

***Drosophila* dosage compensation as a model for combinatorial
transcriptional regulation**

By

William Thomas Jordan, III

B.S., Biological Sciences, Virginia Tech, 2011

B.S., Psychology, Virginia Tech, 2011

M.A., Molecular Biology, Cell Biology, and Biochemistry, Brown University, 2017

A dissertation submitted in partial fulfillment of the requirements for the Degree of
Doctor of Philosophy in the Department of Molecular Biology, Cell Biology and
Biochemistry at Brown University

Providence, Rhode Island

May 2019

© Copyright 2019 by William Thomas Jordan, III

This dissertation by William Thomas Jordan, III is accepted in its present form by
the Department of Molecular Biology, Cell Biology, and Biochemistry as
satisfying the dissertation requirement for the degree of Doctor of Philosophy.

Date _____

Dr. Erica Larschan, Director

Recommended to the Graduate Council

Date _____

Dr. Judith Bender, Reader

Date _____

Dr. Alison DeLong, Reader

Date _____

Dr. Nicola Neretti, Reader

Date _____

Dr. Karen Adelman, Outside Reader

Approved by the Graduate Council

Date _____

Dr. Andrew G. Campbell
Dean of the Graduate School

Curriculum Vitae

William Jordan, III

Providence, RI

E-mail: william_jordan@brown.edu

EDUCATION

Brown University, Providence, RI

PhD Molecular Biology, Cell Biology, and Biochemistry 2019
▪ Advisor: Dr. Erica Larschan (expected)
▪ Dissertation title: *Drosophila* dosage compensation as a model for combinatorial transcriptional regulation

MS Data Science (Requirements completed August 2018) 2019

MA Molecular Biology, Cell Biology, and Biochemistry 2017

Virginia Tech, Blacksburg, VA

BS Biological Sciences 2011

BS Psychology 2011

Southside Virginia Community College, Albemarle, VA

AA Arts and Sciences 2007

Halifax County High School, South Boston, VA 2007

PUBLICATIONS

Jordan, WT & Larschan EN. 2019. CLAMP regulates long-range chromatin interactions on the *Drosophila* X-chromosome that mediate spreading of MSL complex. Submitted.

Jordan, W., Rieder, L.E., and Larschan, E. 2019. Diverse Genome Topologies Characterize Dosage Compensation across Species. *Trends Genet.*, 35(4) 308-315

Rieder, L.E., Koreski, K.P., Boltz, K.A., Kuzu, G., Urban, J.A., Bowman, S.K., Zeidman, A., **Jordan, W.T.,** Tolstorukov, M.Y., Marzluff, W.F. and Duronio, R.J. 2017. Histone locus regulation by the *Drosophila* dosage compensation adaptor protein CLAMP. *Genes and Development*, 31(14), 1494 – 1508

Urban, J. A., Doherty, C. A., **Jordan, W. T.,** Bliss, J. E., Feng, J., Soruco, M. M., and Larschan, E. N. 2016. The essential *Drosophila* CLAMP protein differentially regulates non-coding roX RNAs in males and females. *Chromosome Research*, 25(2) 101 – 113.

Derr, D, G. Deluga, M. Slowe, N. Lohitharn, D.P. MacEachran, J.C. Ulrich, M. Katz, **W. Jordan,** R. Mirani and A. Owen. 2013. A Method for Producing Ketones from Renewable Resources. US Provisional Patent 61/832,249.

In Preparation:

Rieder, LE*, **Jordan WT***, Larschan EN. 2019. Dynamic targeting of the embryonic *Drosophila* male X-chromosome.

*Co-first authors

Jordan, W., Nguyen, E., and Larschan, E. 2019. CLAMP motif sequence variation, density, and proximity promotes recruitment of the *Drosophila* dosage compensation complex to the male X-chromosome.

Jordan, W., Istrail, S., and Larschan, E. 2019. Uncovering factors that promote identification of the *Drosophila* X-chromosome.

PRESENTATIONS

Invited presentations

FASEB: Transcription, Chromatin, and Epigenetics Meeting, Florence, Italy (poster) 2018

▪ An essential transcription factor regulates three-dimensional organization of the *Drosophila* X-chromosome to promote dosage compensation

Cold Spring Harbor Laboratory (CSHL) Biological Data Science, Cold Spring Harbor, NY (poster) 2018

▪ The X in sex: Uncovering factors that promote X-chromosome identification in *Drosophila*

CSHL Systems Biology: Global Regulation of Gene Expression, Cold Spring Harbor, NY (poster) 2017

▪ Clustering of cis-elements open chromatin to promote X-specificity of *Drosophila* dosage compensation

The Allied Genetics Conference, Orlando Florida (poster)	2016
▪ CLAMP: Sex specific or global transcription factor	
Invited presentations: Internal	
Brown University MCB Graduate Program Retreat (speaker), Providence, RI	2017
▪ 3D genome organization promotes dosage compensation in <i>Drosophila melanogaster</i>	
Other presentations	
HHMI Gilliam Fellowship Meeting (poster)	2018
Brown University MCB Data Club (speaker)	2018
HHMI Gilliam Fellowship Meeting (poster)	2017
Brown University MCB Graduate Program Retreat (poster)	2017
Brown University MCB Data Club (speaker)	2017
Brown University MCB Graduate Program Retreat (poster)	2016
HHMI Gilliam Fellowship Meeting (poster)	2016
Brown University MCB Graduate Program Retreat (poster)	2015
American Oil Chemists Society Annual Meeting & Expo	2014
Virginia Tech MAOP & McNair Research Symposium (speaker & poster)	2009

AWARDS AND FELLOWSHIPS

Howard Hughes Medical Institute Gilliam Fellowship Awardee	2016 – 2019
National Science Foundation Graduate Research Fellowship Awardee (on reserve status)	2016 – 2019
Brown University Open Graduate Education Awardee	2017
Brown University Open Graduate Education Awardee	2017
Brown University MCB NIH T32 training grant Awardee	2015
Virginia Tech Phi Sigma Biological Sciences Honor Society Inductee	2008

PROFESSIONAL RESEARCH EXPERIENCE

Logos Technologies, Fairfax, VA	2012 - 2014
Microbiologist	
▪ Performed synthetic biology research funded by DARPA and DOE.	
▪ Developed tools and genetically modified organisms to economically convert cellulosic feedstock into drop-in diesel and kerosene fuels in addition to other high value chemicals.	
▪ Genetically modified a diverse range of pathogenic and non-pathogenic microbes in a variety of ways including creating genetic nulls and inserting marker-less transgenes.	
▪ Gained additional experience in scaling up microbial fermentation process from small laboratory scale proof of concept volumes to larger volumes in preparation for pilot scale testing.	
Intrexon Corporation (UltraVector Division), Blacksburg, VA	2012
Lab Associate I (Fragment isolation group)	
▪ Performed high-throughput restriction enzyme-based cloning to create plasmids.	
▪ Gained experience with a wide variety of automated equipment and liquid handlers.	

DATA SCIENCE MASTER'S DEGREE PROJECTS, BROWN UNIVERSITY

The X in sex: Uncovering factors that promote <i>Drosophila</i> X-chromosome identification	2018
Advisors: Dr. Sorin Istrail (CS Dept.) and Dr. Erica Larschan (MCB Dept.)	
▪ Used publicly available next-generation sequencing (NGS) data to investigate how the <i>Drosophila</i> dosage compensation complex specifically localizes to the X-chromosome when the sequences it appears capable of recognizing <i>in vitro</i> and <i>in silico</i> are found genome-wide.	
▪ I used supervised machine learning in the form of boosted random forests to model this phenomenon.	
▪ I identified novel potential cofactors that may be involved in X-identification for further validation using <i>in vivo</i> experiments.	
Police killings, actual bias or media hype?	2018
▪ Cleaned and integrated publicly available information from the Guardian that cataloged data on Americans killed by police in 2015 and additional data from the FBI's uniform crime reporting database.	
▪ My exploratory analysis showed that overall both whites and blacks made up the highest percentage of deaths; additionally, blacks made up an overwhelming majority of non-gunshot deaths.	
▪ To further investigate which factors may provide insight on police killings I built a multi-level logistic regression model focused on predicting whether an individual killed was armed or unarmed.	
▪ My analysis show that unarmed police shootings of blacks were specifically more likely to occur in areas that have both high black populations and unemployment. Furthermore, killing of blacks was observed to	

occur more frequently in areas that had relatively low violent crime rates in comparison with areas of non-black killings. Together these findings suggest that racial biases may indeed play a pivotal role in police killings.

Neural network flower classification 2018

- Use labeled images of flowers (Kaggle open dataset) and Keras to build and train a convolutional neuronal network (CNN) to classify 5 types of flowers
- Also visualized how the neural network was learning to distinguish among the image classes in order to identify how misclassifications may be occurring.

Twitter sentiment analysis 2018

- Explored the use of neural networks in natural language processing (NLP)
- Used categorized tweets (Kaggle open dataset) and Keras to build a recurrent neural network (RNN) to classify whether a tweet was positive or negative.

Median household income and income per capita in NYC 2018

- Part of a 3-member team tasked with using data from the 2015 American Community Survey to model what factors drive median household income and income per capita in NYC using linear regression.

V1s10n: Using neural networks to classify artist and style of paintings 2017

Advisor: Dr. Thomas Serre (CLPS Dept., Brown University)

- Self-directed team project (3-member team) under the guidance of Dr. Serre, which we set out to build a CNN to classify paintings by style or artist. Images of paintings were largely sourced via web scraping. Images and associated metadata were cleaned and stored in an SQL database.
- I was directly responsible for creating the neural networks that would be used for classification; I coadapted VGG16, a publicly available, highly regarded, deep CNN, for this project. I replaced the top layer of the network, froze the intermediate layers and retrained the network to individually classify images by style or painter.
- I also created a simple python-based GUI program that allows users to submit any image and have the three most similarly "related" artists or styles returned.

Amazon employee access challenge 2017

- Explored the performance of a variety of machine learning models to predict employee resource access needs for new employees at Amazon (Kaggle open dataset).

TEACHING AND MENTORING EXPERIENCE

Teaching Assistant 2017

Brown University, IMSD module: Programming Concepts

- Assisted with planning and teaching a 5-part introduction to programming workshop for graduate students with little to no prior programming experience hoping to develop the computational and analytical skills needed for courses in biomedical informatics and data science.

Graduate Teaching Assistant 2016

Brown University, BIOL 1540/2540: Molecular Genetics

- Led a weekly discussion section, in addition to administering exams/quizzes and assisting with grading.

Mentoring

Larschan Lab

2017 – 2019

- Served as direct laboratory mentor for two undergraduate students and two 1st year PhD rotation students.

Brown University MCB Graduate Program 1st Year Student Mentoring Program

2016 – 2017

- Served as student mentor to 1st year graduate student.

SERVICE AND OUTREACH

Brown University MCB Department Diversity & Inclusion Committee 2016 – 2019

Brown University MCB Graduate Program Admissions Committee 2017

Brown University MCB Graduate Program student liaison 2016 – 2017

Annual Biomedical Research Conference for Minority Students (ABRCMS), Tampa FL 2016

- Brown University MCB Graduate Program student representative
- Poster Judge

Brown University MCB Graduate Program Retreat Planning Committee 2016

PROFESSIONAL MEMBERSHIPS

American Association for the Advancement of Science 2016 – 2018

Genetics Society of America 2016 – 2017

PROFESSIONAL DEVELOPMENT

Center for cell circuits computational genomics workshop, Broad Institute	2017
Effective Performance Workshop, Brown University and Trinity Repertory Theater	2016
Chromatin, Epigenetics and Gene Expression course, Cold Spring Harbor Laboratory	2016
Fermentation Technology, Massachusetts Institute of Technology: Short Programs	2013

ADDITIONAL RESEARCH EXPERIENCE

PhD Rotation Research, Brown University MCB Graduate Program	2014 – 2015
<u>Functional characterization of splicing lariats</u>	
Advisor: Dr. William Fairbrother	
<u>CLAMP: Sex specific or global transcription factor?</u>	
Advisor: Dr. Erica Larschan	
<u>Characterization of motor defects and metabolomic signatures in a <i>Gpt2</i>-null mouse model</u>	
Advisor: Dr. Eric Morrow	
Laboratory of Dr. Richard Veilleux, Virginia Tech, Horticulture Dept.	2010 – 2011
Research Assistant	
▪ Assisted on various graduate research projects characterizing stress response in genetic modified strawberry and potato plants.	
Laboratory of Dr. Dorothea Tholl, Virginia Tech, Biology Dept.	2009
MAOP Summer Research Intern	
▪ Characterized two novel diterpene synthase genes in <i>Arabidopsis thaliana</i> .	
Laboratory of Dr. Karen Brewer, Virginia Tech, Chemistry Dept.	2007 – 2008
Research Assistant	
▪ Synthesized starting materials for graduate research investigating synthesis and application of organometallics.	

Acknowledgements

This dissertation would not be possible without the continued guidance and support of many people that I have had the pleasure of interacting with over the years; to all of my mentors (official and unofficial), friends and family, thank you!

First and foremost, to my doctoral advisor and mentor, Dr. Erica Larschan I would like to say thank you for your thoughtfulness, accessibility, and continued encouragement throughout the years. Thank you for your unwavering confidence in me and providing me with the freedom to fully explore my own interests. Thank you for helping me work through my frustrations. I will miss your infectious energy and positivity for science and life. Thank you for helping me to become a well-rounded scientist; I will lean on my experiences under your guidance throughout the next chapters of my professional development.

To the many other scientific mentors that I have been fortunate to have over the years thank you. I would like to offer a specific thank you to Dr. Michael Campbell and Dr. Daniel MacEachran. Thank you each for helping me to become a better and more confident scientist. Thank you each for helping me to realize my full scientific potential; I am not sure if I would be here if it were not for your continued support and advice.

To the committee members I have had over the years, thank you for your helpful discussions. I would like to offer a specific thank you to the members of current my committee. Dr. Judith Bender and Dr. Alison DeLong, thank you each for your insightful critiques; thank you each for helping me to key in on my deficiencies and expand the breadth of my scientific knowledge. Dr. Nicola Neretti, thank you for agreeing to serve on my committee on short notice, I look forward to your feedback. Dr. Karen Adelman, thank you for your contributions to my scientific development early in my graduate career via your summer course and for serving as my outside reader.

To Dr. Mark Johnson and all of the other faculty, staff, and students of the MCB graduate program, thank you for helping to nurture my intellectual curiosity.

To all of the past and present members of the Larschan laboratory, thank you for providing a brilliant blueprint for me to follow in addition to creating such a fun and intellectually stimulating working environment. I would like to offer a specific thank you to Jacob Bliss, Dr. Jennifer Urban, Dr. Emily Kaye, and Dr. Leila Rieder. Thank you each for your insightful help, ideas, and discussions throughout our time together; you each inspire me in many different ways. I would also like to offer a specific thank you to Erica Nguyen. Thank you for enduring the transition period of the lab alongside me; you have had a greater impact on me than you may think. I look forward to your continued personal and scientific growth.

To the HHMI Gilliam Fellowship program, thank you for the financial support in addition to all of the personal development and networking opportunities. To my fellow Gilliam fellows at Brown and elsewhere, thank you for inspiring me. I would like to offer a specific thank you to Dr. Valerie Zabala who helped me realign my focus during some of the most demanding challenges I faced in my graduate career.

To all of the lifelong friends I have made over the years, thank you for your support and encouragement. To my fellow graduate friends at Brown, thank you for your patience and thoughtful insights; thank you for helping to make my time here more enjoyable.

Most importantly, I would like to thank my family. Though, we are currently more distally apart than we have ever been, coupled with my well-known disdain for non-face-to-face forms of communication, thank each of you for being there when I needed you most. Thank you each for your constant prayers and support. Thank you each for serving as role models and inspiring me. To my sister Shantavea, thank you for stirring my scientific curiosity and for showing me the importance of embracing alternative perspectives. To my sister Shanna, thank you for your constant encouragement. Thank you for showing me the value of perseverance. I am so proud of how far you have come in your own journey. To my mother, Linda,

thank you for everything! I often reflect on my early childhood; you enabled my success in so many subtle and immeasurable ways. Thank you for not giving up on me. Thank you for showing me the power of speaking things into existence when coupled with hard work and making the best of opportunities. Thank you for teaching me how to persevere and building up my self-assurance. Through your hard work and perseverance, you made this possible. I am eternally grateful for the sacrifices that you made.

Lastly, in the words of the urban philosopher Calvin Cordozar Broadus Jr.:

“I want to thank me. I want to thank me for believing in me. I want to thank me for doing all this hard work. I want to thank me for having no days off. I want to thank me for never quitting. I want to thank me for always being a giver and trying to give more than I receive. I want to thank me for trying to do more right than wrong. I want to thank me for just being me at all times...”

Table Of Contents

COPYRIGHT PAGE	II
SIGNATURE PAGE	III
CURRICULUM VITAE	IV
ACKNOWLEDGEMENTS	VIII
LIST OF TABLES	XIV
LIST OF FIGURES	XV
CHAPTER 1 : INTRODUCTION	1
COMPOSITION OF CHROMATIN	1
TRANSCRIPTIONAL REGULATION MEDIATED BY CIS-REGULATORY ELEMENTS	3
OVERVIEW OF DOSAGE COMPENSATION.....	5
Mechanisms of dosage compensation.....	7
TECHNIQUES FOR PROBING CHROMATIN STRUCTURAL DYNAMICS	10
Chromosome conformation capture	10
Micrococcal nuclease sequencing and assay for transposases-accessible chromatin using sequencing	11
Cleavage Under Targets and Release Using Nuclease.....	12
Single-cell techniques	13
DROSOPHILA DOSAGE COMPENSATION AS A MODEL FOR UNDERSTANDING CONTEXT-SPECIFIC TRANSCRIPTIONAL REGULATION	14
X-identification by the Male specific lethal complex	17
CLAMP as a model for context-specific gene regulation	18
CHAPTER 2 : CLAMP MOTIF SEQUENCE VARIATION, DENSITY, AND PROXIMITY PROMOTES RECRUITMENT OF THE DROSOPHILA DOSAGE COMPENSATION COMPLEX TO THE MALE X-CHROMOSOME	21
ABSTRACT	22
INTRODUCTION	23
RESULTS	27
CLAMP binding sequences are necessary but not sufficient for MSLc activity.....	27
MSLc activity is maintained in the context of increasing the density of MREs	30
Close proximity of MRE motifs to each other promotes MSLc activity	31
Flanking sequences can modulate the activity of MREs that are not in close proximity to each other	32
In vivo analysis of MRE proximity and density modifications	37
DISCUSSION	40
METHODS	41
Plasmids and cloning	41
Dual Luciferase Reporter Assay	41
Fly Genetics and Transgenesis.....	42
Immunodetection.....	42
CHAPTER 3 : UNCOVERING FACTORS THAT PROMOTE IDENTIFICATION OF THE DROSOPHILA X-CHROMOSOME	43
ABSTRACT	44
INTRODUCTION	45

MATERIALS AND METHODS	46
Data collection, organization and cleaning	46
Exploratory Analysis	49
Model Building and Validation	51
Model Application and Interpretation	54
CONCLUSIONS AND FUTURE DIRECTIONS	58
SUPPORTING INFORMATION	60
CHAPTER 4 : THE ZINC FINGER PROTEIN CLAMP PROMOTES LONG-RANGE CHROMATIN INTERACTIONS THAT MEDIATE DOSAGE COMPENSATION OF THE DROSOPHILA MALE X-CHROMOSOME.....	61
ABSTRACT	62
INTRODUCTION	63
RESULTS	67
CLAMP regulates long-range three-dimension interactions on the X-chromosome more strongly than on autosomes	67
CLAMP promotes three-dimensional interactions at active chromatin regions including CES and represses three-dimensional interactions at inactive chromatin regions.....	73
CLAMP and MSLc both mediate three-dimensional interactions at CES	76
Insulator proteins have differential occupancy at genomic locations at which CLAMP regulates three-dimensional interactions	81
DISCUSSION	84
ACKNOWLEDGEMENTS	86
AUTHOR CONTRIBUTIONS	86
DECLARATION OF INTERESTS	86
METHODS	87
Cell culture conditions	87
RNAi treatment.....	87
Hi-C experimental procedure	89
4C-seq experimental procedure	92
Datasets	92
Hi-C analysis	92
4C-Seq analysis	94
Comparison of 4C and Hi-C high confidence interactions	95
Chromatin state annotation	96
Chromatin accessibility analysis.....	96
Chip-seq data processing and generation of high confidence peak sets	96
SUPPORTING INFORMATION.....	98
CHAPTER 5 : DISCUSSION.....	103
DIVERSE MODELS OF COMBINATORIAL TRANSCRIPTIONAL REGULATION IN DROSOPHILA.	104
IDENTIFICATION OF ADDITIONAL FEATURES THAT ALLOW THE X-CHROMOSOME TO BE DISTINGUISHED FROM THE AUTOSOMES	106
RESOLVING HETEROGENEITY IS NECESSARY TO BETTER UNDERSTAND COMBINATORIAL TRANSCRIPTIONAL REGULATION.....	108
SUMMARY	109
BIBLIOGRAPHY	112

List of Tables

TABLE S3.1 SEARCH SPACE FOR RANDOMIZED GRID SEARCH	60
--	----

List of Figures

FIGURE 1.1. DOSAGE COMPENSATION STRATEGIES.....	7
FIGURE 1.2. EMERGING TECHNOLOGIES AND THEIR CURRENT IMPACTS WITHIN THE DOSAGE COMPENSATION FIELD	11
FIGURE 2.1. CLAMP SPECIFIC GA-RICH CRES REQUIRED FOR MSLC ACTIVITY.	29
FIGURE 2.2. CLOSE PROXIMITY OF MRES WITHIN CES CONFERS HIGHER MSLC ACTIVITY.	32
FIGURE 2.3. ARTIFICIALLY INCREASING PROXIMITY OF AUTOSOMAL GA-RICH CRES IN RESULTS IN INCREASED MSLC ACTIVITY.	36
FIGURE 2.4. SELECT <i>IN VIVO</i> INSERTIONS VALIDATE LUCIFERASE REPORTER FINDINGS.	39
FIGURE 3.1. EXAMPLE OF VISUAL DATA INTEGRITY CHECKS USING CHIP-SEQ PROFILES.	48
FIGURE 3.2. FEATURE IMPORTANCE RANKINGS FOR THE FULL MODEL.....	54
FIGURE 3.3. FEATURE IMPORTANCE RANKINGS FOR THE NULL MSLC MODEL.....	55
FIGURE 3.4. DISTRIBUTION SHAP VALUES ACROSS TOP FEATURES IDENTIFIED ACROSS VARIOUS SITE CLASSES.	57
FIGURE 4.1. CLAMP REGULATES THE SPAN OF GENOMIC INTERACTIONS ON THE MALE X-CHROMOSOME.	70
FIGURE 4.2. CLAMP PROMOTES THE FORMATION OF LONGER-RANGE CONTACTS MORE STRONGLY ON THE X-CHROMOSOME THAN AUTOSOMES.....	72
FIGURE 4.3. CLAMP REGULATION OF THE THREE-DIMENSIONAL ORGANIZATION IS LINKED TO ITS ROLE IN ALTERING CHROMATIN ACCESSIBILITY.	80
FIGURE 4.4. DIFFERENTIAL OCCUPANCY OF INSULATOR PROTEINS OCCURS AT LOCI WHERE CLAMP PROMOTES OR REPRESSES THE FORMATION OF THREE-DIMENSIONAL INTERACTIONS ON THE X-CHROMOSOME AND AUTOSOMES.....	83
FIGURE S4.1. EXPERIMENTAL QC AND HI-C REPLICATE CONCORDANCE	98
FIGURE S4.2. CLAMP PROMOTES LONG-RANGE INTERACTIONS ON THE X-CHROMOSOME AND GENERALLY PROMOTES INTERACTIONS IN ACTIVE CHROMATIN AND REPRESSES INTERACTIONS IN INACTIVE CHROMATIN	99
FIGURE S4.3. SUMMARY OF HIGH-RESOLUTION 4C-SEQ ANALYSIS.....	100
FIGURE S4.4. OVERVIEW OF DIFFERENTIAL 4C-SEQ ANALYSIS AFTER <i>GFP</i> , <i>MSL2</i> , OR <i>CLAMP</i> RNAI	101

CHAPTER 1 : INTRODUCTION

William Jordan, III¹, Leila E. Rieder^{1,2}, and Erica Larschan¹

Portions of this chapter have been published as a first author review that I wrote for *Trends in Genetics* (Diverse genome topologies characterize dosage compensation across species) and are reproduced here with many changes and additions. I wrote the original manuscript with editorial feedback from Dr. Leila Rieder, who prepared the figures, and Dr. Erica Larschan. I wrote the remainder of this chapter with editorial feedback from Dr. Erica Larschan.

¹ Department of Molecular Biology, Cellular Biology and Biochemistry, Brown University, Providence, RI

² Department of Biology, Emory University, Atlanta, GA

Composition of chromatin

Within the eukaryotic cell nucleus, each chromosome consists of a single molecule of double-stranded DNA. Frequently thought of in linear space, these DNA molecules are compacted within the confines of the nucleus, as much as ten thousand-fold, by adopting diverse three-dimensional conformations. Despite extensive compaction of the chromosomes within the nucleus, specific sequences known as *cis*-regulatory elements need to be precisely identified to regulate genes spatially and temporally.

At the most basic level, chromatin consists of a repeating array of nucleosomes, which are octamers of histone proteins, around which DNA is wrapped. There are two copies of each of the four core histone proteins (H2A, H2B, H3 and H4) within each nucleosome, which is surrounded by 146 bp of DNA (Dixon et al., 2016; Widom, 1989). Chromatin serves as the substrate for essential cellular processes such as transcription, replication, recombination and cell division (Widom, 1989). The idea of crumpled polymer globules, or the “beads on a string” framework for chromatin, was initially introduced forty years ago (Grosberg et al., 1993, 1988). Surprisingly, chromatin only accounts for roughly half of the available nuclear space (Rajapakse and Groudine, 2011): specialized nuclear bodies and diverse chromatin regulators reside in the remaining interchromatin space. Nuclear bodies are dynamic structures that often reflect a cell’s transcriptional state and serve as

sites of transcription or RNA processing (Aumiller Jr et al., 2014; Rajapakse and Groudine, 2011).

Chromatin can be generalized into two contrasting types: euchromatin, which generally resides in the interior of the nucleus, and heterochromatin, which generally resides close to the nuclear envelope (Bonev and Cavalli, 2016). These types of chromatin differ in their accessibility: euchromatin is regarded as “active” because it contains gene-rich regions, and positively correlates with chromatin accessibility; heterochromatin is regarded as “inactive,” is gene poor, and negatively correlates with chromatin accessibility (Bonev and Cavalli, 2016; Gilbert et al., 2004; Solovei et al., 2013). Chromatin may vary dramatically between cell types because histone proteins, histone modifiers (or readers/writers), and other transcription factors are able to recognize and modulate specific loci, further differentiating chromatin compartments and fine-tuning gene expression. Therefore, while chromatin broadly influences gene regulation, it does not fully explain the modulation in activity of individual genes; rather chromatin establishes local environments that are more or less favorable for gene-specific expression (Barutcu et al., 2015; Cavalli and Misteli, 2013; Dixon et al., 2015; Li et al., 2015).

In contrast to the local chromatin environment surrounding a particular gene, topologically associated domains (TADs), are physical megabase-sized partitions of the genome. Within an organism, TAD organization is believed to be largely

invariant and conserved across cell types (Dixon et al., 2016). Smaller domains are known by various names, including sub-TADs, chromatin loops, and neighborhoods. These smaller domains may be highly variable between cell types (Barutcu et al., 2015; Dixon et al., 2015) and thus play a fundamental role in genomic function and regulation.

Across species, TAD boundaries are enriched for insulator elements, which are bound by a specialized class of insulator proteins that regulate three-dimensional genome organization (Bonev and Cavalli, 2016; Ramírez et al., 2015). TAD boundary regions are also enriched for active genes, though recent studies in *Drosophila* suggest three-dimensional chromatin organization is established during zygotic genome activation, independent of transcription (Hug et al., 2017). Few factors, including transcription factors and non-coding RNAs, modulate global genome organization (Bonev and Cavalli, 2016). Therefore, much remains unknown regarding how specific nuclear domains are established and how chromatin domains are targeted to modulate gene regulation.

Transcriptional regulation mediated by *cis*-regulatory elements (CREs)

The evolution of complex biological systems relies on the ability to precisely regulate gene networks which requires many diverse *cis*-regulatory elements (CREs) that recruit DNA-binding transcription factors. Many CREs are thought to have evolved from selective pressures on transposition of transposable elements

(TEs) that rely on the transcriptional machinery of the host cell for gene expression (Chuong et al., 2017). Consequently, many TEs evolved CREs that mimicked those already present in the host or that conferred new benefits (Chuong et al., 2017). CREs can broadly be divided into the four following categories: promoters, enhancers, silencers, and insulators. Promoters are normally positioned upstream of a transcription start site and recruit transcriptional machinery to initiate transcription (Haberle and Stark, 2018; Zabidi and Stark, 2016). Enhancers are CREs that enhance or increase transcriptional activity (Haberle and Stark, 2018; Kolovos et al., 2012; Zabidi and Stark, 2016). Contrarily, silencers are CREs that silence or decrease transcriptional activity (Kolovos et al., 2012). Lastly, insulators are CREs that recruit architectural or insulator proteins that are capable of bridging long-range DNA interactions and thereby regulate three-dimensional genome organization (Sexton et al., 2012; Yang and Corces, 2012). Furthermore, many CREs are capable of regulating nearby genes in *cis* or *trans*; they are able to affect distal genes by taking advantage of three-dimensional organization (Zabidi and Stark, 2016).

Transcription factors may regulate gene programs solely based on their abundance, through post-translational modifications, or the ability to recruit specific cofactors (Bar-Even et al., 2006; Hafner et al., 2017; Reiter et al., 2017). However, in order to mediate their diverse effects on gene regulation, transcription factors must first be able to recognize and bind to specific CREs (Reiter et al.,

2017). Within the highly compacted chromatin environment of the nucleus, many transcription factors are initially incapable of binding the CREs they recognize. These factors rely on a small subset of transcription factors known as pioneer factors that are capable of displacing nucleosomes from DNA (Reiter et al., 2017). Histone modifications and chromatin accessibility also regulate transcription factor binding (Gertz et al., 2013; Reiter et al., 2017). However, it remains unknown how only a small fraction of CREs for a given factor are bound *in vivo*. Furthermore, the subset of bound sites may vary from cell to cell and function differently depending on their genomic context (Bradley et al., 2010; Gertz et al., 2013; Reiter et al., 2017; Slattery et al., 2014; Voss and Hager, 2014). Therefore, little is understood about how transcription factors regulate genes in a context-specific manner.

Overview of dosage compensation

Diverse models of sex chromosome dosage compensation, in which the sex chromosomes are specifically targeted for transcriptional regulation, serve as model systems to study context-specific gene regulation. Moreover, analyses of dosage compensation across species have revealed the mechanisms by which large groups of genes are co-regulated within the context of three-dimensional genome organization.

