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:

_____ _____

Jingping Yang Date

The Role of Insulators and Transcription Factors in Genome Organization and Function in *Drosophila*

By

Jingping Yang
Doctor of Philosophy

Graduate Division of Biological and Biomedical Sciences
Population Biology, Ecology, and Evolution

_____
Victor G. Corces, Ph.D.
Advisor

_____ _____
William G. Kelly, Ph.D. John Lucchesi, Ph.D.
Committee Member Committee Member

_____ _____
James Taylor, Ph.D. Michael E. Zwick, Ph.D.
Committee Member Committee Member

Accepted:

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

_____
Date

The Role of Insulators and Transcription Factors in Genome Organization and Function in *Drosophila*

By

Jingping Yang
M.S., Nanjing University, 2007
B.S., Nanjing University, 2004

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 Sciences
Population Biology, Ecology, and Evolution

2012

Abstract

The Role of Insulators and Transcription Factors in Genome Organization and Function in *Drosophila*

By Jingping Yang

Epigenetic changes can alter the genome function without altering their base composition. These differences can be inherited and can provide an important source of variation within populations that can be acted upon by natural selection. Epigenetic changes in gene expression can take place via covalent modifications of histone or DNA as well as the three-dimensional organization of chromatin in the nucleus. Insulators mediate chromatin interactions in *cis* or *trans* between different regions of the genome and may be important factors regulating the 3D organization of the genome. BEAF-32 is an insulator protein highly conserved in *Drosophila* but not found in other species. Here I describe an analysis of the epigenetic function of BEAF-32 in *Drosophila*. I identify the BEAF-32 insulator as a *cis* regulatory element separating genes arranged in a head-to-head orientation. I then compare the genome-wide binding landscapes of the BEAF-32 in four different *Drosophila* species and highlight the evolutionarily conserved presence of this protein between close adjacent genes. During the formation of new *Drosophila* species, binding of BEAF-32 in the genome is altered along with changes in genome organization caused by DNA re-arrangements. The alterations of BEAF-32 distribution correlate with new gene expression profiles, which in turn translate into specific and distinct phenotypes. Epigenetic information encoded in the 3D organization of the genome mediated by insulators needs to be faithfully transmitted through mitosis and meiosis in order to effect evolutionary change. To address this issue, I have also studied the function of the Myc transcription factor. I found that a subset of Myc sites remain on mitotic chromatin and overlap with aligned insulator proteins binding sites. These sites are enriched at the boundaries of topological chromosome domains, suggesting they may be important for maintaining chromosome structure throughout the cell cycle. Together, these results suggest a mechanism for the establishment of differences in transcription patterns during evolution and may help to decipher the role of epigenetic changes in evolution.

The Role of Insulators and Transcription Factors in Genome Organization and Function in *Drosophila*

By

Jingping Yang
M.S., Nanjing University, 2007
B.S., Nanjing University, 2004

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 Sciences
Population Biology, Ecology, and Evolution

2012

Acknowledgements

I would like to thank my advisor Victor G. Corces for years of endless support, intellectual stimulation and inspiration. I am especially grateful for the right amount of independence he provided for me to grow scientifically and never failing guidance whenever needed.

I would also like to thank my committee members John Lucchesi, William G. Kelly, James Taylor, and Michael E. Zwick for their insight and suggestions.

I would like to acknowledge all the members of the Corces lab past and present especially Eddie Ramos, Naomi Takenaka and Chunhui Hou for years of scientific discussions, encouragement, and friendship.

Finally, I appreciate for having my friends and family who have supported me every step of the way. Very special thanks to my parents for their lifetime of unconditional love and for providing me with a constant source of relief on demand, to my boyfriend for his positive humor and being so reliable there to help me keeping in perspective. I would not be where I am without all of you.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The genome of eukaryotic organisms is packed in the nucleus in a dynamic, yet non-random structure. The DNA is wrapped around nucleosomes to form the 10 nm fiber. Recent results suggest that, contrary to what was previously thought, the 10 nm fiber does not further condense into higher-order structures such as the 30 nm fiber (Fussner et al. 2012). The specific structure of chromatin allows for the replication and transcription of the genome, as well as the safe transmission of the genetic information from mother to daughter cells. Changes in chromatin status are carried out through signals or switches in DNA methylation, histone modifications, histone variants, and non-histone chromatin proteins by altering the local chromatin structure. Some of these changes can be transmitted from mother to daughter cells or even potentially from one generation to the next. Epigenetics is the study of these heritable changes on chromatin which alter genome function without changes to the DNA sequence (Probst et al. 2009).

Evolutionary biology explores the origins and patterns of diversity within populations, the origination of new species, and the divergence between species over long periods of time. Epigenetic changes that contribute to phenotypic variation without altering DNA sequence may represent an important aspect of evolutionary biology, since not all diversity can be explained by differences in base composition. For example, the common European honeybee adopts very specific social behaviors. In the population, there are a few queens overseeing many workers. The honeybee queen and worker differ substantially in morphology, physiology, behavior, and reproductive potential, but they have exactly the same DNA sequence. Although their DNA sequence is identical, the queen honeybees are fed a chemically different diet than the worker bees. Dietary

differences result in significant changes in the methylation of DNA in genes involved in metabolism and RNA synthesis (Lyko et al. 2010). When DNA methyltransferase Dnmt3 levels were significantly reduced in newly hatched honeybee larvae, the relative abundance of queens and workers changed. Reduced Dnmt3 levels also changed DNA methylation patterns at promoters of genes, especially those involved in developmental processes and specific gene expression patterns (Kucharski et al. 2008). The epigenetic marks that bridge diet and genome function make it possible for this species to develop its social organization.

Epigentic mechanisms also play a role in dosage compenstation, a trait that clearly evolves over time and varies between species. Species have developed various epigenetic mechanisms for dosage compensation (Lucchesi et al. 2005; Chess 2012; Conrad and Akhtar 2012). Epigenetic regulations enable the evolution of sex chromosomes. Without changing the DNA sequence, epigenetic phenomena may determine when, where, and how the genetic information should be read and interpreted. Epigenetic changes can rapidly alter the behavior of genomes. Contrary to genetic changes that are highly stable, epigenetic processes have a certain level of plasticity that is inherently reversible. Epigenetic changes can occur in response to environmental like temperature and diet. Then the variations in traits within populations can be acted upon by natural selection to shape evolution.

Changes in epigenetic regulation can take place during the formation of the 10 nm chromatin fiber by different mechanisms such as covalent histone modifications, use of different histone variants, DNA methylation, and alterations induced by ATP-dependent

remodeling complexes. Histone acetylation can directly loosen histone-DNA interactions through charge neutralization to allow binding of other proteins (Bannister and Kouzarides 2011). Acetylation and other histone modification such as methylation or phosphorylation, as well as DNA methylation, can serve as a platform for recruiting other chromatin associated factors, including chromatin-remodeling factors, and factors involved in transcription and replication (Bannister and Kouzarides 2011; Jones 2012). Chromatin-associated factors can also change chromatin modifications. For example, the chromatin-associated factor Males absent On the First (MOF) in the Male-Specific Lethal (MSL) complex induces histone H4 lysine 16 acetylation (H4K16ac) involved in dosage compensation in *Drosophila*. The increase of MOF on chromatin and the resulting histone modifications are required for the increase recruitment of RNA polymerase II (RNAPII) at promoters of X-linked genes (Conrad et al. 2012a; Conrad et al. 2012b).

Genome-wide distribution of histone modifications, non-histone chromatin proteins, Deoxyribonuclease I (DNase I) hypersensitivity sites, as well as Global Run-On sequencing (GRO-seq) reads, revealed chromatin landscapes that characterize different regions of the *Drosophila* genome (Filion et al. 2010; Kharchenko et al. 2011). The combination of chromatin features indicates the function and status of specific regions of the chromosome. For example, regions containing Polycomb (Pc) and histone H3 lysine 27 trimethylation (H3K27me3) and regions containing heterochromatin protein 1 (HP1) and histone H3 lysine 9 dimethylation (H3K9me2) are silenced whereas regions containing histone H3 lysine 4 trimethylation (H3K4me3) are transcribed (Filion et al. 2010; Kharchenko et al. 2011). In the *Drosophila* genome, the domains marked by

histone modifications range from 1 to 52 kb with a median size of 6.5 kb. There are 441 domains larger than 50 kb and 155 domains larger than 100 kb, with the largest domain at 737 kb (Filion et al. 2010). These results suggest that the genome is organized into regions.

Regulatory mechanisms involved in proper expression of the genome required interactions between distant sequences and their bound proteins. These three-dimensional (3D) long-range chromatin interactions are necessary for nuclear processes such as transcription, DNA replication, recombination and DNA repair. For example, Polycomb group (PcG) proteins bound to chromatin tend to cluster in the nucleus to form Pc bodies. These interactions are required for proper silencing of the genes bound by Pc (Bantignies et al. 2011). Similarly to gene silencing, activation of transcription requires interactions between regulatory sequences such as enhancers and promoters. These and other interactions can be detected with Chromosome Conformation Capture (3C) and related techniques. Results from this type of experiments suggest that chromosomes are divided in topological domains, characterized by a high number of interactions, separated by boundaries across which interactions take place infrequently. This topological domain organization has been observed in *Drosophila*, mice and human (Dixon et al. 2012; Hou et al. ; Sexton et al. 2012). Genes in each domain are coordinately regulated during development. Deletion of a boundary between domains leads to ectopic interactions between domains and misregulation of gene expression (Nora et al. 2012).

Recent results suggest that boundaries between topological domains form at regions of high gene density containing actively transcribed genes and clusters of

insulator proteins named “aligned insulators”. Insulators are DNA-bound protein complexes that can mediate interactions in *cis* or *trans* between different regions of the genome (Phillips and Corces 2009) and are good candidates for the factors regulating the 3D arrangement of the chromatin fiber in the nucleus. Insulators have now been found in most eukaryotes, from yeast to humans. They were originally described in *Drosophila* and this organism remains noteworthy for the number of characterized insulators. *Drosophila* insulators share two proteins, modifier of mdg4 (Mod(mdg4)) and Centrosomal protein 190kD (CP190), which interact with a variety of DNA binding proteins named Suppressor of Hairy wing (Su(Hw)), CCCTC-binding Factor (CTCF) and Boundary element-associated factor of 32kD (BEAF-32) whose function appears to be limited to the recruitment of the shared components to different genomic locations, where they may play distinct roles (Bushey et al. 2009). Insulators in *S. cerevisiae* and *S. pombe* are mostly limited to RNA polymerase III (RNAPIII) promoter sequences containing binding sites for Transcription factor IIIC (TFIIIC) (Noma et al. 2006; Valenzuela et al. 2009; Iwasaki et al. 2010). In vertebrates, the most widely studied insulator is CTCF, which requires cohesin for functionality and also associates with other co-factors, although their general requirement for CTCF function is not clear (Parelho et al. 2008; Rubio et al. 2008; Wendt et al. 2008). SINE elements and their associated RNAPIII promoters have been also shown to act as insulators but the relevance of this initial observation has not been pursued in detail (Lunyak et al. 2007). New results suggest that RNAPIII promoters in human tDNA genes can act as both enhancer-blocking and barrier insulators (Raab et al. 2011). It is not clear at this point whether this property can be exclusively assigned to the presence of TFIIIC. Insulator function of these sequences

correlates with their ability to mediate interactions with other tDNAs in the genome. One possible interpretation of these results is that these interactions mediate the formation of RNAPIII transcription factories. Interestingly, TFIIIC sites are often found adjacent to CTCF sites in the genome, suggesting that the two proteins may function together to establish long-range interactions (Carrière et al. 2011).

In *Drosophila*, immunofluorescence microscopy using antibodies against Su(Hw), Mod(mdg4), CTCF and CP190 shows the presence of these proteins in a punctuated pattern. These structures, called insulator bodies, are preferentially located around the nuclear periphery, and it has been suggested that they represent sites where several individual insulator sites coalesce to mediate intra- and/or inter-chromosomal interactions. The morphology of the insulator bodies is disrupted by mutations in lamin, the main component of the nuclear lamina, and various insulator components (Gerasimova et al. 2000; Pai et al. 2004; Gerasimova et al. 2007). The fact that different DNA-binding insulator proteins colocalize at insulator bodies suggests that the various *Drosophila* insulators are able to interact with each other. In support of this conclusion, chromatin immunoprecipitation with microarray (ChIP-chip) analyses indicate that Su(Hw), BEAF and CTCF overlap at 9-24% of sites where only the DNA consensus sequence for one of the proteins is present, suggesting interactions between two or more different insulators at these sites (Bushey et al. 2009). Interactions between insulator sites have been visualized by fluorescence in situ hybridization (FISH), showing that two individual Su(Hw) insulator sites come together to form a loop. Insertion of an additional insulator between the original two Su(Hw) sites leads to the formation of two smaller loops (Byrd and

Corces 2003). Using 3C (Dekker et al. 2002), it has been shown that a *Drosophila* insulator containing the CTCF and CP190 proteins is induced at the *Eip75B* gene after cells are treated with the steroid hormone ecdysone. This insulator prevents an ecdysone enhancer from activating transcription of genes that are not regulated by this hormone (Wood et al. 2011). 3C has also been used to show that two Su(Hw) insulators can interact and loop out the intervening sequences to bring an upstream Polycomb response element (PRE) close to a downstream promoter. When one insulator is deleted, the PRE cannot associate with the promoter and H3K27me3 present at the PRE is lost at the promoter region after the interaction between the two sequences is disrupted (Comet et al. 2011). The Fab7, Fab8 or Mcp insulators, which use CTCF as the DNA binding protein, have also been found to mediate intra-chromosomal interactions. The Fab7 and Mcp insulators target the *abd-B* and *Antp* genes, which are located approximately 10 Mb apart in chromosome 3R. These two loci colocalize in nuclei of cells in which both genes are repressed, but deletion of Fab7 or Mcp results in a reduction of the interaction and colocalization, suggesting an important role of these two insulators in the interaction (Bantignies et al. 2011). The Fab7 and Fab8 insulators have also been shown to interact with a CTCF site located in the *abd-B* promoter region by testing the expression of a reporter (Kyrchanova et al. 2008). Consistent with this result, the *abd-B* promoter and the Mcp, Fab7 and Fab8 elements have been found to cluster in the S2 cells or fly head tissue, where *abd-B* is repressed, but not in tissues where *abd-B* is expressed (Cleard et al. 2006; Lanzuolo et al. 2007). Other work has shown that the insulator sequences present in these regulatory elements, rather than other potential regulatory elements such as PREs, are responsible for these interactions (Xiao et al. 2011). These results suggest a general role

for insulators in mediating intra-chromosomal interactions in order to modulate different transcriptional regulatory processes.

Recently, in mouse ES cells Chromatin Interaction Analysis by Paired-End Tag (ChIA-PET) identified 1480 *cis* interactions mediated by CTCF genome wide (Handoko et al. 2011). Clustering of histone modifications in/around the chromatin loops created by these interactions classified 5 types of distinct patterns. Category I loops contain active chromatin marks such as histone H3 lysine 4 monomethylation (H3K4me1), histone H3 lysine 4 dimethylation (H3K4me2) and histone H3 lysine 36 trimethylation (H3K36me3), whereas repressive marks like methylation at histone H3 lysine 9 (H3K9), histone H3 lysine 20 (H3K20) or histone H3 lysine 27 (H3K27) are depleted. Category II is opposite to the first class with extensive methylated H3K9, H3K20 and H3K27 in the loops, which indicate repressive domains. The two types of interactions may create independent domains for differential regulation of gene expression. Category III loops are suggested as hubs for enhancer and promoter activities. These loops are enriched in H3K4me1 and H3K4me2 at enhancers inside the loop and H3K4me3 at promoter at the end of the loops. These interactions could bring enhancers closer to the target promoters. For genes and their distal enhancers, if they fall into category III loops, the frequency of upregulated genes in Embryonic stem (ES) cells compared to Neural stem (NS) cells is significantly higher than the ones not encompassed by the loops. Category IV loops show opposite chromatin states flanking the ends of the loops, but do not exhibit any specific pattern of histone modifications inside the loops. Category V loops do not show any specific signatures. The function of the last two types of loops is not yet clear.

Interactions mediated by CTCF have been shown to be important for genome function at various loci and for different cellular processes. These interactions can mediate enhancer promoter contacts and direct spatial and temporal promoter selection. Targeting the right promoter is an important pre-requisite for regulatory elements to guide proper transcription. For pairs of regulatory element and single target promoters, the chromosome interactions mediated by insulators have been shown to help promoter targeting by either blocking or facilitating communication between the regulatory elements and the promoter (Hadjur et al. 2009; Comet et al. 2011). Insulators accomplish this function by changing the spatial proximity of enhancers and promoters. In addition to the on/off decision for a single promoter, insulators also play a role in promoter selection for regulatory elements shared by more than one competent target promoter. The three dimensional structure established by insulator-mediated interactions may orchestrate on/off effects for each promoter to contact the right target and/or shift the balance of promoter competition by increasing the targeting potential of certain promoter. The resulting selection of specific promoters may coordinate regional gene expression or even switch transcription programs important for specific differentiation outcomes. One interesting example is the use of CTCF for promoter selection and control of the latency cycle in Epstein-Barr virus. In the virus genome, the enhancer OriP is shared by two downstream promoters, the C promoter (Cp) and the Q promoter (Qp). Cp determines the type III gene expression pattern for latency cycle III and Qp determines the type I expression program for latency cycle I (Chau et al. 2006; Tempera et al. 2010; Tempera et al. 2011). CTCF binds upstream both of Cp and Qp. The default selection for OriP is the proximate promoter Cp. When Cp is active, depletion of CTCF do not affect the

utilization of Cp (Chau et al. 2006; Tempera et al. 2010; Tempera et al. 2011). However, when CTCF dependent interactions bring distal Qp close to OriP, Cp is turned off. At this stage, mutations in the either of the CTCF binding sites disrupt the interaction between OriP and Qp and lead to the reactivation of Cp transcription (Tempera et al. 2011). Although Cp is silent in wild type and active in virus carrying a deletion of the CTCF binding site at Qp, it is also observed that the proximity of Cp to the enhancer is not changed (Tempera et al. 2011). Thus, insulator mediated chromatin 3D structure contributes to the promoter selection probably by increasing the competence of Qp to compete out Cp. When Qp is not competent, Cp is selected as default. It is not known yet whether the Qp is necessary in addition to the 3D structure to silence Cp. The same mechanism may also work in other systems with shared regulator elements and alternatively active promoter like imprinted genes. It is known that CTCF at the imprinting control region (ICR) of the *Igf2/H19* locus determines different chromatin conformations and controls the expression pattern of the two genes in mammals (Kurukuti et al. 2006; Nativio et al. 2009). Thus, for both single promoter or multi promoter systems, insulator interactions facilitate the selection of the appropriate promoter by regulatory elements.

Insulator mediated interactions can also play a role in the regulation of gene networks. In contrast to alternative expression, certain genes need to be controlled by the same regulatory sequences and long-distance physical interactions may be required to accomplish this goal. These interactions have been found to be mediated by insulators. For example, the Major Histocompatibility Class II (MHC-II) genes *HLA-DRB1* and

*HLA-DQA1*are recruited together by CTCF under the induction of cytokine when both genes are activated (Majumder et al. 2008). CTCF also mediates interactions between enhancer sequences present in the *Insulin* (*INS*) and the promoter of the *Synaptotagmin VIII (SYT8)* genes, which is located over 300 kb away. *SYT8* is required for insulin secretion in islets (Xu et al. 2011).

Loops formed by contacts between insulator sites cannot only determine interactions between regulatory sequences but they can also establish distinct domains of chromatin structure defined by the presence of specific histone modifications. Genome-wide studies of CTCF distribution have uncovered a subset of CTCF sites localized at boundaries between active and repressive chromatin domains marked by histone H2A lysine 5 acetylation (H2AK5Ac) and H3K27me3, respectively (Cuddapah et al. 2009). These regions are different between HeLa and CD4+ T cells, suggesting a possible role for CTCF-delimited domains in establishing lineage-specific gene regulation. The functional significance of the chromatin domains demarcated by CTCF is beginning to emerge from studies of specific loci. In human lung fibroblasts IMR90 cells the *HOXA9-13* genes are silenced but other genes in the *HOXA* locus are transcriptionally active (Kim et al. 2011). Consistent with this transcription pattern, the region containing *HOXA9-13* is rich in H3K27me3 while the adjacent region containing the active *HOXA1-7* genes is marked by H3K4me2, H3K4me3 and H3K36me3. The distribution of the repressive marks depends on the interaction between two CTCF insulators flanking *HOXA9-13*. Knockdown of CTCF results in the disruption of loop formation and spreading of silencing histone modifications, which causes downregulation of the adjacent *HOXA6-7*

genes (Kim et al. 2011). A second example of CTCF-delimited chromatin domains in the control of gene expression and the establishment of cell fates is that of the Wilm's tumor protein (Wt1) in the regulation of epithelial-mesenchymal transitions during the development of the epicardium and the kidney. These processes are controlled by Wt1 through the regulation of the expression of the *Wnt4* gene (Essafi et al. 2011). Wt1 binds to the promoter of *Wnt4* in both tissues with opposite outcomes. In kidney cells, Wt1 recruits the CREB-binding protein (CBP)/p300 coactivators and turns on transcription of *Wnt4*. In epicardial cells, Wt1 recruits the Brain acid soluble protein 1 (Basp1) co-repressor to silence *Wnt4*. These Wt1-induced changes in *Wnt4* expression correlate with changes in chromatin structure. Kidney cells, in which the *Wnt4* gene is active, contain H3K4me3, H3K9ac and histone H3 lysine 14 acetylation (H3K14ac) in the *Wnt4* locus. Epicardial cells, where *Wnt4* is silenced, contain H3K27me3 instead. The Wt1-dependent changes of chromatin marks in the *Wnt4* locus are confined to a region of the genome that is delimited by CTCF. The formation of a loop enclosing the *Wnt4* locus may be the basis for the functional chromatin domain established by Wt1. In the absence of CTCF, the chromatin domain created by Wt1 spreads outside of its normal boundaries and alters the transcription of neighboring genes (Essafi et al. 2011).

It is becoming increasingly clear that the role of insulators is to mediate inter- and intra-chromosomal interactions between different regions of the genome. In so doing, insulators can establish a specific 3D organization of the genome within the nucleus of eukaryotic cells. In the cases described above, insulators mediate intra- and inter-chromosomal interactions in order to elicit a specific transcriptional response during

interphase, when cells are transcriptionally active. It is possible that patterns of 3D architecture of the chromatin fiber established by insulators are cell-type specific and responsible for distinct blueprints of gene expression during development. If this is the case, this organization of the chromatin may need to be preserved during mitosis. During mitosis, the compaction of the chromatin fiber is dramatically altered, and it is possible that insulators are involved in regulating this re-arrangement of the 3D organization of the DNA. Examination of the inheritance of insulator proteins and the 3D architecture mediated by insulators remains one of the main challenges for research in the field. Most transcription factors fall off the chromatin during mitosis but a few selective ones remain. The role of these proteins during mitosis is unclear but it has been hypothesized that they remain on chromosomes at specific sites in order to allow early transcription of specific genes at the M/G1 boundary. Two transcription factors that have been shown to persist in mitotic chromatin are Bromodomain-containing protein 4 (Brd4) and Myc.

Myc is a sequence specific DNA binding protein (Blackwell et al. 1990) that can bind to both the canonical (CACGTG or CATGTG) and non-canonical CA--TG E box sequences (Blackwell et al. 1993). It has been suggested as a transcription factor that regulates the expression of genes (O'Donnell et al. 2005; Aguda et al. 2008; Chang et al. 2008; Lovén et al. 2010) and plays critical roles in cancer initiation and metastasis of may different types of tumors (Wolfer and Ramaswamy 2011; Dang 2012). Genomic searches for Myc target genes have uncovered a role for this protein in the regulation of hundreds of genes involved in cell cycle progression, differentiation, apoptosis, DNA repair, angiogenesis, chromosome instability, and ribosome biogenesis(Meyer and Penn 2008; van Riggelen et al. 2010; Dang 2012). In addition to its association with genes Myc is

also found in non-promoter sequences. These non-promoter Myc sites may function as transcriptional regulatory elements, such as enhancers, but their role has not been studied in detail. Myc controls a variety of cellular processes required for cell differentiation and is essential for cellular reprogramming to induce pluripotency and stem cell renewal (Smith et al. 2010; Varlakhanova et al. 2010; Moumen et al. 2012). The role of Myc in these cellular processes may be a consequence of its effects on gene expression at the local level but other evidence suggests that Myc can also affect chromatin more globally (Varlakhanova and Knoepfler 2009). It is possible that those Myc sites at non-promoter regions play roles in regulating chromatin status. An important property of Myc that has been largely ignored when considering potential mechanisms by which this protein can affect gene expression is that it reminds bound to DNA during mitosis (O'Donovan et al. 2010; Ohta et al. 2010), raising the question of whether some of the functions ascribed to this protein are actually a consequence of its presence in mitotic chromosomes.

In this thesis, I investigate the relationship between insulators and gene expression patterns in the context of genome organization in *Drosophila*. There are several types of insulators in *Drosophila* that have been studied in detail. BEAF-32 is one of the insulator proteins that is restricted to *Drosophila* species (Schoborg and Labrador 2010). Analysis of the genomic localization of BEAF-32 in different *Drosophila* species may enable us to analyze the possible role of epigenetic processes in evolution. In addition, analysis of the relationship between Myc and insulator proteins in mitosis may further our understanding of how epigenetic information is transmitted throughout the cell cycle.

