligned dCTCF sites and chromatin organization.

eve is an early pair-rule gene encoding a homeodomain-containing transcription factor involved in segmentation [185]. Expression of *eve* peaks within the first 6 hr of embryogenesis, and is essentially non-expressed in late embryonic/adult *Drosophila* tissues [186], including late embryonic *Drosophila* Kc cells [25]. *eve* is one of several

hundred genes targeted by Polycomb (Pc), and recent data suggest that Pleiohomeotic, a Pc DNA-binding protein, negatively regulates *eve* during embryogenesis [114, 187]. Analysis of the *eve* locus in *Drosophila* Kc cells reveals H3K27me3 mediated repression in the form of a 15 kb H3K27me3 domain (Supplemental Fig. S9). The domain is flanked immediately downstream by dCTCF aligned with both BEAF-32 and SU(HW), and immediately upstream by a dCTCF site aligned with two SU(HW) elements. In both cases, dCTCF binding sites are marked by the secondary target sequence identified by MEME-ChIP and weeder1.3 [168, 172]. These aligned dCTCF sites overlap with ChIP-seq profiles for CP190, MOD(MDG4), L(3)MBT, and Chromator, consistent with genome-wide enrichments for insulator-associated proteins. Disruption of insulator proteins has no effect on the expression of domain flanking genes CG12134 and TER94 (Supplemental Table S2), nor does it significantly affect the expression of *eve*. However, knockdown of dCTCF results in H3K27me3 depletion within the repressed *eve* domain (Supplemental Fig. S9), and knockdown of additional insulator proteins, including MOD(MDG4), produces similar results (Figure 9A,D). Despite loss of insulators and H3K27me3 depletion, *eve* appears to remain repressed, suggesting insulator proteins contribute to appropriate chromatin domain structure but are not essential for maintenance of gene silencing in these domains. Importantly, this model for insulator alignment at H3K27me3 domain borders is consistent throughout the genome, including early stage developmental gene *eyes absent* (*eya*) and hybrid sterility gene *Odysseus-site homeobox* (*OdsH*) (Figure 9B-C, E-F).

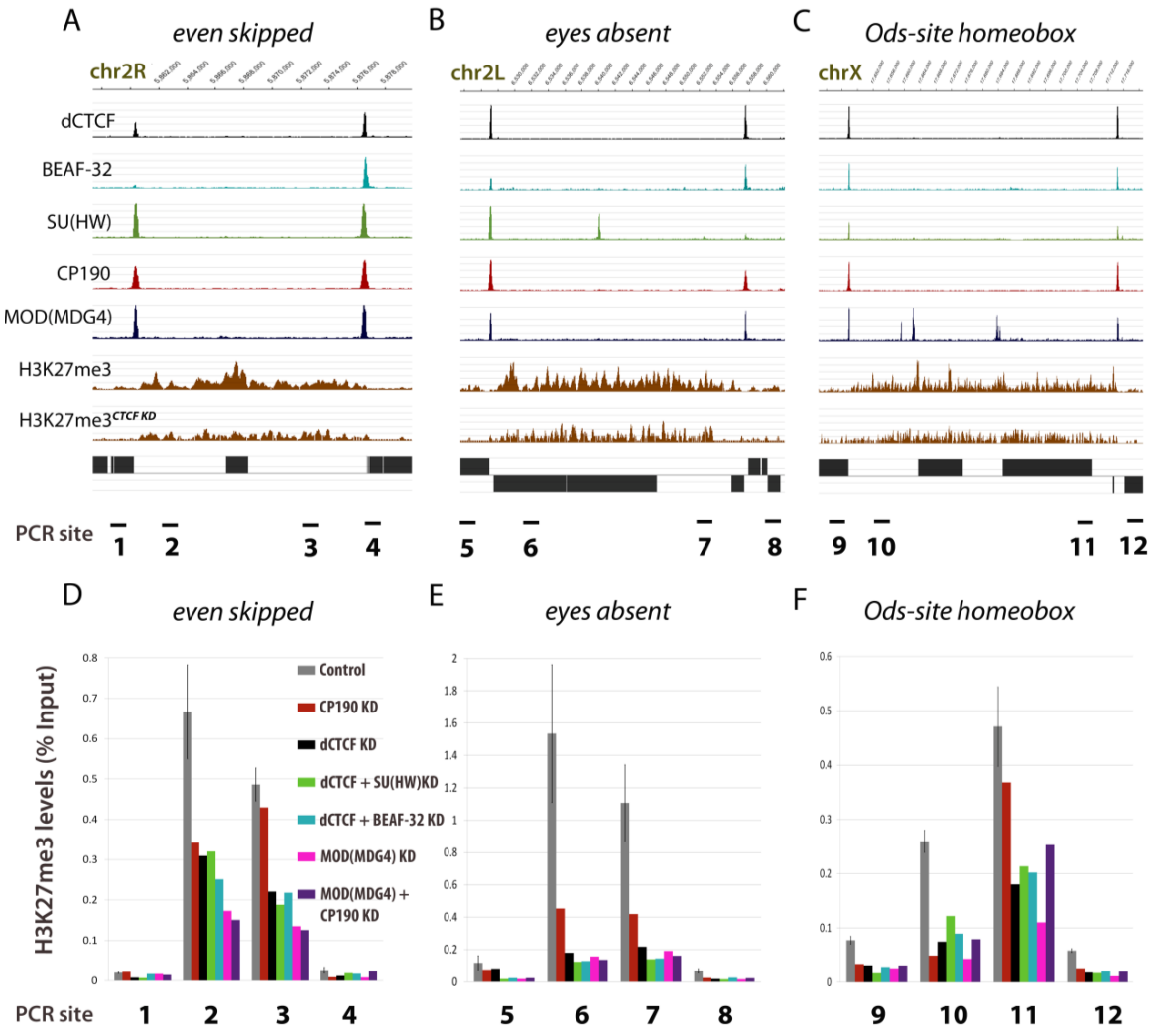


Figure 9. **Insulator proteins, including MOD(MDG4), are necessary for the maintenance of H3K27me3 levels at several loci**. (A-C) ChIP-seq profiles for insulator proteins and H3K27me3 levels before and after RNAi depletion of dCTCF at the *even skipped*, *eyes absent*, and *Odysseus-site homeobox* gene loci respectively. (D-F) ChIP-PCR levels determined by qRT-PCR against H3K27me3 ChIP before and after RNAi depletion of several insulator proteins. Control sample is represented as S.E.M from three independent biological replicates.

DISCUSSION

Improvements in genomic strategies for mapping genome-wide interactions have allowed recent studies to probe basic genome folding principles as well as insulator-mediated chromatin interactions [20, 23, 65, 78]. Results consistently support current models proposing roles for insulator proteins in genome-wide organization [73], and challenge the basic barrier and enhancer-blocking activities that classically defined these proteins. Instead, the ability of insulators to block the spread of heterochromatin and impede enhancer-promoter interactions may simply be consequences of a more paramount role in chromosome organization. New findings in *Drosophila* also suggest that insulators are required to mediate long-range interactions important for Polycomb (Pc) repression [126, 129], and the recent identification of CTCF in transcription factories [140] suggests that insulators may direct the localization of specific genomic loci to discrete nuclear subcompartments for gene regulation [142].

Nevertheless, our finding that heterochromatin does not spread into flanking chromatin domains in response to insulator knockdown is surprising based on numerous examples of insulator mediated barrier function. Though individual insulator elements may indeed serve to prevent the spread of silencing chromatin, our disruption of total insulator protein levels instead significantly affected the levels of H3K27me3 within rather than outside of repressive chromatin domains. Insulator knockdown had no effect on the expression of *E(z)* or total H3K27me3 levels (Supplemental Fig. S7). Therefore, the loss of H3K27me3 within Pc domains genome-wide suggests insulators play a critical role

necessary for the maintenance of appropriate chromatin architecture at these specific loci. Given the requirement for insulators in long-range Pc interactions [126, 129], we speculate that long-range interactions mediated by dCTCF and *Drosophila* insulator proteins are ultimately disrupted by insulator knockdown, and that H3K27me3 depletion likely reflects a defect in Pc mediated compaction and maintenance of H3K27me3 at developmental loci. Interestingly, however, expression of genes within repressive H3K27me3 domains was not significantly affected (Supplemental Fig. S8), suggesting Pc mediated gene silencing was not abrogated, or that additional steps are required to activate these developmental genes. Future studies investigating the role of insulators in Pc-mediated repression, and the effects of insulator knockdown in nuclear organization will provide valuable insight into the relationship between insulator proteins and chromatin architecture.

The diverse activities of CTCF in gene expression and chromatin organization require exploration of the proteins with which it functions and the target sequences associated with specific functions. By combining the resolution conferred by high throughput sequencing (ChIP-seq), with mapping of core target sequences, we provide a stringent but exhaustive map of direct binding sites for *Drosophila* insulators, and extend our previous analyses of dCTCF, SU(HW), BEAF-32, and CP190 to include the insulator protein MOD(MDG4). We show that dCTCF aligns with both the SU(HW) and BEAF-32 insulators, where dCTCF becomes enriched for additional insulator and insulator-associated proteins. The presence of aligned dCTCF sites at the borders of H3K27me3 domains provides an excellent system to query the importance of insulator proteins at the

boundaries of discrete chromatin domains. Recently identified correlations for insulator proteins at the boundaries of physical domains mapped in *Drosophila melanogaster* [65] provides evidence for why only a subset of aligned dCTCF localize to H3K27me3 domain borders (Figure 7), and clearly demonstrate that insulators are also involved in the organization of other, distinct chromatin domains. Whereas Pc-repressed domains are relatively easily identifiable in the form of H3K27me3 signatures, future characterization of discrete physical domains and domain boundaries will require genome-wide interrogation of chromosome interactions in individual cell-types of interest. Nearly 40% of aligned dCTCF sites (~355) localize to physical domain boundaries mapped in late embryos (Figure 7) [65], suggesting physical domains and insulator localization may be conserved at many loci across cell-types.

Interestingly, dCTCF appears to target three different target sequences in *D. melanogaster*, including the highly conserved core motif for which dCTCF has been described as binding in both *Drosophila* and mammals [31]. The secondary motif appears highly similar to the conserved core consensus (AGGNGGC) with an insertion between the first pair of guanines (AGTGTGGC), and average dCTCF levels suggest this represents a low occupancy and potentially lower affinity binding site. These novel dCTCF sites are highly enriched for insulator protein CP190 when compared to its primary target sequence. This finding, combined with previous data indicating CP190 is essential for dCTCF binding to a subset of its binding-sites, suggests that CP190 might facilitate dCTCF binding to these secondary sites. The absence of CP190 in vertebrates may explain why these sequences have not been identified as mammalian target

sequences, raising the possibility that these binding sites are a *Drosophila* specific phenomenon.

Analysis of dCTCF insulator alignment at the *eve* locus and genome-wide uncovers a tight association with BEAF-32 and SU(HW), which may provide dCTCF with numerous advantages for effectively establishing a functional insulator. First, alignment of multiple insulator DNA elements may increase the likelihood of sequence accessibility at important loci, as insulator binding sites have been characterized by reduced nucleosome density [45]. For example, an insulator-binding protein may access its cognate sequence, thereby creating an accessible landscape for other, potentially different insulator proteins to bind their respective targets. Second, by aligning in close proximity, recruitment of essential insulator proteins (i.e. CP190 and MOD(MDG4)) by one insulator-binding protein may facilitate recruitment by others, given that CP190 and MOD(MDG4) may be recruited as multimers. Third, given that dCTCF binds secondary sites that potentially require CP190, recruitment of CP190 by a neighboring insulator (i.e. SU(HW) or BEAF-32) may preclude dCTCF binding, thereby providing a regulatory step in dCTCF recruitment to DNA. Finally, by aligning with SU(HW) and BEAF-32, dCTCF establishes a unique identity compared to independent dCTCF sites, where it becomes enriched for additional cofactors, including L(3)MBT and Chromator (Figure 10).

Though our data shed new and valuable insight into what appears to be cooperative insulator function in *Drosophila melanogaster*, many questions remain. Given current models that insulators function via intra- and inter-chromosomal interactions, it is

plausible that aligned dCTCF sites and their enrichment for CP190 and MOD(MDG4) allow for stable chromosomal interactions. Current locus- and genome-wide interaction assays may effectively answer this question in the near future. Though BEAF-32 has been defined as lineage specific [188], and SU(HW) appears to lack a counterpart in mammals, our results suggest that mammalian CTCF may align with other, unique DNA-binding proteins important for appropriate insulator function at the boundaries of Pc domains.

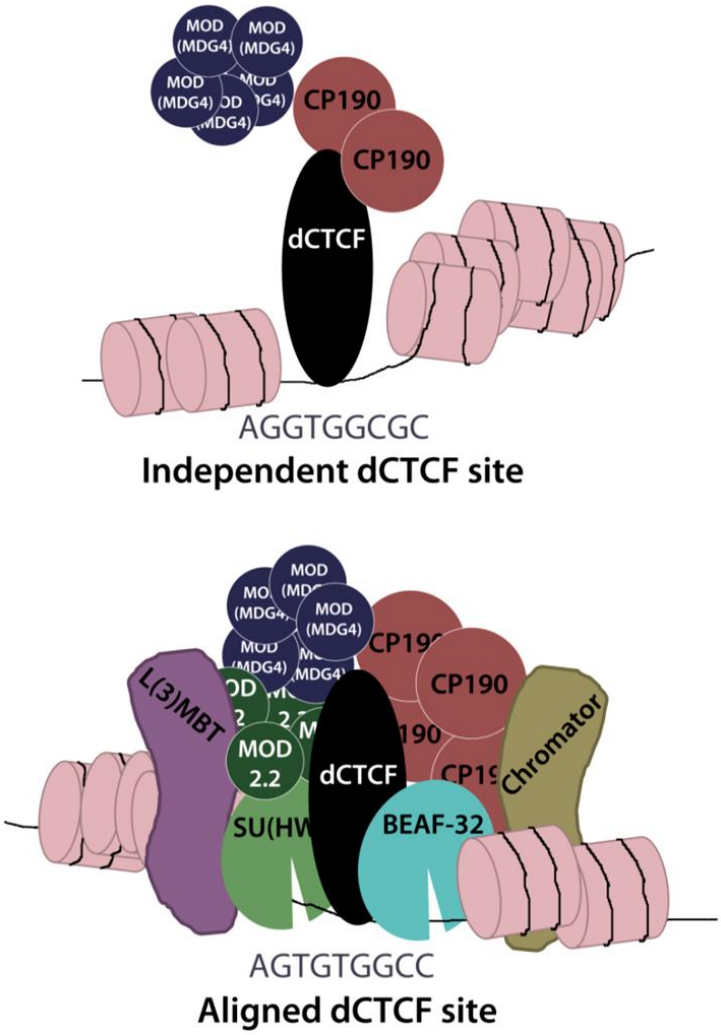


Figure 10. **Diagram comparing independent dCTCF sites and dCTCF sites aligned with BEAF-32 and SU(HW)**. dCTCF, BEAF-32, and SU(HW) function similarly through the recruitment of CP190 and MOD(MDG4). Aligned dCTCF sites are enriched for the secondary DNA sequence and CP190. Ultimately, alignment may allow for cooperative recruitment of CP190 and MOD(MDG4), ensuring dCTCF establishes a functional insulator complex at domain borders. Recruitment of additional proteins, such as L(3)MBT and Chromator, may also contribute to insulator activities at these loci.

EXPERIMENTAL PROCEDURES

ChIP-seq

Chromatin immunoprecipitation was performed as previously described [58]. For Re-ChIP assays, chromatin was eluted in 1% SDS, 0.1 M $NaHCO_3$, diluted 10-fold in IP dilution buffer (0.01% SDS, 1.1% Triton X-100, 1.2 mM EDTA, 16.7 mM Tris-HCl, 167 mM NaCl) and ChIP repeated using antibodies against BEAF-32 or SU(HW). ChIP of MOD(MDG4) was carried out with antibodies against the mod2.2 isoform (α-Rabbit; gift from Elissa Lei) and against the region shared by all isoforms as previously described (Bushey et al., 2009). ChIP for L(3)MBT in *Drosophila* Kc cells was carried out using a previously described antibody (α-Guinea-pig; gift from Jurgen Knoblich) [105], and ChIP against H3K27me3 was performed using a commercially available polyclonal antibody (Millipore Cat# 07-449). To generate sequencing libraries, ChIP DNA was prepared for adaptor ligation by end repair (End-It DNA End Repair Kit, Epicenter Cat# ER0720) and addition of "A" base to 3′ ends (Klenow 3′-5′ exo–, NEB Cat# M0212S). Illumina adaptors (Illumina Cat# PE-102-1001) were titrated according to prepared DNA ChIP sample concentration and ligated with T4 ligase (NEB Cat# M0202S). Ligated ChIP samples were PCR-amplified using Illumina primers and Phusion DNA polymerase (NEB Cat# F-530L) and size selected for 200–300 bp by gel extraction. ChIP libraries were sequenced at the HudsonAlpha Institute for Biotechnology, using an Illumina HiSeq 2000. Sequences were mapped to the dm3 genome with Bowtie 0.12.3 [189] using default settings. Peaks were then called with MACS 1.4.0alpha2 [190] using equal numbers of unique reads for input and ChIP samples and a p value cutoff of 1×10^{-10}.

ChIP-seq and bioinformatics analyses

Previously published ChIP-seq data are available from GEO accession GSE30740 (Wood et al., 2011). DNA sequence motifs present in binding sites for dCTCF, BEAF-32, and SU(HW) were identified using commonly called peaks from three independent biological samples (and thus represent insulator binding sites of highest confidence), *Drosophila* Kc cells treated with ecdysone at 0 hr, 3 hr and 48 hr [77]. Primary motifs were identified by MEME-ChIP using default settings (Machanick and Bailey 2011). dCTCF motif 2 was identified in both MEME-ChIP and Weeder 1.3 [172], and motif 3 in MEME-ChIP by excluding peaks containing the primary conserved motif. Insulator peaks were then trimmed to include only those containing core consensus sequences for each protein using ambiguity codes specified by MEME-ChIP. For dCTCF this included both the motifs described (AG[GA][TG]GGCGC (allowing for one mutation), [AG]GTGT[GT][GA]CC (allowing for one mutation), and GGT[TG][TGC][GA][TA][GA][TA]C[TC][TC][CGT]GCTA (allowing for one mutation). For BEAF-32, this included the previously identified motif [ATG][TGC]CGATA with no mutations allowed and, for SU(HW), the motif GC[AC]TA[CT]TTT allowing for one mutation. Direct insulator binding sites were thus finally called as summits identified by MACS in three independent biological samples +/- 150bp that contain the described consensus sequence specific for each insulator protein. Overlap between insulators and associated proteins were identified using publicly available tools on Galaxy (Goecks et al., 2010; Blankenburg et al., 2010; Giardine et al., 2005).

H3K27me3 domains were called using H3K27me3 ChIP-seq data obtained here in *Drosophila* Kc cells, with comparison to publicly available H3K27me3 ChIP-seq data in late embryos and the added requirement for Polycomb occupancy/signal in *Drosophila* Kc cells [25]. Domain borders were called as 0^{th} nucleotide of peaks called, and organized by K-means clustering by Cluster 3.0 (de Hoon et al., 2004). Genes within H3K27me3 domains were called as intersecting (>300 bp) H3K27me3 domains using publicly available tools on Galaxy [191-193]. Comparisons between histone H3K27me3 before and after dCTCF knockdown were performed after rank order normalization, as recently described [182]. Briefly, these ChIP-seq datasets are rank-ordered in 10bp bins across the *Drosophila* genome, from highest to lowest read intensity. Averages between the two datasets are then assigned to each bin – from highest to lowest read values.

Enrichments for insulator-associated proteins at aligned dCTCF clusters were calculated as percentage of co-occurrence between dCTCF and BEAF-32, SU(HW), CP190, MOD(MDG4), MOD(MDG4)2.2, L(3)MBT, and Chromator at independent dCTCF sites, aligned dCTCF sites, and sites hosting dCTCF + BEAF-32 + SU(HW). Results were hierarchically clustered using cluster 3.0 [194] and visualized by Java Treeview [195].

Immunofluorescence microscopy

Immunofluorescence microscopy of polytene chromosomes was done as previously described [196]. Cells were stained with primary antibodies in antibody dilution buffer (1

× PBS, 0.1% Tween20, 1% BSA) overnight at 4°C (1:100 rabbit α-MOD(MDG4)2.2) (gift from Elissa Lei), 1:100 rat α-MOD(MDG4) (Pai et al., 2004).

Real-time PCR analysis

Real-time PCR analyses for H3K27me3 levels in insulator knockdown experiments and Re-ChIP were performed with independent ChIP samples. Fermentas Life Sciences Maxima qPCR SYBR Green ROX Mix (#K0222) was used and percent input was calculated with a three-point standard curve from the input sample. ChIP DNA and input DNA concentrations were calculated using a Qubit 2.0 fluorometer HS assay (Invitrogen Q32851). ChIP DNA concentrations were consistently lower in insulator knockdown conditions, and thus normalized by equal ChIP/input DNA ratios before qRT-PCR. Primers used for both analyses are provided in Supplemental Table S3.

Gene Expression Analyses

RNAi knockdown in *Drosophila* Kc cells culture was conducted as per the *Drosophila* RNAi Screening Center (DRSC) protocol [197] with the exception that dsRNA was added every day for 3 days and the cells were then collected on the 4^{th} day. Also, multiple amplicons targeting each gene for knockdown were used with the exception of BEAF-32, which only used one. RNA was then isolated from the Kc cells using the Qiagen RNeasy kit (catalog #74104) with on-column DNA digestion (catalog #79254) following manufacture's protocols. cDNA synthesis was then performed using the Applied Biosystems High Capacity cDNA Reverse Transcription Kit (catalog #4368814). cDNA was hybridized to a NimbleGen *D. melanogaster* Gene Expression 12X135K Array at the

Florida State University –NimbleGen Microarray Facility. Expression analysis of variance was performed using Partek® software, version 6.5. A list of primers used for amplicon formation are provided in Supplemental Table S4.

Data Access

Gene expression and ChIP-seq data has been made publicly available under GEO accession numbers GSE36944 (L(3)MBT and MOD(MDG4)2.2/MOD(MDG4) in *Drosophila* Kc cells), GSE37444 (H3K27me3 in *Drosophila* Kc cells – control and dCTCF knockdown) and GSE36393 (Gene expression in *Drosophila* Kc cells before and after insulator knockdown)

ACKNOWLEDGEMENTS

We would like to thank Dr. Elissa Lei and Dr. Jurgen Knoblich for sharing antibodies. We also thank The Genomic Services Lab at the HudsonAlpha Institute for Biotechnology for their help in performing Illumina sequencing of ChIP-Seq samples. This work was supported by U.S. Public Health Service Awards to VC (GM35463) and ER (CA133106) from the National Institutes of Health. Work presented in this chapter is published in: Genome Res. 2012 Nov;22(11):2176-87. doi: 10.1101/gr.136788.111

CHAPTER 3:

INSULATOR FUNCTION AND TOPOLOGICAL DOMAIN BORDER STRENGTH SCALE WITH ARCHITECTURAL PROTEIN OCCUPANCY

Work presented in this chapter is published in:

Van Bortle K, Nichols MH, Li L, Ong CT, Takenaka N, Qin ZS, Corces VG. Genome Biol. 2014. Insulator function and topological domain border strength scale with architectural protein occupancy. Jun 30;15(6)

My contributions include conceiving the project, performing the experiments, computational analysis of the data, interpretation of the results, and writing the paper.

ABSTRACT

Chromosome conformation capture studies suggest that eukaryotic genomes are organized into structures called topologically associating domains. The borders of these domains are highly enriched for architectural proteins with characterized roles in insulator function. However, a majority of architectural protein binding sites localize within topological domains, suggesting sites associated with domain borders represent a functionally different subclass of these regulatory elements. How topologically associating domains are established and what differentiates border-associated from non-border architectural protein binding sites remain unanswered questions. By mapping the genome-wide target sites for several *Drosophila* architectural proteins, including previously uncharacterized profiles for TFIIIC and SMC-containing condensin complexes, we uncover an extensive pattern of colocalization in which architectural proteins establish dense clusters at the borders of topological domains. Reporter-based enhancer-blocking insulator activity as well as endogenous domain border strength scale with the occupancy level of architectural protein binding sites, suggesting co-binding by architectural proteins underlies the functional potential of these loci. Analyses in mouse and human stem cells suggest that clustering of architectural proteins is a general feature of genome organization, and conserved architectural protein binding sites may underlie the tissue-invariant nature of topologically associating domains observed in mammals.

INTRODUCTION

The recent development of high-throughput chromosome conformation capture (3C)-based molecular techniques have propelled our understanding of three-dimensional chromosome organization to new heights. In particular, the organization of eukaryotic genomes into discrete physical domains can now be defined by surveying genome-wide pairwise interaction frequencies. A series of such analyses in *Drosophila*, mice, and humans have shed insights into the hierarchical organization of interphase chromosomes on different length scales, and raise additional questions into the mechanisms governing three-dimensional genome organization [65-67, 198-203]. During interphase, genomes are partitioned into sub-Megabase (Mb) length Topologically Associating Domains (TADs), which are further organized into multi-Mb sized structures called compartments, whose distribution often reflects cell-type specific expression patterns [204]. In contrast, TAD structure is generally consistent between diverse cell types [66], suggesting the sub-Mb scale arrangement of chromosomes may represent a conserved, bottom-up pattern of chromatin organization and genome function. Thus, understanding how TADs are established and maintained between cell-types remains an important question.

Integration of long-range interaction frequencies and domain organization with genomic annotations along the linear genome has revealed a strong relationship between TAD borders and proteins associated with insulator function. For example, CTCF (CCCTC-binding factor) as well as tRNA genes, recently shown by transgene protection assays to possess classical insulator activity in humans [26, 27], are significantly enriched in

regions separating topological domains [66]. Nevertheless, 85% of CTCF binding sites localize within rather than at the borders of TADs, suggesting a majority of CTCF sites are unrelated to the formation of borders that separate TADs. Meanwhile, multiple studies suggest that many insulator elements are not capable of enhancer-blocking or chromatin barrier activity at all [46, 64, 205], and may instead be reserved for other activities such as gene repression, activation, or enhancer-promoter interactions [206-208]. The seemingly contradictory activities of insulators and the dichotomy of border-associated versus non-border target sites suggest that the very use of the name "insulator" is, in most cases, erroneous. To avoid further sustaining this confusion, we here on out refer to proteins associated with insulator function as architectural proteins, and refer to insulators only in the context of elements capable of enhancer-blocking activity.

To date, several architectural proteins have been identified in *D. melanogaster*, including the *Drosophila* homolog of CTCF (dCTCF), Suppressor of Hairy-wing (Su(Hw)), GAGA factor (GAF), and the scs and scs' boundary proteins Boundary Element Associated Factor of 32 kDa (BEAF-32) and Zeste White 5 (Zw5) [209]. Phylogenetic analyses in *Drosophila* suggest however that all but dCTCF and Su(Hw) were successively gained during arthropod evolution [210], and that additional and perhaps unexplored architectural proteins may supplement the highly conserved CTCF protein in vertebrates. Supporting evidence for this possibility comes from recent genome-wide mapping studies of the multisubunit Pol III transcription factor TFIIIC, which is essential for the inherent insulator activity of tRNA genes in yeast [49]. In mammals, TFIIIC often binds to Pol III-independent regions, called extra TFIIIC (ETC) loci, in close proximity to CTCF [55,

56]. TFIIIC binding sites also associate with the cohesin complex in mammals [55], and can also underlie condensin loading onto chromosomes in *S. cerevisiae* [57], strongly suggestive of a role in chromatin organization. Understanding the function of TFIIIC and its relationship to other architectural proteins may therefore shed light on the mechanisms by which these proteins contribute to the three-dimensional organization of the genome in the nucleus.

Here we present the first genome-wide characterization of TFIIIC in *D. melanogaster* and find that this protein localizes to sites combinatorially bound by several *Drosophila* architectural proteins. These high occupancy APBSs localize to the borders of TADs, are enriched for both the cohesin and condensin complexes, and represent highly accessible regions of chromatin that are stable throughout *Drosophila* development, consistent with the tissue-invariant nature of TADs observed in mammals. The relative occupancy of architectural proteins at APBSs scales with the strength of TAD borders, as well as the capacity of these elements to function as enhancer-blocking insulators in transgenic reporter assays, suggesting the composition of these regulatory elements underlies a spectrum of regulatory potential. Finally, we uncover a similar relationship between TFIIIC, CTCF, cohesin, condensin and TADs in mice and humans, suggesting a conserved role for clustered architectural proteins in sub-Mb scale chromatin domain organization.

RESULTS

Characterization and genome-wide mapping of TFIIIC in *D. melanogaster*

TFIIIC targets sequence-specific gene-internal A box and B box promoter elements present in a subset of Pol III-transcribed genes [211], where it then recruits the transcription factor complex TFIIIB. Biochemical and molecular characterization of TFIIIC has revealed evolutionary changes in protein structure and protein-protein interactions between yeast and humans, yet the subunit composition is generally conserved [212]. In *D. melanogaster*, protein-coding gene *CG7099* (Flybase FBgn0032517) is predicted to encode a B box binding subunit of TFIIIC based on protein sequence homology. Immunoblot and immunofluorescence localization of CG7099, which we now refer to as dTFIIIC220, confirms an antigen-specific protein at the predicted molecular weight (~220 kDa), which localizes to numerous binding sites in polytene chromosomes throughout the *Drosophila* genome (Additional File 1 Figure S1A-F).

We performed ChIP-seq against dTFIIIC220 in Kc167 cells as recently carried out for several DNA-binding factors [46, 77]. Genome-wide analysis confirms the localization of dTFIIIC220 to tRNA genes and sites associated with the TFIIIB complex as expected (Figure 11A-C), and MEME-ChIP and CentriMo consensus sequence analysis further demonstrates central motif enrichment for both the *Drosophila* A box and B box elements in our ChIP-seq experiments (Figure 11D-E) [168, 213]. dTFIIIC220 binding sites determined by the commonly used MACS peak calling algorithm [190] are present at a majority of annotated tRNA genes obtained from Flybase (Figure 11F) [214], and

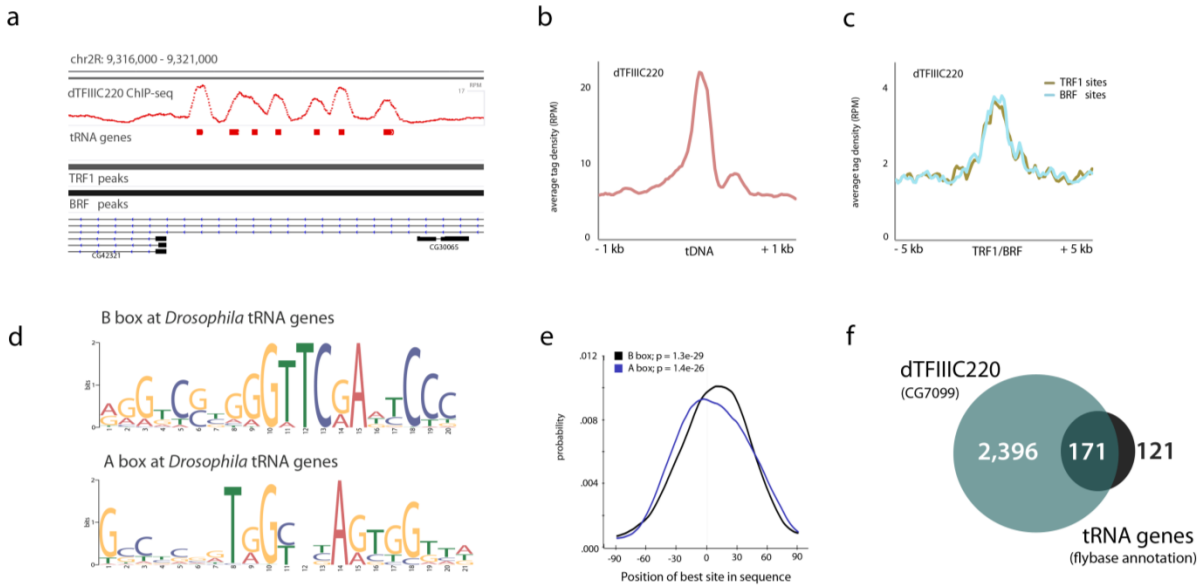


Figure 11. **Genome-wide mapping of dTFIIIC220 in *D. melanogaster*. (A)** Example ChIP-seq profile shown for dTFIIIC220 (red) over a tRNA cluster on *Drosophila* chromosome 2R, co-bound by TFIIIB subunits TRF1 and BRF. **(B-C)** Tag density enrichment profiles for dTFIIIC220 over all annotated tRNA genes and over sites previously identified as bound by TFIIIB complex subunits TRF1 and BRF confirms the expected genome-wide localization patterns for *Drosophila* (overlap significance P < 0.00001, permutation test) **(D)** Consensus sequences identified *de novo* by MEME-ChIP reveals evolutionarily conserved *Drosophila* B box and A box elements present in dTFIIIC220-bound tRNA genes. **(E)** Central motif enrichment (CentriMo) plot for B box and A box sequences with respect to dTFIIIC220 ChIP-seq peaks at tRNA genes. **(F)** Overlap between dTFIIIC220 peaks, independently identified in two biological replicates at FDR 5%, with annotated tRNA genes obtained from Flybase (P < 0.00001, permutation test). Non-overlapping sites indicate thousands of Extra TFIIIC (ETC.) loci

in *D. melanogaster*, of which 348 contain the B box binding motif (14.5%, $P < 0.00001$, permutation test).

---

dTFIIIC220 reads are significantly enriched over all annotated tRNA genes and TFIIIB subunit (TRF1 and BRF) binding sites (Figure 11B-C). In addition to tRNA genes, we identify numerous Extra TFIIIC (ETC) loci (Figure 11F) independent of tRNA gene structure or TFIIIB localization, suggesting *Drosophila* TFIIIC may also function at sites independent of Pol III transcription, as is the case for TFIIIC in mammals [55, 56].

Relationship to SMC-containing cohesin and condensin complexes

Mammalian CTCF recruits and depends on cohesin for functional insulator activity [34 , 35, 63], and original tDNA-based insulator studies in *S. cerevisiae* observed an analogous dependency on SMC proteins [49]. TFIIIC-bound B box elements can also constitute functional loading sites for the condensin complex in *S. cerevisiae* [57], and multiple studies have described a role for condensin in the organization of dispersed Pol III genes in *S. pombe* [215, 216], suggesting TFIIIC activity is tightly associated with SMC complexes. We therefore mapped the genomic binding profile for cohesin and the paralogous condensin complexes via complex specific α-kleisin subunits Rad21 (cohesin), Barren (condensin I), and CAP-H2 (condensin II) to better understand their possible relationship to dTFIIIC220 in *D. melanogaster*.

Analysis of the cohesin and condensin binding profiles in *Drosophila* Kc167 cells reveals substantial overlap between the three SMC-containing complexes (Figure 12A). Further comparison with dTFIIIC220 indicates strong co-localization at ETC loci, particularly for the cohesin and condensin II complexes (Figure 12B), suggesting that association with

cohesin and condensin at TFIIIC sites is conserved in *Drosophila*. Additionally, we find that whereas condensin I is most pronounced at tRNA genes (Figure 12C-D), consistent with recent condensin mapping studies in vertebrate chicken DT40 cells [217], both cohesin and condensin II are present at higher levels at ETC loci (Figure 12C, E). This distinction in cohesin and condensin association suggests a unique specialization of SMC complex recruitment to TFIIIC binding sites, possibly underlying differences in co-factor colocalization patterns and function. We therefore next sought to characterize ETC loci and their potential role in genome function.

Previous genome-wide mapping studies in *Drosophila* and mammalian cells have shown that cohesin often localizes to highly occupied cis regulatory modules that may function as developmental or cell-type specific enhancers [218-221], and both cohesin and condensin II localize to super-enhancers reported to be involved in controlling mammalian cell identity [222]. We thus compared the profiles for TFIIIC, cohesin, and condensin complexes with 1,311 previously reported enhancers characterized by DNase I hypersensitivity and enhancer hallmarks H3K4me1 and H3K27ac in Kc167 cells (Figure 12F) [223]. A large majority of enhancers are bound by the cohesin complex (Figure 12G) and, unlike RNA polymerase II (Pol II), cohesin is more significantly enriched at individual enhancers than transcription start sites (Additional File 2, Figure S2A). However, very few enhancers are bound by dTFIIIC220, and fewer yet associate with the condensin complexes (Figure 12G), suggesting sites co-bound by TFIIIC and SMC complexes generally do not represent active enhancers.