Sex chromosome dosage compensation is an essential, yet diverse, process in many species. There exist many distinct dosage compensation strategies to

equalize transcript levels of genes located on the sex chromosome(s) between the sexes (Deng et al., 2011; Ercan, 2015). A conserved first step in heterogametic XY species, including humans and *Drosophila*, is distinguishing the X-chromosome from the autosomes. For example, in many mammals, one female X-chromosome is singled out for inactivation. One conserved dosage compensation mechanism, first hypothesized by Susumu Ohno (Ohno, 1973), is shared across several species including mammals, *C. elegans* and *Drosophila*: upregulation of transcription of genes on the active X-chromosome to equalize their expression with that of autosomal genes (Deng et al., 2011, 2014; Ercan, 2015) (Figure 1.1).

In contrast to the X-chromosome, there are no known examples in which the Y-chromosome is targeted for dosage compensation. The Y-chromosome is small and heterochromatic, though in some species it does carry specialized genes that have male sex-specific functions (Graves, 2006). X-chromosome upregulation was first proposed by Susumu Ohno (Ohno, 1973), who hypothesized that transcription of the X-chromosome in mammals was upregulated during early evolution of sex chromosomes in order to compensate for degeneration of the Y-chromosome (Veitia et al., 2015). Furthermore, Ohno theorized that X-chromosome upregulation was not limited to males but also occurred in females. Ohno proposed that X-chromosome downregulation evolved in mammals in order to counteract X upregulation in XX females, therefore maintaining equalized dosage with

autosomes (Veitia et al., 2015). This idea is widely known as Ohno's hypothesis. Data exist that both support and oppose the existence of such a mechanism across species (Albritton et al., 2014; Deng et al., 2014; Veitia et al., 2015; Wheeler et al., 2016).

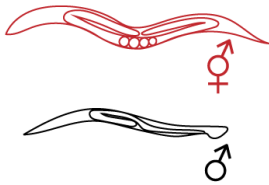
Organism	Sex Chr.	Dosage Compensation	3-D Chromosome Changes
A. <i>C. elegans</i> 	XX, AA	1. X-upregulation in both sexes: XX , AA X O, AA	Dosage compensation results in long-range contacts on—and compaction of—the X-chromosome.
	XO, AA	2. DCC-mediated X-down-regulation in hermaphrodites: XX , AA X O, AA	
B. <i>Drosophila</i> 	XX, AA	MSLc-mediated X-upregulation in males: XX, AA X Y, AA	Long-range contacts on the X-chromosome exist in both sexes.
	XY, AA		
C. Mouse 	XX, AA	Xist-mediated X-inactivation in females: X •, AA XY, AA	Xist spreading leads to additional long-range contacts to the XIST locus. The X-chromosome is highly compacted into the Barr body.
	XY, AA		

Figure 1.1. Dosage compensation strategies: the dosage compensated sex is represented in red.

Mechanisms of dosage compensation

In the nematode worm *C. elegans*, males carry a single X-chromosome and two copies of each autosome (XO, AA), while hermaphroditic females carry two copies of the X-chromosome and each autosome (XX, AA). An unknown mechanism is thought to upregulate expression of the X-chromosome in both sexes. A well-defined dosage compensation complex (DCC) then down regulates gene expression for both X-chromosomes in hermaphrodites to equalize X-linked gene

expression to that of males (Albritton et al., 2017; Deng et al., 2011; Ercan, 2015; Meyer, 2005; Pferdehirt et al., 2011) (Figure 1.1A). The DCC initially localizes to a small number of X-linked X-chromosome “recruitment element on X” (*rex*) sites, which are located in regions of euchromatin and contain binding motifs that cluster in two- and three-dimensional space (Albritton et al., 2017). The process of *C. elegans* dosage compensation is believed to occur through X-chromosome compaction because compaction is lost in the absence of DCC activity and the X-chromosome becomes enlarged. Enlargement of the X-chromosome may contribute to upregulation of X-linked genes by a mechanism that likely involves the function of specific X-linked DNA sequences rather than global regulation of all genes on the X-chromosome (Lau et al., 2014; Wheeler et al., 2016).

In the fruit fly *Drosophila melanogaster*, males carry a single X-chromosome and two copies of each autosome (XY, AA), while females carry two copies of the X-chromosome and each autosome (XX, AA) (Figure 1.1B). A dosage compensation complex known as the Male Specific Lethal complex (MSLc), is specifically assembled in males and upregulates transcription of the single male X-chromosome to approach levels of female X-linked gene expression. Similarly to *C. elegans*, the initial sites of MSLc recruitment, chromatin entry sites (CES), are located in euchromatin and contain binding motifs that cluster in two and three-dimensional space (Alekseyenko et al., 2008; Kuzu et al., 2016; Ramírez et al., 2015). However, the mechanisms by which this clustering occurs remains

unknown. Furthermore, H4K16ac, which is deposited by the MOF component of MSL complex is important for increasing transcript levels on the male X-chromosome (Armstrong et al., 2018; Copur et al., 2018).

In mammals, such as mouse and human, males carry a single X-chromosome and two copies of each autosome (XY, AA), while females carry two copies of the X-chromosome and each autosome (XX, AA) (Figure 1.1C). In females, one copy of the X-chromosome is randomly inactivated in a process known as X-inactivation (XI), which normalizes female X-linked gene expression with that of males (Ercan, 2015). X-inactivation is characterized by extensive DNA methylation that is thought to maintain the inactive state (Nozawa et al., 2013).

In mammals, the inactive female X-chromosome forms two large compact domains (Giorgetti et al., 2016; Jansz et al., 2018). In *C. elegans*, compaction along the entire length of both X-chromosomes in the hermaphrodite is associated with reduced transcription levels (Lau et al., 2014). In *Drosophila*, the X-chromosome exhibits enhanced accessibility in both males and females independent of MSLc (Urban et al., 2017a). In the future, chromosome capture techniques could be applied to reveal which species-specific factors are important for setting up and or maintaining the three-dimensional chromosome conformation many dosage compensation complexes use to target an entire chromosome for specific regulation.

Techniques for probing chromatin structural dynamics

Dosage compensation requires the targeting and regulation of an entire chromosome. Therefore, sex chromosomes represent a model that provides great insights into how smaller chromatin domains are targeted. Many improved techniques have had a substantial impact on the study of chromatin structure and function. While most of our understanding of dosage compensation comes from classical genetics and microscopy, which have been very powerful, these approaches do not address the heterogeneity of dosage compensation across cells and tissue types. Further application of the new approaches described below will allow for even greater insight into how specific parts of the genome may be regulated context-specifically at high resolution in single cells.

Chromosome conformation capture (3C)

The development of Chromosome Conformation Capture (3C) and the subsequent development of the current 3C family of techniques—chromosome conformation capture on chip (4C), chromosome capture carbon-copy (5C), and genome-wide chromosome conformation capture (Hi-C) (Dekker et al., 2002; Lieberman-Aiden et al., 2009; Simonis et al., 2006) have significantly contributed to our understanding of how three-dimensional genome structure contributes to genomic function (Figure 1.2A). There have been many recent improvements in this family of techniques. For example, *in situ* Hi-C may be performed in intact nuclei, which reduces experimental noise due to random ligation events and enables the

production of higher resolution data from fewer reads compared to the original technique (dilution Hi-C) (Rao et al., 2014). 4C has also been improved by the incorporation of unique molecular identifiers, which reduce the impact of PCR-inflated ligation read counts, allowing the generation of 4C profiles with as few as 100,000 reads per bait (Schwartzman et al., 2016) compared with 500,000 to 3 million reads for traditional 4C-seq (Van De Werken et al., 2012).

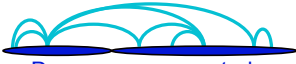
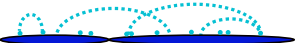
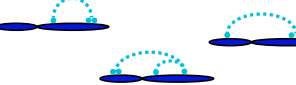
3-D Technique	Insights	DC Advances	Future Advances
A. Chromosome Conformation Capture (3C, 4C, 5C, Hi-C, <i>in situ</i> Hi-C, dilution Hi-C)	High-resolution 3-D contacts  Dosage compensated chromosome	Determine the relationship between structure of the DC chromosome and the dosage compensation machinery.	Determine how the 3-dimensional structure of the dosage compensated X-chromosome differs from that of autosomes.
B. Micrococcal Nuclease (MNase)-seq Assay for Transposon-Accessible Chromatin (ATAC)-seq	Chromatin accessibility and nucleosome positioning 	Determine the relationship between accessibility of the DC chromosome and the dosage compensation machinery.	Determine how dosage compensated X-chromosome accessibility differs from that of autosomes.
C. Cleavage Under Targets & Release Under Nuclease (CUT&RUN)	High-resolution interaction (direct vs. indirect) 3-D contact mapping 	None so far.	Identify factor binding sites and 3-dimensional interactions of dosage compensation regulators.
D. <i>In situ</i> (single cell) Fluorescence Activated Cell Sorting (FACS)	Cell-to-cell heterogeneity 	None so far.	Define how dosage compensation functions differently between cell types and tissues.

Figure 1.2. Emerging technologies and their current impacts within the dosage compensation field

Micrococcal nuclease sequencing (MNase-seq) and assay for transposases-accessible chromatin using sequencing (ATAC-seq)

MNase-seq and ATAC-seq are two popular methods for probing chromatin accessibility genome-wide (Buenrostro et al., 2013; Tolstorukov et al., 2010; Zhang and Pugh, 2011) (Figure 1.2B). ATAC-seq is a quick, cost-effective, two-step

procedure that is able to identify regions of open chromatin (highly accessible regions) and map nucleosome positioning and transcription factor binding sites genome-wide (Buenrostro et al., 2013). The latest version of MNase-seq, which utilizes a titration series, is able to measure relative chromatin accessibility as well as probe nucleosome positioning (lowest to highest regions of accessibility) genome-wide (Mieczkowski et al., 2016). While both techniques offer the ability to determine regions of open chromatin, MNase-seq with titration series offers the added benefit of providing relative chromatin accessibility genome-wide, allowing various regulatory regions throughout the genome to be directly compared.

MNase-seq in *Drosophila* cultured cells revealed that MSLc regulates chromatin accessibility by modulating nucleosome positioning at CES (Ramírez et al., 2015). Furthermore, a non-sex-specific dosage compensation protein, Chromatin-Linked Adaptor for MSL Proteins (CLAMP), regulates chromatin accessibility at CES as well as globally across the entire male X-chromosome (Urban et al., 2017a).

Cleavage Under Targets and Release Using Nuclease (CUT&RUN)

CUT&RUN is a modification of chromatin immunoprecipitation (ChIP) that allows for simultaneous high-resolution chromatin mapping of specific factor localization and probing of the local chromatin environment (Skene and Henikoff, 2017) (Figure 1.2C). CUT&RUN is performed *in situ* under native conditions. It requires only a tenth of the sequencing depth of traditional ChIP-seq and has the ability to map

histone modifications, even within compacted chromatin (Skene and Henikoff, 2017). Furthermore, CUT&RUN may be used to map contact sites at near base-pair resolution and differentiate direct protein binding sites versus indirect sites resulting from long-range interactions (Skene and Henikoff, 2017). This approach has not yet been applied to dosage compensation; however, it has great potential to contribute to better understanding the targeting requirements of various dosage compensation complexes by more accurately identifying binding sites and three-dimensional interactions between transcription factors across species.

Single-cell techniques

Until recently, performing many of the techniques discussed above required large pools of cells. Fluorescence-Activated Cell Sorting (FACS) allows specific marked cell populations to be assayed, but still requires a large number of input cells (Griffiths et al., 2018) (Figure 1.2D). More recently, the modernization of genomic approaches allows them to be performed *in situ*. This has led to the ability to assay the genomic landscape in single cells. Single-cell approaches are powerful; no longer is an average reading recorded from a bulk sample. As a result, cell to cell heterogeneity is uncovered and cells may be clustered into a number of unique groups.

Single-cell techniques are currently being used to probe chromatin accessibility (Buenrostro et al., 2015; Cusanovich et al., 2015), DNA variation (Zong et al.,

2012), transcription factor binding sites (Rotem et al., 2015), methylation (Farlik et al., 2015) and three-dimensional genome structure (Nagano et al., 2013). Additionally, microscopy-based approaches such as single-molecule super-resolution microscopy combined with Oligopaints — specialized FISH probes that allow for super-resolution microscopy and visualization of chromatin organization (Beliveau et al., 2015) — and assay of transposase accessible chromatin with visualization (ATAC-seq) — a modification of ATAC-seq that reveals open chromatin regions targeted by transposons (Chen et al., 2016) — allow for single cell visualization of chromatin architecture. Single-cell methods have not yet been applied to study dosage compensation. However, these approaches have great potential to define how dosage compensation functions in different tissues and non-model organisms.

***Drosophila* dosage compensation as a model for understanding context-specific transcriptional regulation**

Dosage compensation systems provide excellent models for investigating context-specific regulation of genomic loci. In *Drosophila*, dosage compensation occurs via one of the most conserved dosage compensation mechanism across species, which is the X-upregulation. Genes along the single male X-chromosome are upregulated approximately two-fold to equalize with transcription from the two female X-chromosomes. This process is mediated by MSLc.

MSLc is a ribonucleoprotein complex composed of the following male-specific lethal (MSL) proteins: MSL1, MSL2, MSL3, Maleless (MLE), and Males absent on the first (MOF), (Belote and Lucchesi, 1980). One of two functionally redundant long non-coding RNAs (lncRNAs) known as *RNA on the X 1* and *2* (*roX1* and *roX2*) are also required for complex formation (Ilik et al., 2013; Rattner and Meller, 2004). MSL1 is capable of dimerization and along with MSL2 serve as a structural scaffold for the complex (Zheng et al., 2014). A partial complex of MSL1 and MSL2 is sufficient for X-identification but is incapable of spreading (Kelley et al., 1999); the MSL3 component of the complex which contains a chromo domain is required for spreading (Sural et al., 2008). The MSL3 chromodomain has been shown to interact with H3K36me3 (Larschan et al., 2007). H3K36me3 is a histone modification associated with active genes (Consortium, 2007; Roy et al., 2010); therefore, it has been hypothesized that spreading of MSLc to active genes occurs through MSL3 recognition of H3K36me3 (Larschan et al., 2007). Mutations of mammalian homologs of MSL3 have been implicated in many disease states (Basilicata et al., 2018; Leiserson et al., 2013).

In addition to MSL3, the MLE helicase has also been shown to be important for spreading of the complex (Kelley et al., 1999). MLE is a homolog of mammalian RNA helicase A (Lee and Hurwitz, 1993), its association with MSLc is dependent on the *rox* RNAs (Franke and Baker, 1999; Ilik et al., 2013; Quinn et al., 2014).

MOF is a non-sex-specific component of the complex and associates with other transcriptional complexes (Chatterjee et al., 2016). Furthermore, MOF is a histone acetyltransferase that is responsible for acetylating histone H4 at lysine 16 (H4K16ac), another histone modification associated with active genes (Conrad et al., 2012; Taylor et al., 2013). Mammals and *C. elegans* contain homologs of MOF and associations between MOF and H4K16ac are capable of recruiting PolII-S5p, which is associated with transcription initiation at promoters (Deng et al., 2013). The histone acetyltransferase activity of MOF in the context of MSLc is dependent on *roX* lncRNA (Park et al., 2007).

The *roX1* and *roX2* lncRNAs are thought to play additional structural roles in MSLc and interact with MSL2 only after MLE alters their structure (Maenner et al., 2013). The *roX* lncRNAs are functionally redundant, either is sufficient for complex assembly and targeting of the X-chromosome (Ilik and Akhtar, 2009; Ilik et al., 2013). Despite this redundancy these lncRNAs have little similarity in sequence or length. However, the *roX* lncRNAs share a conserved secondary stem-loop structure that is required for the histone acetyltransferase activity of MOF (Conrad et al., 2012; Ilik and Akhtar, 2009; Ilik et al., 2013; Prestel et al., 2010). The genomic loci that encode the *roX* lncRNAs are thought to serve as the initial assembly sites for MSLc assembly (Bai et al., 2004; Kelley et al., 1999). Though the *roX* RNAs are normally transcribed on the X-chromosome it has been shown that this genomic positioning is not required for X-specific targeting of MSLc; the

roX lncRNAs are also capable of acting in *trans* when placed within an autosomal locus (Park et al., 2002). A significant amount is known about the biochemical and *in vivo* activities of the MSLc components; however, much less is understood about how the X-chromosome is specifically targeted for dosage compensation.

X-identification by the Male specific lethal complex (MSLc)

The most strongly enriched sites of MSLc occupancy, which include the *roX* lncRNA loci, are known as CES (Alekseyenko et al., 2008). Nearly all CES contain one or more, approximately, 21-base-pair (bp) GA-rich CREs, known as MSL recognition elements (MREs) (Alekseyenko et al., 2008). MREs are essential for MSLc localization and contain a highly conserved 8-bp GA-rich core motif that is enriched approximately two-fold on the X-chromosome in multiple *Drosophila* species (Alekseyenko et al., 2008, 2013; Ellison et al., 2019; Kuzu et al., 2016). CES are hypothesized to have independently evolved across many *Drosophila* species by GA-dinucleotide repeat expansions and transposition of transposable elements containing GA-rich sequences (Ellison and Bachtrog, 2013; Ellison et al., 2019; Kuzu et al., 2016). Additionally, it has been reported, that mammalian X-chromosomes have also accumulated GA-rich sequences via similar transposition events (D'Souza et al., 2018).

MSLc has been shown to be capable of directly binding MREs with low specificity *in vitro* (Albig et al., 2018). To allow MSLc to bind to MREs within the context of

chromatin, MSLc requires an adapter protein known as Chromatin-Linked Adapter for MSL Proteins (CLAMP) (Larschan et al., 2012; Soruco et al., 2013). CLAMP binds to MREs along the X-chromosome and stabilizes MSLc binding to MREs *in vivo* and *in vitro* (Albig et al., 2018; Soruco et al., 2013). Furthermore, it has been shown that CLAMP and MSLc exhibit cooperative synergy which increases the occupancy of both factors specifically at CES (Soruco et al., 2013). However, CLAMP is not sufficient for X-chromosome specificity because it is not restricted to the X-chromosome and binds to other GA-rich CREs present throughout the genome (Kuzu et al., 2016; Soruco et al., 2013).

Context-specific functions of CLAMP

CLAMP is an essential non-sex-specific protein that has two potential isoforms that differ by only four amino acids (Drysdale et al., 2005; Urban et al., 2017c). Though it is enriched on the X-chromosome, CLAMP binds to GA-rich CREs genome-wide. Furthermore, CLAMP is required for recruitment of the MSLc specifically to the X-chromosome (Soruco et al., 2013). Therefore, CLAMP serves as a model for understanding how transcription factors may function in different contexts such as on the X-chromosome compared with autosomes.

To date there is no evidence that suggests CLAMP is capable of recognizing or binding to any other, non GA-rich, DNA sequences (Albig et al., 2018; Soruco et al., 2013). Within the nucleus, CLAMP has been shown to directly interact, co-

bind, or compete with many other diverse transcription factors (Bag et al., 2019; Kaye et al., 2017; Kyrchanova et al., 2018; Urban et al., 2017b). In addition to its role in dosage compensation (Albig et al., 2018; Soruco et al., 2013), CLAMP has been hypothesized to have additional roles throughout the genome including but not limited to regulation of transcription (Kuzu et al., 2016; Rieder et al., 2017; Urban et al., 2017c), chromatin accessibility (Urban et al., 2017a), and mRNA splicing (Urban et al., 2017b). Understanding the diverse roles of CLAMP at different locations throughout the genome will improve our understanding of the molecular mechanisms underlying the context-specific roles of transcription factors.

I used diverse approaches from fly genetics to machine learning in order to investigate the context-specific function of CLAMP. First, I used luciferase reporter assays and *in vivo* fly genetics to define how subtle variations in CRE sequence composition, distribution, and location influence transcription factor recruitment *in vivo* (Chapter 2). Second, I used machine learning to identify novel candidate cofactors that are likely to influence the context-specific recruitment and function of MSLc on the X-chromosome (Chapter 3). Lastly, I used chromosome conformation capture to reveal new context-specific roles for CLAMP in three-dimensional genome organization and determine that CLAMP specifically regulates long-range interactions between regions of active chromatin including CES (Chapter 4). Overall, I provide key insight into how highly similar

CREs are capable of context-specific transcriptional regulation within a complex biological system.

CHAPTER 2 : CLAMP MOTIF SEQUENCE VARIATION, DENSITY, AND PROXIMITY PROMOTE RECRUITMENT OF THE DROSOPHILA DOSAGE COMPENSATION COMPLEX TO THE MALE X-CHROMOSOME

William Jordan III¹, Erica Nguyen¹, and Erica Larschan¹

I wrote this chapter with editorial feedback from Dr. Erica Larschan.

¹ Department of Molecular Biology, Cell Biology, and Biochemistry, Brown University, Providence, Rhode Island 02912

Abstract

Cis-regulatory elements (CREs) are capable of having many context-specific transcriptional effects. It is not well understood how many identical or highly similar CREs perform different functions dependent on their precise genomic locations. *Drosophila* dosage compensation offers a nice system to interrogate how specific subsets of similar CREs may precisely regulate gene expression in a context-dependent manner. In *Drosophila*, a dosage compensation complex known as MSLc is recruited specifically to the X-chromosome. MSLc is enriched at specific X-linked loci, known as CES. CES often contain a high density of GA-rich CREs known as MREs. MSLc has low affinity for DNA binding *in vivo* and requires the presence of an adapter protein, CLAMP, to stabilize its binding to MREs. Moreover, unlike MSLc, CLAMP is not restricted to the X-chromosome and binds to many other GA-rich CREs. However, in comparison with other similar CREs, MREs are roughly two-fold enriched on the X-chromosome and have the highest genomic density and proximity specifically within CES. Furthermore, it has been demonstrated that the MREs present within CES are necessary and sufficient for MSLc recruitment. No other requirements for MSLc recruitment have been demonstrated *in vivo*. Therefore, we hypothesized that the MRE density and proximity promote accurate MSLc recruitment to the X-chromosome. We made synthetic perturbations of CES and autosomal loci containing various GA-rich CRE arrangements. Our analysis shows that subtle sequence variations, proximity, and density of GA-rich CREs regulate *in vivo* recruitment of MSLc.

Introduction

Transcription factors regulate gene expression by recognizing and binding to specific DNA sequence patterns known as *cis*-regulatory elements (CREs). Many diverse CREs exist; CREs may recruit a single transcription factor or specific transcription factor combinations in order to precisely regulate genes. Only a small percentage of transcription factor recognition sites are actually bound *in vivo* by the transcription factor(s) they are capable of recruiting (Kaplan et al., 2011; Slattery et al., 2014; Wang et al., 2012). Furthermore, the subset of bound sites may vary from cell to cell and function differently depending on their genomic context (Bradley et al., 2010; Slattery et al., 2014; Voss and Hager, 2014). However, little is understood about how transcription factors function to regulate genes in a context-specific manner.

Drosophila sex chromosome dosage compensation offers an excellent model to define how specific subsets of CREs are used to precisely regulate gene expression in a context-specific way because genes along the length of a single chromosome are co-regulated. In *Drosophila*, dosage compensation is mediated by a ribonucleoprotein complex known as the male-specific lethal complex (MSLc) that binds specifically to the X-chromosome. Though MSLc is recruited to specific CREs on the X-chromosome, known as MREs, many highly similar CREs are found throughout the genome (Chapter 1). MSLc is assembled only in males and specifically targets and upregulates transcription of the single male X-chromosome

in order to equalize X-linked gene expression with that of females (Chapter 1). The overall process of *Drosophila* dosage compensation has been proposed to occur in three steps: 1) MSLc recognizes the transcription sites of *RNA on the X (roX) 1* or *2*, which are two functionally redundant long noncoding RNA components of MSLc (Franke and Baker, 1999; Kelley et al., 1999, 2008; Meller and Rattner, 2002). 2) MSLc localizes to specific X-linked GA-rich CREs known as MSLc recruitment elements (MREs). Specific X-linked loci, including the *roX* loci, that have the highest occupancies of MSLc are known as “high-affinity” or “chromatin entry” sites (CES) (Alekseyenko et al., 2008; Straub et al., 2008). 3) MSLc spreads from CES to the bodies of active genes where it increases the rate of transcription elongation at active genes (Hamada et al., 2005; Kelley et al., 2008; Larschan et al., 2007, 2011; Ramírez et al., 2015; Schauer et al., 2017). MSLc requires an essential transcription factor known as chromatin linked adapter for MSL proteins (CLAMP) in order to target CES *in vivo*. Furthermore, synergy between CLAMP and MSLc increases the occupancy of both factors specifically at CES (Albig et al., 2018; Soruco et al., 2013).

CES are enriched for 21-base-pair (bp) MREs (Alekseyenko et al., 2008). MREs are essential for MSLc localization and contain a highly conserved repetitive 8-bp GA-rich core motif that is enriched approximately 2-fold on the X-chromosome in multiple *Drosophila* species (Alekseyenko et al., 2008, 2013; Kuzu et al., 2016). This conservation is striking because the sex chromosomes evolved

independently across many *Drosophila* species (Ellison et al., 2019; Lucchesi, 1978; Zhou et al., 2013). CES also independently evolved in these species by GA-dinucleotide repeat expansions or transposition of transposable elements containing GA-rich sequences (Ellison and Bachtrog, 2013; Ellison et al., 2019; Kuzu et al., 2016). CLAMP binds directly to GA-rich MREs in targeting MSLc specifically to the X-chromosome. However, unlike MSLc, CLAMP is not restricted to the X-chromosome and binds hundreds of similar GA-rich CREs genome-wide (Soruco et al., 2013). GAGA-associated factor (GAF), a similar transcription factor, also binds to GA-rich CREs throughout the genome (Farkas et al., 1994; Kasinathan et al., 2014).

GAF preferentially binds shorter GA-rich sequences of approximately 5-bp, while CLAMP prefers longer GA-rich sequences that are approximately 15-bp in length and also exhibits more tolerance for variations in sequence composition within the GA-repeat (Kaye et al., 2018; Kuzu et al., 2016). Although the conserved core of the MRE motif is just 8-bp, the sequences immediately flanking the MRE core are also required for MSLc and CLAMP recognition because small changes in composition and length of candidate MREs can inhibit CLAMP binding (Alekseyenko et al., 2008; Kuzu et al., 2016). Subtle sequence variation within the GA-rich MREs enhance the binding affinity of CLAMP compared with GAF which is largely excluded from CES (Kaye et al., 2018). Furthermore, no other secondary sequences have been found that contribute or participate in MSLc recruitment

(Alekseyenko et al., 2008). Therefore, it remained unknown how variation within seemingly similar GA-rich CREs allows for X-specific recruitment of MSLc.

The longer GA-rich sequences that CLAMP preferentially binds are two-fold enriched on the X-chromosome (Kaye et al., 2018; Kuzu et al., 2016). Furthermore, given this enrichment, it is not surprising that these sequences are spaced closer to one another on the X-chromosome compared to the autosomes. However, the GA-rich sequences are in highest proximity to one another other specifically within CES, even when normalized for their increased density on the X-chromosome (Kuzu et al., 2016). Therefore, we hypothesized that, in addition to length, motif density (i.e. the number of motifs) and proximity (i.e. the spacing between motifs) promote accurate MSLc recruitment to the X-chromosome.

Here we use a previously validated MSLc-dependent dual luciferase-based reporter assay (Alekseyenko et al., 2008) and *in vivo* insertion of synthetically modified genomic sequences containing GA-rich CREs to better define the precise linear recruitment requirements of MSLc. We showed that the density, as well as proximity, of individual GA-rich CREs drives recruitment of MSLc. Furthermore, we show that recruitment is influenced not only by the overall motif content, but also by variations within DNA sequences flanking MREs. Overall, we provide insight into how slight variations among similar GA-rich CREs found genome-wide promote accurate targeting of MSLc to the male X-chromosome.

Results

CLAMP binding sequences are necessary but not sufficient for MSLc activity

To define the precise targeting requirements of MSLc to CES we used a previously described CES5C2-6 luciferase-based reporter construct (Alekseyenko et al., 2008); this construct holds a 268-bp CES fragment that contains three MREs. Alekseyenko et al. (2008), previously showed that the three MREs present within the CES5C2-6 reporter construct are sufficient for *in vivo* localization of MSLc to an autosomal insertion of CES5C2-6 and for MSLc-dependent luciferase reporter activity, even when the other remaining sequences within the CES are scrambled. Furthermore, mutation of each individual MRE decreases MSLc localization and simultaneous mutation of two or more MREs prevents MSLc localization (Alekseyenko et al., 2008).

We transfected the wild-type CES5C2-6 construct into *Drosophila* male Schneider's line 2 (S2) cells (Schneider, 1972) and compared its activity level following RNAi depletion of *gfp* (control), *clamp*, or *msl2* (Hamada et al., 2005; Larschan et al., 2012; Soruco et al., 2013) in compared with a series of modified versions of the construct (Fig. 2.1A,B): 1) A negative control construct was created by replacing the MREs within the CES with lacO recognition sequences, which do not recruit CLAMP (Rieder et al., 2017) and therefore should not show MSLc-dependent activity (Fig. 2.1A,C). 2) GAF preferentially binds to much shorter GA-

rich sequences than CLAMP (Kaye et al., 2018). Therefore, to test that luciferase reporter activity is dependent on X-linked MREs and the presence of MSLc and CLAMP, we also swapped the conserved core of the MREs with a synthetic motif derived from the consensus GAF sequence (Kasinathan et al., 2014) (Fig. 2.1A).

3) We replaced the entire CES5C2-6 region with the following autosomal loci of similar size that contain multiple GA-rich CREs resembling the MRE where CLAMP or GAF are known to bind: A) the well-studied GA-rich CREs within the *Fab-7* insulator, which are known to bind to CLAMP (Kaye et al., 2017; Kyrchanova et al., 2018; Wolle et al., 2015) (Fig. 2.1B); B) GA-rich CREs present at the *Drosophila* histone locus body (HLB) that are required for histone transcription (Duronio and Marzluff, 2017; Rieder et al., 2017; Salzler et al., 2013).

First, we confirmed that activity of the wild type CES5C2-6 reporter is dependent on MSLc and CLAMP (Fig. 2.1C). Next, we determined that GAF recognition sequences or autosomal locations that contain multiple GA-rich CREs are not capable of driving MSLc-dependent reporter expression (Fig. 2.1D). Interestingly, for the full 1.2-kb length *Fab-7* fragment and smaller subsections, we observed a consistent increase in reporter activity after *clamp* RNAi (Fig. 2.1D). These data suggest that CLAMP may be necessary for the repressive enhancer-blocking activity of *Fab-7*, which is known to be dependent on the presence of a subset of GA-rich motifs present within *Fab-7* (Wolle et al., 2015). Furthermore, CLAMP is

part of an insulator complex called the late boundary complex that is present at *Fab-7* (Kaye et al., 2017) and has insulator activity (Bag et al., 2019).