Chapter 2

The BEAF-32 insulator coordinates genome organization and function during the evolution of *Drosophila* species

This manuscript has been published on Genome Research: Yang J, Ramos E, Corces VG. (2012) The BEAF-32 insulator coordinates genome organization and function during the evolution of Drosophila species. Genome Research 22(11): 2199-2207.

Abstract

Understanding the relationship between genome organization and expression is central to understanding genome function. Closely apposed genes in a head-to-head orientation share the same upstream region and are likely to be co-regulated. Here we identify the *Drosophila* BEAF-32 insulator as a *cis* regulatory element separating close head-to-head genes with different transcription regulation modes. We then compare the binding landscapes of the BEAF-32 insulator protein in four different *Drosophila* genomes and highlight the evolutionarily conserved presence of this protein between close adjacent genes. During the formation of new *Drosophila* species, binding of BEAF-32 to sites in the genome is altered along with changes in genome organization caused by DNA re-arrangements or genome size expansion. The cross talk between BEAF-32 genomic distribution and genome organization contributes to new gene expression profiles, which in turn translate into specific and distinct phenotypes. The results suggest a mechanism for the establishment of differences in transcription patterns during evolution.

Introduction

Eukaryotic genomes are not organized randomly. Rather, genes and their regulatory elements are arranged in a manner that allows for the correct function of the genome. Genes that show similar function or expression tend to be clustered on the chromosomes (Kamath et al. 2003; Pal and Hurst 2003; Hurst et al. 2004; Batada and Hurst 2007), but not all adjacent genes are co-regulated. One interesting feature of eukaryotic genomes is the head-to-head juxtaposition of genes with two adjacent transcription start sites (TSS). Approximately 10% of genes in vertebrates are arranged in a head-to-head orientation and located closer than 1 kb from each other (Adachi and Lieber 2002; Yang et al. 2008). The proportion of close head-to-head gene pairs in the genome correlates with gene density (Li et al. 2006; Yang and Yu 2009), and *Drosophila* shows a higher than expected proportion of genes in this type of arrangement (Koyanagi et al. 2005; Yang and Yu 2009). The intergenic regions of close head-to-head gene pairs are referred to as bidirectional promoters, indicating the two TSSs are close enough to share the same upstream regulatory region (Adachi and Lieber 2002; Koyanagi et al. 2005). Genes positioned in a head-to-head orientation show overall higher correlation of expression than those arranged in other orientations (Herr and Harris 2004; Yang and Yu 2009). However, there are also head-to-head gene pairs whose expression is not correlated or is negatively correlated in both humans and *Drosophila* (Herr and Harris 2004; Li et al. 2006). Interestingly, close head-to-head gene pairs in *Drosophila* species tend to have higher rearrangement rates during evolution (Weber and Hurst 2011), suggesting they are not constrained in their genomic location and they do not share

common regulatory sequences. *Drosophila* must then possess mechanisms to functionally separate closely apposed genes in a head-to-head orientation in order for these genes to be independently regulated.

Insulators have been shown to contribute to the establishment of specific patterns of chromatin organization important for regulation of transcription by, at least in part, regulating interactions between enhancers and promoters (Phillips and Corces 2009; Handoko et al. 2011; Yang and Corces 2011). In *Drosophila* there are several types of insulators differentially distributed throughout the genome in a manner suggestive of distinct functions in gene expression (Bushey et al. 2009; Nègre et al. 2010). BEAF-32 is the DNA-binding protein for one of these insulators with a role in the recruitment of other components to specific sites in the genome. In *D. melanogaster*, BEAF-32 associates preferentially with actively transcribed genes, although the specific mechanism by which it affects gene expression is not known (Bushey et al. 2009; Jiang et al. 2009). Here we identify the BEAF-32 insulator as a *cis* element located between head-to-head genes to attain differential regulation of transcription. Changes in *cis* regulatory sequences represent an important source of variability necessary for divergence between species (Borneman et al. 2007; Odom et al. 2007; Schmidt et al. 2010). A large number of chromosome rearrangements have taken place during *Drosophila* speciation that have resulted in changes in the location of genes in the genome (Consortium 2007). Given the presence of the BEAF-32 insulator between close head-to-head gene pairs, we mapped the binding profiles of BEAF-32 in different *Drosophila* species and investigated changes in the pattern of BEAF-32 localization during the evolution of *Drosophila* species.

Comparison between changes in BEAF-32 insulator distribution and gene location in different *Drosophila* species enabled us to establish correlations between changes in genome organization and function.

Results

BEAF-32 specifically associates with close head-to-head gene pairs

We first used the latest annotation of the *D. melanogaster* genome to examine the frequency of gene pairs. We found that 28% of genes are in a head-to-head orientation with intergenic regions shorter than 1 kb. This fraction is much higher than that found in other eukaryotes, including other insect species, in which the proportion of close head-to-head gene pairs ranges between 8% and 18% (Figure 2-1A, Table 2-1) (Li et al. 2006; Dhadi et al. 2009). It is unlikely that such large numbers of genes are co-regulated in *Drosophila* but not in other species, suggesting the existence of *Drosophila*-specific mechanisms to maintain independent regulation of close head-to-head gene pairs. Insulators are good candidates to perform such function given their ability to regulate enhancer-promoter interactions. More specifically, the BEAF-32 insulator protein is highly conserved in *Drosophila* and its presence appears to be restricted to this genus (Schoborg and Labrador 2010). We therefore examined the genome-wide distribution of BEAF-32 in *D. melanogaster* embryos using ChIP-seq, and found BEAF-32 frequently located between close adjacent genes oriented head-to-head (Figure 2-1B). This is consistent with previous reports suggesting that about 50% of BEAF-32-associated genes are arranged in a head-to-head orientation (Jiang et al. 2009). Based on the genomic distribution of BEAF-32 relative to genes, 50% is significantly greater than expected ($p<1x10^{-4}$) (Figure 2-1C, Figure 2-2). This enrichment is unique to BEAF-32 but not to transcription factors, factors for general transcription, other promoter-associated factors or other insulator proteins (Figure 2-1C and Figure 2-3).

One consequence of the arrangement of genes in pairs in a head-to-head orientation is shorter distances between the TSSs compared to other possible orientations (head-to-tail, tail-to-tail, or tail-to-head) (Figure 2-4A-B). We thus examined the distance between TSSs flanking BEAF-32 binding sites in *D. melanogaster* and confirmed that BEAF-32-associated TSSs have a close neighboring TSS. The distance between the two TSSs peaked at 300-400 bp (Figure 1D and Figure 2-4C). A total of 66% (1042/1563) of close head-to-head gene pairs (distance <500 bp) contain BEAF-32 binding sites while only 36% (506/1400) distant head-to-head gene pairs (distance>1 kb) contain BEAF-32 binding sites between the two genes ($p<1x10^{-4}$). Thus, BEAF-32 preferentially associates with close head-to-head gene pairs.

BEAF-32-associated close head-to-head gene pairs are not co-expressed

Close head-to-head genes tend to be co-regulated in *D. melanogaster* compared to distant head-to-head genes or genes not in a head-to-head orientation, as there is a higher proportion of co-expression for close head-to-head gene pairs (Figure 2-5A and Figure 2-6). However, the correlation in expression for the two genes in close head-to-head gene pairs is spread over a broad range. In addition to a peak of high correlation, the distribution also shows a second peak at a value indicative of no correlation (Figure 2-5A). Therefore, there are close head-to-head gene pairs whose expressions are not correlated or are negatively correlated (correlation < 0.1). For these gene pairs, about 60% (150/251) have BEAF-32 binding sites between the genes. In contrast, less that 20%

(6/33) of highly co-expressed gene pairs (correlation > 0.9) have BEAF-32 ($p<4x10^{-5}$). Over 80% (27/33) of highly correlated head-to-head genes do not harbor BEAF-32 binding sites between them. Herr and colleagues have examined the co-expression of 8 head-to-head gene pairs spatially and temporally during embryonic stages of *Drosophila* development (Brogiolo et al. 2001; Renault et al. 2002; Herr et al. 2003; Herr et al. 2004; Herr and Harris 2004). For the two gene pairs found to be highly co-expressed, we examined the presence of BEAF-32 and found that there is no BEAF-32 binding signal between the genes. A similar analysis shows that BEAF-32 is present in the two gene pairs containing genes that are expressed differently (Figure 2-5C, Table 2-2). These results suggest a correlation between the presence of BEAF-32 between two close adjacent genes and their ability to be independently regulated.

BEAF-32 separates close head-to-head genes with different patterns of transcription regulation

To understand the mechanisms by which the presence of BEAF-32 allows genes to be differentially regulated, we compared the distribution of BEAF-32 binding sites with the mapped landscape of histone modifications in the *D. melanogaster* genome (Kharchenko et al. 2011; Negre et al. 2011). We aligned the map of BEAF-32 binding sites with histone modification data, both obtained in S2 cells (Figure 2-7A and Figure 2-8). Consistent with the association between BEAF-32 and active genes, we found that BEAF-32 clusters with active histone marks and not with repressive marks. Histone

marks for active TSSs, such as H3K4me3, are present on both sides of BEAF-32 binding sites between pairs of genes oriented head to head (Figure 2-7B-C). Interestingly, histone modifications such as histone H4 lysine 8 acetylation (H4K8ac), histone H3 lysine 18 acetylation (H3K18ac) and histone H3 lysine 27 acetylation (H3K27ac) are present at only one of the TSSs of the two genes in each pair, and the signal is reduced to background levels at the other TSS (Figure 2-7D-F). In *Drosophila*, H3K18ac and H3K27ac are thought to be produced by the acetyltransferase CBP, which is present at enhancers and promoters (Tie et al. 2009). The presence of these histone modifications adjacent to TSSs is suggestive of interactions between the enhancer and promoter that lead to activation of transcription. Thus, the asymmetric distribution of histone marks at the two TSSs suggests that BEAF-32 may separate two genes that are differentially transcribed, even though they share the same upstream region.

If this conclusion is true, changing the effect of putative regulatory sequences located in the intergenic region would only affect one of the genes but not the other. However, if the two genes in a pair are not separately regulated, they are likely to respond in the same way to changes in regulation. To test this hypothesis, we examined changes in the transcription profile resulting from mutations in SOX14, which is a *D. melanogaster* transcription factor (Ritter and Beckstead 2010). Among the 271 genes not associated with BEAF-32, 68 (25%) change their transcription in the same direction as their neighbor when SOX14 is mutant. However, only 4 out of 88 (4.5%) of BEAF-32-associated genes change simultaneously with their neighbor ($p<2x10^{-5}$), a five-fold difference with respect to genes not associated with BEAF-32 (Figure 2-7G).

Other histone modifications characteristic of transcription activation, such as histone H4 lysine 5 acetylation (H4K5ac), histone H4 lysine 8 acetylation (H4K8ac) and H4K16ac, are also distributed differently at the two sides of BEAF-32 binding sites (Figure 2-7A,D). H4K5ac has been reported as a histone modification present in genes bookmarked during mitosis (Zhao et al. 2011), and H4K16ac is the product of the acetyltransferase MOF, which functions at enhancers and promoters of X-linked and autosomal genes (Zippo et al. 2009). Both modifications are indicative of transcription activation, enforcing the conclusion that BEAF-32 is present between close head-to-head genes in small genomes, such as those of *Drosophila* species, to separate the TSSs of two different genes that need to be differentially regulated.

Conservation and diversity of BEAF-32 insulators across *Drosophila* species

Since BEAF-32 appears to functionally separate close head-to-head genes, gain or loss of BEAF-32 binding during evolution may prevent or allow adjacent genes to be affected by neighboring regulatory sequences, leading to changes in gene expression. In order to investigate the role of BEAF-32 during the evolution of *Drosophila* species, we systematically compared its binding site distribution in four *Drosophila* genomes, *D. melanogaster*, *D. simulans*, *D. pseudoobscura* and *D. virilis*. For the larger genome of species such as *D. virilis*, we sequenced twice the number of tags as in *D. melanogaster* to reach equal coverage for all the genomes studied (Table 2-3). Genome wide, BEAF-32 shows a similar binding distribution with respect to TSSs, gene bodies, and intergenic

regions across the four species, with a preference for sequences close to TSSs (Figure 2-9A). The association of BEAF-32 with head-to-head gene pairs is conserved in all four species (Figure 2-9B), suggesting a conserved function in *Drosophila*. The consensus motifs identified for BEAF-32 binding sites are virtually identical among the four species (Figure 2-9C), consistent with the protein conservation, particularly in the DNA binding domain.

Since BEAF-32 is significantly associated with gene pairs, in order to investigate changes in the profile of BEAF-32 binding in the genome of different *Drosophila* species we developed a gene-pair-centric analysis pipeline. To do so, we examined whether a BEAF-32 binding site present in the intergenic region of a gene pair in one species was also present in the corresponding intergenic region of the gene pair in the second species (see Material and Methods). With this pipeline, we pooled all the BEAF-32 binding sites from the four species and scored the presence of BEAF-32 at each site in each of the species. Using Cluster 3.0, we then clustered the pattern of BEAF-32 binding among the four species. The results of this clustering agree well with the evolutionary tree of these species (Figure 2-10A).

To quantitatively investigate differences in the distribution of BEAF-32 sites between species using this pipeline, we then analyzed the conservation of BEAF-32 binding between *D. melanogaster,* which has the best-annotated genome, and other species. In this *D. melanogaster*-centric comparison, we examined the occupancy of BEAF-32 on orthologous chromosomal regions between *D. melanogaster* and each of the three other species. The fraction of non-conserved binding sites ranges from 3% in *D.*

simulans to 29% in *D. virilis* (Figure 2-10B). The difference increases appropriately with the molecular distance between these genomes, suggesting the divergence follows the molecular clock. This relationship fits a simple linear regression, with an estimated divergence rate of BEAF-32 binding of 0.6% per Myr (R-squared>0.99) (Figure 2-9D). This estimate may be affected by the quality of the data, although ChIP-seq gives relatively low false positive or negative results. This divergence rate is higher than the non-synonymous nucleotide substitution rate of 0.4% per Myr, but lower than the synonymous nucleotide substitution rate of 6.34% per Myr (Consortium 2007), suggesting that the binding of BEAF-32 in the genome is under selection. We thus examined possible changes in the DNA sequence at BEAF-32 binding sites. Since the motif for BEAF-32 binding is conserved in the four *Drosophila* species (Figure 2-9C), we searched for the presence of this motif at the orthologous regions in their genomes. The results confirm changes in the DNA sequence consistent with the loss of the BEAF-32 binding motif specifically in the species where BEAF-32 binding is lost (Figure 2-9E-F). Thus, the function of BEAF-32 is conserved in *Drosophila* species, but gain or loss of specific binding sites is under selection during the evolution of these species.

Changes of BEAF-32 insulator localization correlate with alterations in genome organization during *Drosophila* evolution

Two obvious changes affecting *Drosophila* genomes during evolution are alterations in genome size and chromosome rearrangements. How does BEAF-32

contribute to the function of the genome after such changes? Since BEAF-32 is preferentially located between close divergently transcribed genes, BEAF-32 binding sites may change along with variations of distance between the genes. The four *Drosophila* species examined show differences in gene density across their genomes. Compared to *D. melanogaster*, the genome size of *D. virilis* is 46% larger and gene density decreases from 116 to 85 genes per Mb (Consortium 2007). At the same time, 32% of all gene pairs have BEAF-32 binding sites in *D. melanogaster* whereas the fraction is reduced to 15% in *D. virilis* (Figure 2-11A). For example, the intergenic region between the genes *myoglianin* and *eyeless* contains a functional BEAF-32 binding site in *D. melanogaster*(Sultana et al. 2011) but not in *D. virilis*. The distance between the two TSSs increased 10 times in *D. virilis*, and this change correlates with the loss of the BEAF-32 binding site in this species or the gain in *D. melanogaster* (Figure 2-11B). For *D. simulans* and *D. pseudoobscura*, the fraction of gene pairs remains around 28% and their gene density is similar to that of *D. melanogaster* (Figure 2-11A). Therefore, BEAF-32 may be recruited to intergenic regions between close TSSs when the distance between the two genes decreases or may be lost when the distance between genes increases.

When we examined the association between non-conserved BEAF-32 sites and chromosome rearrangements we found two types of non-conserved BEAF-32 binding sites. For the first type, the changes of BEAF-32 binding co-occur with chromosomal rearrangements, since the genes flanking these BEAF-32 binding sites have different neighbors in the two species. In this case, BEAF-32 binding is gained or lost when the arrangement between gene pairs is altered. There are 87%, 41% and 55% non-conserved

BEAF-32 binding sites at regions where genes have been rearranged in *D. simulans*, *D. pseudoobscura*, and *D. virilis*, respectively (Figure 2-11C). Most non-conserved binding sites in *D. simulans* are of the first type. For the second type of non-conserved BEAF-32 binding sites, gain/loss of binding sites does not associate with changes in chromosomal organization, as they are located at intergenic regions between the same gene pairs in the two species being compared. There are only 9 (13%) non-conserved binding sites of the second type in *D. simulans*. However, *D. pseudoobscura* and *D. virilis* show a higher frequency of changes in BEAF-32 binding not associated with rearrangements compared to *D. melanogaster*; the number of these events are 145 (59%) and 308 (45%), respectively (Figure 2-11C). Phenotypically, *D. simulans* looks more like *D. melanogaster*, while the other two species are more different. The results may suggest that the first type of non-conserved binding sites may help maintain proper expression patterns in newly rearranged genes, whereas the second type may result in alterations in the regulation of transcription of flanking genes that may contribute to phenotypic differences between the species.

Alterations in BEAF-32 insulator localization correlate with changes of genome function during *Drosophila* evolution

Genes flanking BEAF-32 sites are preferentially involved in metabolic processes that are also known to affect body size (Carreira et al. 2008; Bushey et al. 2009). Thus, it is possible that changes in gene expression arising as a consequence of the gain or loss of

BEAF-32 binding may affect body size, which is also one of the most obvious phenotypic differences among the *Drosophila* species studied. In a screen for mutations that alter body size in *D. melanogaster*, 26 mutations were identified containing P element insertions in intergenic regions (Carreira et al. 2008). We examined these mutations and found that 8 of them map to regions containing BEAF-32 binding sites. Out of these 8 regions, 6 (75%) show loss of BEAF-32 binding in *D. virilis* (Table 2-4, Figure 2-11D). These changes in BEAF-32 binding cannot be explained solely by the increase in genome size and distance between genes that took place in *D. virilis*, as the 75% difference is significantly higher than the overall genome-wide difference for BEAF-32 binding (29%) between the two species ($p<4x10^{-3}$). For these intergenic regions, the distance between genes has not changed appreciably, but BEAF-32 binding is lost in *D. virilis* compared to *D. melanogaster*. Gain or loss of BEAF-32 binding may alter the expression of one or more genes flanking BEAF-32 sites in this region which may lead to changes in body size.

Thus, during *Drosophila* evolution, BEAF-32 binding sites are gained or lost with or without change in gene location, to either maintain transcription or allow for diversity. After the genomic location of genes is altered, genes may be brought close to a new neighboring gene, and the proximity to new regulatory sequences in the adjacent gene may alter their expression pattern. The presence of BEAF-32 binding sites may permit the two new neighboring genes to maintain their original expression patterns (Figure 2-12). In addition, in the absence of chromosome rearrangements, alterations in BEAF-32

binding may result in changes in the expression profile of one or more genes, resulting in the appearance of new complex traits, such as those affecting body size (Figure 2-12).

Discussion

Here we show that the presence of BEAF-32 between close adjacent genes arranged in a head-to-head orientation correlates with different transcription regulatory patterns in the two genes of the pair in *Drosophila*. Close head-to-head gene pairs exist in almost all eukaryotes but it is not known whether other species also use this strategy in order to maintain independent regulation of adjacent genes. In humans, genes present in head-to-head gene pairs also show a bimodal distribution in the correlation of expression (Li et al. 2006). In addition to the peak indicative of high correlation, there is also a peak of enrichment of gene pairs whose expression is not correlated. For these pairs, it is reasonable to predict the existence of regulatory mechanisms that functionally separate the two genes in order to attain the observed differential transcription. BEAF-32 is restricted to *Drosophila* species (Schoborg and Labrador 2010) and mammalian cells may use other insulator proteins to accomplish this goal. In *Drosophila* there are several types of insulator elements that show different genomic distributions with respect to genes (Bushey et al. 2009; Nègre et al. 2010). The distribution of the dCTCF insulator partially overlaps that of BEAF-32. Since CTCF is conserved between *Drosophila* and humans (Moon et al. 2005; Schoborg and Labrador 2010), it is possible that this protein functionally replaces BEAF-32 in maintaining differential transcription programs in genes located in close head-to-head gene pairs. When the human genome was specifically examined for the organization of close head-to-head gene pairs, those containing CTCF showed lower correlation of expression, suggesting that this mechanism may be also conserved in humans (Xie et al. 2007).

The organization of the genome that provides the highest fitness should be selected during evolution. If co-expression of close head-to-head gene pairs provides lower fitness, selection should favor re-arrangements that result in physical or functional separation of the two genes. A comparative analysis of head-to-head gene pairs in different species revealed that these pairs are more conserved in vertebrate lineage than in *Drosophila* species (Yang and Yu 2009; Weber and Hurst 2011). *Drosophila* has more close head-to-head gene pairs than mammals but the conservation of these pairs is 3-fold lower (Yang and Yu 2009). This suggests that some of the head-to-head gene pairs in *Drosophila* arise from genome compaction rather than selection for this specific organization. For these gene pairs, maximum fitness will select for separation of the genes in order to attain differential expression of the two genes in the pair. One strategy to accomplish this is functional separation by recruiting insulator proteins. Alternatively, chromosomal re-arrangements may physically separate the two genes. However, in an already compact genome like that of *Drosophila*, it may be difficult to organize all non-co-expressed genes apart from each other. Thus, a strategy relying on functionally separating the members of head-to-head gene pairs may be more effective. Our analysis has concentrated on close adjacent genes that are divergently transcribed because this arrangement facilitates analysis of the correlation between the location of BEAF-32 and transcription patterns of the two genes. Nevertheless, 38% of BEAF-32 binding sites associate with non head-to-head gene pairs. It is possible that BEAF-32 plays a similar role in this situation in order to control interactions between regulatory sequences located in the 3' region or introns of genes and adjacent promoters from other genes. Although information on the location of regulatory sequences in the *Drosophila* genome is

becoming available, it is not yet known which sequences regulate which genes. In the absence of this information, it is not possible at this time to evaluate the possible role of BEAF-32 in maintaining independent regulation of genes that are far apart and not in a heat-to head orientation.

The organization of head-to-head gene pairs in both humans and *Drosophila* is conserved during evolution, but the two members of each pair are not precisely co-regulated. The distribution of expression correlation suggests that most gene pairs do not show either high correlation or no correlation, but rather a relative level of correlation (Figure2A) (Li et al. 2006), suggesting that they may be co-regulated in certain developmental stages or specific tissues. Co-expression is still important for the genes, but they are not co-regulated all the time. Thus, the head-to-head orientation needs to be maintained for co-expression, but it is also necessary to separate genes when they are not co-regulated. The profiles of genome distribution of different insulator proteins in different cell types suggest a certain degree of cell type specificity in both humans and *Drosophila* (Kim et al. 2007; Bushey et al. 2009; Cuddapah et al. 2009). These observations point to a role for insulators in coordinating genome organization and function during evolution.

Methods

Fly stocks and other reagents

Oregon R was used as the wild type strain for *D. melanogaster*. Strains for other species were obtained from the UC San Diego Drosophila Species Stock Center. Stock numbers are ID 14021-0251.195 for *D. simulans*, ID 14011-0121.94 for *D. pseudoobscura* and ID 15010-1051.87 for *D. virilis*. Flies were grown at 25 °C. BEAF-32B is the main BEAF-32 isoform and its sequence is highly conserved (Figure 2-13). BEAF-32B antibodies were generated against amino acids 1-83 of BEAF-32B in *D. melanogaster*(Bushey et al. 2009). The polyclonal antibody cross-react s with BEAF-32 orthologs in other *Drosophila* species and recognizes specific bands on polytene chromosome squashes from salivary glands of the species examined (Figure 2-14).

ChIP-seq

Chromatin IP was carried out using embryos. To match specific developmental stages for embryos from each species, we determined the collecting time based on the length of the life cycle of the various species (Markow and O'Grady 2005). The age of the embryos used for chromatin immunoprecipitation was 0-8 hr for *D. melanogaster*, 0-8 hr for *D. simulans*, 0-10 hr for *D. pseudoobscura* and 0-12 hr for *D. virilis*. ChIP was performed following published procedures (Sandmann et al. 2007) with the following adjustments. Two grams of embryos were used for two chromatin preparations and extracts were sonicated 20 cycles (10s on/30s off) on a Branson Sonifier 250 with output control set at

1.5. Libraries were prepared with the IlluminaTruSeq DNA Sample Preparation Kit and sequenced at the HudsonAlpha Institute for Biotechnology.

Sequence analysis

For analysis of sequence data, we used genome sequence and annotations released on Flybase, dmel_r5.39, dsim_r1.3, dpse_r2.22 and dvir_r1.2. Sequences were aligned to genomes using Bowtie with indexes built for each genome. The output map files were converted to bed format for each chromosome arm using the VancouverShort package. Only aligned reads on the main chromosomes were used to call peaks, as the small chromosome segments are not well annotated. The main chromosomes include *D. melanogaster*: chr2L, chr2LHet, chr2R, chr2RHet, chr3L, chr3LHet, chr3R, chr3RHet, chr4 and chrX; *D. simulans*: chr2L, chr2R, chr3L, chr3R, chr4, and chrX; *D. pseudoobscura*: chr2, chr3, chr4_group1-5, chrXL_group1a/1e/3a/3b, chrXR_group3a/5/6/8; *D. virilis*: scaffolds with more than 1000 reads. Peaks were called using CCAT3.0 (Xu et al. 2010). BEAF-32 associated genes were defined as genes closest to each peak and BEAF-32 associated pairs were defined as non-overlapping gene pairs flanking each peak. Both genes in a gene pair are defined as BEAF-32 associated genes if they flank BEAF-32 binding sites and are arranged in a head-to-head orientation. For other orientations, only the closest gene is defined as a BEAF-32 associated gene. Overlapping gene pairs were discarded. Associated genes or pairs were called using a custom script (available upon request). Only pairs with well-mapped intergenic regions and a gap of less than 10% of the length of the region or 300 bp were defined as well-mapped pairs.

