

GA-Binding Proteins Regulate Dosage Compensation
in *Drosophila melanogaster*

By

Emily Grace Kaye

B.A., Colby College, Waterville, ME 04901

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 2018

© 2018 Emily Grace Kaye

Signature Page

This dissertation by Emily Grace Kaye is accepted in its present form by the
Molecular Biology, Cell Biology, and Biochemistry Graduate Program as
satisfying the dissertation requirement for the degree of Doctor of Philosophy.

Date _____

Dr. Erica N. Larschan, Advisor

Recommended to the Graduate Council

Date _____

Dr. Richard Freiman, Chair

Date _____

Dr. Nicolas Fawzi, Reader

Date _____

Dr. Kristi Wharton, Reader

Date _____

Dr. Michael Guertin, Reader
University of Virginia

Approved by the Graduate Council

Date _____

Dr. Andrew G. Campbell
Dean of the Graduate School

Curriculum Vitae

Emily G. Kaye

Dept. of Molecular Biology, Cellular Biology, and Biochemistry
Brown University
185 Meeting St, Box G
Providence RI, 02912

Cell Phone: 508-816-5654
e-mail: egkaye99@gmail.com

EDUCATION

Ph.D	Brown University Molecular Biology, Cellular Biology, Biochemistry	August 2017
B.A.	Colby College Chemistry: Biochemistry with Honors	May 2012
	Wayland High School National Honor Society	May 2008

RESEARCH EXPERIENCE

Ph.D Candidate	Brown University Molecular Biology, Cellular Biology, Biochemistry Thesis Research Advisor: Dr. Erica Larschan Thesis Dissertation topic: “GA-Binding Proteins Regulate Dosage Compensation in <i>Drosophila melanogaster</i>”	2012-2017
-------------------	---	-----------

Main projects:

1. We have **revealed an evolutionary mechanism for X-chromosome targeting** (*Kuzu, Kaye et al. 2016*). The Larschan laboratory studies how the master regulator of *Drosophila* dosage compensation, the Male-Specific Lethal (MSL) complex, is recruited to the X-chromosome by the Chromatin-linked adaptor for MSL proteins (CLAMP) transcription factor. We determined that the X-chromosome has rapidly evolved to include expanded GA-repeats that are bound more strongly by CLAMP.
2. In collaboration with Dr. Paul Schedl's laboratory, **we have biochemically determined that CLAMP is part of a GA-targeting insulator complex, the Late Boundary Complex (LBC)** (*Accepted, MCB July, 2017*). We have identified binding of the LBC to a subset of MSL target sites, providing a novel link between insulators and X-chromosome gene regulation.
3. CLAMP is not an X-specific or sex specific protein despite its role in dosage compensation. **To define the context specific functions of CLAMP, we are comparing CLAMP to another GA-motif binding transcription factor, GAGA-factor (GAF).** I am using both *in vitro* and *in vivo* binding competition experiments to uncover their binding interactions at different genomic loci.

Undergraduate Research Assistant	Colby College Chemistry Research Advisor: Dr. Kevin Rice Honors Thesis: “The apoptotic cleavage of poly(ADP-ribose) polymerase (PARP1) by Laromustine and its analogs.”	2009-2012
--	--	-----------

Main project:

We studied the chemotherapeutic mechanism of a prodrug, Laromustine, which is active as two separate subspecies once injected *in vivo*. During my time in the laboratory, I performed drug treatments on mammalian tissue culture cells and collected protein samples to test for apoptotic cell death. **I determined that Laromustine induction of apoptosis is dose-dependent by measuring apoptotic cleavage of Poly-ADP Ribose Polymerase (PARP1).**

ADDITIONAL RESEARCH EXPERIENCE

Graduate Rotation Research	Brown University Completed three rotations in the following laboratories: Dr. Sarah Delaney, Dr. Stephen Helfand, and Dr. Erica Larschan* (*Selected as graduate research laboratory)	2012-2013
----------------------------------	---	-----------

1st Rotation Topic: **“Oxidative DNA damage and its effects on chromatin packaging.”** (Delaney Laboratory)