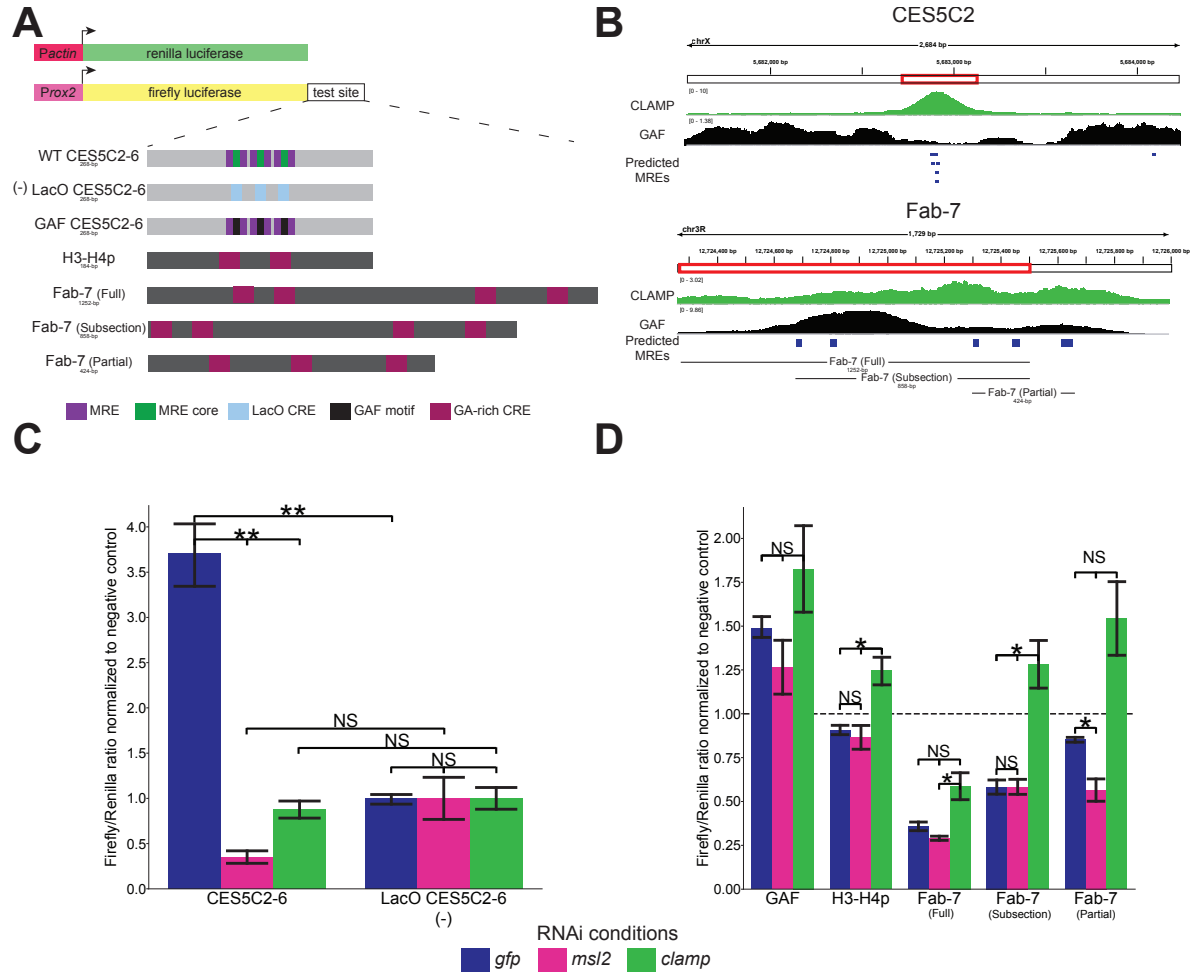


Figure 2.1. CLAMP specific GA-rich CREs required for MSLC activity.

(A) Experimental design of constructs used for dual luciferase reporter assay. **(B)** ChIP-seq profiles for CLAMP and GAF in S2 cells along with predicted MREs for CES5C2 and the *Fab-7* locus. **(C)** Control normalized luciferase activity for wild-type CES5C2-6 and LacO CES5C2-6 that has lacO recognition sequence in place of MREs following RNAi. **(D)** Control normalized luciferase activity for CES5C2-6 that has consensus GAF motif in place of core sequences within MREs along with subsets of autosomal loci containing GA-rich CREs following RNAi. For all bar plots error bars represent +/- standard error of the mean (S.E.M). * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$, Welch's t-test.

MSLc activity is maintained in the context of increasing the density of MREs

On the X-chromosome, MREs are present at the highest density within CES (Kuzu et al., 2016). It has been reported that the number of homo- or heterotypic binding sites for a transcription factor may result in increased activity (Grskovic et al., 2007; Sharon et al., 2012; Smith et al., 2013). Additionally, in some cases, the presence of too many recognition sites for an activating transcription factor has the potential to have an inverse effect and result in repression instead of activation (White et al., 2016). Of the 29 CES tested by Alekseyenko et al. (2008), CES5C2-6 has the strongest activity, correlating with this locus containing three MREs. Therefore, we asked whether transcriptional activity of this system was already saturated or if the presence of additional MRE motifs within CES5C2-6 would result in further MSLc-dependent transcriptional activation. We added six additional MREs with four motifs that were derived by duplicating two of the MREs already present in 5C2; the other two MREs were duplicated from another previously tested CES, CES11D1 (Alekseyenko et al., 2008) (Fig. 2.2A). We observed a significant increase in MSLc activity that is reduced following RNAi knockdown of *msl2* or *clamp* (Fig. 2.2B). However, this construct with additional MREs exhibits increased reporter activity relative to the wild type CES5C2-6 reporter prior to and following RNAi treatment (Fig. 2.2B). This increased residual activity could be due to incomplete RNAi knockdowns and any residual protein remaining localizing to any of the nine recognition sequences present at the synthetic locus. Therefore, it remains inconclusive whether additional MREs lead to saturation of MSLc-

dependent transcriptional activity. However, we show that the presence of additional MREs does not lead to an inverse repressive effect.

Close proximity of MRE motifs to each other promotes MSLc activity

Given their high density within CES, MREs are also in closest genomic proximity to one another at CES (Kuzu et al., 2016). To determine whether or not the close proximity of MREs within CES is required for their ability to recruit MSLc, we tested the spacing requirements for MREs in the context of the CES5C2-6 reporter. Within CES5C2-6, the three MREs are located in close proximity to each other near the center of the fragment with approximately 5-bp separating each motif. We spaced these MREs, approximately, 9X and 18X further apart than their normal spacing; we also placed two, equally spaced, lacO sequences between each MRE in an attempt to avoid the creation of cryptic MSLc recognition sequences (Fig. 2.2A). When the spacing between MREs was increased to 46-bp we observed a 2-fold reduction in reporter activity (Fig. 2.2C). When proximity was further reduced by spacing the motifs 92-bp apart we observed a further loss in reporter activity (Fig. 2.2C). Therefore, close proximity of MREs is more optimal for promoting MSLc activity of the CES5C2-6 reporter.

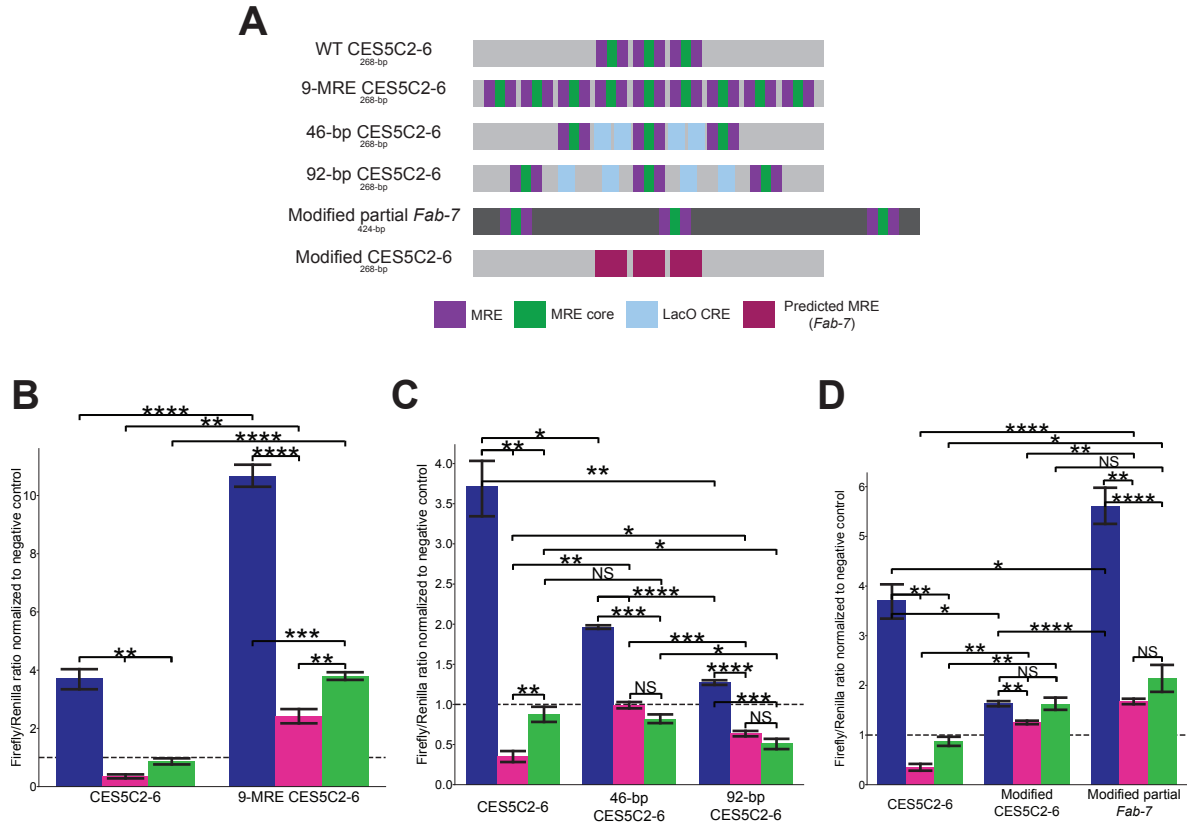


Figure 2.2. Close proximity of MREs within CES confers higher MSLc activity.

(A) Design of sequences tested. **(B)** Control normalized luciferase activity after increasing MRE density within CES5C2-6 following RNAi. **(C)** Control normalized luciferase activity after decreasing the proximity of MREs within CES5C2-6. **(D)** Control normalized luciferase activity after *Fab-7* GA-rich CRE and CES5C2-6 MRE swaps following RNAi. For all bar plots error bars represent $\pm$ standard error of the mean (S.E.M). * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$, Welch's t-test.

Flanking sequences can modulate the activity of MREs that are not in close proximity to each other

To test whether any GA-rich CRE is sufficient for MSLc activation in the context of a CES, we swapped the three MREs present within CES5C2 with three GA-rich CREs present within a 400-bp segment of the *Fab-7* insulator (Fig 2.2A). *Fab-7* acts as a silencer and has a well-established role in enhancer blocking that requires the presence of GA-rich CREs and CLAMP is present at the locus as a member of a large protein complex (Kaye et al., 2017; Kyrchanova et al., 2018;

Wolle et al., 2015) (Fig 2.1B). In contrast to the 5-bp spacing at CES5C2-6, the GA-rich CREs we selected within the 400-bp subsection of *Fab-7* are normally much further from each other (117-bp and 151-bp separation) (Fig 2.2A).

We previously observed that increased MRE proximity promotes MSLc-dependent reporter activity (Fig 2.2C). Therefore, we hypothesized that, given the closer proximity of the GA-rich CREs in the modified CES5C2-6 construct with *Fab-7* GA-rich CREs, we would observe an increase in MSLc reporter activity (Fig 2.2A). Furthermore, given the decreased MRE proximity, we also hypothesized that we would observe a decrease in MSLc activity for the modified partial *Fab-7* sequence (Fig 2.2A).

Consistent with our hypothesis, MSLc-dependent reporter activity was observed in the modified CES5C2-6 construct (Fig 2.2A, D). However, MSLc-dependence of the reporter activity was of a much smaller magnitude than seen with the wild type MREs for CES5C2-6 (Fig 2.2D). Additionally, reporter activity for this construct was not CLAMP-dependent (Fig 2.2D). Therefore, the specific DNA sequences that constitute the MREs within CES are important and autosomal CLAMP binding sites are not fully functional when placed within CES.

Remarkably, in contrast to our hypothesis, significant increases in MSLc and CLAMP-dependent activity were observed for the modified partial *Fab-7* sequence

that contained MREs from CES5C2-6 (Fig. 2.2D). Moreover, this construct which has MREs that are spaced much farther apart than at CES5C2 showed much stronger MSLc and CLAMP-dependent activity than even that of the wild-type CES5C2-6 reporter (Fig. 2.2D). This observation was unexpected because the MREs inserted into *Fab-7* are spaced even further apart than the 92-bp spacing within CES5C2-6 that displayed significant reductions in MSLc activity compared to wild type CES5C2-6 (Fig. 2.2C). We also did not observe any increases in reporter activity following *clamp* RNAi in contrast to the wild type construct (Fig. 2.2D, Fig. 2.1D). It is possible that presence of other CREs results in much more efficient recruitment of CLAMP and MSLc or that repressive sites were ablated. Overall, we find that closer proximity of GA-rich CREs is optimal for MSLc targeting but there are contexts in which flanking DNA sequence can overcome sub-optimal spacing to successfully increase MSL-dependent activity.

Next, we asked whether we could turn an autosomal site containing GA-rich CREs, into a CES by reducing the spacing between the CREs. Therefore, we synthetically modified two autosomal locations of similar size (268-bp) to the CES5C2-6 region in our reporter construct and the previously described 400-bp partial *Fab-7* sequence (Fig 2.1B, D). Each sequence contained three GA-rich CREs, resembling MREs (Soruco et al., 2013) (Fig. 2.3A). We name the selected autosomal sequences “autosomal sequence” 1 and 2 (AS-1 and AS-2). AS-1 is located at chr2L:22,024,319-22,024,586 and AS-2 is located at chr2L:22,377,310-

22,377,579 (Fig. 2.3A). For each sequence we synthetically relocated the GA-rich CREs resembling MREs to the center of the sequence and increased their proximity to one another. We hypothesized that making autosomal sequences more “CES-like” may result in MSLc activation at some sites.

The endogenous AS-1 sequence exhibited low but MSLc-dependent reporter activity. Reporter activity was not CLAMP-dependent, as a small, but not significant, increase was seen in reporter activity following *clamp* RNAi (Fig. 2.3D). When the GA-rich CREs of AS-1 were placed in closer proximity, a significant, approximately two-fold MSLc-dependent increase was observed in reporter activity (Fig. 2.3D). Though not statistically significant, a moderate reduction in reporter activity was observed following *clamp* RNAi (Fig. 2.3D).

In verifying our constructs, we discovered an unannotated insertion within AS-2. This could be due to using S2 cell genomic DNA (gDNA) to clone our wild-type genomic sequences of interest. The modified construct which was created synthetically does not contain this insertion. The AS-2 sequence exhibited moderate MSLc and CLAMP-dependent reporter activity (Fig. 2.3E). The modified AS-2 sequence also exhibited moderate reporter activity but in contrast this activity was solely MSLc dependent and we observed no differences in reporter activity following *clamp* RNAi. Due to differences between AS-2 and modified AS-2 due to

the gDNA insertion, we do not draw any conclusions from this particular set of constructs.

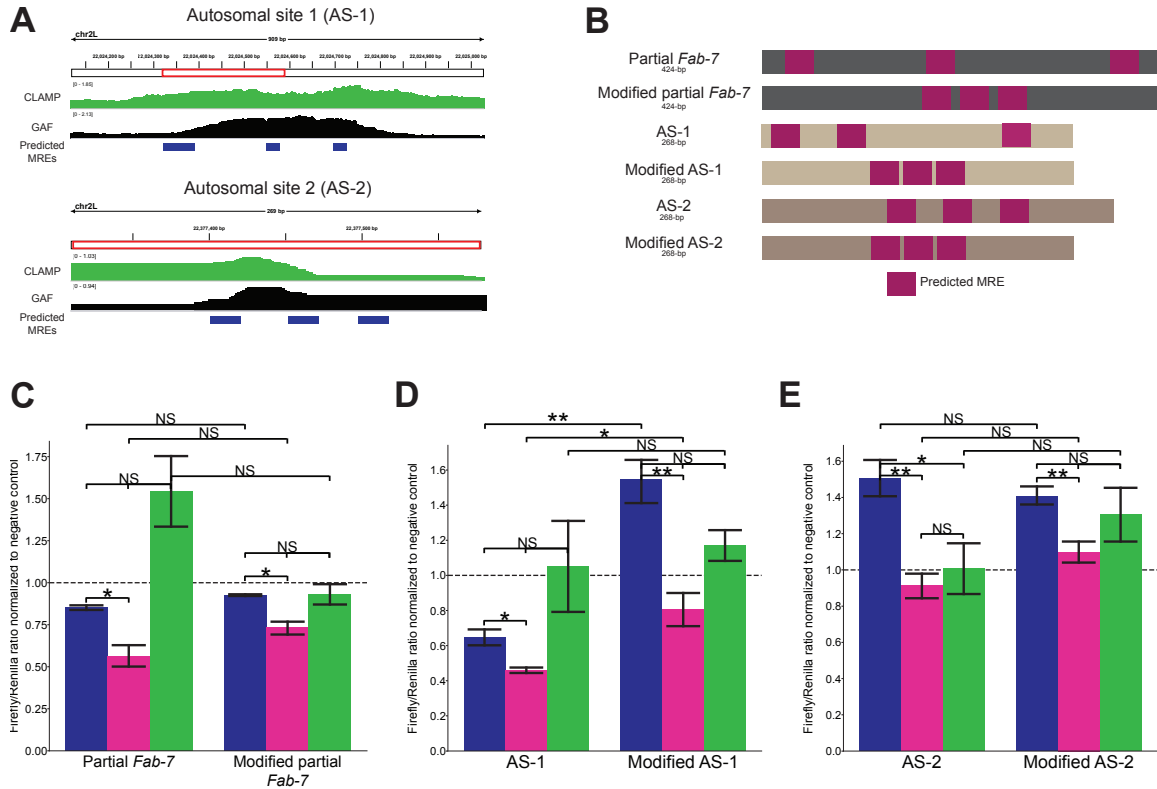


Figure 2.3. Artificially increasing proximity of autosomal GA-rich CREs in results in increased MSLc activity.

(A) ChIP-seq profiles for CLAMP and GAF in S2 cells at selected autosomal loci containing multiple GA-rich CREs with high MRE similarity. (B) Design of sequences tested. (C) Control normalized luciferase activity of modified partial *Fab-7* sequence following RNAi. (D) Control normalized luciferase activity of modified AS-1 following RNAi. (E) Control normalized luciferase activity of modified AS-2 following RNAi. For all bar plots error bars represent +/- standard error of the mean (S.E.M). * p<0.05, ** p<0.01, *** p<0.001, **** p<0.0001, Welch's t-test.

For the wild type partial *Fab-7* sequence, we previously observed modest MSLc-dependent reporter activity (Fig. 2.1D, Fig. 2.3C). We also observed a moderate, non-statistically significant, increase in reporter activity following *clamp* RNAi (Fig 2.1D). The new version of the partial *Fab-7* sequence with closer GA-rich CRE spacing maintained modest MSLc-dependent reporter activity (Fig 2.3C).

However, we no longer observe any increase in reporter activity following *clamp* RNAi (Fig 2.3C). These findings are consistent with the results we obtained after swapping the GA-rich CREs from the partial *Fab-7* sequence in place of the endogenous MREs present at close proximity at CES5C2-6 (Fig. 2.2D). Overall, we conclude that some autosomal GA-rich CREs such as those within AS-1 are sufficient for MSLc-dependent transcriptional activation; however, these sites have other features such as suboptimal proximity or neighboring sequences that reduce their potential for activation by MSLc.

***In vivo* analysis of MRE proximity and density modifications**

In order to test whether modifications of the spacing between MRE or GA-rich CREs lead to loss or gain of MSLc recruitment, we selected the following modified sequences for *in vivo* study: 1) modified CES5C2-6 with MREs (46-bp spacing); 2) modified CES5C2-6 with MREs (95-bp spacing); 3) modified CES5C2-6 with 6 additional MREs and 3) modified AS-1 (Fig. 2.2A). As a negative control for the modified CES5C2-6 sequences, we used the control CES5C2-6 construct in which MREs had been swapped with lacO recognition sequences and we used the wild-type CES5C2-6 sequences as a positive control (Fig. 2.2A). For the modified autosomal construct, we also used the wild-type autosomal sequence as a negative control (Fig. 2.3B). We inserted each construct onto chromosome 2 at cytological location 37B7 using the PhiC31 integrase system in combination with recombinase-mediated cassette exchange (Bateman and Lee, 2006).

We assayed our transgenic insertions at the level of polytene chromosome immunostaining with anti-CLAMP and anti-MSL3 antibodies. If MSLc localizes to any of our autosomal insertions, we would expect to see a single, consistent, autosomal MSL3 band in addition to MSL3 signal on the single male X-chromosome; we would also expect to detect a CLAMP band at that locus. We did not detect any aberrant MSL3 autosomal staining in our control LacO-CES5C2-6 insertion animals (Fig. 2.4A). We also detected consistent, strong, MSL3 autosomal signal in our wild type CES5C2-6 insertion animals (Fig. 2.4B). Our polytene spreads for our modified 9-MRE CES5C2-6 sequence insertion animals consisted of many thin and fragile polytene chromosomes. Despite this fragility we were able to detect autosomal MSL3 signal for one set of chromosomes (Fig. 2.4C). We also were able to detect strong MSL3 signal for the modified CES5C2-6 insertion with 46-bp spacing (Fig. 2.4D). We were unable to detect any strong autosomal MSL3 signal in our polytene spreads for animals corresponding to the modified CES5C2-6 insertion with 92-bp spacing, which exhibited lower MSLc-dependent luciferase activity (Fig. 2.4E).

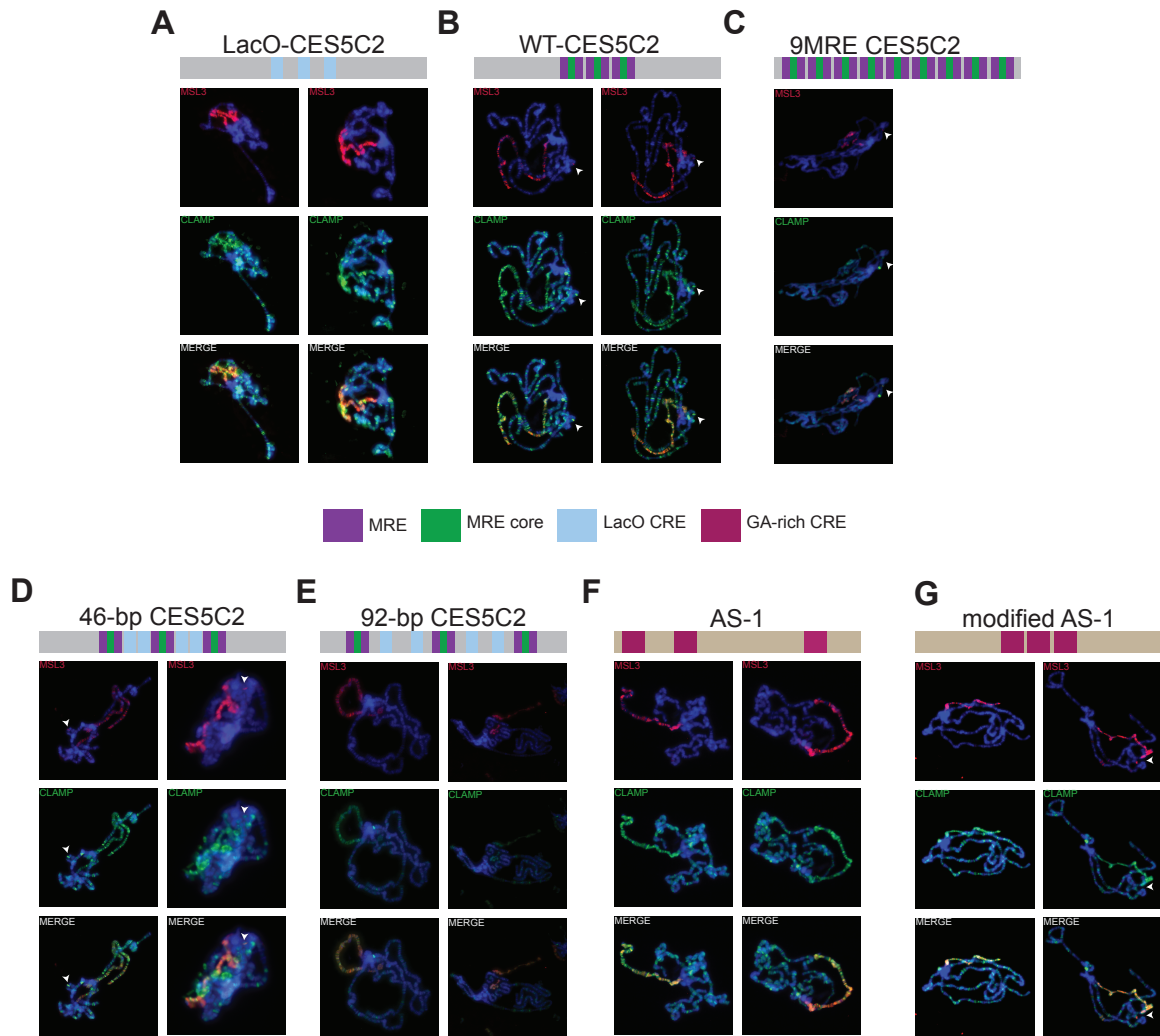


Figure 2.4. Select *in vivo* insertions validate luciferase reporter findings.

(A) No robust aberrant MSL3 staining is detected for negative control LacO-CES5C2-6 insertion. (B) Robust aberrant MSL3 staining is detected for positive control WT-CES5C2-6 insertion. (C) Robust aberrant MSL3 staining is detected for 9MRE CES5C2-6 insertion. (D) Robust aberrant MSL3 staining is detected for 46-bp CES5C2-6 insertion. (E) No robust aberrant MSL3 staining is detected for 92-bp CES5C2-6 insertion. (F) No robust aberrant MSL3 staining is detected for control AS-1 insertion. (G) Variable aberrant MSL3 staining is detected for modified AS-1 insertion.

As expected, we did not detect any strong MSL3 signal in our wild-type AS-1 control animals. Our modified AS-1 sequence with the GA-rich CREs synthetically placed in closer proximity to each other which exhibited modest MSLc-dependent

reporter activity exhibited high background levels of MSL3; in one case we were able to detect a strong autosomal MSL3 band (Fig 2.4G). We plan to repeat the polytene staining for this line. Overall, our *in vivo* analysis recapitulates the findings of our luciferase reporter experiments and demonstrates that close proximity of GA-rich CREs is optimal for MSLc targeting *in vivo*. Furthermore, sequences flanking GA-rich CREs can also influence MSLc targeting in the context of sub-optimal spacing between GA-rich CREs.

Discussion

We determined that precise sequence content, proximity, and density of MREs regulates MSL complex recruitment. However, much remains to be discovered in investigating how precise sequence variations of MRE flanking sequences contribute to MSLc recruitment in the presence of sub-optimal motif proximity or sequence. In addition to dosage compensation, GA-rich CREs have been shown to have many other functions throughout the genome including functions at insulators (Kaye et al., 2017; Wolle et al., 2015) and at the histone locus body (Rieder et al., 2017). We currently provide insight that simple spacing arrangements between certain GA-rich CREs is sufficient to promote MSLc recruitment in some genomic contexts. We also determine that not all GA-rich CREs are sufficient for MSLc recruitment, even when they are optimally spaced. This likely occurs because small variations in sequence composition can perturb recruitment of CLAMP or MSLc. Our current work is limited because luciferase

activity and MSL3 staining are not sensitive enough to determine the level of enrichment or spreading of MSLc in the context of our cell culture and *in vivo* experiments. Therefore, future work will require performing ChIP-qPCR on our transgenic animals to better quantify how MSLc localization is perturbed in the context of sub-optimal MRE proximity. This will also allow us to determine whether the three native MREs present in CES5C2-6 are sufficient for saturation of MSLc recruitment or if the presence of additional MREs leads to higher local enrichment of MSLc or spreading of the complex into neighboring regions. We will also retest our luciferase reporter constructs in female cells; a system that will remove the potential for incomplete RNAi knockdown and enable better assessment of MSLc-dependent reporter activity. We will also generate additional constructs to further define the requirements for the context-specific recruitment of MSLc to X-linked sequences.

Methods

Plasmids and cloning

Wild-type sequences were generated from S2 cell genomic DNA, otherwise all plasmids and predicted MRE sequences were obtained or generated as previously described (Alekseyenko et al., 2008).

Dual Luciferase Reporter Assay

Luciferase reporter assays were conducted as previously described (Alekseyenko et al., 2008). Generation of dsRNA targeting *gfp* (control), *msl2*, and *clamp* for RNAi have been previously validated and described (Hamada et al., 2005; Larschan et al., 2012; Soruco et al., 2013). PCR product was used as template to generate dsRNA with an ambion T7 MEGAscript kit (ThermoFisher Scientific). dsRNAs were purified following DNase treatment with a Qiagen RNeasy kit (Qiagen).

Fly Genetics and Transgenesis

Transgenic animals were generated using recombination-mediated cassette exchange as previously described (Alekseyenko et al., 2008; Bateman and Lee, 2006).

Immunodetection

Polytene squashes and subsequent immunodetection were performed as previously described (Kaye et al., 2018).

CHAPTER 3 : UNCOVERING FACTORS THAT PROMOTE IDENTIFICATION OF THE *DROSOPHILA* X-CHROMOSOME

William Jordan III¹, Sorin Istrail², and Erica Larschan¹

I wrote this chapter with editorial feedback from Dr. Erica Larschan.

¹ Department of Molecular Biology, Cell Biology, and Biochemistry, Brown University, Providence, Rhode Island 02912

² Center for Computational Molecular Biology and Computer Science Department, Brown University, Providence, Rhode Island 02912

Abstract

The mechanism by which transcript levels of X-linked genes are equalized between XY males and XX females is known as dosage compensation. A key first step for dosage compensation mechanisms across species is identifying the X-chromosome. In *Drosophila*, the Male Specific Lethal complex (MSLc) mediates dosage compensation by binding to the X-chromosome and increasing gene expression roughly two-fold. However, MSLc lacks the ability to strongly bind DNA and instead relies on interactions with a zinc finger adapter protein, Chromatin-linked Adapter for MSL proteins (CLAMP). In contrast with MSLc, CLAMP is not restricted to the X-chromosome and binds loci throughout the genome. Therefore, it remains a mystery how MSLc localizes specifically to the X-chromosome. To further investigate this phenomenon, I sought to identify candidate genomic features or transcriptional regulators that may have roles in directing the X-specific recruitment of MSLc. I used many different types of next-generation sequencing based data and a supervised machine learning approach in order to find novel candidates that may have novel roles in X-identification. My work identifies many features that have been tangentially linked to X-identification in *Drosophila* or other species as well as novel candidates without established roles in X-identification.