Fraction of genes in head-to-head gene pairs in different species

Genome annotations for each species were downloaded from the UCSC genome browser. For genes with alternative transcripts, only the longest transcript was considered for analysis. We then created a list of all possible non-overlapping gene pairs and counted the number of unique genes in all gene pairs as A. We then selected the gene pairs in head-to-head combination and closer than 1 kb. We counted the number of unique genes in these head-to-head gene pairs as B. The fraction of genes in head-to-head gene pairs is B/A. For *D. melanogaster*, we used four different annotation versions. The flybase-dmel5.39 annotation includes both coding and non-coding genes, and the others include only coding genes. The results are the same for all different versions. We carried out a similar analysis with the latest genome annotation for *H. sapien*s and *M. musculus*. The results are comparable to the values previously reported (Table 2-1). Values for genome size and percentage of genes in head-to-head orientation shown in figure 1A for *H. sapiens, M. musculus, O. sativa* and *A. thaliana* were obtained from the literature.

Calculation of the fraction of gene pair combinations associated with various proteins

To calculate the expected fraction, we call P_1 the fraction of TSSs containing binding sites for a specific protein located in the 500 bp upstream region of a gene in the genome, and P_2 the fraction of protein binding sites 500 bp downstream of the TTS. For all the

gene pairs in the genome, there are N_1, N_2, N_3 and N_4 pairs for the head-to-tail, head- to-head, tail- to-tail and tail- to-head combinations, respectively. The expected number of gene pairs bound by a specific protein is $N_1 \times (1-(1-P_1)\times(1-P_2))$, $N_2 \times (1-(1-P_1)\times(1-P_1))$, $N_3 \times (1-(1-P_2)\times(1-P_2))$, $N_4 \times (1-(1-P_2)\times(1-P_1))$ for each combination. The expected fraction for each combination is then determined as the expected numbers divided by the total number of gene pairs. To calculate the observed fraction, we count the number of gene pairs (X) in each category (head-to-tail, head- to-head, tail- to-tail and tail- to-head) present in the pool of total gene pairs associated with the different proteins (T). Then the observed fraction is obtained by dividing X by T for each category.

Gene co-expression analysis

The expression value for each gene in each gene pair was extracted from the table in modENCODE_3305 (http://submit.modencode.org/submit/public/download/modENCODE_3305?root=data), which includes expression scores for different cell lines and developmental stages from embryo to adult. Pearson correlations were calculated for the expression scores for the two genes in each pair across the cell lines and developmental stages.

Alignment of BEAF-32 and histone modifications

The clustering of BEAF-32 sites and alignment of BEAF-32 with histone modifications were carried out using ChromaSig (Hon et al. 2008). Since *Drosophila* genomes are

smaller than the mammalian genomes for which this program was originally written, we changed several parameters as follows: STAT_HALF_WINDOW_SIZE = 1000 and OVERLAP_HALF_WINDOW_SIZE = 1000. The output of ChromaSig was viewed using custom scripts (available upon request) and TreeView (Saldanha 2004). To distinguish the differences between the two sides flanking BEAF-32 binding sites, the direction information from the ChromaSig output was also incorporated for graphical viewing.

Gene-pair-centric conservation analysis

BEAF-32 associated pairs are two non-overlapping genes flanking a BEAF-32 binding site. For a BEAF-32 associated pair composed of gene1 and gene2 in species A, orthologous genes are found in table gene_orthologs_fb_2011_07.tsv from Flybase. Then the BEAF-32 binding signal is examined for the corresponding intergenic region for gene1 or gene2 in the second species-species B. The term "corresponding intergenic region" signifies that this region should be downstream of gene1 or upstream of gene2 in species B if it is downstream of gene1 and upstream of gene2 in species A. If BEAF-32 is found at the corresponding intergenic region in species B, it is determined to be conserved. For clustering analysis, all BEAF-32 binding sites from the four species were pooled together. Each site in each species is assigned a value of 1 if BEAF-32 is present, -1 if BEAF-32 is not present, 0 if no ortholog is found and NA if the site is not mapped by ChIP-seq. The created matrix is then clustered using hierarchical clustering in Cluster 3.0 and the results were viewed using TreeView. For comparisons among species, the conservation score was calculated based on the peaks called by CCAT 3.0 using default parameters (enrichment value of 5). For the quantitative comparison between *D.*

melanogaster and other species, peaks used were called with an enrichment of 10 for *D. melanogaster* and an enrichment of 3 for other species. Thus, the regions called as non-conserved are the ones with at least 10-fold enrichment in *D. melanogaster* and at most 3-fold enrichment in other species. At least a 3-fold difference was required to call a gain or loss of protein binding. To count the co-occurrence of non-conserved BEAF-32 sites and chromosome rearrangements, gene pairs flanking non-conserved BEAF-32 sites in *D. melanogaster* are searched for their orthologous presence in other species. If the two genes in the gene pair are still next to each other and in the other species, it is counted as non rearranged. Otherwise it is counted as having undergone a rearrangement.

Motif analysis

Consensus sequences were discovered using Weeder (Pavesi et al. 2004) to analyze BEAF-32 binding sequences obtained from peak files called with CCAT 3.0. Changes of sequences in the BEAF-32 binding sites were determined based on the 5 bp motif sequence CGATA or its reverse complementary sequence TATCG in intergenic regions.

Other datasets

ChIP-chip results for BEAF-32, other insulator proteins, JIL1 and histone modifications in S2 cells were obtained from modENCODE (www.modencode.org/publications/integrative_fly_2010/) (Consortium et al. 2010). ChIP-seq results for Twist and Snail in embryo were obtained from the EMBL-EBI

website under accession code E-MTAB-376 (He et al. 2011). ChIP-chip data for Smc1 was obtained from GEO under accession number GSE9248 (Misulovin et al. 2008). Expression data for *Sox14* mutant animals is from GSE23355 (Ritter and Beckstead 2010).

Data access

ChIP-seq data are deposited in NCBI's Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE35648.

Acknowledgments

We would like to thank Dr. Nicolas Gompel for the images of *Drosophila* species shown in Figure 6B. We thank Chunhui Hou for enlightening discussions, Naomi Takenaka for help with the ChIP-seq libraries and the Corces lab for moral support. We also thank the Genomic Services Lab at the HudsonAlpha Institute for Biotechnology for their help in performing Illumina sequencing of ChIP-Seq samples. Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R01GM035463 to VGC and National Cancer Institute award number K01CA133106 to ER. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Table 2-1. Percentage of genes in head-to-head (<1kb) gene pairs for different species.

For each species, the denoted genome annotation was used to calculate the percentage of genes in head-to-head gene pairs. The annotations were obtained from Flybase or from the UCSC genome browser. The name also indicates the version and track of the annotation. The term "all pairs" indicates all possible non-overlapping gene pairs found using each genome annotation. hh<1 kb pairs are gene pairs in which the two genes are in head-to-head orientation and the two are separated by less than 1 kb. All genes are unique genes in all pairs and hh<1 kb genes are unique genes in head-to-head orientation separated by less than 1 kb. Percentage indicates the fraction of hh<1 kb genes out of all genes. Percentage in literature is the value previously reported in the literature as mentioned in text.

Species	Annotation	All pairs	hh<1 kb pairs	All genes	hh<1 kb genes	%	% from literature
D. melanogaster	flybase-dmel5.39	12563	2125	14888	4250	28.5 %	
D. melanogaster	UCSC-dm3-flyBaseGene	12395	2038	14596	4076	27.9 %	
D. melanogaster	UCSC-dm3-refGene	11950	1997	14421	3994	27.7 %	
D. melanogaster	UCSC-dm3-ensGene	12395	2038	14596	4076	27.9 %	
A. gambiae	UCSC-anoGam1-ensGene	13686	644	14330	1288	9.0%	
A. mellifera	UCSC-apiMel2-ensGene	11955	473	13264	946	7.1%	
C. elegans	UCSC-ce10-refGene	18695	1609	19835	3218	16.2 %	
H. sapiens	UCSC-hg19-GencodeV12	27858	1462	34454	2924	8.5%	9.4%
H. sapiens	UCSC-hg19-refGene	20488	1030	23208	2060	8.9%	9.4%
M. musculus	UCSC-mm10-refGene	21155	1017	23052	2034	8.8%	8.2%
O.sativa							9%
A.thaliana							18%

Table 2-2. Expression information for genes in Figure 2-5C.

Pearson score 1 is the correlation of expression across developmental stages for the two genes. The value is calculated as described in methods for co-expression analysis. Pearson score 2 is the correlation score reported previously (Herr and Harris 2004). Co-expression level and expression pattern in embryos have also been reported previously (Herr and Harris 2004).

		Pearson score 1	Pearson score 2	Co-expression	Embryonic expression	Embryonic expression
FBgn0035390	FBgn0052484	0.77	N/A	+++	Syncytial blastoderm, stage 7–9 PMG/AMG; stage 12–13 mesoderm	Syncytial blastoderm, stage 7–9 PMG/AMG; stage 12–13 mesoderm
FBgn0035420	FBgn0035421	0.72	0.855	++++	Syncytial blastoderm	Syncytial blastoderm
FBgn0039773	FBgn0039774	0.08	−0.640	−	Stage 10–13 prohemocytes; stage 16 PMG	Stage 10–13 prohemocytes; stage 16 PMG
FBgn0037315	FBgn0051542	-0.15	N/A	−	Absent	Absent

Table 2-3. Summary of sequence data.

Species	Aligned Reads		Genome size	Gene number	BEAF-32 peaks
	ChIP	Input			
D. melanogaster	11590739	15801184	129316289	15021	5121
D. simulans	10261595	11124982	109695738	14496	3383
D. pseudoobscura	20683221	20884796	127291806	15305	2070
D. virilis	20915794	17095302	173705494	14886	1974

Table 2-4. Summary of BEAF-32 binding sites at intergenic regions affecting body size.

Test line indicates the ID for the fly strain tested by Carreira and colleagues (Carreira et al. 2008). P-element insertions in these lines affect body size of *D. melanogaster*. In these strains, P-elements are inserted at intergenic regions that also contain BEAF-32 binding sites. Homologous gene pairs are found in the four *Drosophila* species and BEAF-32 binding between the gene pairs was analyzed for the gene pairs in each species. If BEAF-32 binds between the two genes or close to one of the genes, it is denoted as y. If BEAF-32 is not found in the gene pairs, it is denoted as n. The largest difference is observed between *D. melanogaster* and *D. virilis*.

	D. melanogaster			D. simulans		
BG02199a	FBgn0011577	FBgn0020633	y	FBgn0185823	FBgn0185824	y
BG01011	FBgn0010905	FBgn0010909	y	FBgn0185100	FBgn0185099	y
BG02435a	FBgn0036490	FBgn0029114	y	FBgn0186213	FBgn0084138	y
BG02131b	FBgn0032078	FBgn0032079	y	FBgn0193781	FBgn0193780	n
BG01613b	FBgn0259212	FBgn0037296	y		FBgn0191130	y
BG01563c	FBgn0250746	FBgn0037315	y	FBgn0191116	FBgn0191115	y
BG01713	FBgn0053100	FBgn0039149	y	FBgn0189856	FBgn0192519	y
BG02118	FBgn0020386	FBgn0035099	y	FBgn0185255	FBgn0185256	y
	D. psuedoobscura			D. virilis		
BG02199a	FBgn0078566	FBgn0078570	y	FBgn0200305	FBgn0200306	y
BG01011	FBgn0074159	FBgn0245725	n	FBgn0203234	FBgn0199618	n
BG02435a	FBgn0249992	FBgn0079928	n	FBgn0199157	FBgn0201068	n
BG02131b	FBgn0081836	FBgn0076562	y	FBgn0201353	FBgn0201349	n
BG01613b		FBgn0248710	y		FBgn0200551	n
BG01563c	FBgn0071377	FBgn0074124	y	FBgn0197337	FBgn0197336	y
BG01713	FBgn0247893	FBgn0074959	y	FBgn0201748	FBgn0201414	n
BG02118	FBgn0249759		y	FBgn0199737	FBgn0199738	n

Figure 2-1. BEAF-32 specifically associates with close head-to-head gene pairs.

(A) Genome size and percentage of genes in head-to-head gene pairs in different eukaryotic genomes. There is a high proportion of head-to-head gene pairs in the compact *D. melanogaster* genome compared to other species. (B) Snapshot of two regions of the *D. melanogaster* genome showing BEAF-32 binding sites associate with close head-to-head gene pairs. The top track represents genes. Genes above the line are transcribed from the plus strand and genes below the line are transcribed from the minus strand. The bottom track represents sites of BEAF-32 localization in the region; signal corresponds to the number of raw reads from ChIP-seq analysis. (C) Percentage of head-to-head gene pairs flanking different proteins. BEAF-32 associated pairs are significantly enriched for head-to-head gene pairs compared to the genome wide expectation as well as compared to other proteins. The error bars are from the results of different datasets. The expected and observed fraction of gene pairs was calculated independently for each dataset, and the mean and standard deviation were then determined. For BEAF-32, we used datasets obtained using embryos from this study and modEncode, and S2 cells. For twi or sna, we used datasets from different biological repeats. For Smc1, we used datasets for cell lines Kc, S2 and Bg3. (D) Distribution of distances between TSSs for genes flanking BEAF-32 and transcription factors. The number in parentheses is the total number of gene pairs in each category. BEAF-32 frequently associates with adjacent gene pairs close to each other.

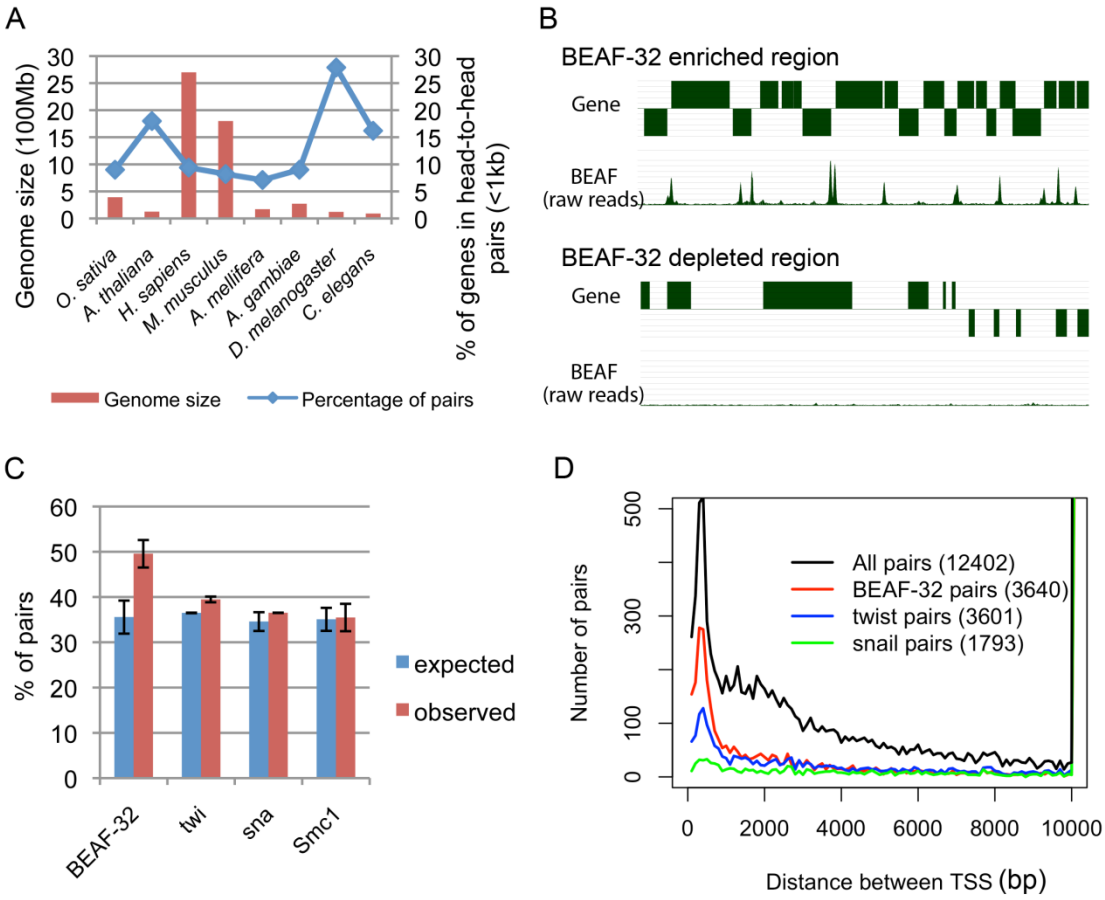
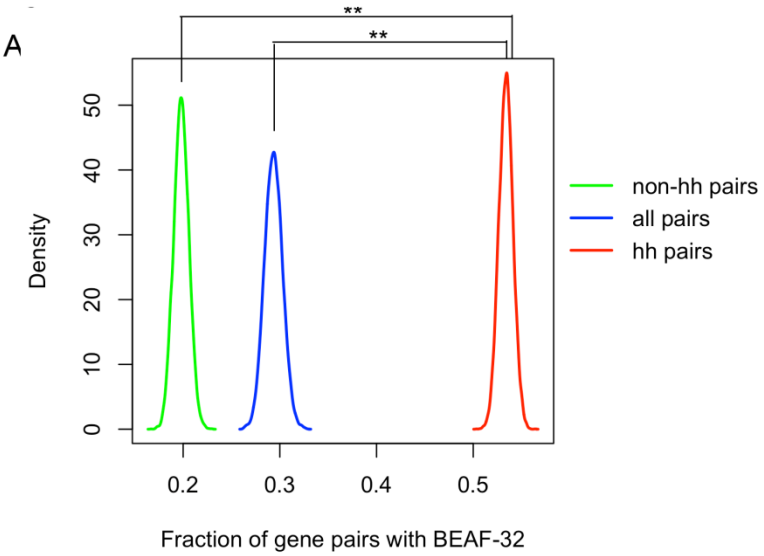
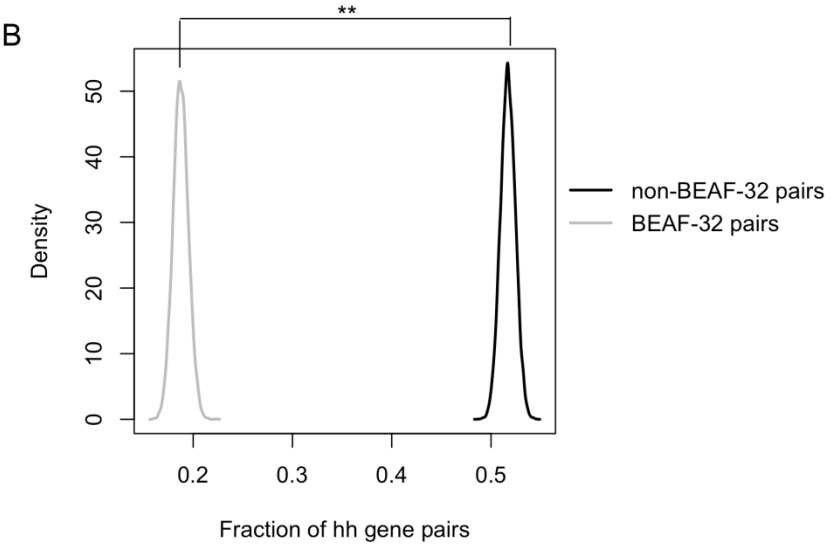

A
Genome size (100Mb)
% of genes in head-to-head pairs (<1kb)
O. sativa
A. thaliana
H. sapiens
M. musculus
A. mellifera
A. gambiae
D. melanogaster
C. elegans
Genome size
Percentage of pairs
B
BEAF-32 enriched region
Gene
BEAF (raw reads)
BEAF-32 depleted region
Gene
BEAF (raw reads)
C
% of pairs
expected
observed
BEAF-32
twi
sna
Smc1
D
Number of pairs
All pairs (12402)
BEAF-32 pairs (3640)
twist pairs (3601)
snail pairs (1793)
Distance between TSS (bp)

Figure 2-2. BEAF-32 is enriched between head-to-head gene pairs.

(A) A simulation was carried out to determine the fraction of gene pairs that are associated with BEAF-32. Three groups were investigated, all gene pairs in the genome, head-to-head gene pairs (hh pairs), and non-hh pairs. For each group, 2000 gene pairs were randomly chosen, and the fraction of gene pairs that associate with BEAF-32 was determined. The process was repeated 10,000 times to obtain the distribution shown. Differences in the distribution was tested by t-test. (B) A second simulation was carried out to determine the fraction of hh gene pairs for two groups, BEAF-32 associated gene pairs and non-BEAF-32 associated gene pairs. For each group, 2000 gene pairs were randomly selected and the fraction of hh gene pairs was determined. The process was repeated 10,000 times to create the distribution. Differences in distribution were tested by t-test. (C) Different window sizes were used to determine the fraction of hh orientation in BEAF-32 associated gene pairs as described in Figure 2-1C.


A
**
**
Density
0
10
20
30
40
50
0.2
0.3
0.4
0.5
Fraction of gene pairs with BEAF-32
non-hh pairs
all pairs
hh pairs


B
**
Density
0
10
20
30
40
50
0.2
0.3
0.4
0.5
Fraction of hh gene pairs
non-BEAF-32 pairs
BEAF-32 pairs

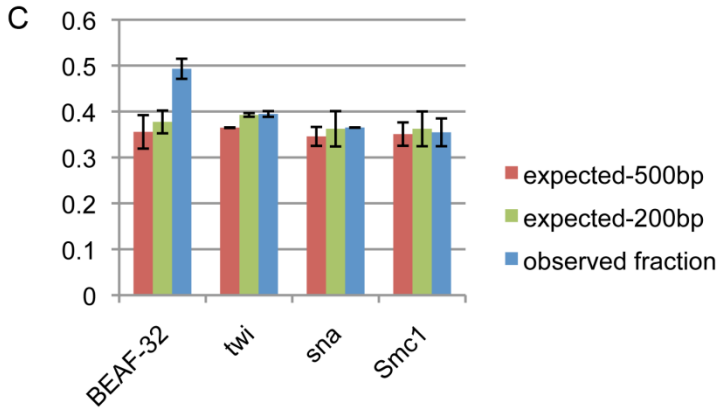

C
0.6
0.5
0.4
0.3
0.2
0.1
0
BEAF-32
twi
sna
Smc1
expected-500bp
expected-200bp
observed fraction

Figure 2-3. Percentage of head-to-head gene pairs in protein associated gene pairs. Observed and expected percentage of all gene pair combinations flanking different protein factors. In addition to BEAF-32, fractions are examined for various other proteins. JIL1 is a general transcription factor, twist and snail are transcription factors, and Smc1 is a protein associated with TSSs.