2nd Rotation Topic: **“Changes in RNAi machinery in aging *Drosophila melanogaster*.”** (Helfand Laboratory)

3rd Rotation Topic: **“Identifying interactions between the CLAMP zinc-fingers and the MSL complex.”** (Larschan Laboratory)

Summer Research	Dana Farber Cancer Institute Cell Biology/Pediatric Oncology Department Laboratory of Dr. David Pellman Assisted postdoctoral researcher Anna Selmecki on her project: “Aneuploidy as an evolutionary mechanism in Yeast.”	Summer, 2011
--------------------	--	--------------

PUBLICATIONS

Published:

*Kuzu, G, ***Kaye, EG**, Chery, J, Siggers, T, Yang, L, Dobson, JR, Boor, S, Bliss, J, Liu, W, Jogl, G, Rohs, R, Singh, ND, Bulyk, ML, Tolstorukov, MY, Larschan, E. “Expansion of GA Dinucleotide Repeats Increases the Density of CLAMP Binding Sites on the X-Chromosome to Promote *Drosophila* Dosage Compensation.” *PLOS Genetics*, July 14, 2016. ***Co-first authors**

Highlighted in: “Sex Chromosomes Do It Differently”, *PLOS Biology*, October 13, 2016.

***Kaye, EG**, *Kurbidaeva, A, Wolle, D, Aoki, T, Schedl, P, Larschan, E. "*Drosophila* dosage compensation loci associate with boundary forming insulator complexes". *Accepted at MCB July, 2017. *Co-first authors*

In Preparation:

Kaye, EG, Booker, M, Kurland, JV, Conicella, A, Fawzi, NL, Bulyk, ML, Tolstorukov, MY, Larschan, EN. "Differential competition between the essential *Drosophila* GAF and CLAMP transcription factors for GA-dinucleotide repeats promotes context-specific gene regulation." *In preparation for submission in September 2017.*

GRANTS and HONORS

Grants:

NSF Travel grant, CSHL Nuclear Organization and Function Meeting	2016
NIH Aging T-32 training grant	2013-2015

Poster Prize:

Brown MCB Graduate Program Retreat (Providence, RI)	August 2015
Award includes student talk at Retreat the following year.	

TEACHING EXPERIENCE

Teaching Assistant:

Brown University	
<i>Intro Genetics</i>	Fall, 2013
Colby College	
<i>Organic Chemistry I (Fall) and II (Spring)</i>	2010-2012
<i>Introductory Chemistry Laboratory</i>	Spring, 2009
<i>Chemistry of Life</i>	January 2009

Tutoring:

Colby College	
Chemistry Science Center Tutor	2011-2012

MENTORING EXPERIENCE

Larschan Laboratory (Brown University undergraduates)

Emma Thygesen	Summer, 2016
Sonia Boor	2013-2015

MCB Graduate program (1st year students are matched with 3rd or 4th year graduate students)

Abigail Brown	2015-2016
Sun Kim	2014-2015

SERVICE and OUTREACH

Brown University MCB Graduate Program:

MCB graduate recruitment weekend volunteer	2013-2017
MCB graduate recruitment student admissions committee member	2016
Student coordinator for the MCB program summer retreat	2014

Publication Service:

Ad hoc reviewer: *PLOS Genetics* [under mentorship]

Colby College:

Chemistry Club President	2011-2012
Relay for Life: Team development and Recruitment subcommittee leader	2011-2012
Publicity subcommittee leader	2010-2011
Top fundraiser	2009

PROFESSIONAL DEVELOPMENT

Research Courses and Workshops:

Computational Genomics Workshop by the Cell Circuits Center
September 22-23, 2014
Broad Institute, Boston MA

Professional Memberships:

American Association for Cancer Research (AACR).	2012-2013
American Association for the Advancement of Science (AAAS)	2017-present

RESEARCH COMMUNICATIONS

Presentations: National and International Conferences

Kaye, EG, Booker, M, Kurland, JV, Bulyk, ML, Tolstorukov, MY, Larschan, EN.
"Understanding GA-repeat targeting by the essential *Drosophila* transcription factors
GAF and CLAMP." Gordon Research Conference: Cancer Genetics and Epigenetics.
Apr. 23-28, 2017. Lucca (Barga), Italy. (POSTER)