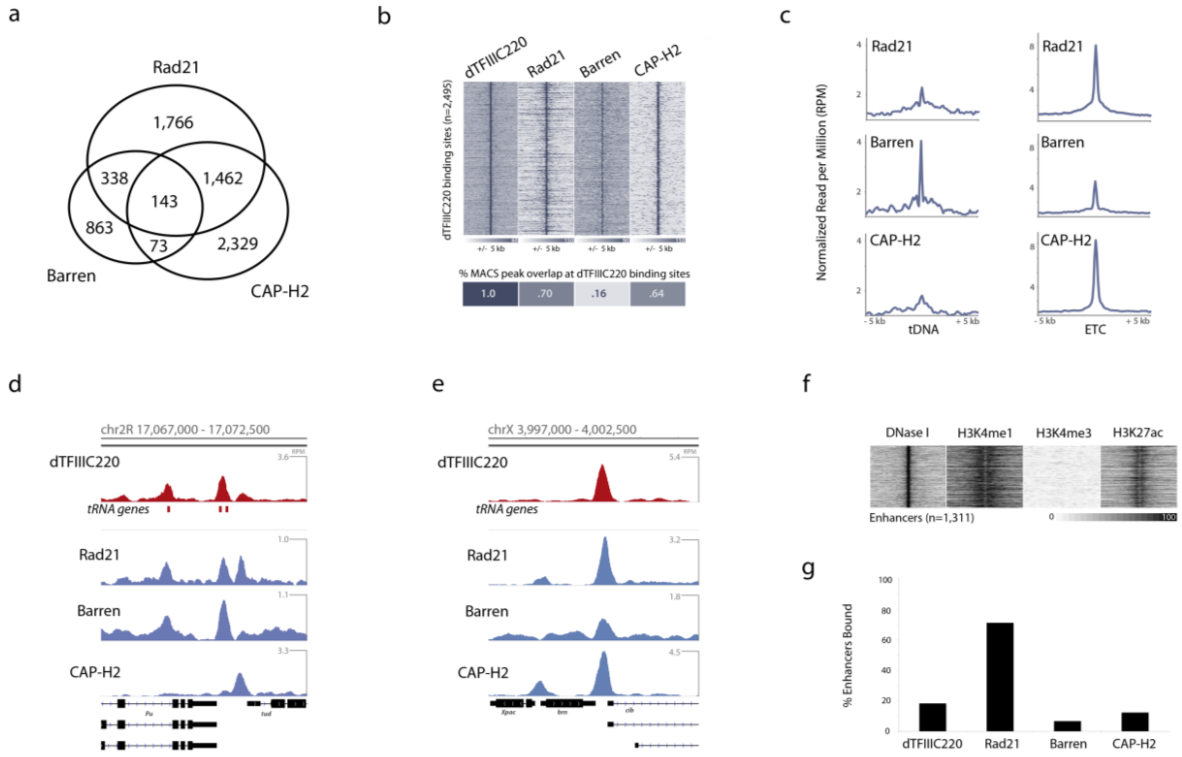


Figure 12. **SMC-containing cohesin and condensin complexes localize to a subset of tDNAs and extra TFIIIC (ETC.) loci. (A)** Number of overlapping peaks identified by ChIP-seq against α-kleisen subunits Rad21 (cohesin), Barren (Condensin I), and CAP-H2 (Condensin II) in Kc167 cells. ($P < 0.00001$ for overlap between Rad21 with CAP-H2 or Barren, permutation test) **(B)** Heatmap representation shown for ChIP-seq read intensities of SMC-containing complexes and TFIIIC subunit dTFIIIC220, anchored across all dTFIIIC220 peaks (top), plus or minus 5 kb. Heatmap representation (bottom) of overlap frequencies between dTFIIIC220 peaks and those of SMC-containing complexes (overlap significance for dTFIIIC220 with each factor $P < 0.00001$, permutation test). **(C)** Read intensity plots for Rad21, Barren, and CAP-H2 at TFIIIC-bound tDNAs (left) and ETC. loci (right) plus or minus 5 kb. Tag density is represented as rank-order normalized reads per million (RPM) across all three ChIP-seq experiments. **(D-E)** Example genomics

viewer profiles of overlapping dTFIIIC220 sites at tRNA genes and ETC. loci. **(F)** Heatmap representation shown for DNase-seq and ChIP-seq read intensities at 1,311 active enhancers previously defined by STARR-seq, and marked by active enhancer characteristics in the Kc167 cell line – including DNase I hypersensitivity, H3K4me1 and H3K27ac. **(G)** Percentage of enhancers bound by dTFIIIC220 and SMC-containing complexes.

TFIIIC clusters with CTCF and other *Drosophila* architectural proteins

Visual inspection of dTFIIIC220 ChIP-seq data instead suggests that TFIIIC target regions coincide with sites marked by previously characterized architectural proteins. In particular, dTFIIIC220 binding sites often localize to regions combinatorially bound by several factors shown to associate with insulator activity (Figure 13A). These high occupancy architectural protein binding sites (APBSs) also correlate with SMC-containing cohesin and condensin complexes, consistent with the strong correlation observed with TFIIIC. This finding is surprising, however, as previous ChIP-chip studies mapping an ancillary cohesin subunit, Scc3, observed a relatively weak overlap with dCTCF [68] which, like BEAF-32 and Su(Hw), recruit BTB-containing proteins CP190 and Mod(mdg4) essential for insulator activity [43, 44, 167]. These original observations have led to speculation that *Drosophila* CTCF functions through a unique mechanism compared to its mammalian counterpart, yet our genome-wide high resolution profile of Rad21 suggests a more extensive co-localization between CTCF and cohesin in *Drosophila*. For example, nearly half of all high confidence CTCF binding sites identified in three biological replicates correlate with Rad21, similar to numbers originally identified in vertebrate HeLa cells [35], and Rad21 ChIP enrichment is significantly greater at APBSs than at independent loci (Additional file 2, Figure S2A-D). Furthermore, depletion of dCTCF by RNAi in Kc167 cells disrupts Rad21 localization specifically to dCTCF binding sites (Additional File 2, Figure S2E-G), suggesting recruitment of cohesin is conserved from *Drosophila* to mammals.

Genome-wide, dTFIIIC220, Rad21, and CAP-H2 strongly associate with combinatorially-bound APBSs, independently determined to contain four or more previously characterized architectural proteins. Hierarchical clustering of overlap frequencies observed between TFIIIC, SMC-complexes and defined transcription factor binding sites in Kc167 cells illustrates this relationship, wherein dTFIIIC220, Rad21, and CAP-H2 cluster with architectural proteins at these loci (Figure 13C). For example, out of 3,728 combinatorially bound APBSs, 1,489 (40%), 2,124 (57%), and 1,830 (49%) are associated with dTFIIIC220, CAP-H2, or Rad21, respectively ($P<0.00001$, permutation test). We observe a comparatively weak overlap with transcription factor binding sites identified in Kc167 cells (Figure 13C), suggesting colocalization patterns observed for these architectural proteins are different from transcription factor hotspots. However, the enrichment of Rad21 at high occupancy APBSs is intriguing, as cohesin was recently shown to maintain high occupancy transcription factor clusters in mammals [219, 220].

In order to determine whether TFIIIC might directly interact with *Drosophila* architectural proteins, dTFIIIC220-associated complexes were isolated by immunoaffinity purification (Figure 13D). Western analysis of control preimmune and α-dTFIIIC220 immunoaffinity purifications suggests that TFIIIC associates with both CP190 and Mod(mdg4), as is the case for other *Drosophila* architectural proteins [43, 44, 46, 167]. Although a comparatively weak interaction is detected with BEAF-32, dTFIIIC220 does not appear to directly associate with dCTCF or Su(Hw), and we could not detect an interaction with Rad21, suggesting the dTFIIIC220 subunit may not directly recruit cohesin via α-kleisin subunit Rad21. Nevertheless, interactions with CP190 and

Mod(mdg4) extend a common theme observed for proteins associated with insulator function in *D. melanogaster* to TFIIIC, suggesting BTB-containing proteins may also represent a unifying mechanism for both long-range interactions as well as co-occupancy at these sites.

Clustering of architectural proteins scales with TAD border strength

Analyses of TADs in *D. melanogaster* consistently demonstrate that architectural proteins are highly enriched at boundary regions flanked by two adjacent domains [65, 198]. We therefore sought to define high occupancy APBSs by cross-analyzing ChIP-seq data against dTFIIIC220, Rad21, CAP-H2, dCTCF, BEAF-32, Su(Hw), CP190, Mod(mdg4), the transcription factor DREF [224], the chromo-domain protein Chromator, previously shown to colocalize and coimmunoprecipitate with BEAF-32 [179], and the tumor suppressor L(3)mbt protein, recently shown to localize specifically to *Drosophila* APBSs [105] (List provided in Additional File 3). We further classified overlapping binding sites based on the number of overlapping proteins into sites with high, medium, or low occupancy (See Additional File 4, Fig. S3A). High occupancy APBSs correlate with regions associated with strong DNase I hypersensitive sites (DHSs) [225], and associate with increasing DHS intensity as measured by DNase-seq in Kc167 cells (Additional File 4, Figure S3B), suggesting APBSs represent open chromatin regions whose accessibility increases with increasing cofactor occupancy. Analysis of the location of APBSs with respect to gene structure indicates that high occupancy sites are more likely to reside in regions that are upstream and proximal to transcription start sites,

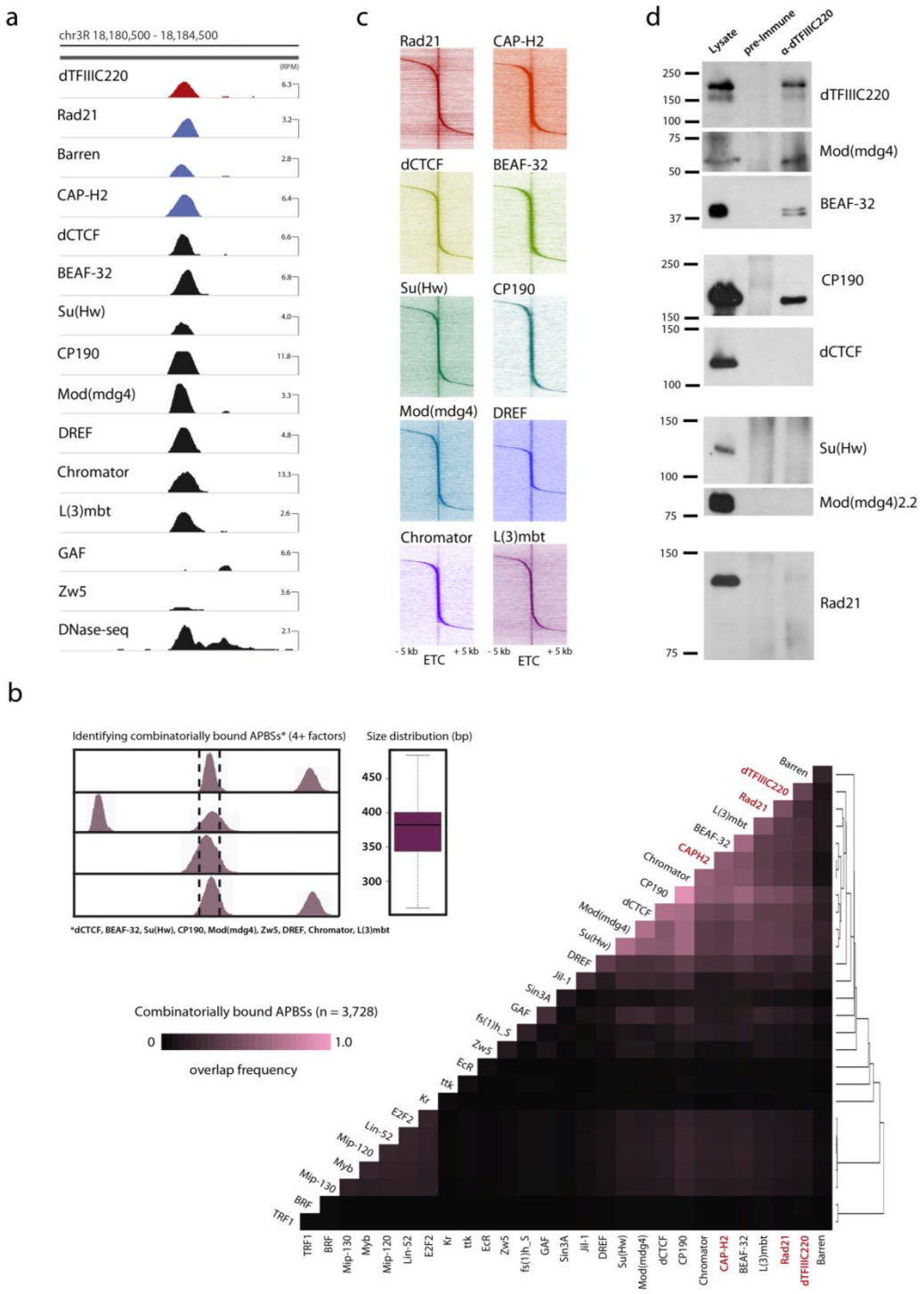


Figure 13. ***Drosophila*** **TFIIIC clusters with CTCF at sites combinatorially bound by architectural proteins, cohesin, and condensin II. (A)** Example genomics viewer profile of a combinatorially bound APBS, co-bound by dTFIIIC220, SMC-containing

cohesin and condensin complexes, dCTCF, BEAF-32, Su(Hw), CP190, Mod(mdg4), DREF, Chromator, L(3)mbt, and marked by strong DNase I hypersensitivity. **(B)** Heatmap representation of co-factor co-localization at 3,728 genomic loci combinatorially bound by architectural proteins. Overlap frequency is the fraction of combinatorially bound loci bound by each individual factor. Inset: sites were identified as genomic fragments having 4 or more proteins in Kc167 cells using MACS called summits +/− 200 bp for factors dCTCF, BEAF-32, Su(Hw), CP190, Mod(mdg4), Zw5, DREF, Chromator, and L(3)mbt, and mapped independently of TFIIIC and SMC-complexes; size distribution (bp) of combinatorially bound loci. $P < 0.00001$ for overlap between combinatorially bound loci with dTFIIIC220, Rad21, and CAP-H2, permutation test. Overlap frequency matrix hierarchically clustered (absolute centered, single linkage). **(C)** Heatmaps depict ChIP-seq tag densities for each *Drosophila* architectural protein as a function of distance, +/− 5 kb, from ETC. loci. **(D)** Western analysis of control preimmune and α-dTFIIIC220 immunoaffinity purifications detect interactions between dTFIIIC220 and CP190, Mod(mdg4), and BEAF-32.

analogous to colocalization patterns recently observed for overlapping mammalian factors [219]. Nevertheless, DNase I hypersensitivity centers on APBSs and is independent of the proximity of these regions with gene promoters.

Comparison of protein occupancy with respect to TAD localization further reveals a significant enrichment for high occupancy APBSs near TAD borders previously identified by high-throughput chromosome conformation capture [226]. For example, a strong domain border can be observed at 7 x 10^6 bp on *Drosophila* chromosome 3L in the form of two TADs defined by high intra-domain interaction frequencies and low inter-domain interaction frequencies (Figure 14A). The single fragment resolution TAD boundary identified corresponds to a region containing a high occupancy APBS bound by all queried proteins, including dTFIIIC220, suggesting strong chromatin domain separation may be collectively orchestrated by several architectural proteins. Genome-wide, protein occupancy is a strong predictor of TAD border localization, wherein 49% of TAD boundaries defined in Kc167 cells [198] are delineated within one restriction cut site by a high occupancy APBS, 35% by a medium occupancy APBS, and 12% by a low occupancy APBS (Figure 14B). We find similar enrichment profiles at TAD borders defined by Hi-C in embryos [65], and that localization to domain borders is independent of gene structure (Additional file 4, Figure S3E-F and Additional file 5).

TADs are defined by the compartmentalization of interaction frequencies, yet they also show varying degrees of compartmentalization. In other words, the borders that separate TADs appear to vary in terms of strength. We therefore quantified the degree of domain

separation, or “border strength”, by measuring the ratio of intra- vs. inter-TAD interaction frequencies (see Methods). Comparison of APBSs with border strength reveals a striking relationship, wherein chromatin domain separation scales incrementally with architectural protein occupancy (Figure 14C), providing strong evidence that combinatorial binding of these factors underlies a spectrum of functional capacity. In addition to differences in domain border strength, TADs also vary widely in size, ranging from only a few to several hundred kb in length. Visualization of pairwise interaction frequencies on a Mb scale illustrates this heterogeneity, which scales with the density and occupancy of APBSs (Additional File 6, Figure S5A-C). Whereas dense regions of high occupancy APBSs associate with very small TADs (median size ~55 kb), genomic regions characterized by low densities of clustered architectural proteins are comparatively much larger (median size 145-180 kb), consistent with a role for high occupancy APBSs in chromatin domain separation.

High occupancy APBSs associate with robust enhancer blocking activity

The role and function of insulator elements in genome biology has remained difficult to describe, despite extensive characterization and analyses. Though first defined by their ability to insulate genes from position effects and to prevent enhancer-promoter communication in transgenic reporter assays, many endogenous APBSs appear to lack these defining characteristics [64], suggesting they do not represent “insulators” in the classical sense. In agreement with this, recent work in mammals suggests that many CTCF sites fail to interfere with enhancer-promoter interactions and that their role may

be to facilitate interactions between these regulatory sequences instead [206, 207]. We therefore analyzed the relative occupancy level of architectural proteins in DNA fragments previously tested for enhancer-blocking activity using reporter assays, wherein specific regions of the genome were shown to be capable of robust or context-dependent enhancer blocking, incapable of enhancer-blocking activity, or act instead as transcriptional repressors [64, 184].

Insulator sequences capable of robust enhancer-blocking activity indeed correlate with high occupancy APBSs, with an average occupancy of 7.1 factors (Figure 14D). We find an intermediate level of protein occupancy at context-dependent insulators (5.2 factors), and comparatively low occupancy at fragments that did not possess enhancer-blocking activity (3.5 factors). The gradient of insulator activity correlates with DNase I hypersensitivity, consistent with the observed occupancy level and suggesting that robust enhancer-blocking insulators represent chromatin bound by several architectural proteins (Figure 14D).

Analysis of these sequences with respect to TAD border strength within their endogenous contexts further confirms that reporter-based assays reflect the functional capacity of these elements *in vivo*. For example, robust enhancer-blocking sequences correspond with genomic regions associated with strong TAD border strength, whereas non- or weak enhancer-blocking elements associate with weak border strength (Figure 14E). These data suggest that highly occupied APBSs enriched at the borders of TADs represent strong insulators involved in chromatin domain organization, whereas sites bound

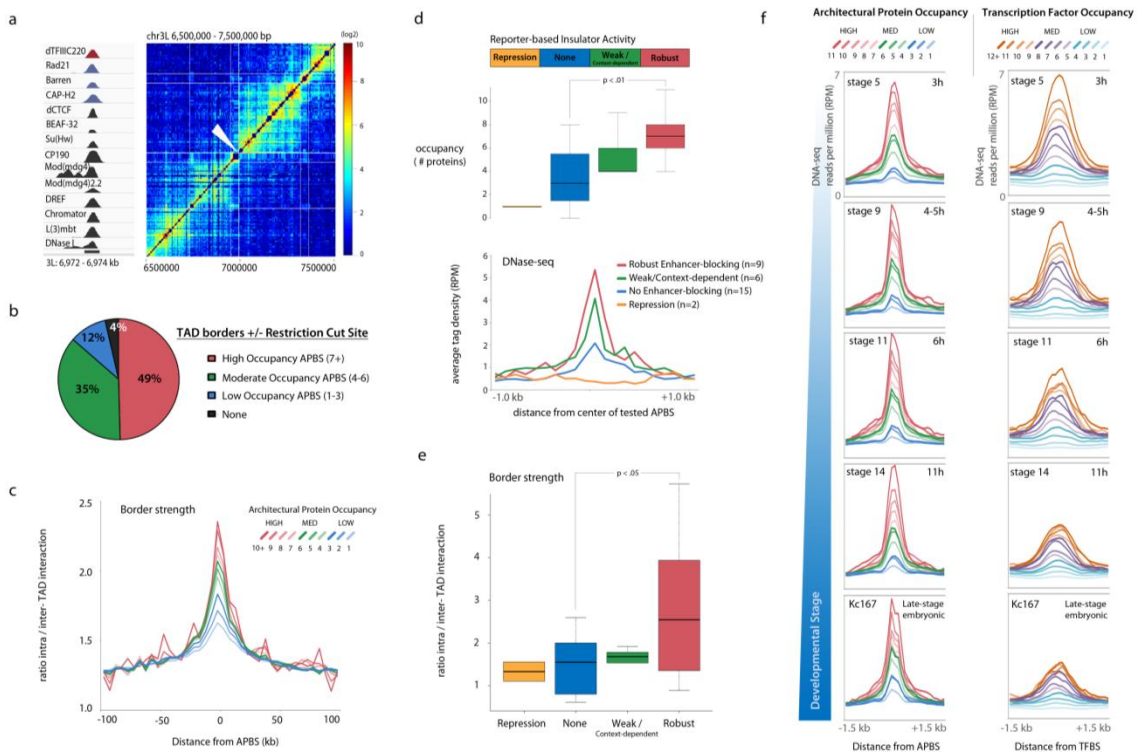


Figure 14. **High occupancy APBSs delineate TADs and associate with robust enhancer-blocking activity**. **(A)** Heatmap representing Hi-C interaction frequencies at single fragment resolution for a 1 Mb region across *Drosophila* chromosome 3L in Kc167 cells. White lines demarcate previously defined TAD boundaries [1]. A high occupancy APBS (left) is present at a single fragment topological domain border strongly separating two TADs (white arrow). Colorbar represents (log2) interaction frequencies observed between restriction fragments, ranging from low (blue) to high (red) **(B)** Percentage of TADs defined in Kc167 cells delineated by a high, medium, or low occupancy APBSs +/− one restriction cut site. (TAD borders n = 1,110, high occupancy APBSs n = 1,638 P < 0.00001, permutation test) **(C)** Topological border strength defined by the ratio of intra- versus inter-TAD interaction frequencies scales with the occupancy (number of bound proteins) at APBSs. **(D)** Architectural protein occupancy and DNase I

hypersensitivity at DNA fragments previously tested for enhancer-blocking activity in transgenic reporter assays [13,51,52]. Sequences shown to possess robust activity (red) correlate with both the highest occupancy and DNase I activity, whereas sites incapable of insulator activity are marked by low occupancy ($P < 0.01$, Wilcoxon rank sum test, two-sided). **(E)** Quantification of topological domain border strengths at sequences tested for insulator function within their endogenous context. Robust insulator sequences are characterized by significantly greater topological border strength than non-enhancer-blocking sequences ($P < 0.05$, Wilcoxon rank sum test, two-sided). **(F)** Tag density plots of rank-order normalized DNase-seq profiles throughout embryonic stages of development at APBSs [53], and at TFBSs shown to function as developmental enhancers during early embryogenesis. The progressive loss of DNase accessibility at highly bound TFBSs (right) contrasts with the combinatorially bound APBSs (left), which are marked by strong DNase I hypersensitivity throughout each stage of development.

individually or by few architectural proteins reside within TADs and may be reserved for specific regulation of genes.

High occupancy APBSs are characterized by DNase I hypersensitivity throughout *Drosophila* development

Genome-wide chromosome conformation capture studies provide evidence that a majority of topological domains are tissue invariant [66], suggesting sub-Mb scale domain structure may represent a common framework for higher order organizational dynamics. If clustered architectural proteins function to establish or maintain TADs, then high occupancy APBSs too must be largely tissue invariant and present throughout *Drosophila* development. We therefore compared APBSs defined in *Drosophila* Kc167 cells with DNase I hypersensitivity (DHS) profiles captured throughout stages of embryogenesis as a proxy for both chromatin accessibility and protein occupancy [225, 227]. DNase-seq profiles were rank-order normalized (see methods) across five embryonic stages, including the late-stage Kc167 cell line, and plotted with respect to protein occupancy at APBSs (Figure 14F).

High occupancy APBSs show a remarkably consistent pattern of DHS intensity, even at the earliest embryonic stages of development tested – just three hours post fertilization (Figure 14F), suggesting they are indeed stably occupied. Importantly, DNase I hypersensitivity is consistent across both promoter and non-promoter associated clusters (Additional File 5), supporting the use of chromatin accessibility as a measure of protein occupancy. The consistently open chromatin status at high occupancy APBSs starkly

contrasts with the DNase I profiles of previously characterized transcription factor HOT regions, which instead gradually lose DNase accessibility across embryonic stages (Figure 14F). The loss of DHS intensity at sites co-bound by several early transcription factors is consistent with data suggesting HOT regions function as spatiotemporal specific developmental enhancers during early embryogenesis [218]. These findings suggest that unlike HOT sites, clustered APBSs remain highly occupied throughout *Drosophila* development, and thus denote stable hubs for architectural protein association that may underlie the conserved topological domain structure observed across diverse cell types.

Mammalian TFIIIC and CTCF cluster at TAD borders

The observation that architectural proteins form large clusters and scale with the strength of TAD borders is made possible by the large repertoire of factors characterized to be essential for insulator function and mapped by ChIP-seq in *Drosophila*. This phenomenon has not been studied in mammals however, due to our limited understanding of what factors, besides CTCF, are capable of insulator function in vertebrates. Recent discovery that tRNA genes possess insulator activity in humans [27] suggest that TFIIIC may be responsible for this function, and raise the possibility that clustering of architectural proteins may have functional significance in mammals as well. For example, ETC loci often localize near CTCF sites in both human cells and mouse embryonic stem (ES) cells, and similarly associate with the cohesin complex as well [55, 56]. We therefore asked whether TFIIIC and CTCF cluster together at topological domain borders

by analyzing recent Hi-C data from mouse and human ES cells and IMR90 fibroblasts [66].

Comparison of ChIP-seq data mapping CTCF, cohesin, and three subunits of TFIIIC (TFIIIC220, -110, and -90) in mouse ES cells (mESCs) indicates strong overlap among these proteins (Figure 15A). Furthermore, we find enrichment for condensin II subunits CAP-H2 and CAP-D3, consistent with colocalization patterns in *Drosophila*, as well as PRDM5, a SET domain protein recently shown to interact and co-occupy genomic loci with CTCF, TFIIIC, and cohesin [228]. Binding of these five distinct factors was therefore used as a proxy for occupancy at CTCF sites analogous to APBSs in *Drosophila*. Analysis of CTCF occupancy with respect to TAD borders in mESCs again demonstrates a strong correlation between architectural protein clustering and chromatin organization. For example, a strong TAD border mapped to chromosome 5 in mESCs corresponds to a region bound by CTCF, TFIIIC (-220, -110, -90) Rad21, Condensin II (CAP-H2 and CAP-D3) and PRDM5, and marked by strong DNase I hypersensitivity (Figure 15B). Occupancy at CTCF sites is a strong predictor of both TAD border localization (Figure 15C) and TAD border strength (Figure 15D) as observed for APBSs in *D. melanogaster*, suggesting that clustering of architectural proteins is a general feature of genome organization conserved between *Drosophila* and mammals.

Genome-wide mapping of human architectural proteins associated with insulator activity has, to date, been limited to CTCF, TFIIIC-110, and cohesin. Nevertheless, we find that sites occupied by all three factors are significantly enriched within TAD borders mapped

in human ES cells and IMR90 fibroblasts, particularly at borders shown to be conserved between these two cell-types (Additional File 7, Figure S6A). To gain better insight into the occupancy of CTCF binding sites, we took advantage of recent large-scale mapping studies in which > 100 transcription factors and DNA-binding proteins were mapped by ChIP-seq in the human K562 cell line [229, 230]. In agreement with a machine learning approach [229], we find strong co-localization patterns between CTCF and DNA-binding proteins Znf143 (29%), JunD (40%), and the myc-associated zinc finger protein Maz (48%) (Full list provided in Additional file 7). The occupancy of CTCF binding sites again scales with TAD border strength as defined in *Drosophila* (Figure 15E), suggesting that a gradient of combinatorial binding by architectural proteins scales with topological structure and regulatory potential in human cells as well.

In addition to mapping hundreds of distinct factors, the human Encyclopedia of DNA Elements (ENCODE) project has mapped CTCF across dozens of human cell lines and diverse tissues [231], providing a powerful advantage for analyzing cell-type specific versus tissue-invariant CTCF binding sites. We therefore compared CTCF cell-type specificity and TAD border localization patterns by analyzing CTCF binding profiles reported across thirty-one cell lines (62 biological replicates; Additional File 8, Figure S6B). Whereas cell-type specific CTCF binding sites show relatively no enrichment at TAD borders, a striking trend toward TAD border localization is observed with increasing ubiquity, wherein ubiquitous CTCF sites present in all cell lines and biological replicates are most significantly enriched at TAD borders (Figure 15F). These results support Hi-C data proposing that a majority of topological domains are conserved among

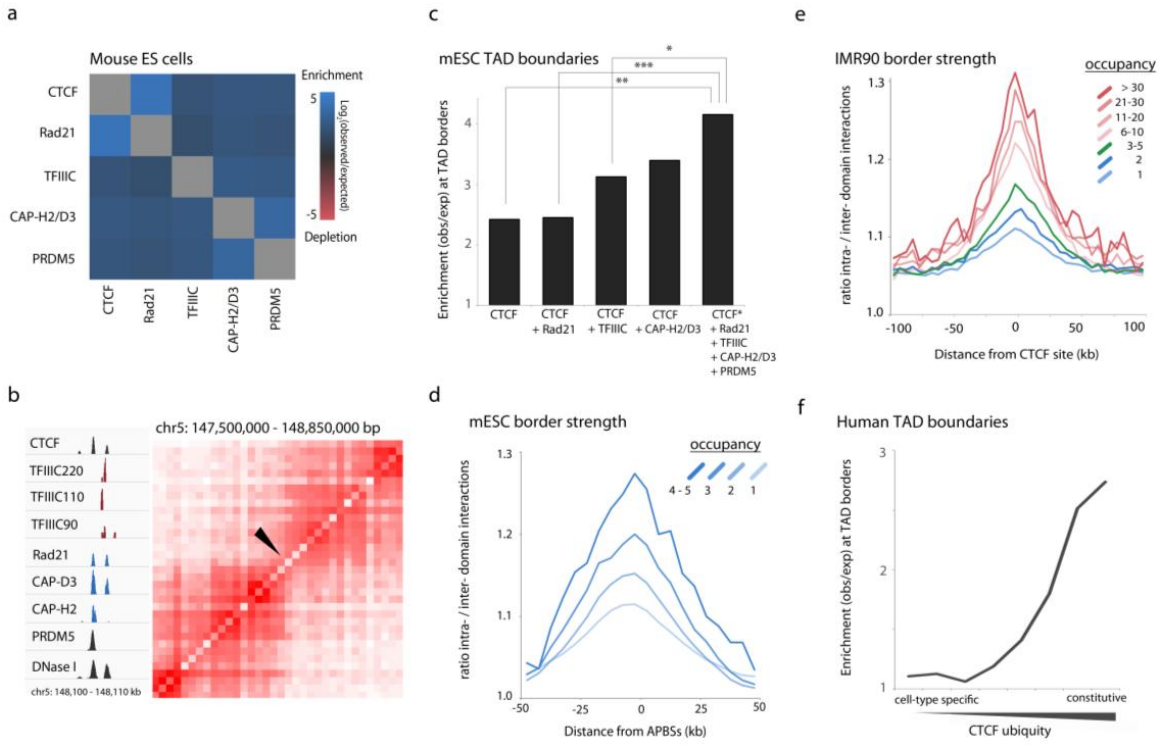


Figure 15. **Clustering of architectural proteins is a conserved feature of genome organization. (A)** Heatmap depicting overlap enrichment between architectural proteins mapped by ChIP-seq in mouse ES cells. Red to blue squares represent depletion (red) or enrichment (blue), determined as the $\log_2$ (observed/expected) frequency of overlap when compared to randomized, simulated data. **(B)** Example genomics viewer profile (left) of a high occupancy APBS in mouse ES cells, bound by CTCF, TFIIIC (−220, −110, and −90), Rad21, Condensin II (CAP-D3 and CAP-H2), PRDM5, and marked by strong DNase I hypersensitivity. Hi-C interaction matrix (right) illustrates the corresponding TAD separation observed *in vivo*. **(C)** Sites combinatorially bound by CTCF and other factors (* CTCF + 3 or more proteins) are significantly enriched at TAD borders in mouse ES cells. P values (* <0.05, ** <0.01, *** < 0.001) were calculated using permutation tests. **(D)** Relationship between protein occupancy, defined by the presence

CTCF, Rad21, PRDM5, TFIIIC (any or all subunits −220, −110, −90) and Condensin II (CAP-H2 and/or CAP-D3), and topological domain border strength in mouse ES cells. **(E)** Parallel analysis of topological domain border strength in human IMR90 fibroblasts as a function of protein occupancy at CTCF binding sites. Co-binding determined by cross-comparison of ChIP-seq datasets for transcription factors and DNA binding proteins in human K562 cells. **(F)** Relationship between cell-type specificity of CTCF binding sites and localization to TAD borders. CTCF ubiquity determined by cross-comparison of 62 CTCF ChIP-seq datasets across 31 human cell lines. X-axis represents CTCF sites grouped into 8 bins (~15,000 sites each) of increasing ubiquity ranging from cell type specific to constitutive. For a list of human cell lines, ubiquity scores and exact number of CTCF binding sites in each bin, see Materials and methods and Additional file 8

cell types and even species [66], and further suggest that this tissue-invariant structure may be determined by the constitutive genomic landscape of architectural proteins.

DISCUSSION

Insulators have been described as regulatory elements capable of activating and repressing transcription [208], able to block enhancer-promoter interactions and, more recently, to facilitate enhancer-promoter communication [232], yet multiple studies in *Drosophila* suggest that many architectural protein binding sites (APBSs) are not capable of insulator activity at all [46, 64]. Architectural proteins are enriched at the borders of topologically associating domains (TADs) [65, 66, 198], but why a majority of APBSs localize within topological domains and what differentiates border-associated from non-border binding sites have remained important and unresolved questions. By characterizing and mapping the genome-wide binding profiles for several architectural proteins, including the B box binding subunit of TFIIIC in *D. melanogaster*, we uncover a widespread spectrum of combinatorial binding by architectural proteins that offers an explanation for the diversity of localization patterns and function.

We find that clustering of architectural proteins scales with the tissue-invariant topological domain structure recently described by high throughput chromosome conformation capture studies. High occupancy APBSs are strongly enriched at TAD borders, and the number of architectural proteins present at a TAD border directly correlates with its strength, as measured by the ratio of inter-TAD versus intra-TAD interaction frequencies. TAD border-associated APBSs represent highly accessible DNase I hypersensitive regions present throughout *Drosophila* embryonic development, suggesting the binding of architectural proteins at these sites is constitutive across diverse

cell types and may underlie the conserved topological domain structure between different tissues. In support of this conclusion, comparison of CTCF ChIP-seq data across thirty-one human cell lines suggests that, whereas cell-type specific APBSs are virtually unrelated to domain structure, ubiquitous CTCF binding strongly predicts TAD boundary localization.

The clustering of architectural proteins is reminiscent but distinct from the clustering of transcription factors at highly occupied *cis* regulatory modules, similar to *Drosophila* HOT regions, recently shown to form around the cohesin complex [219, 220]. Though CTCF co-occurs with cohesin at a majority of binding sites, it does not localize to cohesin sites associated with mediator and dozens of other transcription factors in humans [219]. Nevertheless, Rad21 is necessary for stabilizing dense transcription factor clusters [220], suggesting the cohesin complex may serve an analogous role at clustered APBSs. Our finding that *Drosophila* architectural proteins, including CTCF, associate with Rad21 further suggests that this role may be evolutionarily conserved.

Genome-wide mapping of condensin complexes extends the relationship between APBSs and SMC-containing complexes even further. High occupancy APBSs are significantly enriched for the condensin II complex, most significantly at a subset of sites bound by Chromator and BEAF-32 (Figure 13C). Comparison of condensin II subunits CAP-H2 and CAP-D3 with the genome-wide CTCF profile in mouse ES cells further suggests that this relationship, like that with cohesin, may be a common feature of high occupancy APBSs. Mammalian CTCF was recently shown to interact with the condensin complex,

particularly CAP-D3, both *in vitro* and *in vivo* [233], suggesting CTCF may be responsible for recruiting condensin II to these clustered elements. However, whereas RNAi depletion of dCTCF leads to reduced cohesin localization at low, moderate, and high occupancy APBSs, we find no disruption of CAP-H2 localization to high occupancy APBSs (Additional File 2), suggesting additional factors may play a role in the recruitment of condensin II to these regulatory elements.

What role condensin II plays at APBSs will require future exploration, but many intriguing possibilities arise from its regulated activity throughout the cell cycle. For example, though defined for its involvement in chromosome assembly and segregation, condensin II has been shown to promote the formation of chromosome territories and to be tightly regulated during interphase [234], wherein phosphorylated CAP-H2 is targeted by the ubiquitin ligase complex SCF^{Slimb} for ubiquitin-mediated degradation [235]. CAP-H2 accumulates upon Slimb disruption, leading to chromosome reorganization and nuclear envelope defects, suggesting condensin II levels are tightly regulated for appropriate interphase chromatin organization. Meanwhile, *Drosophila* architectural proteins tightly associate with DNA and remain bound during mitosis [236], particularly at sites aligned with multiple factors, suggesting that condensin-bound APBSs may function as chromatin bookmarks for organized compaction and re-establishment of epigenetic regulation throughout the cell cycle.

The distinct localization of low versus high occupancy APBSs with respect to TAD borders suggests that function is often context-dependent and modulated by protein

composition (Figure 16). Whereas high occupancy APBSs are present at TAD borders and represent genomic loci capable of robust enhancer-blocking activity in transgenic reporter assays, low occupancy APBSs exhibit weak or virtually no enhancer-blocking function, or in the case of Su(Hw), gene repression [64]. These assays are commonly approached using the *gypsy* insulator, composed of 12 clustered Su(Hw) binding sites, as a positive control for such insulator activity, but nevertheless suggest that most APBSs do not represent "insulators" in the classical sense. Instead, low occupancy binding sites localize within TADs and may be reserved for locus-specific gene regulation, such as facilitating enhancer-promoter interactions.