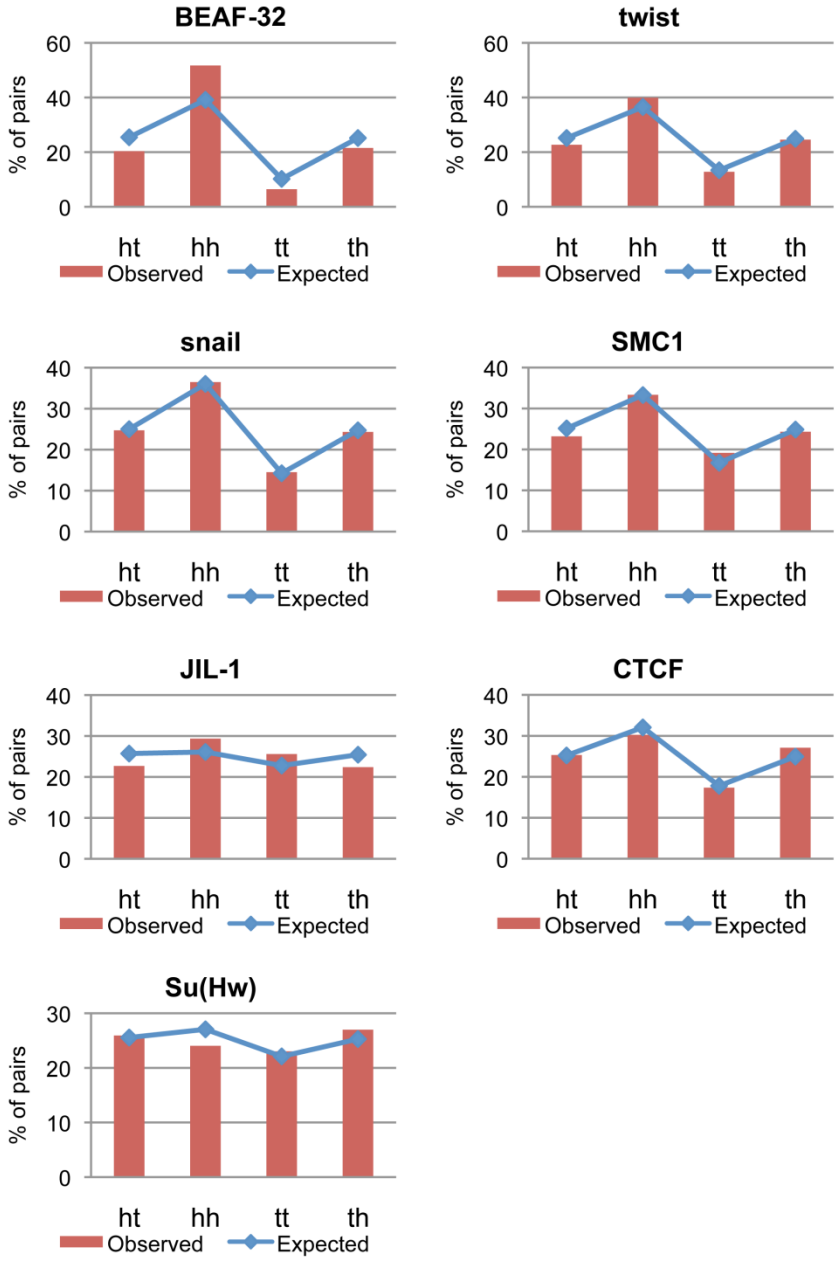
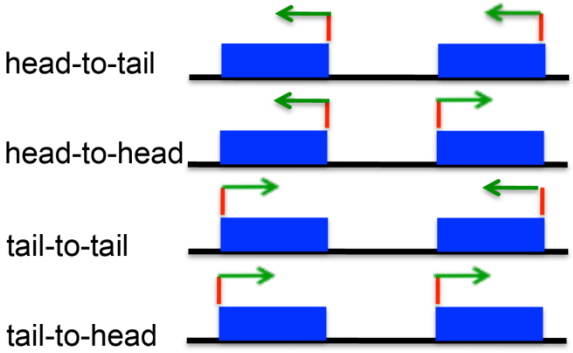

BEAF-32
% of pairs
60
40
20
0
ht hh tt th
Observed
Expected
twist
% of pairs
60
40
20
0
ht hh tt th
Observed
Expected
snail
% of pairs
40
30
20
10
0
ht hh tt th
Observed
Expected
SMC1
% of pairs
40
30
20
10
0
ht hh tt th
Observed
Expected
JIL-1
% of pairs
40
30
20
10
0
ht hh tt th
Observed
Expected
CTCF
% of pairs
40
30
20
10
0
ht hh tt th
Observed
Expected
Su(Hw)
% of pairs
30
20
10
0
ht hh tt th
Observed
Expected

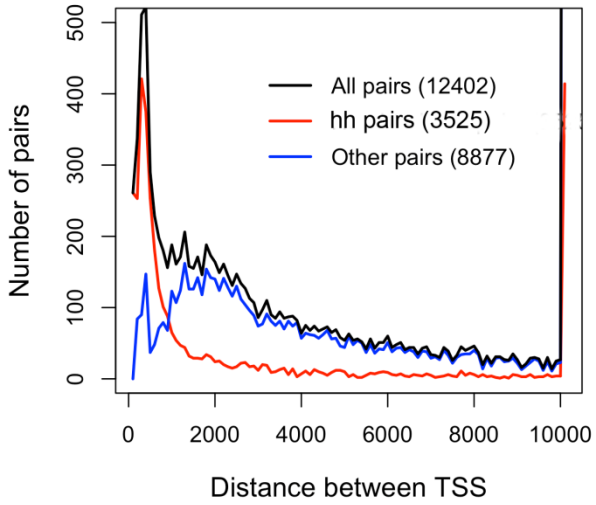
Figure 2-4. Distance between TSSs of gene pairs.

(A) Distance between the two TSSs in different gene pair combinations. The green arrow indicates the direction of transcription and the red line indicates the location of the TSS. The distance between the two TSSs is shortest in the head-to-head (hh) combination. (B) Distribution of the distances between head-to-head and other gene pairs. (C) Distribution of the distances between genes flanking BEAF-32 and other insulator proteins.

A


head-to-tail
head-to-head
tail-to-tail
tail-to-head

B


All pairs (12402)
hh pairs (3525)
Other pairs (8877)
Number of pairs
0
100
200
300
400
500
0
2000
4000
6000
8000
10000
Distance between TSS

C

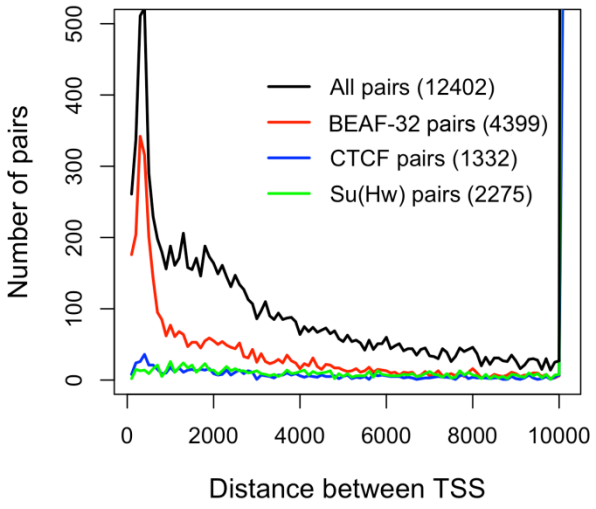

All pairs (12402)
BEAF-32 pairs (4399)
CTCF pairs (1332)
Su(Hw) pairs (2275)
Number of pairs
0
100
200
300
400
500
0
2000
4000
6000
8000
10000
Distance between TSS

Figure 2-5. BEAF-32-associated close head-to-head genes are not co-expressed.

(A) Distribution of the correlation of expression for the two genes in close head-to-head gene pairs (distance < 500bp). Red arrow indicates a secondary peak for enrichment of genes that are not co-regulated. (B) Percentage of gene pairs associated or not-associated with BEAF-32 binding sites present between co-expressed and non-co-expressed genes in close head-to-head gene pairs. (C) Examples of BEAF-32 location in co-expressed and non-co-expressed gene pairs. The blocks indicate genes with flybase IDs. Blocks on top of the track are transcribed from the plus strand, and blocks at the bottom of the track are transcribed from the minus strand. The tracks under the gene tracks show the location of BEAF-32 signal with raw reads from ChIP-seq. The symbol ‘co-ex’ represents the level of co-expression between the two genes. Detailed information about the expression of these genes is presented in Table 2-2.

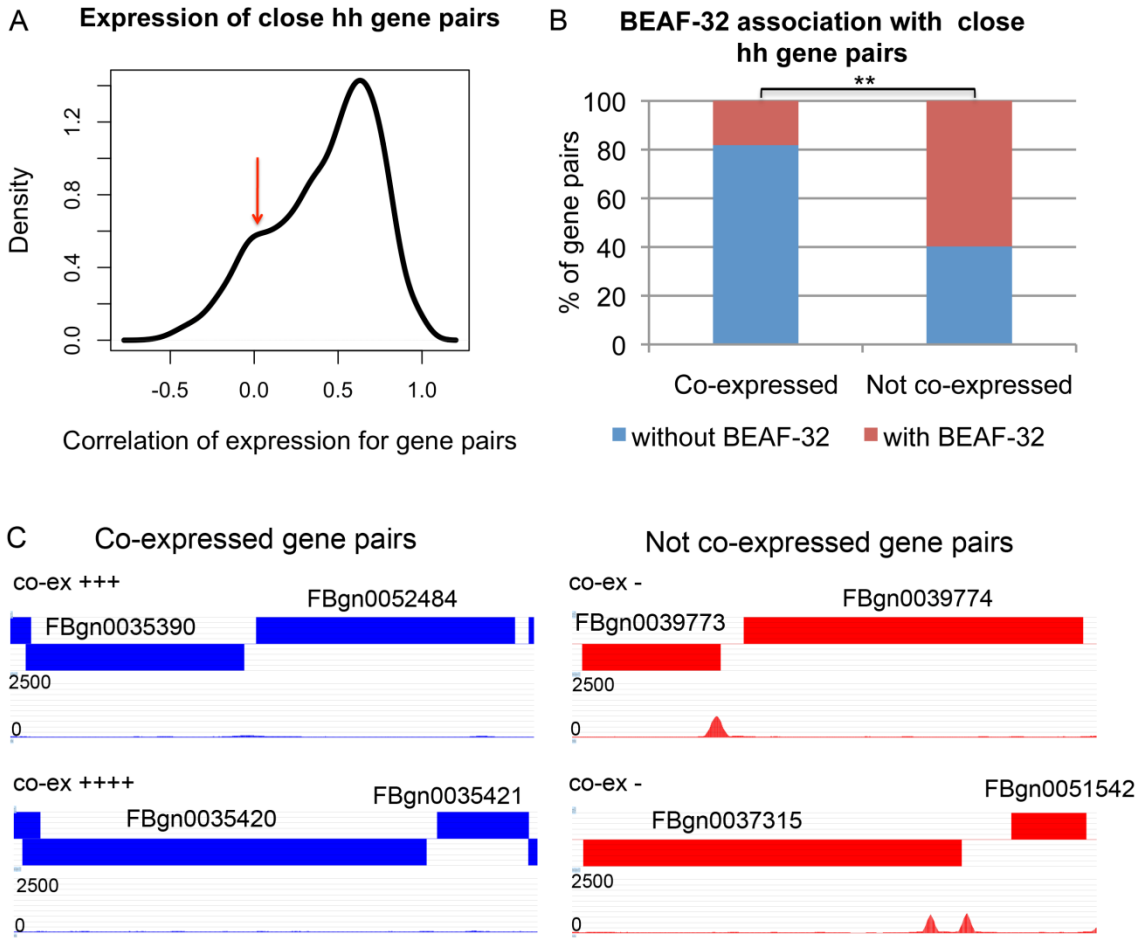
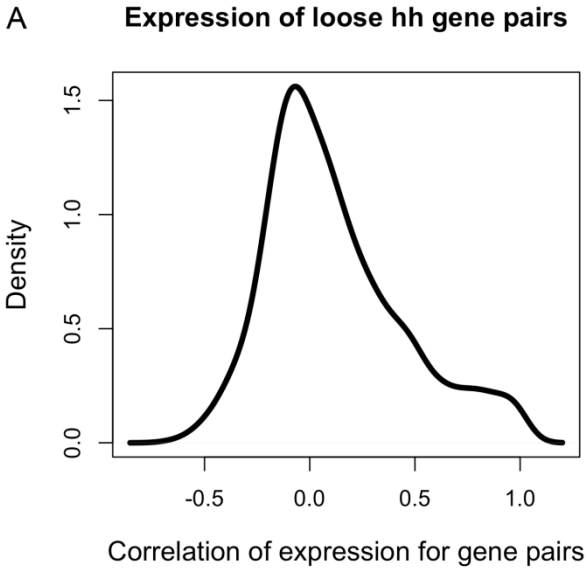

A
Expression of close hh gene pairs
Density
0.0
0.4
0.8
1.2
-0.5
0.0
0.5
1.0
Correlation of expression for gene pairs
B
BEAF-32 association with close hh gene pairs
**
% of gene pairs
100
80
60
40
20
0
Co-expressed
Not co-expressed
without BEAF-32
with BEAF-32
C
Co-expressed gene pairs
Not co-expressed gene pairs
co-ex +++
FBgn0035390
FBgn0052484
2500
0
co-ex -
FBgn0039773
FBgn0039774
2500
0
co-ex ++++
FBgn0035420
FBgn0035421
2500
0
co-ex -
FBgn0037315
FBgn0051542
2500
0

Figure 2-6. Correlation of expression for gene pairs.

(A) Distribution of the correlation of expression for distant head-to-head gene pairs (distance >1kb). (B) Distribution of the correlation of expression for non-head-to-head gene pairs.

A


Expression of loose hh gene pairs
Density
0.0
0.5
1.0
1.5
-0.5
0.0
0.5
1.0
Correlation of expression for gene pairs

B

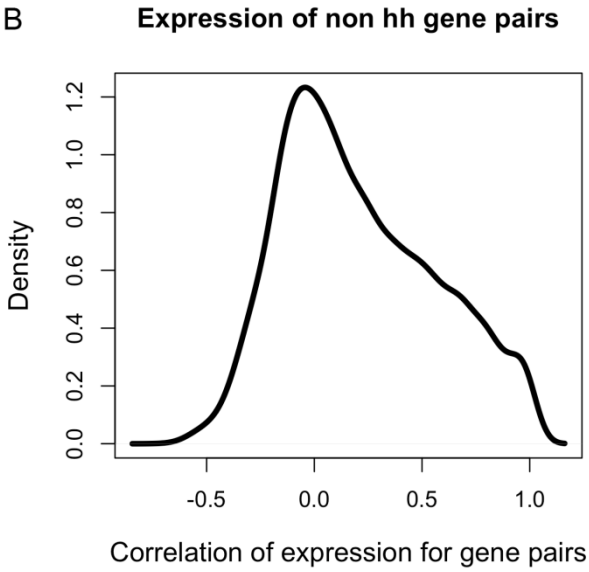

Expression of non hh gene pairs
Density
0.0
0.2
0.4
0.6
0.8
1.0
1.2
-0.5
0.0
0.5
1.0
Correlation of expression for gene pairs

Figure 2-7. BEAF-32 separates close head-to-head gene pairs to achieve differential regulation of transcription.

(A) Alignment and clustering of BEAF-32 and histone modifications in *D. melanogaster* S2 cells. All sites were identified and aligned using ChromaSig. Each vertical stripe represents a 3 kb region. Clusters I-V grouped here are clusters with BEAF-32 binding from Figure 2-8. (B-F) Mean value of enrichment for sites in cluster I. Site 0 is the site where BEAF-32 is enriched. Each figure represents a 3 kb region flanking site 0. Each colored line represents a different type of gene pair arrangement case in cluster I: head-to-head (red), tail-to-head (blue) and head-to-tail (green). (G) Fraction of gene pairs whose expression changes significantly in the same up or down direction in *Sox14* mutant animals. Changes are considered significant when the difference is at least 3-fold.

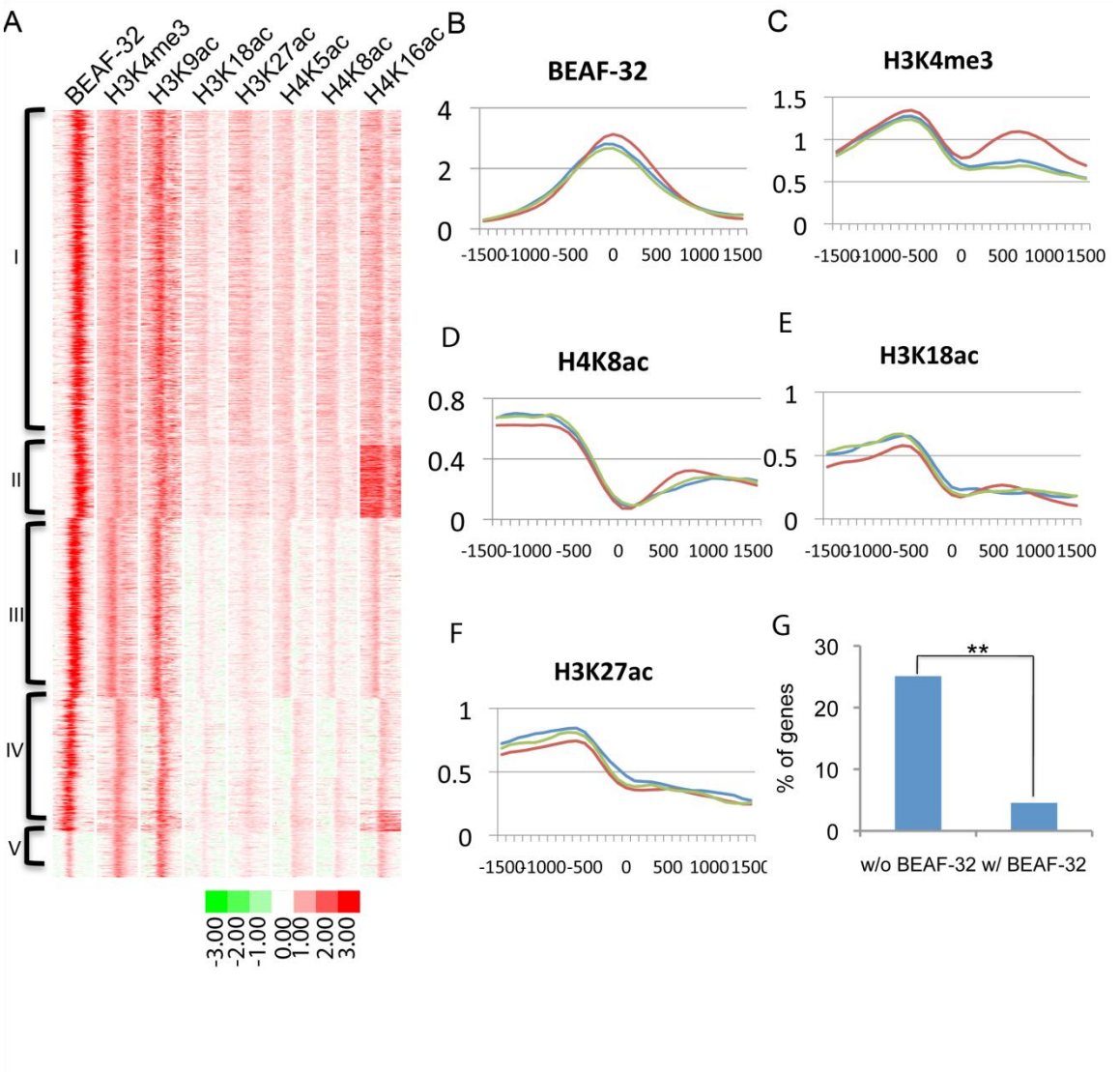

A
BEAF-32
H3K4me3
H3K9ac
H3K18ac
H3K27ac
H4K5ac
H4K8ac
H4K16ac
I
II
III
IV
V
-3.00
-2.00
-1.00
0.00
1.00
2.00
3.00
B
BEAF-32
C
H3K4me3
D
H4K8ac
E
H3K18ac
F
H3K27ac
G
% of genes
**
w/o BEAF-32
w/ BEAF-32

Figure 2-8. Clustering of BEAF-32, other *Drosophila* insulator proteins, and various histone modifications in *D. melanogaster* S2 cells. Clusters containing BEAF-32 are displayed in Figure 2-7A. The numbers indicate the corresponding cluster in Figure 2-7A.

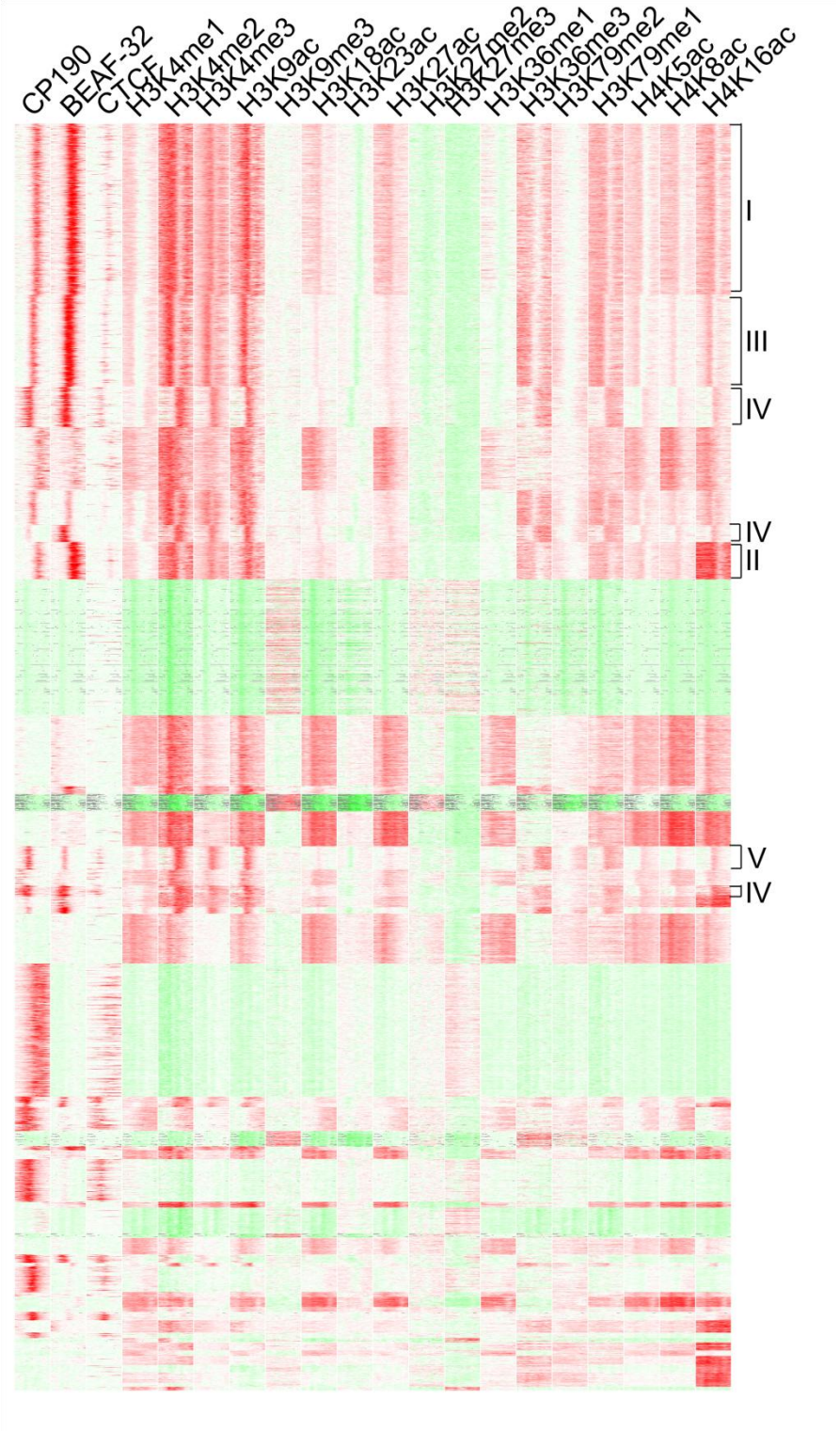

CP190
BEAF-32
CTCF
H3K4me1
H3K4me2
H3K4me3
H3K9ac
H3K9me3
H3K18ac
H3K23ac
H3K27ac
H3K27me2
H3K27me3
H3K36me1
H3K36me3
H3K79me2
H3K79me1
H4K5ac
H4K8ac
H4K16ac
I
III
IV
IV
II
V
IV

Figure 2-9. Conservation and divergence of BEAF-32 sites in *Drosophila* species.

(A) Distribution of BEAF-32 binding sites with respect to various gene landmarks. (B) Percentage of various gene pair combinations flanking BEAF-32 binding sites; ht, head-to-tail; hh, head-to-head; tt, tail-to-tail; th, tail-to-head. BEAF-32 association with head-to-head gene pairs is conserved. (C) Consensus motif for BEAF-32 occupied sequences in the four species. (D) Correlation of BEAF-32 binding divergence with evolutionary distance between *D. melanogaster* and other species. The Y axis represents the percentage of BEAF-32 binding sites that are not conserved in the other three species with respect to all BEAF-32 binding sites in *D. melanogaster*. (E-F) Species-specific loss of BEAF-32 binding associates with species-specific loss of the BEAF-32 motif in the DNA sequence. Dark blue bars represent the background absence of BEAF-32 motif for all BEAF-32 associated gene pairs in each species. Light blue bars represent absence of the BEAF-32 motif for a group of gene pairs with BEAF-32 binding lost only in *D. pseudoobscura*(E) or only in *D. virilis*(F).