Introduction

Sexual reproduction is critical for maintaining genetic diversity within species. In X/Y heterogametic species, females have two X-chromosome (XX genotype) and males have one X-chromosome and one Y-chromosome (XY genotype). Many essential genes are encoded on the X-chromosome, however, there exists an imbalance in the number of X-chromosomes in females versus males. As a result, the expression (dosage) of X-chromosome genes must be adjusted to balance gene expression between the sexes. This regulatory mechanism is broadly known as dosage compensation. Many dosage compensation mechanisms have evolved in different species to compensate for sex chromosome imbalanced gene expression. Despite differences, a key first step exists for all dosage compensation mechanisms: identifying the X-chromosome for precise regulation. How this conserved initial step of X-identification is achieved is not well understood.

In the fruit fly *Drosophila melanogaster*, the Male Specific Lethal (MSLc) complex mediates dosage compensation by first localizing to specific sites on the single male X-chromosome known as “chromatin entry” sites (CES) (Alekseyenko et al., 2008; Soruco et al., 2013). The complex then spreads out from these sites, eventually coating the entire X-chromosome, ultimately resulting in an approximately two-fold increase in gene expression of male X-linked genes, balancing with that of the two active X-chromosomes in females (Conrad et al., 2012; Hamada et al., 2005; Straub and Becker, 2007). MSLc lacks the ability to

strongly bind DNA and instead relies on interactions with a DNA-binding zinc finger protein, Chromatin-linked Adapter for MSL proteins (CLAMP), to target the X-chromosome (Albig et al., 2018; Soruco et al., 2013; Villa et al., 2016). However, unlike MSLc, CLAMP targets sites throughout the genome. There is a modest enrichment of CLAMP on the X-chromosome compared to autosomes which is not sufficient for specifically targeting MSLc to the male X-chromosome (Kuzu et al., 2016; Soruco et al., 2013). Therefore, I used machine learning to uncover new candidate factors that may allow MSLc to distinguish the X-chromosome from autosomes.

Materials and methods

Data collection, organization and cleaning

To further investigate the process of X-identification, I used supervised machine learning, to model how MSLc specifically targets the *Drosophila* male X-chromosome. In order to build a comprehensive model, a variety of *Drosophila* derived data sources and types from a variety of cell lines or embryonic stages was obtained. The data was primarily comprised of public next-generation sequencing (NGS) data available through either the NCBI Gene Expression Omnibus (Barrett et al., 2012) or the Encyclopedia of DNA Elements (Consortium, 2007), a complete list of the data initially examined can be found in supplemental file 1. Some of the types of data initially examined include gene expression (Wang et al., 2009), measures of chromatin accessibility (Buenrostro et al., 2013;

Mieczkowski et al., 2016), and transcription factor binding intensity (Davie et al., 2015; Davies et al., 2017; Meyer and Liu, 2014). To date, NGS data have been utilized in building successful machine learning models in predicting DNA binding motifs, enhancers, DNA shape, etc. (Chiu et al., 2015; Greenbaum et al., 2007; Kumar and Bucher, 2016; Liu et al., 2016; Mourad et al., 2018; Whitaker et al., 2014).

In total over 380 potentially useful NGS datasets were identified (supplemental file 1). In order to determine which data would be used for model building, datasets were limited to those that had already been processed and normalized by the depositing authors because it would require a considerable computational resources to reprocess each dataset from the raw data. The tradeoff is that the data were in various text or binary formats including, but not limited to, fasta, bed, bedgraph, wiggle, bigWig, etc. Each dataset was inspected using the integrative genomics viewer (IGV) (Thorvaldsdóttir et al., 2013) as a simple visual check to ensure that the data at least appeared to be properly input-normalized and there were no visually noticeable X-chromosome differences compared to the autosomes. Some exceptions were made for features that are known to exhibit an X-specific bias, such as MSLc members and other associated features (Figure 3.1). Such features were initially included to serve as controls. A combination of tools including, but not limited to, UCSC genome browser's bigWigToBedGraph utility (Kent et al., 2002, 2010), bedtools (Quinlan, 2014; Quinlan and Hall, 2010),

and BEDOPS (Neph et al., 2012) were utilized in standardizing the NGS datasets into an easy to handle, tab separated, table format.

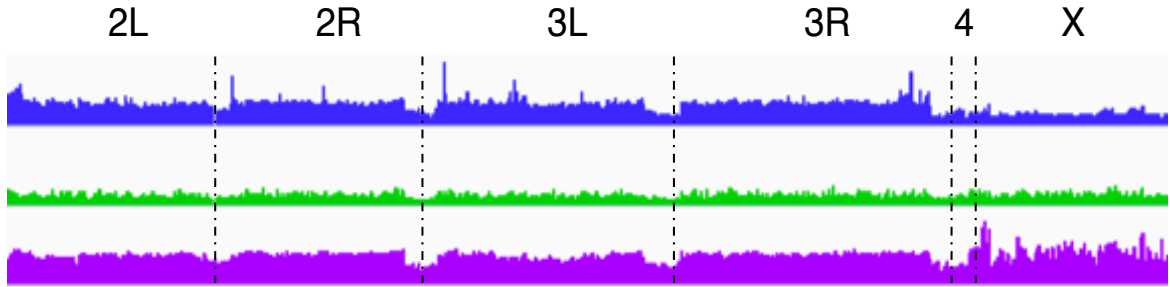


Figure 3.1. Example of visual data integrity checks using ChIP-seq profiles.

The top track (blue), Chromator, represents a perspective feature that has deficient signal on both the 4th and X-chromosomes which could be due to improper input normalization. The middle track (green), H3K27ac, serves as an example of a ideal dataset that does not appear to have any specific chromosome biases. The bottom track (purple), MSL3, which is an example of a feature known to exhibit X-chromosome enrichment.

The *Drosophila* genome, release 5 (dm3) (Hoskins et al., 2007), was sectioned into 300 base-pair (bp) non-overlapping bins and the NGS datasets were mapped into the appropriate bins. Signals spanning multiple bins were averaged where necessary. In addition to the NGS data, primary DNA sequence information obtained from the UCSC genome browser was also utilized (Kent et al., 2002). DNA shape analysis was performed on the sequences represented within each bin using DNASHapeR to predict the following structural features: minor groove width, helix twist, propeller twist, Roll, and electrostatic potential (Chiu et al., 2015). DNASHapeR was also used to one-hot encode the primary DNA sequence associated with each bin. Lastly, the occurrence of mono-, di-, tri- and penta-nucleotides within the sequence represented by each bin was calculated. If the corresponding DNA sequence of a position within a bin contained an N, meaning

the actual DNA base pair (A, G, T, or C) was unknown in the reference genome, that particular k-mer was excluded from the final count.

Exploratory Analysis

Initially there were a total of 432,283 bins or rows in the dataset. Missing values in each bin were imputed separately for each chromosome arm using forward and backward linear interpolation with a distance limit of 20 bins (i.e. 6-kb). Data features originating from different datasets targeting the same transcription factor were scaled and averaged. Any remaining rows containing null values were removed. Finally, the correlation space for the NGS features in the data was examined. Because the feature space contained a few known MSLc components or other features displayed an X-specific bias, the correlation threshold was set in such a way as to ensure that at least one feature known to exhibit X-specific bias would be discarded from the dataset. Using this methodology, a correlation threshold of 0.75 was set, meaning that any feature that exceeded a Pearson correlation of 0.75 (absolute value) with any other feature or the X-chromosome was to be discarded. As a result, at this step a total of three features were discarded including just one of the six known MSLc components in the feature space. No additional MSLc components were excluded at this step in order to serve as positive controls and also set a baseline for evaluating model performance.

The final *Drosophila* dataset contained a little over half the original number of bins for total of 236,473 bins each containing 3,076 total features. An 80/20 split was used to divide the data into a training and test set respectively. This resulted in 189,179 bins in the training set and 47,294 bins in the held-out test set. The X-chromosome represents just one of the possible chromosomes and therefore the dataset has a very heavy class imbalance. Overall, the X-chromosome represents roughly 13% of the *Drosophila* genome, however in the cleaned dataset X-chromosome sequences account for just 7.13% of the bins in the training set and 7.27% in the test set. Due to this heavy class skew, a naive model that simply predicts “False”, i.e. not belonging to the X-chromosome, 100% of the time, would seemingly perform rather well with approximately 93% accuracy. Therefore, the following criteria were used to evaluate model performance: recall, precision and area under the receiver operating characteristics curve (AUC). Recall is the number of true positives divided by the sum of true positives and false negatives. Precision is the number of true positives divided by the sum of true positives and false positives (Buckland and Gey, 1994; Davis and Goadrich, 2006). AUC is a measure of a model’s probability to distinguish between two groups (Buckland and Gey, 1994; Davis and Goadrich, 2006)..

Training a model on the full training set with such a large feature space was very time intensive. Even after incorporating GPUs, the first initial rounds of training required over 16 hours. In order to speed up training, the dimensionality of the

dataset was reduced. A covariance matrix was created using the DNA shape prediction data. From this matrix, eigenvectors were calculated to determine the optimal number of principal components after which no additional gains in variance would be obtained. This resulted in the collapse of the original 1,497 DNA shape features to 471 features. This much smaller feature space retained 93.4% of the overall variance of the original feature space. Similar analysis was performed with the one-hot encoded DNA sequence that represented a total of 1200 features. After determining that including an optimal number of principal components would only reduce the feature space by 300, I decided not to reduce the feature space. Any slight gain in computational time resulting from a small reduction in feature space would not be worth the potential trade off of being able to determine if any specific combinations of DNA nucleotides proved important for X-identification. After a few trials during the initial phase of model testing and selection, I decided to include only penta-nucleotide counts in the final dataset as this group was consistently more predictive than other k-mer features (data not shown). The penta-nucleotide counts represented an additional 256 features, for a total of 1,977 features in the final dataset.

Model Building and Validation

In order to select a supervised learning model, I performed a literature search on the performance of supervised learning (Camacho et al., 2018; Caruana and Niculescu-Mizil, 2006; Chen and Guestrin, 2016; Chiu et al., 2015; Greenbaum et

al., 2007; Liu et al., 2016; Mourad et al., 2018; Nielsen, 2016; Whitaker et al., 2014). I decided to use a well-known boosted tree-based model, eXtreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016; Mitchell and Frank, 2017). XGBoost is a widely used algorithm and has had widespread success in many machine learning competitions. Furthermore, the algorithm is invariant to input scaling, so there is less concern for feature normalization; it is also able to capture higher order interactions within very high dimensional feature spaces. When applying this model, binary target prediction was utilized. For example, even though *Drosophila melanogaster* has one X-chromosome and three autosomes, a sequence is labeled as 1, i.e. “True”, if it originates from the X-chromosome and as 0, i.e. False, if it originates from any of the autosomes.

For the first full training rounds on the finalized dataset, I utilized a Python implementation of XGBoost (Chen and Guestrin, 2016; Millman and Aivazis, 2011; Mitchell and Frank, 2017; Oliphant, 2006, 2007; Pedregosa et al., 2011; Pérez and Granger, 2007; Van Der Walt et al., 2011). The parameters used were estimators: 1000, tree method: gpu_hist, evaluation metric: mean average precision and early stopping rounds: 25 (stops training if the evaluation metric score has not improved after 25 rounds); otherwise default parameters and settings were used. When evaluated using the held-out test set, this base model achieved an overall accuracy of 97.3% with a precision of 0.919, recall of 0.694 and AUC of 0.983. Using the training data, fifty iterations were performed with five-fold cross validation and

randomized grid search to determine more optimal model parameters. The full parameter search space can be seen in Table S3.1. The following are the parameters that were selected as optimal: subsample 0.8, scale_pos_weight 10, reg_lambd 0.1, reg_alpha 0.1, min_child_weight 1. On the held-out test this tuned full model achieved an overall accuracy of 99.1% with a precision of 0.93, recall of 0.798, and AUC of 0.991. The gain in recall demonstrates that this tuned model is much better than the base model at discriminating the X-chromosome from the autosomes. Examining the rankings of gain (improvement in training loss or the accuracy of tree splits that use a given feature), among the top 15 features related to gain, all but 1 of the known MSLc components is captured as being important in the model's decision-making process (Figure 3.2A). Many of the other features identified have been previously associated with X-chromosome identification in *Drosophila* or other heterogametic species (Brown and Bachtrog, 2014; Deng et al., 2014; Disteché, 2012; Kaye et al., 2018; Mendjan and Akhtar, 2007; Soruco et al., 2013; Wells et al., 2012).

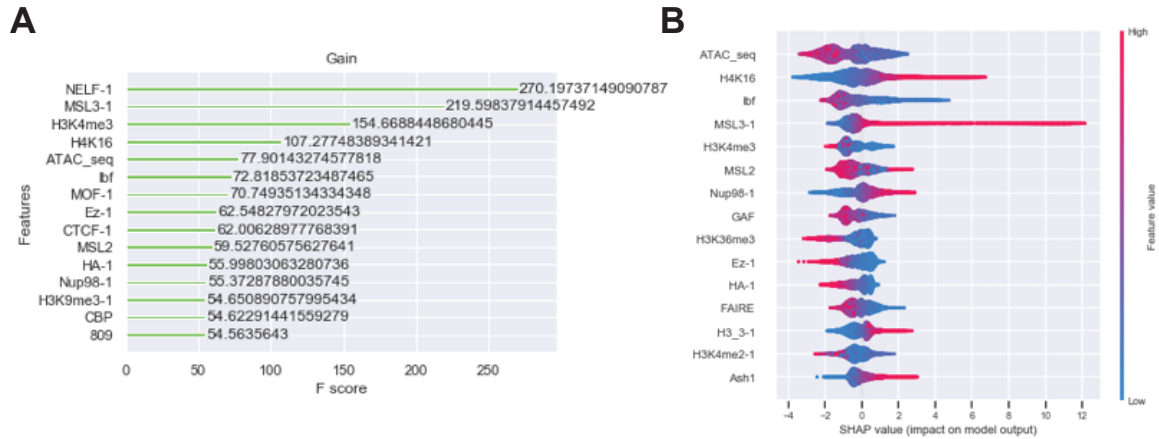


Figure 3.2. Feature importance rankings for the full model

(A) Top predictive feature rankings as measured by gain. The higher the gain the more a given feature contributes to predictive accuracy. **(B)** Top predictive feature rankings as measured by SHAP. Red represents a high value for a given feature compared to other bins and blue a low value. A negative SHAP signifies that a given feature value contributes to reducing the probability a given genomic locus is predicted to belong to the X-chromosome. A positive SHAP value signifies that a given feature value contributes to increasing the probability a given genomic locus is predicted to belong to the X-chromosome.

To further interrogate the model's output, SHapley Additive exPlanations (SHAP), was utilized to rank feature importance (Lundberg and Lee, 2017). SHAP analysis of the model's performance also captures many features with known roles in X-identification (Figure 3.2B).

Model Application and Interpretation

Once the chosen modeling approach was demonstrated to accurately identify known features that contribute to X-identification, those features known to directly make up the MSLc were removed, as well as H4K16ac that is deposited by the MOF component of MSLc. Overall the following features were removed: MSL2, MSL3, MLE, MOF, and H4K16ac (MSL1 was previously removed with correlation cutoff). Another round of 5-fold cross-validated grid search was performed to

ensure optimal parameters were again utilized for the next round of this null MSLc training. Nearly identical parameters were selected as optimal, therefore the null MSLc model was trained with the same parameters of the full MSLc model.

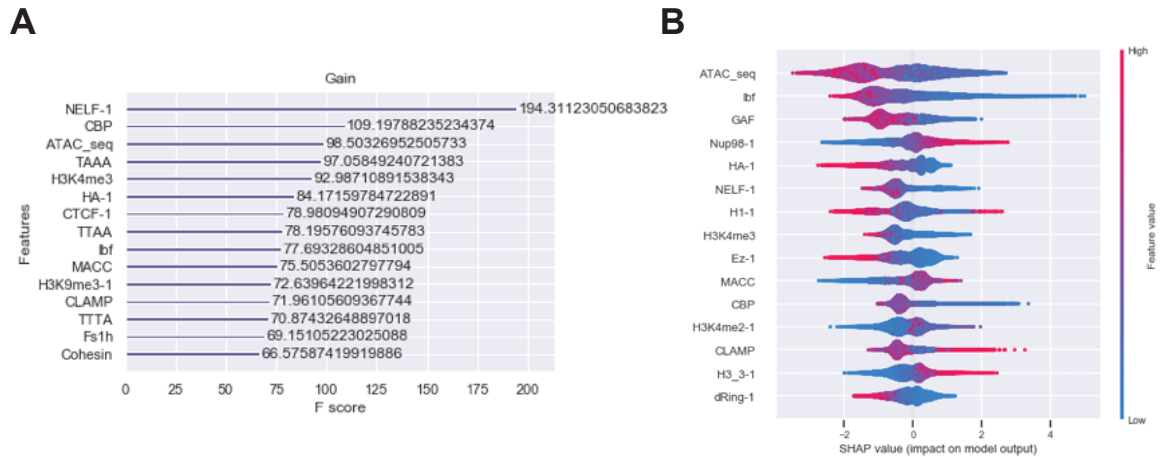


Figure 3.3. Feature importance rankings for the null MSLc model

(A) Top predictive feature rankings as measured by gain. **(B)** Top predictive feature rankings as measured by SHAP.

On the held-out test set, the null MSLc model achieved an overall accuracy of 97.2% with a precision of 0.885, recall of 0.709, and AUC of 0.981. Although this model now lacks the known key players in X-identification, it still does an excellent job of discriminating the X-chromosome from autosomes. Again, examining the top 15 gain contributors of the null MSLc model, the top feature hit for gain, NELF (Negative Elongation Factor) which contributed nearly 2-fold more predictive power than the second scoring feature (Figure 3.3A). Examining the top 15 contributors for SHAP value, I also again identified some known features that have established roles in X-identification (Figure 3.3B) (Brown and Bachtrog, 2014; Buenrostro et al., 2013; Deng et al., 2014; Disteche, 2012; Kaye et al., 2018; Mendjan and Akhtar, 2007; Soruco et al., 2013; Wells et al., 2012). The top SHAP

hit, ATAC-seq, is an indicator of chromatin accessibility (Buenrostro et al., 2013). A few of the other features identified, such as GAF, are associated with a class of proteins called as insulators, which are well known to play roles in chromatin organization (Bohla et al., 2014; Cuartero et al., 2014; Nègre et al., 2011; Philip et al., 2015; Wolle et al., 2015; Yang and Corces, 2012).

Another useful property of SHAP is that it allows for detailed interrogation of feature importance within singular sites or groups of interests. As previously mentioned, MSLc first localizes to CES and then spreads across the remainder of the X-chromosome. Therefore, I decided to further interrogate two classes of X-chromosome sites, CES and non-CES, along with autosomal sites, to more closely examine how a given feature impacts predictions at different site classes (Figure 3.4). Because there are just a few instances of CES in the test set, SHAP analysis was repeated using both the test and training datasets. A total of 57 unique CES remained in the final null MSLc model dataset. Therefore, a random sampling of 57 autosomal sites was performed in addition to 57 random X-chromosome sites in order to compare the distribution of average SHAP values across the classes defined above (Figure 3.4).

Overall, the top predictive features were preserved across site classes with a few potentially meaningful shifts in importance rankings (Figure 3.4). In particular, features such as ATAC-seq and MACC are indicators of chromatin accessibility

and the X-chromosome has been shown to exhibit higher levels of chromatin accessibility. Another feature, GAF, has been previously observed to be depleted at CES (Kaye et al., 2018). Furthermore, the known adapter protein for MSLC, CLAMP, plays a much larger role at CES, which are the initial MSLC recruitment sites (Figure 3.4). This finding confirms previously published results and serves as another positive indication that the feature importance rankings tools are providing meaningful results (Kaye et al., 2018; Kuzu et al., 2016; Soruco et al., 2013). Another feature identified, NELF, has never been directly associated with X-identification but exhibits a similar relationship as GAF in model analysis. Therefore, NELF-depletion on the X-chromosome, compared with autosomes, is a candidate for discriminating the X-chromosome. NELF has been observed to interact with GAF (Fay et al., 2011; Li et al., 2013) and has recently shown to also be associated with CLAMP using proteomic-based approaches (Urban et al., 2017b).

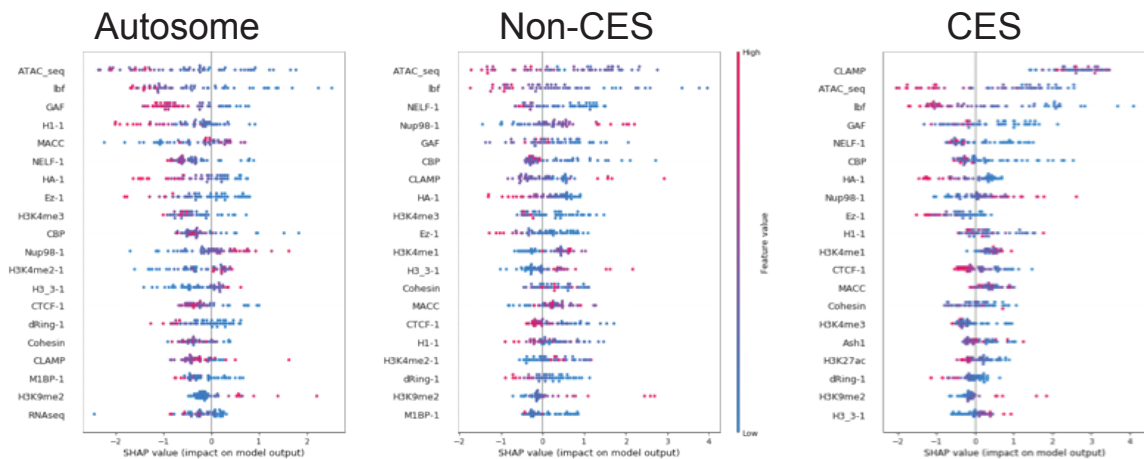


Figure 3.4. Distribution SHAP values across top features identified across various site classes.

Conclusions and Future Directions

The full model, which included MSLc components known to localize solely to the X-chromosome, served as an excellent proof-of-concept model. This model was able to identify 4 out of the 5 MSLc members remaining in the feature set as well as other features that have been previously linked to MSLc or X-identification. This full model also greatly helped in setting suitable evaluation criteria for the models that followed that did not contain MSLc members. The null MSLc model performed remarkably well, scoring slightly below that of the full model. The null MSLc model was readily able to identify features such as chromatin accessibility, GAF, CLAMP, NELF, etc. that have been previously linked to, or are, factors known to have some involvement in X-identification. However, it was not known to what extent these factors were predictive for distinguishing the X-chromosome from the autosomes.

I determined that enrichment of CLAMP and depletion of NELF and GAF are predictive of the X-chromosome compared with autosomes even without including MSLc components in the model. CLAMP is known to outcompete GAF at CES compared to other sites in the genome, consistent with depletion of GAF being predictive of X-chromosome identification (Kaye et al., 2018). Furthermore, NELF promotes pausing of RNA Polymerase II (Adelman and Lis, 2012; Gilchrist et al., 2008) and pausing is known to be decreased on the hyper-active X-chromosome compared with autosomes (Larschan et al., 2011). Therefore, a depletion of NELF on the X-chromosome is consistent with decreased pausing on the hyper-active

X-chromosome. Moreover, my model provides the first unbiased approach to identify and rank factors that distinguish the X-chromosome from the autosomes.

My two machine learning models, the full and null MSLc models, provide promising candidate features for further examination. It would be useful to perform automated feature selection to specifically determine a smaller range of NGS derived features. This could aide in creating a more focused list of possible targets for further bioinformatic interrogation. As a next step, it would be interesting to perform clustering on different classes of factors that the model predicts on the X-chromosome at CES and non-CES, as well as locations that resemble CES throughout the genome. The most promising set of novel features could then be more closely interrogated with more in depth bioinformatic analysis. The most promising candidate factors can then be experimentally validated *in vivo*. Overall, I have generated the first machine learning model to predict which factors distinguish the X-chromosome from the autosomes.

Supporting Information

Parameter	Inputs
learning_rate	[0.01, 0.05, 0.1, 0.3]
max_depth	[3, 5, 7, 10]
scale_pos_weight	[1, 5, 10]
min_child_weight	[1, 2, 5]
gamma	[0, 0.1, 0.2, 0.3, 0.5]
subsample	[0.5, 0.6, 0.7, 0.8, 0.9]
colsample_bytree	[0.5, 0.6, 0.7, 0.8, 0.9]
reg_alpha	[0.01, 0.1, 0, 1]
reg_lambda	[0.01, 0.1, 0, 1]

Table S3.1 Search space for randomized grid search

Parameters selected after 50 iterations of grid search are highlighted in red.

Supplemental Files

Available upon request

CHAPTER 4 : THE ZINC FINGER PROTEIN CLAMP PROMOTES LONG-RANGE CHROMATIN INTERACTIONS THAT MEDIATE DOSAGE COMPENSATION OF THE *DROSOPHILA* MALE X-CHROMOSOME

William Jordan, III¹ and Erica Larschan^{1,2}

This chapter has been submitted for publication and is reproduced here with minor formatting changes. I wrote the manuscript with editorial feedback from Dr. Leila Rieder and Dr. Erica Larschan.

¹ Department of Molecular Biology, Cell Biology, and Biochemistry, Brown University, Providence, Rhode Island 02912

² Corresponding author: erica_larschan@brown.edu

Abstract

Drosophila dosage compensation is an important model system for defining how active chromatin domains are formed. The male-specific lethal dosage compensation complex (MSLc) increases transcript levels of genes along the length of the single male X-chromosome to equalize with that on the two female X-chromosomes. The strongest binding sites for MSLc cluster together in three-dimensional space independently of MSLc because clustering occurs in both sexes. CLAMP, a non-sex-specific, ubiquitous, zinc finger protein, binds synergistically with MSLc to enrich the occupancy of both factors on the male X-chromosome. Here, we demonstrate that CLAMP promotes the observed clustering of MSLc binding sites. Genome-wide, CLAMP promotes interactions between active chromatin regions and represses interactions between inactive chromatin regions. Moreover, the X-enriched CLAMP protein promotes longer-range interactions on the active X-chromosome than autosomes. Overall, we define how long-range interactions, mediated by a locally-enriched ubiquitous transcription factor, generate a three-dimensional active chromatin domain.

Introduction

Three-dimensional chromatin domains are important for coordinating gene regulation. Recent work has provided new insight into how silent chromatin domains are formed, for example, through phase separation (Sexton and Cavalli, 2015; Strom et al., 2017), but less is understood regarding the formation of active chromatin domains. Dosage compensation in *Drosophila* provides one of the few model systems for studying the formation of a large active chromatin domain: approximately one thousand genes along the length of the single male X-chromosome are coordinately upregulated two-fold (Conrad et al., 2012; Hamada et al., 2005; Straub and Becker, 2007).

In heterogametic species, dosage compensation is essential to correct transcriptional imbalance of X-linked genes between the sexes (Jordan et al., 2019) and to correct for dosage imbalance between the single X-chromosome and paired autosomes. Diverse dosage compensation mechanisms have evolved across species, but an essential conserved step is distinguishing the X-chromosome from autosomes for specific regulation.

In *Drosophila*, the male-specific lethal complex (MSLc) forms specifically in males and is responsible for increasing transcript levels of X-linked genes along the length of the single male X-chromosome roughly two-fold, equalizing gene

expression with that of females (Conrad et al., 2012; Hamada et al., 2005; Straub and Becker, 2007). MSLc consists of five proteins MSL1, MSL2, MSL3, maleless (MLE) (Belote and Lucchesi, 1980; Kuroda et al., 1991; Lucchesi, 1996), and males absent on the first (MOF) (Hilfiker et al., 1997), and one of two functionally redundant long non-coding RNAs known as RNA on the X 1 and 2 (*roX1* and *roX2*) (Franke and Baker, 1999; Meller and Rattner, 2002). MSLc first targets X-linked genomic elements known as “high-affinity” or “chromatin entry” sites (CES), which also include the *roX* loci (Alekseyenko et al., 2008; Kelley et al., 1999; Meller and Rattner, 2002; Straub et al., 2008).

Within CES, MSLc is recruited to GA-rich 21-bp elements known as MSL recognition elements (MREs) (Alekseyenko et al., 2008). Accumulation of MRE sequences on the X-chromosome occurred by expansion of GA-rich sequences and transposon transposition (Ellison et al., 2019; Kuzu et al., 2016). However, MRE sequences are not X-chromosome specific and are only approximately two-fold enriched on the X-chromosome compared with autosomes (Alekseyenko et al., 2008), suggesting that they are not sufficient for X-chromosome targeting. Although the MSL2 component of MSLc has a low affinity for MREs, the complex requires synergy with an essential, non-sex-specific, zinc finger adapter protein known as chromatin-linked adapter for MSL proteins (CLAMP) in order to stabilize its binding to MREs (Albig et al., 2018; Larschan et al., 2012; Soruco et al., 2013).

Synergy between CLAMP and MSLc, which has been demonstrated both *in vivo* and *in vitro* (Albig et al., 2018; Soruco et al., 2013), enhances the occupancy of both factors on the male X-chromosome. Maternally deposited CLAMP is present on chromatin before MSL complex assembles at the maternal-zygotic transition (Franke et al., 1996; Lott et al., 2011; Rieder et al., 2017) and regulates chromatin accessibility of the X-chromosome and transcription of X-linked genes (Larschan et al., 2012; Urban et al., 2017a). Therefore, it is likely that CLAMP functions as an early transcription factor to enhance X-chromosome accessibility and promote MSLc targeting.

After initial targeting to CES by CLAMP, MSLc generates an active chromatin domain by localizing to the bodies of active genes and increasing their transcript levels by modulating transcription elongation (Alekseyenko et al., 2006, 2008; Kelley et al., 1999; Larschan et al., 2011; Urban et al., 2017a). MSLc is hypothesized to take advantage of pre-existing three-dimensional chromatin organization to target the X-chromosome (Ramírez et al., 2015; Schauer et al., 2017). Chromosome conformation capture techniques have demonstrated that CES cluster three-dimensionally in both males and females and form long-range interactions with other X-linked active chromatin regions within the nucleus independently of MSLc (Ramírez et al., 2015; Schauer et al., 2017). However, the mechanism by which CES cluster remained unknown.

We hypothesized that CLAMP promotes clustering of CES based on the following lines of evidence: 1) In contrast to MSLc, CLAMP is required to increase the accessibility of the entire male X-chromosome (Urban et al., 2017a); 2) CLAMP is part of an insulator complex (Kaye et al., 2017) and functions as an insulator protein in several enhancer-blocking assays that are dependent on the Su(Hw) insulator complex (Bag et al., 2019). Insulator proteins mediate chromatin interactions across the genome to regulate specialized chromatin domains throughout development (Darrow et al., 2016; Kieffer-Kwon et al., 2017; Muller et al., 1999; Noordermeer et al., 2011; Vian et al., 2018).

We used genome-wide chromosome conformation capture (Hi-C) analysis complemented by circular chromosome conformation capture with high-throughput sequencing (4C-seq) to test the hypothesis that CLAMP regulates clustering of CES and three-dimensional organization of the X-chromosome. We discover that CLAMP promotes long-range interactions on the male X-chromosome more strongly than on autosomes. Furthermore, we demonstrate that CLAMP primarily represses interactions at silent chromatin regions and promotes long-range interactions within active chromatin regions, including CES. We also show that the enrichment of several insulator proteins is increased at loci where CLAMP promotes or represses genomic interactions. Overall, we show that the X-enriched CLAMP protein regulates long-range three-dimensional interactions between CES to target MSLc to the male X-chromosome. Synergy between CLAMP and MSLc

(Albig et al., 2018; Soruco et al., 2013) increases the occupancy of both factors to generate a three-dimensional active chromatin domain that specifically upregulates transcription of approximately one thousand X-linked genes in males.