Kaye, EG, Kuzu, G, Kurland, JV, Wolle, D, Kurbidaeva, A, Schedl, P, Bulyk, ML,
Tolstorukov, MY, Larschan, EN. "Competition between two essential transcription factors
regulates dosage compensation in *D. melanogaster*." Cold Spring Harbor Laboratory
Nuclear Organization and Function Meeting. May 3-7, 2016, Cold Spring Harbor, NY.
(SPEAKER)

Kaye, EG, Kuzu, G, Kurland, JV, Bulyk, ML, Tolstorukov, MY, Larschan, EN.
"Understanding GA-repeat targeting by the essential *Drosophila* transcription factors
GAF and CLAMP." Chromatin Snapshot: All about polycomb meeting. Oct. 15, 2015,
Boston, MA. (POSTER and POSTER TEASER PRESENTATION).

Kaye, EG, Kuzu, G, Kurland, JV, Bulyk, ML, Tolstorukov, MY, Larschan, EN.
"Competition between the essential transcription factors GAF and CLAMP for GA-
repeats defines their binding patterns in *Drosophila Melanogaster*." Penn State

Chromatin and Epigenetic Regulation of Transcription conference, 34th annual summer symposium in molecular biology. Jul. 21-24, 2015, University Park, PA. (POSTER)

Kaye, EG, Gross, KG, Bellairs, J, Dale, LE, Gallagher, C, Sher, EV, Robinson, KN, and Rice, KP. "Changes in apoptotic gene expression in human leukemia cells treated with the anticancer drug Laromustine." 103rd Annual meeting of the American Association for Cancer Research, Mar. 31-Apr. 4, 2012, Chicago, IL. (POSTER)

Paine, AN, Tepper, JL, Coleman, C, Praggastis, VA, Guerra, RM, **Kaye, EG**, Gallagher, C, and Rice, KP. "Cell death mechanisms associated with the carbamoylating activity of the anticancer prodrug Laromustine." 2010 National IDeA Symposium of Biomedical Research Excellence, Jun. 16-18, 2010, Bethesda, MD. (POSTER)

Presentations: Internal

Kaye, EG. "The X-factor: GA-repeat density on the X-chromosome recruits the CLAMP transcription factor." 28th Annual MCB Retreat, Aug. 31, 2016, Bristol, RI. (SPEAKER)

Kaye, EG. "The X-factor: A battle between two transcription factors for X-chromosome dosage compensation." MCB Graduate Program Recruitment Weekend Data Club. Jan. 29, 2016, Providence, RI. (SPEAKER)

Kaye, EG, Kuzu, G, Kurland, JV, Bulyk, ML, Tolstorukov, MY, Larschan, EN. "Competition between the essential transcription factors GAF and CLAMP for GA-repeats defines their binding patterns in *Drosophila melanogaster*." 27th Annual MCB Retreat, Sept. 2, 2015, Bristol, RI. (POSTER)

Kaye, EG. "CLAMP vs GAF: Transcription factor competition throughout the Genome." Providence Area Aging Research Forum, Mar. 17, 2015, Providence, RI. (SPEAKER)

Kaye, EG. "CLAMP vs GAF: Transcription factor competition throughout the genome." MCB Graduate Program Data Club. Nov. 21, 2014, Providence, RI. (SPEAKER)

Kaye, EG, Larschan, EN. "Defining the role of CLAMP in targeting the MSL complex to the X-chromosome." 26th Annual MCB Retreat, Aug. 27, 2014, Bristol, RI. (POSTER)

Kaye, EG, Chery, J, and Larschan, EN. "Identifying interactions between CLAMP zinc fingers and the MSL complex." 25th Annual MCB Retreat, Aug. 28, 2013, Bristol, RI. (POSTER)

Kaye, EG, Paine, AN, and Rice, KP. "Diagnosing Cellular Suicide: Looking for PARP's Smoking Gun." Colby Undergraduate Summer Research Retreat, Jul. 22, 2010, The Forks, ME. (POSTER)

Acknowledgements

Before I acknowledge the many people who have supported me throughout the evolving stages of my scientific career, I must first thank one very important person who could not be there:

To my father, Steven M. Kaye (1946-2009), I am forever grateful for the time that I had with you, and that I was able to share a hint of what was to come when I told you I was accepted into my first research laboratory position as an undergraduate at Colby. You always encouraged me and provided me with every opportunity possible as a result of your own hard work and success. You continue to be my role model, whether in science or comedic timing, and I hope to honor your memory as I continue in life.