We propose that the spectrum of TAD border strengths accompanied by differences in protein occupancy reflect the role of architectural proteins in long-range interactions (Figure 16) [73, 78]. For example, combinatorial binding of architectural proteins and recruitment of SMC-containing cohesin and condensin complexes may increase both their propensity to interact and the stability of interactions with other regulatory elements, strengthened by synergistic protein-protein and protein-DNA interactions. Furthermore, the very nature of high occupancy APBSs may indirectly reflect interactions with proteins bound to discrete genomic loci. In either case, the strong TAD separation defined by clustered APBSs is determined by the likelihood and/or stability of long-range interactions with other regulatory elements. Higher inter-TAD interaction frequencies observed across a comparatively weaker TAD border bound by fewer architectural proteins may be less likely to interact or exhibit weaker, more transient interactions that allow for greater inter-TAD interaction frequencies. A recent study further suggests that

APBSs are regulated by poly(ADP-ribosyl)ation of CP190, particularly at low occupancy, independent APBSs [237], whereas high occupancy APBSs more often remain unaffected. The synergy of several factors at clustered APBSs may contribute to this apparent immunity to certain post-translational regulatory mechanisms, which may be directed toward a subset of architectural proteins, and thereby represent a means for establishing stable chromatin domain organization in interphase cells.

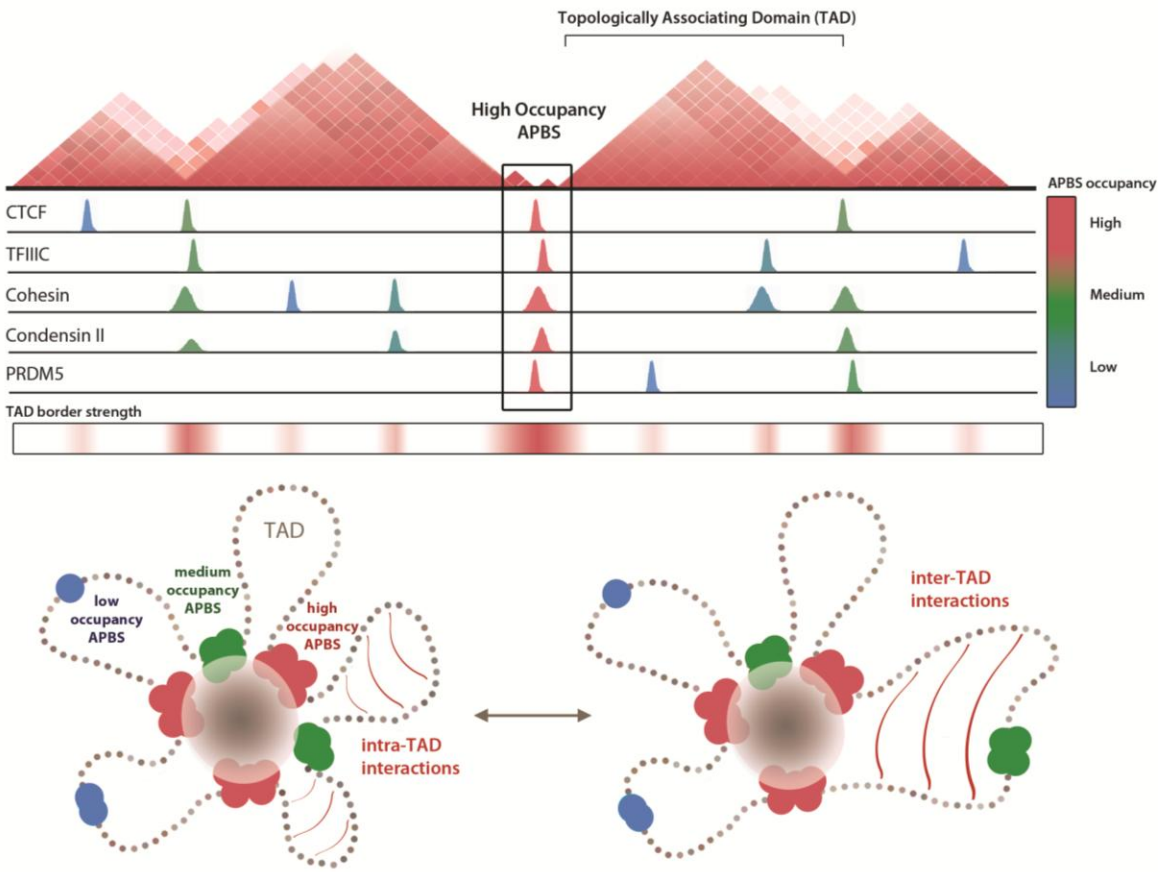


Figure 16. **Combinatorial binding of architectural proteins shapes topological domain structure.** Model illustrating the relationship between protein occupancy at APBSs and observed heterogeneity in TAD border strengths. We uncover a spectrum of architectural protein co-localization, ranging from low (blue) to high (red), which scales with the strength of TAD border formation. We propose that differences in TAD border strength reflect the role of architectural proteins in mediating long-range interactions. Interaction frequencies and/or interaction stability are greatest at high occupancy APBSs (red), whereas fewer or less stable interactions at intermediate APBSs (green) allows for inter-TAD interactions resulting in comparatively weaker TAD borders observed by Hi-C

EXPERIMENTAL PROCEDURES

dTFIIIC220 antibody generation

cDNA corresponding to CG7099 amino acids 1,357-1,907 was obtained from the *Drosophila* Genomics Resource Center (DGRC clone LD46862), PCR-amplified introducing a BglII restriction site upstream of the coding sequence, and subcloned into a pET-23a vector containing a GST and His tag at the C' and N' termini respectively. CG7099 protein fragment expression was induced by IPTG (0.5 mM) in BL21-CodonPlus® Competent Cells grown to a culture density ~ OD_{600}: 0.5, and shaken ~ 100 rpm for 2 hr. Cells were subsequently pelleted and proteins extracted via the B-PER protein extraction reagent (ThermoScientific product # 78243). Polyclonal rabbit antibodies were generated against the isolated CG7099 fragment at the Pocono Rabbit Farm and Laboratory. Quality control and antigen specificity were tested by peptide competition assays against Kc167 lysate with rabbit polyclonal α-dTFIIIC220 antibody preincubated with bacterial extract expressing GST empty construct or GST-CG7099 construct expressing a fragment corresponding to amino acids 1,357-1,907 (Additional File 1).

Immunoprecipitation and Western Analysis

All steps were performed at 4°C. Kc167 cells were harvested and washed once with ice-cold PBS. Cells (0.1 g) were lysed by incubating 10 min with 1 ml of ice-cold PBSMT (2.5 mM $MgCl_2$, 3mM KCl, and 0.3% Triton X-100 in PBS) plus protease inhibitors (1

mM PMSF and Complete protease inhibitor tablet cocktail [Roche]). Lysates were clarified by centrifugation at 16000 g for 10 min and protein concentrations were determined by Bradford assays (Bio-Rad). Packed Protein A Sepharose (15 µl bead volume [GE Healthcare]) was washed three times in PBSMT and pre-incubated with 3 µl of rabbit polyclonal anti-dTFIIIC220 or preimmune serum for 1 hr. Lysate was added to the antibody-conjugated Protein A Sepharose and incubated with agitation for 1 hr. Beads were washed three times with 1 ml PBSMT and once with 1 ml PBS. For comparing interaction between dTFIIIC220 and other insulator proteins, 50 µl of 1 M $MgCl_2$ was added to the beads and incubated for 5 min. Supernatant containing the eluted proteins was isolated by centrifugation. Laemmli SDS buffer was then added to the eluted proteins and boiled for 5 min. Samples were resolved by 6% SDS-PAGE and transferred to PVDF membrane (Millipore) in Tris-glycine transfer buffer and 20% methanol for 2 hr at 100 volts.

For Western blotting, membranes were blocked in TBST (20 mM Tris, pH7.4, 150 mM NaCl, 0.05% Tween 20) with 5% nonfat milk powder and incubated overnight with the following antibodies: rabbit-anti-dTFIIIC220 (1:2000), rabbit-anti-CP190 (1:10000), rabbit-anti-Su(Hw) (1:3000), rabbit-anti-Mod(mdg4)2.2 (1:3000), guinea pig-anti-dCTCF (1:1000), mouse-anti-BEAF-32 (1:100 [DSHB]), rabbit-anti-Rad21 (1:1000; gift from Dr. Dale Dorsett) and rabbit-anti-histone H3 (1:3000 [Abcam]). Membranes were washed three times with TBST and probed with secondary antibodies-conjugated to HRP (1:5000, Jackson ImmunoResearch Laboratories) for 1 hr. After three more washes, the

presence of different proteins was detected using SuperSignal West Pico/Dura Chemiluminescent substrate (Thermo Scientific).

ChIP-seq and Reference Data

Chromatin immunoprecipitation was performed as previously described [58]. In addition to dTFIIIC220, ChIP for Rad21, Barren, and CAP-H2 in *Drosophila* Kc cells were carried out using previously described antibodies (Rad21: α-Rabbit [221]; Barren: α-Rabbit [238], CAP-H2: α-Rabbit; gifts from Dr. Dale Dorsett, Dr. Hugo Bellen, and Dr. Giovanni Bosco, respectively). Sequences were mapped to the dm3 genome with Bowtie 0.12.3 [189], using default settings. To account for the repetitive nature of tRNA genes, multimapping sequences were filtered out for all dTFIIIC220 ChIP-seq experiments. Peaks were then called with MACS 1.4.0alpha2 [190] using equal numbers of unique reads for input and ChIP samples, a p value cutoff of 1×10^{-10}, and a False Discovery Rate (FDR) threshold of 5% (Additional File 9, Figure S7). For classification of overlapping APBSs, MACS-identified peaks (pval 1e-10, FDR 5%) are further refined as the MACS calculated summit +/- 200 bp. For dTFIIIC220, peaks used for analyses were independently identified by MACS in 2 out of 3 biological replicates. For visualization, mapped sequence reads were loaded on to the Integrated Genomics Viewer [239], [240]. Previously published ChIP-seq data for *Drosophila* architectural proteins were obtained from GEO accessions GSE30740 [77] and GSE36944 [46], and ChIP-chip data corresponding to TFIIIB subunits TRF1 and BRF from [241]. Raw DNase-seq in Kc167 cells was obtained from http://compbio.med.harvard.edu/flychromatin/data.html [225];

DNase-seq in HeLa S3 cells from GEO series GSE32970; and DNase-seq in mouse ES cells from ENCODE dataset wgEncodeUwDgfEscj7129 [231]. ChIP-seq data for architectural proteins in mouse ES cells were obtained from GEO series GSE29218 (CTCF), GSE33346 (Rad21, CAP-H2, CAP-D3) [222], GSE51816 (PRDM5), and from ArrayExpress accession E-MTAB-767 (TFIIIC-110 -220 -90). ChIP-seq data for architectural proteins in HeLa S3 and K562 cells was obtained from GEO series GSE31477 (TFIIIC, Rad21) and from publicly available ENCODE data [229, 230, 242].

Bioinformatics analyses

Sequence alignment for the dTFIIIC220 B box binding domain with analogous proteins in yeast and humans was generated using the Conserved Domains Database [243] and visualized using C3nD v4.3 [244-246] and Jalview [247]. For ChIP-seq comparisons, DNA sequence motifs were identified by MEME-ChIP using default settings [28]. Overlap between dTFIIIC220 and annotated tRNA genes were identified using publicly available tools on Galaxy [191-193]. Comparison of APBSs with respect to Pol II-transcribed genes employed gene structure (TSSs, exons, introns, TTSs) obtained using the UCSC genome browser [248], [249]. Enrichment profiles for architectural protein co-occurrence and localization to TAD boundaries were defined as the observed overlapping frequencies over expected frequencies determined by shuffling datasets, while controlling for the number of peaks and start/stop location of peaks on each chromosome. P-values were determined as the chance of observing an equal or greater co-occurrence across 100,000 Monte Carlo Permutation tests. Results were visualized using Java Treeview

[195]. Unless otherwise noted, read intensity plots were generated by binning ChIP-seq reads into 100 bp bins and extracting read counts in bins surrounding described anchor points (eg. dTFIIIC220 summits), and visualized using Java Treeview [195]. Rank-order normalization of DNase-seq and/or ChIP-seq data was carried out as recently described [182]. Briefly, datasets are rank-ordered in 10-bp bins across the reference genome, descending from high to low read intensity, and at each level, bins are re-assigned the average read value across samples used for comparison.

Overlap matrices and classification of APBSs

D. melanogaster

ChIP-seq peaks, defined as 400bp centered around MACS calculated summits, were cross-analyzed using BED tools MultiIntersectBed [250], creating a matrix of unique genomic loci bound by architectural proteins. In *Drosophila* this includes ChIP-seq data for dCTCF, BEAF-32, Su(Hw), CP190, Mod(mdg4), DREF, Chromator, L(3)mbt, dTFIIIC220, Rad21, and CAP-H2. The number of target motifs and the relative level of ChIP tag density were not considered when generating this list. Adjacent output peaks were merged and the largest occupancy region and associated factors isolated for further analyses, (i.e. directly adjacent regions bound by 4, 5, then 4 proteins were merged into one peak centered on the highest (5 proteins) occupied region) see additional files 3 and 7. Each Architectural Protein Binding Site (APBS) was then classified as being either low occupancy (1-3 proteins), moderate occupancy (4-6 proteins), or high occupancy (7+ proteins). Co-localization frequencies for factors depicted in Figure 13C were calculated

similarly and correspond to sites combinatorially bound by 4 or more architectural proteins identified independently of dTFIIIC220, Cohesin, Condensin I or II.

Mouse ES cells

ChIP-seq peak data for CTCF (GSE 29218), Cohesin (GSE 33346), TFIIIC, Condensin (GSE 33346), and PRDM5 (GSE 51816) in mESCs was obtained from published sources. ChIP-seq experiments for multiple subunits were available for TFIIIC (-220, -110, and -90) and Condensin (CAP-H2 and CAP-D3). For these cases data from all available subunits was combined into a single set. Being in proximity to any subunit of TFIIIC or Condensin was considered sufficient for co-localization. CTCF sites were classified as in proximity to Rad21 or TFIIIC if there was a Rad21 or TFIIIC peak within 500 bases of the center of the CTCF site. As a majority of CTCF sites had Rad21, the number of CTCF sites with TFIIIC and without Rad21 was very small and therefore was not shown. Each of these unique subsets was then assayed for its prevalence near TAD borders. Sites within 20 kb of a border were considered at a TAD border and sites outside of these windows were considered not at a TAD border. Expected values were calculated using a random distribution of sites with site type, size, and chromosomes conserved and locations randomized. We performed a Monte Carlo permutation test in order to calculate significance. The classifications of the sites were randomized and the number of permutations that resulted in a result as extreme as the observed over the total number of permutations was taken as the p value.

Humans

Occupancy at CTCF binding sites were determined by cross-comparison with publicly available genomewide binding datasets for DNA-binding/transcription factors CTCF, Rad21, TF3C, Yy1, Smc3, Znf143, Myc, Max, Maz, JunD, Arid3a, Atf1, Atf3, Bach1, Bcl, Bcl3, Bdp1, Bhlhe40, Brf1, Brf2, Brg1, Cbx3, CCNT2, CEBPbeta, CHD2, Corest, CTCFL, E2F4, E2F6, Egr1, Elf1, Elk1, Ets1, Ezh2, fos, FosL, GATA1, GATA2, HDAC1, HDAC2, HDAC6, HDAC8, HMGN3, Ini1, cJun, MafF, MafK, Mef2a, MXI1, Nelfe, Nfe2, Nfya, Nfyb, Nr2f2, Nrf1, P300, Phf8, Plu1, Rbbp5, Rfx5, Sap30, Setdb1, Sin3A, Sirt6, Six5, Sp1, Sp2, Srf, Stat5, Taf1, Taf7, Tal1, Tblr1, Tbp, Tead4, TFIIB, TFIIF, Thap1, Tr4, Trim28, Ubtfs, Usf1, Usf2, Xrcc4, Zbtb7, Zbtb33, Znf263, Znf274 [231], [56, 230]. Overlap matrices were generated as described for *Drosophila*.

Topologically Associating Domains and calculation of TAD border strength

Hi-C analysis and definition of TADs in *Drosophila* Kc167 cells were used as previously reported [226]. To measure the degree of separation of chromatin between two sides of a specific enzyme cutting site S, we analyze region A, which is adjacent to S on one side, and region B flanking cut site S on the opposite side. Intra-TAD Hi-C interaction counts within A and intra-TAD Hi-C interaction counts within B are calculated and compared with inter-TAD Hi-C interaction counts between regions A and B. The difference is defined as local contrast and centered to have median value of approximately 1. High value of local contrast corresponds to enriched intra-domain contact frequencies relative to inter-domain contacts. Thus, TAD borders generally exhibit strong measures of local contrast. TADs defined in mouse ES cells and humans were obtained from published data

[66]. TAD borders were taken from hESC and IMR90 lines and a common subset of borders found in both was used to form the conserved dataset. TAD border strengths in mESCs and humans were calculated as described for *Drosophila*.

Comparison of APBS occupancy and Insulator function from transgenic reporter assays

Enhancer-blocking results reported for several tested insulator elements were obtained from work by Negre et al 2011 [184] and Schwartz et al. 2012 [64], and categorized as either capable of robust enhancer blocking, weak/context dependent enhancer blocking, no enhancer blocking, or in the case of Schwartz et al, two suppressor of hairy wing independent loci capable of gene repression. The occupancy of each insulator element was then extracted by comparison with ChIP-seq peaks and overlap matrices.

	chr	start	stop	reported insulator activity	occupancy
1	chr2R	5428851	5429464	(robust) enhancer blocking	7
2	chr2R	10028181	10029180	(robust) enhancer blocking	10
3	chr2R	20199584	20201894	(robust) enhancer blocking	7
4	chr2R	20486926	20487926	(robust) enhancer blocking	11
5	chr3R	2696003	2697000	(robust) enhancer blocking	4
6	chr3R	7774458	7775524	(robust) enhancer blocking	6
7	chr3R	17231264	17234459	(robust) enhancer blocking	7
8	chrX	255313	255772	(robust) enhancer blocking	4
9	chrX	13180486	13183206	(robust) enhancer blocking	8
10	chr2L	18347655	18348655	no enhancer blocking	1

11	chr2L	19989037	19990037	no enhancer blocking	1
12	chr2R	4091291	4092348	no enhancer blocking	2
13	chr2R	10272132	10274060	no enhancer blocking	3
14	chr2R	17837219	17838527	no enhancer blocking	3
15	chr3L	7241880	7242880	no enhancer blocking	6
16	chr3L	8321816	8322816	no enhancer blocking	4
17	chr3L	14664327	14665327	no enhancer blocking	5
18	chr3L	15975461	15976460	no enhancer blocking	3
19	chr3R	2718919	2719623	no enhancer blocking	8
20	chr3R	10133401	10134400	no enhancer blocking	7
21	chr3R	12683547	12683918	no enhancer blocking	0
22	chr3R	12810442	12811667	no enhancer blocking	0
23	chr3R	20850521	20851520	no enhancer blocking	7
24	chrX	20953657	20955060	no enhancer blocking	7
25	chr2R	18327070	18328070	Repression	1
26	chr3R	13006890	13007890	Repression	1
27	chr2L	8463984	8464984	weak/context-dependent	4
28	chr2R	18021026	18022026	weak/context-dependent	4
29	chr3R	11318599	11320846	weak/context-dependent	4
30	chr3R	27206661	27207660	weak/context-dependent	4
31	chrX	7827021	7828020	weak/context-dependent	9
32	chrX	9904070	9904636	weak/context-dependent	6

CTCF site ubiquity

Existing CTCF ChIP-seq data was obtained from the ENCODE project for analysis. 31 cell lines with 2 replicates each were chosen for a total of 62 unique ChIP-seq

experiments in a wide range of human cell lines. These 62 data sets were combined into a composite list of all CTCF sites classified by the number of experiments each was found in. Sites that were found in only 1 of the 62 experiments were discarded as they failed to replicate. Sites less than a thousand bases from a site present in over twice as many cell lines were merged into the more ubiquitous site. To create an expected distribution, CTCF sites were shuffled. The ubiquity, size, and chromosome of each site were conserved, but the locations were randomized to a position between the first and the last CTCF sites on the chromosome. Sites were then separated into 8 bins of ~15,000 sites by their ubiquity. The ubiquity scores of each bin and number of CTCF sites are as follows: bin 1: 2 replicates, 15,568 sites; bin 2: 3-4 replicates, 14,328 sites; bin 3: 5-8 replicates, 14,326 sites; bin 4: 9-17 replicates, 15,240 sites; bin 5: 18-33 replicates, 14,536 sites; bin 6: 34-52 replicates, 15,707 sites; bin 7: 53-61 replicates, 15,213 sites; bin 8: 62 replicates, 15,582 sites. To analyze localization to human embryonic stem cell TAD borders, each site in the observed and expected data sets was classified as within 20 kb of a TAD border or not. The resulting frequencies were used to calculate observed over expected values.

Human cell lines and corresponding GEO accession numbers

	Human cell line:	**replicate 1**	**replicate 2**
1	A549	GSM1022640	GSM1022639
2	Ag04449	GSM749695	GSM749678
3	Ag04450	GSM749769	GSM1022635
4	Ag09309	GSM749750	GSM749680
5	Ag09319	GSM749728	GSM749723

6	Ag10803	GSM749714	GSM749759
7	Aoaf	GSM749666	GSM749736
8	Be2c	GSM1022653	GSM1022650
9	Gm12866	GSM849305	GSM849305
10	Gm12867	GSM849301	GSM849301
11	Gm12868	GSM849300	GSM849300
12	Gm12869	GSM849303	GSM849303
13	Gm12870	GSM849302	GSM849302
14	Gm12871	GSM849304	GSM849304
15	Hac	GSM1022661	GSM1022662
16	Hasp	GSM749696	GSM1022668
17	Hbmec	GSM749743	GSM749710
18	Hcm	GSM1022657	GSM1022677
19	Hcpe	GSM749735	GSM749745
20	Hct116	GSM1022652	GSM1022651
21	Hee	GSM749712	GSM749745
22	Hffmyc	GSM1022671	GSM1022669
23	Hmf	GSM749665	GSM749675
24	Hpaf	GSM749681	GSM749751
25	Hpf	GSM749699	GSM749717
26	Hrpe	GSM749673	GSM1022665
27	Hvmf	GSM1022630	GSM1022628
28	Mcf7	GSM1022658	GSM1022663
29	Nhdfneo	GSM1022675	GSM1022676
30	Rptec	GSM1022667	GSM1022666
31	Wi38	GSM1022637	GSM1022634

CTCF, Cohesin, and TFIIIC analysis in HeLa S3 cells

Enrichment of CTCF, TFIIIC, and Rad21 at human TAD borders (Additional File 7, Figure S6) was performed using CTCF, TFIIIC, and Rad21 datasets commonly mapped in HeLa S3 cells. Published ChIP-exo experiments were used as HeLa CTCF sites without any additional modification [242]. Rad21 and TFIIIC sites were determined from previously published ChIP-seq experiments (GSE31477). TAD borders were taken from hESC and IMR90 lines and a common subset of borders found in both was used to form the conserved dataset. CTCF sites were classified as in proximity to Rad21 or TFIIIC if there was a Rad21 or TFIIIC peak within 500 bases of the center of the CTCF site. As a majority of CTCF sites had Rad21, the number of CTCF sites with TFIIIC and without Rad21 was very small and therefore was not shown. Each of these unique subsets was then assayed for its prevalence near TAD borders. Sites within 20 kb of a border were considered at a TAD border and sites outside of these windows were considered not at a TAD border. Expected values were calculated using a random distribution of sites with site type, size, and chromosomes conserved and locations randomized. We performed a Monte Carlo permutation test in order to calculate significance. The classifications of the sites were randomized and the number of permutations that resulted in a result as extreme as the observed over the total number of permutations was taken as the p value.

Accession numbers

All ChIP-seq data is publicly available under GEO accession number GSE54529

ACKNOWLEDGEMENTS

We are especially thankful to Drs. Dale Dorsett, Hugo Bellen, Giovanni Bosco, and Jørgen Johansen for graciously sharing antibodies against Rad21, Barren, CAP-H2, and Chromator. We thank the *Drosophila* Genomic Resource Center (supported by NIH grant OD010949-10) for reagents, and The Genomic Services Lab at the HudsonAlpha Institute for Biotechnology for their help in performing Illumina sequencing of ChIP-Seq samples. This work was supported by U.S. Public Health Service Award R01GM035463 from the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Work presented in this chapter is published in: Genome Biol. 2014 Jun 30;15(6):R82. doi: 10.1186/gb-2014-15-5-r82

CHAPTER 4:

CTCF-DEPENDENT CO-LOCALIZATION OF CANONICAL SMAD SIGNALING FACTORS AT ARCHITECTURAL PROTEIN BINDING SITES IN D. MELANOGASTER

Work presented in this chapter is part of a collaboration with Aidan J. Peterson and Michael O'Connor and the University of Minnesota

Van Bortle K, Peterson AJ, Takenaka N, O'Conor M, Corces VG. CTCF-dependent co-localization of canonical smad signaling factors at architectural protein binding sites in D. melanogaster. *In preparation.*

My contributions include conceiving the project, performing the experiments, computational analysis of the data, interpretation of the results, and writing the manuscript.

ABSTRACT

Eukaryotic chromosomes are organized into highly self-interacting modules, called topologically associating domains, which are inherently defined by long-range interactions relevant to transcription. We have recently shown that the boundaries of topological domains are delimited by high occupancy architectural protein binding sites that possess robust insulator activity. However, architectural proteins also target low occupancy sites that facilitate interactions within topological domains important for gene activation and repression. CTCF, an architectural protein conserved from flies to humans, was recently shown to recruit receptor-regulated Smad proteins to the H19 imprinting control region and the Alzheimer beta-amyloid precursor gene. However, to what degree CTCF recruits these canonical TGF-β signaling proteins genome-wide and whether CTCF-dependent co-localization is conserved remains unknown. By mapping TGF-β signaling factors Mad, dSmad2, Medea, and Schnurri genome-wide, we identify numerous overlapping sites dependent on CTCF in *D. melanogaster*. CTCF-RNAi sensitive Smad binding sites are enriched within topological domains, whereas sites overlapping high occupancy topological domain borders remain unaffected, suggesting a potential redundancy in recruitment to architectural protein binding sites. Surprisingly, we show that whereas Mad, Medea, and Schnurri binding is dynamically altered in response to DPP signaling, dCTCF occupancy remains static, suggesting a limited role for CTCF in this TGF-β response.

INTRODUCTION

Architectural proteins mediate long-range interactions that contribute to the three-dimensional organization of interphase chromosomes [251, 252]. We have recently shown that topologically associating domains (TADs), which represent highly self-interacting regions of eukaryotic chromosomes [65-67], are partitioned by high occupancy architectural protein binding sites (APBSs) [46, 226, 253]. Both TADs and high occupancy APBSs appear to be largely tissue-invariant [66], suggesting modular chromatin domains are conserved throughout development. However, intra-TAD interactions, which may be facilitated by low occupancy APBSs within domains, often consist of enhancer-promoter and promoter-promoter interactions that are likely to be dynamically regulated in response to cell signaling events and between cell types to produce cell-type specific transcription patterns [78, 199, 200, 206]. We have previously demonstrated that *Drosophila* architectural proteins exhibit moderate changes in genome-wide localization in response to the insect steroid hormone 20-hydroxyecdyone [77], which is bound by the nuclear ecdysone receptor, a ligand-activated transcription factor that in turn activates specific genes [254]. Recruitment of CP190, a BTB-containing protein essential for insulator activity, is dynamically targeted to DNA-binding architectural proteins, leading to ecdysone-induced changes in chromosome organization. However, to what degree architectural protein binding contributes to additional signaling events remains poorly addressed.

Here we probe the relationship between *Drosophila* architectural proteins, including the *Drosophila* homolog of CTCF (dCTCF), and several transforming growth factor beta

(TGF-β) and bone morphogenic protein (BMP) signaling effector proteins, which direct the transcriptional response to TGF-β pathways involved in controlling cellular homeostasis, proliferation, differentiation, and apoptosis [255, 256]. In the canonical TGF-β signaling pathway, serine-threonine kinase transmembrane receptors bind to extracellular TGF-β ligands, and in turn phosphorylate receptor-regulated Smad proteins which are then able to form complexes that translocate into the nucleus to promote transcriptional activation and repression [257]. Several recent studies have focused attention on uncovering the chromatin determinants of Smad localization. Master regulatory transcription factors, which control the transcription of key cellular identity genes and are themselves expressed in a cell-type specific manner, were shown to direct the localization of BMP and Wnt signaling factors in hematopoietic progenitor cells, and of TGF-β effector Smad proteins in embryonic stem cells (ESCs) [258, 259]. The co-localization of Smads at cell-type specific master transcription factor binding sites in multi-potent cells provides an attractive model for how TGF-β signaling events can produce diverse, tissue-specific responses.

In addition to master transcription factors, TRIM33, a multifunctional Smad-interacting protein that recognizes the dual histone mark motif of H3K9me3 and H3K18ac, also creates a platform for Smad localization in response to signaling events in ESCs [260]. Nevertheless, the mechanisms driving recruitment of Smad proteins to DNA in non-stem cells remain largely unexplored. Limited studies probing the mammalian Alzheimer amyloid precursor protein (APP) gene promoter [261] and the H19 imprinting control region [262] have identified sites in which Smad recruitment to DNA is mediated by architectural protein CTCF. These studies led us to ask whether CTCF also recruits Smad

proteins to DNA on a global scale, to what degree these interactions are related to the TGF-β response, and whether CTCF-directed Smad localization is conserved.

Here, we report the genome-wide landscape of the receptor-regulated Smad proteins Mothers against DPP (Mad) and dSmad2, the co-Smad Medea, and the co-repressor Schnurri in *Drosophila melanogaster*. Indeed, we identify numerous sites in which Smad binding co-localizes with the architectural protein dCTCF and further demonstrate that dCTCF binding is required for co-localization at a subset of sites. These data support a role for CTCF as a conserved determinant of Smad localization from *Drosophila* to humans. However, we also demonstrate that dynamic binding of Smad proteins in response to the *Drosophila* bone morphogenic protein Decapentaplegic (DPP), a member of the TGF-β superfamily, occurs in the context of dCTCF-independent binding sites and that dCTCF binding itself remains unchanged. Together, these results suggest that CTCF may play a relatively minor role in directing the genomic response to DPP, but nevertheless appears to act as an important player in defining the occupancy landscape of Smad proteins.

RESULTS

Genome-wide mapping of *Drosophila* Smad proteins

TGF-β superfamily signaling in *D. melanogaster* is traditionally broken into two branches based on the ligand-receptor interaction and the class of receptor-regulated Smad proteins (R-Smads) that are subsequently phosphorylated [263]. In the TGF-β/Activin branch, binding between the Activin-like ligand Dawdle and the type I serine-threonine kinase receptor Baboon results in phosphorylation of the R-Smad dSmad2 (Smad on X) [264]. BMP signaling, on the other hand, occurs through transmembrane receptors Thickveins, Saxophone, and Punt, and induces phosphorylation of the R-Smad Mothers against DPP (Mad) [265, 266]. The phosphorylated forms of dSmad2 and Mad are able to form heterotrimeric complexes with the co-Smad Medea, and upon which translocate into the nucleus to direct changes in transcription [267]. The Mad/Medea complex has been shown to target activation regulatory elements in response to DPP, as well as repressive regulatory elements to which the transcriptional repressor Schnurri is recruited [268]. Therefore, in order to effectively capture the chromatin landscape of TGF-β signaling pathways, we carried out ChIP-seq experiments against dSmad2, Mad, Medea, and Schnurri in the late-embryonic *Drosophila* hemocyte cell line Kc167 (Figure 17).

Initial genome-wide analysis of Smad binding reveals extensive overlap between R-Smads and the co-Smad Medea (Mad/Medea – 85%; dSmad2/Medea – 81%; $p<0.0001$) as well as with the co-repressor Schnurri (Mad/Schnurri – 67%; dSmad2/Schnurri – 70%; $p<0.0001$) (Figure 17A). Nearly half of all binding sites for either Mad or dSmad2 also

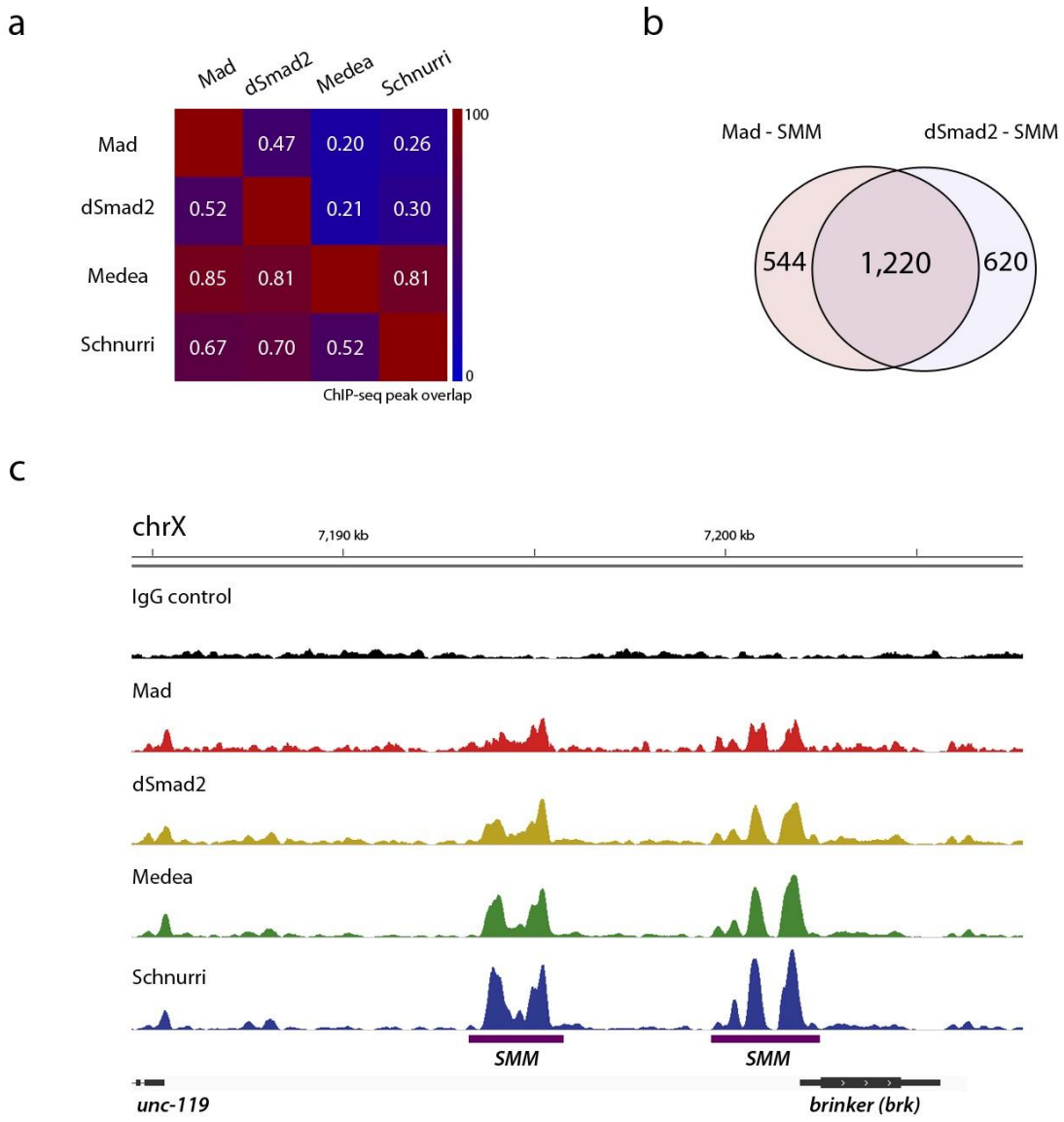


Figure 17. **Genome-wide mapping of BMP and TGF-β signaling proteins Mad, dSmad2, Medea, and Schnurri in *Drosophila* Kc167 cells**. (a) Heatmap representation of ChIP-seq peak overlap between individual Smad proteins. Corresponding values represent percentage of total binding sites for factors along the horizontal axis that overlap peaks identified for factors along the vertical axis, ranging from 0 (blue) to 100 (red). (b) Overlap of Schnurri-Mad-Medea modules (SMM) individually identified for receptor-regulated Smad proteins Mad or dSmad2. Overlap is statistically significant

($p<0.0001$, permutation test). (c) Illustration of ChIP-seq experiments and Smad protein overlap at SMM regulatory elements present in previously characterized promoter-proximal modules of the *brinker* gene locus.

overlap with the distinct R-Smad (Mad/dSmad2 – 52%; dSmad2/Mad – 47%; $p<0.0001$), suggesting a potential interaction between TGF-β/Activin and BMP signaling pathways at the genomic level. Recent studies do suggest that TGF-β/Activin activation can cross-talk with BMP signaling upstream of transcription. For example, the BMP R-Smad Mad can become phosphorylated by the TGF-β receptor Baboon in response to ligand stimulation and independent of BMP receptors [269], and mutation of the Activin R-Smad dSmad2 leads to altered BMP pathway signaling [270]. This suggests a potentially strong degree of interaction between TGF-β and BMP signaling pathways both upstream and downstream of nuclear translocation. Interestingly, numerous Medea and Schnurri binding sites are present independently of either Mad or dSmad2, though a majority of Schnurri binding sites overlap with Medea (Figure 17A), suggesting additional regulatory factors may be involved in recruiting Medea and Schnurri to DNA.