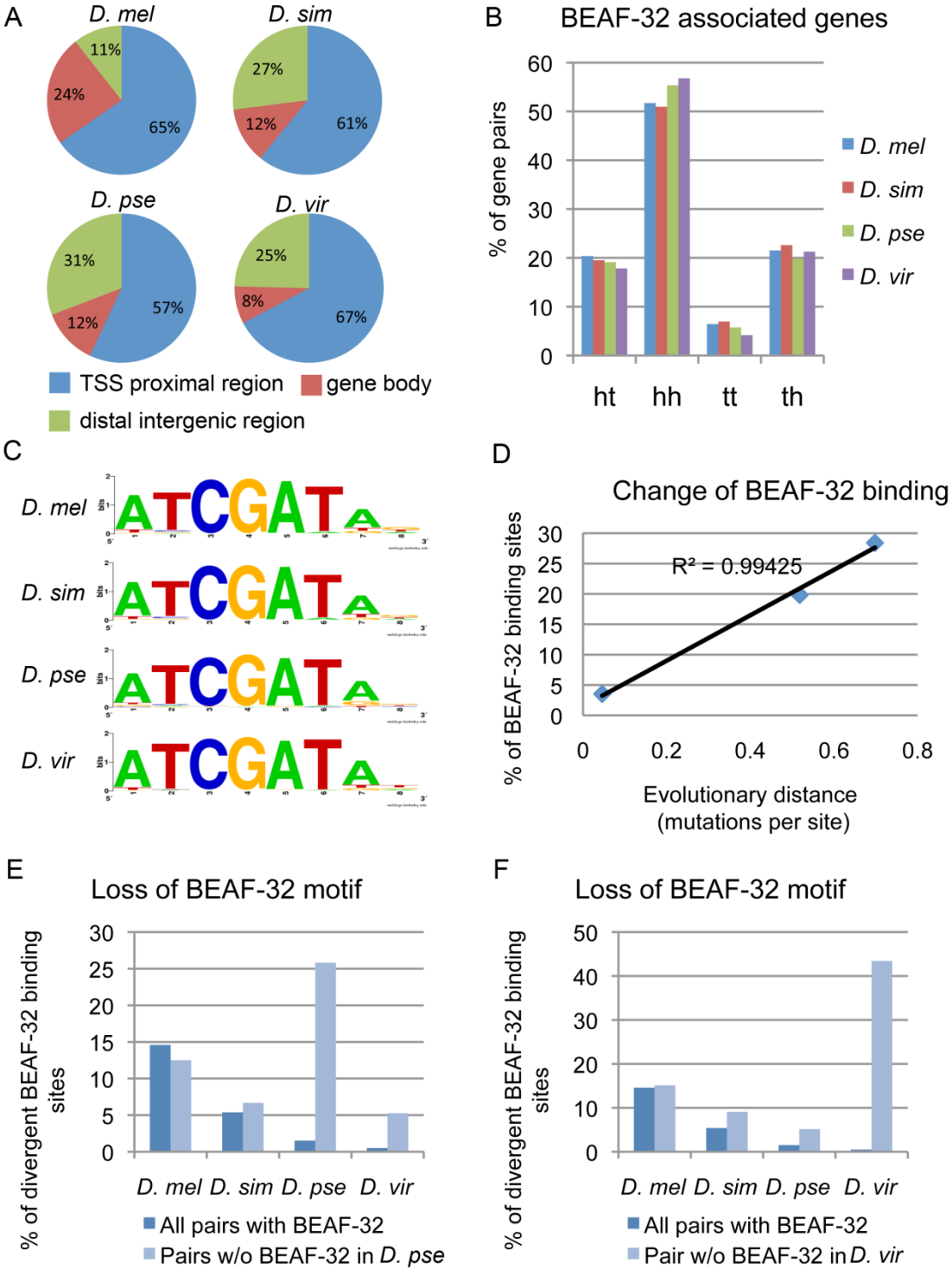
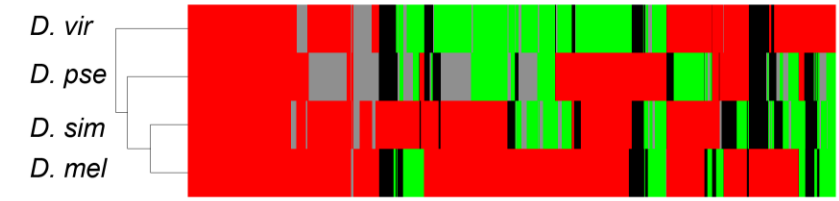

A
D. mel
11%
24%
65%
D. sim
27%
12%
61%
D. pse
31%
12%
57%
D. vir
25%
8%
67%
TSS proximal region
gene body
distal intergenic region
B
BEAF-32 associated genes
% of gene pairs
60
50
40
30
20
10
0
ht
hh
tt
th
D. mel
D. sim
D. pse
D. vir
C
D. mel
D. sim
D. pse
D. vir
D
Change of BEAF-32 binding
% of BEAF-32 binding sites
30
25
20
15
10
5
0
R² = 0.99425
0
0.2
0.4
0.6
0.8
Evolutionary distance
(mutations per site)
E
Loss of BEAF-32 motif
% of divergent BEAF-32 binding sites
30
25
20
15
10
5
0
D. mel D. sim D. pse D. vir
All pairs with BEAF-32
Pairs w/o BEAF-32 in D. pse
F
Loss of BEAF-32 motif
% of divergent BEAF-32 binding sites
50
40
30
20
10
0
D. mel D. sim D. pse D. vir
All pairs with BEAF-32
Pair w/o BEAF-32 in D. vir

Figure 2-10. BEAF-32 binding across *Drosophila* species.

(A) Clustering of BEAF-32 binding pattern. Red is binding, green is not binding, black indicates lack of an ortholog, and grey indicates intergenic regions that are not well mapped. (B) Distribution of BEAF-32 binding diversity between *D. melanogaster* and each of the other three species.

A


D. vir
D. pse
D. sim
D. mel

B

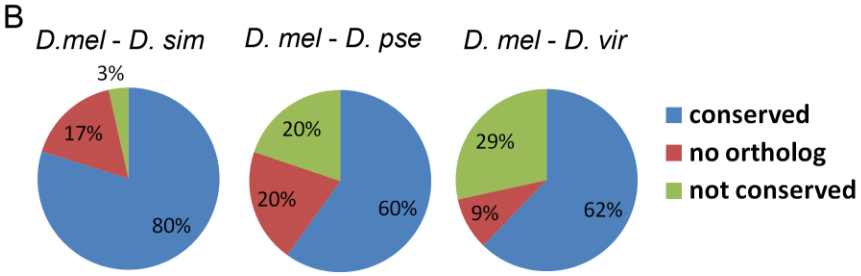

D.mel - D. sim
D. mel - D. pse
D. mel - D. vir
3%
17%
80%
20%
20%
60%
29%
9%
62%
conserved
no ortholog
not conserved

Figure 2-11. Changes in BEAF-32 binding correlate with changes in genome organization and function.

(A) BEAF-32 density decreases as gene density declines. The left Y axis represents the percentage of gene pairs containing BEAF-32 with respect to all well mapped gene pairs and is a measure of BEAF-32 density. The right Y axis indicates the number of genes per Mb as a measure of gene density. (B) Arrangement of the *myoglianin* and *eyeless* genes and location of BEAF-32 binding sites. Light green shadowing indicates the orthologous genes in *D. virilis*. (C) Percentage of divergent BEAF-32 binding sites either associate or not associate with chromosome rearrangement between *D. melanogaster* and the species listed. The numbers above the bar indicate the number of cases in each category. (D) An example of gene arrangement and location of BEAF-32 binding sites in a region whose mutation affects body size in *D. melanogaster* is shown for four different species. The mutation affecting body size results in alteration of sequences in the intergenic region encompassing the BEAF-32 binding site in *D. melanogaster*. Light green shadowing represents the four orthologous regions in each of the *Drosophila* species.

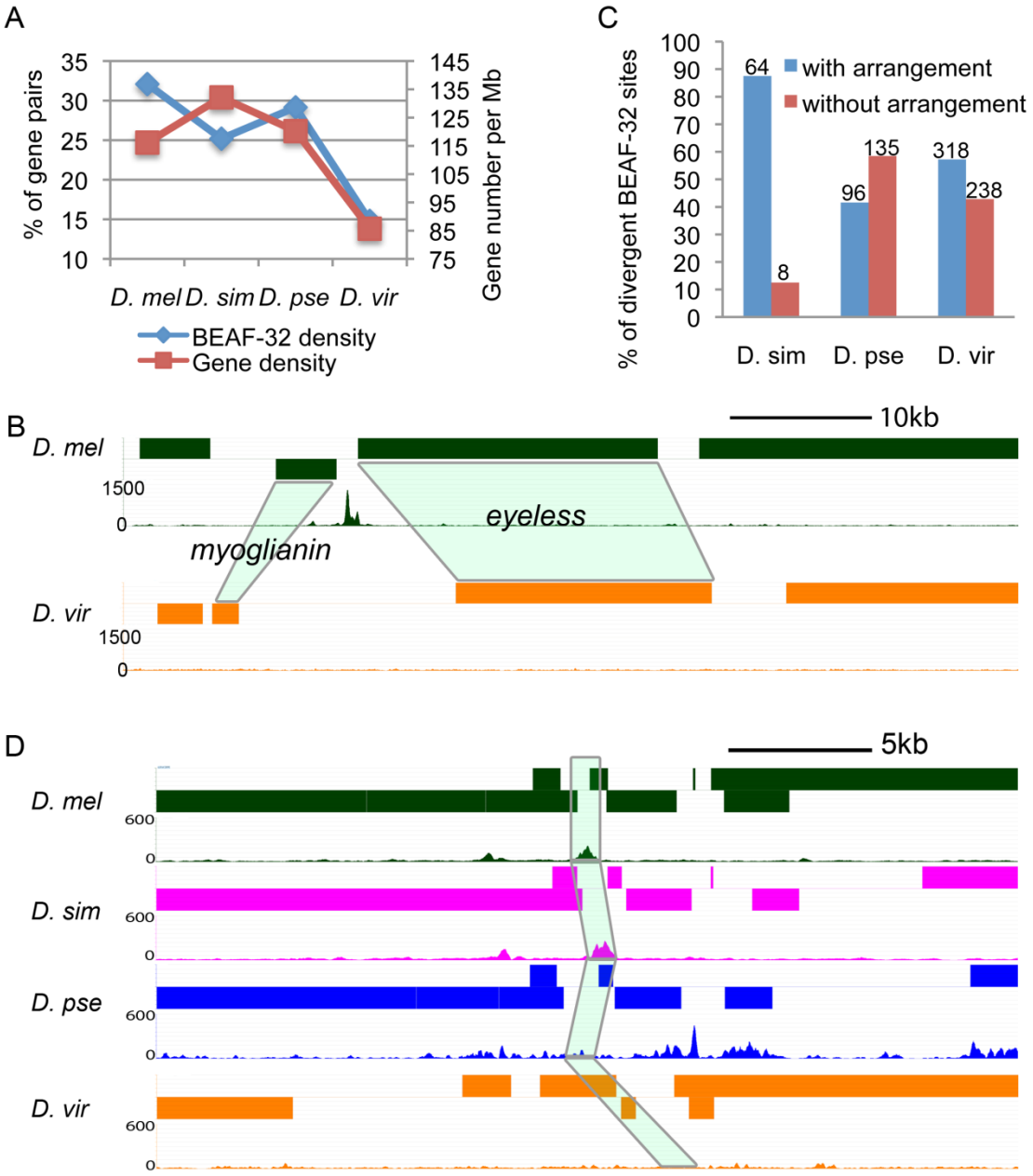

A
% of gene pairs
35
30
25
20
15
10
145
135
125
115
105
95
85
75
Gene number per Mb
D. mel
D. sim
D. pse
D. vir
BEAF-32 density
Gene density
C
% of divergent BEAF-32 sites
100
90
80
70
60
50
40
30
20
10
0
64
8
96
135
318
238
with arrangement
without arrangement
D. sim
D. pse
D. vir
B
10kb
D. mel
1500
0
myoglianin
eyeless
D. vir
1500
0
D
5kb
D. mel
600
0
D. sim
600
0
D. pse
600
0
D. vir
600
0

Figure 2-12. Simplified models for the role of BEAF-32 during evolution of *Drosophila* species. (A) Two alternative possibilities explaining how alterations in BEAF-32 binding may affect transcription. Blocks indicate genes. Black and white indicate different transcription regulatory modes of the genes. Grey means converged regulation for the two genes. Gain or loss of BEAF-32 binding when gene pairs are reorganized to maintain proper transcription (top).Gain or loss of BEAF-32 binding when gene organization does not change to create transcription diversity (bottom). (B) Phylogeny and phenotype of *Drosophila* species analyzed in this study.

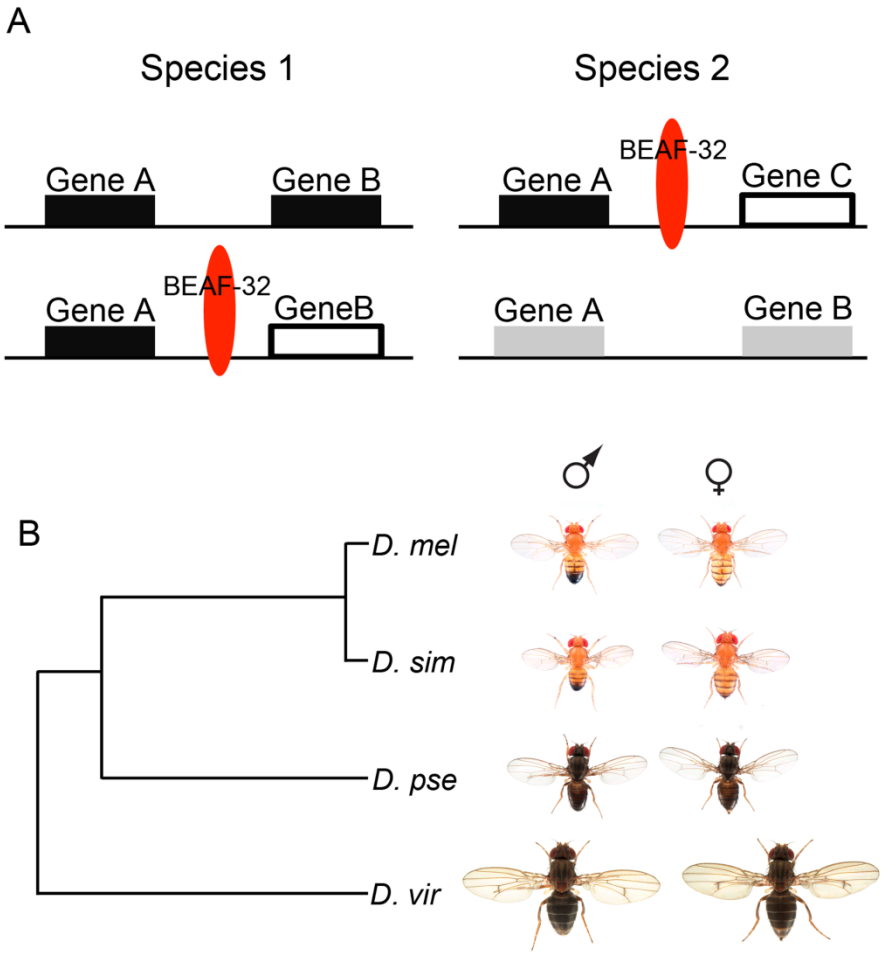

A
Species 1
Species 2
Gene A
Gene B
BEAF-32
Gene A
GeneB
BEAF-32
Gene A
Gene C
Gene A
Gene B
B
♂
♀
D. mel
D. sim
D. pse
D. vir

Figure 2-13. Conservation of BEAF-32 protein sequences in the four *Drosophila* species analyzed. The blue and green lines indicate the two functional regions of the BEAF-32 protein. The red box indicates the conserved region used to generate the BEAF-32 antibody.

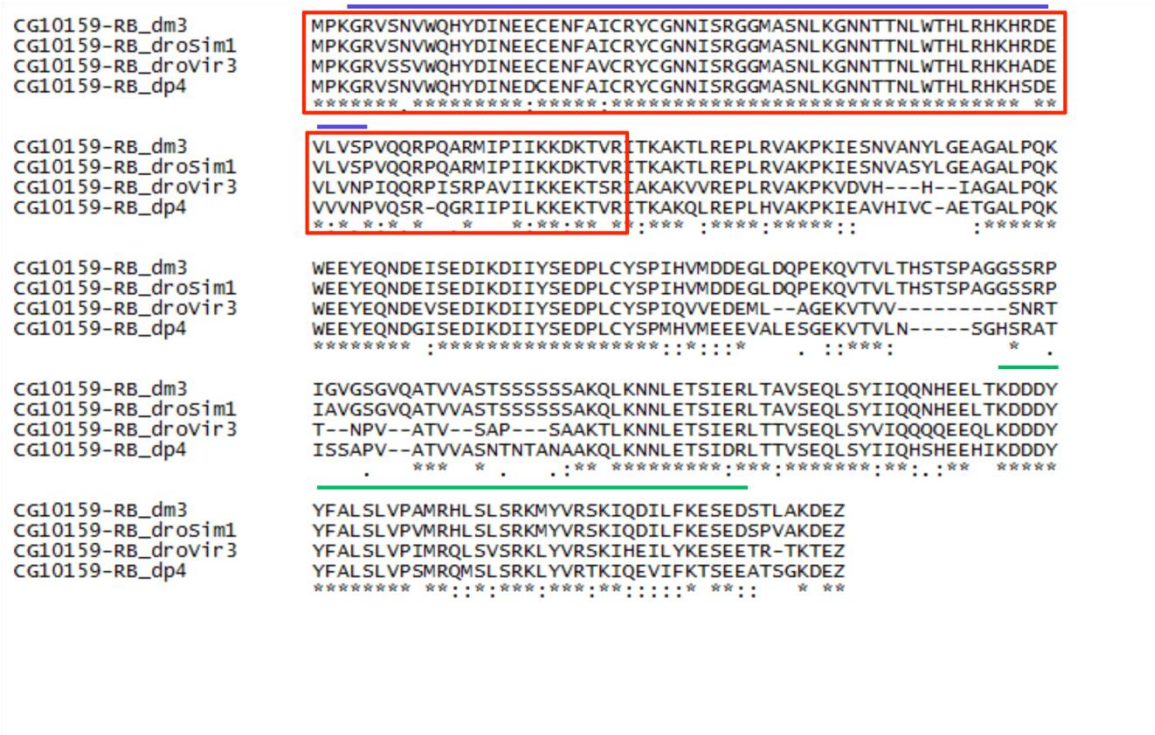
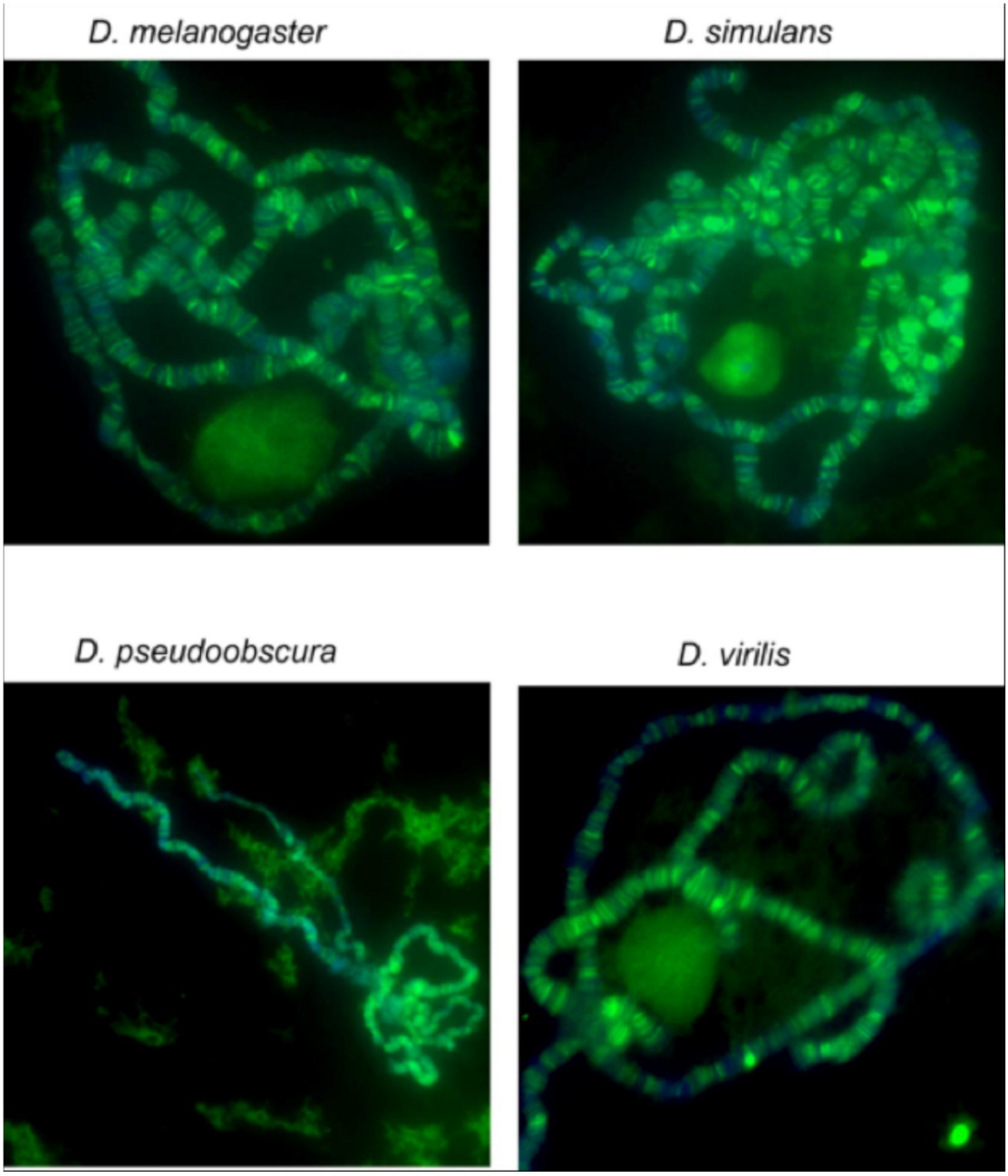

CG10159-RB_dm3 MPKGRVSNVWQHYDINEECENFAICRYCGNNISRGGMASNLKGNNTTNLWTHLRHKHRDE
CG10159-RB_droSim1 MPKGRVSNVWQHYDINEECENFAICRYCGNNISRGGMASNLKGNNTTNLWTHLRHKHRDE
CG10159-RB_droVir3 MPKGRVSSVWQHYDINEECENFAVCRYCGNNISRGGMASNLKGNNTTNLWTHLRHKHADE
CG10159-RB_dp4 MPKGRVSNVWQHYDINEDCENFAICRYCGNNISRGGMASNLKGNNTTNLWTHLRHKHSDE
CG10159-RB_dm3 VLVSPVQQRPQARMIPIIKKDKTVRITKAKTLREPLRVAKPKIESNVANYLGEAGALPQK
CG10159-RB_droSim1 VLVSPVQQRPQARMIPIIKKDKTVRITKAKTLREPLRVAKPKIESNVASYLGEAGALPQK
CG10159-RB_droVir3 VLVNPIQQRPISRPAVIIKKEKTSRIAKAKVVREPLRVAKPKVDVH---H--IAGALPQK
CG10159-RB_dp4 VVVNPVQSR-QGRIIPILKKEKTVRITKAKQLREPLHVAKPKIEAVHIVC-AETGALPQK
CG10159-RB_dm3 WEEYEQNDEISEDIKDIIYSEDPLCYSPIHVMDDEGLDQPEKQVTVLTHSTSPAGGSSRP
CG10159-RB_droSim1 WEEYEQNDEISEDIKDIIYSEDPLCYSPIHVMDDEGLDQPEKQVTVLTHSTSPAGGSSRP
CG10159-RB_droVir3 WEEYEQNDEVSEDIKDIIYSEDPLCYSPIQVVEDEML--AGEKVTVV---------SNRT
CG10159-RB_dp4 WEEYEQNDGISEDIKDIIYSEDPLCYSPMHVMEEEVALESGEKVTVLN-----SGHSRAT
CG10159-RB_dm3 IGVGSGVQATVVASTSSSSSSAKQLKNNLETSIERLTAVSEQLSYIIQQNHEELTKDDDY
CG10159-RB_droSim1 IAVGSGVQATVVASTSSSSSSAKQLKNNLETSIERLTAVSEQLSYIIQQNHEELTKDDDY
CG10159-RB_droVir3 T--NPV--ATV--SAP---SAAKTLKNNLETSIERLTTVSEQLSYVIQQQQEEQLKDDDY
CG10159-RB_dp4 ISSAPV--ATVVASNTNTANAAKQLKNNLETSIDRLTTVSEQLSYIIQHSHEEHIKDDDY
CG10159-RB_dm3 YFALSLVPAMRHLSLSRKMYVRSKIQDILFKESEDSTLAKDEZ
CG10159-RB_droSim1 YFALSLVPVMRHLSLSRKMYVRSKIQDILFKESEDSPVAKDEZ
CG10159-RB_droVir3 YFALSLVPIMRQLSVSRKLYVRSKIHEILYKESEETR-TKTEZ
CG10159-RB_dp4 YFALSLVPSMRQMSLSRKLYVRTKIQEVIFKTSEEATSGKDEZ

Figure 2-14. Immunofluorescence microscopy of polytene chromosomes of different *Drosophila* species using an antibody against the *D. melanogaster* BEAF-32B sequence shown in Figure 2-13. The antibody recognizes BEAF-32B protein in all four species.


D. melanogaster
D. simulans
D. pseudoobscura
D. virilis

Chapter 3

A specific subset of *Drosophila* Myc sites remains associated with mitotic chromosomes co-localized with insulator proteins

Abstract

Myc has been characterized as a transcription factor that activates expression of genes involved in pluripotency and cancer, and as a component of the replication complex. Here we find that Myc is present at promoters and enhancers of *D. melanogaster* genes during interphase. Myc co-localizes with Orc2, which is part of the pre-replication complex during G1. As is the case in mammals, Myc associates preferentially with paused genes, suggesting that it may also be involved in the release of RNAPII from promoter proximal pausing in *Drosophila*. Interestingly, about 40% of Myc sites present in interphase persists during mitosis. None of the Myc mitotic sites correspond to enhancers and only some correspond to promoters. The rest of mitotic Myc sites overlap with multiple insulator proteins that are also maintained in mitosis. These results suggest alternative mechanisms to explain the role of Myc in pluripotency and cancer.

Introduction

Myc has been extensively studied as an oncogene that plays critical roles in cancer initiation and metastasis of may different types of tumors (Wolfer and Ramaswamy 2011; Dang 2012). Myc is a sequence specific DNA binding protein (Blackwell et al. 1990) that can bind to both the canonical (CACGTG or CATGTG) and non-canonical CA--TG E box sequences (Blackwell et al. 1993). Although Myc has been found to regulate various cellular processes including cell growth, cell proliferation, and cell differentiation, the mechanisms by which it elicits neoplastic transformation are not well understood.