To my mother Carole and my brother Michael, thank you for your support and interest in my science projects. I look forward to more litters of puppies, watching football games, and to new adventures as you embrace farm life.

To Alex, my new family, thank you for helping manage the stress of graduate school. You have been there from the beginning, supporting me in both my new hobbies as well as in science listening to hours of practice talks. I am lucky to have you, and I look forward to sharing my life with you.

To my scientific advisors over the years, thank you for inspiring me, for teaching me, and for allowing me to contribute ideas and new directions to the laboratory.

Kevin: Thank you for welcoming me into the Rice Laboratory and getting me started in research. Two of my favorite memories from Colby College are of attending the Colby undergraduate research retreat and the American Association for Cancer Research national meeting: Both very different, but equally impactful science experiences I will have forever.

Erica: Thank you, thank you, thank you! I'm not sure it is possible to list everything you have done to help me learn and improve on my skills as a scientist in your laboratory. You have offered guidance and support throughout the entire process of graduate school, from the victories, to the hard fought trouble shooting successes, to knowing when to move on from the lost causes, potentially the most valuable skill of all. Finally, thank you for managing the many collaborations that helped to generate some great science!

To my thesis committee, thank you for helping to shape my science and make sure I stayed on focused on the right directions. Rich: Thank you for your constant optimism and your insight during the Scientific Communication grant writing class. Kristi: Thank you for sharing your fly expertise. I have enjoyed being lab neighbors, especially when your dogs visit. Nick: Thank you for allowing me to visit your lab and learn some new skills. And to my newest "committee" member, my outside reader Dr. Mike Guertin: Thank you for taking the time to look over my research and I look forward to your input!

To my graduate school friends, thank you for my new ice coffee addiction, the many hours of cards against humanity, the second floor SFH 250 lunch group with the Gerbi/Larschan/Mowry/Johnson labs (and guest appearances from others as well), and for the memories of adventures to the GCB as first years.

To my non-graduate school friends (and in some cases, non-science friends!), though I do not see you as often I love getting updates from you as we go through major life events.

Finally, to the Larschan lab past and present, I have so many great memories with you both science and non-science. Thank you for keeping me active, with our many 5K races and gym sessions. Thank you for reading drafts of grants, papers, posters, and of course, this final thesis draft, and for inviting me to do the same for you. Most importantly, thank you for embracing the many “meow”s, “rawr”s, and other assorted noises I offer instead of normal people communication.

Table of Contents

Signature Page	iii
Curriculum Vitae	iv
Acknowledgements	ix
Table of Contents	xii
Table of Tables	xv
List of Illustrations	xvi
Chapter 1 : Introduction to Transcription Factor Gene Regulation	
Mechanisms	1
Transcription factor regulation, targeting, and interactions.	2
Regulation of transcription initiation	3
Regulation of pausing, elongation and termination	4
DNA accessibility and transcription	6
Insulators and three-dimensional architecture	9
TF binding to target sequences	12
Dosage compensation: a model for TF targeting and formation of a domain of coordinated gene expression	14
The diverse functions of CLAMP and GAF reveal insight into context-specific transcriptional regulation	21
Chapter 2 : Expansion of GA Dinucleotide Repeats Increases the Density of CLAMP Binding Sites on the X-Chromosome to Promote <i>Drosophila</i>	
Dosage Compensation	24
Abstract	25
Author Summary	26
Introduction	27
Results	30
CLAMP directly binds to a long GA-rich motif that is enriched at in vivo targets	30
CLAMP binding requires DNA elements flanking the 8-bp core part of the motif ...	33
CLAMP binds more efficiently to long GA-dinucleotide repeats in vivo and in vitro	37
The X-chromosome has increased density of GA-repeats and CLAMP occupancy compared to autosomes	42
The CLAMP protein and X-enrichment of its GA-repeat binding sites are conserved across species	45
Discussion	51
Materials and Methods	54
Acknowledgments	64
Supporting Information	65
Supplemental Figures	65
Supplemental Tables	78