Results

CLAMP regulates long-range three-dimension interactions on the X-chromosome more strongly than on autosomes

In order to understand how CLAMP regulates the three-dimensional organization of the genome, we performed *in situ* chromosome conformation capture with high-throughput sequencing (*in situ* Hi-C) (Rao et al., 2014) using HindIII, a 6-bp cutter restriction enzyme, in *Drosophila* male Schneider's line 2 (S2) (Schneider, 1972) cultured cells after validated RNAi depletion of either *gfp* (control) or *clamp* (Hamada et al., 2005; Larschan et al., 2012; Soruco et al., 2013). We performed two biological replicates for each experimental condition. We confirmed depletion of CLAMP protein after *clamp* RNAi by Western blot (Fig. S4.1A) and used GenomeDISCO (Ursu et al., 2018) to determine replicate concordance. Reproducibility between replicates was high at multiple resolutions (Figure S4.1B).

We generated Hi-C interaction maps combining replicates for each condition (Fig. 4.1A; File S4.1). To compare the *clamp* RNAi and control *gfp* RNAi contact maps, we also generated a differential interaction map (Fig. 4.1A). In the differential map for the two conditions (*clamp/gfp* RNAi), we observed an increase in interaction frequency directly along the diagonal (i.e. shorter-range interactions). Many of the

more pronounced off-diagonal differences are in regions proximal to centromeres, which will be discussed later. In general, we observed decreased interaction frequency moving away from the diagonal (i.e. longer-range interactions) on all chromosomes, with the exception of chromosome 4, when CLAMP was depleted. To quantify these perceived differences between our *gfp* and *clamp* RNAi interaction matrices, we calculated the $\log_2$ ratio of distal to local interactions (DLR) (Heinz et al., 2018) (Fig. 4.1B). We defined local interactions as those that span less than 250 kb and distal as interactions that span more than 250 kb. After depleting CLAMP, we observed a widespread increase in the DLR of X-linked genomic interactions across the entire chromosome, signifying that there are more local than distal interactions when CLAMP is depleted. In contrast, the DLR was more varied on autosomes except for the heterochromatic fourth chromosome, suggesting that the X-chromosome specifically becomes more three-dimensionally compact upon depletion of CLAMP. Moreover, this X-specific change in three-dimensional interactions (Fig. 4.1B) is consistent with previous linear analysis demonstrating that chromatin accessibility of the male X-chromosome but not autosomes or the female X-chromosome decreases upon RNAi depletion of CLAMP (Urban et al., 2017a). Therefore, our Hi-C data support a model in which CLAMP decompacts the X-chromosome more than autosomes.

Previous Hi-C studies in multiple cell lines and embryos found that the strongest MSLc binding sites (CES) interact with each other and other genomic regions

along the X-chromosome more frequently than expected by chance (Ramírez et al., 2015; Schauer et al., 2017). To confirm these observations within our own data sets, we investigated whether CES interact frequently with other genomic locations in our Hi-C data. We used Fit-Hi-C (Ay et al., 2014) to determine high-confidence intrachromosomal contacts (see methods) from our control *gfp* RNAi Hi-C maps (Fig. S4.1C; File S4.3). Consistent with previous reports (Ramírez et al., 2015; Schauer et al., 2017), approximately 60% of the high-confidence X-chromosome interactions identified involve CES and 16% of such interactions link regions containing CES at 20 kb resolution (Fig. S4.1D). Therefore, our data are consistent with prior reports of CES frequently interacting with each other and other regions of the genome.

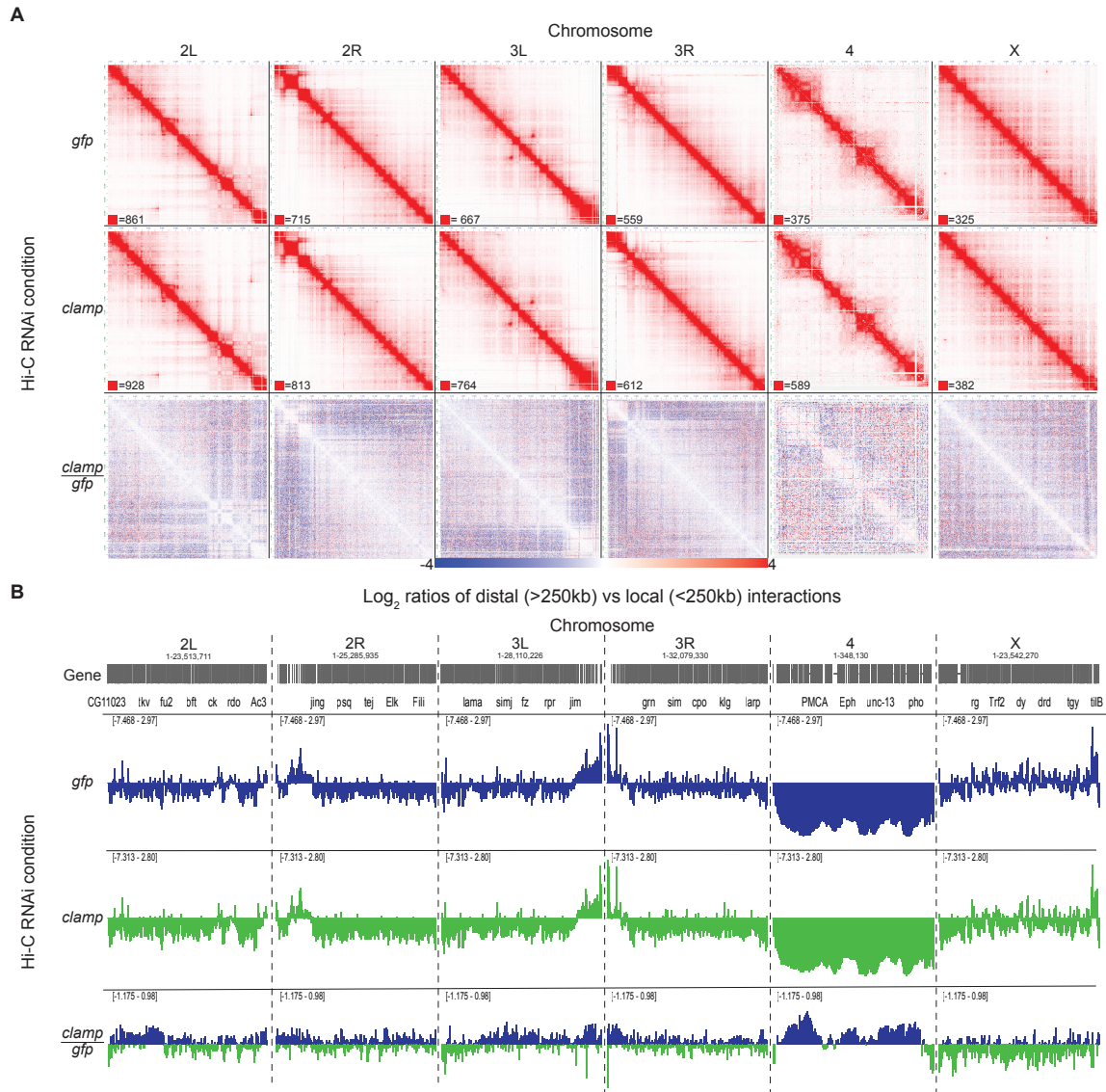


Figure 4.1. CLAMP regulates the span of genomic interactions on the male X-chromosome.

A. Per chromosome KR balanced matrices for *gfp* RNAi (top), *clamp* RNAi (middle), and *clamp* vs *gfp* RNAi (bottom). Each chromosome is shown at 50 kb resolution, with the exception of chromosome 4 which is shown at 5 kb resolution due to its small size.

B. Per chromosome distal vs local ratio (DLR) for *gfp* RNAi (blue), *clamp* RNAi (green), and *clamp* vs *gfp* RNAi (bottom). For the *clamp* vs *gfp* RNAi comparison, a positive number (blue) indicates the ratio of distal vs. local interactions becomes higher following *clamp* RNAi. A negative number (green) indicates the ratio of distal vs. local interactions becomes lower following *clamp* RNAi.

To more quantitatively define specific regions throughout the genome where

CLAMP regulates three-dimensional interactions, we compared intra-

chromosomal interaction frequencies after *clamp* RNAi with those after *gfp* RNAi using diffHic (Lun and Smyth, 2015). We found 2,552 significantly (FDR < 0.05) differential interactions (DIs) at 30 kb resolution (Fig. S4.2A; File S4.4). We classified these DIs by the directionality of their $\log_2$ fold change. Notches on all box plots represent 95% confidence intervals. Interactions that decrease in contact probability after *clamp* RNAi are defined as CLAMP-promoted (55%), whereas interactions that increase in contact probability after *clamp* RNAi are defined as CLAMP-repressed (45%). The $\log_2$ fold change was significantly larger for CLAMP-promoted X-linked interactions (Fig. 4.2A; Fig. S4.2B) than other remaining interaction types throughout the genome. Next, we divided the DIs into quantiles based on absolute $\log_2$ fold change and normalized for chromosome size and ploidy. We found that as absolute $\log_2$ fold change increases, the frequency of both CLAMP-promoted and CLAMP-repressed interactions on the X-chromosome increases, while those on autosomes decrease (Fig. 4.2B). Therefore, CLAMP functions more strongly to promote three-dimensional interactions on the X-chromosome compared to on autosomes.

To define the properties of DIs mediated by CLAMP, we measured the genomic distance between DI anchors. The linear distance spanned by CLAMP-promoted X-linked interactions is significantly longer than those that CLAMP promotes on

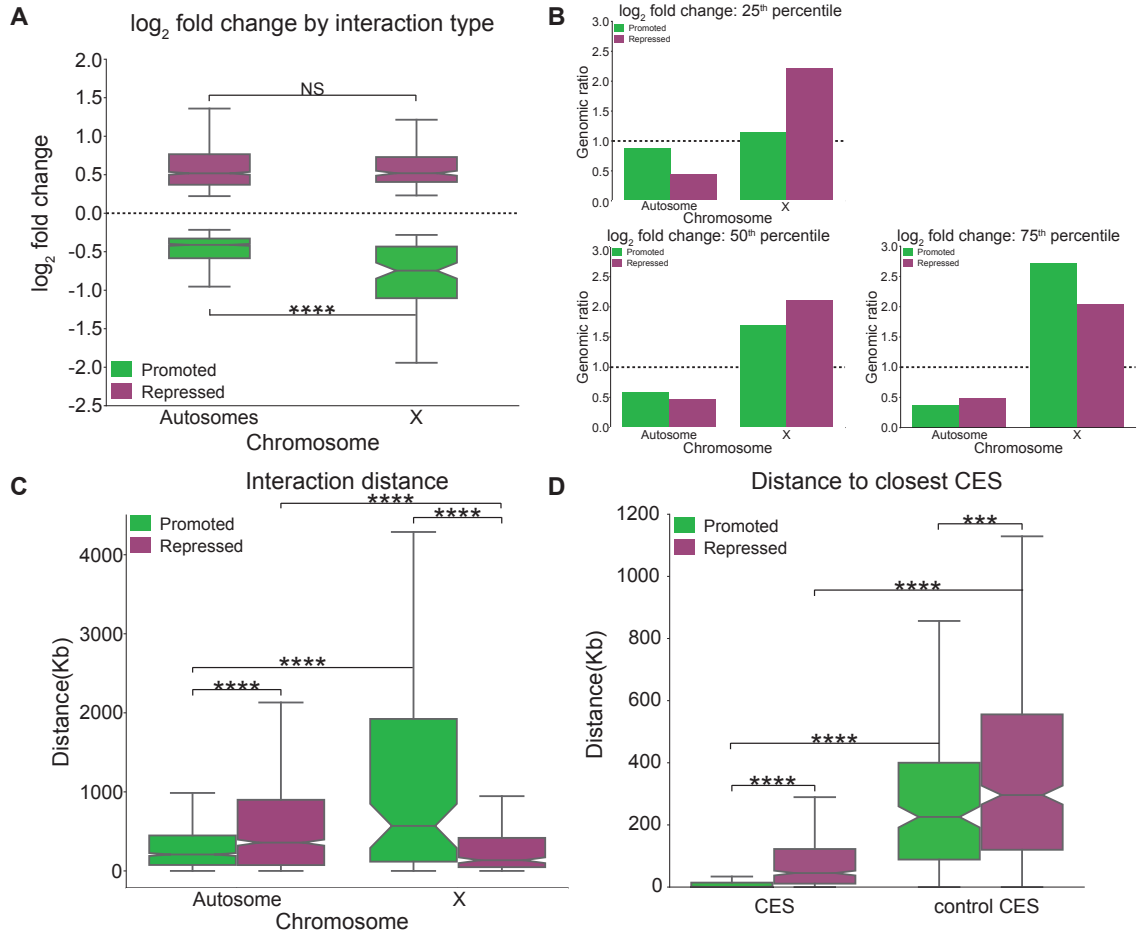


Figure 4.2. CLAMP promotes the formation of longer-range contacts more strongly on the X-chromosome than autosomes.

A. $\log_2$ fold changes per chromosome of interactions that are weakened after *clamp* RNAi (CLAMP-promoted) or strengthen (CLAMP-repressed).

B. 25th, 50th, and 75th percentiles of absolute $\log_2$ fold change. Genomic ratio is normalized to account for chromosome length and ploidy. The dotted line at 1 represents the expected ratio if there is no bias towards a particular chromosome. As absolute fold change increases both CLAMP-promoted and CLAMP-repressed interactions show a bias toward the X-chromosome.

C. Distribution of distances between DI anchors. On the autosomes, CLAMP-promoted interactions are shorter-range compared to autosomal CLAMP-repressed interactions. On the X-chromosome however CLAMP-promoted interactions are much longer-range than CLAMP-repressed.

D. Distribution of distances to nearest CES or control CES for CLAMP-promoted and CLAMP-repressed interactions.

For all box and whisker plots, the 95% confidence interval is shown with a notch around the median line. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$, two-sided Mann-Whitney test.

autosomes (Fig. 4.2C; Fig. S4.2C). In contrast, the length span of interactions repressed by CLAMP is significantly shorter on the X-chromosome than on autosomes (Fig. 4.2C; Fig. S4.2C), consistent with the X-biased decrease in DLR after depleting CLAMP (Fig. 4.1B). Therefore, CLAMP promotes long-range interactions and prevents the formation of shorter-range interactions on the X-chromosome. We also measured the distance between X-linked DI anchors and CES. We determined that CLAMP regulates three-dimensional interactions at locations that are in closer to proximity to CES, compared with previously defined control CES-like regions that do not recruit MSLc (Soruco et al., 2013) (Fig. 4.2D). Overall, CLAMP promotes long-range contacts on the X-chromosome that enhance three-dimensional interactions involving CES.

CLAMP promotes three-dimensional interactions at active chromatin regions including CES and represses three-dimensional interactions at inactive chromatin regions

Next, we defined the relationship between DIs that are regulated by CLAMP and the enrichment of the CLAMP protein. We used available CLAMP ChIP-seq data (Kaye et al., 2018) from S2 cells to generate a high-confidence list of CLAMP peaks and average profiles of CLAMP peak enrichment DI anchors (see methods). We found that CLAMP-promoted DI anchors more frequently contain CLAMP peaks than CLAMP-repressed DI anchors on both the X-chromosome and autosomes (Fig. 4.4A). Therefore, CLAMP-promoted contacts are more likely to be directly linked to CLAMP function. Furthermore, we found that 60% of CLAMP-

promoted DI anchors on the X-chromosome are occupied by CLAMP, which is greater than other classes of DIs. While CLAMP frequently binds to DI anchors, the ChIP-seq enrichment of CLAMP is not enhanced at genomic locations where CLAMP regulates three-dimensional genomic interactions. Therefore, we hypothesize that the presence of additional cofactors within active or inactive chromatin modulates the ability of CLAMP to regulate three-dimensional interactions.

To test this hypothesis, we first measured the chromatin states, as defined by the *Drosophila* 9-state chromatin model (Kharchenko et al., 2011), that are present at CLAMP-promoted and CLAMP-repressed DI anchors. We found that across all chromosomes, 60.5% of the chromatin states present within CLAMP-promoted DI anchors represent active chromatin and the remaining 39.5% of chromatin states represent inactive chromatin (Fig S4.2E). In contrast, 70% of the chromatin states present within CLAMP-repressed DI anchors represent inactive states, while the remaining 30% of chromatin states represent active chromatin across all chromosomes (Fig S4.2F). Additionally, at CLAMP-repressed DI anchors, there is an enrichment of the chromatin state corresponding to pericentromeric heterochromatin (Fig. S4.2F), which correlates with our visual observation that many pronounced off-diagonal changes in our differential Hi-C maps occur in regions proximal to centromeres (Fig. 4.1A).

To further quantify the relationship between DI anchors and chromatin states, we computed Jaccard similarity coefficients as ratios ranging from 0 to 1 (Quinlan, 2014): the larger the Jaccard coefficient (i.e. closer to 1), the more similar two sets of genomic regions are to each other. CLAMP-promoted DI anchors are more correlated with active chromatin states (Jaccard coefficient: 0.328) than inactive chromatin states (Jaccard coefficient: 0.220). In contrast, CLAMP-repressed DI anchors are more correlated with inactive chromatin states (Jaccard coefficient: 0.297) than active chromatin states (Jaccard coefficient: 0.090). Therefore, CLAMP primarily promotes interactions within active regions and represses interactions within inactive regions of the genome. We also measured the proximity of CLAMP-promoted and CLAMP-repressed DI anchors to CES and two randomized classes of sites within regions of either active or inactive chromatin based on the 9-state chromatin state model for S2 cells (Kharchenko et al., 2011) (Fig. S4.2D). We found that CLAMP-promoted regions are closer to active chromatin regions than CLAMP-repressed regions. In contrast, CLAMP-repressed regions are closer to inactive chromatin than active chromatin regions. Also, CES are in closer proximity to CLAMP-promoted DI anchors and more distal from CLAMP-repressed DI anchors than other active chromatin regions. Moreover, the ability of CLAMP to promote genomic interactions within active chromatin regions (Fig. S4.2D; Fig. S4.2E) is consistent with its ability to activate gene expression on the active male X-chromosome more frequently than on autosomes (Urban et al., 2017a).

CLAMP and MSLc both mediate three-dimensional interactions at CES

We and others previously reported that CLAMP and MSLc function synergistically to target each other to the X-chromosome and increase occupancy of both factors (Albig et al., 2018; Soruco et al., 2013). MSLc modulates chromatin accessibility specifically within CES, in contrast to CLAMP, which not only functions at CES but also enhances chromatin accessibility of the entire male X-chromosome (Ramírez et al., 2015; Urban et al., 2017a). Furthermore, long-range interactions at several CES occur in an MSL2-dependent manner (Grimaud and Becker, 2009; Schauer et al., 2017). Therefore, we hypothesized that both CLAMP and MSL complex function to regulate three-dimensional interactions at CES. We performed differential interaction analysis with available Hi-C data (GSE58821) from replicated dilution Hi-C experiments that compare RNAi depletion of two MSLc components (MSL2 and MSL3) with a matched control (*gfp* RNAi) experiment (Ramírez et al., 2015). In contrast to our *clamp* RNAi experiments, we did not identify significant DIs following *msl2* RNAi, and *msl3* RNAi resulted in only 1 significant DI when compared to matched *gfp* controls (File S4.4). Therefore, MSLc does not modulate three-dimensional interactions detectable by dilution Hi-C.

To validate our Hi-C findings that CLAMP regulates three-dimensional organization and generate a higher resolution subset of DIs, we performed circularized chromosome conformation capture with high-throughput sequencing (4C-seq) at four CES in close proximity to regions containing many differential

interactions identified by Hi-C. In addition, we assessed the function of MSL2 and the GAF protein (encoded by the *trl* gene), which is a GA-binding zinc-finger protein similar to CLAMP. CLAMP and GAF are present in the same insulator complex (Farkas et al., 1994; Kasinathan et al., 2014; Kaye et al., 2017, 2018). However, in contrast to CLAMP, GAF has a modest function in MSLc recruitment and is depleted within CES because CLAMP outcompetes GAF for binding to long GA-rich sequences (Greenberg et al., 2004; Kaye et al., 2018). Therefore, we hypothesized that CLAMP but not GAF regulates three-dimensional interactions at CES.

To test this hypothesis, we performed 4C-seq experiments in biological duplicate in *Drosophila* S2 cells following previously validated RNAi depletion of *gfp* (control), *msl2*, *clamp* and *trl* (Kaye et al., 2018; Soruco et al., 2013) and confirmed depletions by qRT-PCR (Fig. S4.3A; File S4.1; File S4.2). We identified high frequency *cis*-interacting regions and performed differential interaction analysis for each viewpoint using the 4C-ker pipeline (Raviram et al., 2016) (Fig. S4.4). Consistent with our Hi-C analysis, we found that 59% of high-confidence *cis*-interactions link our four CES viewpoints to regions that also contain a CES (Fig S4.3B; File S4.3). After visualizing differential interactions for each viewpoint (Fig. S4.4; File S4.4), we pooled all identified DI anchors from each 4C-seq viewpoint to increase the number of DIs for further analysis.

Next, we compared the number of DIs obtained after *clamp* and *trl* RNAi. We identified 199 total DIs after *clamp* RNAi and only 17 DIs after *trl* RNAi (Fig S4.3C). Therefore, even though both CLAMP and GAF are part of the same insulator complex (Kaye et al., 2017), CLAMP has a stronger role than GAF in modulating three-dimensional interactions involving CES, consistent with an enrichment of CLAMP versus GAF at CES (Kaye et al., 2018). Next, we measured chromatin state occurrence within CLAMP-promoted and CLAMP-repressed 4C-seq DI anchors and found similar chromatin states to those identified in our differential Hi-C analysis (Fig. S4.3F and S4.2E-F). Therefore, our 4C-seq data demonstrate that CLAMP has a stronger role than GAF in regulating three-dimensional interactions at CES and validate a role for CLAMP in promoting contacts between active chromatin regions and repressing contacts with inactive chromatin regions.

In addition, we measured the role of CLAMP and MSL2 in regulating three-dimensional interactions at CES by comparing our 4C-seq data after *clamp* and *msl2* RNAi treatments. In contrast to our differential Hi-C interaction analysis which did not identify any DIs after *msl2* RNAi, we observed 285 DIs after *msl2* RNAi from our CES viewpoints, consistent with prior findings (Grimaud and Becker, 2009; Schauer et al., 2017) (Fig S4.3C). The discrepancy between the Hi-C and 4C-seq analyses for MSL2 may be due to the dilution Hi-C technique performed by Ramirez et al., which contains more technical noise because there is higher potential for spurious contacts during in-solution ligation compared with *in situ*

ligation (Gavrilov et al., 2013; Rao et al., 2014; Van De Werken et al., 2012). We also measured X-specific chromatin state occurrence for 4C-seq DI anchors identified after *msl2* RNAi and we found similar chromatin states to those identified at CLAMP-dependent 4C-seq DI anchors (Fig. S4.3 E-F). Overall, both CLAMP and MSL2 regulate three-dimensional interactions at CES, consistent with prior observations that CLAMP and MSLc modulate chromatin accessibility (Ramírez et al., 2015; Urban et al., 2017a) and promote each other's occupancy (Albig et al., 2018; Soruco et al., 2013) locally at CES.

To further define the relationship between CLAMP and MSLc in mediating three-dimensional interactions, we integrated our Hi-C and 4C data with previously-generated chromatin accessibility data from a Micrococcal Nuclease (MNase)-seq experiment performed in S2 cells under the same RNAi conditions (GSE99894) (Urban et al., 2017a). In MNase-seq experiments, a MNase accessibility score (MACC) greater than zero indicates that a region of chromatin is relatively accessible while a MACC score less than zero indicates that a region is relatively inaccessible compared to the average genomic accessibility. We found that CLAMP-promoted and CLAMP-repressed DIs from both Hi-C and 4C occur within regions where the chromatin is relatively accessible (MACC >0) under control *gfp* and *msl2* RNAi conditions (Fig. 4.3A, B). However, significant decreases in chromatin accessibility after *clamp* RNAi, are observed at CLAMP-promoted and

CLAMP-repressed DI anchors on the X-chromosome and CLAMP-repressed DIs on autosomes (Fig. 4.3A, B).

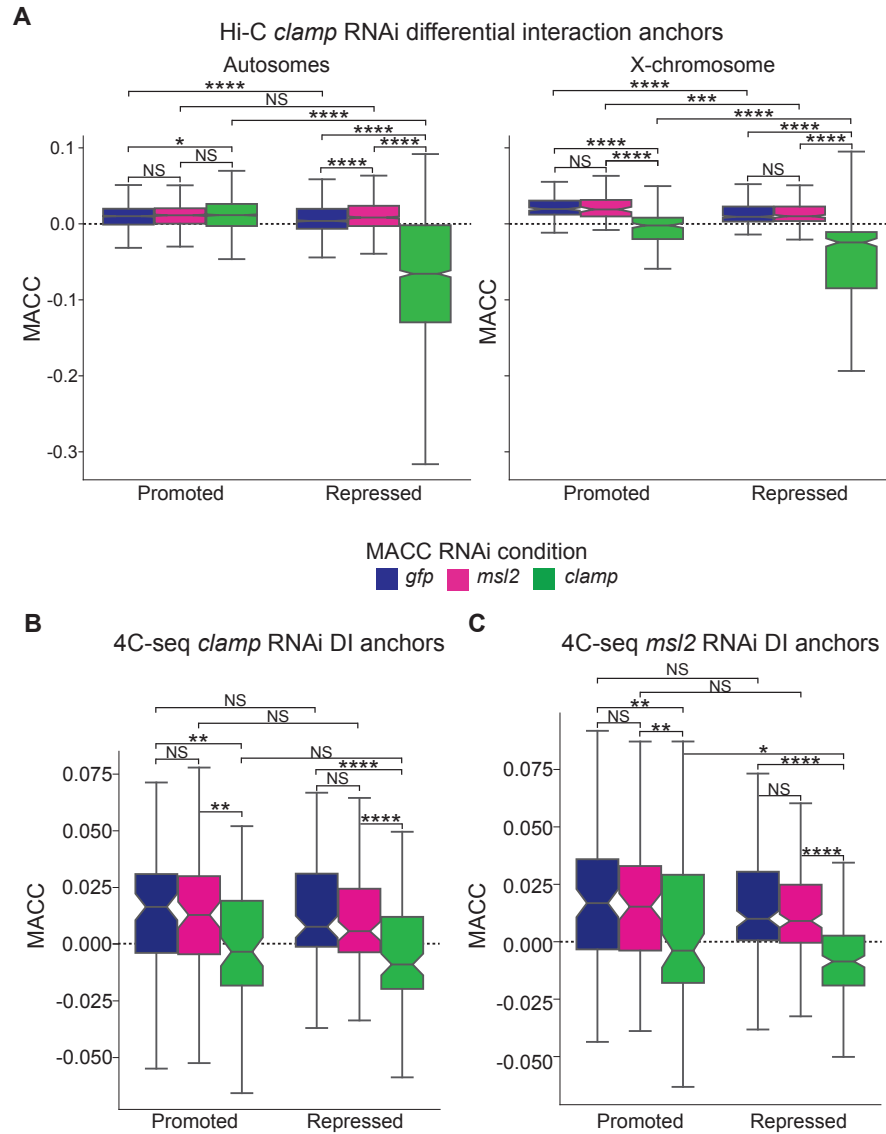


Figure 4.3. CLAMP regulation of the three-dimensional organization is linked to its role in altering chromatin accessibility.

A. Distribution of chromatin accessibility MACC values within CLAMP Hi-C DI anchors as measured when MACC was previously calculated under RNAi conditions of *gfp* (blue), *msl2* (pink) and *clamp* (green). A positive MACC value indicates accessible chromatin whereas a negative MACC value indicates inaccessible chromatin (Urban et al., 2017a). On the autosomes, CLAMP promoted interactions occur in regions that are accessible under all three RNAi conditions. In contrast, autosomal CLAMP-repressed interactions occur in regions that have lowered chromatin accessibility after *clamp* RNAi. On the X-chromosome both CLAMP-promoted and repressed interactions occur in regions which are accessible after *gfp* and *msl2* RNAi but become inaccessible after *clamp* RNAi.

B. Distribution of MACC values within CLAMP 4C-seq DI anchors. CLAMP 4C-seq identified DIs occur in regions which have significantly lowered chromatin accessibility after *clamp* RNAi (green) compared to control *gfp* RNAi (blue). No significant changes are observed for these CLAMP 4C-seq identified DI regions are observed following *msl2* RNAi (pink).

C. Distribution of MACC values within MSL2 4C-seq DI anchors. MSL2 4C-seq identified DIs occur in regions that do not have significant changes in chromatin accessibility after *msl2* RNAi (pink) compared to control *gfp* RNAi (blue). However, chromatin accessibility is significantly decreased following *clamp* RNAi (green).

For all box and whisker plots, the 95% confidence interval is shown with a notch around the median line.

Next, we asked whether CLAMP or MSL2 regulates chromatin accessibility at sites where MSL2 regulates three-dimensional interactions. We determined that MSL2-dependent 4C-seq DIs show significant decreases in chromatin accessibility after *clamp* RNAi and modest but not statistically significant changes after *msl2* RNAi (Fig. 4.3C). Therefore, MSL2-dependent 4C-seq DIs require CLAMP to promote chromatin accessibility. Overall, the function of CLAMP in modulating three-dimensional interactions is linked to its role in altering chromatin accessibility.

Insulator proteins have differential occupancy at genomic locations at which CLAMP regulates three-dimensional interactions

CLAMP has been physically and functionally linked with two different insulator complexes containing either the insulator proteins GAF (Kaye et al., 2017) or Su(Hw) (Bag et al., 2019). Therefore, we hypothesized that the function of CLAMP in promoting or repressing three-dimensional interactions is mediated by differential occupancy of insulator proteins at DI anchors. To test this hypothesis, we generated high confidence peaks and average profiles for insulator proteins from the following publicly available ChIP-seq data generated in S2 cells for the

insulator proteins GAF (GSE107059), CP190, Su(Hw), Mod(mdg4) and dCTCF (GSE41354). We restricted our analysis to CLAMP-dependent DIs identified by Hi-C in order to make genome-wide comparisons and compared ChIP-seq enrichment of each factor at peaks that occur within DI anchors to enrichment at the remaining set of peaks that occur outside of DI anchors (non-DIs).