Identification of genomic loci co-bound by Medea, Schnurri, and either R-Smad reveals even greater overlap between the BMP and TGF-β signaling factors Mad and dSmad2 (Figure 17B; 69% for Mad; 66% for dSmad2; $p<0.0001$), suggesting both R-Smads recruit Medea and Schnurri to similar target sequences. The 1,220 overlapping Schnurri, Mad, Medea modules (SMM) bound by both Mad and dSmad2 provide a high confidence list of signaling target regions commonly identified in four independent ChIP-seq experiments, that we consider for further analysis. Visual inspection of high confidence SMM modules confirms co-localization of these signaling proteins at well characterized target genes. For example, Brinker, a nuclear repressor that antagonizes DPP signaling by binding similar Mad/Medea target sequences, is transcriptionally repressed by Mad, Medea, and Schnurri in response to DPP signaling events. Accordingly, the *brinker* gene

was recently shown to include several promoter modules targeted by Mad, Medea, and Schnurri [271]. Although *brinker* in modestly expressed in *Drosophila* Kc167 cells, our ChIP-seq data provide evidence that these SMM modules are highly occupied in cell culture even prior to induction with DPP (Figure 17C), suggesting either a minimal level of signaling occurs in Kc167 cells or that Smad-mediated repression is regulated downstream of sequence binding.

SMM modules overlap *Drosophila* CTCF and other architectural proteins

De novo motif analysis on SMM modules using MEME-ChIP [168] identifies significant enrichment for two sequences that show significant similarity to putative Mad and Medea consensus sequences (Figure 18). Previous comparisons of specific Mad/Medea-binding elements suggest that Smads target GCCGNC as a consensus binding sequence [272, 273], and, similarly, MEME-ChIP identifies significant enrichment for GCYGSC at SMM modules (Figure 18A). SMM modules are nearly equally enriched for a GC rich consensus sequence with strong similarity to putative Medea binding sites, together suggesting that our ChIP-seq profiles for dSmad2, Mad, Medea, and Schnurri provide an accurate means for identifying Smad-signaling response elements.

Interestingly, motif analysis also identifies a consensus sequence previously identified as being enriched at *Drosophila* CTCF binding sites that overlap with additional architectural proteins, such as Boundary Element Associated Factor of 32 kDa (BEAF-32), Centrosomal Protein 190 (CP190), Modifier of mdg4 (Mod(mdg4)), Chromator, and the cohesin and condensin complex kleisen subunits Rad21 and CAP-H2 respectively [46, 253]. In fact, we find significant motif enrichment for DNA binding architectural

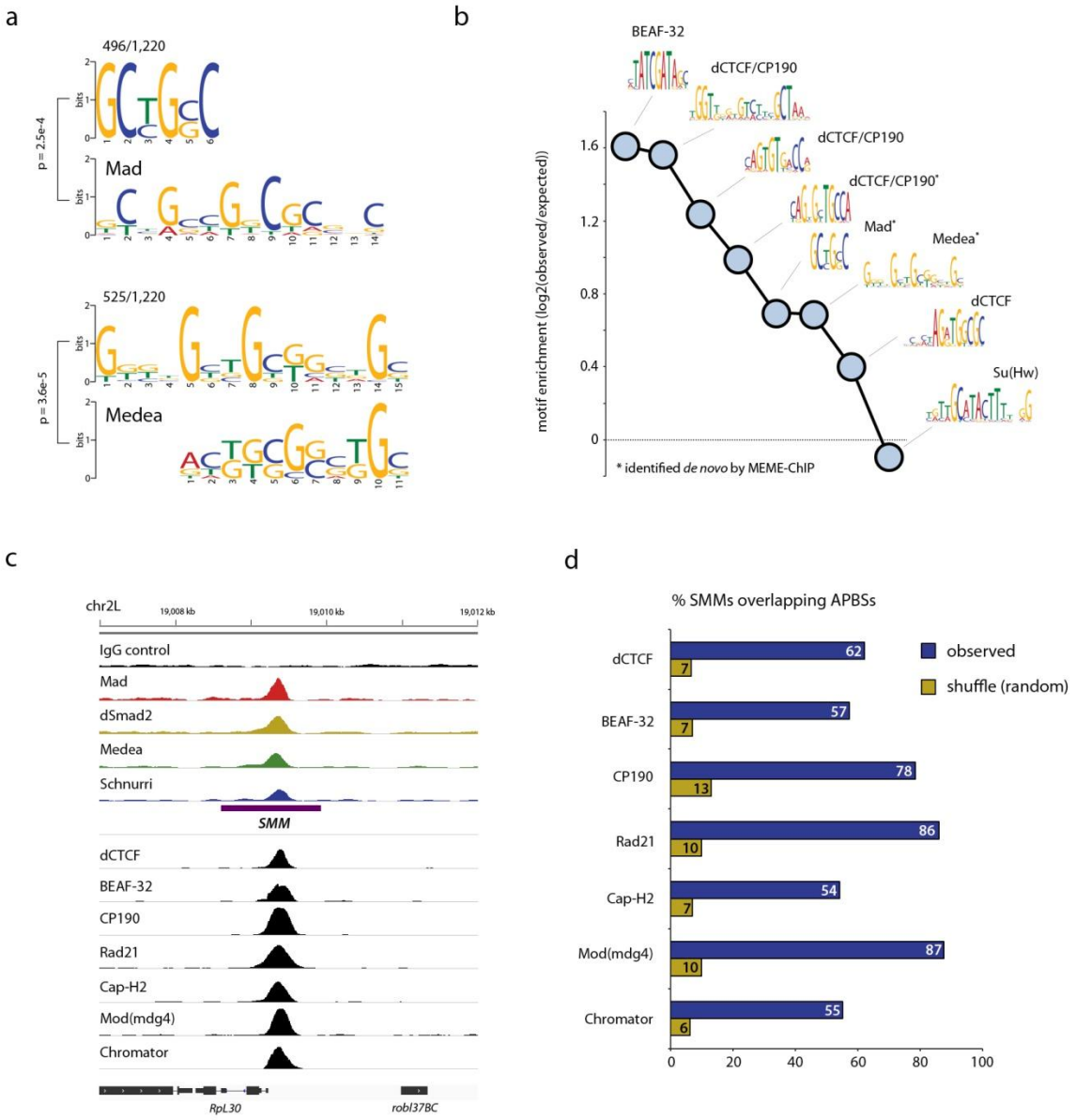


Figure 18. **SMM module motif enrichment and overlap at architectural protein binding sites.** (a) Putative Mad and Medea motifs identified *de novo* by MEME-ChIP and the fraction of SMM modules containing the identified consensus sequence. Sequences are aligned with database motifs for Mad (top) and Medea (bottom) predicted based on previous DNA-binding experiments. (b) Consensus sequence enrichment ($\log_2$(observed/expected frequency)) for Mad, Medea, and dCTCF/CP190 motifs

identified *de novo* by MEME-ChIP (*) along with previously characterized DNA-binding architectural protein consensus sequences. (c) Example genomics viewer illustration of SMM module overlap at architectural protein binding sites at the Ribosomal Protein L30 (*RpL30*) gene promoter. (d) Percentage of SMM modules overlapping architectural protein binding sites for dCTCF, BEAF-32, CP190, Rad21, Cap-H2, Mod(mdg4), and Chromator. Comparison of observed overlap frequencies (blue) with randomized shuffle control peaks (yellow) illustrates the strong enrichment of SMM modules at APBSs.

protein consensus sequences targeted by dCTCF, BEAF-32, and CP190, whereas sequences targeted by the gypsy-binding architectural protein, Suppressor of Hairy-wing (Su(Hw)), are depleted in SMM modules (Figure 18B). This result suggests that genomic elements bound by R-Smads, Medea, and Schnurri often occur in close spatial proximity with APBSs, and that architectural proteins may play an important role in either creating an accessible chromatin landscape for SMM occupancy, or may themselves directly recruit Smads to DNA.

Visualization of SMM modules overlapping APBSs illustrates that ChIP-seq read densities for BMP and TGF-β Smad proteins tightly intersect those of dCTCF and other architectural proteins (Figure 18C). Genome-wide, 62% of SMM modules co-localize with dCTCF ($p<0.0001$), and a majority of modules also co-localize with additional architectural proteins (Figure 18D). These data, together with motif enrichment profiles for architectural proteins in Smad binding sites, provides supporting evidence that architectural proteins may play an important role in Smad localization.

dCTCF-dependent co-localization of Smad proteins at APBSs

The strong overlap of BMP and TGF-β effectors with *Drosophila* CTCF suggest that previous reports of CTCF-dependent recruitment of Smad proteins in humans may be an important, highly conserved genome-wide phenomenon. In order to test whether Smad proteins also target *Drosophila* APBSs in a dCTCF-dependent manner, we repeated ChIP-seq experiments for Mad, dSmad2, and Medea in cell culture depleted for dCTCF by RNAi [46, 253]. In all cases, disruption of dCTCF levels significantly perturbs the levels of Smad ChIP-seq read density at a subset of sites (Figure 19A-C). For example,

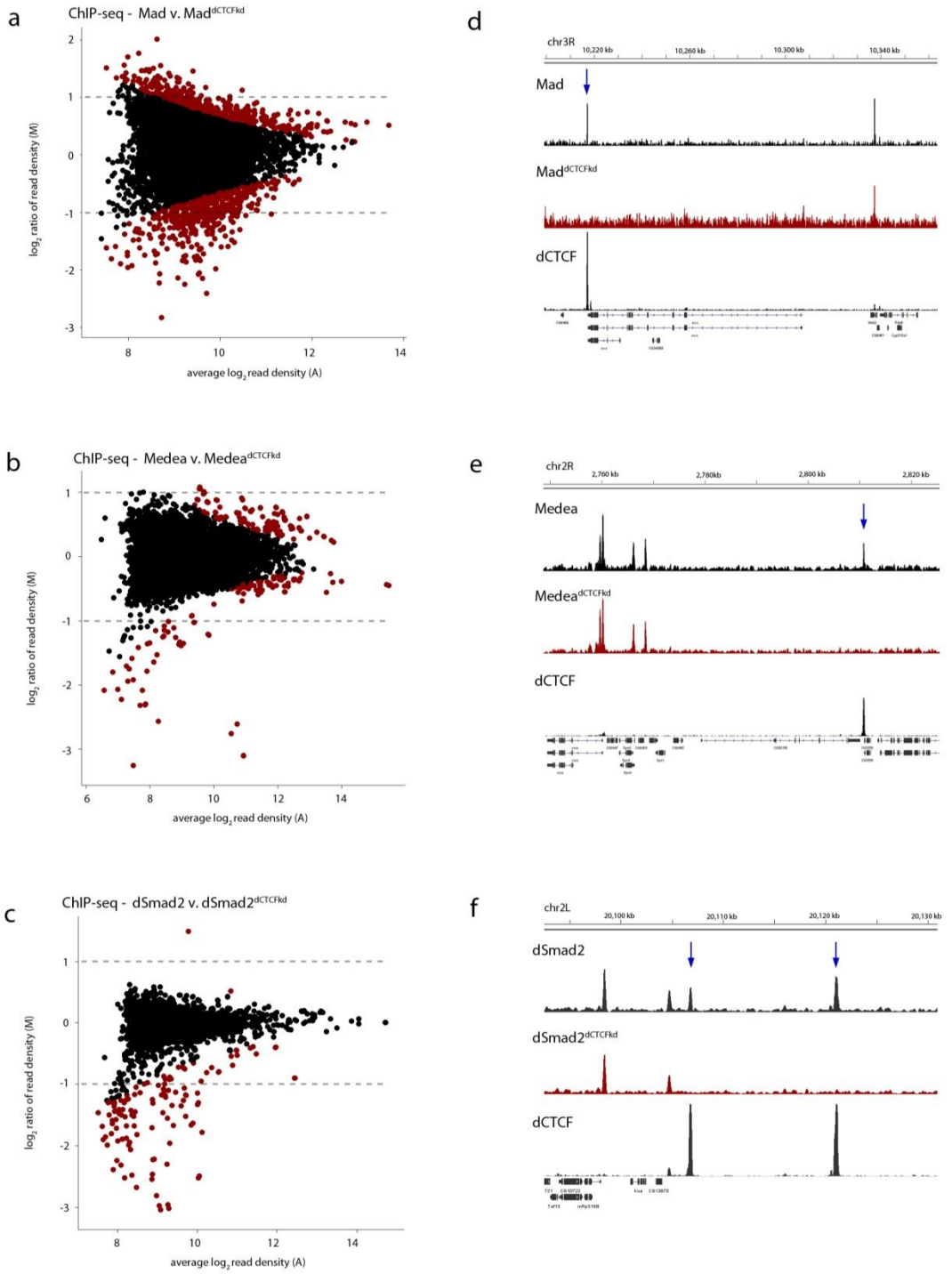


Figure 19. **RNAi depletion of *Drosophila* CTCF abrogates Smad localization to a subset of dCTCF binding sites.** (a-c) MA plots depicting changes in ChIP-seq read density as a function of average peak read densities for Mad, Medea, and dSmad2

respectively. dCTCF RNAi and control ChIP-seq experiments were normalized using the MAnorm package for quantitative comparison of ChIP-seq data sets [274]. An M value (log_2 ratio of ChIP-seq read density) threshold cutoff of 1 (dotted line) and a p value cutoff of 0.02 (red) were chosen for consideration of significantly changing peaks. (d-e) Genomic viewer comparison of dCTCF RNAi (red) and control ChIP-seq experiments (black) for Mad, Medea, and dSmad2 respectively. In all cases, several significant ChIP-seq peaks overlapping dCTCF are lost in response to knockdown of dCTCF (blue arrow), whereas non-overlapping Smad binding sites remain comparatively unchanged.

Mad occupancy significantly decreases at more than 200 binding sites (Figure 19A), and these sites are significantly enriched for the dCTCF motif. Whereas fewer dSmad2 and Medea peaks are significantly downregulated using the same threshold (absolute M value cutoff of 1, $p<0.02$), all Medea and dSmad2 peaks identified as being lost after dCTCF RNAi indeed overlap dCTCF ($p<0.0001$), suggesting loss of Smad binding is a direct consequence of dCTCF depletion.

Strikingly, Mad, Medea, and dSmad2 binding sites that are affected by dCTCF depletion are most often entirely lost after dCTCF RNAi (Figure 19D-E), suggesting recruitment of Smad proteins to these loci is entirely dependent on dCTCF. These results mirror the dependence of Smad localization to both the H19 imprinting control region and the Alzheimer β-amyloid precursor protein proximal promoter region on human CTCF. However, despite the significant drop in Smad occupancy at a subset of sites in *Drosophila* Kc167 cells, ChIP-seq read density for Mad, Medea, and dSmad2 is not otherwise depleted at all dCTCF sites (Supplemental Figure 1), suggesting SMM elements are differentially affected by, and perhaps differentially dependent on, dCTCF.

dCTCF-dependent Smad binding occurs at low occupancy APBSs within topological domains

Drosophila CTCF and other architectural proteins target thousands of regulatory elements that play unique roles in shaping chromosome organization and transcription. For example, long-range interactions mediated by architectural proteins can facilitate either active enhancer-promoter interactions or repressive polycomb response element interactions, whereas sites bound by numerous architectural proteins are involved in

establishing discrete topological domains and are capable of robust insulator function for which these proteins were originally characterized. We therefore asked whether Smad binding sites that are sensitive or insensitive to dCTCF RNAi occur within unique contexts. Here we compare Mad, Medea, and dSmad2 binding sites overlapping dCTCF that are lost in response to dCTCF depletion with overlapping SMM modules that do not change in response to RNAi.

Motif analysis reveals significant enrichment for the dCTCF core consensus sequence at sites in which Mad, Medea, and/or dSmad2 are lost after dCTCF knockdown (Figure 20A). In contrast, Smad binding sites in which dCTCF co-localization occurs independently of dCTCF are enriched for the BEAF-32 motif and consensus sequences known to be enriched at high occupancy APBSs. *Drosophila* CTCF-independent, overlapping Smad peaks are also enriched for the putative Mad and Medea motifs identified by MEME-ChIP, whereas sites sensitive to dCTCF knockdown are depleted for Mad and Medea consensus sequences. These differences in motif enrichment suggest potentially unique binding modes for Smad proteins at dCTCF-independent versus dCTCF-dependent peaks. Intuitively, Smad peaks containing direct target sequences are not sensitive to dCTCF depletion, whereas sites lacking direct binding motifs appear to be entirely dependent on dCTCF for recruitment to these loci.

Enrichment of the BEAF-32 consensus sequence and dCTCF/CP190 motifs present at high occupancy APBSs also suggests that, in addition to dCTCF, other architectural proteins may provide some redundancy in recruiting Smad proteins to APBSs. Comparison of DNase I hypersensitivity at dCTCF-sensitive and dCTCF-insensitive Smad peaks by DNase-seq in *Drosophila* Kc167 cells provides additional evidence that

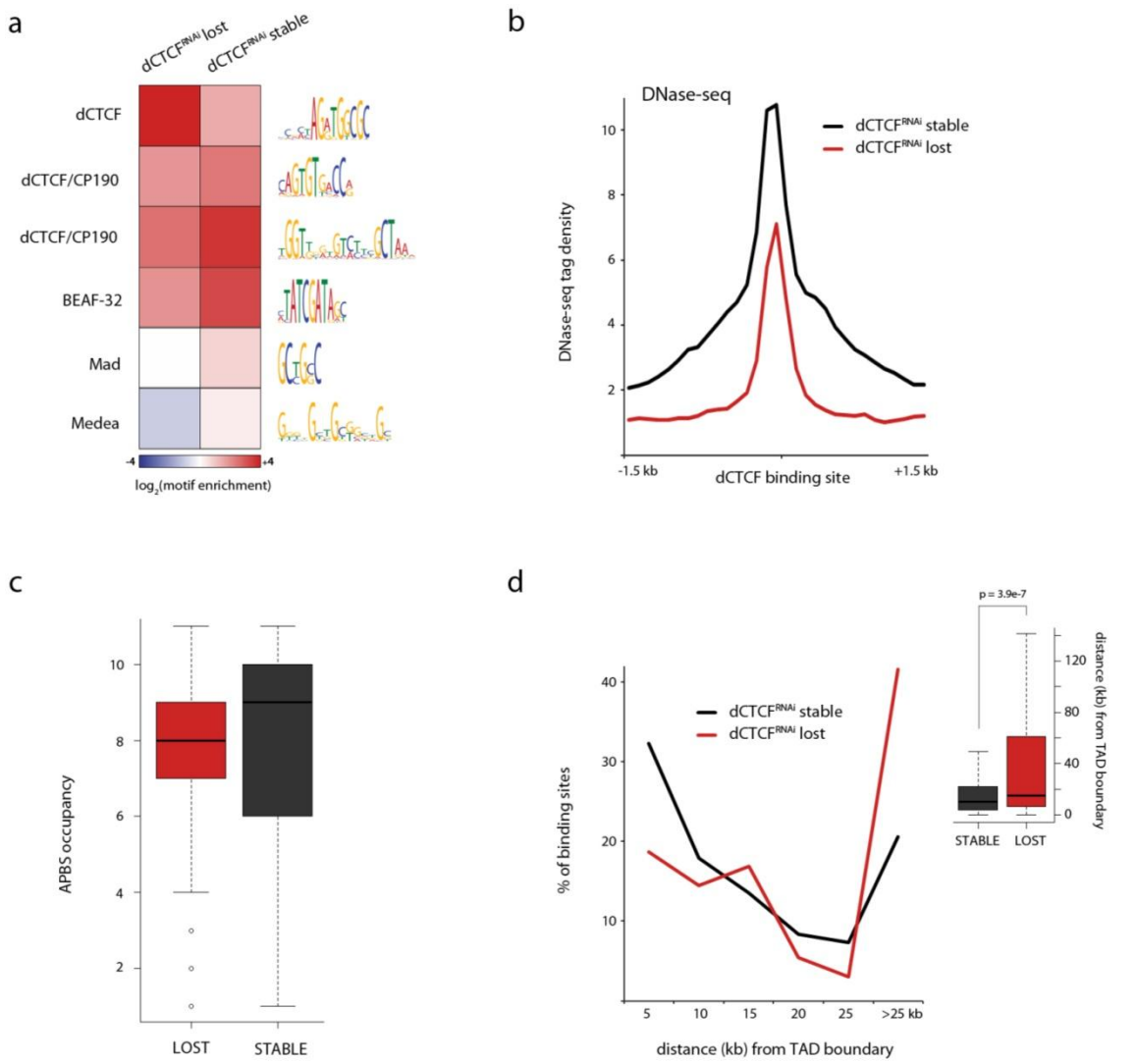


Figure 20. **Differential motif enrichment, architectural protein occupancy, and chromosomal location of dCTCF-dependent versus independent Smad binding sites**. (a) Motif enrichment (log_2(observed/expected frequency)) for architectural protein consensus sequences and putative Mad/Medea motifs at dCTCF^{RNAi} lost (left) and dCTCF^{RNAi} stable Smad binding sites. Comparison includes only Smad peaks which overlap dCTCF. (b) DNase-seq read densities centered on dCTCF binding sites overlapping Mad, Medea, and dSmad2 that are stable (black) or lost in response to dCTCF RNAi (red). (c) APBS occupancy, determined as the number of overlapping

MACS-called peaks for architectural proteins, at Smad binding sites that are stable (black) or lost after dCTCF knockdown (red) ($p = 0.061$, wilcoxon rank-sum test). (d) Distance (kb) from topologically associating domain (TAD) boundaries, at which high occupancy APBSs are highly enriched [253]. $dCTCF^{RNAi}$ lost Smad peaks are most abundant within domains, whereas $dCTCF^{RNAi}$ stable peaks are enriched near TAD boundaries (median distance from TAD borders: 14.9 kb and 9.9 kb respectively, $p = 3.9e\text{-}7$, wilcoxon rank-sum test).

dCTCF-independent binding occurs at high occupancy APBSs. For example, whereas Smad peaks that are lost after dCTCF knockdown exhibit sharp but comparatively weaker DNase I hypersensitivity, dCTCF-independent Smad binding sites are characterized by broad, robust DNase sensitivity commonly observed at high occupancy APBSs (Figure 20B). Accordingly, dCTCF-independent Smad peaks overlap on average more architectural proteins than dCTCF-dependent loci (Figure 20C). Though the difference in overlap does not reach statistical significance ($p = 0.061$, wilcoxon rank-sum test), this comparison draws on binary peak identification by MACS and does not take into account the relative ChIP-seq tag enrichment.

The observed differences in consensus sequence enrichment, DNase activity, and total architectural protein occupancy at differentially affected Smad binding sites suggest that dCTCF-dependent and dCTCF-independent Smad peaks are likely present in unique chromosomal contexts. Integration of ChIP-seq with genome-wide chromosomal interaction mapping has recently shown that interphase chromosomes are organized into discrete, self-interacting domains that are separated by high occupancy APBSs [253]. Accordingly, we find that Mad, Medea, and dSmad2 peaks that overlap dCTCF but are not significantly affected by dCTCF RNAi are enriched near the boundaries of topological domains (Figure 20D). In contrast, dCTCF-dependent Smad binding sites do not predominantly localize near TAD borders, consistent with the nature of low occupancy dCTCF binding sites residing within topological domains. Recent enhancer-trap assays suggest that individual topological domains functionally limit the activity of enhancers to genes that reside within the same TAD [275]. We speculate that whereas high occupancy APBSs mediate long-range higher order chromosomal domain

organization, low occupancy dCTCF binding sites may facilitate short-range enhancer-promoter interaction relevant to transcription. However, to what degree dCTCF binding plays a role in the TGF-β signaling response has not been previously addressed.

dCTCF binding remains constant in response to DPP-induced signaling events

In order to assay whether dCTCF binding plays a role in the actual TGF-β/Activin and BMP signaling pathways, we repeated ChIP-seq experiments for dCTCF as well as Mad, Medea, and Schnurri in response to the *Drosophila* bone morphogenic protein DPP. Treatment with 30 nM DPP for 6 hrs robustly activates phosphorylation of Mad in *Drosophila* Kc167 cells, whereas phospho-Mad is undetectable in untreated control cells (Figure 21A). Surprisingly, ligand-mediated activation of the BMP cascade does not dynamically alter the DNA-binding profile for dCTCF (Figure 21B), suggesting the architectural protein occupancy landscape remains relatively static. However, DPP does induce upregulation and downregulation of Mad, Medea, and Schnurri at several hundred sites (Figure 21C-E), suggesting Smad localization is dynamically altered in response to signal transduction.

We next performed gene ontology (GO) analysis on genes that are in close spatial proximity to SMM modules and near sites that increase or decrease in response to DPP (< 4kb from TSS). Whereas Mad, Medea, and Schnurri peaks that decrease are enriched near genes associated with imaginal disc development, consistent with SMM modules prior to induction with DPP, dynamically upregulated binding sites occur in close spatial proximity to genes associated with neurogenesis (Figure 21F), suggesting DPP may actively regulate neuronal cell fate decisions in this cell line. However, comparison of

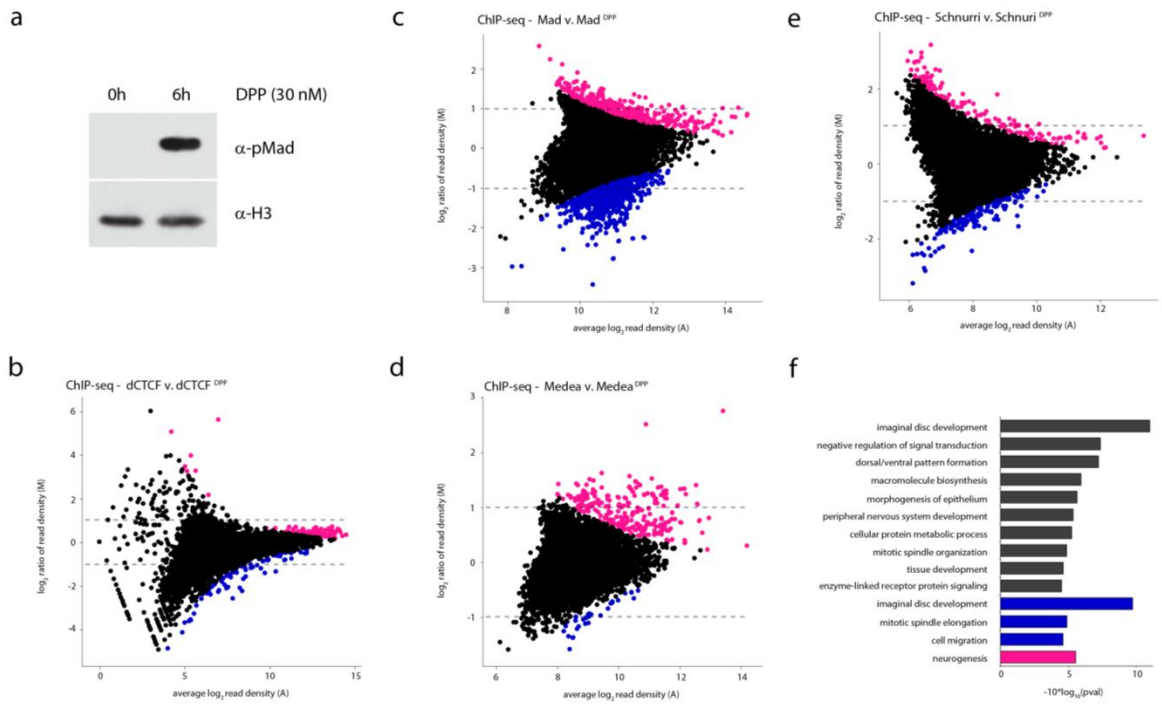


Figure 21. **Dynamic Smad localization in response to DPP activated phosphorylation of Mad**. (a) Western blot analysis of phosphorylated Mad levels before and after 6h treatment with DPP (30 nM), using phospho-specific p-Mad antibody. Loading control staining of histone H3. (b) MA plot depicting changes in dCTCF ChIP-seq read density as a function of average peak read density. DPP treatment and control ChIP-seq experiments were normalized using the MAnorm package for quantitative comparison of ChIP-seq data sets [274]. An M value (log_2 ratio of ChIP-seq read density) threshold cutoff of 1 (dotted line) and a p value cutoff of 0.02 (upregulated peaks – pink; downregulated peaks – blue) were chosen for consideration of significantly changing peaks. Only 40 peaks decrease and 7 peaks increase significantly in response to DPP. (c-e) Analogous MA plots for Mad, Medea, and Schnurri ChIP-seq experiments before and after treatment with DPP. Number of significantly changing peaks: Mad 161 increasing, 488 decreasing; Medea: 63 increasing, 9 decreasing; Schnurri: 90 increasing, 109

decreasing. (f) Gene ontology (GO) analysis for SMM modules and significantly changing peaks in response to DPP treatment. GO term enrichment was performed for closest genes (TSS) within 4 kb of Mad, Medea, and Schnurri binding sites before treatment (black), that decrease (blue), or increase (pink) significantly after treatment with DPP. Whereas decreasing Smad binding sites are enriched near genes associated with imaginal disc development and signaling, increasing binding sites occur near genes involved in neurogenesis.

upregulated Smad peaks with respect to dCTCF binding suggests that this signaling transcriptional response occurs in a non-dCTCF context. For example, upregulated Mad binding sites are depleted for dCTCF, and co-localization of Medea and Schnurri upregulated peaks with dCTCF is no greater than expected by random chance (Supplemental Figure 2). DPP-induced downregulation of Smad binding, on the other hand, does occur at dCTCF-binding sites, many of which coincide with low occupancy dCTCF APBSs, suggesting loss of Smad proteins occurs at architectural protein-bound regulatory elements within topological domains. The dynamic localization of Mad, Medea, and Schnurri from dCTCF sites to non-dCTCF sites suggests that DPP signaling may actively redirect Smad localization to unique chromatin contexts relevant to cell differentiation.

DISCUSSION

TGF-β effector proteins have been shown to co-localize with mammalian CTCF in a CTCF-dependent manner at just two individual loci. We now extend this observation to *Drosophila* using a genome-wide approach, providing evidence that architectural protein CTCF and canonical Smad signaling proteins, both highly conserved from fly to humans, co-localize on a global scale. Smad co-binding at dCTCF sites is abrogated at low occupancy dCTCF target sequences for which Smad consensus sequences are depleted, whereas high occupancy dCTCF binding sites co-bound by additional architectural proteins remain unaffected. The dCTCF-independent recruitment of Smads to high occupancy APBSs suggests that additional architectural proteins may redundantly recruit Smads, or simply provide an accessible chromatin landscape to which Mad, Medea, and dSmad2 can associate. Nevertheless, dCTCF-dependent localization of Smad proteins to specific low occupancy elements is consistent with the CTCF-dependent nature of Smad binding at both the APP and H19 promoters in humans [261, 262].

Surprisingly, DPP-activated phosphorylation of Mad does not lead to significant changes in dCTCF binding, whereas Mad, Medea, and Schnurri levels increase at regulatory elements away from dCTCF. These results suggest that TGF-β signaling in our cell line redirects Smad binding to genomic loci independent of architectural proteins, and that architectural proteins may facilitate binding of nuclear Smad proteins in the absence of signaling. In this regard, we indeed find Mad, dSmad2, Medea, and Schnurri bound to previously characterized response elements even in the absence of DPP ligand, in which levels of phosphorylated Mad are undetectable. The signal-independent clustering of these proteins suggests that the genomic TGF-β signaling response is not as simple as

regulating binary "off versus on" states, dependent on phosphorylated Mad. Conceivably, phosphorylation of Mad might instead regulate the resident time of DNA-binding, the recruitment of additional regulatory partners, or the ability to establish functional long-range interactions.

We speculate that dCTCF-dependent Smad localization to low occupancy APBSs within topological domains may represent regulatory elements involved in enhancer-promoter interactions, whereas dCTCF-independent high occupancy APBSs are involved in establishing higher-order chromosome organization. What role Smads might play in establishing or maintaining such long-range interactions relevant to chromosome architecture, or whether Smads and other transcription factors simply localize to high occupancy APBSs due to chromatin accessibility, remains difficult to address. However, we have recently shown that high occupancy APBSs are distinct from analogous transcription factor hotspots, suggesting some level of specificity, most likely governed by protein-protein interactions, decides which factors can associate and where. Alternatively, the enrichment of ChIP-seq signal at high occupancy APBSs may, to some degree, reflect indirect association via long-range interactions with regulatory elements directly bound by Smad proteins. This possibility raises a possible explanation for why Smad ChIP signal is independent of dCTCF binding at high occupancy APBSs.

The complete loss of Smad ChIP signal at numerous dCTCF binding sites enriched for the core dCTCF consensus sequence, nevertheless provides compelling evidence that recruitment of Smad proteins is directly governed by *Drosophila* CTCF at a subset of binding sites. These results establish CTCF as an important determinant of Smad localization and, depending on the cell-type specific binding patterns of CTCF, suggest

that CTCF might also influence the tissue-specific localization of Smad proteins analogous to master regulatory transcription factors in multi-potent stem cells.

EXPERIMENTAL PROCEDURES

Cell Culture and DPP Treatment

Drosophila Kc167 cells were grown in HyClone SFX cell culture medium. For treatment with DPP, cells were split overnight to a density of 0.5 x 10^6 cells/mL and treated with 30 nM DPP the follow morning. DPP treatment and control lysates were extracted after 6hr incubations. RNAi knockdown of *Drosophila* CTCF in cell culture was conducted as per the *Drosophila* RNAi Screening Center (DRSC) protocol [197] with the exception that dsRNA was added every day for 3 days and the cells were then collected on the 4^{th} day. Chromatin isolation from dCTCF depleted cell culture were performed as part of previously published knockdown experiments and described protocols [46].

Immunoprecipitation and Western Analysis

Chromatin immunoprecipitation was performed as previously described [58]. Protein A Sepharose beads (GE Healthcare) were pre-washed in PBSMT and pre-incubated with 6 µl of rabbit polyclonal anti-Mad [276], sheep polyclonal anti-dSmad2 (R&D systems cat#AF7948), rabbit polyclonal anti-Medea [277], or rabbit polyclonal anti-Schnurri [278] for 1 hr prior to pull-down of sheared DNA. To generate sequencing libraries, ChIP DNA was prepared for adaptor ligation by end repair (End-It DNA End Repair Kit, Epicenter Cat# ER0720) and addition of "A" base to 3′ ends (Klenow 3′-5′ exo–, NEB Cat# M0212S). Illumina adaptors (Illumina Cat# PE-102-1001) were titrated according to prepared DNA ChIP sample concentration and ligated with T4 ligase (NEB Cat# M0202S). Ligated ChIP samples were PCR-amplified using Illumina primers and Phusion DNA polymerase (NEB Cat# F-530L) and size selected for 200–300 bp by gel

extraction. ChIP libraries were sequenced at the HudsonAlpha Institute for Biotechnology, using an Illumina HiSeq 2000.

For Western blotting, membranes were blocked in TBST (20 mM Tris, pH7.4, 150 mM NaCl, 0.05% Tween 20) with 5% nonfat milk powder and incubated overnight with phospho-specific Mad antibodies (Gift from Lan Xu [276]). Membranes were washed three times with TBST and probed with secondary antibodies-conjugated to HRP (1:5000, Jackson ImmunoResearch Laboratories) for 1 hr. After three more washes, the presence of different proteins was detected using SuperSignal West Pico/Dura Chemiluminescent substrate (Thermo Scientific).

ChIP-seq and Reference Data

In addition to novel Mad, Medea, and Schnurri ChIP experiments, architectural protein ChIP-seq data was obtained from previously published sources [46, 253]. Sequences were mapped to the dm3 genome with Bowtie 0.12.3 [189], using default settings. Peaks were then called with MACS 1.4.0alpha2 [190] using equal numbers of unique reads for IgG control and ChIP samples, using a p value cutoff of 1×10^{-10}. For classification of overlapping binding sites, MACS-identified peaks were intersected using publicly available tools on Galaxy [191-193]. For visualization, mapped sequence reads were loaded on to the Integrated Genomics Viewer [239, 240]. Previously published ChIP-seq data for *Drosophila* architectural proteins were obtained from GEO accessions GSE30740 [77] and GSE36944 [46]. Raw DNase-seq in Kc167 cells was obtained from published resources [225];

Bioinformatics analyses

DNA sequence motifs were identified by MEME-ChIP set to identify 12 motifs and otherwise using default settings [28]. Comparison of APBSs with respect to Pol II-transcribed genes employed gene structure (TSSs) obtained using the UCSC genome browser [248, 249]. Significantly changing ChIP-seq peaks in response to dCTCF RNAi and DPP signaling were determined using the MAnorm package for quantitative comparison of ChIP-seq data sets [274]. An M value ($\log_2$ ratio of ChIP-seq read density) threshold cutoff of 1 and a p value cutoff of 0.02 were chosen for consideration of significantly changing peaks. Enrichment profiles for Mad, Medea, Schnurri, and dSmad2 protein co-occurrence and overlap with architectural proteins were defined as the observed overlapping frequencies over expected frequencies determined by shuffling datasets, while controlling for the number of peaks and start/stop location of peaks on each chromosome. P-values were determined as the chance of observing an equal or greater co-occurrence across 100,000 Monte Carlo Permutation tests. Unless otherwise noted, read intensity plots were generated by binning ChIP-seq reads into 100 bp bins and extracting read counts in bins surrounding described anchor points (eg. SMM modules or dCTCF peaks), and visualized using Java Treeview [195]. Gene Ontology analysis was performed by running functional annotation tools on DAVID [279] and biological process (level 5) GO terms were subsequently summarized using REVIGO [280].

ACKNOWLEDGEMENTS

We are especially thankful to Drs. Lan Xu, Laurel Raftery, and Mark Biggin for graciously sharing antibodies against Mad (and p-Mad), Medea, and Schnurri. We thank the Genomic Services Lab at the HudsonAlpha Institute for Biotechnology for help in performing Illumina sequencing of ChIP-Seq samples. This work was supported by U.S. Public Health Service Award R01GM035463 from the National Institutes of Health. The content is solely the responsibility of the authors and does not represent the official views of the National Institutes of Health.