Chapter 3 : Differential occupancy of two GA-binding proteins promotes targeting of the <i>Drosophila</i> dosage compensation complex to the male X-chromosome	90
Abstract	91
Introduction	93
Results	98
MSL complex binding sites have high CLAMP occupancy and low GAF occupancy	98
CLAMP and GAF compete for the same binding sites <i>in vitro</i>	105
CLAMP and GAF alter the other's occupancy to promote specific targeting of MSL complex to CES	109
CLAMP reduces GAF occupancy at genes where CLAMP, GAF, and NELF regulate transcription	112
Discussion	115
Materials and Methods	119
Supporting Information	125
 Chapter 4 : <i>Drosophila</i> dosage compensation loci associate with a boundary forming insulator complex.	131
Abstract	132
Keywords	132
Introduction	133
Results	137
CLAMP binds to the <i>Fab-7</i> boundary <i>in vivo</i> and <i>in vitro</i>	137
CLAMP is an integral LBC component	140
The LBC binds to the <i>roX1</i> and <i>roX2</i> CES	142
LBC binding to CES is MRE dependent	147
Knockdowns of CLAMP and GAF reduce MSL3 binding to the <i>roX1</i> , <i>roX2</i> and 5C2 CES	149
CES 11D1: an example of a CES with only a single MRE	151
The LBC binds to topologically linked CES	152
Discussion	154
LBC and boundaries in BX-C	154
LBC and CES	156
Boundaries versus CES	158
Acknowledgements	162
Materials and Methods	162
Supporting Information	166
Supplemental Methods	166
Supplemental Figures	170
Supplemental Tables	176
 Chapter 5 : Discussion of the Function of GAF and CLAMP in Transcriptional Regulation	181
Discussion	182
Different binding specificities for similar GA-binding proteins	182
GAF and CLAMP regulate gene expression through chromatin	185

Differential interaction between CLAMP and GAF promotes context-specific functions of GA-repeat <i>cis</i> -elements	188
The specific three-dimensional organization of the X-chromosome likely promotes targeting by MSL complex.....	191
Summary.....	192
 A. Appendix: Tools for mapping CLAMP and MSL interaction domains ..	194
Abstract:	195
Introduction:	196
Results and Discussion:.....	198
Materials and Methods:.....	204
 References	208

Table of Tables

Table 2.1 EMSA probe sequences.....	59
Table S2.1 The PBM probes. A description of the PBM probe classes and the number of probes in each class.	78
Table S2.2 Kolmogorov–Smirnov test was applied to the PBM intensities of probes. ...	79
Table S2.3 Kolmogorov–Smirnov test was applied to the enrichments on the sequences with different similarities to the motif.	80
Table S2.4 Euclidean distance between motifs.....	80
Table S2.5 Kolmogorov–Smirnov test to compare binding to different flanking sequence categories.....	81
Table S2.6 The count of probes that contain each number of GA-repeats.	81
Table S2.7 Number of GA-repeats in <i>D. melanogaster</i> autosomes (A) and the X-chromosome (X).....	82
Table S2.8 Number of GA repeats in <i>D. melanogaster</i> chromosomes and individual chromosomal arms. Repeats with different lengths may be overlapping.....	83
Table S2.9 Kolmogorov–Smirnov test was applied to the average distance values between S2 CLAMP ChIP-seq peaks.	84
Table S2.10 Kolmogorov–Smirnov test was applied to the average distance values between Kc CLAMP ChIP-seq peaks.	84
Table S2.11 Numbers of S2 and KC ChIP-seq peaks used in the average distance calculation are given.	85
Table S2.12 Kolmogorov–Smirnov test was applied to the normalized distance values between S2 CLAMP ChIP-seq peaks.	85
Table S2.13 Kolmogorov–Smirnov test was applied to the normalized distance values between Kc CLAMP ChIP-seq peaks.	86
Table S2.14 Conservation of CLAMP across species.	86
Table S2.15 Number of matches to the CLAMP PBM motif per Mb on individual chromosomal arms.....	87
Table S2.16 Number of GA repeats in <i>D. miranda</i> chromosomes. Repeats may be overlapping.....	88
Table S2.17 Number of GA repeats in <i>A. gambiae</i> chromosomes. Repeats may be overlapping.....	89
Table S3.1 P-values from student's t-test of PEV data assuming unequal variance....	129
Table S3.2 Aligned reads from each S2 ChIP-seq sample.	129
Table S4.1 DNA sequences used for the EMSA experiments.	176
Table S4.2 Sequences (5' - 3') used to generate dsRNA.....	180
Table S4.3 Sequences (5' – 3') used for ChIP-qPCR targets.	180