Overall, we observed enhanced differential occupancy of all insulator proteins at genomic anchor points where CLAMP regulates three dimensional interactions (DIs) compared with regions where it does not regulate three-dimensional interactions (non-DIs). Furthermore, we discovered the following differences in factor enrichment: 1) Overall the enrichment patterns of GAF are similar to that of CLAMP, which correlates with their presence in the same insulator complex (Kaye et al., 2018). However, GAF enrichment is enhanced at CLAMP-promoted DI anchors and depleted at CLAMP-repressed DIs. Also, GAF is depleted at CLAMP-promoted autosomal DI anchors (Fig. 4.4B). 2) Su(Hw) is bound to fewer CLAMP-repressed DI anchors on autosomes compared with those on the X-chromosome, although its average enrichment at bound sites is similar. Su(Hw) is also modestly enriched at peaks occurring within CLAMP-promoted and -repressed sites, compared to peaks that occur at non-DIs (Fig. 4.4C). 3) Mod(mdg4), dCTCF and CP190 are all more frequently bound at CLAMP-promoted DI anchors. However, their enrichment is higher at CLAMP-repressed DI anchor sites versus CLAMP-promoted anchors. Furthermore, CP190 binds to a larger majority of CLAMP-

promoted DIs than any of the other factors tested (Fig. 4.4D, E, F). Overall, we observed different patterns of insulator protein occupancy within CLAMP-promoted and CLAMP-repressed DI anchors, compared to peak locations where we did not detect CLAMP-dependent three-dimensional interactions (non-DIs). Therefore, CLAMP is likely to modulate recruitment or function of insulator protein complexes to regulate three-dimensional interactions at DI anchors. In fact, very recent work demonstrated that CLAMP regulates the occupancy of the CP190 protein (Bag et al., 2019).

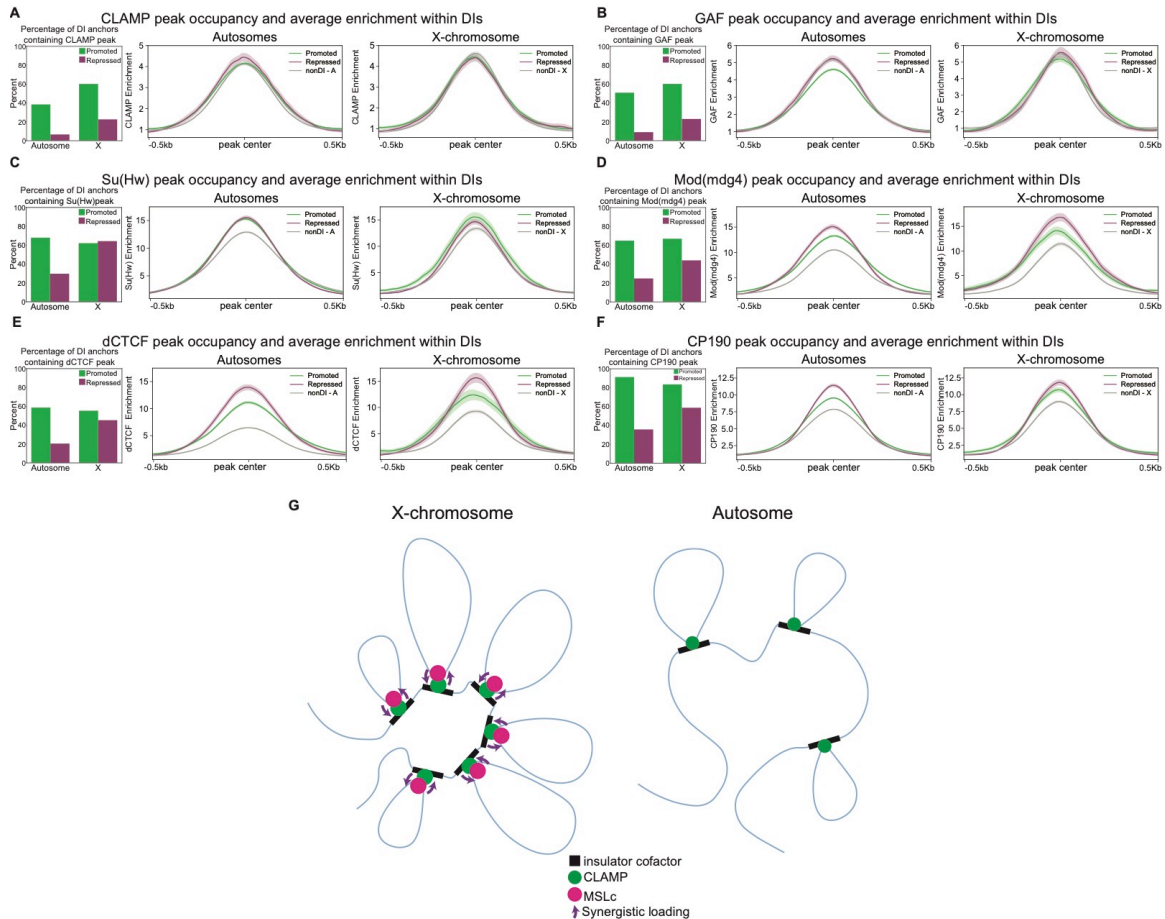


Figure 4.4. Differential occupancy of insulator proteins occurs at loci where CLAMP promotes or represses the formation of three-dimensional interactions on the X-chromosome and autosomes.

A. Left: Percentage of DI anchors containing a high confidence CLAMP ChIP-seq peak. Right: average fold enrichment over input for CLAMP at peaks within CLAMP-promoted DI anchors (green) and CLAMP-repressed DI anchors (purple) versus the remaining set of peaks falling outside of DIs (gray). Light shading on average profiles represents standard error.

The same analysis was performed for GAF (**B**), Su(Hw) (**C**), Mod(mdg4) (**D**), dCTCF (**E**), and CP190 (**F**).

G. CLAMP promotes long-range interactions on the X-chromosome that are responsible for the clustering of CES. Synergy between CLAMP and MSL complex increases the local concentration of both factors on the male X-chromosome. On autosomes, the local concentration of CLAMP is not as high due to the lack of synergy with MSL complex (Sorucu et al., 2013) which reduces the number of long-range interactions that are mediated by CLAMP.

Discussion

Overall, our data provide key insight into how an active chromatin domain is formed in three-dimensions. CLAMP de-compacts the genomic architecture of the X-chromosome by promoting long-range interactions and preventing the formation of shorter-range interactions (Fig. 4.4G). CLAMP not only tethers MSLc to the X-chromosome (Sorucu et al., 2013) but also promotes long-range three-dimensional interactions involving MSLc binding sites. (Fig. 4.4G). Previously, we demonstrated that CLAMP and MSLc function synergistically at CES to increase transcript levels of X-linked genes (Urban et al., 2017a). Here, we show that both CLAMP and MSLc regulate the three-dimensional organization of the X-chromosome: CLAMP regulates three-dimensional interactions along the length of the entire X-chromosome while MSLc acts locally at several CES.

Throughout the genome, CLAMP promotes long-range interactions between active chromatin regions, including CES, and represses three-dimensional interactions

at inactive chromatin regions. However, we observe that CLAMP has a stronger function in mediating three-dimensional interactions on the X-chromosome than on autosomes. CLAMP is an ancient zinc-finger protein that is highly conserved across *Diptera* and is present on chromatin in the earliest stages of development before the more recently evolved MSLc (Franke et al., 1996; Kuzu et al., 2016; Rieder et al., 2017; Soruco et al., 2013). Transposons containing CLAMP recognition sequences transposed onto the ancient X-chromosome and GA-repeats expanded, which together increased the density of CLAMP occupancy at CES (Ellison et al., 2019; Kuzu et al., 2016). The enhanced role of CLAMP on the X-chromosome compared with autosomes is likely due to the increased density of CLAMP binding sites within CES on the X-chromosome (Kuzu et al., 2016).

Furthermore, CLAMP interacts with several insulator protein complexes (Bag et al., 2019; Kaye et al., 2017), which are known to modulate three-dimensional genomic interactions. We observe that the enrichment of insulator proteins is increased at sites where CLAMP functions to regulate three-dimensional interactions, compared to non-interacting control regions. Therefore, it is likely that CLAMP regulates three-dimensional organization by modulating the ability of insulator proteins to drive the formation of three-dimensional interactions. For example, CLAMP is known to alter the occupancy of CP190, a component of several different insulator complexes (Bag et al., 2019). It is also possible that insulator proteins alter the ability of CLAMP to regulate three-dimensional

interactions. Overall, we demonstrate that CLAMP, a transcription factor present at high density on the male X-chromosome due to synergy with MSLc (Albig et al., 2018; Soruco et al., 2013), regulates long-range genomic interactions on the male X-chromosome including those involving CES. Therefore, CLAMP-mediated three-dimensional interactions create the three-dimensional active chromatin domain critical for dosage compensation. Further work will be required to define the functional relationships between CLAMP and all of the known insulator proteins.

Acknowledgements

We thank Dr. Leila Rieder for critical reading of the manuscript and Dr. John Urban and Dr. Nicola Neretti for advice on computational analysis. This research was supported by NIH grant: 5R35GM126994. WJ was funded by a NSF GRFP Fellowship and HHMI Gilliam Fellowship.

Author contributions

WJ performed all experiments, analysis and wrote the manuscript. EL advised on experiments and writing of the manuscript.

Declaration of interests

The authors declare no competing interests.

Methods

Cell culture conditions

Drosophila S2 cells were maintained at 25C in Gibco Schneider's *Drosophila* media (ThermoFisher Scientific) supplemented with 10% heat-inactivated fetal bovine serum (FBS) and 1.4X antibiotic-antimycotic (ThermoFisher Scientific). Cells were passaged every 2-3 days to maintain appropriate density.

RNAi treatment

Generation of dsRNA for RNAi treatment

Generation of dsRNA targeting *gfp* (control), *clamp*, *msl2* and *trl* for RNAi have been previously validated and described (Fuda et al., 2015; Hamada et al., 2005; Larschan et al., 2012; Soruco et al., 2013). PCR product was used as template to generate dsRNA with an ambion T7 MEGAscript kit (ThermoFisher Scientific). dsRNAs were purified following DNase treatment with a Qiagen RNeasy kit (Qiagen).

RNAi treatment

RNAi was performed in T75 tissue culture flasks. A total of 7×10^6 S2 cells were suspended in 6mL of Schneider's *Drosophila* media without FBS were added to a T75 culture flask containing 67.5ug of *gfp*, *clamp*, *msl2*, or *trl* dsRNA suspended in 3mL of Invitrogen UltraPure water (ThermoFisher Scientific). The cells were serum starved for 45 minutes at room temperature (RT), then 10.5mL of

Schneider's *Drosophila* media supplemented with 10% FBS was added to the flask. Cells were incubated for a total of 6 days as described previously (Kaye et al., 2018; Soruco et al., 2013). After 6 days, samples were collected and RNAi treatment was validated via western blot or qRT-PCR. The remaining cells were collected by centrifugation and resuspended to 5×10^6 cells/ml in fresh non-FBS media. Fresh formaldehyde solution (36.5-38% in H₂O; Sigma Aldrich) was then added to obtain a final concentration of approximately 1% formaldehyde. Cell suspensions were then incubated at RT on a rocking platform for 10 minutes. 2.5M glycine, to reach a final concentration of 0.125M, was then added to quench the formaldehyde; the cells were incubated as before for an additional 5 minutes. Afterwards, the cell suspension was immediately placed on ice for 15 minutes. The cell suspension was then aliquoted into individual tubes (5 million cells per tube). The cell suspensions were centrifuged at 4C for 5 minutes at 3000 RPM, supernatant was removed, and cell suspensions were flash frozen in liquid nitrogen and stored at -80C for future processing.

Knockdown validation

For Hi-C experiments knockdown of CLAMP was validated using the Western Breeze kit (Invitrogen). Antibodies used for detection were a previously described custom rabbit anti-CLAMP (1:1,000, Abcam) (Rieder et al., 2017). Mouse anti-actin (1:400,000, Sigma Aldrich) was used as a loading control.

Quantitative real-time PCR

To determine transcript abundance of *clamp*, *msl2*, or *trl* RNA was calculated using the $2^{-\Delta\Delta C_t}$ method (Livak and Schmittgen, 2001) using RNA extracted from 500ul of cells following the 6 day incubation using an RNeasy Plus RNA extraction kit (Qiagen). Gene targets were amplified from cDNA using previously validated primers for *clamp*, *msl2*, *trl*, and three internal control genes (*gapdh*, *rpl32*, and *ras64b*) using triplicate technical replicates for each biological replicate per condition. Samples were then normalized to the *gfp* RNAi (control) condition.

Hi-C experimental procedure

Hi-C libraries of two independent biological replicates per RNAi condition (*clamp* and *gfp*) were generated as follows:

Cell lysis and restriction enzyme digestion

Approximately 10 million formaldehyde crosslinked (crosslinking procedure described above) S2 cells were resuspended in 500ul of fresh cold lysis buffer (10mM Tris-HCL pH 8.0, 10mM NaCl, 10ul of 50X Protease inhibitors cocktail, 0.2% Igepal CA630 in UltraPure water) and incubated on ice for 30 minutes. Cells were then centrifuged for 5 minutes at 3000 RPM at 4C and the supernatant removed. The pellet was resuspended with 300ul of cold 1x NEBuffer2. Cells were again centrifuged at 3000 RPM at 4C and the supernatant was removed. Cells were resuspended in 95ul of 1X NEBuffer2 and 5ul of 10% SDS was added. The cell solution was homogenized by gentle pipetting and then incubated at 65C for

10 minutes. The cell solution was then placed on ice and 200ul of 1X NEBuffer2 along with 60ul of 10% Triton X-100 (to quench the SDS) was added. The cell solution was incubated at 37C for 15 minutes. Lysis efficiency and nuclei integrity was checked via microscope. The samples were centrifuged for 5 minutes at 3000 RPM at 4C and the supernatant was removed. The pellet was resuspended in 300ul of 1X NEBuffer2 and 400 U of HindIII restriction enzyme was added. The samples were incubated overnight at 37C. The following morning an additional 200U of HindIII was added to each of the samples and they were incubated an additional 2 hours at 37C.

End-repair, labeling, in-nuclei ligation, and crosslink reversal

The samples were centrifuged for at 3000 RPM for 5 minutes at 4C and the supernatant removed. The samples were resuspended with 250ul of 1X NEBuffer and 50ul of the following mix was added: 1.5ul of 10mM dATP, 1.5ul of 10mM dGTP, 1.5ul of 10mM dTTP, 37.5ul of 0.4mM Biotin-11-dCTP, 1ul of 50U/ul DNA Polymerase I Large (Klenow) fragment, and 7ul of UltraPure water. The sample were incubated at 37C for 45 minutes and then incubated at 65C for 15 minutes. The samples were centrifuged for 5 minutes at 3000 RPM and the supernatant discarded. Samples were resuspended in 1.195mL of the following ligation mix: 120ul of 10X T4 DNA Ligase Buffer, 100ul of 10% Triton X-100, 12ul of 10mg/mL BSA and 963ul of UltraPure water. 5ul of 2000U/ul T4 DNA ligase was then added to each sample and the samples were incubated at 16C overnight. The samples

were centrifuged for 5 minutes at 3,000 RPM and supernatant removed. The samples were resuspended in 400ul of 1X NEBuffer2 and 10ul of 10mg/mL RNase A was added and the samples were incubated for 15 minutes at 37C at 300 RPM shaking. 20ul of 10mg/ml Proteinase K was added and the samples were incubated overnight at 65C to reverse crosslinks at 300 RPM shaking. The next morning an additional 20ul of Proteinase K was added again and the sample were incubated for 2 hours at 65C.

DNA purification

The samples were cooled to room temperature and 400ul of an equal volume of Phenol/Chloroform/Isoamyl Alcohol was added to each sample and the solution was mixed vigorously. The samples were centrifuged for 5 minutes at 13,000 RPM and the upper aqueous phase was transferred to a new tube. DNA was precipitated by adding 40ul of NaAc pH 5.2 and 1mL of 100% ethanol. The samples were incubated for 30 minutes at -80C and then centrifuged at 13,000 RPM for 30 minutes at 4C. The supernatant was discarded and pellet was washed with 1mL of Ethanol 70%. The sample was centrifuged for 15 minutes at 13,000 RPM at 4C. The supernatant was discarded and the pellet was air-dried for 5 minutes. The pellet was resuspended in 1X TE buffer. DNA shearing and size selection, biotin pull-down and sequencing library preparation were performed as described in (Rao '14). Multiplexed libraries were sequenced on an Illumina Hi-Seq 2500 configured for 150-bp paired-end reads.

4C-seq experimental procedure

4C-seq libraries were generated from S2 cells. Nuclear extraction and crosslinking were carried out as described in Hi-C experimental procedure above. The remainder of the 4C-seq procedure was carried out as previously described (Helm et al '14). Csp6I, DpnII or NlaIII were used as primary or secondary restriction enzymes. Primer sequences for each viewpoint and condition were generated using 4C primer design (<https://mnlab.uchicago.edu/4Cpd/>) and are listed in supplemental file S4.2. Independent biological replicates of multiplexed libraries were sequenced on separate lanes of an Illumina Hi-Seq 2500 configured for 150-bp paired-end reads.

Datasets

CES locations were obtained from GSE39271. Hi-C following RNAi against *msl2*, *msl3*, or *gfp* was obtained from GSE58821. MACC data was obtained from GSE99894. S2 cell chip-seq data sets for CP190, SuHw, Mod, and dCTCF GSE41354. S2 cell chip-seq data sets for GAF and CLAMP were obtained from GSE107059. Raw and processed sequencing data generated are deposited at NCBI GEO GSE130546.

Hi-C analysis

Hi-C data was processed using the HiC-Pro pipeline version 2.7.8 (Servant et al., 2015) for read mapping (MIN_MAPQ = 15), filtering and quality checks to generate

valid read pairs and interaction matrices. GenomeDisco version 1.0.0 (Ursu et al., 2018) was utilized to determine concordance between biological replicates. Custom scripts were used to convert Hi-C-Pro interaction matrices to a format suitable for input into GenomeDisco. For matrix visualizations valid pairs were processed into .hic files using the hicpro2juicebox script provided with HiC-Pro. Matrices were visualized using Juicebox (Durand et al., 2016) with KR matrix balancing.

PCA and DLR analysis

Valid pairs for biological replicates were merged and imported into Homer version 4.10 (Heinz et al., 2010) using the command makeTagDirectory with the parameter -format HiCsummary. Principal component analysis (PCA) was performed using Homer with the command runHiCpca.pl at resolutions of 10kb, 20kb and 50kb. Distal to local log2 ratio analysis was performed using Homer with the command analyzeHiC with the following parameters: -res 5000 -window 15000 -compactionStats auto -dlrDistance 250000. Distal to local log2 ratio differential analysis was performed using the command subtractBedGraphsDirectory.pl with the parameter -center.

Identification of high confidence long-range interactions

20kb resolution contact matrices corresponding to the control (*gfp* RNAi) conditions were generated and iterative corrected using HiC-Pro. The resultant

contact matrices were then adapted for Fit-Hi-C version 2.0.5 (Ay et al., 2014) using the hicpro2Fit-Hi-C.py script provided with Hi-C-Pro. Raw contact matrices and ICE biases were used as input into Fit-Hi-C to identify significant contacts with the following parameters: -L 5000000 -U 15000000 -v -b 100 -p2. Significant contacts were filtered for those with $q < 0.01$ and the intersection of significant contacts ($q < 0.01$) from the replicates was used as the high confidence interaction set.

Identifying Differential interactions (DIs)

diffHiC version 1.14.0 (Lun and Smyth, 2015) which uses the edgeR (Robinson et al., 2010) package for differential statistics was utilized to identify differential interactions. Valid pairs from HiC-Pro for each replicate was imported into diffHiC using the savePairs function and custom scripts. Low-abundance read pairs were filtered out and the resulting data was normalized with TMM normalization and trended biases were removed. Differential interactions were identified using an FDR target of 0.05 and the functions diClusters, combineTests, and getBestTest.

4C-Seq analysis

4C sequencing reads were demultiplexed based on index sequences and inline barcodes. Sequencing reads corresponding to the reading primer were aligned to a reduced genome derived from dm6 unique sequences adjacent to restriction

enzyme sites using Bowtie2 version 2.3.0 (Langmead and Salzberg, 2012) with the following parameters: -N 0 -5 24 -3 101 --very-sensitive, which removes barcode and primer sequences and trim the read length to 25bp prior to mapping. The aligned reads were then processed as described for use with the 4C-Ker pipeline (Raviram et al., 2016) which uses a three state Hidden-Markov Model to find chromosome-wide interactions and DESeq2 (Love et al., 2014) to perform differential analysis. Quality checks were performed within the 4C-Ker pipeline. *Cis* analysis was performed for each viewpoint using k=10 and p-value cutoff of 0.05 was used for differential analysis. Viewpoints were visualized using IGV (Thorvaldsdóttir et al., 2013).

Identification of high confidence long-range interactions

A set of high confidence interactions for each viewpoint was determined by intersecting the significant interactions individually detected in each replicate of the control (*gfp* RNAi) condition using BEDTools version 2.27.1 (Quinlan, 2014).

Comparison of 4C and Hi-C high confidence interactions

High confidence Hi-C interactions involving each 4C-seq viewpoint was determined by intersecting the viewpoint coordinate sequence with the Hi-C high confidence interactions set. This set of interactions was then intersected with the 4C-seq high confidence interactions set to determine overlapping pairwise high

confidence interactions identified in both experiments. All intersections were performed using custom scripts or BEDTools.

Chromatin state annotation

Chromatin states were from the modENCODE project (DCC id: modENCODE_3363) (Kharchenko et al., 2011; Roy et al., 2010) and were lifted over to dm6 with the UCSC liftOver tool (<http://genome.ucsc.edu>) (Kent et al., 2002). Ratios were normalized to account for the per chromosome abundance of each individual chromatin state. Jaccard similarity scores were computed using BEDTools.

Chromatin accessibility (MACC) analysis

MACC data generated following RNAi knockdown of *gfp* (control), *clamp*, or *msl2* was lifted over to dm6 with the UCSC liftover tool. MACC data was intersected with genomic regions of interest using custom scripts and BEDTools to determine corresponding MACC scores. Other analyses were performed in the python programming environment (<https://www.python.org>).

Chip-seq data processing and generation of high confidence peak sets

For each ChIP-seq dataset used, the respective sequencing reads were downloaded and mapped to release 6 *D. melanogaster* genome (dm6) (Hoskins et al., 2015) using Bowtie2 with parameter -N 1. Reads with a MAPQ <30 and PCR

duplicate reads identified using Picard MarkDuplicates version 2.9.2 (Van der Auwera et al., 2013) were removed using SAMtools version 1.9 (Li et al., 2009). Reads mapped to dm3 blacklisted regions (<https://sites.google.com/site/anshulkundaje/projects/blacklists>) lifted over to dm6, using UCSC liftOver, were also removed. In cases where biological replicates were not available the aligned reads were split into pseudoreplicates. MACS2 version 2.1.1 (Zhang et al., 2008) was used to identify peaks with the following parameters: --nomodel -B --SPMR --keep-dup all -f AUTO -g dm -p 0.01. MACS2 was also used generate fold enrichment (default parameters) bedGraphs for each factor. In order to reduce the number of false positive peaks the irreproducible discovery rate (IDR) was calculated using IDR version 2.0.3 (Li et al., 2011) using the MACS2 peak score calculated for each replicate experiment (biological replicate in the case of GAF and CLAMP; pseudoreplicate in the case of CP190, SuHw, Mod, dCTCF) as input to IDR. Peaks with an $IDR < 0.01$ were retained (In cases where there were more than two biological replicates pairwise IDR comparison of each replicate was made and the longest resulting peak list was used). BEDTools and UCSC bedGraphToBigWig (Kent et al., 2010) tool was used to convert bedGraphs to bigwig format. Average profiles were generated using deepTools version 3.1.0 (Ramírez et al., 2014).

Supporting Information

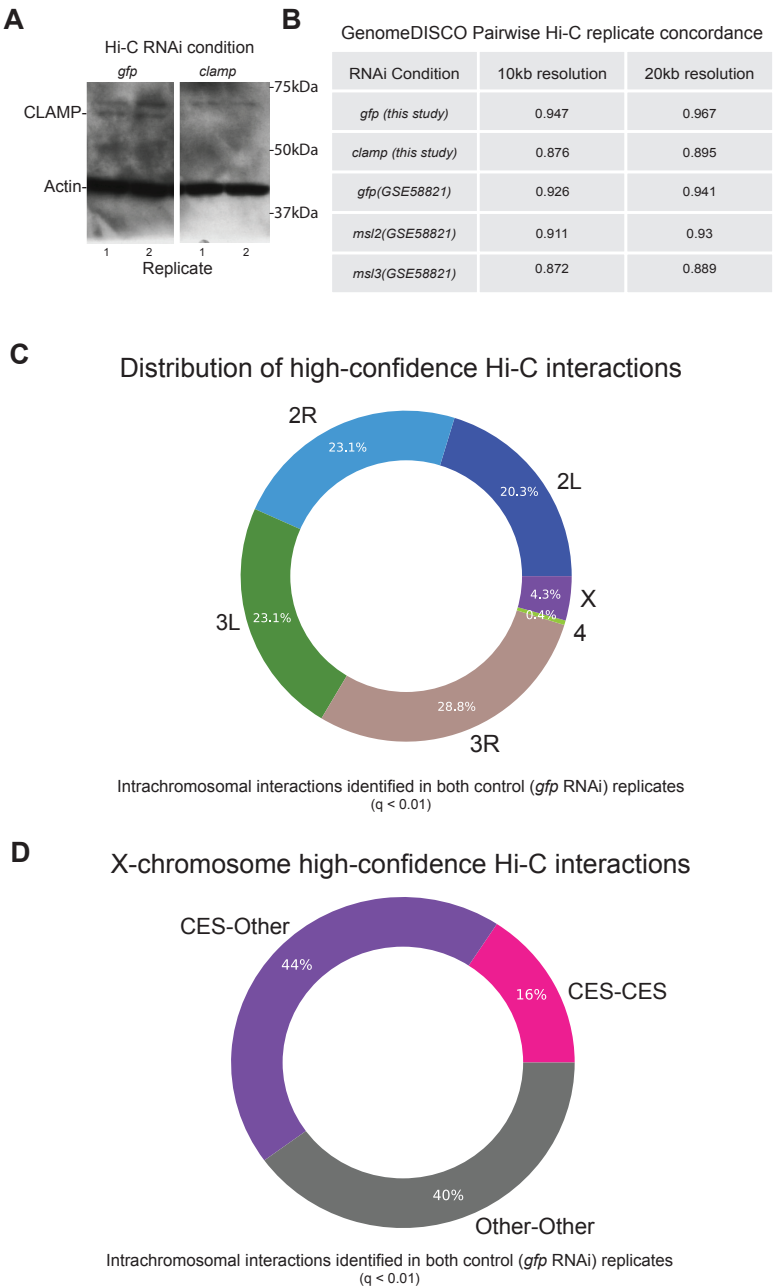


Figure S4.1. Experimental QC and Hi-C Replicate Concordance

- A. Western blot for *gfp* and *clamp* RNAi Hi-C experiments indicating successful knockdown of CLAMP protein.
- B. Pairwise Hi-C replicate concordance scores as measured by GenomeDISCO.
- C. The per chromosome distribution of high confidence Hi-C interactions (see also File S4.3).
- D. The distribution of X-chromosome high confidence Hi-C interactions that involve a CES.

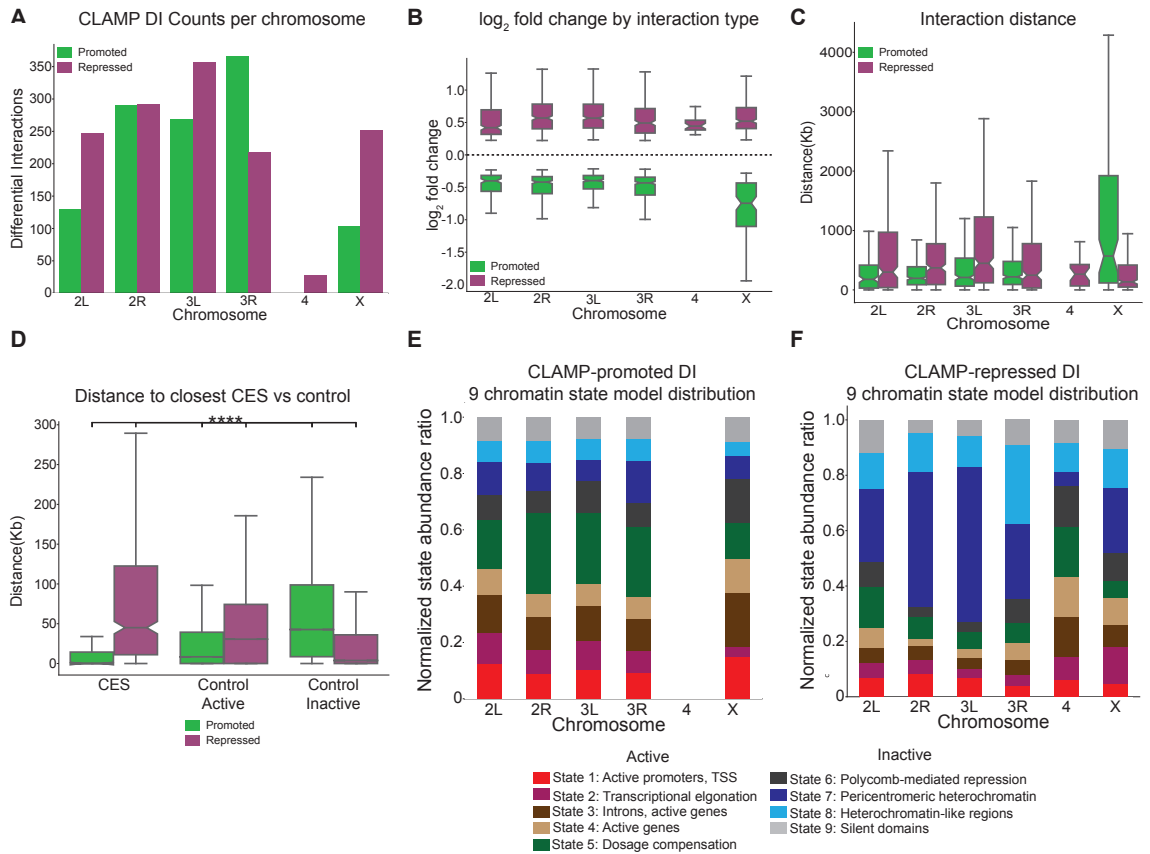


Figure S4.2. CLAMP promotes long-range interactions on the X-chromosome and generally promotes interactions in active chromatin and represses interactions in inactive chromatin

A. Per chromosome differential interaction count of CLAMP-promoted and CLAMP-repressed interactions (see also file S4.4).

B. Per chromosome $\log_2$ fold changes by interaction type for CLAMP-promoted and CLAMP-repressed interactions.

C. Per chromosome distribution of distances between differential interaction anchors for CLAMP-promoted and CLAMP-repressed interactions.

D. Distribution of distance for Hi-C identified CLAMP-promoted and CLAMP-repressed interactions to nearest CES, randomized active region control, or randomized inactive control. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$, two-sided Mann-Whitney test.

E. Per chromosome normalized ratio of chromatin states occurring at CLAMP-promoted interactions.

F. Per chromosome normalized ratio of chromatin states occurring at CLAMP-repressed interactions.

For all box and whisker plots, the 95% confidence interval is shown with a notch around the median line.