CHAPTER 5:

DYNAMIC RECRUITMENT OF REGULATORY FACTORS TO ARCHITECTURAL PROTEIN BINDING SITES DURING STEROID-HORMONE SIGNALING EVENTS

Data discussed in this chapter were generated in a previously published research article:

Wood AM, **Van Bortle** K, Ramos E, Takenaka N, Rohrbaugh M, Jones BC, Jones KC, Corces VG. Mol Cell. 2011. Regulation of chromatin organization and inducible gene expression by a Drosophila insulator. Oct 7;44(1):29-38

My contributions include performing the ChIP-seq experiments and computational analysis of the data.

ABSTRACT

Architectural proteins mediate long-range interactions that contribute to the three-dimensional organization of interphase chromosomes. We have shown that recently identified topologically associating domains (TADs), which represent self-interacting regions of the chromosome, are partitioned by high occupancy architectural protein binding sites (APBSs). Both TADs and high occupancy APBSs appear to be largely tissue-invariant, suggesting modular chromatin domains are conserved throughout development. However, intra-TAD interactions, which may be facilitated by low occupancy APBSs within domains, often consist of enhancer-promoter and promoter-promoter interactions that are likely to be dynamically regulated in response to cell signaling events and between cell types to produce cell-type specific transcription patterns. Here we assay the binding profiles for architectural proteins in response to steroid-hormone signaling events in *Drosophila melanogaster* and discover dynamic binding of DNA-binding proteins within topological domains. Ecdysone-induced architectural protein binding correlates with changes in chromosome organization, suggesting a role for architectural proteins in mediating the transcriptional response to signaling.

INTRODUCTION

Spatiotemporal pulses of the steroid hormone 20-hydroxyecdysone (ecdysone, 20HE) occur during insect metamorphosis, creating tissue-specific physiological responses critical for development [254]. Ecdysone is bound by the nuclear ecdysone receptor, a ligand-activated transcription factor that in turn activates specific genes. In *Drosophila* Kc167 cells, ecdysone induces differentiation by activating and repressing transcription with temporal specificity [281]. Whereas only a couple dozen genes directly targeted by the ecdysone receptor are differentially expressed after three hours, several hundred genes are dynamically regulated after forty-eight hours, producing significant changes in cell morphology, cytoplasmic organization, and cellular identity. Ecdysone therefore provides an excellent model system for assaying the organization and stability of architectural protein binding in response to cellular differentiation and changes in transcription.

Here we assay architectural protein binding profiles by ChIP-seq in response to ecdysone induced differentiation of *Drosophila* Kc167 plasmatocytes. We find that a small subset of regulatory elements targeted by DNA-binding architectural proteins dCTCF, BEAF-32, and Su(Hw) are up- and down-regulated after 3hrs and 48hrs treatment. More drastically, BTB-containing protein CP190, which is essential for insulator activity in *D. melanogaster*, is dynamically targeted to several hundred loci specifically after 48hrs. Ecdysone induced CP190 recruitment correlates with changes in long-range interaction frequencies, suggesting dynamic recruitment of regulatory factors to architectural protein binding sites may functionally influence chromosome organization and transcription.

RESULTS

Ecdysone treatment leads to changes in the genome-wide distribution of architectural proteins

To determine whether architectural protein localization is modulated concomitant to changes in ecdysone-induced transcription patterns, we performed ChIP-seq analysis for DNA-binding proteins dCTCF, BEAF-32, and Su(Hw), as well as for BTB-containing architectural protein CP190, which is recruited by each DNA-binding protein and required for insulator activity. Binding profiles were assayed at 3hrs and 48hrs post ecdysone treatment as well as in control treated cells. Differential binding among each architectural protein was determined by normalizing ChIP experiments by read count, filtering reads to remove random background noise, and calling significant changes in occupancy using CMARRT, a modified correlation and moving average algorithm [282], taking into account log2() ratios between samples within ChIP peaks containing a p value less than 1e-10.

Comparison of architectural protein binding before and after ecdysone treatments suggests that a majority of APBSs remain intact despite the robust alteration in transcription induced by 20HE. This finding is consistent with our recent observation in which medium-high occupancy APBSs are conserved throughout *Drosophila* development and between cell types [253]. Nevertheless, many APBSs are significantly up- or down-regulated (Figure 22A). For example, compared to control treated cells, BEAF -32 binding decreases at 135 sites (2%) at 3hrs and at 65 sites (1%) after 48hrs, of which 58 binding sites are common between the two time points. Similarly, dCTCF

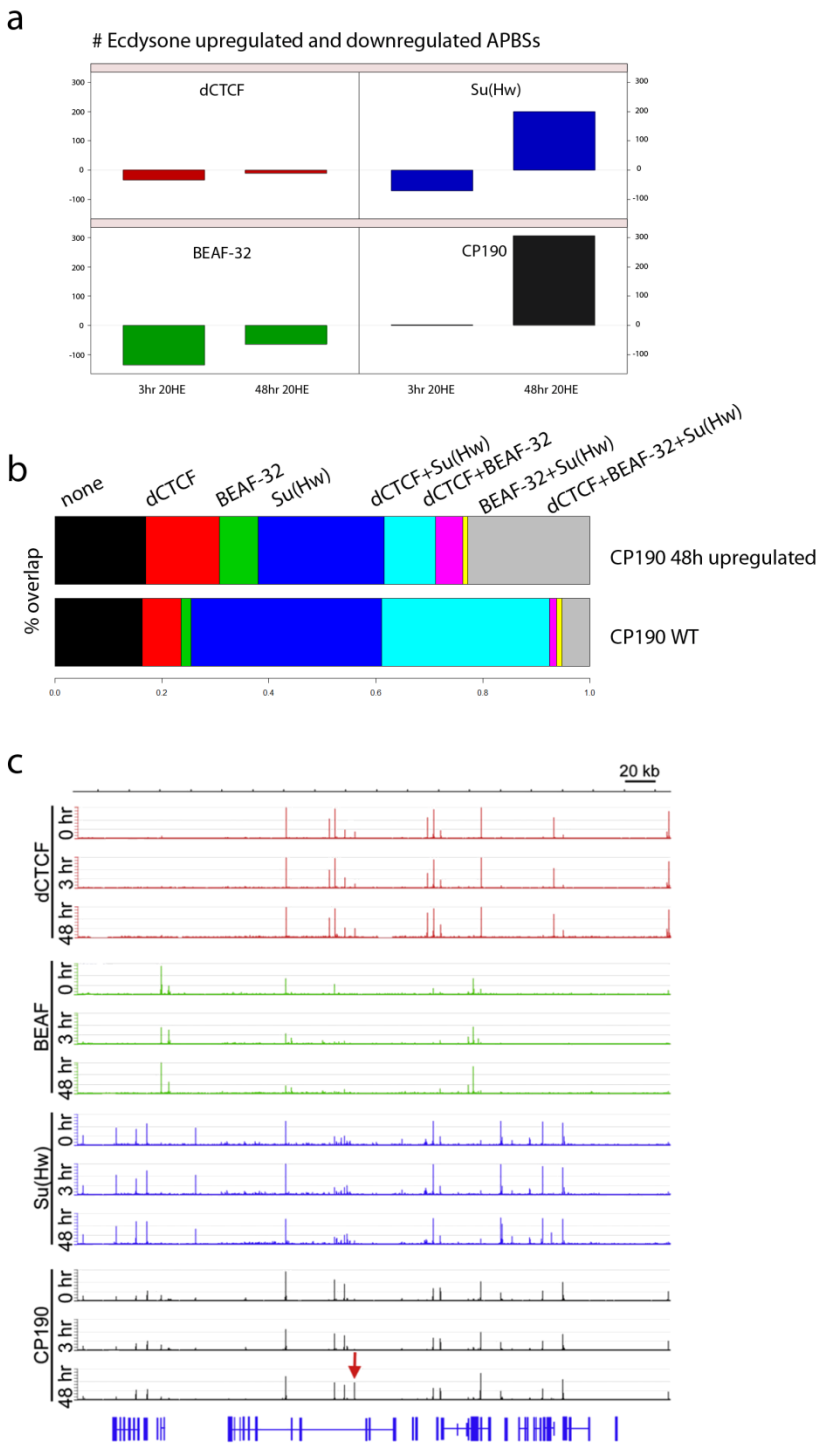


Figure 22. **Ecdysone treatment leads to changes in the genome-wide distribution of architectural proteins.** (a) Number of architectural protein binding sites (APBSs)

regulated for each protein at each time point after ecdysone treatment. (b) Percentage of CP190 upregulated sites (306, 48hrs from panel A) and percentage of all CP190 binding sites that overlap DNA binding architectural proteins dCTCF, BEAF-32, and Su(Hw). (c) Example genomics viewer of ChIP-seq tag density at 0hrs, 3hrs, and 48hrs for each architectural protein. Eip75B gene contains significantly upregulated CP190 binding site (red arrow), overlapping dCTCF.

binding is downregulated at 34 sites (0.6%) at 3hrs, and at 11sites (0.2%) at 48hrs after treatment with ecdysone. Su(Hw), on the other hand, shows reduced binding at 70 sites (1%) at 3hrs, then an increase in binding at 200 sites (2%) after 48hrs. Most dramatically, CP190 binding is upregulated at 306 binding sites specifically after 48hrs, suggesting 20HE modulates the recruitment of CP190 to DNA-binding proteins in a context dependent manner.

Further analysis of upregulated CP190 binding sites after 48hrs reveals that these sites show enriched colocalization for dCTCF, BEAF-32, dCTCF+BEAF-32, and dCTCF+BEAF-32+Su(Hw) (Figure 22B), suggesting a non-random pattern for CP190 upregulation which may relate to the interaction and transcriptional status of the cell. To test whether CP190 dynamics influence interactions, we next focused our attention on a dCTCF APBS with induced CP190 colocalization specifically after 48hrs (Figure 22C, red arrow). This dynamic binding site is present within the ecdysone-induced protein 75B (Eip75B) gene, which increases dramatically in response to 20HE. Chromosome conformation capture (3C) analysis, using this region as an anchor, further reveals strengthened interactions with downstream HindIII fragments specifically after 48hrs [77]. These interactions are significantly weakened in response to CP190 depletion, suggesting ecdysone-regulated chromatin conformation surrounding Eip75B requires the architectural protein CP190.

DISCUSSION

By mapping the binding profiles for architectural proteins in response to extracellular signaling events, we identify specific architectural protein binding sites (APBSs) that are dynamically bound, suggesting a role in regulating chromatin organization. Supporting evidence for this possibility comes from differential interaction frequencies measured by chromosome conformation capture at a dCTCF binding site where CP190 recruitment is induced by ecdysone [77]. However, in contrast to the hundreds of differential binding sites observed for BTB-containing protein CP190 after 48hr 20HE, sites bound by DNA-binding proteins Su(Hw), BEAF-32, and dCTCF remain relatively stable, consistent with similar analyses of dCTCF binding during BMP signaling in response to DPP. These data suggests that DNA-binding architectural protein binding is not significantly influenced by signaling events, whereas recruitment of co-factors involved in long-range interactions and transcription may provide a mechanism for dynamically regulating both chromosome organization and transcription.

We interpret the comparatively static binding of DNA-binding architectural proteins to mean that these factors instead create stably accessible chromatin environments that are regulated by the recruitment of additional factors, such as CP190. We speculate that although treatment of Kc167 cells with 20HE and DPP leads to relatively minor changes in dCTCF binding, a more robust differentiation induction or comparison of diverse cell types may likely shed light on the degree to which dCTCF binding is tissue-specific. Given the role of dCTCF and other DNA-binding architectural proteins in recruiting non-DNA-binding co-factors, it is likely that tissue-specific APBSs delineate the ability or

direction for which a given cell-type can respond to extracellular stimuli. In the case of ecdysone signaling, for example, one can imagine a scenario in which recruitment of CP190 and other regulatory factors facilitated by dCTCF is very different in two unique cell-types, dependent on where dCTCF is bound. The transcriptional induction would then be, to some degree, directed by the cell-type specific dCTCF-interactome.

EXPERIMENTAL PROCEDURES

Cell culture and ecdysone/DPP treatments

Kc167 cells were grown in CCM3 serum-free insect media (HyClone SH30065.01) at 25 C. For ecdysone treatment, cells were plated at $0.5*10^{-6}$ cells/ml and grown overnight. Cells were treated with 20-HE (Sigma) at a final concentration of $0.5*10^{-7}$ M for 3 or 48 hrs prior to cellular lysis. For control (WT) experiments, cells were treated with EtOH.

Chromatin immunoprecipitation and sequencing

Chromatin immunoprecipitation was performed as described [58]. To generate sequencing libraries, ChIP DNA was prepared for adaptor ligation by end repair (End-It DNA End Repair Kit, Epicenter Cat# ER0720) and addition of ‘‘A’’ base to 30 ends (Klenow 30-50 exo–, NEB Cat# M0212S). Illumina adaptors (Illumina Cat# PE-102-1001) were titrated according to prepared DNA ChIP sample concentration and ligated with T4 ligase (NEB Cat# M0202S). Ligated ChIP samples were PCR-amplified using Illumina primers and Phusion DNA polymerase (NEB Cat# F-530L) and size selected for 200–300 bp by gel extraction. ChIP libraries were sequenced at the HudsonAlpha Institute for Biotechnology, using an Illumina GAII sequencer. Sequences were mapped to the dm3 genome with Bowtie 0.12.3 [189] using default settings. Peaks were then called with MACS 1.4.0alpha2 [190] using equal numbers of unique reads for input and ChIP samples and a p value cutoff of 1e-10. To determine up- and downregulated insulator sites at each time point, ChIP-Seq data for each insulator protein were normalized by total read count between the different time points. The data were then

filtered by a second normalization for lane effect differences of the mean plus 1/2 standard deviation of lane-specific read counts. The ChIP-time point-A/ChIP-time point-B Log2() ratio was determined and assigned for every 100 bp. Significant changes in insulator occupancy were then called as up- or downregulated using CMARRT, a modified correlation and moving average algorithm [282], and a stringent p value cutoff of 1e-10.

ACKNOWLEDGEMENTS

We thank Naomi Takenaka for library preparation of ChIP-seq libraries, Ashley Wood for performing CP190 ChIP-seq experiments, and Eddie Ramos for assistance with the analysis of differential binding architectural protein binding during the ecdysone response. Also, we would like to thank the Genomic Services Lab at HudsonAlpha for assistance with high-throughput sequencing. Ecdysone-induced architectural protein dynamic ChIP-seq datasets are published in: Mol Cell. 2011 Oct 7;44(1):29-38. doi: 10.1016/j.molcel.2011.07.035.

CHAPTER 6:

DYNAMIC CODON USAGE AND tRNA ABUNDANCE IN RESPONSE TO ECDYSONE INDUCED DIFFERENTIATION

Work discussed in this chapter is currently in preparation:

Van Bortle K, Nichols M, Ramos E, Corces VG. Dynamic codon usage and tRNA abundance in response to ecdysone induced differentiation. *In preparation*.

My contributions include conceiving the project, performing the experiments, computational analysis of the data, interpretation of the results, and writing the manuscript.

ABSTRACT

The insect steroid hormone 20-hydroxyecdysone (ecdysone, 20HE) controls *Drosophila* metamorphosis by transducing extracellular signals into tissue-specific expression patterns through the ligand-activated transcription factor ecdysone receptor (EcR). By profiling the transcriptional, small RNA, and proteomic response to ecdysone in Kc167 cells, a late-stage embryonic cell line, we discover a shift in codon usage and codon usage bias associated with differentiation. In particular, messenger RNAs (mRNAs) and proteins upregulated by 20HE show a bias towards codons enriched in differentiation genes, and a depletion for codons enriched in proliferation genes. Cellular tRNA pools concomitantly shift in response to 20HE, albeit moderately, in favor of codons associated with differentiation. Interestingly, $tRNA^{SeC}$, a single copy tRNA gene required for selenoprotein synthesis, is upregulated by ecdysone, suggesting 20HE may regulate the biosynthesis of the antioxidant selenoproteins. Finally, we show that tRNA abundance predicts preferential codon usage throughout the Kc167 proteome compared to cellular mRNA codon usage, and that significantly changing tRNA isoacceptors become enriched or depleted accordingly in proteins after treatment with ecdysone. These results support the translational efficiency hypothesis in which cellular tRNA levels influence the translational output of the cell, and further provide evidence of mRNA and tRNA dynamics that favor balanced codon usage and tRNA abundance.

INTRODUCTION

Of the twenty amino acids encoded by the standard genetic code, eighteen are defined by more than one codon. Despite this redundancy, synonymous codons are often used unequally, creating a bias for preferential codons and against suboptimal codons. The resulting codon usage bias (CUB) for specific codons has been suggested to play a role in favoring translational accuracy and/or translational efficiency [283]. By comparing the frequency of codon usage with tRNA gene copy number as a proxy for tRNA abundance in a cell, studies have drawn support for the translational efficiency hypothesis [284]. For example, across many organisms, preferential codons often correlate with high abundance tRNA isoacceptors, whereas suboptimal codons correlate with low level tRNAs [285]. These findings suggest that highly expressed genes, which are often enriched for preferential codons, may be rapidly translated by pairing with highly abundant cognate tRNAs [283].

Despite circumstantial evidence supporting the translational efficiency hypothesis, recent studies profiling ribosome occupancy on mRNAs suggest that codon selection time, or the time a ribosome must wait for the correct tRNA isoacceptor, is actually similar for both high frequency and low frequency codons [286, 287]. As a consequence, a recent alternative hypothesis was proposed in which balancing tRNA abundance with codon usage creates an equal concentration for each tRNA isoacceptor [288]. The major assumption of this model requires that aminoacylated tRNAs be in short supply to translating ribosomes, yet to what degree this is true in different organisms and cell types is currently unknown.

Regardless of whether tRNAs influence translational efficiency through differential abundance or by creating an equal pool when balanced against codon frequency, tRNA levels are finely tuned with the codon usage across diverse organisms and all domains of life. This strongly suggests that tRNA abundance plays an important role in optimizing the translational output of cell. Taking the next step, it is conceivable that within an organism, different cell types may express a unique codon usage that would require dynamic regulation of the tRNA pool. The increasing demand for one codon, for example, would benefit by increasing the relative abundance of the cognate tRNA, and so forth.

A recent survey of tRNA abundance and codon usage across numerous proliferative and differentiated cell types suggests that tRNA levels are indeed dynamically tuned towards the codon usage in a given cell [289]. RNA expression analysis shows that proliferative cancer cell lines and cells overexpressing the transcription factor Myc tend to have a unique codon bias compared to differentiated cells or those with induced senescence. Accordingly, tRNA abundances in differentiated cells were better matched for differentiated codon bias, and vice versa. However, to what degree tRNA expression is modulated in a single cell type in response to differentiation, and whether or not changes in tRNA abundance produce differences in translational output remain poorly addressed.

The insect steroid hormone 20-hydroxyecdysone (ecdysone, 20HE) provides an advantageous model system for studying the effects of differentiation on codon usage and translation efficiency. Here we show that in *Drosophila* Kc167 cells, 20HE induces plastmatocyte-to-macrophage differentiation and changes in codon usage that mirror the

differentiation status of the cell. By further profiling the transcriptional, small RNA, and proteomic response to ecdysone we show that cellular tRNA pools concomitantly shift in response to 20HE in favor of codons associated with differentiation. tRNA abundance scales with preferential codon usage in proteins when compared to cellular mRNA codon usage, and significantly changing tRNA isoacceptors become enriched or depleted accordingly in proteins after treatment with ecdysone. These results support the translational efficiency hypothesis in which cellular tRNA levels influence the translational output of the cell, and further provide evidence of mRNA and tRNA dynamics that favor balanced codon usage and tRNA abundance.

RESULTS

Ecdysone induces differentiation of *Drosophila* Kc167 plasmatocyte cells into macrophages

In order to assay both codon usage and tRNA abundance, as well as the effect of ecdysone-induced differentiation, we profiled small RNAs by RNA-seq, mRNAs by microarray analysis, and protein abundance using label-free quantitative mass spectrometry (Figure 23A). Transcriptome analyses were performed in control and at both 3hrs and 48hrs after treatment with 20HE, whereas proteomic samples were compared between control and 48hr time points. Quantitative real-time PCR was used to validate the ecdysone response prior to performing transcriptomic and proteomic assays (Figure 23B).

Though *Drosophila* Kc167 cells are often utilized for their robust morphological response to ecdysone, to our knowledge, no previous study has defined the cellular identity of these differentiated cells. Taking advantage of microarray analyses, the expression of plasmatocyte markers before and after treatment with 20HE suggests that Kc167 cells are a type of hemocyte as recently proposed [290]. Studies suggest that plasmatocytes may differentiate into several unique immune cell types, including lamellocytes, crystal cells, nephrocytes, and macrophages [291]. Comparison of expected expression markers for these cell types provides evidence that Kc167 cells differentiate into activated macrophages in response to ecdysone. For example, expression of Croquemort (crq), a macrophage receptor protein important for phagocytosis, and Singed (aka Fascin, sn), a highly expressed factor critical for macrophage motility and

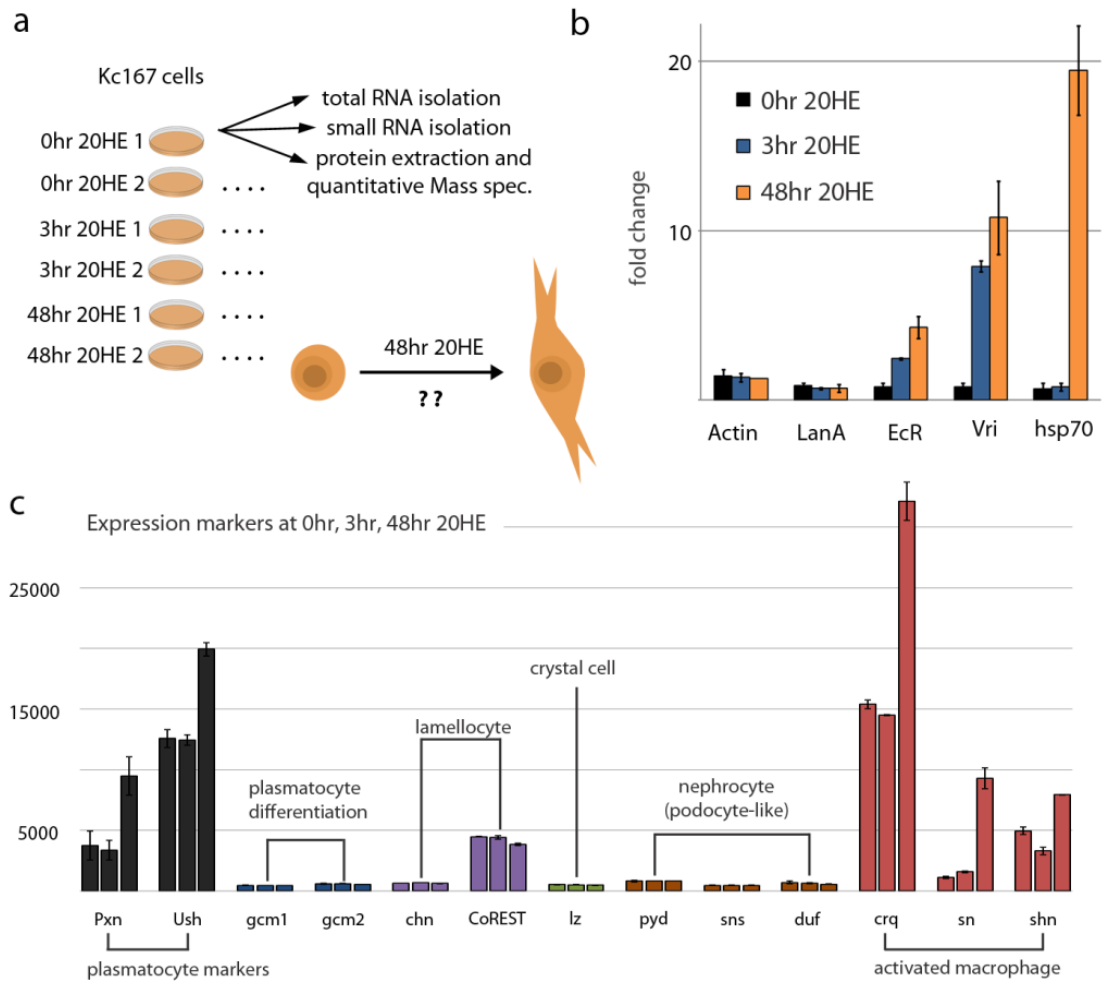


Figure 23. **Ecdysone induces differentiation of *Drosophila* Kc167 plasmatocyte cells into macrophages.** (a) Experimental setup: transcriptomic and proteomic profiling before and after treatment with 20HE; two biological replicates per time point. (b) Quantitative real-time PCR validation of 20HE response in Kc167 after 3hrs, 48hrs 20HE. (c) Key expression markers for potential cell-types that are derived from the plasmatocyte (hemocyte) lineage. Expression of Croquemort (crq) and singed (sn) suggest plasmatocyte-to-macrophage differentiation in response to ecdysone.

developmental migration, significantly increase after 48hrs 20HE (Figure 23C). Likewise, expression of Schnurri (shn), a transcriptional repressor recently shown to induce macrophage-like morphology in retinal hemocytes [292], also increases. Together, these data suggest that after 48hrs, ecdysone has induced Kc167 differentiation into macrophage cells.

Transcriptome and quantitative proteomics after 48h ecdysone treatment

Temporal transcriptional analyses in Kc167 cells has demonstrated an early and late response to 20HE, wherein only a few dozen genes are upregulated within the first couple hours and several hundred genes are differentially expressed 1-2 days after treatment with ecdysone [281]. Our microarray analysis of gene expression at 3hrs and 48hrs independently confirm these unique responses to 20HE. Specifically, 53 transcripts increase significantly after 3hrs, whereas 612 transcripts increase and 240 decrease more than 2 fold after 48hrs (Figure 24A-B). Quantitative proteomic analysis similarly identifies several hundred proteins that increase or decrease after 48hrs with an FDR cutoff of 5% (Figure 24C), with 276 proteins increasing and 129 proteins decreasing. Correlation analysis of the proteomic response with mRNA levels before and after ecdysone-induced differentiation shows a strong relationship between changes in protein levels and transcript abundance (Figure 24D, Spearman's rank correlation coefficient = 0.627, $p<2.2e-16$), suggesting our transcriptomic and proteomic profiles accurately reflect cellular dynamics in response to ecdysone signaling. Together, these results illustrate the robust molecular response to ecdysone in *Drosophila* Kc167 cells.

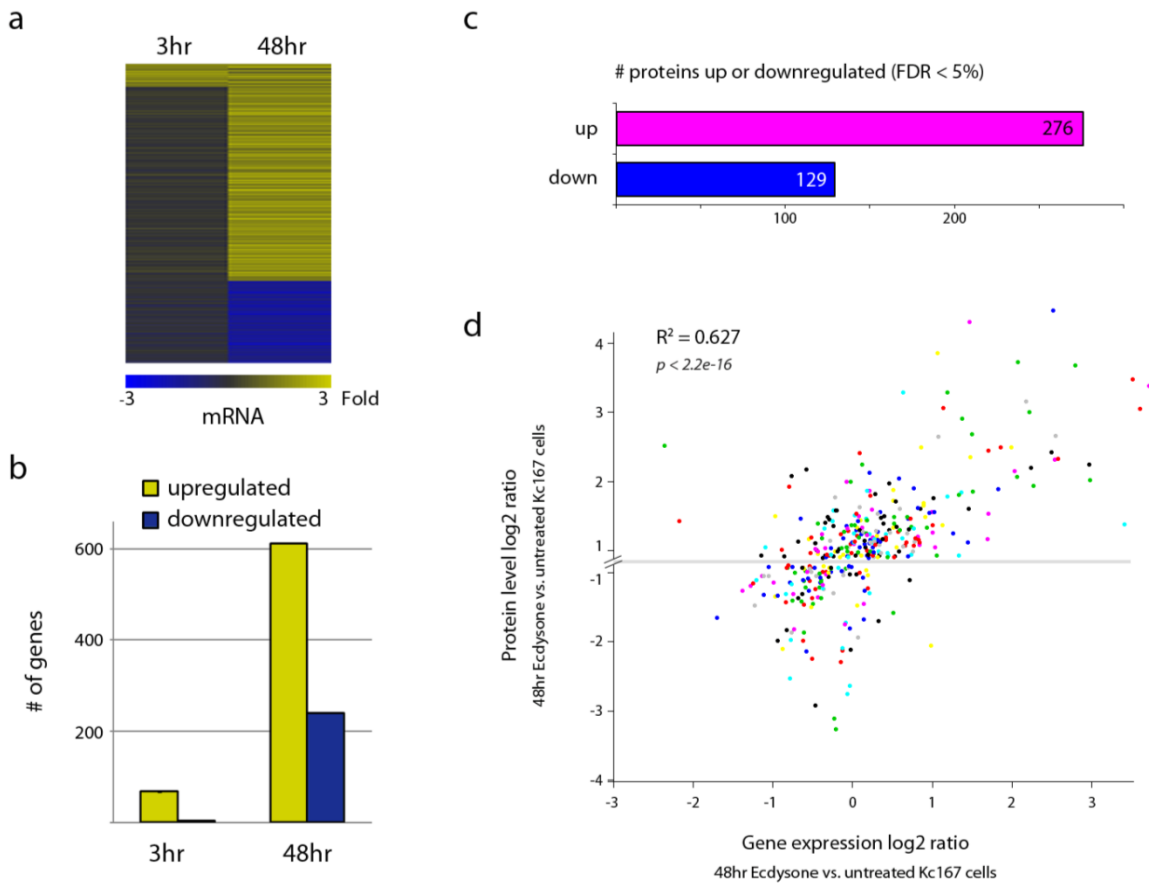


Figure 24. **Gene expression and label-free quantitative mass spectrometry analysis after 48hrs ecdysone treatment.** (a) Gene expression heatmap for genes significantly changing at 3hrs and/or 48hrs after ecdysone treatment. (b) Number of genes increasing or decreasing more than 2 fold after 3hrs and 48hrs 20HE. (c) Number of proteins significantly increasing or decreasing (FDR <5%) after 48hrs 20HE. (d) Correlation plot for protein (y-axis) and mRNA (x-axis) dynamics after 48hrs 20HE. Spearman's rank correlation coefficient = 0.627, p<2.2e-16.

Codon usage after 48hrs 20HE mirrors preferential codon incorporation in differentiation genes

In order to determine whether ecdysone induces Kc167 cells to change the frequency in which specific codons are used, we normalized codon usage in *D. melanogaster* by the expression level of a gene or protein. In other words, codons are counted by the number of times they are observed in the transcriptome or proteome, rather than the reference genome (Figure 25A-B). In this way, mRNA levels provide an accurate measure of the comparative frequency in which a codon is used, and protein levels provide a readout of the number of times a codon was translated.

Comparison of codon usage within the transcriptome of cells treated or untreated with ecdysone reveals small but significant changes in codon frequency. For example, tyrosine codon UAU and lysine codon AAG are the most dynamic codons after 48hrs, with a 2.2% increase and 2.5% decrease in mRNA codon usage respectively. Moreover, these codons are similarly enriched and depleted in genes grouped by ontology under differentiation. Correlation analysis of codon usage dynamics over 48hrs with codon usage in differentiation versus proliferation genes indeed shows a strong relationship (Figure 25C, Spearman's rank correlation coefficient = 0.773, $p<2.2e-16$), suggesting changes in codon usage after 48hrs mirror the differentiation status of the cell.

Although changes in codon usage after 48hrs are similar to preferences in differentiation genes, many of the increasing or decreasing codons cluster by amino acid. Comparison of amino acid usage confirms that the ecdysone treated transcriptome prefers amino acids that are similarly enriched in differentiation proteins (Figure 25C, Spearman's rank

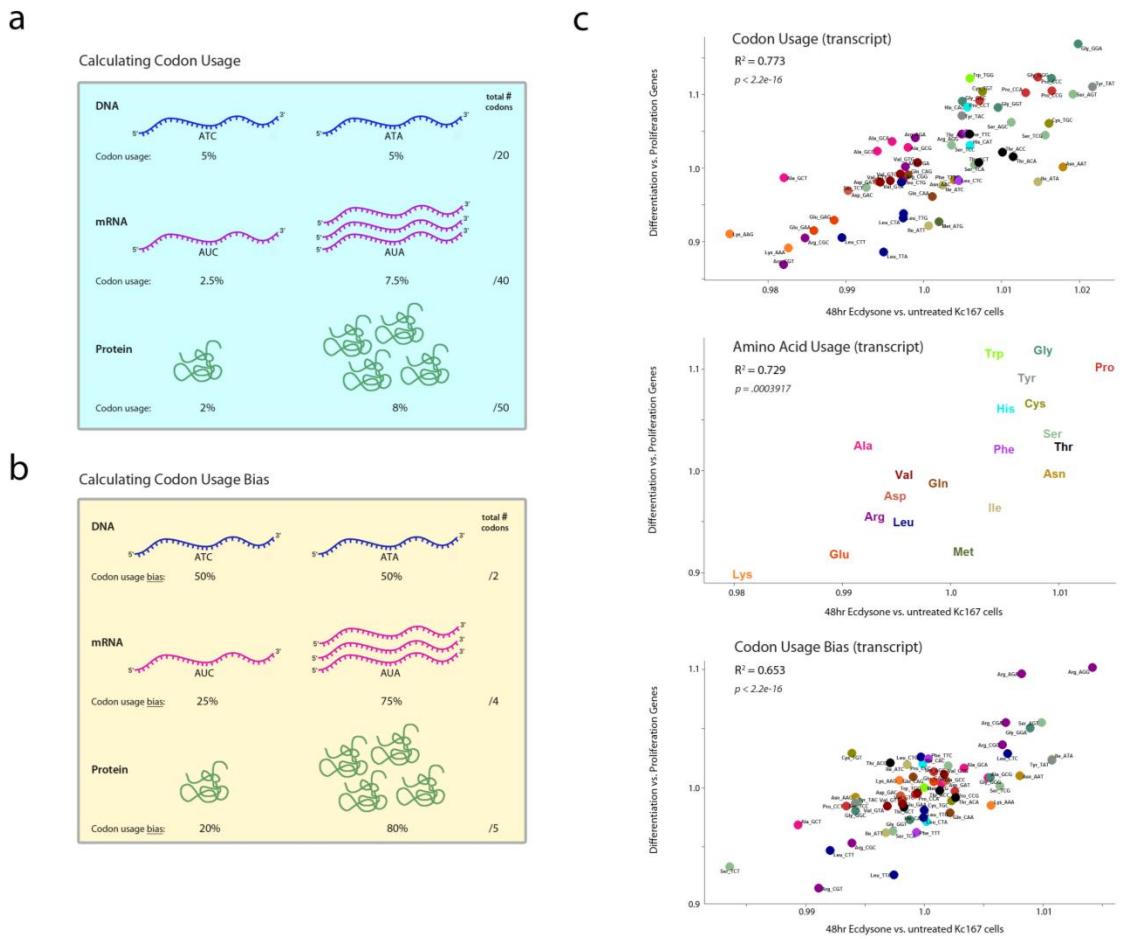


Figure 25. **Change in codon usage after 48hrs ecdysone treatment mirrors preferential codon incorporation in differentiation genes.** (a) Illustration of codon usage calculation as a function of mRNA levels rather than genomic content. Similarly, normalization by protein levels provides a read out for the number of times a codon is translated (b) Analogous representation for codon usage bias calculation. (c) Correlation plots for codon usage (top), amino acid usage (middle), and codon usage bias (bottom) throughout the transcriptome before and after 48hrs treatment with ecdysone (x-axis), and in genes assigned under differentiation GO terms versus genes associated with proliferation (cell-cycle) GO terms (y-axis). R^2 = Spearman's rank correlation coefficient.

correlation coefficient = 0.729, p=3.9e-4). Nevertheless, analysis of codon usage bias, that is the relative increases or decreases in synonymous codon usage for a given amino acid, similarly shows strong correlation between ecdysone treatment and differentiation genes (Figure 25C, Spearman's rank correlation coefficient = 0.653, p<2.2e-16). For example, arginine codons AGG and AGA are significant enriched, whereas codons CGU and CGC are significantly depleted, in both the ecdysone treated transcriptome and in differentiation genes. These findings together suggest that after 48hrs, ecdysone has dynamically modified the codon usage (and codon usage bias) in Kc167 cells in a manner that reflects the differentiation status of the cell.

Small RNA profiling identifies changes in tRNA and miRNA abundance

Small RNA-seq analysis before and after 48hr treatment with 20HE reveals significant changes in miRNAs previously shown to be induced by ecdysone, providing an internal control for our datasets [293]. Specifically, the Let-7 miRNA cluster, which encodes highly conserved developmental miRNAs critical for cell fate decisions, becomes significantly upregulated after 48hrs. Let-7 has been previously characterized for its delayed response to 20HE [294], and thus is not present during the early, 3hr 20HE response (Figure 26A). On the other hand, miRNAs known to target the ecdysone receptor and ecdysone responsive genes, such as miR-14 and miR-34 [295], become downregulated after 48hrs.