List of Illustrations

Figure 1.1. Nucleosome spacing dictates accessibility of genes.	8
Figure 1.2. Overview of diverse insulator functions.	11
Figure 1.3 The MSL complex localizes to specific sites on the X-chromosome.	17
Figure 1.4 GAF interacts with different proteins or protein complexes to perform a broad range of functions.....	18
Figure 1.5 Step-wise model for the role of CLAMP in MSL in dosage compensation....	21
Figure 2.1 CLAMP CLAMP directly binds to a long GA-rich motif that is enriched at <i>in vivo</i> targets.....	31
Figure 2.2 A long 15-bp motif increases CLAMP binding to DNA.....	34
Figure 2.3 Increasing the number of GA-dinucleotide repeats increases CLAMP occupancy.....	39
Figure 2.4 The X-chromosome has more GA-repeats than autosomes and the CES have clustering of CLAMP occupancy.	41
Figure 2.5 CLAMP and enrichment of the CLAMP motif are highly conserved across species.....	46
Figure 2.6 CLAMP binds to similar <i>in vivo</i> X-enriched binding sites in <i>D. miranda</i> and <i>D. melanogaster</i>	48
Figure S2.1	65
Figure S2.2	66
Figure S2.3	67
Figure S2.4	68
Figure S2.5	69
Figure S2.6	70
Figure S2.7	71
Figure S2.8	71
Figure S2.9	72
Figure S2.10	73
Figure S2.11	74
Figure S2.12	75
Figure S2.13	76
Figure S2.14	77
Figure 3.1 GAF and CLAMP open chromatin, a feature of the dosage compensated male X-chromosome, but only CLAMP co-localizes with MSL.	100
Figure 3.2 ChIP-seq determines CLAMP overlaps better with MSL than GAF.....	102
Figure 3.3 GAF and CLAMP zinc-finger proteins bind to similar motifs.	104
Figure 3.4 GAF and CLAMP compete for GA-repeat sequences <i>in vitro</i>	106
Figure 3.5 Effects of RNAi on GAF, CLAMP and MSL localization.....	110
Figure 3.6 Enrichment of GAF and CLAMP at genes sorted by transcriptional changes due to <i>clamp</i> , <i>nelf</i> , and <i>trl</i> RNAi treatments.	114
Figure S3.1	125
Figure S3.2	126
Figure S3.3	127
Figure S3.4	128
Figure S3.5	128
Figure 4.1 CLAMP colocalizes with several boundaries in BX-C.	139
Figure 4.2 LBC binds to <i>Fab-7</i> GAGA3+4, <i>roX1</i> and <i>roX2</i>	141
Figure 4.3 The LBC is a >1000 kDa complex that contains CLAMP, GAF, Mod(mdg4).	143