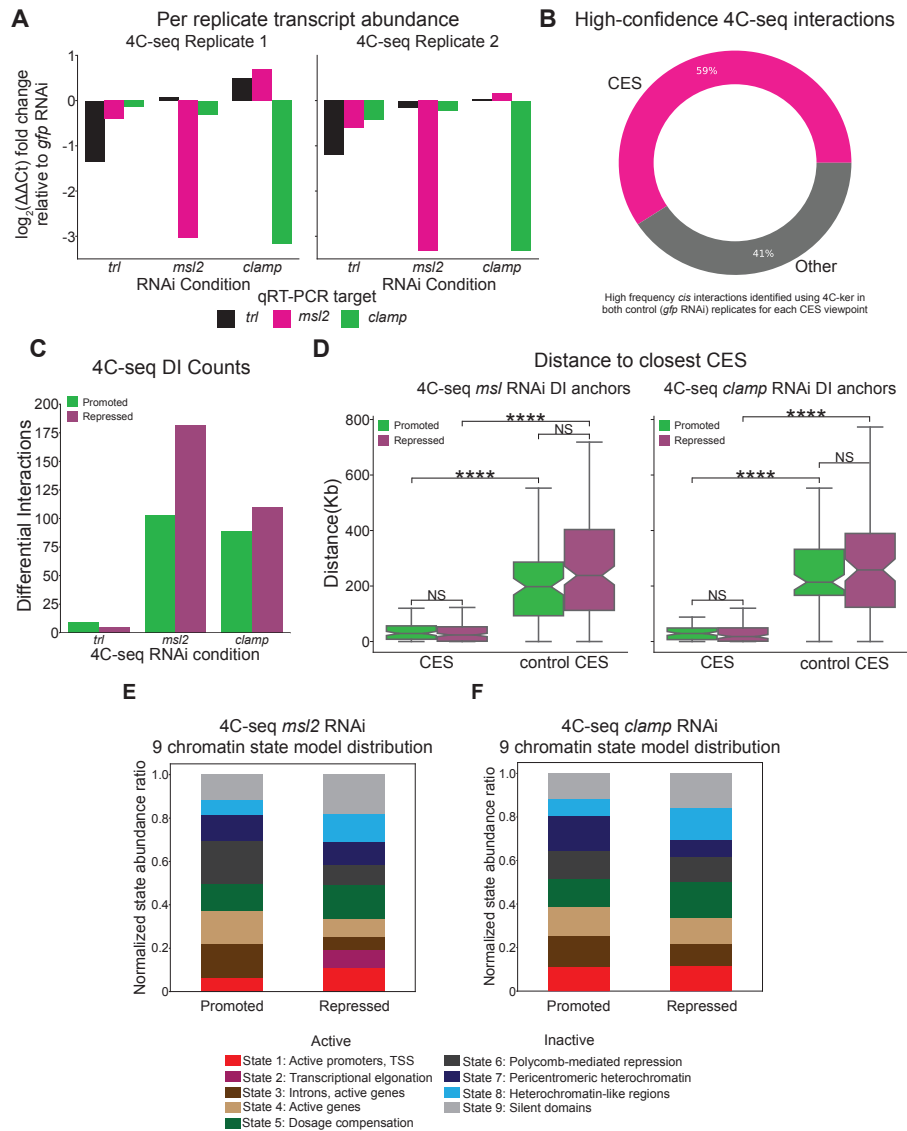


Figure S4.3. Summary of high-resolution 4C-seq analysis

A. Quantitative real-time PCR indicates successful RNAi knockdown of each target gene. Plotted is the log₂ fold change (ΔΔCt) for each biological replicate after internal normalization to three control genes (*gapdh*, *rpl32*, and *ras64b*). Samples are normalized to the *gfp* RNAi condition.

B. Percentage of high-confidence *cis* interactions for all 4C-seq viewpoints (control *gfp* RNAi condition) that correspond to a region containing a CES.

C. 4C-seq differential interaction counts per RNAi condition identified using 4C-ker.

D. Distribution of distances to nearest CES or control CES for MSL2 and CLAMP-promoted and CLAMP-repressed interactions. For all box and whisker plots, the 95% confidence interval is shown with a notch around the median line. * p<0.05, ** p<0.01, *** p<0.001, **** p<0.0001, two-sided Mann-Whitney test.

E. Ratio of chromatin states, normalized for X-chromosome state abundance, occurring at MSL2- promoted and MSL2-repressed interactions

F. Ratio of chromatin states, normalized for X-chromosome state abundance, occurring at CLAMP- promoted and CLAMP-repressed interactions

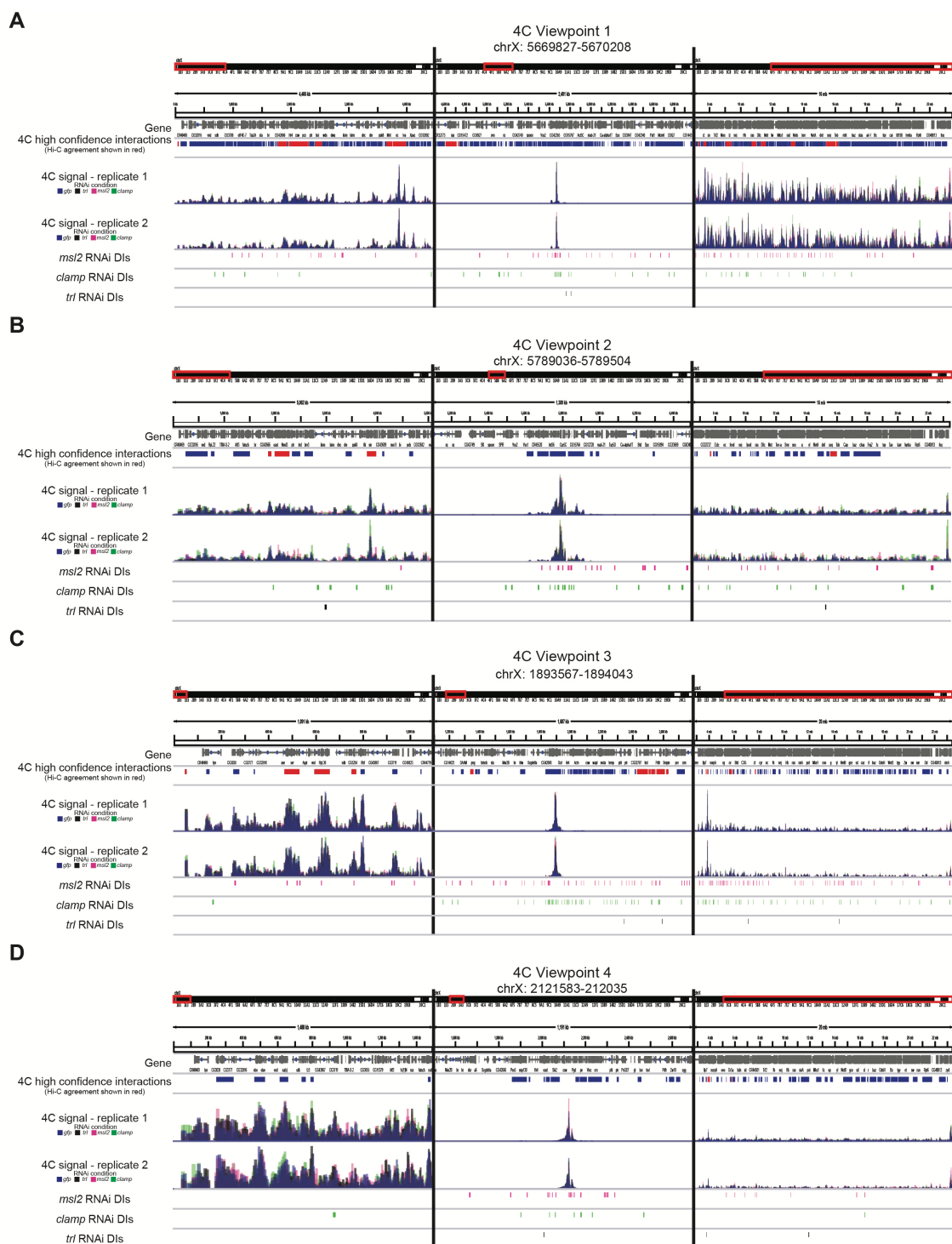


Figure S4.4. Overview of differential 4C-seq analysis after *gfp*, *msl2*, or *clamp* RNAi
Top track represents high-confidence 4C interactions (see also file S4.3) along with those that are also found in the high-confidence Hi-C interactions (red). Autoscaled signal tracks are shown per replicate for each of the following RNAi conditions: *gfp*(blue), *tr*(black), *msl2*(pink), and *clamp*(green). Followed by their respective differential interactions compared to control (*gfp*) as identified by 4C-ker (see also file S4.4).

- A. 4C-seq viewpoint 1 - chrX:5669827-5670208)
- B. 4C-seq viewpoint 2 - chrX:5789036-5789504
- C. 4C-seq viewpoint 3 - chrX:1893567-1894043
- D. 4C-seq viewpoint 4 - chrX:2121583-2122035

Supplemental Files

Available upon request

File S4.1. Processing of Hi-C and 4C sequencing reads. This spreadsheet contains mapping and filtering data from processing Hi-C data with HiC-Pro and 4C-seq data with 4C-ker (see methods for analysis description).

File S4.2. 4C Viewpoints. This spreadsheet contains the viewpoint coordinates, primers, and primary and secondary restriction enzymes.

File S4.3. High-confidence interactions. This spreadsheet contains the high-confidence interactions identified for Hi-C and 4C-seq (see methods for analysis description).

File S4.4. Differential Analysis of Hi-C and 4C-seq. This spreadsheet contains the results of differential analysis performed for Hi-C using diffHiC and 4C-seq using 4C-ker (see methods for analysis description).

CHAPTER 5 : DISCUSSION

William Jordan, III¹ and Erica Larschan¹

I wrote this chapter with editorial feedback from Dr. Erica Larschan.

¹ Department of Molecular Biology, Cell Biology, and Biochemistry, Brown University, Providence, Rhode Island 02912

Diverse models of combinatorial transcriptional regulation in *Drosophila*

The existence of diverse *cis*-regulatory elements (CREs) and their interplay with equally diverse transcription factors allows for precise modulation of gene expression; ultimately this interplay gives rise to the highly specialized, phenotypically diverse, cell types found within complex biological systems. Specific subsets of bound CREs may vary widely from cell to cell. Furthermore, CRE density (i.e. number), proximity (i.e. spacing between multiple CREs), or sequence variations confer the potential to recruit a single transcription factor or combinations of multiple transcription factors to specific genomic loci. These characteristics further increase the diversity of potential transcription factor binding patterns. Moreover, transcription factors are capable of directly recruiting other factors that may lack direct DNA binding capabilities such as coactivators or corepressors. Distinct combinations of transcription factors may act synergistically to enhance one another's ability to bind genomic loci containing their requisite CREs. Alternatively, they may also act antagonistically and compete with one another for binding at the same, or similar, CREs. Furthermore, it is also important to consider that many CREs are capable of acting in *trans*; therefore, having the potential to regulate transcriptional expression of nearby or distal genes. These combinatorial relationships between transcription factors and CREs allow for the encoding of complex genomic regulatory programs.

It remains a challenge to accurately identify and characterize the molecular mechanisms responsible for the context-specific functions a transcription factor may exhibit upon recruitment to specific CRE subsets. Additionally, transcription factor binding in various genomic contexts within the same or different cells may have seemingly divergent effects, further confounding characterization efforts.

Drosophila dosage compensation is an excellent model to better define the relationships between CREs and transcription factors because genes along a single chromosome are specifically targeted for co-regulation. The male-specific lethal complex (MSLc) localizes to specific GA-rich CREs, known as MSLc recruitment elements (MREs), along the single male X-chromosome to mediate dosage compensation. Furthermore, MSLc requires a non sex-specific adapter protein, Chromatin Linked Adapter for MSL Proteins (CLAMP), in order to bind chromatin. CLAMP is capable of binding the GA-rich MREs within CES but many other GA-rich CREs found throughout the genome. It has been shown at CES, the strongest recruitment sites of MSLc, MREs typically occur in greater density and are positioned in much closer proximity compared to other sites along the X-chromosome and the autosomes. In addition to this linear bunching of CREs, it has also been shown that many CES cluster three-dimensionally.

Identification of additional features that allow the X-chromosome to be distinguished from the autosomes

In Chapter 2, I show that the observed patterns of increased MRE proximity and density within CES are necessary but not sufficient for MSLc recruitment *in vivo*. To date, no secondary motifs have been found that may provide insight into the X-specific recruitment of MSLc to a small subset of CLAMP CREs. Therefore, in Chapter 3, I used machine learning to identify candidate genomic features that may play a novel role in X-identification or MSLc recruitment to X-linked MREs. One of the top candidates I identified is negative elongation factor (NELF). It has been previously reported that NELF is physically associated with CLAMP via mass spectrometry experiments; CLAMP has also been shown to inhibit NELF recruitment at highly paused genes (Urban et al., 2017a, 2017b). Therefore, it will be of interest to further define the relationship between MSLc, CLAMP, and NELF in addition to other factors (Chapter 3).

In Chapter 4, I show that CLAMP is responsible for the observed three-dimensional clustering of CES. Furthermore, CLAMP more strongly regulates three-dimensional interactions on the X-chromosome versus autosomes. CLAMP also promotes the formation of longer-range genomic interactions on the X-chromosome compared to the autosomes. More specifically, CLAMP promotes interactions that predominately involve CES and other regions of active chromatin. However, it remains unknown how CLAMP differentially mediates genomic

interactions on the X-chromosome compared with the autosomes. Given that CES cluster in three-dimensions and many CREs are capable of acting in *trans* over long distances, it will be interesting to explore the possibility of specific, long-distance, CRE combinations that may have been missed when exclusively examining CES in linear space. It is also of interest to examine the differences between CES or MREs that appear to exhibit varying degrees of CLAMP-dependent three-dimensional clustering patterns.

Outside of the context of its relationship with MSLc on the male X-chromosome, CLAMP has been reported to have seemingly conflicting roles in both transcriptional activation and repression throughout the genome. Furthermore, CLAMP has been shown to mediate changes in chromatin accessibility on the X-chromosome as well as many other specific loci throughout the genome. Interestingly, GAGA associated factor (GAF), also binds to GA-rich CREs bound by CLAMP and CLAMP outcompetes GAF at CES. Furthermore, many other factors in addition to MSLc and GAF have also been linked to CLAMP including insulator proteins, chromatin remodelers (Kaye et al., 2017; Urban et al., 2017b). Therefore, it will be essential to define how specific combinations of CLAMP-interacting factors influence the observed context-specific roles throughout the genome.

Resolving heterogeneity is necessary to better understand combinatorial transcriptional regulation

CLAMP exhibits synergy with MSLc; at CES, CLAMP and MSLc cooperatively increase the occupancy of one another (Albig et al., 2018; Soruco et al., 2013). In contrast, CLAMP and GAF exhibit the opposite relationship and compete for binding at certain genomic loci. The residence time of transcription factors bound to DNA *in vivo* is thought to be very short lived, in the time scale of mere seconds. Given the short but varying times factors are bound to DNA, it is incredibly difficult to know if factors are legitimately interacting with one another other. Therefore, it is also difficult to access if prospective interactions occur synergistically, antagonistically, or neutrally. Currently no method exists that allow for multiple transcription factors to be spatiotemporally assessed in a high-throughput genome-wide manner. Furthermore, crosslinking based approaches may trap nearby, but non-interacting factors resulting in false positives. The development of methods that allow simultaneous probing of multiple factors under native conditions will allow for better discrimination of legitimate chromatin interaction events versus false positives that occur as a result of genomic proximity.

Many experimental approaches rely on the use of bulk cell populations where cell-to-cell heterogeneity is lost; averaging over a population of cells results in the loss of information that may identify which CRE subsets are bound context-specifically from cell to cell. High-throughput single-cell methods that have the potential to

resolve some cellular heterogeneity issues exist, however they are costly to adopt. One possible alternative is to use lower throughput single-molecule imaging approaches to complement genomic approaches and simultaneously measure the binding dynamics of specific transcription factor combinations (Mir et al., 2018). This approach has the potential to offer additional insight into whether two or more factors are capable of interacting directly on chromatin. Additionally, many analysis pipelines discard potentially valuable information that corresponds to repetitive genomic regions or rely on signal averaging (i.e. average profiles). The use of single-cell approaches when economically feasible along with the development of more advanced analysis pipelines that avoid discarding reads aligning to repetitive regions will prove useful in the future for higher throughput interrogation of genomic interactions between multiple transcription factors. New computational methods that are able to intrinsically account for the potential of multiple distinct classes of genomic profiles within either a single cell, or bulk cell populations, will also provide new insight into combinatorial transcriptional regulation in diverse genomic contexts.

Summary

In comparison to the X-chromosome, the autosomes have decreased proximity and density of MRE like, GA-rich, CREs; this arrangement may largely prevent aberrant autosomal recruitment of MSLc to the thousands of MRE-like CREs found on the autosomes (Chapter 2). My machine learning analysis of X-identification

generated many new leads to explore factors that may directly or indirectly help further promote MSLc recruitment specifically to the X-chromosome (Chapter 3). My analysis also reveals that though CLAMP binds to GA-rich CREs genome-wide, it specifically promotes long-range interactions on the dosage compensated X-chromosome. These long-range interactions promote the clustering of CES with one another and other active regions of chromatin (Chapter 4). Lastly, I demonstrate that genome-wide CLAMP is capable of promoting interactions between active chromatin regions and repressing interactions in inactive chromatin regions throughout the genome (Chapter 4). My dissertation provides key insight into how a single transcription factor performs a diverse repertoire of context-specific gene regulation.

Understanding the diverse context-specific roles of CLAMP and the cofactors it recruits or interacts with at specific genomic loci will contribute to the following: 1) A better understanding of the interaction dynamics between CREs and transcription factors. 2) The discovery of molecular mechanisms responsible for the context-specific effects of transcription factors at various genomic loci. Using a combination of *in vivo* and *in silico* approaches, I have provided key insights into many of the diverse mechanisms CLAMP uses to modulate context-specific gene regulation at the seemingly similar GA-rich CREs it binds throughout the genome. In the future, it will be exciting to use single-cell genomic methods along with advanced analysis tools that avoid averaging and allow for the discovery of multi-

modal signals to gain further insight into the dynamic processes underlying context-specific gene regulation.

BIBLIOGRAPHY

- Adelman, K., and Lis, J.T. (2012). Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat. Rev. Genet.* *13*, 720.
- Albig, C., Tikhonova, E., Krause, S., Maksimenko, O., Regnard, C., and Becker, P.B. (2018). Factor cooperation for chromosome discrimination in *Drosophila*. *Nucleic Acids Res.*
- Albritton, S.E., Kranz, A.-L.L., Rao, P., Kramer, M., Dieterich, C., and Ercan, S. (2014). Sex-biased gene expression and evolution of the X chromosome in nematodes. *Genetics* *197*, 865–883.
- Albritton, S.E., Kranz, A.-L., Winterkorn, L.H., Street, L.A., and Ercan, S. (2017). Cooperation between a hierarchical set of recruitment sites targets the X chromosome for dosage compensation. *Elife* *6*, e23645.
- Alekseyenko, A.A., Larschan, E., Lai, W.R., Park, P.J., and Kuroda, M.I. (2006). High-resolution ChIP-chip analysis reveals that the *Drosophila* MSL complex selectively identifies active genes on the male X chromosome. *Genes Dev.* *20*, 848–857.
- Alekseyenko, A.A., Peng, S., Larschan, E., Gorchakov, A.A., Lee, O.-K.K., Kharchenko, P., McGrath, S.D., Wang, C.I., Mardis, E.R., Park, P.J., et al. (2008). A Sequence Motif within Chromatin Entry Sites Directs MSL Establishment on the *Drosophila* X Chromosome. *Cell* *134*, 599–609.
- Alekseyenko, A.A., Ellison, C.E., Gorchakov, A.A., Zhou, Q., Kaiser, V.B., Toda, N., Walton, Z., Peng, S., Park, P.J., and Bachtrog, D. (2013).

- Conservation and de novo acquisition of dosage compensation on newly evolved sex chromosomes in *Drosophila*. *Genes Dev.* *27*, 853–858.
- Armstrong, R.L., Penke, T.J.R., Strahl, B.D., Matera, A.G., McKay, D.J., MacAlpine, D.M., and Duronio, R.J. (2018). Chromatin conformation and transcriptional activity are permissive regulators of DNA replication initiation in *Drosophila*. *Genome Res.* *28*, 1688–1700.
- Aumiller Jr, W.M., Davis, B.W., and Keating, C.D. (2014). Phase separation as a possible means of nuclear compartmentalization. In *International Review of Cell and Molecular Biology*, (Elsevier), pp. 109–149.
- Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., and Thibault, J. (2013). From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* *43*, 10–11.
- Ay, F., Bailey, T.L., and Noble, W.S. (2014). Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res.*
- Bag, I., Dale, R.K., Palmer, C., and Lei, E.P. (2019). The zinc-finger protein CLAMP promotes gypsy chromatin insulator function in *Drosophila*. *J Cell Sci* jcs-226092.
- Bai, X., Alekseyenko, A.A., and Kuroda, M.I. (2004). Sequence-specific targeting of MSL complex regulates transcription of the roX RNA genes. *23*, 2853–2861.
- Bar-Even, A., Paulsson, J., Maheshri, N., Carmi, M., O’Shea, E., Pilpel, Y., and

- Barkai, N. (2006). Noise in protein expression scales with natural protein abundance. *Nat. Genet.* *38*, 636.
- Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., and Holko, M. (2012). NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* *41*, D991–D995.
- Barutcu, A.R., Lajoie, B.R., McCord, R.P., Tye, C.E., Hong, D., Messier, T.L., Browne, G., van Wijnen, A.J., Lian, J.B., and Stein, J.L. (2015). Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells. *Genome Biol.* *16*, 214.
- Basilicata, M.F., Bruel, A.-L., Semplicio, G., Valsecchi, C.I.K., Aktaş, T., Duffourd, Y., Rumpf, T., Morton, J., Bache, I., and Szymanski, W.G. (2018). De novo mutations in MSL3 cause an X-linked syndrome marked by impaired histone H4 lysine 16 acetylation. *Nat. Genet.* *50*, 1442.
- Bateman, J.R., and Lee, A.M. (2006). Site-specific transformation of *Drosophila* via ϕ C31 integrase-mediated cassette exchange. *Genetics* *173*, 769–777.
- Beliveau, B.J., Boettiger, A.N., Avendano, M.S., Jungmann, R., McCole, R.B., Joyce, E.F., Kim-Kiselak, C., Bantignies, F., Fonseka, C.Y., Erceg, J., et al. (2015). Single-molecule super-resolution imaging of chromosomes and in situ haplotype visualization using Oligopaint FISH probes. *Nat. Commun.* *6*, 7147.

- Belote, J.M., and Lucchesi, J.C. (1980). Male-specific lethal mutations of *Drosophila melanogaster*. *Genetics* *96*, 165–186.
- Bohla, D., Herold, M., Panzer, I., Buxa, M.K., Ali, T., Demmers, J., Krüger, M., Scharfe, M., Jarek, M., Bartkuhn, M., et al. (2014). A functional insulator screen identifies NURF and dREAM components to be required for enhancer-blocking. *PLoS One* *9*, 1–13.
- Bonev, B., and Cavalli, G. (2016). Organization and function of the 3D genome. *Nat. Rev. Genet.* *17*, 661.
- Bradley, R.K., Li, X.-Y., Trapnell, C., Davidson, S., Pachter, L., Chu, H.C., Tonkin, L.A., Biggin, M.D., and Eisen, M.B. (2010). Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related *Drosophila* species. *PLoS Biol.* *8*, e1000343.
- Brown, E.J., and Bachtrog, D. (2014). The chromatin landscape of *Drosophila*: comparisons between species, sexes, and chromosomes. *Genome Res.* *24*, 1125–1137.
- Buckland, M., and Gey, F. (1994). The relationship between recall and precision. *J. Am. Soc. Inf. Sci.* *45*, 12–19.
- Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* *10*, 1213.
- Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder,

- M.P., Chang, H.Y., and Greenleaf, W.J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* 523, 486.
- Camacho, D.M., Collins, K.M., Powers, R.K., Costello, J.C., and Collins, J.J. (2018). Next-generation machine learning for biological networks. *Cell*.
- Caruana, R., and Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd International Conference on Machine Learning, (ACM)*, pp. 161–168.
- Cavalli, G., and Misteli, T. (2013). Functional implications of genome topology. *Nat. Struct. Mol. Biol.* 20, 290.
- Chatterjee, A., Seyfferth, J., Lucci, J., Gilsbach, R., Preissl, S., Böttinger, L., Mårtensson, C.U., Panhale, A., Stehle, T., and Kretz, O. (2016). MOF acetyl transferase regulates transcription and respiration in mitochondria. *Cell* 167, 722–738.
- Chen, T., and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, (ACM)*, pp. 785–794.
- Chen, X., Shen, Y., Draper, W., Buenrostro, J.D., Litzenburger, U., Cho, S.W., Satpathy, A.T., Carter, A.C., Ghosh, R.P., and East-Seletsky, A. (2016). ATAC-se reveals the accessible genome by transposase-mediated imaging and sequencing. *Nat. Methods* 13, 1013.
- Chiu, T.-P., Comoglio, F., Zhou, T., Yang, L., Paro, R., and Rohs, R. (2015). DNASHapeR: an R/Bioconductor package for DNA shape prediction and

- feature encoding. *Bioinformatics* 32, 1211–1213.
- Chuong, E.B., Elde, N.C., and Feschotte, C. (2017). Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.* 18, 71.
- Conrad, T., Cavalli, F.M.G., Holz, H., Hallaceli, E., Kind, J., Ilik, I., Vaquerizas, J.M., Luscombe, N.M., and Akhtar, A. (2012). The MOF chromobarrel domain controls genome-wide H4K16 acetylation and spreading of the MSL complex. *Dev. Cell* 22, 610–624.
- Consortium, E.P. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799.
- Copur, Ö., Gorchakov, A., Finkl, K., Kuroda, M.I., and Müller, J. (2018). Sex-specific phenotypes of histone H4 point mutants establish dosage compensation as the critical function of H4K16 acetylation in *Drosophila*. *Proc. Natl. Acad. Sci.* 115, 13336–13341.
- Cuartero, S., Fresán, U., Reina, O., Planet, E., and Espinàs, M.L. (2014). Ibf1 and Ibf2 are novel CP190-interacting proteins required for insulator function. *EMBO J.* 33, 637–647.
- Cusanovich, D.A., Daza, R., Adey, A., Pliner, H.A., Christiansen, L., Gunderson, K.L., Steemers, F.J., Trapnell, C., and Shendure, J. (2015). Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* (80-.). 348, 910–914.
- D’Souza, E., Hosage, E., Weinand, K., Gisselbrecht, S., Marktein, V., Markstein, P., Bulyk, M., and Markstein, M. (2018). GA-repeats on mammalian X

chromosomes support Ohno's hypothesis of dosage compensation by transcriptional upregulation. *BioRxiv* 485300.

- Darrow, E.M., Huntley, M.H., Dudchenko, O., Stamenova, E.K., Durand, N.C., Sun, Z., Huang, S.-C., Sanborn, A.L., Machol, I., and Shamim, M. (2016). Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture. *Proc. Natl. Acad. Sci.* *113*, E4504–E4512.
- Davie, K., Jacobs, J., Atkins, M., Potier, D., Christiaens, V., Halder, G., and Aerts, S. (2015). Discovery of transcription factors and regulatory regions driving in vivo tumor development by ATAC-seq and FAIRE-seq open chromatin profiling. *PLoS Genet.* *11*, e1004994.
- Davies, J.O.J., Oudelaar, A.M., Higgs, D.R., and Hughes, J.R. (2017). How best to identify chromosomal interactions: a comparison of approaches. *Nat. Methods* *14*, 125.
- Davis, J., and Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, (ACM), pp. 233–240.
- Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. *Science* (80-.). *295*, 1306–1311.
- Deng, X., Hiatt, J.B., Nguyen, D.K., Ercan, S., Sturgill, D., Hillier, L.W., Schlesinger, F., Davis, C.A., Reinke, V.J., Gingeras, T.R., et al. (2011). Evidence for compensatory upregulation of expressed X-linked genes in mammals, *Caenorhabditis elegans* and *Drosophila melanogaster*. *Nat.*

- Genet. 43, 1179–1185.
- Deng, X., Berletch, J.B., Ma, W., Nguyen, D.K., Hiatt, J.B., Noble, W.S., Shendure, J., and Disteche, C.M. (2013). Mammalian X upregulation is associated with enhanced transcription initiation, RNA half-life, and MOF-mediated H4K16 acetylation. *Dev. Cell* 25, 55–68.
- Deng, X., Berletch, J.B., Nguyen, D.K., and Disteche, C.M. (2014). X chromosome regulation: diverse patterns in development, tissues and disease. *Nat. Rev. Genet.* 15, 367.
- Disteche, C.M. (2012). Dosage compensation of the sex chromosomes. *Annu. Rev. Genet.* 46, 537–560.
- Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., and Xie, W. (2015). Chromatin architecture reorganization during stem cell differentiation. *Nature* 518, 331.
- Dixon, J.R., Gorkin, D.U., and Ren, B. (2016). Chromatin domains: the unit of chromosome organization. *Mol. Cell* 62, 668–680.
- Drysdale, R.A., Consortium, F., Crosby, M.A., and Consortium, F. (2005). FlyBase: genes and gene models. *Nucleic Acids Res.* 33, D390–D395.
- Durand, N.C., Robinson, J.T., Shamim, M.S., Machol, I., Mesirov, J.P., Lander, E.S., and Aiden, E.L. (2016). Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* 3, 99–101.
- Duronio, R.J., and Marzluff, W.F. (2017). Coordinating cell cycle-regulated histone gene expression through assembly and function of the Histone

- Locus Body. *RNA Biol.* *14*, 726–738.
- Ellison, C.E., and Bachtrog, D. (2013). Dosage compensation via transposable element mediated rewiring of a regulatory network. *Science* (80-.). *342*, 846–850.
- Ellison, C., Bachtrog, D., and Ellison, C. (2019). Contingency in the convergent evolution of a regulatory network: Dosage compensation in *Drosophila*. *PLOS Biol.* *17*, e3000094.
- Ercan, S. (2015). Mechanisms of X Chromosome Dosage Compensation. *J. Genomics* *3*, 1–19.
- Farkas, G., Gausz, J., Galloni, M., Reuter, G., Gyurkovics, H., and Karch, F. (1994). The *Trithorax*-like gene encodes the *Drosophila* GAGA factor. *Nature* *371*, 806–808.
- Farlik, M., Sheffield, N.C., Nuzzo, A., Datlinger, P., Schönegger, A., Klughammer, J., and Bock, C. (2015). Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell Rep.* *10*, 1386–1397.
- Fay, A., Misulovin, Z., Li, J., Schaaf, C.A., Gause, M., Gilmour, D.S., and Dorsett, D. (2011). Cohesin selectively binds and regulates genes with paused RNA polymerase. *Curr. Biol.* *21*, 1624–1634.
- Franke, A., and Baker, B.S. (1999). The *rox1* and *rox2* RNAs are essential components of the compensasome, which mediates dosage compensation in *Drosophila*. *Mol. Cell* *4*, 117–122.