Our small RNA-seq assay also provides a sensitive measure of tRNA abundance. To

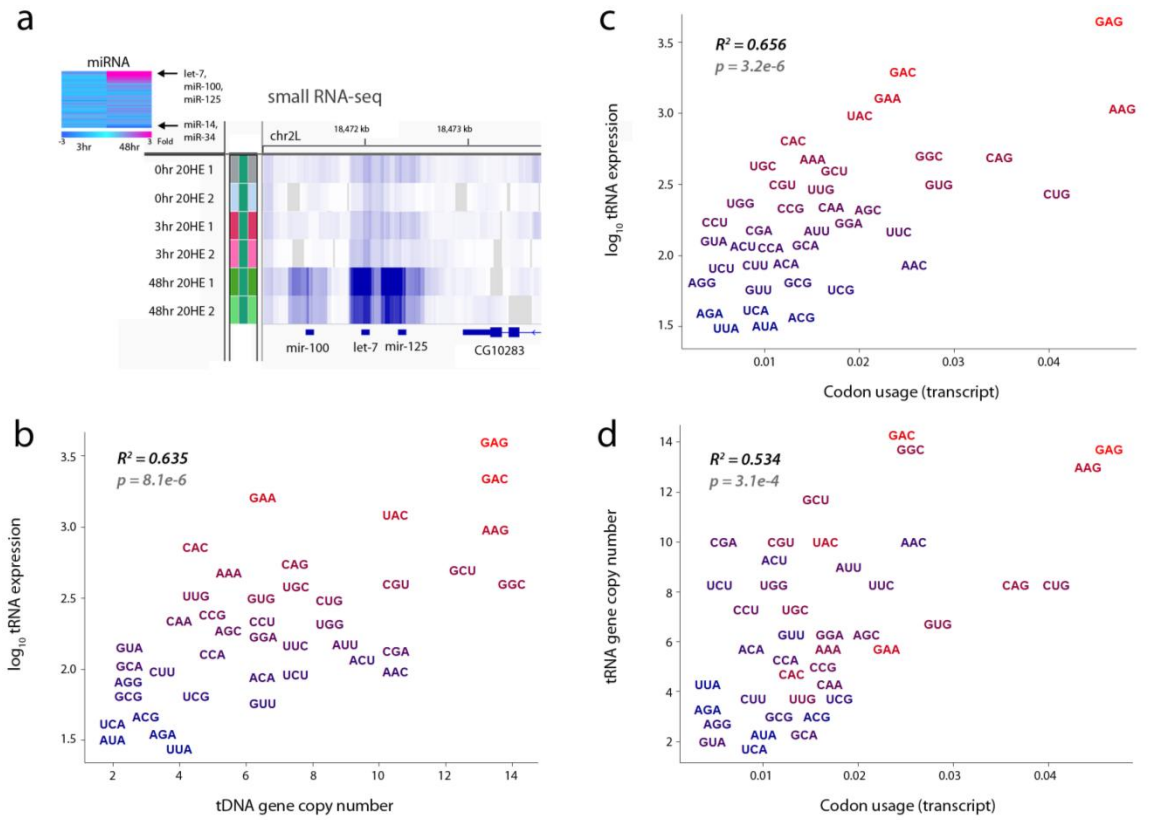


Figure 26. **Validation of small RNA profiles and tRNA abundance.** (a) Small RNA-seq heatmap at the Let-7 miRNA cluster reveals high expression specifically in 48hrs ecdysone replicates. inset: heatmap of miRNA fold changes validates the late small RNA-seq response. In addition to upregulation of the Let-7 miRNAs, miR-14 and miR-34, two miRNAs known to target the ecdysone receptor and ecdysone responsive genes, are downregulated after 48hrs. (b) Correlation plot for tRNA expression measurements from small RNA-seq (y-axis) with tRNA gene copy number (x-axis). (c) Correlation plot for tRNA expression (y-axis) with transcriptome codon usage (x-axis). (d) Correlation plot for tRNA gene copy number (y-axis) with transcriptome codon usage (x-axis). R^2 = Spearman's rank correlation coefficient.

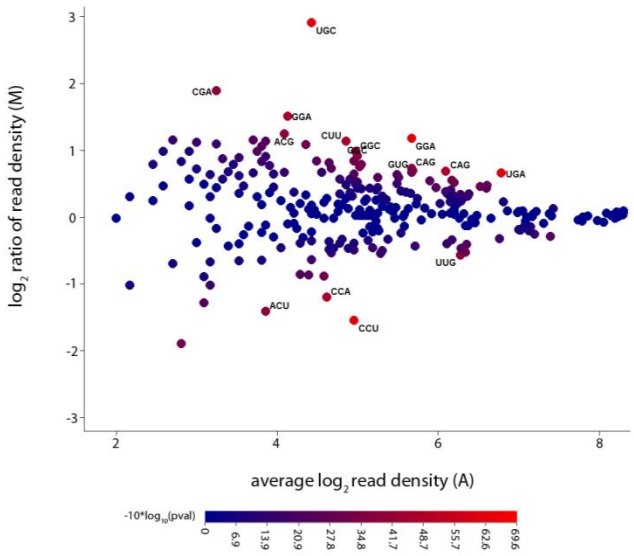
validate our estimation of tRNA levels, we show strong correlation between tRNA levels measured by RNA-seq with tRNA gene copy number (Figure 26B) and with transcriptome codon usage (Figure 26C). Importantly, whereas tRNA gene copy numbers are often used as a proxy for cellular tRNA levels and have been shown to correlate well with codon usage [288], our RNA-seq defined estimation of tRNA abundance correlates significantly better with codon usage. These comparisons strongly suggest that tRNA abundance can be accurately defined by small RNA-seq in our model system.

Analysis of tRNA abundance before and after 48hrs reveals moderate changes in the expression of specific tRNA isoacceptors (Figure 27A-B). For example, expression of the tRNA isoacceptor for glycine codon GGA (anticodon UCC) increases significantly after 48hrs (cumulative fold change 1.75), whereas the tRNA isoacceptor for proline codon CCA (anticodon UGG) decreases (cumulative fold change 0.75) (Figure 27B). Interestingly, the single copy tRNA gene encoding $tRNA^{SeC}$, which recodes UGA stop codons as selenocysteine insertion sites specifically in selenoprotein mRNAs, increases significantly after 48hrs. $tRNA^{SeC}$ expression is often repressed in proliferative cancer cells [296], and a recent GRO-seq study also uncovered increased $tRNA^{SeC}$ expression in response to estrogen signaling [297], suggesting $tRNA^{SeC}$ expression relates to the differentiation status of the cell.

tRNA abundance predicts preferential codon incorporation in polypeptides

We next asked whether increasing or decreasing tRNA isoacceptors relate to changes

a



b

Significantly changing tRNAs by codon
(p value < 0.02)

	Codon	Amino Acid	Fold Change
upregulated	UGC	Cys	1.17
	GGA	Gly	1.75
	UGA	SeC	1.60
	CUU	Leu	1.78
	CAG	Gln	1.39
	GGC	Gly	1.52
	GUG	Val	1.29
	ACG	Thr	2.19
downregulated	CCU	Pro	0.81
	CCA	Pro	0.75
	ACU	Thr	0.70
	UUG	Leu	0.76

Figure 27. **Ecdysone-induced tRNA dynamics.** (a) MA plot depicting significantly changing tRNA gene expression levels. Log2 read density ratio (48hrs/control) as a function of the average log2 read density per tRNA gene. Colorbar represents statistical significance ($-10*\log_{10}(\text{pval})$). (b) Significantly changing tRNAs by codon (p value cutoff of 0.02). Fold change = cumulative fold change for tRNA abundance as a function of codon (i.e. from multiple tRNA gene copies).

in codon usage bias after 48hrs (Figure 25C), and whether these changes lead to differences in codons incorporated into proteins. Unfortunately, however, of the approximately 300 currently annotated *Drosophila* tRNA genes, nearly one-third of synonymous codons are entirely missing their tRNA gene(s), likely due to deficiencies in the current approach for annotating tRNAs in *D. melanogaster*. As a consequence, tRNA dynamics with respect to individual amino acids can only accurately be quantified for glutamine, glutamic acid, and lysine, each of which are encoded by two synonymous codons. Nevertheless, in all cases, the increasing and decreasing tRNA isoacceptors correspond to codons with higher or lower codon usage bias in differentiation genes and in frequency of translated codons (Figure 28A), suggesting tRNA dynamics after 48hrs are tuned for differentiation-associated codon usage bias.

In addition to codon usage bias, significantly changing tRNA isoacceptor pools associate with increased or decreased codon frequency in proteins, depending on the direction of tRNA levels (Figure 28B). Significantly upregulated tRNAs, for example, associate with greater codon frequencies in the *Drosophila* proteome after 48hrs, whereas downregulated tRNAs are most often observed less frequently. These data suggest that tRNA abundance is an important player in defining the translational output of the cell. To further test this hypothesis, we compared codon usage in the Kc167 transcriptome with the number of translated codons in the proteome (Spearman's rank correlation coefficient = 0.982, $p<2.2e-16$). Analysis of codons enriched or depleted in proteins reveals a strong relationship between tRNA abundance and the frequency of codon translation (Figure 28C-E). Whereas the amino acids from low abundance tRNAs are observed less frequently in proteins, high abundance tRNAs are observed more frequently, suggesting

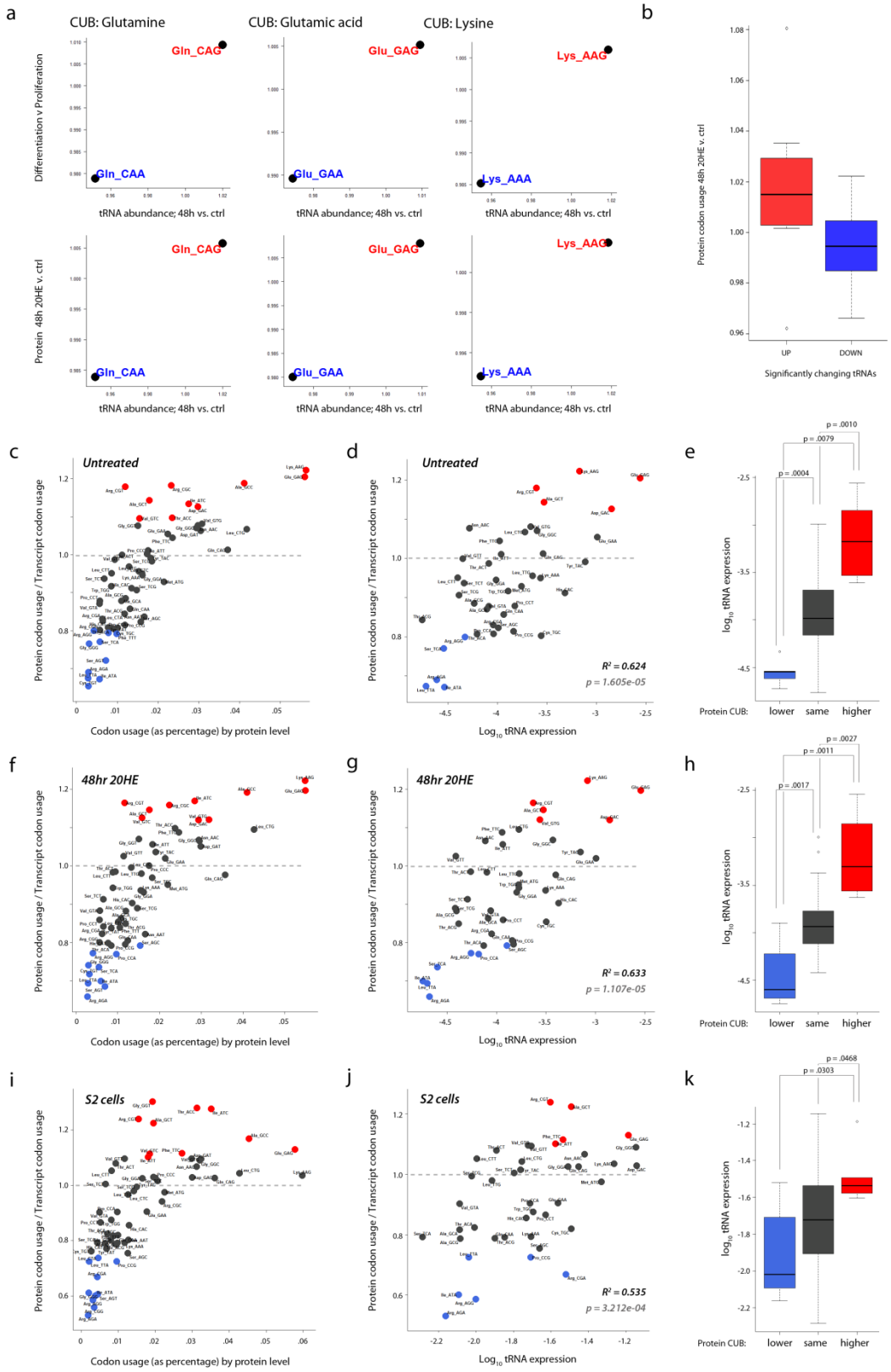

a
CUB: Glutamine
CUB: Glutamic acid
CUB: Lysine
Differentiation v Proliferation
Protein 48h 20HE v. ctrl
Gln_CAG
Gln_CAA
Glu_GAG
Glu_GAA
Lys_AAG
Lys_AAA
tRNA abundance; 48h vs. ctrl
b
Protein codon usage 48h 20HE v. ctrl
UP
DOWN
Significantly changing tRNAs
c
Untreated
Protein codon usage / Transcript codon usage
Codon usage (as percentage) by protein level
d
Untreated
R² = 0.624
p = 1.605e-05
Log10 tRNA expression
e
p = .0010
p = .0079
p = .0004
log10 tRNA expression
Protein CUB:
lower
same
higher
f
48hr 20HE
Codon usage (as percentage) by protein level
g
48hr 20HE
R² = 0.633
p = 1.107e-05
Log10 tRNA expression
h
p = .0027
p = .0011
p = .0017
log10 tRNA expression
Protein CUB:
lower
same
higher
i
S2 cells
Codon usage (as percentage) by protein level
j
S2 cells
R² = 0.535
p = 3.212e-04
Log10 tRNA expression
k
p = .0468
p = .0303
log10 tRNA expression
Protein CUB:
lower
same
higher

Figure 28. **tRNA abundance predicts preferential codon incorporation in polypeptides.** (a) Relative change in tRNA isoacceptor abundance for synonymous codons encoding glutamine (left), glutamic acid (middle), and lysine (right), and correlation with codon usage bias in differentiation-associated genes (top) and with changes in proteomic codon frequency (bottom). (b) Changes in proteomic codon frequency for codons corresponding to significantly upregulated (red) and downregulated (blue) tRNA isoacceptors. (c) proteomic codon frequency divided by transcriptomic codon usage (y-axis) as a function of the observed codon frequency in the proteome (x-axis). (d) correlation of enriched and depleted proteomic codon frequencies with tRNA abundance (R^2 = Spearman's rank correlation coefficient). (e) Expression of tRNA isoacceptors for codons enriched or depleted in the Kc167 proteome are significantly higher or lower, respectively. (f-h) parallel analyses described for (c-d) in Kc167 cells after 48hr 20HE treatment. (i-k) identical analyses described for (c-d) in Drosophila S2 cells. S2 small RNA-seq data was obtained from GEO accession GSE40016, S2 proteomic data was obtained from the *Drosophila* peptide atlas.

highly abundant tRNAs provide an advantage for protein biosynthesis (Figure 28E). Parallel analyses after 48hrs similarly confirm this phenomenon (Figure 28F-H), and identical analysis of small RNA, transcriptome, and proteomic data from *Drosophila* S2 cells independently support these observations (Figure 28I-K). Together, these data are consistent with the translational efficiency hypothesis, in which cellular tRNA levels influence the translational output of the cell, and further provide evidence of mRNA and tRNA dynamics that favor balanced codon usage and tRNA abundance.

Upregulation of selenocysteine tRNA and an experimental system to determine whether increasing $tRNA^{SeC}$ regulates selenoprotein synthesis

Selenoproteins are common to all domains of life and play key roles in antioxidant pathways as glutathione peroxidases, thioredoxin reductases, and by regulating the synthesis of other antioxidant enzymes [298]. Meanwhile, ecdysone-induced differentiation of *Drosophila* Kc167 cells has been previously shown to protect cells against oxidative stress [299], and small RNA-seq analysis after 48hrs 20HE reveals significant upregulation of $tRNA^{SeC}$ levels. Together, it is tempting to suggest that ecdysone may increase the antioxidant capacity of Kc167 cells by increasing selenoprotein biosynthesis. Importantly, gene expression for all *Drosophila* selenoproteins remains constant throughout the ecdysone response, and thus increases in protein synthesis would occur independently of transcription.

To assay whether ecdysone indeed increases selenoprotein biosynthesis, we have designed an experimental system for monitoring selenoprotein levels in control cells and

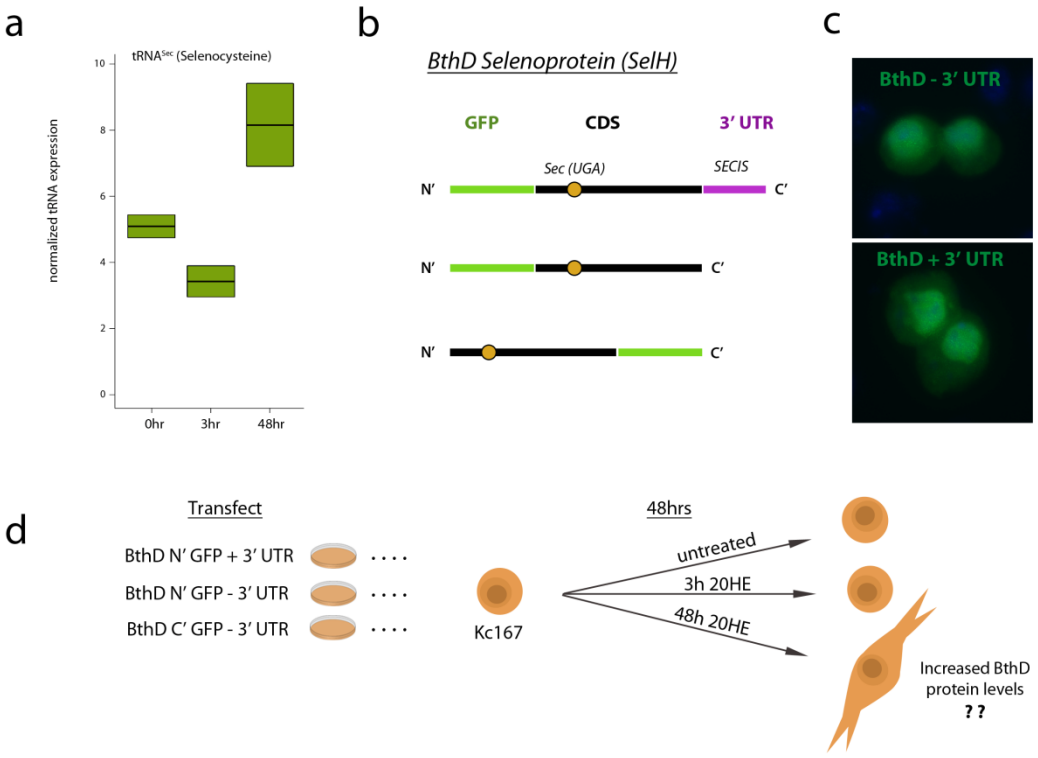


Figure 29. **An assay to determine whether upregulation of tRNA^{SeC} influences the synthesis of selenoproteins.** (a) relative abundance of tRNA^{SeC} in control, 3hrs, and 48hrs ecdysone-treated Kc167 cells. (b) experimental constructs wherein selenoprotein BthD is expressed with either N' or C' GFP tags. Control constructs are lacking the 3' UTR, which contains the selenocysteine insertion sequence (SECIS) essential for UGA read-through. (c) Life cell imaging validation of transfection and expression of GFP constructs in Kc167 cells for both constructs containing and lacking the 3' UTR. (d) experimental design: Kc167 cells are transfected with BthD constructs and subsequently split into three treatment conditions: control, 3hrs 20HE, and 48hrs 20HE.

in cells treated for 48hrs with 20HE (Figure 29). First, we insert the coding sequence for the selenoprotein BthD (aka SelH) into a constitutively expressed construct expressing either an N' or C' tag (currently GFP). Constructs were designed to contain the 3' UTR, which contains the selenocysteine insertion sequence (SECIS) element essential for translation termination read-through [298]. Experimental control vectors, on the other hand, do not contain this critical sequence (Figure 29B). BthD was selected for this assay because of the early incorporation of selenocysteine, allowing one to discriminate between premature termination and UGA read-through products based on molecular weight (See UGA marker, Figure 29B). These constructs are transfected into Kc167 cells; transfection efficiency is subsequently monitored by live-cell imaging (Figure 29C). Finally, Kc167 cells are either treated with control (ethanol), ecdysone for 3hrs, or ecdysone for 48hrs prior to cell lysis and protein quantification by western blot (Figure 29D). At the current stage of writing this thesis, it has become apparent that GFP antibodies produce a non-specific band at the same molecular weight as the putative BthD-GFP construct, and thus whether ecdysone-induced tRNA dynamics influence the biosynthesis of selenoproteins remains inconclusive. In order to further test this hypothesis, BthD-myc tag constructs will be used instead of GFP.

DISCUSSION

By profiling both the transcriptomic and proteomic response to steroid-hormone signaling in *Drosophila*, our data provide a novel, integrated measurement of the relationship between tRNA abundance, codon usage, and protein synthesis. Whereas recent studies have drawn parallels between codon usage and tRNA repertoires [289], our analysis suggests that tRNA abundance itself influences the translational output of a cell and that tRNA dynamics lead to changes in the frequency of codon translation. We show that ecdysone induces plasmatocyte-to-macrophage differentiation in Kc167 cells, and in so doing shifts codon usage (and codon usage bias) in favor of codons preferred by differentiation genes. tRNA pools are adjusted concomitantly, mirroring differentiation-associated codon usage bias and changing the frequency of codon translation accordingly. As a population, tRNAs that are high in supply correlate with codons that are translated with greater frequency than the overall transcriptome codon usage. Conversely, tRNAs that are in low supply correlate with codons that are translated with lower frequency than the transcriptome codon usage.

Our observed relationship between tRNA abundance and protein codon frequency supports the translation efficiency hypothesis, which argues that tRNA abundance influences the rate of translation. However, recent ribosome profiling measurements do not identify greater occupancy at suboptimal versus optimal synonymous codons, suggesting codon selection time is equal for tRNA isoacceptors with seemingly different relative abundances. We speculate that the observed differences in transcriptomic and proteomic codon frequencies are generated cumulatively by small differences in

translational efficiency that may not be captured by ribosome profiling measurements. We cannot, however, rule out the possibility that preferred codons with high abundance tRNAs have a lower rate of misfolding and subsequent degradation, causing differences in protein levels and codon frequencies. Similarly, our assay does not account for potential bias in mRNAs secondary structure that may promote or antagonize translation, or for differences in tRNA aminoacylation efficiency. Nevertheless, the fact that increasing the abundance of specific tRNA isoacceptors leads to increased frequency of the corresponding codon across the proteome, and vice versa, does lend credence to a model in which tRNA abundance drives translational efficiency.

Though we cannot currently draw conclusions about the ecdysone response and selenoprotein synthesis, it is tempting to suggest that increased $tRNA^{SeC}$ levels and antioxidant capacity after 48hrs are related to increasing levels of selenoproteins. This non-ectopic, biologically relevant system would provide direct evidence of whether tRNA abundance alone can drive differential protein synthesis. We describe an experimental system to test this hypothesis, taking advantage of the early incorporation of selenocyteine in the *Drosophila* selenoprotein BthD, a redox-sensing DNA binding protein involved in glutathione synthesis [300].

EXPERIMENTAL PROCEDURES

Small RNA-seq

Small RNA was purified from Kc cells with the mirVana isolation kit (Life Technologies). Small RNA library preparation and sequencing was performed by the Genomic Services Lab at HudsonAlpha. Small RNA-sequencing reads were trimmed using ngsShort and subsequently mapped to the Dm6 release of the *Drosophila* genome, allowing one mismatch. Multi-mapping reads are mapped once to the highest scoring region. Technical replicates (2) within each biological replicate were combined. tRNA read counts for each tRNA gene were then extracted using bedtools [301], and differential expression normalization and analysis performed by DEGseq [302]. Cumulative fold change for each codon was calculated by summing normalized read counts across each tRNA isoacceptor pool and taking the average fold change between conditions across all biological replicates.

Microarray analysis

RNA was purified from Kc cells with the RNeasy kit (QIAGEN), and cDNA synthesis was performed with the High Capacity cDNA Reverse Transcription kit (Applied Biosystems). cDNA was then digested with RNase A for 30 min at 37 C and cleaned by running through a PCR purification column (QIAGEN). Sample labeling, hybridization, and data extraction were performed by the FSU-Nimblegen Microarray facility using Nimblegen 12 3 135K arrays. Two biological replicates were performed for each sample.

Proteomics

Whole cell lysates for each biological replicate, matched with small RNA-seq samples, were submitted for label-free quantitative mass spectrometry analysis by the Emory University Proteomics Core as recently described [303]. Briefly, $\log_2$ fold change cutoffs (48h/WT) were set at the 95% confidence interval for a Gaussian fit to the experimental data. Filtering criteria requires that one or more peptides be sequenced across four samples, signal to noise ratio maximum measurements for each protein be greater than 10, and coefficient of variation less than 50%. Using these filtering criteria, FDR for significantly changing proteins is estimated to be 5.2%

Codon usage and codon usage bias calculations

Codon usage was calculated by assigning the coding sequences for all transcripts and proteins detected in transcriptomic and proteomic analyses, and counting the occurrence of every nucleotide triplet for each gene, requiring that coding sequences end with a stop codon and are a multiple of three. Transcriptomic codon usage and codon frequencies in proteomic data were normalized by the abundance of mRNAs and proteins respectively. Specifically, codon counts for each coding sequence were multiplied by $\log_2$ microarray expression values or by spectral counts. Codon counts were subsequently pooled for all coding sequences within specific biological replicates and conditions, and percent codon usage calculated as the frequency of using any given codon with respect to all codons. Codon usage bias was similarly calculated, and represented as the frequency of using any given codon with respect to all synonymous codons for a specific amino acid.

ACKNOWLEDGEMENTS

We thank Ashley Wood and Eddie Ramos for preparation and analysis of microarray data, Eddie Ramos for assistance with small RNA quantification, and Mike Nichols for assistance with codon usage calculation. Also, we would like to thank the Genomic Services Lab at HudsonAlpha for assistance with small RNA library preparations and for high-throughput sequencing, and the Emory University Proteomics Core for performing label-free quantitative mass spectrometry, especially Duc Duong and Eric Dammer for assistance with analysis.

CHAPTER 7:

DISCUSSION

Building a new model for insulator function and chromosome domain organization

Several important questions concerning the organization and function of insulator elements remained unanswered when I began my research described in this dissertation. For one, a majority of sites bound by insulator proteins are not capable of classical insulator activities, such as enhancer-blocking, in transgenic reporter assays. In addition, a majority of insulator binding sites localize within chromosome domains, rather than at the boundaries of these domains for which they are proposed to function as barriers. Thus, what differentiates boundary from non-boundary insulators, and sites capable versus incapable of enhancer-blocking, remained unresolved. In this context, the very use of the word "insulator" to describe sites occupied by insulator-binding proteins became erroneous as most insulators show no real insulator activity. Thus a transition was made in our lab to instead refer to insulator proteins as architectural proteins, and sites bound by architectural proteins as architectural protein binding sites (APBSs) rather than insulators. This transition is most evident between chapters two and three. We thereafter refer to insulators only as elements shown to possess classical insulator activity.

The use of the name architectural protein stems from building evidence supporting the role of these factors in facilitating long-range interactions, establishing chromatin loops that tend to separate genes based on transcriptional activities [78]. Building off of these results, I first set out to characterize the genome-wide binding profile for architectural proteins in *Drosophila* by ChIP-seq and determine the relationship between individual factors. Not surprisingly, ChIP-seq analysis revealed extensive overlap between architectural proteins, a result we first interpreted as an artifact of long-range interactions

and indirect association with DNA. However, when attempting to annotate which architectural protein directly occupied each site through consensus sequence analysis, it became evident that motifs for unique architectural proteins often clustered together at high occupancy APBSs, a phenomenon that first appeared to be enriched at the boundaries of repressive Polycomb H3K27me3 domains [46]. Surprisingly, depletion of architectural proteins by RNAi resulted in loss of H3K27me3 within domains, rather than significant spreading, suggesting APBSs may not function as chromatin barriers in their endogenous context, but rather mediate interactions important for transcription regulation.

Concurrent with the finding that architectural proteins form high occupancy elements, the age of chromatin loops gave way to the Topologically Associating Domain (TAD). High-throughput chromosome conformation capture studies (Hi-C) in fly, mouse, and humans consistently revealed that interactions along an individual chromosome are highly concentrated within regions of strong self-interaction [65-67, 198, 199], defined as TADs. Enhancer-trap analyses suggest that enhancers are restricted to regulating resident genes within the same TAD [275], and that depletion of the SMC-containing cohesin complex, an architectural protein critical for long-range interactions in mammals, allows enhancers to break out of these defined TADs [200, 202, 203], forming new aberrant interactions that cause misregulation of neighboring genes. Meanwhile, studies in mammals suggest that TADs are conserved between cell-types, suggesting they represent a modular, pre-defined, bottom-up pattern of chromatin organization. Though architectural proteins were shown to be enriched at the borders of TADs [65, 66, 198],

suggesting a role in directly or indirectly partitioning these self-interacting chromatin domains, a majority of APBSs reside within TADs and an explanation for what differentiated these sites remained elusive.

Though the fruit fly provides an advantageous system for studying architectural proteins, building off of decades of insulator-related studies characterizing factors critical for classical insulator activity in transgenic reporter-assay, our analyses were still limited to just three DNA-binding proteins: Su(Hw), BEAF-32, and the *Drosophila* homolog of CTCF; and BTB-containing proteins CP190 and Mod(mdg4) [46]. Meanwhile, studies in yeast and humans had demonstrated that the Pol III transcription factor, TFIIIC, is capable of classical insulator activity, and that SMC-containing complexes cohesin and condensin were, in certain contexts, critical components [27, 49, 50, 304]. However, both the TFIIIC and condensin complexes remained poorly studied in *D. melanogaster*, leading us to focus our efforts on characterizing the b-box binding subunit of TFIIIC and both cohesin and condensin complexes by ChIP-seq.

My genome-wide analysis for TFIIIC, cohesin, and condensin revealed strong overlap, both among each other, as well as with architectural proteins Su(Hw), BEAF-32, and dCTCF [253]. With the increasing knowledge of overlapping factors, an accurate map of relative occupancy by architectural proteins could be made, ranging from low, medium, to high – where most or all architectural proteins overlap. Comparison of APBSs occupancy with TAD border colocalization revealed a striking trend: TAD borders are defined by high occupancy APBSs, and whereas low occupancy APBSs reside within

TADs. Furthermore, calculation of TAD border strength with respect to APBS occupancy demonstrates that the strength of physical domain separation, or in other words the degree to which interactions between separate TADs are restricted, scales with the number of overlapping architectural proteins. These results suggest that domain-border associated APBSs are indeed distinct from non-border associated APBSs based on the occupancy of these factors. Comparison of APBS occupancy with the capacity of specific genomic elements to function as insulators, previously reported in transgenic assays, also reveals that high occupancy APBSs are capable of robust enhancer-blocking activity, whereas low occupancy APBSs are not [253], suggesting APBS occupancy, domain border localization, and insulator activity are intrinsically related.

The occupancy of architectural proteins at high, medium, and low APBSs is highly reminiscent of analogous high occupancy transcription factor "hotspots", recently shown to function as spatiotemporal specific enhancers [218] and similarly associated with the SMC-containing cohesin complex [219, 220]. However, comparison of DNase I hypersensitivity at HOT and high occupancy APBSs throughout *Drosophila* development revealed that whereas HOT sites progressively lose chromatin accessibility during development, APBSs are consistently stable, consistent with the tissue-invariant nature of TAD borders and suggesting APBS occupancy may play an important role. Finally, by taking advantage of hundreds of genome-wide mapping studies in humans and mouse, we were able to show that clustering of architectural proteins at strong TAD borders is a conserved feature of genome organization [253].

Together, my graduate insulator studies have revealed a spectrum of insulator co-localization that scales with domain boundary strength and regulatory capacity, establishing a novel framework for understanding the diverse roles of insulators in gene regulation (Figure 30). However, several questions remain with regards to the mechanism by which this spectrum of architectural protein occupancy contributes to the parallel spectrum of topological domain separation and insulator function. As the Hi-C data is merely a representation of interaction frequencies, we speculate that differences in TAD border strengths result from the stability and/or likelihood for an APBS to interact with another region. Whereas high occupancy APBSs are either more likely to interact with other genomic elements or to form more stable interactions, local interactions across these regulatory elements will occur with very low frequency. The very occupancy of APBSs may, to some degree, reflect interactions themselves, as the presence of one or more factors may be an indirect artifact of its presence at an interacting locus. Medium-to-low occupancy APBSs, on the other hand, participate in fewer or more transient interactions, allowing greater interaction frequencies across these regions which, in the Hi-C data, will be represented as weak TAD borders that may or may not be called using an arbitrary threshold.

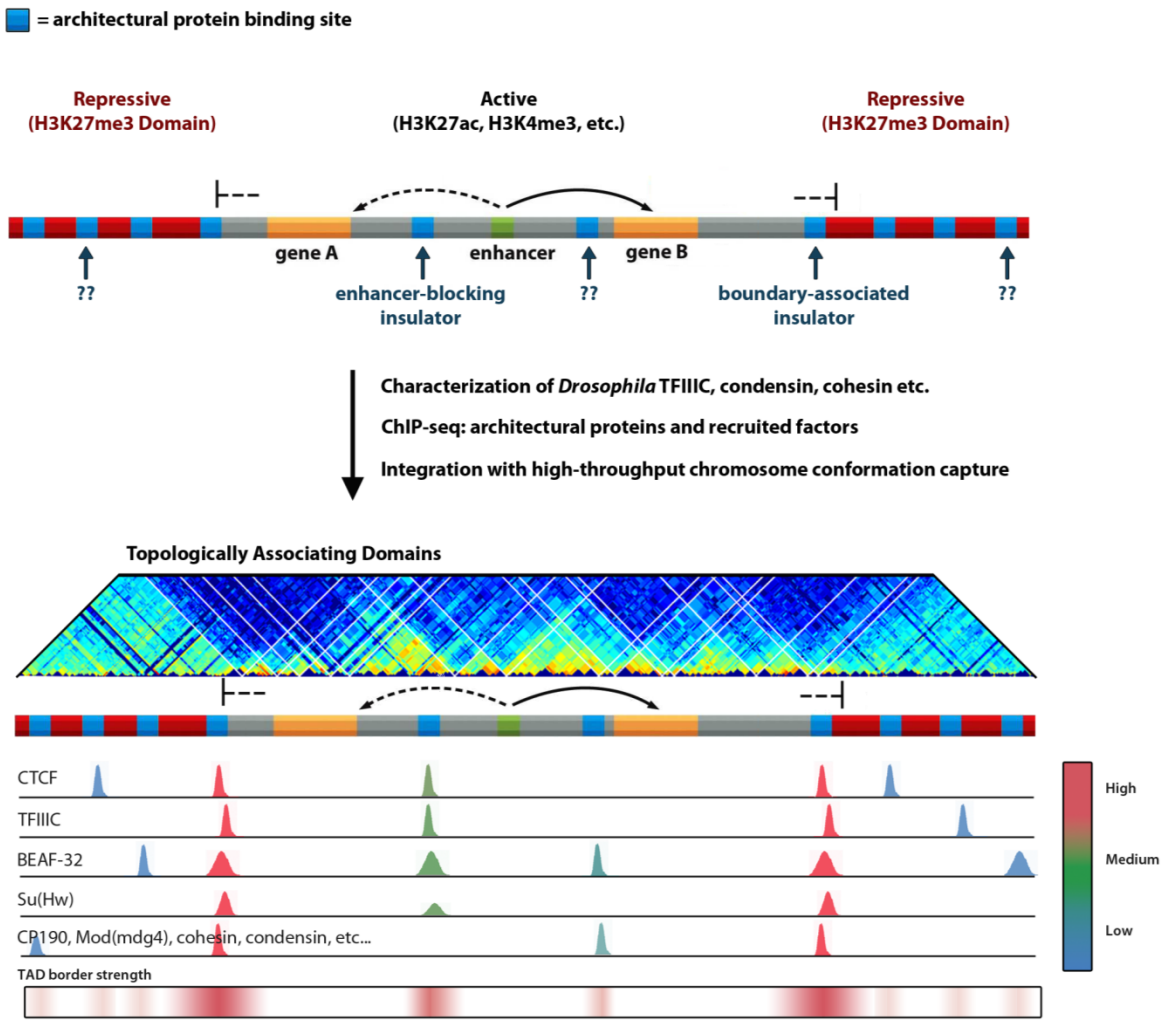


Figure 30. **A novel, integrative model for architectural protein binding, insulator function, and topological domain organization.** Top: Unresolved issues concerning insulator organization and function. (1) the majority of sites bound by architectural proteins are not capable of classical insulator activities, such as enhancer-blocking, in transgenic reporter assays. (2) despite enrichment at the borders of H3K27me3 domains, the majority of architectural protein binding sites localize within chromosome domains. What differentiates boundary from non-boundary insulators, and sites capable versus incapable of enhancer-blocking? Bottom: A new model integrating genome-wide

interaction and DNA-binding data explains the enigmatic role of architectural proteins in insulator function and chromosome organization. A spectrum of architectural protein occupancy scales with topologically associating domain (TAD) border localization and the strength of physical domain separation, as well as enhancer-blocking insulator activity. High occupancy architectural protein binding sites (APBSs) localize to strong TAD boundaries and possess robust enhancer-blocking activity. Low occupancy APBSs are enriched with TADs and appear reserved for local interactions and regulation of specific genes.

tRNAs: an important player in cellular identity and differentiation

Our transcriptional, small RNA, and proteomic profiling experiments during the ecdysone response in Kc167 cells uncovers a shift in codon usage and codon usage bias associated with differentiation. Both messenger RNAs (mRNAs) and proteins upregulated by 20HE show a bias towards codons enriched in differentiation genes, and a depletion for codons enriched in proliferation genes. We provide evidence that cellular tRNA pools are dynamically modulated concomitantly to changes in codon usage, favoring codons associated with differentiation. Though we could not, to date, show that changes in $tRNA^{SeC}$, a single copy tRNA gene required for selenoprotein synthesis and upregulated by ecdysone, lead to changes in selenoprotein levels, we do provide evidence of a role for tRNA levels in regulating protein synthesis. Specifically, tRNA abundance predicts preferential codon usage throughout the Kc167 proteome compared to cellular mRNA codon usage, and tRNA isoacceptors levels that significantly change after ecdysone treatment become enriched or depleted accordingly in proteins after 48hrs.