Figure 4.4 The LBC can bind to <i>roX1</i> and <i>roX2</i> CES.....	144
Figure 4.5 LBC binding is MRE dependent.	146
Figure 4.6 GAF and CLAMP contribute to MSL complex recruitment.....	149
Figure 4.7 The LBC binds to topologically linked CES.....	150
Figure S4.1 GAF and CLAMP bind to the <i>Fab-7</i> boundary <i>in vivo</i>	170
Figure S4.2 CLAMP is part of LBC binding to <i>Fab-7</i> GAGA5+6 sequence.	170
Figure S4.3 CLAMP, GAF and Mod(mdg4) co-fractionate with the LBC.	171
Figure S4.4 Competition experiments with wild type and mutant <i>roX1</i> and <i>roX2</i> DNAs.	172
Figure S4.5 Cross competition of CES 5C2, CES 5C2 mut 1+2+3 and interacting CES.	173
Figure S4.6 Supershift experiments demonstrating that the LBC binds to three additional CES.	174
Figure S4.7 LBC shift of CES CG14446 by Superose 6 column fractions.	175
Figure 5.1 The X-chromosome evolved longer, more numerous GA-rich sequences compared to autosomes.....	186
Figure 5.2 Overview of the function of CLAMP and GAF interactions on the X- chromosome.	189
Figure A.1 Expression and binding of GST-MSL2-CXC fusion protein to glutathione beads.	199
Figure A.2 GST-MSL2-CXC interacts with MBP-CLAMP in a GST pull-down assay. .	200
Figure A.3 CLAMP truncation construct design.	202

Chapter 1 : Introduction to Transcription Factor Gene Regulation Mechanisms

I wrote this introductory chapter with edits and suggestions incorporated from the Larschan Laboratory including Dr. Erica Larschan, Dr. Leila Rieder, Dr. Jennifer Urban, and William Jordan III.

Transcription factor regulation, targeting, and interactions.

Cells function through the concerted effort of three types of macromolecules: DNA, RNA, and proteins. The central dogma states that DNA is converted to RNA via transcription, and RNA is then translated into protein. This progression from DNA to RNA to protein requires many different forms of regulation. One major class of regulatory proteins, called Transcription Factors (TFs), bind directly to DNA to regulate the process of transcription. Transcriptional regulation is critical to normal cellular function, demonstrated by the observation that the majority of genetic mutations leading to cancer encode TFs and epigenetic regulators (Berdasco and Esteller 2010).

TFs can interact directly with the RNA Polymerase transcriptional machinery, other TFs, or the histone proteins that package and organize DNA into chromatin. TFs also recruit co-factors that alter the chromatin accessibility of genes. These co-factors include chromatin-remodeling enzymes that reposition nucleosomes, or proteins that deposit histone modifications to alter how tightly the histone core associates with DNA (Kingston and Narlikar 1999). A subset of TFs can also shape the genome by binding architectural elements called insulators (Kellum and Schedl 1991; Kellum and Schedl 1992). TFs bound to insulators can interact with one another to bring specific DNA sequences closer together for regulation. These mechanisms of TF regulation lead to transcriptional activation or repression. Here, I provide an overview of these mechanisms, followed by a discussion of TF targeting. Finally, I present the system of dosage compensation as an ideal opportunity to study how TF targeting and co-factor recruitment initiates gene regulatory mechanisms at specific genomic locations.

Regulation of transcription initiation

TFs regulate transcription by binding DNA *cis*-elements at different genetic locations, such as promoters and enhancers. TFs can bind to promoters and enhancers separately or link them together through three-dimensional interactions. Promoters are upstream of the transcription start site (TSS), and contain specific sequences such as the TATA box that recruit transcriptional machinery (Patikoglou et al. 1999). Enhancers, *cis*-regulatory DNA elements named for their ability to “enhance” transcriptional activity, are sometimes separated from the promoter by several linear kilobases (kb) along a chromosome (Banerji, Rusconi, and Schaffner 1981). Enhancers often perform temporal or tissue specific regulatory functions, rather than constitutive activity thus providing an additional level of transcriptional regulation (Villar et al. 2015; Farley et al. 2015; Yanez-Cuna et al. 2012). Although not fully characterized, recently work in *Drosophila* has mapped thousands of enhancers using a transcription reporter assay to identify sequences with enhancer activity (Arnold et al. 2013), providing the groundwork for many future investigations.