- Franke, A., Dernburg, A., Bashaw, G.J., and Baker, B.S. (1996). Evidence that MSL-mediated dosage compensation in *Drosophila* begins at blastoderm. *Development* *122*, 2751–2760.
- Fuda, N.J., Guertin, M.J., Sharma, S., Danko, C.G., Martins, A.L., Siepel, A., and Lis, J.T. (2015). GAGA Factor Maintains Nucleosome-Free Regions and Has a Role in RNA Polymerase II Recruitment to Promoters. *PLOS Genet.* *11*, e1005108.
- Gavrilov, A.A., Gushchanskaya, E.S., Strelkova, O., Zhironkina, O., Kireev, I.I., Iarovaia, O. V, and Razin, S. V (2013). Disclosure of a structural milieu for the proximity ligation reveals the elusive nature of an active chromatin hub. *Nucleic Acids Res.* *41*, 3563–3575.
- Gertz, J., Savic, D., Varley, K.E., Partridge, E.C., Safi, A., Jain, P., Cooper, G.M., Reddy, T.E., Crawford, G.E., and Myers, R.M. (2013). Distinct properties of cell-type-specific and shared transcription factor binding sites. *Mol. Cell* *52*, 25–36.
- Gilbert, N., Boyle, S., Fiegler, H., Woodfine, K., Carter, N.P., and Bickmore, W.A. (2004). Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. *Cell* *118*, 555–566.
- Gilchrist, D.A., Nechaev, S., Lee, C., Ghosh, S.K.B., Collins, J.B., Li, L., Gilmour, D.S., and Adelman, K. (2008). NELF-mediated stalling of Pol II can enhance gene expression by blocking promoter-proximal nucleosome assembly. *Genes Dev.* *22*, 1921–1933.

- Giorgetti, L., Lajoie, B.R., Carter, A.C., Attia, M., Zhan, Y., Xu, J., Chen, C.J., Kaplan, N., Chang, H.Y., and Heard, E. (2016). Structural organization of the inactive X chromosome in the mouse. *Nature* *535*, 575.
- Graves, J.A.M. (2006). Sex chromosome specialization and degeneration in mammals. *Cell* *124*, 901–914.
- Greenbaum, J.A., Parker, S.C.J., and Tullius, T.D. (2007). Detection of DNA structural motifs in functional genomic elements. *Genome Res.* *17*, 940–946.
- Greenberg, A.J., Yanowitz, J.L., and Schedl, P. (2004). The *Drosophila* GAGA factor is required for dosage compensation in males and for the formation of the male-specific-lethal complex chromatin entry site at 12DE. *Genetics* *166*, 279–289.
- Griffiths, J.A., Scialdone, A., and Marioni, J.C. (2018). Using single-cell genomics to understand developmental processes and cell fate decisions. *Mol. Syst. Biol.* *14*, e8046.
- Grimaud, C., and Becker, P.B. (2009). The dosage compensation complex shapes the conformation of the X chromosome in *Drosophila*. *Genes Dev.* *23*, 2490–2495.
- Grosberg, A., Rabin, Y., Havlin, S., and Neer, A. (1993). Crumpled globule model of the three-dimensional structure of DNA. *EPL (Europhysics Lett.)* *23*, 373.
- Grosberg, A.Y., Nechaev, S.K., and Shakhnovich, E.I. (1988). The role of

- topological constraints in the kinetics of collapse of macromolecules. *J. Phys.* **49**, 2095–2100.
- Grskovic, M., Chaivorapol, C., Gaspar-Maia, A., Li, H., and Ramalho-Santos, M. (2007). Systematic identification of cis-regulatory sequences active in mouse and human embryonic stem cells. *PLoS Genet.* **3**, e145.
- Haberle, V., and Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. *Nat. Rev. Mol. Cell Biol.* **1**.
- Hafner, A., Stewart-Ornstein, J., Purvis, J.E., Forrester, W.C., Bulyk, M.L., and Lahav, G. (2017). p53 pulses lead to distinct patterns of gene expression albeit similar DNA-binding dynamics. *Nat. Struct. Mol. Biol.* **24**, 840.
- Hamada, F.N., Park, P.J., Gordadze, P.R., and Kuroda, M.I. (2005). Global regulation of X chromosomal genes by the MSL complex in *Drosophila melanogaster*. *Genes Dev.* **19**, 2289–2294.
- Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589.
- Heinz, S., Texari, L., Hayes, M.G.B., Urbanowski, M., Chang, M.W., Givarkes, N., Rialdi, A., White, K.M., Albrecht, R.A., and Pache, L. (2018). Transcription elongation can affect genome 3D structure. *Cell* **174**, 1522–1536.
- Hilfiker, A., Hilfiker-Kleiner, D., Pannuti, A., and Lucchesi, J.C. (1997). mof, a

putative acetyl transferase gene related to the Tip60 and MOZ human genes and to the SAS genes of yeast, is required for dosage compensation in *Drosophila*. *EMBO J.* *16*, 2054–2060.

Hoskins, R.A., Carlson, J.W., Kennedy, C., Acevedo, D., Evans-Holm, M., Frise, E., Wan, K.H., Park, S., Mendez-Lago, M., and Rossi, F. (2007).

Sequence finishing and mapping of *Drosophila melanogaster* heterochromatin. *Science* (80-.). *316*, 1625–1628.

Hoskins, R.A., Carlson, J.W., Wan, K.H., Park, S., Mendez, I., Galle, S.E., Booth, B.W., Pfeiffer, B.D., George, R.A., and Svirskas, R. (2015). The Release 6 reference sequence of the *Drosophila melanogaster* genome. *Genome Res.* *25*, 445–458.

Hug, C.B., Grimaldi, A.G., Kruse, K., and Vaquerizas, J.M. (2017). Chromatin architecture emerges during zygotic genome activation independent of transcription. *Cell* *169*, 216–228.

Ilik, I., and Akhtar, A. (2009). roX RNAs: non-coding regulators of the male X chromosome in flies. *RNA Biol.* *6*, 113–121.

Ilik, I.A., Quinn, J.J., Georgiev, P., Tavares-Cadete, F., Maticzka, D., Toscano, S., Wan, Y., Spitale, R.C., Luscombe, N., Backofen, R., et al. (2013). Tandem stem-loops in roX RNAs act together to mediate X Chromosome dosage compensation in *Drosophila*. *Mol. Cell* *51*, 156–173.

Jansz, N., Keniry, A., Trussart, M., Bildsoe, H., Beck, T., Tonks, I.D., Mould, A.W., Hickey, P., Breslin, K., and Iminoff, M. (2018). Smchd1 regulates

- long-range chromatin interactions on the inactive X chromosome and at Hox clusters. *Nat. Struct. Mol. Biol.* **25**, 766.
- Jordan III, W., Rieder, L.E., and Larschan, E. (2019). Diverse Genome Topologies Characterize Dosage Compensation across Species. *Trends Genet.*
- Kaplan, T., Li, X.-Y., Sabo, P.J., Thomas, S., Stamatoyannopoulos, J.A., Biggin, M.D., and Eisen, M.B. (2011). Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early *Drosophila* development. *PLoS Genet.* **7**, e1001290.
- Kasinathan, S., Orsi, G.A., Zentner, G.E., Ahmad, K., and Henikoff, S. (2014). High-resolution mapping of transcription factor binding sites on native chromatin. *Nat. Methods* **11**, 203.
- Kaye, E.G., Kurbidaeva, A., Wolle, D., Aoki, T., Schedl, P., and Larschan, E. (2017). *Drosophila* dosage compensation loci associate with a boundary forming insulator complex. *Mol. Cell. Biol.* MCB-00253.
- Kaye, E.G., Booker, M., Kurland, J. V, Conicella, A.E., Fawzi, N.L., Bulyk, M.L., Tolstorukov, M.Y., and Larschan, E. (2018). Differential Occupancy of Two GA-Binding Proteins Promotes Targeting of the *Drosophila* Dosage Compensation Complex to the Male X Chromosome. *Cell Rep.* **22**, 3227–3239.
- Kelley, R.L., Meller, V.H., Gordadze, P.R., Roman, G., Davis, R.L., and Kuroda, M.I. (1999). Epigenetic spreading of the *Drosophila* dosage compensation

- complex from roX RNA genes into flanking chromatin. *Cell* **98**, 513–522.
- Kelley, R.L., Lee, O.K., and Shim, Y.K. (2008). Transcription rate of noncoding roX1 RNA controls local spreading of the *Drosophila* MSL chromatin remodeling complex. *Mech. Dev.* **125**, 1009–1019.
- Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. *Genome Res.* **12**, 996–1006.
- Kent, W.J., Zweig, A.S., Barber, G., Hinrichs, A.S., and Karolchik, D. (2010). BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204–2207.
- Kharchenko, P. V, Alekseyenko, A.A., Schwartz, Y.B., Minoda, A., Riddle, N.C., Ernst, J., Sabo, P.J., Larschan, E., Gorchakov, A.A., Gu, T., et al. (2011). Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. *Nature* **471**, 480–485.
- Kieffer-Kwon, K.-R., Nimura, K., Rao, S.S.P., Xu, J., Jung, S., Pekowska, A., Dose, M., Stevens, E., Mathe, E., and Dong, P. (2017). Myc regulates chromatin decompaction and nuclear architecture during B cell activation. *Mol. Cell* **67**, 566–578.
- Kolovos, P., Knoch, T.A., Grosveld, F.G., Cook, P.R., and Papantonis, A. (2012). Enhancers and silencers: an integrated and simple model for their function. *Epigenetics Chromatin* **5**, 1.
- Kumar, S., and Bucher, P. (2016). Predicting transcription factor site occupancy

- using DNA sequence intrinsic and cell-type specific chromatin features. *BMC Bioinformatics* 17, 4.
- Kuroda, M.I., Kernan, M.J., Kreber, R., Ganetzky, B., and Baker, B.S. (1991). The maleless protein associates with the X chromosome to regulate dosage compensation in *Drosophila*. *Cell* 66, 935–947.
- Kuzu, G., Kaye, E.G., Chery, J., Siggers, T., Yang, L., Dobson, J.R., Boor, S., Bliss, J., Liu, W., and Jögl, G. (2016). Expansion of GA dinucleotide repeats increases the density of CLAMP binding sites on the X-chromosome to promote *Drosophila* dosage compensation. *PLoS Genet.* 12, e1006120.
- Kyrchanova, O., Kurbidaeva, A., Sabirov, M., Postika, N., Wolle, D., Aoki, T., Maksimenko, O., Mogila, V., Schedl, P., and Georgiev, P. (2018). The bithorax complex *iab-7* Polycomb response element has a novel role in the functioning of the Fab-7 chromatin boundary. *PLoS Genet.* 14, e1007442.
- Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357.
- Larschan, E., Alekseyenko, A.A., Gortchakov, A.A., Peng, S., Li, B., Yang, P., Workman, J.L., Park, P.J., and Kuroda, M.I. (2007). MSL Complex Is Attracted to Genes Marked by H3K36 Trimethylation Using a Sequence-Independent Mechanism. *Mol. Cell* 28, 121–133.
- Larschan, E., Bishop, E.P., Kharchenko, P. V, Core, L.J., Lis, J.T., Park, P.J.,

- and Kuroda, M.I. (2011). X chromosome dosage compensation via enhanced transcriptional elongation in *Drosophila*. *Nature* *471*, 115–118.
- Larschan, E., Soruco, M.M.L., Lee, O.K., Peng, S., Bishop, E., Chery, J., Goebel, K., Feng, J., Park, P.J., and Kuroda, M.I. (2012). Identification of chromatin-associated regulators of MSL complex targeting in *Drosophila* dosage compensation. *PLoS Genet.* *8*.
- Lau, A.C., Nabeshima, K., and Csankovszki, G. (2014). The *C. elegans* dosage compensation complex mediates interphase X chromosome compaction. *Epigenetics Chromatin* *7*, 31.
- Lee, C.-G., and Hurwitz, J. (1993). Human RNA helicase A is homologous to the maleless protein of *Drosophila*. *J. Biol. Chem.* *268*, 16822–16830.
- Leiserson, M.D.M., Blokh, D., Sharan, R., and Raphael, B.J. (2013). Simultaneous identification of multiple driver pathways in cancer. *PLoS Comput. Biol.* *9*, e1003054.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* *25*, 2078–2079.
- Li, J., Liu, Y., Rhee, H.S., Ghosh, S.K.B., Bai, L., Pugh, B.F., and Gilmour, D.S. (2013). Kinetic competition between elongation rate and binding of NELF controls promoter-proximal pausing. *Mol. Cell* *50*, 711–722.
- Li, L., Lyu, X., Hou, C., Takenaka, N., Nguyen, H.Q., Ong, C.-T.T., Cubeñas-Potts, C., Hu, M., Lei, E.P., Bosco, G., et al. (2015). Widespread

- Rearrangement of 3D Chromatin Organization Underlies Polycomb-Mediated Stress-Induced Silencing. *Mol. Cell* 58, 216–231.
- Li, Q., Brown, J.B., Huang, H., and Bickel, P.J. (2011). Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* 5, 1752–1779.
- Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., and Dorschner, M.O. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* (80-.). 326, 289–293.
- Liu, F., Li, H., Ren, C., Bo, X., and Shu, W. (2016). PEDLA: predicting enhancers with a deep learning-based algorithmic framework. *Sci. Rep.* 6, 28517.
- Livak, K.J., and Schmittgen, T.D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. *Methods* 25, 402–408.
- Lott, S.E., Villalta, J.E., Schroth, G.P., Luo, S., Tonkin, L.A., and Eisen, M.B. (2011). Noncanonical compensation of zygotic X transcription in early *Drosophila melanogaster* development revealed through single-embryo RNA-Seq. *PLoS Biol.* 9, e1000590.
- Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550.
- Lucchesi, J.C. (1978). Gene dosage compensation and the evolution of sex chromosomes. *Science* (80-.). 202, 711–716.

- Lucchesi, J.C. (1996). Dosage compensation in *Drosophila* and the 'complex' world of transcriptional regulation. *BioEssays* *18*, 541–547.
- Lun, A.T.L., and Smyth, G.K. (2015). diffHic: a Bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinformatics* *16*, 258.
- Lundberg, S.M., and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pp. 4765–4774.
- Maenner, S., Müller, M., Fröhlich, J., Langer, D., and Becker, P. (2013). ATP-dependent roX RNA remodeling by the Helicase maleless enables specific association of MSL proteins. *Mol. Cell* *51*, 174–184.
- Meller, V.H., and Rattner, B.P. (2002). The roX genes encode redundant male-specific lethal transcripts required for targeting of the MSL complex. *EMBO J.* *21*, 1084–1091.
- Mendjan, S., and Akhtar, A. (2007). The right dose for every sex. *Chromosoma* *116*, 95–106.
- Meyer, B. (2005). X chromosome dosage compensation. *WormBook* 1–14.
- Meyer, C.A., and Liu, X.S. (2014). Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat. Rev. Genet.* *15*, 709.
- Mieczkowski, J., Cook, A., Bowman, S.K., Mueller, B., Alver, B.H., Kundu, S., Deaton, A.M., Urban, J.A., Larschan, E., and Park, P.J. (2016). MNase titration reveals differences between nucleosome occupancy and

- chromatin accessibility. *Nat. Commun.* **7**, 11485.
- Millman, K.J., and Aivazis, M. (2011). Python for scientists and engineers. *Comput. Sci. Eng.* **13**, 9–12.
- Mir, M., Stadler, M.R., Ortiz, S.A., Hannon, C.E., Harrison, M.M., Darzacq, X., and Eisen, M.B. (2018). Dynamic multifactor hubs interact transiently with sites of active transcription in *Drosophila* embryos. *Elife* **7**, e40497.
- Mitchell, R., and Frank, E. (2017). Accelerating the XGBoost algorithm using GPU computing. *PeerJ Comput. Sci.* **3**, e127.
- Mourad, R., Ginalski, K., Legube, G., and Cuvier, O. (2018). Predicting double-strand DNA breaks using epigenome marks or DNA at kilobase resolution. *Genome Biol.* **19**, 34.
- Muller, M., Hagstrom, K., Gyurkovics, H., Pirrotta, V., and Schedl, P. (1999). The mcp element from the *Drosophila melanogaster* bithorax complex mediates long-distance regulatory interactions. *Genetics* **153**, 1333–1356.
- Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59.
- Nègre, N., Brown, C.D., Ma, L., Bristow, C.A., Miller, S.W., Wagner, U., Kheradpour, P., Eaton, M.L., Loriaux, P., and Sealfon, R. (2011). A cis-regulatory map of the *Drosophila* genome. *Nature* **471**, 527.
- Neph, S., Kuehn, M.S., Reynolds, A.P., Haugen, E., Thurman, R.E., Johnson, A.K., Rynes, E., Maurano, M.T., Vierstra, J., and Thomas, S. (2012).

- BEDOPS: high-performance genomic feature operations. *Bioinformatics* 28, 1919–1920.
- Nielsen, D. (2016). Tree Boosting With XGBoost-Why Does XGBoost Win" Every" Machine Learning Competition?
- Noordermeer, D., Leleu, M., Splinter, E., Rougemont, J., De Laat, W., and Duboule, D. (2011). The dynamic architecture of Hox gene clusters. *Science* (80-.). 334, 222–225.
- Nozawa, R.-S., Nagao, K., Igami, K.-T., Shibata, S., Shirai, N., Nozaki, N., Sado, T., Kimura, H., and Obuse, C. (2013). Human inactive X chromosome is compacted through a PRC2-independent SMCHD1-HBiX1 pathway. *Nat. Struct. Mol. Biol.* 20, 566.
- Ohno, S. (1973). Ancient linkage groups and frozen accidents. *Nature* 244, 259.
- Oliphant, T.E. (2006). A guide to NumPy (Trelgol Publishing USA).
- Oliphant, T.E. (2007). Python for scientific computing. *Comput. Sci. Eng.* 9, 10–20.
- Park, S.W., Yool, I.K., Sypula, J.G., Choi, J., Oh, H., and Park, Y. (2007). An evolutionarily conserved domain of roX2 RNA is sufficient for induction of H4-Lys16 acetylation on the *Drosophila* X chromosome. *Genetics* 177, 1429–1437.
- Park, Y., Kelley, R.L., Oh, H., Kuroda, M.I., and Meller, V.H. (2002). Extent of chromatin spreading determined by roX RNA recruitment of MSL proteins. *Science* (80-.). 298, 1620–1623.

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* *12*, 2825–2830.
- Pérez, F., and Granger, B.E. (2007). IPython: a system for interactive scientific computing. *Comput. Sci. Eng.* *9*, 21–29.
- Pferdehirt, R.R., Kruesi, W.S., and Meyer, B.J. (2011). An MLL/COMPASS subunit functions in the *C. elegans* dosage compensation complex to target X chromosomes for transcriptional regulation of gene expression. *Genes Dev.* *25*, 499–515.
- Philip, P., Boija, A., Vaid, R., Churcher, A.M., Meyers, D.J., Cole, P.A., Mannervik, M., and Stenberg, P. (2015). CBP binding outside of promoters and enhancers in *Drosophila melanogaster*. *Epigenetics Chromatin* *8*, 48.
- Prestel, M., Feller, C., Straub, T., Mitlöhner, H., and Becker, P.B. (2010). The activation potential of MOF is constrained for dosage compensation. *Mol. Cell* *38*, 815–826.
- Quinlan, A.R. (2014). BEDTools: the Swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinforma.* *47*, 11–12.
- Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* *26*, 841–842.
- Quinn, J.J., Ilik, I. a, Qu, K., Georgiev, P., Chu, C., Akhtar, A., and Chang, H.Y. (2014). Revealing long noncoding RNA architecture and functions using

- domain-specific chromatin isolation by RNA purification. *Nat. Biotechnol.* **32**, 1–9.
- Rajapakse, I., and Groudine, M. (2011). On emerging nuclear order. *J. Cell Biol.* **192**, 711–721.
- Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, W187–W191.
- Ramírez, F., Lingg, T., Toscano, S., Lam, K.C.C., Georgiev, P., Chung, H.-R., Lajoie, B.R., de Wit, E., Zhan, Y., de Laat, W., et al. (2015). High-Affinity Sites Form an Interaction Network to Facilitate Spreading of the MSL Complex across the X Chromosome in *Drosophila*. *Mol. Cell* **60**, 146–162.
- Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., and Lander, E.S. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680.
- Rattner, B.P., and Meller, V.H. (2004). *Drosophila* Male-Specific Lethal 2 Protein Controls Sex-Specific Expression of the roX Genes. *Genetics* **166**, 1825–1832.
- Raviram, R., Rocha, P.P., Müller, C.L., Miraldi, E.R., Badri, S., Fu, Y., Swanzey, E., Proudhon, C., Snetkova, V., and Bonneau, R. (2016). 4C-ker: a method to reproducibly identify genome-wide interactions captured by 4C-Seq experiments. *PLoS Comput. Biol.* **12**, e1004780.

- Reiter, F., Wienerroither, S., and Stark, A. (2017). Combinatorial function of transcription factors and cofactors. *Curr. Opin. Genet. Dev.* *43*, 73–81.
- Rieder, L.E., Koreski, K.P., Boltz, K.A., Kuzu, G., Urban, J.A., Bowman, S.K., Zeidman, A., Jordan, W.T., Tolstorukov, M.Y., and Marzluff, W.F. (2017). Histone locus regulation by the *Drosophila* dosage compensation adaptor protein CLAMP. *Genes Dev.*
- Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* *26*, 139–140.
- Rotem, A., Ram, O., Shores, N., Sperling, R.A., Goren, A., Weitz, D.A., and Bernstein, B.E. (2015). Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.* *33*, 1165.
- Roy, S., Ernst, J., Kharchenko, P. V, Kheradpour, P., Negre, N., Eaton, M.L., Landolin, J.M., Bristow, C.A., Ma, L., and Lin, M.F. (2010). Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* (80-.). *330*, 1787–1797.
- Salzler, H.R., Tatomer, D.C., Malek, P.Y., McDaniel, S.L., Orlando, A.N., Marzluff, W.F., and Duronio, R.J. (2013). A sequence in the *Drosophila* H3-H4 Promoter triggers histone locus body assembly and biosynthesis of replication-coupled histone mRNAs. *Dev. Cell* *24*, 623–634.
- Schauer, T., Ghavi-Helm, Y., Sexton, T., Albig, C., Regnard, C., Cavalli, G., Furlong, E.E.M., and Becker, P.B. (2017). Chromosome topology guides

- the Drosophila Dosage Compensation Complex for target gene activation. *EMBO Rep.* **18**, 1854–1868.
- Schneider, I. (1972). Cell lines derived from late embryonic stages of *Drosophila melanogaster*. *Development* **27**, 353–365.
- Schwartzman, O., Mukamel, Z., Oded-Elkayam, N., Olivares-Chauvet, P., Lubling, Y., Landan, G., Izraeli, S., and Tanay, A. (2016). UMI-4C for quantitative and targeted chromosomal contact profiling. *Nat. Methods* **13**, 685.
- Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.-J., Vert, J.-P., Heard, E., Dekker, J., and Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259.
- Sexton, T., and Cavalli, G. (2015). The role of chromosome domains in shaping the functional genome. *Cell* **160**, 1049–1059.
- Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F.F.F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell* **148**, 458–472.
- Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L., Yakhini, Z., Weinberger, A., and Segal, E. (2012). Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.* **30**, 521.
- Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., De Wit, E., Van

- Steensel, B., and De Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture—on-chip (4C). *Nat. Genet.* *38*, 1348.
- Skene, P.J., and Henikoff, S. (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife* *6*, e21856.
- Slattery, M., Zhou, T., Yang, L., Machado, A.C.D., Gordân, R., and Rohs, R. (2014). Absence of a simple code: how transcription factors read the genome. *Trends Biochem. Sci.* *39*, 381–399.
- Smith, R.P., Taher, L., Patwardhan, R.P., Kim, M.J., Inoue, F., Shendure, J., Ovcharenko, I., and Ahituv, N. (2013). Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat. Genet.* *45*, 1021.
- Solovei, I., Wang, A.S., Thanisch, K., Schmidt, C.S., Krebs, S., Zwerger, M., Cohen, T. V, Devys, D., Foisner, R., and Peichl, L. (2013). LBR and lamin A/C sequentially tether peripheral heterochromatin and inversely regulate differentiation. *Cell* *152*, 584–598.
- Soruco, M.M.L., Chery, J., Bishop, E.P., Siggers, T., Tolstorukov, M.Y., Leydon, A.R., Sugden, A.U., Goebel, K., Feng, J., Xia, P., et al. (2013). The CLAMP protein links the MSL complex to the X chromosome during *Drosophila* dosage compensation. *Genes Dev.* *27*, 1551–1556.
- Straub, T., and Becker, P.B. (2007). Dosage compensation: the beginning and end of generalization. *Nat. Rev. Genet.* *8*, 47.

- Straub, T., Grimaud, C., Gilfillan, G.D., Mitterweger, A., and Becker, P.B. (2008). The chromosomal high-affinity binding sites for the *Drosophila* dosage compensation complex. *PLoS Genet.* *4*, 1–14.
- Strom, A.R., Emelyanov, A. V, Mir, M., Fyodorov, D. V, Darzacq, X., and Karpen, G.H. (2017). Phase separation drives heterochromatin domain formation. *Nature* *547*, 241.
- Sural, T.H., Peng, S., Li, B., Workman, J.L., Park, P.J., and Kuroda, M.I. (2008). The MSL3 chromodomain directs a key targeting step for dosage compensation of the *Drosophila melanogaster* X chromosome. *Nat. Struct. Mol. Biol.* *15*, 1318–1325.
- Taylor, G., Eskeland, R., and Hekimoglu-balkan, B. (2013). H4K16 acetylation marks active genes and enhancers of embryonic stem cells , but does not alter chromatin compaction H4K16 acetylation marks active genes and enhancers of embryonic stem cells , but does not alter chromatin compaction. 2053–2065.
- Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* *14*, 178–192.
- Tolstorukov, M.Y., Kharchenko, P. V, and Park, P.J. (2010). Analysis of the primary structure of chromatin with next-generation sequencing. *Epigenomics* *2*, 187–197.
- Urban, J., Kuzu, G., Bowman, S., Scruggs, B., Henriques, T., Kingston, R.,

- Adelman, K., Tolstorukov, M., and Larschan, E. (2017a). Enhanced chromatin accessibility of the dosage compensated *Drosophila* male X-chromosome requires the CLAMP zinc finger protein. *PLoS One* *12*, e0186855.
- Urban, J.A., Doherty, C.A., Jordan, W.T., Bliss, J.E., Feng, J., Soruco, M.M., Rieder, L.E., Tsiarli, M.A., and Larschan, E.N. (2017c). The essential *Drosophila* CLAMP protein differentially regulates non-coding roX RNAs in male and females. *Chromosom. Res.* *25*, 101–113.
- Urban, J.A., Urban, J.M., Kuzu, G., and Larschan, E.N. (2017b). The *Drosophila* CLAMP protein associates with diverse proteins on chromatin. *PLoS One* *12*, e0189772.
- Ursu, O., Boley, N., Taranova, M., Wang, Y.X.R., Yardimci, G.G., Noble, W.S., and Kundaje, A. (2018). GenomeDISCO: A concordance score for chromosome conformation capture experiments using random walks on contact map graphs. *Bioinformatics* *1*, 7.
- Veitia, R.A., Veyrunes, F.F., Bottani, S., and Birchler, J.A. (2015). X chromosome inactivation and active X upregulation in therian mammals: Facts, questions, and hypotheses. *J. Mol. Cell Biol.* *7*, 2–11.
- Vian, L., Pękowska, A., Rao, S.S.P., Kieffer-Kwon, K.-R., Jung, S., Baranello, L., Huang, S.-C., El Khattabi, L., Dose, M., and Pruett, N. (2018). The energetics and physiological impact of cohesin extrusion. *Cell* *173*, 1165–1178.

- Villa, R., Schauer, T., Smialowski, P., Straub, T., and Becker, P.B. (2016). PionX sites mark the X chromosome for dosage compensation. *Nature* 537, 244–248.
- Voss, T.C., and Hager, G.L. (2014). Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nat. Rev. Genet.* 15, 69.
- Van Der Walt, S., Colbert, S.C., and Varoquaux, G. (2011). The NumPy array: a structure for efficient numerical computation. *Comput. Sci. Eng.* 13, 22.
- Wang, J., Zhuang, J., Iyer, S., Lin, X., Whitfield, T.W., Greven, M.C., Pierce, B.G., Dong, X., Kundaje, A., and Cheng, Y. (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* 22, 1798–1812.
- Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57.
- Wells, M.B., Csankovszki, G., and Custer, L.M. (2012). Finding a balance: how diverse dosage compensation strategies modify histone h4 to regulate transcription. *Genet. Res. Int.* 2012.
- Van De Werken, H.J.G., Landan, G., Holwerda, S.J.B., Hoichman, M., Klous, P., Chachik, R., Splinter, E., Valdes-Quezada, C., Öz, Y., and Bouwman, B.A.M. (2012). Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nat. Methods* 9, 969.
- Wheeler, B.S., Anderson, E., Frøkjær-Jensen, C., Bian, Q., Jorgensen, E., and Meyer, B.J. (2016). Chromosome-wide mechanisms to decouple gene

- expression from gene dose during seX-chromosome evolution. *Elife* 5, e17365.
- Whitaker, J.W., Chen, Z., and Wang, W. (2014). Predicting the human epigenome from DNA motifs. *Nat. Methods* 12, 265.
- White, M.A., Kwasnieski, J.C., Myers, C.A., Shen, S.Q., Corbo, J.C., and Cohen, B.A. (2016). A simple grammar defines activating and repressing cis-regulatory elements in photoreceptors. *Cell Rep.* 17, 1247–1254.
- Widom, J. (1989). Toward a Unified Model of Chromatin Folding. *Annu. Rev. Biophys. Biophys. Chem.*
- Wolle, D., Cleard, F., Aoki, T., Deshpande, G., Schedl, P., and Karch, F. (2015). Functional requirements for Fab-7 boundary activity in the Bithorax Complex. *Mol. Cell. Biol.* 35, 3739–3752.
- Yang, J., and Corces, V.G. (2012). Insulators, long-range interactions, and genome function. *Curr. Opin. Genet. Dev.* 22, 86–92.
- Zabidi, M.A., and Stark, A. (2016). Regulatory enhancer–core-promoter communication via transcription factors and cofactors. *Trends Genet.* 32, 801–814.
- Zhang, Z., and Pugh, B.F. (2011). High-resolution genome-wide mapping of the primary structure of chromatin. *Cell* 144, 175–186.
- Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., and Li, W. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 9, R137.

- Zheng, S., Villa, R., Wang, J., Feng, Y., Wang, J., Becker, P.B., and Ye, K. (2014). Structural basis of X chromosome DNA recognition by the MSL2 CXC domain during *Drosophila* dosage compensation. *Genes Dev.* *28*, 2652–2662.
- Zhou, Q., Ellison, C.E., Kaiser, V.B., Alekseyenko, A.A., Gorchakov, A.A., and Bachtrög, D. (2013). The epigenome of evolving *Drosophila* neo-sex chromosomes: dosage compensation and heterochromatin formation. *PLoS Biol.* *11*, e1001711.
- Zong, C., Lu, S., Chapman, A.R., and Xie, X.S. (2012). Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* (80-.). *338*, 1622–1626.