Our observed relationship between tRNA abundance and codon frequency in proteins supports the translational efficiency hypothesis, which proposes that cellular tRNA levels influence the translational output of the cell [283]. However, our results also provide evidence of mRNA and tRNA dynamics that favor balanced codon usage and tRNA abundance, consistent with a recent model in which balancing codon usage optimizes translational efficiency [288]. We speculate that differences in tRNA isoacceptor availability lead to very small but cumulative changes in translational efficiency that may not be detected using current ribosome profiling assays [286, 287]

We build off of our results by proposing a model in which tRNAs are a key player in cellular identity and differentiation (Figure 31). Tuning tRNA concentrations to favor balanced codon usage, or in other words changing the cellular tRNA supply to accommodate demand in response to changes in a given cell's transcriptome will optimize the translational efficiency of that cell, regardless of whether the translational efficiency hypothesis or, instead, the balance codon usage model is correct. Meanwhile, small adjustments in tRNA abundance are likely to have small but significant effects in the large-scale proteomic output of the cell. We find evidence for this, as small but significant changes in tRNA levels induced by ecdysone lead to changes in codon frequencies observed in the Kc167 proteome.

Several important questions with regards to how tRNA dynamics are regulated remain unanswered. For example, although our small RNA-seq analysis captures steady state levels of tRNA isoacceptor molecules, we possess no direct evidence of the transcription level for any particular tRNA gene, and thus cannot say whether ecdysone is directly affecting the transcription of these genes. In this regard, it remains difficult to say whether changes in tRNA levels are transcriptionally driven and thus a cause of dynamic codon usage, or rather a response to changes induced by other mechanisms. To address tRNA transcription dynamics, global run-on assays combined with high-throughput sequencing could be used to measure the level of nascent transcripts at tRNA genes. However, tRNAs are heavily modified postranscriptionally in a manner that affects both the stability and maturity of the tRNA isoacceptor molecule [305], and thus knowledge of the transcription status at tRNA genes alone is not sufficient to quantify cellular tRNA levels. Thus future studies querying both the transcription, modification, and stability of

tRNAs will provide important insight into dynamic tRNA regulation under unique conditions.

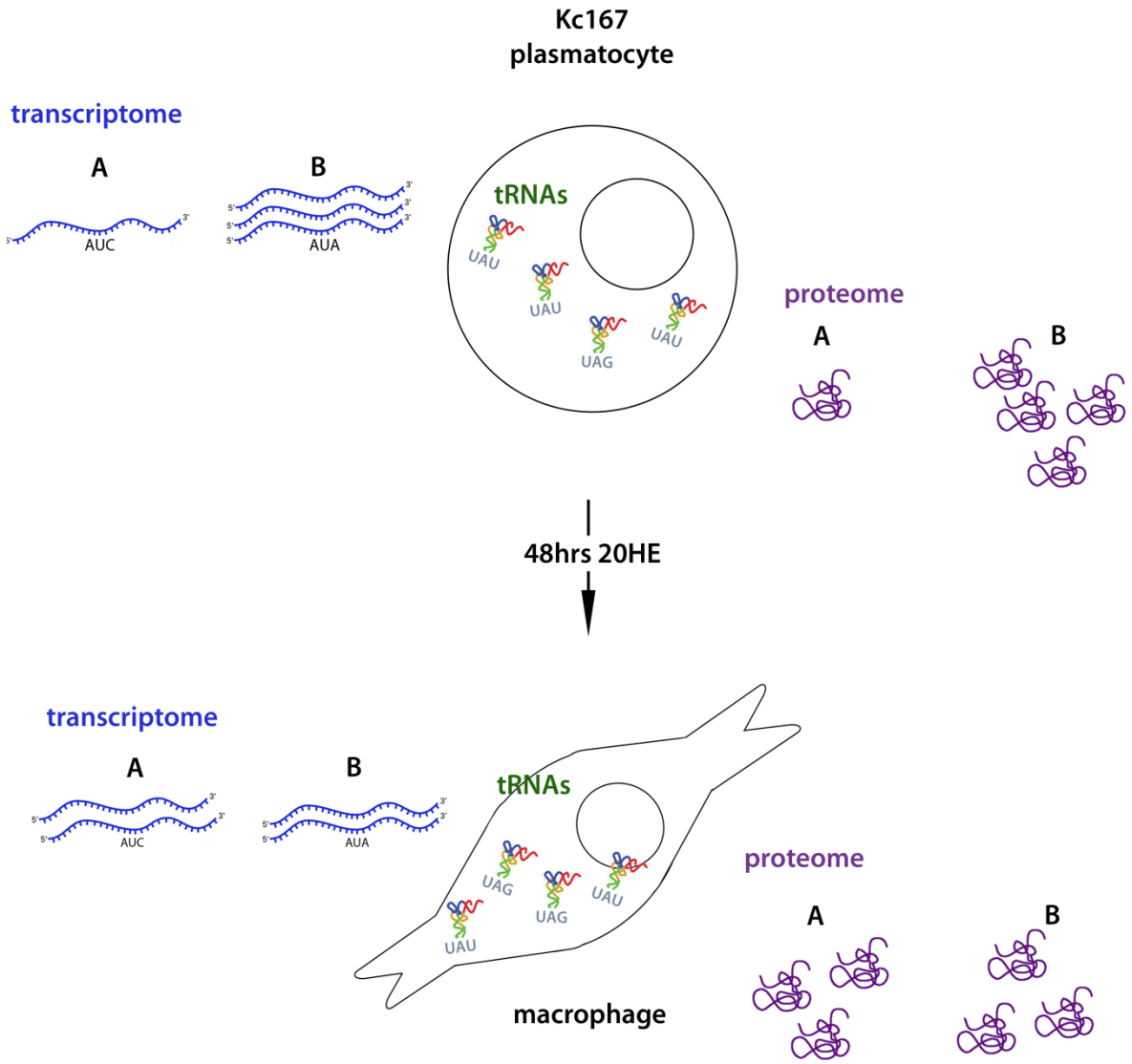


Figure 31. **A model for the role of tRNAs in cellular identity and differentiation.** Ecdysone-induced differentiation of *Drosophila* Kc167 plasmatocytes into macrophages is accompanied by changes in transcriptomic codon usage, tRNA abundance, and observed codon frequencies across the proteome. Ecdysone-altered codon usage reflects the differentiation status of the cell. Highly abundant tRNA isoacceptor molecules correlate with codons that are observed with greater frequency in proteins than across the transcriptome, suggesting tRNA levels influence protein synthesis. Changes in tRNA

abundance after 48hrs 20HE lead to corresponding changes in proteomic codon frequency after treatment with ecdysone.

CURRICULUM VITAE

KEVIN THOMAS VAN BORTLE

Department of Biology, Emory University, Atlanta, GA 30322

585-880-4017

kevin.vanbortle@gmail.com

I. Education

2009 – Present	Ph.D. Candidate. Interdepartmental program in Biochemistry, Cell, and Developmental Biology (BCDB). Department of Biology, Emory University, Atlanta, GA.
2005 – 2009	Bachelors of Science in Biochemistry *cum laude*. Distinction in research; minor in clinical psychology. University of Rochester, Rochester, NY

II. Research Experience

2009 – Present	Graduate student research, PI - Victor Corces, Department of Biology, Emory University.
2008 – 2009	de Kiewiet undergraduate research – Jeffrey Hayes lab, Department of Biochemistry and Biophysics, University of Rochester.

III. Publications

Van Bortle, K., Nichols, M., Li, L., Ong, C.T., Takenaka, N., Qin, ZS., Corces, VG. Insulator function and topological domain border strength scale with architectural protein occupancy. **Genome Biol**., 15(6):R82. doi: 10.1186/gb-2014-15-5-r82

**Featured in research highlight – Genome Biol. 2014. 15:121*

Van Bortle, K., Corces, V. Lost in transition: dynamic enhancer organization across naïve and primed stem cell states. **Cell Stem Cell**. 2014. 14(6):693-694.

Ong, C.T., **Van Bortle, K.**, Ramos, E., Corces, VG. Poly(ADP-ribosyl)ation regulates insulator function and intrachromosomal interactions in *Drosophila*. **Cell**. 2013. 155(1):148-159.

**Featured in leading edge preview – Cell. 2013. 155:15-16*

Kellner, WA., **Van Bortle, K.**, Li, L., Ramos, E., Takenaka, N., Corces, VG. Distinct isoforms of the *Drosophila* Brd4 homologue are present at enhancers, promoters and insulator sites. **Nucleic Acids Res**. 2013. 41(20):9274-9283.

Gurudatta BV., Yang J., **Van Bortle, K.**, Donlin-Asp, PG., Corces, V. Dynamic changes in the genomic localization of DNA replication-related element binding factor during the cell cycle. **Cell Cycle**. 2013. 12(10).

Van Bortle, K., Corces, V. Spinning the web of cell fate. **Cell**. 2013. 152(6):1213-1217.

Van Bortle, K., Corces, V. The role of chromatin insulators in nuclear architecture and genome function. **Curr Opin Genet Dev**. 2013. 10.1016/j.gde.2012.11.003.

Kellner, WA., Ramos, E., **Van Bortle, K.**, Takenaka, N., Corces, VG. Genome-wide phosphoacetylation of histone H3 at *Drosophila* enhancers and promoters. **Genome Res**. 2012. 22(6):1081-1088.

Van Bortle, K., Ramos, E., Takenaka, N., Yang, J., Wahi, J., Corces, V. *Drosophila* CTCF tandemly aligns with other insulator proteins at the borders of H3K27me3 domains. **Genome Res**. 2012. 22(11)2176-2187.

**Selected by Faculty of 1000 Biology*

Van Bortle, K., Corces, V. tDNA insulators and the emerging role of TFIIIC in genome organization. **Transcription**. 2012. 3(6).

Van Bortle, K., Corces, V. Nuclear organization and Genome Function. **Annu. Rev. Cell Dev. Biol**. 2012. 28: 163-187.

Wood, A., **Van Bortle, K**., Ramos, E., Takenaka, N., Rohrbaugh, M., Jones, B., Jones, K., Corces, V. Regulation of chromatin organization and inducible gene expression by a Drosophila insulator. **Mol. Cell**. 2011. 44(1):29-38

**Featured in Molecular Cell preview – Mol. Cell. 2011. 44:1-2*

IV. Abstracts; Conferences

Van Bortle, K. Nichols, M., Li, L., Ong, C.T., Takenaka, N., Qin, ZS., Corces, VG. A Subclass of Insulator Protein Binding Sites Establish Dense Clusters to Delineate Topologically Associating Domains. Janelia Farms Research Conference 2014; Poster presentation. Ashburn, VA

Van Bortle, K. Nichols, M., Li, L., Ong, C.T., Takenaka, N., Qin, ZS., Corces, VG. Combinatorial Binding of Insulator Proteins Shapes Topological Domain Structure and Regulatory Function. Emory GDBBS Symposium 2014; Oral Presentation. Atlanta, GA

Van Bortle, K. Ong, C.T., Nichols, M., Li, L., Takenaka, N., Qin, ZS., Corces, VG. The Role of Architectural Cluster Elements (ACEs) in Nuclear Organization and Genome Function. ASCB Annual Meeting 2013; Oral Presentation. New Orleans, LA.

**Featured in MBoC ASCB annual meeting highlights - Mol. Biol. Cell 2014 25:6 735*

Bolhy, S., Geoffroy, M.C., Bouhlel, I., **Van Bortle, K.**, Powers, M. and Doye, V. Intranuclear dynamics of the Nup107-160 complex and links with Nup98/CRM1-containing bodies. Dynamic Organization of Nuclear Function Meeting. 2012. Cold Spring Harbor, NY

Van Bortle, K. Liu, N. Hayes, J.J. Development of a cysteine reactivity assay to monitor H2B:H4 interface stability in a nucleosome. UPBM Poster Session. 2008. Rochester, NY

V. Teaching Experience

Spring 2014	Guest Lecturer: Eukaryotic Chromosome Function (IBS 561); Guest Lecturer: Developmental Biology (Biology 223)
Fall 2013	Graduate Methods Workshop on "Analysis of Sequencing-based Genomics Data"
Fall 2012	Graduate Methods Workshop on "ChIP-seq; the evolution of chromatin immunoprecipitation"
Spring 2011	Teaching Assistant: Basic Biomedical and Bio Sciences (IBS 556)

2011 - 2014	Supervision and training of several undergraduate and graduate rotation students

VI. Academic Awards and Acknowledgements

2009 – 2014	Emory Division Scholar Fellow – Emory University
2012	Program Scholar of the Year Award – BCDB, Emory University
2010 – 2011	NIH Training grant in Biochemistry, Cellular, and Molecular Biology, Emory University (Project # 5T32GM008367)
2010, 2011	Achievement rewards for college scientists (ARCS) nominee
2008	de Kiewiet research fellowship, University of Rochester
2005	Dexter Perkins History Award, University of Rochester

LITERATURE CITED

1. **Cremer T, Cremer M: Chromosome territories. *Cold Spring Harbor perspectives in biology* 2010, 2(3):a003889.**

2. **Fraser P, Bickmore W: Nuclear organization of the genome and the potential for gene regulation. *Nature* 2007, 447(7143):413-417.**

3. **Mao YS, Zhang B, Spector DL: Biogenesis and function of nuclear bodies. *Trends in genetics : TIG* 2011, 27(8):295-306.**

4. **Branco MR, Pombo A: Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS biology* 2006, 4(5):e138.**

5. **Chambeyron S, Bickmore WA: Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription. *Genes & development* 2004, 18(10):1119-1130.**

6. **Chakalova L, Fraser P: Organization of transcription. *Cold Spring Harbor perspectives in biology* 2010, 2(9):a000729.**

7. **Ahmed K, Dehghani H, Rugg-Gunn P, Fussner E, Rossant J, Bazett-Jones DP: Global chromatin architecture reflects pluripotency and lineage commitment in the early mouse embryo. *PloS one* 2010, 5(5):e10531.**

8. **Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP *et al*: Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 2007, 448(7153):553-560.**

9. **Vastenhouw NL, Zhang Y, Woods IG, Imam F, Regev A, Liu XS, Rinn J, Schier AF: Chromatin signature of embryonic pluripotency is established during genome activation. *Nature* 2010, 464(7290):922-926.**

10. **Wiblin AE, Cui W, Clark AJ, Bickmore WA: Distinctive nuclear organisation of centromeres and regions involved in pluripotency in human embryonic stem cells. *Journal of cell science* 2005, 118(Pt 17):3861-3868.**

11. **Peric-Hupkes D, van Steensel B: Role of the nuclear lamina in genome organization and gene expression. *Cold Spring Harbor symposia on quantitative biology* 2010, 75:517-524.**

12. **Meister P, Mango SE, Gasser SM: Locking the genome: nuclear organization and cell fate. *Current opinion in genetics & development* 2011, 21(2):167-174.**

13. Berezney R, Dubey DD, Huberman JA: Heterogeneity of eukaryotic replicons, replicon clusters, and replication foci. *Chromosoma* 2000, 108(8):471-484.

14. Kruhlak MJ, Celeste A, Dellaire G, Fernandez-Capetillo O, Muller WG, McNally JG, Bazett-Jones DP, Nussenzweig A: Changes in chromatin structure and mobility in living cells at sites of DNA double-strand breaks. *The Journal of cell biology* 2006, 172(6):823-834.

15. Lisby M, Mortensen UH, Rothstein R: Colocalization of multiple DNA double-strand breaks at a single Rad52 repair centre. *Nature cell biology* 2003, 5(6):572-577.

16. Mekhail K, Moazed D: The nuclear envelope in genome organization, expression and stability. *Nature reviews Molecular cell biology* 2010, 11(5):317-328.

17. Nagai S, Davoodi N, Gasser SM: Nuclear organization in genome stability: SUMO connections. *Cell research* 2011, 21(3):474-485.

18. Dekker J, Rippe K, Dekker M, Kleckner N: Capturing chromosome conformation. *Science* 2002, 295(5558):1306-1311.

19. Tolhuis B, Palstra RJ, Splinter E, Grosveld F, de Laat W: Looping and interaction between hypersensitive sites in the active beta-globin locus. *Molecular cell* 2002, 10(6):1453-1465.

20. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO *et al*: Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 2009, 326(5950):289-293.

21. Grosberg A, Nechaev SK, Shakhnovich EI: [The role of topological limitations in the kinetics of homopolymer collapse and self-assembly of biopolymers]. *Biofizika* 1988, 33(2):247-253.

22. Mirny LA: The fractal globule as a model of chromatin architecture in the cell. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology* 2011, 19(1):37-51.

23. Yaffe E, Tanay A: Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature genetics* 2011, 43(11):1059-1065.

24. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE *et al*: Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007, 447(7146):799-816.

25. Celniker SE, Dillon LA, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM *et al*: Unlocking the secrets of the genome. *Nature* 2009, 459(7249):927-930.

26. Ebersole T, Kim JH, Samoshkin A, Kouprina N, Pavlicek A, White RJ, Larionov V: tRNA genes protect a reporter gene from epigenetic silencing in mouse cells. *Cell Cycle* 2011, 10(16):2779-2791.

27. Raab JR, Chiu J, Zhu J, Katzman S, Kurukuti S, Wade PA, Haussler D, Kamakaka RT: Human tRNA genes function as chromatin insulators. *The EMBO journal* 2011.

28. Klenova EM, Nicolas RH, Paterson HF, Carne AF, Heath CM, Goodwin GH, Neiman PE, Lobanenkov VV: CTCF, a conserved nuclear factor required for optimal transcriptional activity of the chicken c-myc gene, is an 11-Zn-finger protein differentially expressed in multiple forms. *Molecular and cellular biology* 1993, 13(12):7612-7624.

29. Filippova GN, Fagerlie S, Klenova EM, Myers C, Dehner Y, Goodwin G, Neiman PE, Collins SJ, Lobanenkov VV: An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Molecular and cellular biology* 1996, 16(6):2802-2813.

30. Moon H, Filippova G, Loukinov D, Pugacheva E, Chen Q, Smith ST, Munhall A, Grewe B, Bartkuhn M, Arnold R *et al*: CTCF is conserved from Drosophila to humans and confers enhancer blocking of the Fab-8 insulator. *EMBO reports* 2005, 6(2):165-170.

31. Holohan EE, Kwong C, Adryan B, Bartkuhn M, Herold M, Renkawitz R, Russell S, White R: CTCF genomic binding sites in Drosophila and the organisation of the bithorax complex. *PLoS genetics* 2007, 3(7):e112.

32. Weth O, Renkawitz R: CTCF function is modulated by neighboring DNA binding factors. *Biochemistry and cell biology = Biochimie et biologie cellulaire* 2011, 89(5):459-468.

33. Zlatanova J, Caiafa P: CTCF and its protein partners: divide and rule? *Journal of cell science* 2009, 122(Pt 9):1275-1284.

34. Nativio R, Wendt KS, Ito Y, Huddleston JE, Uribe-Lewis S, Woodfine K, Krueger C, Reik W, Peters JM, Murrell A: Cohesin is required for higher-order chromatin conformation at the imprinted IGF2-H19 locus. ***PLoS genetics*** **2009, 5(11):e1000739.**

35. Wendt KS, Yoshida K, Itoh T, Bando M, Koch B, Schirghuber E, Tsutsumi S, Nagae G, Ishihara K, Mishiro T ***et al*****: Cohesin mediates transcriptional insulation by CCCTC-binding factor.** ***Nature*** **2008, 451(7180):796-801.**

36. Donohoe ME, Zhang LF, Xu N, Shi Y, Lee JT: Identification of a Ctcf cofactor, Yy1, for the X chromosome binary switch. ***Molecular cell*** **2007, 25(1):43-56.**

37. Wilkinson FH, Park K, Atchison ML: Polycomb recruitment to DNA in vivo by the YY1 REPO domain. ***Proceedings of the National Academy of Sciences of the United States of America*** **2006, 103(51):19296-19301.**

38. Ishihara K, Oshimura M, Nakao M: CTCF-dependent chromatin insulator is linked to epigenetic remodeling. ***Molecular cell*** **2006, 23(5):733-742.**

39. Yao H, Brick K, Evrard Y, Xiao T, Camerini-Otero RD, Felsenfeld G: Mediation of CTCF transcriptional insulation by DEAD-box RNA-binding protein p68 and steroid receptor RNA activator SRA. ***Genes & development*** **2010, 24(22):2543-2555.**

40. Gurudatta BV, Corces VG: Chromatin insulators: lessons from the fly. ***Briefings in functional genomics & proteomics*** **2009, 8(4):276-282.**

41. Misulovin Z, Schwartz YB, Li XY, Kahn TG, Gause M, MacArthur S, Fay JC, Eisen MB, Pirrotta V, Biggin MD ***et al*****: Association of cohesin and Nipped-B with transcriptionally active regions of the Drosophila melanogaster genome.** ***Chromosoma*** **2008, 117(1):89-102.**

42. Bonchuk A, Denisov S, Georgiev P, Maksimenko O: Drosophila BTB/POZ domains of "ttk group" can form multimers and selectively interact with each other. ***Journal of molecular biology*** **2011, 412(3):423-436.**

43. Gerasimova TI, Lei EP, Bushey AM, Corces VG: Coordinated control of dCTCF and gypsy chromatin insulators in Drosophila. ***Molecular cell*** **2007, 28(5):761-772.**

44. Ghosh D, Gerasimova TI, Corces VG: Interactions between the Su(Hw) and Mod(mdg4) proteins required for gypsy insulator function. ***The EMBO journal*** **2001, 20(10):2518-2527.**

45. Negre N, Brown CD, Shah PK, Kheradpour P, Morrison CA, Henikoff JG, Feng X, Ahmad K, Russell S, White RA *et al*: A comprehensive map of insulator elements for the Drosophila genome. *PLoS genetics* 2010, 6(1):e1000814.

46. Van Bortle K, Ramos E, Takenaka N, Yang J, Wahi J, Corces V: Drosophila CTCF tandemly aligns with other insulator proteins at the borders of H3K27me3 domains. *Genome research* 2012.

47. Lei EP, Corces VG: RNA interference machinery influences the nuclear organization of a chromatin insulator. *Nature genetics* 2006, 38(8):936-941.

48. Moshkovich N, Nisha P, Boyle PJ, Thompson BA, Dale RK, Lei EP: RNAi-independent role for Argonaute2 in CTCF/CP190 chromatin insulator function. *Genes & development* 2011, 25(16):1686-1701.

49. Donze D, Adams CR, Rine J, Kamakaka RT: The boundaries of the silenced HMR domain in Saccharomyces cerevisiae. *Genes & development* 1999, 13(6):698-708.

50. Scott KC, Merrett SL, Willard HF: A heterochromatin barrier partitions the fission yeast centromere into discrete chromatin domains. *Current biology : CB* 2006, 16(2):119-129.

51. Noma K, Cam HP, Maraia RJ, Grewal SI: A role for TFIIIC transcription factor complex in genome organization. *Cell* 2006, 125(5):859-872.

52. Donze D, Kamakaka RT: RNA polymerase III and RNA polymerase II promoter complexes are heterochromatin barriers in Saccharomyces cerevisiae. *The EMBO journal* 2001, 20(3):520-531.

53. Valenzuela L, Dhillon N, Kamakaka RT: Transcription independent insulation at TFIIIC-dependent insulators. *Genetics* 2009, 183(1):131-148.

54. Moqtaderi Z, Struhl K: Genome-wide occupancy profile of the RNA polymerase III machinery in Saccharomyces cerevisiae reveals loci with incomplete transcription complexes. *Molecular and cellular biology* 2004, 24(10):4118-4127.

55. Carriere L, Graziani S, Alibert O, Ghavi-Helm Y, Boussouar F, Humbertclaude H, Jounier S, Aude JC, Keime C, Murvai J *et al*: Genomic binding of Pol III transcription machinery and relationship with TFIIS transcription factor distribution in mouse embryonic stem cells. *Nucleic acids research* 2012, 40(1):270-283.

56. Moqtaderi Z, Wang J, Raha D, White RJ, Snyder M, Weng Z, Struhl K: Genomic binding profiles of functionally distinct RNA polymerase III transcription complexes in human cells. *Nature structural & molecular biology* 2010, 17(5):635-640.

57. D'Ambrosio C, Schmidt CK, Katou Y, Kelly G, Itoh T, Shirahige K, Uhlmann F: Identification of cis-acting sites for condensin loading onto budding yeast chromosomes. *Genes & development* 2008, 22(16):2215-2227.

58. Bushey AM, Ramos E, Corces VG: Three subclasses of a Drosophila insulator show distinct and cell type-specific genomic distributions. *Genes & development* 2009, 23(11):1338-1350.

59. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K: High-resolution profiling of histone methylations in the human genome. *Cell* 2007, 129(4):823-837.

60. Cuddapah S, Jothi R, Schones DE, Roh TY, Cui K, Zhao K: Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome research* 2009, 19(1):24-32.

61. Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, Green RD, Zhang MQ, Lobanenkov VV, Ren B: Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* 2007, 128(6):1231-1245.

62. Parelho V, Hadjur S, Spivakov M, Leleu M, Sauer S, Gregson HC, Jarmuz A, Canzonetta C, Webster Z, Nesterova T *et al*: Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell* 2008, 132(3):422-433.

63. Rubio ED, Reiss DJ, Welcsh PL, Disteche CM, Filippova GN, Baliga NS, Aebersold R, Ranish JA, Krumm A: CTCF physically links cohesin to chromatin. *Proceedings of the National Academy of Sciences of the United States of America* 2008, 105(24):8309-8314.

64. Schwartz YB, Linder-Basso D, Kharchenko PV, Tolstorukov MY, Kim M, Li HB, Gorchakov AA, Minoda A, Shanower G, Alekseyenko AA *et al*: Nature and function of insulator protein binding sites in the Drosophila genome. *Genome research* 2012.

65. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G: Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome. *Cell* 2012.

66. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B: Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 2012, 485(7398):376-380.

67. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL, Meisig J, Sedat J *et al*: Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* 2012, 485(7398):381-385.

68. Bartkuhn M, Straub T, Herold M, Herrmann M, Rathke C, Saumweber H, Gilfillan GD, Becker PB, Renkawitz R: Active promoters and insulators are marked by the centrosomal protein 190. *The EMBO journal* 2009, 28(7):877-888.

69. Soshnev AA, He B, Baxley RM, Jiang N, Hart CM, Tan K, Geyer PK: Genome-wide studies of the multi-zinc finger Drosophila Suppressor of Hairy-wing protein in the ovary. *Nucleic acids research* 2012.

70. Recillas-Targa F, Pikaart MJ, Burgess-Beusse B, Bell AC, Litt MD, West AG, Gaszner M, Felsenfeld G: Position-effect protection and enhancer blocking by the chicken beta-globin insulator are separable activities. *Proceedings of the National Academy of Sciences of the United States of America* 2002, 99(10):6883-6888.

71. Yao S, Osborne CS, Bharadwaj RR, Pasceri P, Sukonnik T, Pannell D, Recillas-Targa F, West AG, Ellis J: Retrovirus silencer blocking by the cHS4 insulator is CTCF independent. *Nucleic acids research* 2003, 31(18):5317-5323.

72. Barkess G, West AG: Chromatin insulator elements: establishing barriers to set heterochromatin boundaries. *Epigenomics* 2012, 4(1):67-80.

73. Phillips JE, Corces VG: CTCF: master weaver of the genome. *Cell* 2009, 137(7):1194-1211.

74. Kurukuti S, Tiwari VK, Tavoosidana G, Pugacheva E, Murrell A, Zhao Z, Lobanenkov V, Reik W, Ohlsson R: CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2. *Proceedings of the National Academy of Sciences of the United States of America* 2006, 103(28):10684-10689.

75. Splinter E, Heath H, Kooren J, Palstra RJ, Klous P, Grosveld F, Galjart N, de Laat W: CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes & development* 2006, 20(17):2349-2354.

76. Kim YJ, Cecchini KR, Kim TH: Conserved, developmentally regulated mechanism couples chromosomal looping and heterochromatin barrier activity at the homeobox gene A locus. ***Proceedings of the National Academy of Sciences of the United States of America*** **2011, 108(18):7391-7396.**

77. Wood AM, Van Bortle K, Ramos E, Takenaka N, Rohrbaugh M, Jones BC, Jones KC, Corces VG: Regulation of chromatin organization and inducible gene expression by a Drosophila insulator. ***Molecular cell*** **2011, 44(1):29-38.**

78. Handoko L, Xu H, Li G, Ngan CY, Chew E, Schnapp M, Lee CW, Ye C, Ping JL, Mulawadi F ***et al*****: CTCF-mediated functional chromatin interactome in pluripotent cells.** ***Nature genetics*** **2011, 43(7):630-638.**

79. Hadjur S, Williams LM, Ryan NK, Cobb BS, Sexton T, Fraser P, Fisher AG, Merkenschlager M: Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. ***Nature*** **2009, 460(7253):410-413.**

80. Majumder P, Boss JM: Cohesin regulates MHC class II genes through interactions with MHC class II insulators. ***J Immunol*** **2011, 187(8):4236-4244.**

81. Duan Z, Andronescu M, Schutz K, McIlwain S, Kim YJ, Lee C, Shendure J, Fields S, Blau CA, Noble WS: A three-dimensional model of the yeast genome. ***Nature*** **2010, 465(7296):363-367.**

82. Thompson M, Haeusler RA, Good PD, Engelke DR: Nucleolar clustering of dispersed tRNA genes. ***Science*** **2003, 302(5649):1399-1401.**

83. Gerasimova TI, Byrd K, Corces VG: A chromatin insulator determines the nuclear localization of DNA. ***Molecular cell*** **2000, 6(5):1025-1035.**

84. MacPherson MJ, Beatty LG, Zhou W, Du M, Sadowski PD: The CTCF insulator protein is posttranslationally modified by SUMO. ***Molecular and cellular biology*** **2009, 29(3):714-725.**

85. Yusufzai TM, Tagami H, Nakatani Y, Felsenfeld G: CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species. ***Molecular cell*** **2004, 13(2):291-298.**

86. Pickersgill H, Kalverda B, de Wit E, Talhout W, Fornerod M, van Steensel B: Characterization of the Drosophila melanogaster genome at the nuclear lamina. ***Nature genetics*** **2006, 38(9):1005-1014.**

87. Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, Talhout W, Eussen BH, de Klein A, Wessels L, de Laat W ***et al*****: Domain organization of human**

chromosomes revealed by mapping of nuclear lamina interactions. *Nature* 2008, 453(7197):948-951.

88. Zullo JM, Demarco IA, Pique-Regi R, Gaffney DJ, Epstein CB, Spooner CJ, Luperchio TR, Bernstein BE, Pritchard JK, Reddy KL *et al*: DNA sequence-dependent compartmentalization and silencing of chromatin at the nuclear lamina. *Cell* 2012, 149(7):1474-1487.

89. van Bemmel JG, Pagie L, Braunschweig U, Brugman W, Meuleman W, Kerkhoven RM, van Steensel B: The insulator protein SU(HW) fine-tunes nuclear lamina interactions of the Drosophila genome. *PloS one* 2010, 5(11):e15013.

90. Ruben GJ, Kirkland JG, MacDonough T, Chen M, Dubey RN, Gartenberg MR, Kamakaka RT: Nucleoporin mediated nuclear positioning and silencing of HMR. *PloS one* 2011, 6(7):e21923.

91. Sawarkar R, Paro R: Interpretation of developmental signaling at chromatin: the Polycomb perspective. *Developmental cell* 2010, 19(5):651-661.

92. Paro R: Imprinting a determined state into the chromatin of Drosophila. *Trends in genetics : TIG* 1990, 6(12):416-421.

93. Simon J, Chiang A, Bender W, Shimell MJ, O'Connor M: Elements of the Drosophila bithorax complex that mediate repression by Polycomb group products. *Developmental biology* 1993, 158(1):131-144.

94. Woo CJ, Kharchenko PV, Daheron L, Park PJ, Kingston RE: A region of the human HOXD cluster that confers polycomb-group responsiveness. *Cell* 2010, 140(1):99-110.

95. Beisel C, Paro R: Silencing chromatin: comparing modes and mechanisms. *Nature reviews Genetics* 2011, 12(2):123-135.

96. Simon JA, Kingston RE: Mechanisms of polycomb gene silencing: knowns and unknowns. *Nature reviews Molecular cell biology* 2009, 10(10):697-708.

97. Beck DB, Bonasio R, Kaneko S, Li G, Margueron R, Oda H, Sarma K, Sims RJ, 3rd, Son J, Trojer P *et al*: Chromatin in the nuclear landscape. *Cold Spring Harbor symposia on quantitative biology* 2010, 75:11-22.

98. Margueron R, Reinberg D: The Polycomb complex PRC2 and its mark in life. *Nature* 2011, 469(7330):343-349.

99. Kim H, Kang K, Kim J: AEBP2 as a potential targeting protein for Polycomb Repression Complex PRC2. *Nucleic acids research* 2009, 37(9):2940-2950.

100. Li G, Margueron R, Ku M, Chambon P, Bernstein BE, Reinberg D: Jarid2 and PRC2, partners in regulating gene expression. *Genes & development* 2010, 24(4):368-380.

101. Sarma K, Margueron R, Ivanov A, Pirrotta V, Reinberg D: Ezh2 requires PHF1 to efficiently catalyze H3 lysine 27 trimethylation in vivo. *Molecular and cellular biology* 2008, 28(8):2718-2731.

102. Cao R, Tsukada Y, Zhang Y: Role of Bmi-1 and Ring1A in H2A ubiquitylation and Hox gene silencing. *Molecular cell* 2005, 20(6):845-854.

103. Wang H, Wang L, Erdjument-Bromage H, Vidal M, Tempst P, Jones RS, Zhang Y: Role of histone H2A ubiquitination in Polycomb silencing. *Nature* 2004, 431(7010):873-878.

104. Trojer P, Cao AR, Gao Z, Li Y, Zhang J, Xu X, Li G, Losson R, Erdjument-Bromage H, Tempst P *et al*: L3MBTL2 protein acts in concert with PcG protein-mediated monoubiquitination of H2A to establish a repressive chromatin structure. *Molecular cell* 2011, 42(4):438-450.

105. Richter C, Oktaba K, Steinmann J, Muller J, Knoblich JA: The tumour suppressor L(3)mbt inhibits neuroepithelial proliferation and acts on insulator elements. *Nature cell biology* 2011, 13(9):1029-1039.

106. Schwartz YB, Kahn TG, Nix DA, Li XY, Bourgon R, Biggin M, Pirrotta V: Genome-wide analysis of Polycomb targets in Drosophila melanogaster. *Nature genetics* 2006, 38(6):700-705.

107. Tolhuis B, de Wit E, Muijrers I, Teunissen H, Talhout W, van Steensel B, van Lohuizen M: Genome-wide profiling of PRC1 and PRC2 Polycomb chromatin binding in Drosophila melanogaster. *Nature genetics* 2006, 38(6):694-699.

108. Negre N, Hennetin J, Sun LV, Lavrov S, Bellis M, White KP, Cavalli G: Chromosomal distribution of PcG proteins during Drosophila development. *PLoS biology* 2006, 4(6):e170.

109. Oktaba K, Gutierrez L, Gagneur J, Girardot C, Sengupta AK, Furlong EE, Muller J: Dynamic regulation by polycomb group protein complexes controls pattern formation and the cell cycle in Drosophila. *Developmental cell* 2008, 15(6):877-889.

110. Boyer LA, Plath K, Zeitlinger J, Brambrink T, Medeiros LA, Lee TI, Levine SS, Wernig M, Tajonar A, Ray MK *et al*: Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* 2006, 441(7091):349-353.

111. Bracken AP, Dietrich N, Pasini D, Hansen KH, Helin K: Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions. *Genes & development* 2006, 20(9):1123-1136.

112. Lee TI, Jenner RG, Boyer LA, Guenther MG, Levine SS, Kumar RM, Chevalier B, Johnstone SE, Cole MF, Isono K *et al*: Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* 2006, 125(2):301-313.

113. Schuettengruber B, Ganapathi M, Leblanc B, Portoso M, Jaschek R, Tolhuis B, van Lohuizen M, Tanay A, Cavalli G: Functional anatomy of polycomb and trithorax chromatin landscapes in Drosophila embryos. *PLoS biology* 2009, 7(1):e13.

114. Kwong C, Adryan B, Bell I, Meadows L, Russell S, Manak JR, White R: Stability and dynamics of polycomb target sites in Drosophila development. *PLoS genetics* 2008, 4(9):e1000178.

115. Schwartz YB, Kahn TG, Stenberg P, Ohno K, Bourgon R, Pirrotta V: Alternative epigenetic chromatin states of polycomb target genes. *PLoS genetics* 2010, 6(1):e1000805.

116. Zhao J, Ohsumi TK, Kung JT, Ogawa Y, Grau DJ, Sarma K, Song JJ, Kingston RE, Borowsky M, Lee JT: Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Molecular cell* 2010, 40(6):939-953.

117. Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, Shi Y, Segal E, Chang HY: Long noncoding RNA as modular scaffold of histone modification complexes. *Science* 2010, 329(5992):689-693.