Myc is a basic-helix-loop-helix leucine zipper (bHLH-Zip) transcription factor that regulates the expression of protein coding genes and microRNAs (O'Donnell et al. 2005; Aguda et al. 2008; Chang et al. 2008; Lovén et al. 2010). Genomic searches for Myc target genes have uncovered a role for this protein in the regulation of hundreds of genes involved in cell cycle progression, differentiation, apoptosis, DNA repair, angiogenesis, chromosome instability, and ribosome biogenesis (Meyer and Penn 2008; van Riggelen et al. 2010; Dang 2012). Results suggest that Myc may regulate expression of these genes, at least in part, by interacting with P-TEFb to release RNA polymerase II (RNAPII) from promoter proximal pausing and bring about productive elongation (Kanazawa et al. 2003; Gargano et al. 2007; Rahl et al. 2010). Most genome-wide studies of Myc in mammalian cells have focused on its presence at promoter regions (Li et al. 2003; Guccione et al. 2006; Kidder et al. 2008). However, Myc is also found in non-promoter sequences. For example, Myc is enriched in the first intron of genes, and about

10% of Myc sites are present in intergenic regions (>100 kb from genes) in human B cells (Zeller et al. 2006). In mouse fibroblasts, 30.4% of Myc sites are intergenic (>1 kb from genes) and 22.4% are intragenic (1 kb downstream of TSS to 3' end) (Perna et al. 2012). These non-promoter Myc sites may function as transcriptional regulatory elements, such as enhancers, but their role has not been studied in detail.

In addition to its role in transcription of genes encoding proteins involved in DNA replication, Myc may also regulate this process directly. Cells overexpressing Myc become polyploid but do not enter mitosis (Li and Dang 1999). In *Drosophila*, Myc is required for endoreplication (Maines et al. 2004; Pierce et al. 2004), and it has been suggested that the role of Myc in replication is independent of transcription (Dominguez-Sola et al. 2007). In human cells, the Myc protein interacts with the pre-replication complex and it has been shown to be required for recruitment of Mcm proteins at specific loci (Dominguez-Sola et al. 2007; Swarnalatha et al. 2012), but whether this is a general phenomenon, and Myc co-localizes genome-wide with the origin recognition complex, has not been investigated.

Myc controls a variety of cellular processes required for cell differentiation and is essential for cellular reprogramming to induce pluripotency and stem cell renewal (Smith et al. 2010; Varlakhanova et al. 2010; Moumen et al. 2012). The role of Myc in these cellular processes may be a consequence of its effects on gene expression at the local level but other evidence suggests that Myc can also affect chromatin more globally (Varlakhanova and Knoepfler 2009). An important property of Myc that has been largely ignored when considering potential mechanisms by which this protein can affect gene

expression is that it remains bound to DNA during mitosis (O'Donovan et al. 2010; Ohta et al. 2010), raising the question of whether some of the functions ascribed to this protein are actually a consequence of its presence in mitotic chromosomes. However, the location and function of Myc in mitotic cells has not been explored. Here we examine the distribution of Myc during interphase and mitosis in *Drosophila* Kc cells. We find that Myc co-localizes extensively with Orc2 during interphase, supporting a generalized role for Myc in the pre-replication complex. In addition to promoters, Myc is also present at enhancers of *Drosophila* genes. Interestingly, only a specific subset of interphase Myc sites remain in mitotic chromosome. Mitotic Myc sites include a fraction of promoter regions and aligned insulators, where several insulator proteins co-localize within a 300 bp region. These results suggest that Myc may have an as of yet unappreciated role in the maintenance of chromosome structure and epigenetic information during the cell cycle that may explain some of its effects in tumorigenesis and pluripotency.

Results

Myc is present at the promoters of paused genes

In order to study changes in the distribution of Myc, we performed chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) in *D. melanogaster* Kc cells using an antibody against Myc. To distinguish possible roles of Myc in mitosis versus other stages of the cell cycle, we partially synchronized the cells, labeled them with antibodies to Lamin Dm0, and separated the mitotic and interphase populations as previously described (Gurudatta et al, 2012). In order to determine whether the role of Myc in transcription and replication is conserved in *Drosophila*, we first mapped the binding sites for this protein in interphase when the cells are undergoing a process of biomass accumulation and preparing for the next cell cycle. We identified approximately 4000 Myc binding sites across the genome of interphase cells (Table 2-1). Analysis of these data indicates that Myc associates with coding and non-coding genes (Figure 3-1A, left and right panels, respectively). Myc binds preferentially to promoter proximal regions (between TSSs and -200 bp). Only 8% of Myc sites fall in exons, including 5' and 3' UTRs. In addition to promoter regions, Myc also binds significantly in introns (21% of sites) and intergenic regions (17% of sites) (Figure 3-1B).

Myc has been extensively characterized as a transcription factor in mammals. In *Drosophila* 53% of Myc sites are located in the promoter regions of genes and 4% in the 5'UTR (Figure 3-1B). These sites of Myc associate with genes involved in rRNA synthesis, cell cycle, and development (Table 2-2). To understand how Myc

regulates adjacent genes, we examined their expression status. Generally Myc associates with genes that also have RNAPII at the TSS, suggesting they have undergone transcription initiation. However, the difference in RNAPII levels between the promoter region and the gene body is larger for genes with Myc than for genes without Myc (Figure 3-1A). These observations suggest that Myc-associated genes have high pausing indexes. In mammalian cells, Myc plays a role in the release of RNAPII from promoter-proximal pausing (Rahl et al. 2010). We therefore examined whether *Drosophila* Myc is also associated with paused genes by determining the pausing index, which is a measure of the difference in RNAPII levels between the promoter and gene body. The results indicate that Myc preferentially associates with paused genes in *Drosophila*. More Myc-associated genes show high pausing index than the average genome level. Around 41% of Myc associated genes show pausing indexes higher than 1, while only around 16% of all genes show a pausing index larger than 1 (K-S test, $p<2x10^{-16}$) (Figure 3-1C). Although Myc-associated genes show high pausing indexes, they are also highly expressed. About 50% of Myc-associated genes belong to the group with the highest transcription levels (Figure 3-1D), suggesting that RNAPII in Myc-associated paused genes is quickly released into productive elongation.

The role of Myc at non-promoter regions

As previously observed in mammalian cells (Zeller et al. 2006; Perna et al. 2012), *Drosophila* Myc also binds to non-promoter regions (Figure 3-1B). Genome-wide studies

of Myc distribution have been carried out in different types of mammalian cells but no detailed analysis of the function of these non-promoter sites is available. We therefore examined the distribution of histone modifications and RNAPII enrichment at these non-promoter Myc sites. Results indicate that a subset of these Myc sites show chromatin signatures characteristic of enhancers (Figure 3-2A and 2B). Myc sites at promoter regions are enriched in H3K4me3 and RNAPII. In contrast, a subset of Myc sites at non-promoter regions show H3K4me1 while H3K4me3 and RNAPII are absent (Figure 3-2B). Enrichment in H3K4me1 without H3K4me3 is characteristic of enhancers (Heintzman et al. 2007). These Myc sites also contain H3K27ac, indicating they correspond to active enhancers (Figure 3-2B). We have previously identified all enhancers in *Drosophila* Kc cells (Kellner et al. 2012) and we used this information here to compare enhancers with and without Myc. The results suggest that most active enhancers containing H3K27ac also have Myc, while inactive enhancers lacking H3K27ac are depleted of Myc (Figure 3-2C). Thus, a subset of non-promoter Myc sites appears to be present at active enhancers. However, not all the non-promoter Myc sites show enhancer-like chromatin features. A second subset of non-promoter Myc sites (denoted with a question mark in Figure 3-2B) is depleted of H3K4me1, H3K4me3, and H3K27ac and, therefore, it does not correspond to enhancers or promoters.

Myc associates with Orc2 genome-wide in *D. melanogaster*

In mammalian cells, the Myc protein interacts with the pre-replication complex and it has been found at specific DNA replication origins with Origin recognition

complex (Orc) proteins (Dominguez-Sola et al. 2007). To test whether this is also the case in *Drosophila*, and some of the sites with unknown function in Figure 3-2B correspond to replication origins, we compared the binding profiles of Myc and the Orc2 component of the complex. The signals for the two proteins show significant overlap in specific regions of the genome (Figure 3-3A). We then examined genome-wide correlations between the two proteins using heatmaps to visualize the information. The results indicate that Orc2 is present at most Myc sites in the genome and vice versa (Figure 3-3B). Interestingly, the correlation between the amount of Myc and Orc2 in the genome is higher at Orc2 sites than at Myc sites (Figures 3-3C and 3-3D). Thus, Myc co-localizes with Orc2 genome-wide in *Drosophila* cells at both promoter and non-promoter regions.

A distinct subset of Myc sites remains bound to chromosomes during mitosis

Myc has been shown to be present in mitotic chromosomes (O'Donovan et al. 2010; Ohta et al. 2010) but its specific distribution in chromatin during mitosis has never been analyzed. The presence of Myc in mitotic chromosomes may be critical for its role in transcription. To gain further insights into mechanisms by which Myc affects gene expression, we mapped Myc binding sites in mitotic cells and compared the distribution of this protein between interphase and mitosis. The results indicate that all Myc mitotic sites are also occupied by this protein during interphase, but not all Myc sites in interphase are retained in mitosis (Figure 3-4A). Myc sites in the genome can therefore

be classified as interphase-specific (Class I) or common to mitosis and interphase (Class II). We will refer to this second group as mitotic sites, although they are also present in interphase (Figure 3-4A). In interphase, the average enrichment of Myc at Class II sites is only about half of the average enrichment at Class I sites, but the enrichment is significantly higher than background (Figure 3-4B, top panel). In mitosis, there is no significant enrichment for Myc at Class I sites (Figure 3-4B, bottom panel). The mechanism by which Myc persists at only a subset of interphase sites may depend on the specific recognition sequence present at each class of sites. We therefore examined potential differences in the consensus motif at Class I and Class II sites. E boxes can be found in about 70% of Myc sites in either class, but the two classes show different preferences for specific sequences. Myc sites in Class I preferentially contain the canonical E box (CATGTG/CACGTG). In contrast, Myc sites in Class II sites are depleted of the canonical E box and, instead, show enrichment for the non-canonical E box (Figure 3-4C). The preference in utilization of each E box type is significant (chi square $p<0.0001$) and may represent the underlying mechanism to select Myc sites that will be maintained during the cell cycle.

The two classes of Myc sites may have different roles in gene expression

To gain insights into possible functional differences between interphase-specific and mitotic Myc sites, we first performed GO analysis for genes associated with each class. Class I Myc sites are enriched at genes involved in ribosome biogenesis, which is important for biomass accumulation in G1. This is consistent with reports for Myc-

regulated genes in interphase cells in mammals (van Riggelen et al. 2010). However, genes associated with Class II Myc sites are not enriched for this category. Instead, there is a higher enrichment for cell cycle or developmental genes (Figure 3-5A). In addition to their presence at different target genes, the two groups of Myc sites may also affect gene expression by different mechanisms. In interphase cells, Myc-associated genes generally show higher pausing index than the average gene in the genome. We then parsed genes into two groups based on their association with Class I or Class II Myc sites. The results suggest that genes associated with Class I Myc sites (interphase-specific) still show a high pausing index, whereas genes associated with Class II Myc sites (those also present in mitosis) have lower pausing indexes (K-S test, $p=2x10^{-6}$) (Figure 3-5B). Class I Myc sites may then be involved in the release of paused RNAPII for productive elongation in interphase cells, whereas Class II Myc sites may play a different regulatory function on transcription that is dependent on the presence of Myc protein in mitotic chromosomes.

Mitotic Myc sites are present at a subset of promoters but not enhancers

To further explore functional differences between the two classes of Myc sites, we clustered all the sites with histone modifications characteristic of enhancers and promoters using *k*-means clustering. The results reveal 5 clusters of Myc sites (Figure 3-5C). Class I Myc sites are present in three different clusters whereas class II Myc sites associate with two clusters. Class I sites are present at enhancers (Cluster I), promoters (Cluster III) and a cluster lacking either characteristic (Cluster II). Class II sites are

present at promoters (Cluster IV) and a cluster of unknown function (Cluster V). Therefore, Myc is present at enhancers only during interphase and persists during mitosis at only a specific subset of all promoters occupied during interphase (Figure 3-5C). Interestingly, Myc-associated promoters in both interphase and mitosis appear to cluster in two groups with high or low levels of H3K4me3. In *Drosophila*, enhancers defined as sequences enriched in H3K27ac and H3K4me1 but lacking H3K4me3, are typically found within intronic regions (Kharchenko et al. 2011). Consistent with the clustering results, 55% of Class I non-promoter Myc sites are in introns while 70% of Class II non-promoter Myc sites fall in intergenic regions ($p<0.0001$) (Figure 3-5D). These results suggest that Class II Myc sites do not function as enhancers and they may play a different role in the genome independent of transcription. The possibility of a different role for Class II sites is supported by the observation that these sites are further apart from each other compared to Class I sites (Figure 3-5E).

Myc sites of unknown function associate with insulators

A subset of Myc sites in both Class I and Class II are not present at either enhancers or promoters. A third type of regulatory sequences found in eukaryotic cells is represented by insulators, which mediate long-range interactions between different sites in the genome. To test whether the Myc sites of unknown function are present at insulators, we parsed ChIP-seq datasets of *Drosophila* insulator proteins BEAF-32, dCTCF, Su(Hw), GAF and CP190 with the clusters shown in Figure 3-5C. The results

show a dramatic difference between Class I and Class II sites (Figure 3-6A). Class I sites associate preferentially with GAF, both at enhancers and promoters where this protein has been shown to be present (Negre et al. 2011), as well as in Cluster II containing sites not present at these two types of regulatory sequences. A subset of interphase-specific Class I promoter sites present in Cluster III, those containing high levels of H3K4me3 and presumably actively transcribed, contain BEAF-32 instead of GAF. Class II sites, on the other hand, associate with insulator proteins other than GAF (Figure 3-6A). In particular, all Class II sites, including those in Cluster V, contain all four insulator proteins tested, Su(Hw), BEAF-32, dCTCF and CP190 (Figure 3-6A and 6B).

Myc mitotic sites associate with mitotic insulator sites

The results presented above suggest a strong association between Class II Myc sites and sites of specific insulator proteins from interphase cells. Since Class II sites persist during mitosis, we wondered whether insulator proteins also remain at these sites during mitosis. To test this possibility, we compared the distribution of Class II Myc sites with datasets of insulator protein localization in mitotic chromosomes (Gurudatta et al. 2012b). The results indicate that Myc overlaps extensively with insulator proteins during mitosis (Figure 3-6C). All mitotic Myc sites contain dCTCF, DREF and CP190, and a subset also contains BEAF-32. A fifth insulator protein, Su(Hw), is not present in chromosomes during mitosis (Gurudatta et al. 2012b). Interestingly, a subset of the sites

where Myc and Orc2 co-localize during interphase are sites where Myc persists during mitosis (Figure 3-6C).

Mitotic Myc sites are enriched at the borders of topological chromosomal domains

The role of Myc during mitosis may be local i.e. to mark a subset of promoters or origins of replication for rapid resumption of transcription or assembly of the pre-replication complex at the beginning of G1. Alternatively, Myc may play a more global role in chromatin organization. Recent work suggests that eukaryotic chromosomes during interphase are organized into topological domains, characterized by high frequency of interactions, and separated by domain borders (Lieberman-Aiden et al. 2009; Dixon et al. 2012; Hou et al. 2012; Nora et al. 2012; Sexton et al. 2012). These borders are enriched in insulator proteins, which may contribute to the formation of boundaries that separate topological domains. It is possible that some of this organization persist during mitosis, and that insulator proteins contribute to the maintenance of chromosome architecture during the cell cycle. The fact that Myc persists at the same genomic sites as insulator proteins in mitotic chromosomes suggest that it may also be present at domain borders. To test this hypothesis, we compared the distribution of Class I and Class II Myc sites with respect to domain borders previously defined in *Drosophila* embryonic nuclei (Sexton et al. 2012). The results suggest that this is indeed the case (Figure 3-6D). Class II Myc sites that remain on chromosomes during mitosis are significantly enriched at domain borders, whereas interphase-specific Class II sites are significantly enriched

inside domains. These results could be interpreted to suggest that a specific subset of Myc sites may remain bound to chromosomes during mitosis to organize the higher order structure of chromatin. Alternatively, the presence of Myc at domain borders may be a consequence, rather than a cause, of chromosome organization. We have recently shown (Hou et al. 2012) that domain boundaries are more open and accessible than the interior of domains, suggesting that Myc may remain bound to these sequences during mitosis because of their higher accessibility.

Discussion

Myc is a bHLH-Zip sequence-specific DNA binding protein that plays a crucial role in the regulation of critical cellular processes such as cell growth, cell division and cell differentiation. Importantly, disruption of Myc levels in the cell lead to oncogenic transformation (Dang 2012). Since Myc interacts with DNA in a sequence-specific manner, its role in these various processes has been explained based on its ability to control the expression of specific genes by activation or repression of transcription. The effects of Myc in transcription and replication have been rationalized on the basis of its involvement in the control of promoter-proximal pausing of RNAPII and its effects on chromatin structure at the level of histone covalent modifications. Myc can induce H4 acetylation (Frye et al. 2007; Swarnalatha et al. 2012), which correlates with an increase of H4K20me2 and a transient increase of H4K20me1 (Frye et al. 2007). H4K20me1 can function at the crossroad of genome integrity, cell cycle, and transcription (Beck et al. 2012), and H4K20me2 is recognized by Orc1, which is a component of the Orc complex mediating pre-DNA replication licensing. The bromo adjacent homology (BAH) domain of Orc1 specifically recognizes H4K20me2, a property common to BAH domains present within diverse metazoan Orc1 proteins(Kuo et al. 2012). The sole enzyme that catalyzes H4K20me1 is Setd8 (also known as PR-Set7 or KMT5a), which is an essential mediator of Myc-induced epidermal differentiation. Deletion of Setd8 in Myc-overexpressing skin cells blocks proliferation and differentiation (Driskell et al. 2012).

Although the ability of Myc to act as a sequence-specific transcription factor and elicit changes in the 10 nm chromatin fiber may account for many of its effects on cell

function, the finding of Myc in the proteome of mitotic chromosomes (Ohta et al. 2010) represents an interesting puzzle. It is possible that Myc persistence on chromatin during mitosis has no relevance to its role in nuclear biology. On the other hand, several aspects of the distribution of Myc in mitotic chromosomes offer tantalizing explanations for some of its effects on transcription and replication. By comparing Myc binding sites in cells at interphase and mitosis we find two distinct groups of Myc sites. Class I sites only harbor Myc during interphase but become devoid of this protein during mitosis. These sites are adjacent to genes involved ribosome biogenesis, which have been reported to be cell type and species independent Myc targets (Ji et al. 2011). In contrast, Class II Myc sites that persist during mitosis associate with genes that play roles in cell cycle or cell differentiation. In mammalian cells, this includes genes important for maintaining pluripotency and reprogramming (Ji et al. 2011). It is possible that the presence of Myc at these genes during mitosis serves to preserve epigenetic memory of their expression necessary for the maintenance of cell identity.

The striking overlap of Myc and insulator sites during mitosis points to a more complex role for this protein in mitotic chromatin. Insulators have been shown to mediate long-range intra- and inter-chromosomal interactions (Phillips and Corces 2009). Although the role of some of these interactions may be to regulate enhancer-promoter contacts, the finding of insulators at the boundaries of topological chromosome domains points to a larger and more complex function of these proteins in higher-order chromatin organization. The presence of Myc together with insulator proteins at these sites in mitotic chromosomes may explain some Myc-dependent phenotypes, including its effects

on genome integrity. Mouse induced pluripotent stem (iPS) cell lines induced with Myc show a significantly higher frequency of translocation than those induced without Myc (Chen et al. 2011). This result is Myc dependent, as deletion of Myc box II reduces the translocation frequency (Guffei et al. 2007). Myc overexpression also induces telomeric aggregation in the interphase nucleus (Louis et al. 2005). These effects of Myc on genomic integrity suggest that Myc may play a role in chromosome higher order structure that may depend on its presence at insulator sites during interphase and/or mitosis. This conclusion is further supported by the observation that mitotic Myc sites are enriched at the borders of topological chromosome domains, which are also enriched in insulator proteins. Domain boundaries are more accessible to the insertion of transposable elements and allow higher expression of transgenes, suggesting that they represent regions of the genome with more open higher-order chromatin (Hou et al. 2012). Together, these observations agree with a model by which insulators organize the chromatin in the interphase nucleus by mediating interactions that create chromosomal domains. Transition from interphase to mitosis involves a condensation of the chromatin that nevertheless maintains this organization via the persistence of insulator proteins at domain boundaries. The boundary regions contain more open chromatin that may become accessible to components of the transcription and replication apparatus earlier at the end of M phase. The maintenance of Myc at these domain boundaries may ensure that adjacent genes are transcribed early at the M/G1 transition and a subset of replication origins assemble pre-replication complexes by recruiting Orc2 and, perhaps, determining replication timing. Additional experiments will be necessary to test this speculative but plausible model.

Methods

Cell culture and flow cytometry

Drosophila Kc167cells were grown at 25^{o}C in CCM3 media (Hyclone) to a density of 2X10^{6}. Cells were synchronized and sorted as previously described (Gurudatta et al. 2012b). Briefly, cells were treated with hydroxyurea (1 mg/ml in ethanol to a final concentration of 15 ng/ml) for 16 hr, incubated for 8 hr with nocodazole (5 mg/ml in DMSO, to a final concentration of 2 ng/ml) and harvested. For flow cytometry, cells were fixed for 10 min in 1% formaldehyde, blocked in suspension for 30 min in blocking buffer, incubated overnight with rabbit α-H3S10ph at 1:5000 or mouse α-Lamin Dm0 at 1:500, washed 3 x 15 min in blocking buffer, and then incubated with secondary antibody Alexa Fluor 488 α-rabbit at 1:5000. After a 30 min incubation in blocking buffer plus propidium iodide (0.1 mg/ml), samples were passed several times through a 25-gauge syringe to reduce clumping and sorted on a FACSAria II cell sorter. Enrichment of the mitotic and interphase cell populations was carried out by visualization of the mitotic marker H3S10ph by immunofluorescence microscopy, showing 97-99% purity (Gurudatta et al. 2012b).

ChIP-seq analysis

ChIP was performed with ~4 × 10^{7} cells. Cells were cross-linked with 1% formaldehyde for 10 min at room temperature. Nuclear lysates were sonicated to generate 200-1000 bp

DNA fragments. ChIP was then performed with 6 μL of *Drosophila* α-Myc antibody (Santa Cruz Biotechnology, sc-28208). Libraries are prepared with the IlluminaTruSeq DNA Sample Preparation Kit. Fragments in the 200-300 bp ranged were selected and sequenced in an IlluminaHiSeq sequencer at the HudsonAlpha Institute for Biotechnology.

Bioinformatics analyses

Sequences were aligned to *Drosophila* dm3 using Bowtie. The output map files were converted to bed format for each chromosome arm using the VancouverShort package. Peaks were called using CCAT3.0 (Xu et al. 2010) with the enrichment parameter set to 15. Myc-associated genes were defined as genes with Myc binding sites between -200 bp and the TSS or in the 5'UTR region.

In addition to the *Drosophila* Myc data obtained in this study, we used several datasets obtained from public sources. Orc2 ChIP-seq (modENCODE_2755), RNAPIIChIP-chip (modENCODE_328) and RNA expression in Kc cells (modENCODE_3305) were obtained from modENCODE. ChIP-seq data sets for H3K4me1, H3K4me3 and H3K27ac are from GSE36374. ChIP-seq data sets for insulator proteins are from GSE30740, GSE32584 and GSE39664. To build heatmaps, values for each ChIP-seq dataset were extracted for the 2000 bp region around the summit of peaks using custom R scripts (available upon request) and heatmap graphs were created using TreeView(Saldanha 2004). Clusters in Figures 3-5C and 3-6A were created by *k*-means clustering using Cluster3.0 based on the mean values of the 300 bp around Myc sites for the samples listed.

The pausing index of genes was calculated using ChIP-chip data sets of RNAPII in Kc cells obtained from modENCODE. RNAPII at TSSs (P_{TSS}) was calculated as the mean enrichment of RNAPII at the 200 bp region around each TSS. RNAPII in the gene body (P_{body}) was calculated as the mean enrichment of RNAPII from +200 bp to the end of the gene. The pausing index is defined as the different between the P_{TSS} and P_{body}. Motif analysis of Myc binding sites was performed using Myc peak summits extend 50 bp on either side. The resulting 100 bp sequence for each peak was used to search for E boxes using a custom Perl script available upon request. Gene ontology analysis for Myc associated genes was performed with DAVID (http://david.abcc.ncifcrf.gov). Flybase IDs were used to determine statistically enriched biological process categories on the basis of a background list of all annotated genes in the *Drosophila* genome.

Data access

Sequence data have been deposited in NCBI's Gene Expression Omnibus (GEO) under accession number GSE39521.

Acknowledgements

We would like to thank members of the lab for helpful discussions and suggestions during this study. We also thank The Genomic Services Lab at the HudsonAlpha Institute for Biotechnology for their help in performing Illumina sequencing of ChIP-seq samples. Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R01GM035463. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Figure 3-1. Characteristics of Myc-associated genes.

(A) Examples of Myc-associated genes. The signal for Myc is represented as the number of raw reads from ChIP-seq and the signal for RNAPII is represented by the log2 enrichment from ChIP-chip. On the gene track, the genes above the line are transcribed from the plus strand and the genes below the line are transcribed from the minus strand. The two arrows on the left point to two protein coding genes associated with Myc that also show high accumulation of RNAPII at the TSS. The two arrows on the right point to two non protein-coding microRNA genes associated with Myc. (B) Genome wide distribution of Myc binding sites with respect to various gene landmarks. Distal intergenic region means regions that are at least 200 bp away from genes. (C) Cumulative curve of pausing index for all coding genes in the genome (black) or coding genes associated with Myc (green). (D) Distribution of expression levels of Myc target genes. All genes in the genome were sorted according to their expression score and binned into five groups (Group 1 with the lowest expression and Group 5 with the highest expression). Myc target genes were assigned to one of the groups if their expression scores fall into the range of expression levels for that group.