118. Chu C, Qu K, Zhong FL, Artandi SE, Chang HY: Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Molecular cell* 2011, 44(4):667-678.

119. Ficz G, Heintzmann R, Arndt-Jovin DJ: Polycomb group protein complexes exchange rapidly in living Drosophila. *Development* 2005, 132(17):3963-3976.

120. Grimaud C, Bantignies F, Pal-Bhadra M, Ghana P, Bhadra U, Cavalli G: RNAi components are required for nuclear clustering of Polycomb group response elements. *Cell* 2006, 124(5):957-971.

121. Lanzuolo C, Roure V, Dekker J, Bantignies F, Orlando V: Polycomb response elements mediate the formation of chromosome higher-order structures in the bithorax complex. *Nature cell biology* 2007, 9(10):1167-1174.

122. Tiwari VK, Cope L, McGarvey KM, Ohm JE, Baylin SB: A novel 6C assay uncovers Polycomb-mediated higher order chromatin conformations. *Genome research* 2008, 18(7):1171-1179.

123. Tiwari VK, McGarvey KM, Licchesi JD, Ohm JE, Herman JG, Schubeler D, Baylin SB: PcG proteins, DNA methylation, and gene repression by chromatin looping. *PLoS biology* 2008, 6(12):2911-2927.

124. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, van Steensel B, de Laat W: Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature genetics* 2006, 38(11):1348-1354.

125. Tolhuis B, Blom M, Kerkhoven RM, Pagie L, Teunissen H, Nieuwland M, Simonis M, de Laat W, van Lohuizen M, van Steensel B: Interactions among Polycomb domains are guided by chromosome architecture. *PLoS genetics* 2011, 7(3):e1001343.

126. Li HB, Muller M, Bahechar IA, Kyrchanova O, Ohno K, Georgiev P, Pirrotta V: Insulators, not Polycomb response elements, are required for long-range interactions between Polycomb targets in Drosophila melanogaster. *Molecular and cellular biology* 2011, 31(4):616-625.

127. Gruzdeva N, Kyrchanova O, Parshikov A, Kullyev A, Georgiev P: The Mcp element from the bithorax complex contains an insulator that is capable of pairwise interactions and can facilitate enhancer-promoter communication. *Molecular and cellular biology* 2005, 25(9):3682-3689.

128. Zhou J, Barolo S, Szymanski P, Levine M: The Fab-7 element of the bithorax complex attenuates enhancer-promoter interactions in the Drosophila embryo. *Genes & development* 1996, 10(24):3195-3201.

129. Comet I, Schuettengruber B, Sexton T, Cavalli G: A chromatin insulator driving three-dimensional Polycomb response element (PRE) contacts and Polycomb association with the chromatin fiber. *Proceedings of the National Academy of Sciences of the United States of America* 2011, 108(6):2294-2299.

130. Sigrist CJ, Pirrotta V: Chromatin insulator elements block the silencing of a target gene by the Drosophila polycomb response element (PRE) but allow trans interactions between PREs on different chromosomes. *Genetics* 1997, 147(1):209-221.

131. Bantignies F, Roure V, Comet I, Leblanc B, Schuettengruber B, Bonnet J, Tixier V, Mas A, Cavalli G: Polycomb-dependent regulatory contacts between distant Hox loci in Drosophila. ***Cell*** **2011, 144(2):214-226.**

132. Grau DJ, Chapman BA, Garlick JD, Borowsky M, Francis NJ, Kingston RE: Compaction of chromatin by diverse Polycomb group proteins requires localized regions of high charge. ***Genes & development*** **2011, 25(20):2210-2221.**

133. Bell O, Schwaiger M, Oakeley EJ, Lienert F, Beisel C, Stadler MB, Schubeler D: Accessibility of the Drosophila genome discriminates PcG repression, H4K16 acetylation and replication timing. ***Nature structural & molecular biology*** **2010, 17(7):894-900.**

134. Noordermeer D, Leleu M, Splinter E, Rougemont J, De Laat W, Duboule D: The dynamic architecture of Hox gene clusters. ***Science*** **2011, 334(6053):222-225.**

135. Werner F, Grohmann D: Evolution of multisubunit RNA polymerases in the three domains of life. ***Nature reviews Microbiology*** **2011, 9(2):85-98.**

136. Thiry M, Lafontaine DL: Birth of a nucleolus: the evolution of nucleolar compartments. ***Trends in cell biology*** **2005, 15(4):194-199.**

137. Iborra FJ, Pombo A, Jackson DA, Cook PR: Active RNA polymerases are localized within discrete transcription "factories' in human nuclei. ***Journal of cell science*** **1996, 109 (Pt 6):1427-1436.**

138. Osborne CS, Chakalova L, Brown KE, Carter D, Horton A, Debrand E, Goyenechea B, Mitchell JA, Lopes S, Reik W ***et al*****: Active genes dynamically colocalize to shared sites of ongoing transcription.** ***Nature genetics*** **2004, 36(10):1065-1071.**

139. Tanizawa H, Iwasaki O, Tanaka A, Capizzi JR, Wickramasinghe P, Lee M, Fu Z, Noma K: Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. ***Nucleic acids research*** **2010, 38(22):8164-8177.**

140. Melnik S, Deng B, Papantonis A, Baboo S, Carr IM, Cook PR: The proteomes of transcription factories containing RNA polymerases I, II or III. ***Nature methods*** **2011, 8(11):963-968.**

141. Yang L, Lin C, Liu W, Zhang J, Ohgi KA, Grinstein JD, Dorrestein PC, Rosenfeld MG: ncRNA- and Pc2 methylation-dependent gene relocation between nuclear structures mediates gene activation programs. ***Cell*** **2011, 147(4):773-788.**

142. Pirrotta V, Li HB: A view of nuclear Polycomb bodies. *Current opinion in genetics & development* 2011.

143. Gerasimova TI, Corces VG: Polycomb and trithorax group proteins mediate the function of a chromatin insulator. *Cell* 1998, 92(4):511-521.

144. Banerji J, Rusconi S, Schaffner W: Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* 1981, 27(2 Pt 1):299-308.

145. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA *et al*: Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics* 2007, 39(3):311-318.

146. Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh TY, Peng W, Zhang MQ *et al*: Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature genetics* 2008, 40(7):897-903.

147. Ong CT, Corces VG: Enhancers: emerging roles in cell fate specification. *EMBO reports* 2012, 13(5):423-430.

148. Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S *et al*: Widespread transcription at neuronal activity-regulated enhancers. *Nature* 2010, 465(7295):182-187.

149. Wang D, Garcia-Bassets I, Benner C, Li W, Su X, Zhou Y, Qiu J, Liu W, Kaikkonen MU, Ohgi KA *et al*: Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* 2011, 474(7351):390-394.

150. Bulger M, Groudine M: Functional and mechanistic diversity of distal transcription enhancers. *Cell* 2011, 144(3):327-339.

151. Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, Poh HM, Goh Y, Lim J, Zhang J *et al*: Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* 2012, 148(1-2):84-98.

152. Maston GA, Evans SK, Green MR: Transcriptional regulatory elements in the human genome. *Annual review of genomics and human genetics* 2006, 7:29-59.

153. Zippo A, Serafini R, Rocchigiani M, Pennacchini S, Krepelova A, Oliviero S: Histone crosstalk between H3S10ph and H4K16ac generates a histone code that mediates transcription elongation. *Cell* 2009, 138(6):1122-1136.

154. Vernimmen D, Lynch MD, De Gobbi M, Garrick D, Sharpe JA, Sloane-Stanley JA, Smith AJ, Higgs DR: Polycomb eviction as a new distant enhancer function. *Genes & development* 2011, 25(15):1583-1588.

155. Drissen R, Palstra RJ, Gillemans N, Splinter E, Grosveld F, Philipsen S, de Laat W: The active spatial organization of the beta-globin locus requires the transcription factor EKLF. *Genes & development* 2004, 18(20):2485-2490.

156. Vakoc CR, Letting DL, Gheldof N, Sawado T, Bender MA, Groudine M, Weiss MJ, Dekker J, Blobel GA: Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1. *Molecular cell* 2005, 17(3):453-462.

157. Schoenfelder S, Sexton T, Chakalova L, Cope NF, Horton A, Andrews S, Kurukuti S, Mitchell JA, Umlauf D, Dimitrova DS *et al*: Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nature genetics* 2010, 42(1):53-61.

158. Noordermeer D, de Wit E, Klous P, van de Werken H, Simonis M, Lopez-Jones M, Eussen B, de Klein A, Singer RH, de Laat W: Variegated gene expression caused by cell-specific long-range DNA interactions. *Nature cell biology* 2011, 13(8):944-951.

159. Gillen AE, Harris A: The role of CTCF in coordinating the expression of single gene loci. *Biochemistry and cell biology = Biochimie et biologie cellulaire* 2011, 89(5):489-494.

160. Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, Ebmeier CC, Goossens J, Rahl PB, Levine SS *et al*: Mediator and cohesin connect gene expression and chromatin architecture. *Nature* 2010, 467(7314):430-435.

161. Chien R, Zeng W, Kawauchi S, Bender MA, Santos R, Gregson HC, Schmiesing JA, Newkirk DA, Kong X, Ball AR, Jr. *et al*: Cohesin mediates chromatin interactions that regulate mammalian beta-globin expression. *The Journal of biological chemistry* 2011, 286(20):17870-17878.

162. Gaszner M, Felsenfeld G: Insulators: exploiting transcriptional and epigenetic mechanisms. *Nature reviews Genetics* 2006, 7(9):703-713.

163. Bushey AM, Dorman ER, Corces VG: Chromatin insulators: regulatory mechanisms and epigenetic inheritance. *Molecular cell* 2008, 32(1):1-9.

164. Chao W, Huynh KD, Spencer RJ, Davidow LS, Lee JT: CTCF, a candidate trans-acting factor for X-inactivation choice. *Science* 2002, 295(5553):345-347.

165. Guerrero PA, Maggert KA: The CCCTC-binding factor (CTCF) of Drosophila contributes to the regulation of the ribosomal DNA and nucleolar stability. *PloS one* 2011, 6(1):e16401.

166. Guo C, Yoon HS, Franklin A, Jain S, Ebert A, Cheng HL, Hansen E, Despo O, Bossen C, Vettermann C *et al*: CTCF-binding elements mediate control of V(D)J recombination. *Nature* 2011, 477(7365):424-430.

167. Pai CY, Lei EP, Ghosh D, Corces VG: The centrosomal protein CP190 is a component of the gypsy chromatin insulator. *Molecular cell* 2004, 16(5):737-748.

168. Machanick P, Bailey TL: MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* 2011, 27(12):1696-1697.

169. Ramos E, Ghosh D, Baxter E, Corces VG: Genomic organization of gypsy chromatin insulators in Drosophila melanogaster. *Genetics* 2006, 172(4):2337-2349.

170. Ohlsson R, Renkawitz R, Lobanenkov V: CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends Genet* 2001, 17(9):520-527.

171. Xie X, Mikkelsen TS, Gnirke A, Lindblad-Toh K, Kellis M, Lander ES: Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proceedings of the National Academy of Sciences of the United States of America* 2007, 104(17):7145-7150.

172. Pavesi G, Mereghetti P, Mauri G, Pesole G: Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic acids research* 2004, 32(Web Server issue):W199-203.

173. Essien K, Vigneau S, Apreleva S, Singh LN, Bartolomei MS, Hannenhalli S: CTCF binding site classes exhibit distinct evolutionary, genomic, epigenomic and transcriptomic features. *Genome biology* 2009, 10(11):R131.

174. Mohan M, Bartkuhn M, Herold M, Philippen A, Heinl N, Bardenhagen I, Leers J, White RA, Renkawitz-Pohl R, Saumweber H *et al*: The Drosophila insulator proteins CTCF and CP190 link enhancer blocking to body patterning. *The EMBO journal* 2007, 26(19):4203-4214.

175. Oliver D, Sheehan B, South H, Akbari O, Pai CY: The chromosomal association/dissociation of the chromatin insulator protein Cp190 of Drosophila melanogaster is mediated by the BTB/POZ domain and two acidic regions. *BMC cell biology* 2010, 11:101.

176. Gerasimova TI, Gdula DA, Gerasimov DV, Simonova O, Corces VG: A Drosophila protein that imparts directionality on a chromatin insulator is an enhancer of position-effect variegation. *Cell* 1995, 82(4):587-597.

177. Dorn R, Krauss V: The modifier of mdg4 locus in Drosophila: functional complexity is resolved by trans splicing. *Genetica* 2003, 117(2-3):165-177.

178. Guruharsha KG, Rual JF, Zhai B, Mintseris J, Vaidya P, Vaidya N, Beekman C, Wong C, Rhee DY, Cenaj O *et al*: A protein complex network of Drosophila melanogaster. *Cell* 2011, 147(3):690-703.

179. Gan M, Moebus S, Eggert H, Saumweber H: The Chriz-Z4 complex recruits JIL-1 to polytene chromosomes, a requirement for interband-specific phosphorylation of H3S10. *Journal of biosciences* 2011, 36(3):425-438.

180. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E *et al*: A protein interaction map of Drosophila melanogaster. *Science* 2003, 302(5651):1727-1736.

181. Eggert H, Gortchakov A, Saumweber H: Identification of the Drosophila interband-specific protein Z4 as a DNA-binding zinc-finger protein determining chromosomal structure. *Journal of cell science* 2004, 117(Pt 18):4253-4264.

182. Whyte WA, Bilodeau S, Orlando DA, Hoke HA, Frampton GM, Foster CT, Cowley SM, Young RA: Enhancer decommissioning by LSD1 during embryonic stem cell differentiation. *Nature* 2012, 482(7384):221-225.

183. Pauler FM, Sloane MA, Huang R, Regha K, Koerner MV, Tamir I, Sommer A, Aszodi A, Jenuwein T, Barlow DP: H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome. *Genome research* 2009, 19(2):221-233.

184. Negre N, Brown CD, Ma L, Bristow CA, Miller SW, Wagner U, Kheradpour P, Eaton ML, Loriaux P, Sealfon R *et al*: A cis-regulatory map of the Drosophila genome. *Nature* 2011, 471(7339):527-531.

185. Macdonald PM, Ingham P, Struhl G: Isolation, structure, and expression of even-skipped: a second pair-rule gene of Drosophila containing a homeo box. *Cell* 1986, 47(5):721-734.

186. Gelbart WM, Emmert DB: FlyBase High Throughput Expression Pattern Data Beta Version. In.; 2010.

187. Kim SN, Shim HP, Jeon BN, Choi WI, Hur MW, Girton JR, Kim SH, Jeon SH: The pleiohomeotic functions as a negative regulator of Drosophila even-skipped gene during embryogenesis. ***Molecules and cells*** **2011.**

188. Schoborg TA, Labrador M: The phylogenetic distribution of non-CTCF insulator proteins is limited to insects and reveals that BEAF-32 is Drosophila lineage specific. ***Journal of molecular evolution*** **2010, 70(1):74-84.**

189. Langmead B: Aligning short sequencing reads with Bowtie. ***Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]*** **2010, Chapter 11:Unit 11 17.**

190. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W ***et al*****: Model-based analysis of ChIP-Seq (MACS).** ***Genome biology*** **2008, 9(9):R137.**

191. Goecks J, Nekrutenko A, Taylor J: Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. ***Genome biology*** **2010, 11(8):R86.**

192. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J ***et al*****: Galaxy: a platform for interactive large-scale genome analysis.** ***Genome research*** **2005, 15(10):1451-1455.**

193. Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J: Galaxy: a web-based genome analysis tool for experimentalists. ***Current protocols in molecular biology / edited by Frederick M Ausubel [et al]*** **2010, Chapter 19:Unit 19 10 11-21.**

194. de Hoon MJ, Imoto S, Nolan J, Miyano S: Open source clustering software. ***Bioinformatics*** **2004, 20(9):1453-1454.**

195. Saldanha AJ: Java Treeview--extensible visualization of microarray data. ***Bioinformatics*** **2004, 20(17):3246-3248.**

196. Ivaldi MS, Karam CS, Corces VG: Phosphorylation of histone H3 at Ser10 facilitates RNA polymerase II release from promoter-proximal pausing in Drosophila. ***Genes & development*** **2007, 21(21):2818-2831.**

197. Armknecht S, Boutros M, Kiger A, Nybakken K, Mathey-Prevot B, Perrimon N: High-throughput RNA interference screens in Drosophila tissue culture cells. ***Methods in enzymology*** **2005, 392:55-73.**

198. Hou C, Li L, Qin Z, Corces V: Gene density, transcription and insulators contribute to partitioning the Drosophila genome into physical domains. ***Molecular cell*** **2012, In press.**

199. Phillips-Cremins JE, Sauria ME, Sanyal A, Gerasimova TI, Lajoie BR, Bell JS, Ong CT, Hookway TA, Guo C, Sun Y *et al*: Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* 2013, 153(6):1281-1295.

200. Seitan VC, Faure AJ, Zhan Y, McCord RP, Lajoie BR, Ing-Simmons E, Lenhard B, Giorgetti L, Heard E, Fisher AG *et al*: Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments. *Genome research* 2013.

201. Naumova N, Imakaev M, Fudenberg G, Zhan Y, Lajoie BR, Mirny LA, Dekker J: Organization of the mitotic chromosome. *Science* 2013, 342(6161):948-953.

202. Zuin J, Dixon JR, van der Reijden MI, Ye Z, Kolovos P, Brouwer RW, van de Corput MP, van de Werken HJ, Knoch TA, van Ijcken WF *et al*: Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proceedings of the National Academy of Sciences of the United States of America* 2014, 111(3):996-1001.

203. Sofueva S, Yaffe E, Chan WC, Georgopoulou D, Vietri Rudan M, Mira-Bontenbal H, Pollard SM, Schroth GP, Tanay A, Hadjur S: Cohesin-mediated interactions organize chromosomal domain architecture. *The EMBO journal* 2013, 32(24):3119-3129.

204. Dekker J, Mirny L: Biological techniques: Chromosomes captured one by one. *Nature* 2013, 502(7469):45-46.

205. Schuettengruber B, Cavalli G: Polycomb domain formation depends on short and long distance regulatory cues. *PloS one* 2013, 8(2):e56531.

206. Sanyal A, Lajoie BR, Jain G, Dekker J: The long-range interaction landscape of gene promoters. *Nature* 2012, 489(7414):109-113.

207. Xu Z, Wei G, Chepelev I, Zhao K, Felsenfeld G: Mapping of INS promoter interactions reveals its role in long-range regulation of SYT8 transcription. *Nature structural & molecular biology* 2011, 18(3):372-378.

208. Soshnev AA, Baxley RM, Manak JR, Tan K, Geyer PK: The insulator protein Suppressor of Hairy-wing is an essential transcriptional repressor in the Drosophila ovary. *Development* 2013, 140(17):3613-3623.

209. Van Bortle K, Corces VG: Nuclear Organization and Genome Function. *Annual review of cell and developmental biology* 2012.

210. Heger P, George R, Wiehe T: Successive gain of insulator proteins in arthropod evolution. *Evolution; international journal of organic evolution* 2013, 67(10):2945-2956.

211. White RJ: Transcription by RNA polymerase III: more complex than we thought. *Nature reviews Genetics* 2011, 12(7):459-463.

212. Dumay-Odelot H, Marck C, Durrieu-Gaillard S, Lefebvre O, Jourdain S, Prochazkova M, Pflieger A, Teichmann M: Identification, molecular cloning, and characterization of the sixth subunit of human transcription factor TFIIIC. *The Journal of biological chemistry* 2007, 282(23):17179-17189.

213. Bailey TL, Machanick P: Inferring direct DNA binding from ChIP-seq. *Nucleic acids research* 2012, 40(17):e128.

214. Marygold SJ, Leyland PC, Seal RL, Goodman JL, Thurmond J, Strelets VB, Wilson RJ: FlyBase: improvements to the bibliography. *Nucleic acids research* 2013, 41(Database issue):D751-757.

215. Iwasaki O, Tanaka A, Tanizawa H, Grewal SI, Noma K: Centromeric localization of dispersed Pol III genes in fission yeast. *Molecular biology of the cell* 2010, 21(2):254-265.

216. Iwasaki O, Noma K: Global genome organization mediated by RNA polymerase III-transcribed genes in fission yeast. *Gene* 2012, 493(2):195-200.

217. Kim JH, Zhang T, Wong NC, Davidson N, Maksimovic J, Oshlack A, Earnshaw WC, Kalitsis P, Hudson DF: Condensin I associates with structural and gene regulatory regions in vertebrate chromosomes. *Nature communications* 2013, 4:2537.

218. Kvon EZ, Stampfel G, Yanez-Cuna JO, Dickson BJ, Stark A: HOT regions function as patterned developmental enhancers and have a distinct cis-regulatory signature. *Genes & development* 2012, 26(9):908-913.

219. Yan J, Enge M, Whitington T, Dave K, Liu J, Sur I, Schmierer B, Jolma A, Kivioja T, Taipale M *et al*: Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell* 2013, 154(4):801-813.

220. Faure AJ, Schmidt D, Watt S, Schwalie PC, Wilson MD, Xu H, Ramsay RG, Odom DT, Flicek P: Cohesin regulates tissue-specific expression by stabilizing highly occupied cis-regulatory modules. *Genome research* 2012, 22(11):2163-2175.

221. Schaaf CA, Kwak H, Koenig A, Misulovin Z, Gohara DW, Watson A, Zhou Y, Lis JT, Dorsett D: Genome-wide control of RNA polymerase II activity by cohesin. *PLoS genetics* 2013, 9(3):e1003382.

222. Dowen JM, Bilodeau S, Orlando DA, Hubner MR, Abraham BJ, Spector DL, Young RA: Multiple structural maintenance of chromosome complexes at transcriptional regulatory elements. *Stem cell reports* 2013, 1(5):371-378.

223. Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A: Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* 2013, 339(6123):1074-1077.

224. Gurudatta BV, Yang J, Van Bortle K, Donlin-Asp PG, Corces VG: Dynamic changes in the genomic localization of DNA replication-related element binding factor during the cell cycle. *Cell Cycle* 2013, 12(10):1605-1615.

225. Kharchenko PV, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, Ernst J, Sabo PJ, Larschan E, Gorchakov AA, Gu T *et al*: Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature* 2011, 471(7339):480-485.

226. Hou C, Li L, Qin ZS, Corces VG: Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. *Molecular cell* 2012, 48(3):471-484.

227. Thomas S, Li XY, Sabo PJ, Sandstrom R, Thurman RE, Canfield TK, Giste E, Fisher W, Hammonds A, Celniker SE *et al*: Dynamic reprogramming of chromatin accessibility during Drosophila embryo development. *Genome biology* 2011, 12(5):R43.

228. Galli GG, Carrara M, Francavilla C, Honnens de Lichtenberg K, Olsen JV, Calogero RA, Lund AH: Genomic and proteomic analyses of prdm5 reveal interactions with insulator binding proteins in embryonic stem cells. *Molecular and cellular biology* 2013, 33(22):4504-4516.

229. Xie D, Boyle AP, Wu L, Zhai J, Kawli T, Snyder M: Dynamic trans-acting factor colocalization in human cells. *Cell* 2013, 155(3):713-724.

230. Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M: An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012, 489(7414):57-74.

231. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 2004, 306(5696):636-640.

232. Phillips-Cremins JE, Corces VG: Chromatin insulators: linking genome organization to cellular function. *Molecular cell* 2013, 50(4):461-474.

233. Huang K, Jia J, Wu C, Yao M, Li M, Jin J, Jiang C, Cai Y, Pei D, Pan G *et al*: Ribosomal RNA gene transcription mediated by the master genome regulator protein CCCTC-binding factor (CTCF) is negatively regulated by the condensin complex. *The Journal of biological chemistry* 2013, 288(36):26067-26077.

234. Bauer CR, Hartl TA, Bosco G: Condensin II promotes the formation of chromosome territories by inducing axial compaction of polyploid interphase chromosomes. *PLoS genetics* 2012, 8(8):e1002873.

235. Buster DW, Daniel SG, Nguyen HQ, Windler SL, Skwarek LC, Peterson M, Roberts M, Meserve JH, Hartl T, Klebba JE *et al*: SCFSlimb ubiquitin ligase suppresses condensin II-mediated nuclear reorganization by degrading Cap-H2. *The Journal of cell biology* 2013.

236. Yang J, Sung E, Donlin-Asp PG, Corces VG: A subset of Drosophila Myc sites remain associated with mitotic chromosomes colocalized with insulator proteins. *Nature communications* 2013, 4:1464.

237. Ong CT, Van Bortle K, Ramos E, Corces VG: Poly(ADP-ribosyl)ation Regulates Insulator Function and Intrachromosomal Interactions in Drosophila. *Cell* 2013, 155(1):148-159.

238. Bhat MA, Philp AV, Glover DM, Bellen HJ: Chromatid segregation at anaphase requires the barren product, a novel chromosome-associated protein that interacts with Topoisomerase II. *Cell* 1996, 87(6):1103-1114.

239. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: Integrative genomics viewer. *Nature biotechnology* 2011, 29(1):24-26.

240. Thorvaldsdottir H, Robinson JT, Mesirov JP: Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics* 2013, 14(2):178-192.

241. Isogai Y, Takada S, Tjian R, Keles S: Novel TRF1/BRF target genes revealed by genome-wide analysis of Drosophila Pol III transcription. *The EMBO journal* 2007, 26(1):79-89.

242. Rhee HS, Pugh BF: Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* 2011, 147(6):1408-1419.

243. Marchler-Bauer A, Zheng C, Chitsaz F, Derbyshire MK, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Lanczycki CJ *et al*: CDD: conserved domains and protein three-dimensional structure. *Nucleic acids research* 2013, 41(Database issue):D348-352.

244. Porter SG, Day J, McCarty RE, Shearn A, Shingles R, Fletcher L, Murphy S, Pearlman R: Exploring DNA structure with Cn3D. *CBE life sciences education* 2007, 6(1):65-73.

245. Wang Y, Geer LY, Chappey C, Kans JA, Bryant SH: Cn3D: sequence and structure views for Entrez. *Trends in biochemical sciences* 2000, 25(6):300-302.

246. Hogue CW: Cn3D: a new generation of three-dimensional molecular structure viewer. *Trends in biochemical sciences* 1997, 22(8):314-316.

247. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ: Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 2009, 25(9):1189-1191.

248. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: The human genome browser at UCSC. *Genome research* 2002, 12(6):996-1006.

249. Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B *et al*: The UCSC Genome Browser database: extensions and updates 2013. *Nucleic acids research* 2013, 41(Database issue):D64-69.

250. Quinlan AR, Hall IM: BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010, 26(6):841-842.

251. Van Bortle K, Corces VG: The role of chromatin insulators in nuclear architecture and genome function. *Current opinion in genetics & development* 2013.

252. Gomez-Diaz E, Corces VG: Architectural proteins: regulators of 3D genome organization in cell fate. *Trends in cell biology* 2014.

253. Van Bortle K, Nichols MH, Li L, Ong CT, Takenaka N, Qin ZS, Corces VG: Insulator function and topological domain border strength scale with architectural protein occupancy. *Genome biology* 2014, 15(6):R82.

254. Li TR, White KP: Tissue-specific gene expression and ecdysone-regulated genomic networks in Drosophila. *Developmental cell* 2003, 5(1):59-72.

255. Liu F: Receptor-regulated Smads in TGF-beta signaling. ***Frontiers in bioscience : a journal and virtual library*** **2003, 8:s1280-1303.**

256. Ross S, Hill CS: How the Smads regulate transcription. ***The international journal of biochemistry & cell biology*** **2008, 40(3):383-408.**

257. Massague J: How cells read TGF-beta signals. ***Nature reviews Molecular cell biology*** **2000, 1(3):169-178.**

258. Trompouki E, Bowman TV, Lawton LN, Fan ZP, Wu DC, DiBiase A, Martin CS, Cech JN, Sessa AK, Leblanc JL ***et al*****: Lineage regulators direct BMP and Wnt pathways to cell-specific programs during differentiation and regeneration.** ***Cell*** **2011, 147(3):577-589.**

259. Mullen AC, Orlando DA, Newman JJ, Loven J, Kumar RM, Bilodeau S, Reddy J, Guenther MG, DeKoter RP, Young RA: Master transcription factors determine cell-type-specific responses to TGF-beta signaling. ***Cell*** **2011, 147(3):565-576.**

260. Xi Q, Wang Z, Zaromytidou AI, Zhang XH, Chow-Tsang LF, Liu JX, Kim H, Barlas A, Manova-Todorova K, Kaartinen V ***et al*****: A poised chromatin platform for TGF-beta access to master regulators.** ***Cell*** **2011, 147(7):1511-1524.**

261. Burton T, Liang B, Dibrov A, Amara F: Transforming growth factor-beta-induced transcription of the Alzheimer beta-amyloid precursor protein gene involves interaction between the CTCF-complex and Smads. ***Biochemical and biophysical research communications*** **2002, 295(3):713-723.**

262. Bergstrom R, Savary K, Moren A, Guibert S, Heldin CH, Ohlsson R, Moustakas A: Transforming growth factor beta promotes complexes between Smad proteins and the CCCTC-binding factor on the H19 imprinting control region chromatin. ***The Journal of biological chemistry*** **2010, 285(26):19727-19737.**

263. Peterson AJ, O'Connor MB: Strategies for exploring TGF-beta signaling in Drosophila. ***Methods*** **2014, 68(1):183-193.**

264. Brummel T, Abdollah S, Haerry TE, Shimell MJ, Merriam J, Raftery L, Wrana JL, O'Connor MB: The Drosophila activin receptor baboon signals through dSmad2 and controls cell proliferation but not patterning during larval development. ***Genes & development*** **1999, 13(1):98-111.**

265. Brummel TJ, Twombly V, Marques G, Wrana JL, Newfeld SJ, Attisano L, Massague J, O'Connor MB, Gelbart WM: Characterization and relationship

of Dpp receptors encoded by the saxophone and thick veins genes in Drosophila. *Cell* 1994, 78(2):251-261.

266. Letsou A, Arora K, Wrana JL, Simin K, Twombly V, Jamal J, Staehling-Hampton K, Hoffmann FM, Gelbart WM, Massague J *et al*: Drosophila Dpp signaling is mediated by the punt gene product: a dual ligand-binding type II receptor of the TGF beta receptor family. *Cell* 1995, 80(6):899-908.

267. Gao S, Steffen J, Laughon A: Dpp-responsive silencers are bound by a trimeric Mad-Medea complex. *The Journal of biological chemistry* 2005, 280(43):36158-36164.

268. Weiss A, Charbonnier E, Ellertsdottir E, Tsirigos A, Wolf C, Schuh R, Pyrowolakis G, Affolter M: A conserved activation element in BMP signaling during Drosophila development. *Nature structural & molecular biology* 2010, 17(1):69-76.

269. Peterson AJ, Jensen PA, Shimell M, Stefancsik R, Wijayatonge R, Herder R, Raftery LA, O'Connor MB: R-Smad competition controls activin receptor output in Drosophila. *PloS one* 2012, 7(5):e36548.

270. Peterson AJ, O'Connor MB: Activin receptor inhibition by Smad2 regulates Drosophila wing disc patterning through BMP-response elements. *Development* 2013, 140(3):649-659.

271. Yao LC, Phin S, Cho J, Rushlow C, Arora K, Warrior R: Multiple modular promoter elements drive graded brinker expression in response to the Dpp morphogen gradient. *Development* 2008, 135(12):2183-2192.

272. Xu X, Yin Z, Hudson JB, Ferguson EL, Frasch M: Smad proteins act in combination with synergistic and antagonistic regulators to target Dpp responses to the Drosophila mesoderm. *Genes & development* 1998, 12(15):2354-2370.

273. Kusanagi K, Inoue H, Ishidou Y, Mishima HK, Kawabata M, Miyazono K: Characterization of a bone morphogenetic protein-responsive Smad-binding element. *Molecular biology of the cell* 2000, 11(2):555-565.

274. Shao Z, Zhang Y, Yuan GC, Orkin SH, Waxman DJ: MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. *Genome biology* 2012, 13(3):R16.

275. Symmons O, Uslu VV, Tsujimura T, Ruf S, Nassari S, Schwarzer W, Ettwiller L, Spitz F: Functional and topological characteristics of mammalian regulatory domains. *Genome research* 2014, 24(3):390-400.

276. Xu L, Yao X, Chen X, Lu P, Zhang B, Ip YT: Msk is required for nuclear import of TGF-{beta}/BMP-activated Smads. *The Journal of cell biology* 2007, 178(6):981-994.

277. Sutherland DJ, Li M, Liu XQ, Stefancsik R, Raftery LA: Stepwise formation of a SMAD activity gradient during dorsal-ventral patterning of the Drosophila embryo. *Development* 2003, 130(23):5705-5716.

278. MacArthur S, Li XY, Li J, Brown JB, Chu HC, Zeng L, Grondona BP, Hechmer A, Simirenko L, Keranen SV *et al*: Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome biology* 2009, 10(7):R80.

279. Huang da W, Sherman BT, Zheng X, Yang J, Imamichi T, Stephens R, Lempicki RA: Extracting biological meaning from large gene lists with DAVID. *Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]* 2009, Chapter 13:Unit 13 11.

280. Supek F, Bosnjak M, Skunca N, Smuc T: REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS one* 2011, 6(7):e21800.

281. Gauhar Z, Sun LV, Hua S, Mason CE, Fuchs F, Li TR, Boutros M, White KP: Genomic mapping of binding regions for the Ecdysone receptor protein complex. *Genome research* 2009, 19(6):1006-1013.

282. Kuan PF, Chun H, Keles S: CMARRT: a tool for the analysis of ChIP-chip data from tiling arrays by incorporating the correlation structure. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing* 2008:515-526.

283. Novoa EM, Ribas de Pouplana L: Speeding with control: codon usage, tRNAs, and ribosomes. *Trends in genetics : TIG* 2012, 28(11):574-581.

284. Ikemura T: Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *Journal of molecular biology* 1981, 151(3):389-409.

285. Ikemura T: Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs. *Journal of molecular biology* 1982, 158(4):573-597.

286. Ingolia NT: Genome-wide translational profiling by ribosome footprinting. *Methods in enzymology* 2010, 470:119-142.

287. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS: Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 2009, 324(5924):218-223.

288. Qian W, Yang JR, Pearson NM, Maclean C, Zhang J: Balanced codon usage optimizes eukaryotic translational efficiency. *PLoS genetics* 2012, 8(3):e1002603.

289. Gingold H, Tehler D, Christoffersen NR, Nielsen MM, Asmar F, Kooistra SM, Christophersen NS, Christensen LL, Borre M, Sorensen KD *et al*: A dual program for translation regulation in cellular proliferation and differentiation. *Cell* 2014, 158(6):1281-1292.

290. Cherbas L, Willingham A, Zhang D, Yang L, Zou Y, Eads BD, Carlson JW, Landolin JM, Kapranov P, Dumais J *et al*: The transcriptional diversity of 25 Drosophila cell lines. *Genome research* 2011, 21(2):301-314.

291. Stofanko M, Kwon SY, Badenhorst P: Lineage tracing of lamellocytes demonstrates Drosophila macrophage plasticity. *PloS one* 2010, 5(11):e14051.

292. Kelsey EM, Luo X, Bruckner K, Jasper H: Schnurri regulates hemocyte function to promote tissue recovery after DNA damage. *Journal of cell science* 2012, 125(Pt 6):1393-1400.

293. Luhur A, Chawla G, Sokol NS: MicroRNAs as components of systemic signaling pathways in Drosophila melanogaster. *Current topics in developmental biology* 2013, 105:97-123.

294. Bashirullah A, Pasquinelli AE, Kiger AA, Perrimon N, Ruvkun G, Thummel CS: Coordinate regulation of small temporal RNAs at the onset of Drosophila metamorphosis. *Developmental biology* 2003, 259(1):1-8.

295. Varghese J, Cohen SM: microRNA miR-14 acts to modulate a positive autoregulatory loop controlling steroid hormone signaling in Drosophila. *Genes & development* 2007, 21(18):2277-2282.

296. Almondes KG, Leal GV, Cozzolino SM, Philippi ST, Rondo PH: The role of selenoproteins in cancer. *Rev Assoc Med Bras* 2010, 56(4):484-488.

297. Hah N, Danko CG, Core L, Waterfall JJ, Siepel A, Lis JT, Kraus WL: A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. *Cell* 2011, 145(4):622-634.

298. Hatfield DL, Gladyshev VN: How selenium has altered our understanding of the genetic code. *Molecular and cellular biology* 2002, 22(11):3565-3576.

299. Roesijadi G, Rezvankhah S, Binninger DM, Weissbach H: Ecdysone induction of MsrA protects against oxidative stress in Drosophila. *Biochemical and biophysical research communications* 2007, 354(2):511-516.

300. Panee J, Stoytcheva ZR, Liu W, Berry MJ: Selenoprotein H is a redox-sensing high mobility group family DNA-binding protein that up-regulates genes involved in glutathione synthesis and phase II detoxification. *The Journal of biological chemistry* 2007, 282(33):23759-23765.

301. Quinlan AR: BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al]* 2014, 47:11 12 11-11 12 34.

302. Wang L, Feng Z, Wang X, Zhang X: DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* 2010, 26(1):136-138.

303. Donovan LE, Dammer EB, Duong DM, Hanfelt JJ, Levey AI, Seyfried NT, Lah JJ: Exploring the potential of the platelet membrane proteome as a source of peripheral biomarkers for Alzheimer's disease. *Alzheimer's research & therapy* 2013, 5(3):32.

304. Van Bortle K, Corces V: tDNA insulators and the emerging role of TFIIIC in genome organization. *Transcription* 2012, 3(6).

305. Phizicky EM, Hopper AK: tRNA biology charges to the front. *Genes & development* 2010, 24(17):1832-1860.