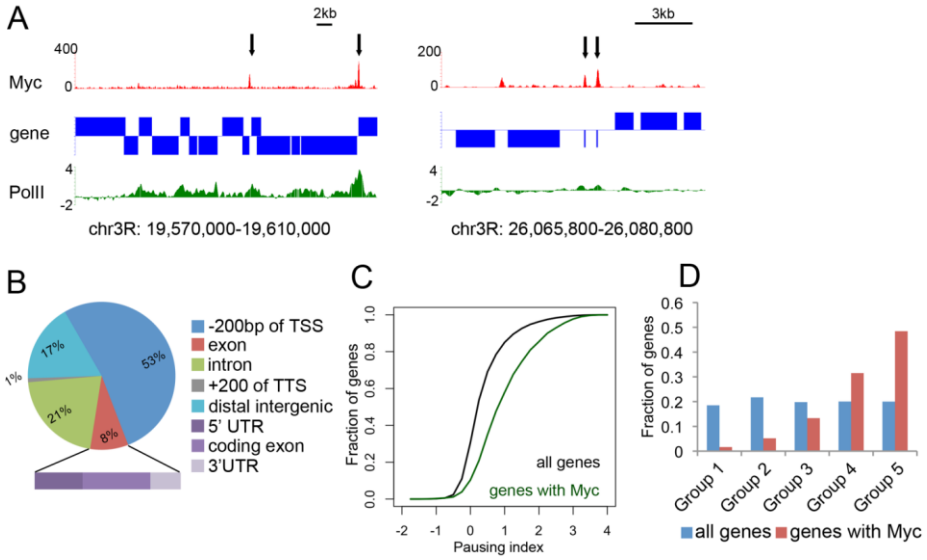

A
2kb
3kb
400
0
Myc
gene
4
-2
PolII
200
chr3R: 19,570,000-19,610,000
chr3R: 26,065,800-26,080,800
B
17%
53%
1%
21%
8%
-200bp of TSS
exon
intron
+200 of TTS
distal intergenic
5' UTR
coding exon
3'UTR
C
Fraction of genes
all genes
genes with Myc
Pausing index
D
Fraction of genes
Group 1
Group 2
Group 3
Group 4
Group 5
all genes
genes with Myc

Figure 3-2. A subset of Myc sites at non-promoter regions have characteristics of enhancers.

(A) Examples of non-promoter Myc sites in the genome. The signals on the tracks of Myc, H3K4me1, H3K27ac and H3K4me3, are represented by the raw reads from ChIP-seq. The signal for RNAPII is represented by the log2 value from ChIP-chip. The arrows represent two Myc sites that display enhancer chromatin signatures (presence of H3K4me1/H3K27ac and absence of H3K4me3). (B) Heatmaps showing the chromatin features at promoter and non-promoter Myc sites. Each panel represents 2 kb upstream and downstream of the Myc sites. The sites are ordered by signal of H3K4me1. (C) Heatmaps showing chromatin features at all identified enhancers in Kc cells. The sites are ordered by signal of Myc.

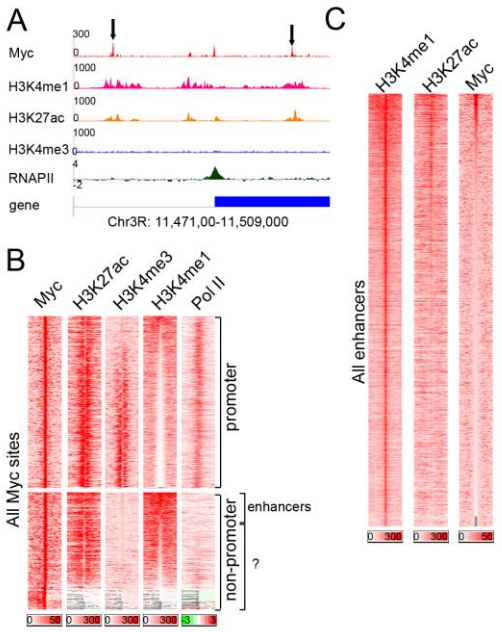

A
Myc
H3K4me1
H3K27ac
H3K4me3
RNAPII
gene
Chr3R: 11,471,00-11,509,000
B
Myc
H3K27ac
H3K4me3
H3K4me1
Pol II
All Myc sites
promoter
non-promoter
enhancers
?
C
H3K4me1
H3K27ac
Myc
All enhancers

Figure 3-3. Myc associates with Orc2 genome wide.

(A) Snapshot of a region of the *Drosophila* genome showing the distribution of Orc2 sites compared with Myc. The signals represent the number of raw reads from ChIP-seq data sets across the region. (B) Heatmaps of Myc and Orc2 signal at all Myc and Orc2 binding sites. Each panel represents 2 kb upstream and downstream of the anchor sites. The two panels on the left are the signals at all Myc binding sites in Kc cells discovered in this study ordered by Orc2 signal intensity. The two panels on the right are the signals at all Orc2 binding sites obtained from modENCODE ordered by Myc signal. (C) Correlation between the intensities of Orc2 and Myc at all Myc sites. (D) Correlation between the intensities of Orc2 and Myc at all Orc2 sites.

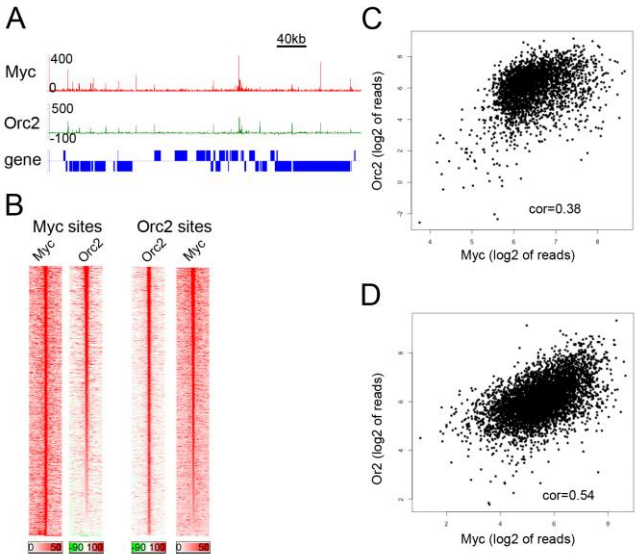

A
40kb
Myc
400
0
Orc2
500
-100
gene
B
Myc sites
Orc2 sites
Myc
Orc2
Orc2
Myc
C
Orc2 (log2 of reads)
cor=0.38
Myc (log2 of reads)
D
Or2 (log2 of reads)
cor=0.54
Myc (log2 of reads)

Figure 3-4. Properties of Myc sites in interphase and mitotic chromosomes.

(A) Heatmap showing signals of Myc in interphase or mitosis at all the Myc site in the genome. The information indicates the existence of two groups of Myc sites in the genome, one is interphase-specific (Class I) and the second one is common to interphase and mitosis (Class II). (B) Binding intensity at Myc sites during interphase or mitosis plotted from the information displayed in panel A. The X axis represents distance from Myc sites and '0' is the summit of Myc sites. Negative values indicate upstream and positive values indicate downstream of the Myc sites. (C) Usage of different binding motifs by the Myc protein in interphase (Class I) or mitosis (Class II).

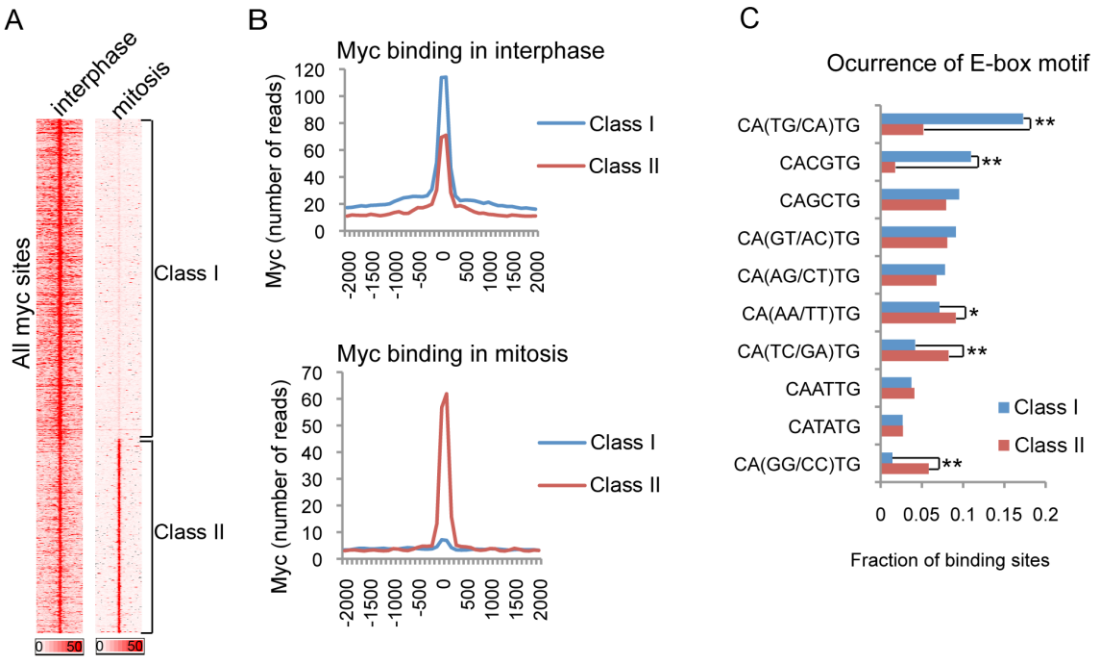

A
interphase
mitosis
All myc sites
Class I
Class II
0
50
0
50
B
Myc binding in interphase
Myc (number of reads)
120
100
80
60
40
20
0
-2000
-1500
-1000
-500
0
500
1000
1500
2000
Class I
Class II
Myc binding in mitosis
Myc (number of reads)
70
60
50
40
30
20
10
0
-2000
-1500
-1000
-500
0
500
1000
1500
2000
Class I
Class II
C
Ocurrence of E-box motif
CA(TG/CA)TG
CACGTG
CAGCTG
CA(GT/AC)TG
CA(AG/CT)TG
CA(AA/TT)TG
CA(TC/GA)TG
CAATTG
CATATG
CA(GG/CC)TG
**
**
*
**
**
Class I
Class II
0
0.05
0.1
0.15
0.2
Fraction of binding sites

Figure 3-5. Myc sites occupied during interphase and mitosis have different characteristics.

(A) Gene ontology of protein coding genes associated with the two classes of Myc sites. (B) Cumulative curve of pausing index for protein coding genes associated with Class I (red) or Class II (blue) Myc sites. (C) Heatmaps showing chromatin features of Class I and Class II Myc sites. Each panel represents 2 kb upstream or downstream of the Myc sites. Clusters were created using Cluster 3.0 based on the signal value for the listed features at the Myc sites. (D) Distribution of Class I and Class II non-promoter Myc sites in introns or intergenic regions. (E) Distance between Myc site pairs for Class I and Class II Myc sites.

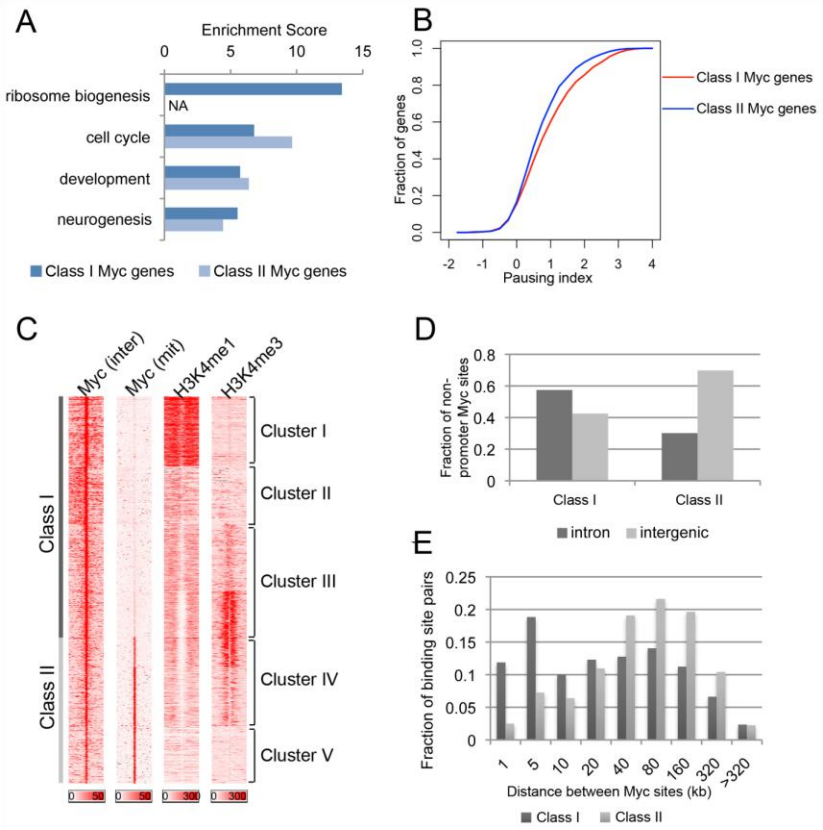
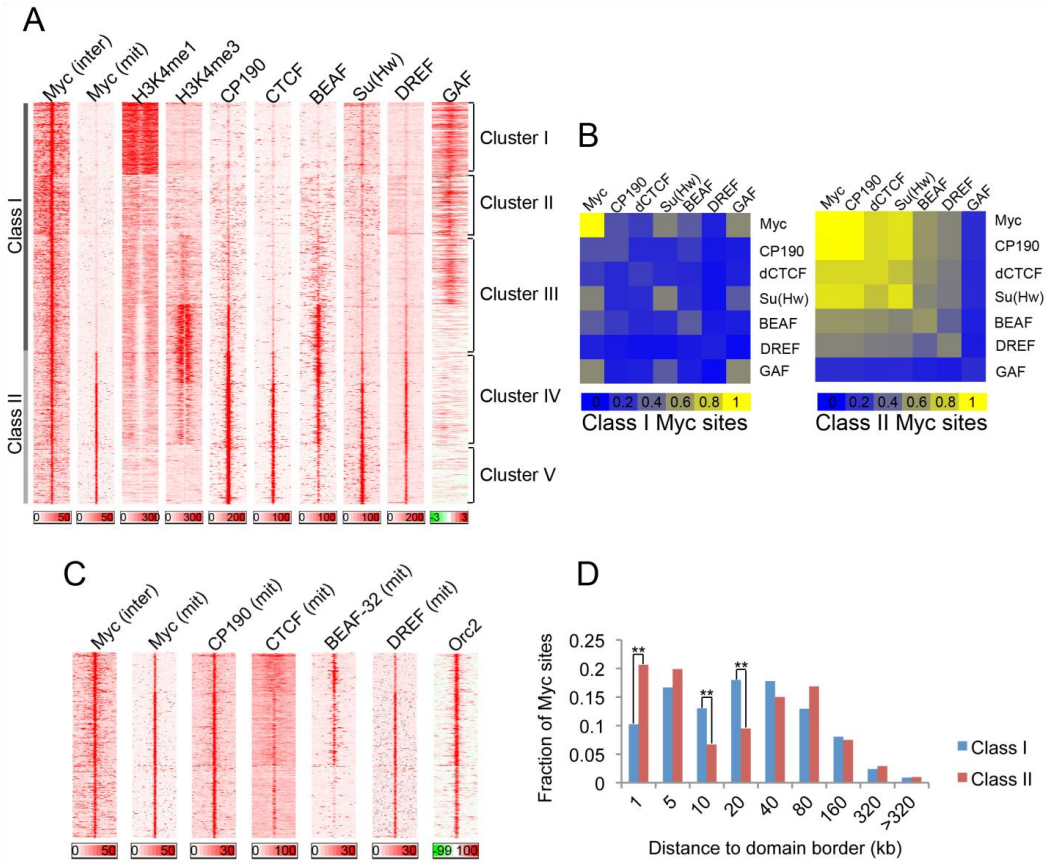

A
Enrichment Score
0
5
10
15
ribosome biogenesis
NA
cell cycle
development
neurogenesis
Class I Myc genes
Class II Myc genes
B
Fraction of genes
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
3
4
Pausing index
Class I Myc genes
Class II Myc genes
C
Myc (inter)
Myc (mit)
H3K4me1
H3K4me3
Class I
Class II
Cluster I
Cluster II
Cluster III
Cluster IV
Cluster V
0
50
0
50
0
300
0
300
D
Fraction of non-promoter Myc sites
0
0.2
0.4
0.6
0.8
Class I
Class II
intron
intergenic
E
Fraction of binding site pairs
0
0.05
0.1
0.15
0.2
0.25
1
5
10
20
40
80
160
320
>320
Distance between Myc sites (kb)
Class I
Class II

Figure 3-6. Myc sites present in mitotic chromosomes associate with insulator proteins.

(A) Heatmaps showing chromatin features at Class I and Class II Myc sites. Each panel represents 2 kb upstream or downstream of the Myc sites. Signals are represented by the number of raw reads from ChIP-seq or enrichment log2 value from ChIP-chip for the listed proteins or histone modifications. (B) Heatmaps showing the overlap of the listed protein at the two classes of Myc sites. (C) Insulator proteins and Myc persist on chromatin during mitosis at Class II Myc sties. (D) Distribution of Class I and Class II Myc sites with respect to topological domain boundaries.


A
Myc (inter)
Myc (mit)
H3K4me1
H3K4me3
CP190
CTCF
BEAF
Su(Hw)
DREF
GAF
Class I
Class II
Cluster I
Cluster II
Cluster III
Cluster IV
Cluster V
B
Myc
CP190
dCTCF
Su(Hw)
BEAF
DREF
GAF
0 0.2 0.4 0.6 0.8 1
Class I Myc sites
Class II Myc sites
C
Myc (inter)
Myc (mit)
CP190 (mit)
CTCF (mit)
BEAF-32 (mit)
DREF (mit)
Orc2
D
Fraction of Myc sites
0.25
0.2
0.15
0.1
0.05
0
**
**
**
1
5
10
20
40
80
160
320
>320
Distance to domain border (kb)
Class I
Class II

Chapter 4

Discussion and future directions

Chromatin architecture and genome function

It is becoming increasingly evident that the genetic material is arranged in the nucleus of eukaryotic cells in a non-random fashion. The eukaryotic genome is arranged in the three-dimensional nuclear space by interactions between different types of proteins that bring together distant genomic sequences. An important question arising from this concept is whether this 3D arrangement plays exclusively a structural role than determines function or whether the structure is a consequence of the functional output of the genome. Current evidence appears to suggest that a combination of these two processes may cooperate to establish the 3D organization of the genome and that this arrangement may be cell type-specific and may carry epigenetic information.

Enhancers activate gene expression by looping out intervening sequences and contacting the promoters of their target genes. In addition, enhancers appear to also recruit gene promoters to transcription factories where genes are co-regulated. Therefore, these two types of interactions may represent examples of processes where function gives rise to structure in the nucleus. Insulators, on the other hand, may represent examples of sequences whose primary role is to mediate contacts between distant sequences to affect genome function. It is becoming increasingly clear that, in addition to their original role in interfering with enhancer-promoter interactions, insulators can actually tether enhancers to their target promoters. Therefore, insulators may be responsible for a subset of the 3D organization of the genome that determines its functional outcome. In addition, recent evidence suggests that a subset of insulator sites may have a purely structural role and mediate a subset of interactions that are conserved during mitosis. These insulator

sites, called “aligned insulators”, contain binding sites for several insulator proteins within a 200-300 bp region and they may represent specially strong insulator sites whose function is to maintain chromosome structure throughout the cell cycle. Work described here indicates that, in addition to insulator proteins, these sites are also enriched for Myc, which until now has been considered a classical transcription factor.

As I have described above, the organization of the chromatin in the nucleus is established by different factors that contribute to this process through their participation in various aspects of genome function. For example, transcription factors such as ER-alpha function by tethering regulatory elements to gene promoters to initiate transcription, or bring coordinately regulated genes together (Fullwood et al. 2009). On the other hand, insulator proteins like CTCF may mediate a subset of interactions between distant sequences in the genome in order to separate differentially regulated genes into separate domains (Handoko et al. 2011). *Drosophila* has at least four different types of insulators and it is not known whether they play similar or distinct roles in chromatin organization and gene regulation. Here I have taken an evolutionary approach to dissect the role of the BEAF-32 insulator in nuclear biology. BEAF-32 has been previously described as a protein whose role is to facilitate high levels of transcription (Jiang et al. 2009). Mutation of BEAF-32 results in down-regulation of the expression of some genes and up-regulation of others (Gurudatta et al. 2012a). The work described here helps explain this apparently dual role of BEAF-32, suggesting that this protein acts as a *cis* regulatory element that separates close head-to-head genes with different transcription regulation modes. The mechanisms by which BEAF-32 performs this function are

probably similar to those used by other insulator proteins such as mammalian CTCF. One hypothesis to explain the role of BEAF-32 in transcription is that it can separate the two adjacent genes into two different functional domains via interactions with other insulators located in the vicinity. The domains formed by each of the two loops may compartmentalize regulatory sequences from adjacent promoters to insure independent regulation of the two genes. Additional experiments will be necessary to test this speculative but plausible model.

Inheritance of chromatin higher-order structure

In interphase, chromatin architecture depends on all the interactions mediated by various chromatin associated factors, including enhancers and insulators. When the cells enter mitosis, most of these factors are removed from chromosome while only a few of the proteins persist on the chromosomes (O'Donovan et al. 2010; Ohta et al. 2010). Thus, some of the interactions will be disrupted, since the factors mediating these contacts are no longer bound to chromatin. Nevertheless, the mother cell needs to transmit epigenetic information to daughter cells to maintain cell identity. How can then cells precisely remember the chromatin architecture present during G1 and recover this information in the following interphase? Proteins retained on chromatin during mitosis are reasonable candidates to play a role in epigenetic memory transmitted throughout the cell cycle. Myc and insulator proteins CTCF/BEAF-32/CP190 are good candidates to play a role in this process, since they persist on mitotic chromosomes (O'Donovan et al. 2010; Ohta et al.

2010; Gurudatta et al. 2012b). Both Myc and insulator proteins have been associated with different types of chromatin interactions (Handoko et al. 2011; Lin et al. 2012). Here, I have described results indicating that Myc stays bound in mitotic chromosomes at sites where it overlaps with insulator proteins. The role of Myc at these sites is unclear at this time. Aligned insulator sites that persist during mitosis are preferentially located at the borders of topological domains defined by Hi-C. These domain boundaries appear to represent more accessible regions of chromatin. Therefore, it is possible that these regions become de-condensed earlier during the M/G1 transition. It is possible that Myc contributes to the maintenance of an open chromatin structure at these regions. Alternatively, the presence of Myc at these sites may ensure transcription of adjacent genes at the end of mitosis.

Chromatin architecture and evolution

It is clear from the previous discussion that chromatin interactions contribute to the genome function. As a consequence, regulation of higher-order chromatin structure may represent a strategy for the organism to re-program transcription and adapt to new environmental conditions. For example, latency of Epstein-Barr Virus is controlled via chromatin interactions that mediate crosstalk between a distal enhancer and different promoters to start distinct transcription programs (Chau et al. 2006; Tempera et al. 2010; Tempera et al. 2011). By regulating chromatin architecture, the virus can select when to exit latency in order to take advantage of new environmental situations. More complex

eukaryotic organisms with larger genomes use similar strategies to respond to environmental cues. These alterations in chromatin organization ensure that changes in transcription are confined to specific regions without affecting the normal function of the rest of the genome. For example, hormone treatment can induce chromatin interactions that limit the impact to hormone responding genes without changing the expression of surrounding genes (Wood et al. 2011).

Changes in the three-dimensional arrangement of chromatin may represent an efficient strategy to control the transcription program of a cell or organism in order to adapt to new environmental conditions. The change on chromatin architecture may contribute to the phenotypic variations within species. These phenotypic variations could provide resources for selection to shape evolution. It is possible that many nonsense single-nucleotide polymorphisms (SNPs) observed within or between populations do not alter protein coding regions but they may affect chromatin high order structure and change the levels or developmental timing of genes. Analyses of possible effects of these SNPs on the 3D arrangement of the chromatin may represent an important avenue for future research in the evolution field.

Reference

Adachi N, Lieber MR. 2002. Bidirectional Gene Organization: A Common Architectural Feature of the Human Genome. *Cell* **109**(7): 807-809.

Aguda BD, Kim Y, Piper-Hunter MG, Friedman A, Marsh CB. 2008. MicroRNA regulation of a cancer network: Consequences of the feedback loops involving miR-17-92, E2F, and Myc. *Proceedings of the National Academy of Sciences* **105**(50): 19678-19683.

Bannister AJ, Kouzarides T. 2011. Regulation of chromatin by histone modifications. *Cell Res* **21**(3): 381-395.

Bantignies F, Roure V, Comet I, Leblanc B, Schuettengruber B, Bonnet J, Tixier V, Mas A, Cavalli G. 2011. Polycomb-Dependent Regulatory Contacts between Distant Hox Loci in Drosophila. *Cell* **144**(2): 214-226.

Batada NN, Hurst LD. 2007. Evolution of chromosome organization driven by selection for reduced gene expression noise. *Nat Genet* **39**(8): 945-949.

Beck DB, Oda H, Shen SS, Reinberg D. 2012. PR-Set7 and H4K20me1: at the crossroads of genome integrity, cell cycle, chromosome condensation, and transcription. *Genes & Development* **26**(4): 325-337.

Blackwell T, Kretzner L, Blackwood E, Eisenman R, Weintraub H. 1990. Sequence-specific DNA binding by the c-Myc protein. *Science* **250**(4984): 1149-1151.

Blackwell TK, Huang J, Ma A, Kretzner L, Alt FW, Eisenman RN, Weintraub H. 1993. Binding of myc proteins to canonical and noncanonical DNA sequences. *Molecular and Cellular Biology* **13**(9): 5216-5224.

Borneman AR, Gianoulis TA, Zhang ZD, Yu H, Rozowsky J, Seringhaus MR, Wang LY, Gerstein M, Snyder M. 2007. Divergence of Transcription Factor Binding Sites Across Related Yeast Species. *Science* **317**(5839): 815-819.

Brogiolo W, Stocker H, Ikeya T, Rintelen F, Fernandez R, Hafen E. 2001. An evolutionarily conserved function of the Drosophila insulin receptor and insulin-like peptides in growth control. *Current Biology* **11**(4): 213-221.

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

Byrd K, Corces VG. 2003. Visualization of chromatin domains created by the gypsy insulator of Drosophila. *The Journal of Cell Biology* **162**(4): 565-574.

Carreira VP, Mensch J, Fanara JJ. 2008. Body size in Drosophila: genetic architecture, allometries and sexual dimorphism. *Heredity* **102**(3): 246-256.

Carrière L, Graziani S, Alibert O, Ghavi-Helm Y, Boussouar F, Humbertclaude H, Jounier S, Aude J-C, Keime C, Murvai J et al. 2011. Genomic binding of Pol III transcription machinery and relationship with TFIIS transcription factor distribution in mouse embryonic stem cells. *Nucleic Acids Research*.

Chang T-C, Yu D, Lee Y-S, Wentzel EA, Arking DE, West KM, Dang CV, Thomas-Tikhonenko A, Mendell JT. 2008. Widespread microRNA repression by Myc contributes to tumorigenesis. *Nat Genet* **40**(1): 43-50.

Chau CM, Zhang X-Y, McMahon SB, Lieberman PM. 2006. Regulation of Epstein-Barr Virus Latency Type by the Chromatin Boundary Factor CTCF. *J Virol* **80**(12): 5723-5732.

Chen Q, Shi X, Rudolph C, Yu Y, Zhang D, Zhao X, Mai S, Wang G, Schlegelberger B, Shi Q. 2011. Recurrent trisomy and Robertsonian translocation of chromosome 14 in murine iPS cell lines. *Chromosome Research* **19**(7): 857-868.

Chess A. 2012. Mechanisms and consequences of widespread random monoallelic expression. *Nat Rev Genet* **13**(6): 421-428.

Cleard F, Moshkin Y, Karch F, Maeda RK. 2006. Probing long-distance regulatory interactions in the Drosophila melanogaster bithorax complex using Dam identification. *Nat Genet* **38**(8): 931-935.

Comet I, Schuettengruber B, Sexton T, Cavalli G. 2011. 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* **108**(6): 2294-2299.

Conrad T, Akhtar A. 2012. Dosage compensation in Drosophila melanogaster: epigenetic fine-tuning of chromosome-wide transcription. *Nat Rev Genet* **13**(2): 123-134.

Conrad T, Cavalli Florence MG, Holz H, Hallacli E, Kind J, Ilik I, Vaquerizas Juan M, Luscombe Nicholas M, Akhtar A. 2012a. The MOF Chromobarrel Domain Controls Genome-wide H4K16 Acetylation and Spreading of the MSL Complex. *Developmental Cell* **22**(3): 610-624.

Conrad T, Cavalli FMG, Vaquerizas JM, Luscombe NM, Akhtar A. 2012b. Drosophila Dosage Compensation Involves Enhanced Pol II Recruitment to Male X-Linked Promoters. *Science*.

Consortium DG. 2007. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**(7167): 203-218.

Consortium Tm, Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L et al. 2010. Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE. *Science* **330**(6012): 1787-1797.

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

Dang Chi V. 2012. MYC on the Path to Cancer. *Cell* **149**(1): 22-35.

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

Dhadi SR, Krom N, Ramakrishna W. 2009. Genome-wide comparative analysis of putative bidirectional promoters from rice, Arabidopsis and Populus. *Gene* **429**(1–2): 65-73.

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

Dominguez-Sola D, Ying CY, Grandori C, Ruggiero L, Chen B, Li M, Galloway DA, Gu W, Gautier J, Dalla-Favera R. 2007. Non-transcriptional control of DNA replication by c-Myc. *Nature* **448**(7152): 445-451.

Driskell I, Oda H, Blanco S, Nascimento E, Humphreys P, Frye M. 2012. The histone methyltransferase Setd8 acts in concert with c-Myc and is required to maintain skin. *EMBO J* **31**(3): 616-629.

Essafi A, Webb A, Berry RL, Slight J, Burn SF, Spraggon L, Velecela V, Martinez-Estrada OM, Wiltshire JH, Roberts SGE et al. 2011. A Wt1-controlled chromatin switching mechanism underpins tissue-specific Wnt4 activation and repression. *Developmental Cell* **21**.

Filion GJ, van Bemmel JG, Braunschweig U, Talhout W, Kind J, Ward LD, Brugman W, de Castro IJ, Kerkhoven RM, Bussemaker HJ et al. 2010. Systematic Protein Location Mapping Reveals Five Principal Chromatin Types in Drosophila Cells. *Cell* **143**(2): 212-224.

Frye M, Fisher AG, Watt FM. 2007. Epidermal Stem Cells Are Defined by Global Histone Modifications that Are Altered by Myc-Induced Differentiation. *PLoS ONE* **2**(8): e763.

Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH et al. 2009. An oestrogen-receptor-[agr]-bound human chromatin interactome. *Nature* **462**(7269): 58-64.

Fussner E, Strauss M, Djuric U, Li R, Ahmed K, Hart M, Ellis J, Bazett-Jones DP. 2012. Open and closed domains in the mouse genome are configured as 10-nm chromatin fibres. *EMBO Rep* **advance online publication**.

Gargano B, Amente S, Majello B, Lania L. 2007. P-TEFb is a Crucial Co-Factor for Myc Transactivation. *Cell Cycle* **6**(16): 2031-2037.

Gerasimova TI, Byrd K, Corces VG. 2000. A Chromatin Insulator Determines the Nuclear Localization of DNA. *Molecular Cell* **6**(5): 1025-1035.

Gerasimova TI, Lei EP, Bushey AM, Corces VG. 2007. Coordinated Control of dCTCF and gypsy Chromatin Insulators in Drosophila. *Molecular Cell* **28**(5): 761-772.

Guccione E, Martinato F, Finocchiaro G, Luzi L, Tizzoni L, Dall' Olio V, Zardo G, Nervi C, Bernard L, Amati B. 2006. Myc-binding-site recognition in the human genome is determined by chromatin context. *Nat Cell Biol* **8**(7): 764-770.

Guffei A, Lichtensztejn Z, Goncalves Dos Santos Silva A, Louis SF, Caporali A, Mai S. 2007. c-Myc-dependent formation of Robertsonian translocation chromosomes in mouse cells. *Neoplasia (New York, NY)* **9**(7): 578-588.

Gurudatta BV, Ramos E, Corces VG. 2012a. The BEAF insulator regulates genes involved in cell polarity and neoplastic growth. *Developmental Biology* **369**(1): 124-132.

Gurudatta BV, Sung ER, Yang J, Bryant D, Donlin-Asp PG, Van Bortle K, Corces VG. 2012b. Drosophila insulator proteins persist in mitotic chromosomes where they may mark sites for pre-replication complex assembly and gene bookmarking. *Genome Research* **submitted**.

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

Hall BK, Hallgrímsson B. 2008. *Strickberger's Evolution*.

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

He Q, Bardet AF, Patton B, Purvis J, Johnston J, Paulson A, Gogol M, Stark A, Zeitlinger J. 2011. High conservation of transcription factor binding and evidence for combinatorial regulation across six Drosophila species. *Nat Genet* **43**(5): 414-420.

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

Herr DR, Fyrst H, Creason MB, Phan VH, Saba JD, Harris GL. 2004. Characterization of the Drosophila Sphingosine Kinases and Requirement for Sk2 in Normal Reproductive Function. *Journal of Biological Chemistry* **279**(13): 12685-12694.

Herr DR, Fyrst H, Phan V, Heinecke K, Georges R, Harris GL, Saba JD. 2003. Sply regulation of sphingolipid signaling molecules is essential for Drosophila development. *Development* **130**(11): 2443-2453.

Herr DR, Harris GL. 2004. Close head-to-head juxtaposition of genes favors their coordinate regulation in Drosophila melanogaster. *FEBS Letters* **572**(1–3): 147-153.

Hon G, Ren B, Wang W. 2008. ChromaSig: A Probabilistic Approach to Finding Common Chromatin Signatures in the Human Genome. *PLoS Comput Biol* **4**(10): e1000201.

Hou C, Li L, Qin Z, Corces VG. 2012. Gene Density, Transcription and Insulators Contribute to the Partition of the Drosophila Genome into Physical Domains. *Molecular Cell* **submitted**.

Hurst LD, Pal C, Lercher MJ. 2004. The evolutionary dynamics of eukaryotic gene order. *Nat Rev Genet* **5**(4): 299-310.

Iwasaki O, Tanaka A, Tanizawa H, Grewal SIS, Noma K-i. 2010. Centromeric Localization of Dispersed Pol III Genes in Fission Yeast. *Molecular Biology of the Cell* **21**(2): 254-265.

Ji H, Wu G, Zhan X, Nolan A, Koh C, De Marzo A, Doan HM, Fan J, Cheadle C, Fallahi M et al. 2011. Cell-Type Independent MYC Target Genes Reveal a Primordial Signature Involved in Biomass Accumulation. *PLoS ONE* **6**(10): e26057.

Jiang N, Emberly E, Cuvier O, Hart CM. 2009. Genome-Wide Mapping of Boundary Element-Associated Factor (BEAF) Binding Sites in Drosophila melanogaster Links BEAF to Transcription. *Molecular and Cellular Biology* **29**(13): 3556-3568.

Jones PA. 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet* **13**(7): 484-492.

Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M et al. 2003. Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature* **421**(6920): 231-237.

Kanazawa S, Soucek L, Evan G, Okamoto T, Peterlin BM. 2003. c-Myc recruits P-TEFb for transcription, cellular proliferation and apoptosis. *Oncogene* **22**(36): 5707-5711.

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

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

Kidder BL, Yang J, Palmer S. 2008. Stat3 and c-Myc Genome-Wide Promoter Occupancy in Embryonic Stem Cells. *PLoS ONE* **3**(12): e3932.

Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, Green Roland D, Zhang MQ, Lobanenkov VV, Ren B. 2007. Analysis of the Vertebrate Insulator Protein CTCF-Binding Sites in the Human Genome. *Cell* **128**(6): 1231-1245.

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

Koyanagi KO, Hagiwara M, Itoh T, Gojobori T, Imanishi T. 2005. Comparative genomics of bidirectional gene pairs and its implications for the evolution of a transcriptional regulation system. *Gene* **353**(2): 169-176.

Kucharski R, Maleszka J, Foret S, Maleszka R. 2008. Nutritional Control of Reproductive Status in Honeybees via DNA Methylation. *Science* **319**(5871): 1827-1830.

Kuo AJ, Song J, Cheung P, Ishibe-Murakami S, Yamazoe S, Chen JK, Patel DJ, Gozani O. 2012. The BAH domain of ORC1 links H4K20me2 to DNA replication licensing and Meier-Gorlin syndrome. *Nature* **484**(7392): 115-119.

Kurukuti S, Tiwari VK, Tavoosidana G, Pugacheva E, Murrell A, Zhao Z, Lobanenkov V, Reik W, Ohlsson R. 2006. 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* **103**(28): 10684-10689.

Kyrchanova O, Toshchakov S, Podstreshnaya Y, Parshikov A, Georgiev P. 2008. Functional Interaction between the Fab-7 and Fab-8 Boundaries and the Upstream Promoter Region in the Drosophila Abd-B Gene. *Mol Cell Biol* **28**(12): 4188-4195.

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

Li Q, Dang CV. 1999. c-Myc Overexpression Uncouples DNA Replication from Mitosis. *Molecular and Cellular Biology* **19**(8): 5339-5351.

Li Y-Y, Yu H, Guo Z-M, Guo T-Q, Tu K, Li Y-X. 2006. Systematic Analysis of Head-to-Head Gene Organization: Evolutionary Conservation and Potential Biological Relevance. *PLoS Comput Biol* **2**(7): e74.

Li Z, Van Calcar S, Qu C, Cavenee WK, Zhang MQ, Ren B. 2003. A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. *Proceedings of the National Academy of Sciences* **100**(14): 8164-8169.

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

Lin YC, Benner C, Mansson R, Heinz S, Miyazaki K, Miyazaki M, Chandra V, Bossen C, Glass CK, Murre C. 2012. Global changes in the nuclear positioning of genes and intra- and interdomain genomic interactions that orchestrate B cell fate. *Nat Immunol* **advance online publication**.

Louis SF, Vermolen BJ, Garini Y, Young IT, Guffei A, Lichtensztejn Z, Kuttler F, Chuang TCY, Moshir S, Mougey V et al. 2005. c-Myc induces chromosomal rearrangements through telomere and chromosome remodeling in the interphase nucleus. *Proceedings of the National Academy of Sciences of the United States of America* **102**(27): 9613-9618.

Lovén J, Zinin N, Wahlström T, Müller I, Brodin P, Fredlund E, Ribacke U, Pivarcsi A, Påhlman S, Henriksson M. 2010. MYCN-regulated microRNAs repress estrogen receptor-α (ESR1) expression and neuronal differentiation in human neuroblastoma. *Proceedings of the National Academy of Sciences* **107**(4): 1553-1558.

Lucchesi JC, Kelly WG, Panning B. 2005. CHROMATIN REMODELING IN DOSAGE COMPENSATION. *Annual Review of Genetics* **39**(1): 615-651.

Lunyak VV, Prefontaine GG, NÃºÃ±ez E, Cramer T, Ju B-G, Ohgi KA, Hutt K, Roy R, GarcÃ-a-DÃ-az A, Zhu X et al. 2007. Developmentally Regulated Activation of a SINE B2 Repeat as a Domain Boundary in Organogenesis. *Science* **317**(5835): 248-251.

Lyko F, Foret S, Kucharski R, Wolf S, Falckenhayn C, Maleszka R. 2010. The Honey Bee Epigenomes: Differential Methylation of Brain DNA in Queens and Workers. *PLoS Biol* **8**(11): e1000506.

Maines JZ, Stevens LM, Tong X, Stein D. 2004. Drosophila dMyc is required for ovary cell growth and endoreplication. *Development* **131**(4): 775-786.

Majumder P, Gomez JA, Chadwick BP, Boss JM. 2008. The insulator factor CTCF controls MHC class II gene expression and is required for the formation of long-

distance chromatin interactions. *The Journal of Experimental Medicine* **205**(4): 785-798.

Markow TA, O'Grady P, ed. 2005. *Drosophila: A guide to species identification and use*. Academic Press.

Meyer N, Penn LZ. 2008. Reflecting on 25 years with MYC. *Nat Rev Cancer* **8**(12): 976-990.

Misulovin Z, Schwartz Y, Li X-Y, Kahn T, Gause M, MacArthur S, Fay J, Eisen M, Pirrotta V, Biggin M et al. 2008. Association of cohesin and Nipped-B with transcriptionally active regions of the <i>Drosophila melanogaster</i> genome. *Chromosoma* **117**(1): 89-102.

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

Moumen M, Chiche A, Deugnier M-A, Petit V, Gandarillas A, Glukhova MA, Faraldo MM. 2012. The Proto-Oncogene Myc Is Essential for Mammary Stem Cell Function. *STEM CELLS* **30**(6): 1246-1254.

Nativio R, Wendt KS, Ito Y, Huddleston JE, Uribe-Lewis S, Woodfine K, Krueger C, Reik W, Peters J-M, Murrell A. 2009. Cohesin Is Required for Higher-Order Chromatin Conformation at the Imprinted *IGF2-H19* Locus. *PLoS Genet* **5**(11): e1000739.

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

Nègre N, Brown CD, Shah PK, Kheradpour P, Morrison CA, Henikoff JG, Feng X, Ahmad K, Russell S, White RAH et al. 2010. A Comprehensive Map of Insulator Elements for the <italic>Drosophila</italic> Genome. *PLoS Genet* **6**(1): e1000814.

Noma K-i, Cam HP, Maraia RJ, Grewal SIS. 2006. A Role for TFIIIC Transcription Factor Complex in Genome Organization. *Cell* **125**(5): 859-872.

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

O'Donnell KA, Wentzel EA, Zeller KI, Dang CV, Mendell JT. 2005. c-Myc-regulated microRNAs modulate E2F1 expression. *Nature* **435**(7043): 839-843.

O'Donovan KJ, Diedler J, Couture GC, Fak JJ, Darnell RB. 2010. The Onconeural Antigen cdr2 Is a Novel APC/C Target that Acts in Mitosis to Regulate C-Myc Target Genes in Mammalian Tumor Cells. *PLoS ONE* **5**(4): e10045.

Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK, Fraenkel E. 2007. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet* **39**(6): 730-732.

Ohta S, Bukowski-Wills J-C, Sanchez-Pulido L, Alves FdL, Wood L, Chen ZA, Platani M, Fischer L, Hudson DF, Ponting CP et al. 2010. The Protein Composition of

Mitotic Chromosomes Determined Using Multiclassifier Combinatorial Proteomics. *Cell* **142**(5): 810-821.

Pai C-Y, Lei EP, Ghosh D, Corces VG. 2004. The Centrosomal Protein CP190 Is a Component of the gypsy Chromatin Insulator. *Molecular Cell* **16**(5): 737-748.

Pal C, Hurst LD. 2003. Evidence for co-evolution of gene order and recombination rate. *Nat Genet* **33**(3): 392-395.

Parelho V, Hadjur S, Spivakov M, Leleu M, Sauer S, Gregson HC, Jarmuz A, Canzonetta C, Webster Z, Nesterova T et al. 2008. Cohesins Functionally Associate with CTCF on Mammalian Chromosome Arms. *Cell* **132**(3): 422-433.

Pavesi G, Mereghetti P, Mauri G, Pesole G. 2004. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Research* **32**(suppl 2): W199-W203.

Perna D, Faga G, Verrecchia A, Gorski MM, Barozzi I, Narang V, Khng J, Lim KC, Sung WK, Sanges R et al. 2012. Genome-wide mapping of Myc binding and gene regulation in serum-stimulated fibroblasts. *Oncogene* **31**(13): 1695-1709.

Phillips JE, Corces VG. 2009. CTCF: Master Weaver of the Genome. *Cell* **137**(7): 1194-1211.

Pierce SB, Yost C, Britton JS, Loo LWM, Flynn EM, Edgar BA, Eisenman RN. 2004. dMyc is required for larval growth and endoreplication in Drosophila. *Development* **131**(10): 2317-2327.

Probst AV, Dunleavy E, Almouzni G. 2009. Epigenetic inheritance during the cell cycle. *Nat Rev Mol Cell Biol* **10**(3): 192-206.

Raab JR, Chiu J, Zhu J, Katzman S, Kurukuti S, Wade PA, Haussler D, Kamakaka RT. 2011. Human tRNA genes function as chromatin insulators. *EMBO J* **31**(2): 330-350.

Rahl PB, Lin CY, Seila AC, Flynn RA, McCuine S, Burge CB, Sharp PA, Young RA. 2010. c-Myc Regulates Transcriptional Pause Release. *Cell* **141**(3): 432-445.

Renault AD, Starz-Gaiano M, Lehmann R. 2002. RETRACTED: Metabolism of sphingosine 1-phosphate and lysophosphatidic acid: a genome wide analysis of gene expression in Drosophila. *Gene Expression Patterns* **2**(3–4): 337-345.

Ritter AR, Beckstead RB. 2010. Sox14 is required for transcriptional and developmental responses to 20-hydroxyecdysone at the onset of drosophila metamorphosis. *Developmental Dynamics* **239**(10): 2685-2694.

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

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

Sandmann T, Jakobsen JS, Furlong EEM. 2007. ChIP-on-chip protocol for genome-wide analysis of transcription factor binding in Drosophila melanogaster embryos. *Nat Protocols* **1**(6): 2839-2855.

Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S et al. 2010. Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding. *Science* **328**(5981): 1036-1040.

Schoborg T, Labrador M. 2010. The Phylogenetic Distribution of Non-CTCF Insulator Proteins Is Limited to Insects and Reveals that BEAF-32 Is <i>Drosophila</i> Lineage Specific. *Journal of Molecular Evolution* **70**(1): 74-84.

Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G. 2012. Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome. *Cell* **148**(3): 458-472.

Smith KN, Singh AM, Dalton S. 2010. Myc Represses Primitive Endoderm Differentiation in Pluripotent Stem Cells. *Cell Stem Cell* **7**(3): 343-354.

Sultana H, Verma S, Mishra RK. 2011. A BEAF dependent chromatin domain boundary separates myoglianin and eyeless genes of Drosophila melanogaster. *Nucleic Acids Research* **39**(9): 3543-3557.

Swarnalatha M, Singh AK, Kumar V. 2012. The epigenetic control of E-box and Myc-dependent chromatin modifications regulate the licensing of lamin B2 origin during cell cycle. *Nucleic Acids Research*.

Tempera I, Klichinsky M, Lieberman PM. 2011. EBV Latency Types Adopt Alternative Chromatin Conformations. *PLoS Pathog* **7**(7): e1002180.

Tempera I, Wiedmer A, Dheekollu J, Lieberman PM. 2010. CTCF Prevents the Epigenetic Drift of EBV Latency Promoter Qp. *PLoS Pathog* **6**(8): e1001048.

Tie F, Banerjee R, Stratton CA, Prasad-Sinha J, Stepanik V, Zlobin A, Diaz MO, Scacheri PC, Harte PJ. 2009. CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing. *Development* **136**(18): 3131-3141.

Valenzuela L, Dhillon N, Kamakaka RT. 2009. Transcription Independent Insulation at TFIIIC-Dependent Insulators. *Genetics* **183**(1): 131-148.

van Riggelen J, Yetil A, Felsher DW. 2010. MYC as a regulator of ribosome biogenesis and protein synthesis. *Nat Rev Cancer* **10**(4): 301-309.

Varlakhanova NV, Cotterman RF, deVries WN, Morgan J, Donahue LR, Murray S, Knowles BB, Knoepfler PS. 2010. myc maintains embryonic stem cell pluripotency and self-renewal. *Differentiation* **80**(1): 9-19.

Varlakhanova NV, Knoepfler PS. 2009. Acting locally and globally: Myc's ever-expanding roles on chromatin. *Cancer Res* **69**(19): 7487-7490.

Weber C, Hurst L. 2011. Support for multiple classes of local expression clusters in Drosophila melanogaster, but no evidence for gene order conservation. *Genome Biology* **12**(3): R23.

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

Wolfer A, Ramaswamy S. 2011. MYC and Metastasis. *Cancer Research* **71**(6): 2034-2037.

Wood Ashley M, Van Bortle K, Ramos E, Takenaka N, Rohrbaugh M, Jones Brian C, Jones Keith C, Corces Victor G. 2011. Regulation of Chromatin Organization and Inducible Gene Expression by a Drosophila Insulator. *Molecular Cell* **44**(1): 29-38.

Xiao T, Wallace J, Felsenfeld G. 2011. Specific Sites in the C Terminus of CTCF Interact with the SA2 Subunit of the Cohesin Complex and Are Required for Cohesin-Dependent Insulation Activity. *Mol Cell Biol* **31**(11): 2174-2183.

Xie X, Mikkelsen TS, Gnirke A, Lindblad-Toh K, Kellis M, Lander ES. 2007. 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* **104**(17): 7145-7150.

Xu H, Handoko L, Wei X, Ye C, Sheng J, Wei C-L, Lin F, Sung W-K. 2010. A signal–noise model for significance analysis of ChIP-seq with negative control. *Bioinformatics* **26**(9): 1199-1204.

Xu Z, Wei G, Chepelev I, Zhao K, Felsenfeld G. 2011. Mapping of INS promoter interactions reveals its role in long-range regulation of SYT8 transcription. *Nat Struct Mol Biol* **18**(3): 372-378.

Yang J, Corces VG. 2011. Chromatin Insulators: A Role in Nuclear Organization and Gene Expression. *Adv Cancer Res* **110**: 43-76.

Yang L, Yu J. 2009. A comparative analysis of divergently-paired genes (DPGs) among Drosophila and vertebrate genomes. *BMC Evolutionary Biology* **9**(1): 55.

Yang M, Taylor J, Elnitski L. 2008. Comparative analyses of bidirectional promoters in vertebrates. *BMC Bioinformatics* **9**(Suppl 6): S9.

Zeller KI, Zhao X, Lee CWH, Chiu KP, Yao F, Yustein JT, Ooi HS, Orlov YL, Shahab A, Yong HC et al. 2006. Global mapping of c-Myc binding sites and target gene networks in human B cells. *Proceedings of the National Academy of Sciences* **103**(47): 17834-17839.

Zhao R, Nakamura T, Fu Y, Lazar Z, Spector DL. 2011. Gene bookmarking accelerates the kinetics of post-mitotic transcriptional re-activation. *Nat Cell Biol* **13**(11): 1295-1304.

Zippo A, Serafini R, Rocchigiani M, Pennacchini S, Krepelova A, Oliviero S. 2009. Histone Crosstalk between H3S10ph and H4K16ac Generates a Histone Code that Mediates Transcription Elongation. *Cell* **138**(6): 1122-1136.