TFs function as transcriptional activators or repressors, and there are some TFs that have context specific-functions, meaning that they function to activate transcription in certain contexts and repress transcription in other contexts. Often, the effect on transcription of a particular TF is the result of a TF recruiting a co-factor rather than the activity of the TF itself. For example, This co-factor could be another TF, or it could be a protein that does not possess a DNA-binding domain and requires a TF for targeting.

There are several stages of transcription that can be altered by TFs: initiation, pausing, pause-release, elongation, and termination. The first, initiation, is dependent on

recruitment and assembly of the pre-initiation complex. The pre-initiation complex includes RNA polymerase II (RNA Pol II) and a set of General Transcription Factors (GTFs) that are conserved across many eukaryotic species, including yeast, humans, rats, and *Drosophila melanogaster*. TFIIA, TFIIB, TFIID, TFIIIE, TFIIF, and TFIIH (Lue and Kornberg 1987; Matsui et al. 1980; Conaway and Conaway 1990). Together, RNA Pol II, the GTFs, and a large enhancer-binding multi-protein complex referred to as the Mediator complex, are required for transcriptional initiation (Kelleher, Flanagan, and Kornberg 1990; Flanagan et al. 1991). Proteins bound to a promoter or enhancer(s) can influence transcriptional initiation through either promoting or preventing pre-initiation complex recruitment and assembly.

Regulation of pausing, elongation and termination

Once transcription is initiated, TFs further regulate subsequent transcriptional steps, including pausing, transition from pausing into elongation, and termination. Pausing occurs when RNA Pol II is recruited and transcription is initiated, but RNA Pol II pauses after progressing only 30-60 nucleotides into the coding sequence (Rougvie and Lis 1988; Jonkers and Lis 2015). TFs interact with RNA pol II, GTFs, or other associated proteins in the paused state. Several sequence-specific pausing factors have been identified in *Drosophila* and mammals. These include GAGA Factor (GAF) and M1BP in *Drosophila* (Lee et al. 2008; Li et al. 2013; Farkas, Leibovitch, and Elgin 2000), and CCAAT-box binding TF (CTF), SP1, adenovirus E1A, and myoblast determination protein 1 (MYOD1) in mammals (Blau et al. 1996; Krumm, Hickey, and Groudine 1995). These factors can either promote the paused state, or initiate release from pausing by interacting with other co-factors. For example, the pausing factors Negative Elongation Factor (NELF) and 5,6-Dichloro-1- β -D-ribofuranosylbenzimidazole (DRB)-Sensitivity-

Inducing Factor (DSIF) bind and stabilize paused polymerase (Kwak and Lis 2013; Adelman and Lis 2012).

In mammals, pause-release occurs through the activity of Positive Transcription Elongation Factor-b (P-TEFb) kinase, which phosphorylates NELF, DSIF, and the carboxy-terminal domain (CTD) of RNA Pol II at Ser2, allowing productive elongation to occur (Peterlin and Price 2006). Evidence for the critical role of pausing and pause-release in broad transcriptional regulation is that nearly all transcription is blocked when cells are treated with a P-TEFb inhibitor (Rahl et al. 2010; Henriques et al. 2013; Jonkers, Kwak, and Lis 2014).

Similar to initiation, pausing, and pause-release into elongation, termination of transcription requires factors that recognize specific sequences. The “AAUAAA” sequence element was discovered to be associated with the 3′ polyadenylation step of termination (Proudfoot and Brownlee 1974; Fitzgerald and Shenk 1978). Studies in yeast of the cytochrome c gene revealed that this sequence signaling polyadenylation is required for transcription termination when a deletion mutation including the “AAUAAA” site led to extension of transcripts (Zaret and Sherman 1982). After the identification of a sequence element that regulates termination, it was determined that the chromatin structure of the region 3′ of genes, which alters access to the poly(A) signal sequence, requires careful regulation for proper transcriptional termination. For example, the chromatin remodeling Chd1p and imitation switch (ISWI) proteins were identified in a genetic screen in yeast for mutants disrupting termination (Alén et al. 2002). Chromatin remodeling regulates many stages of transcription in addition to termination, as discussed below.

DNA accessibility and transcription

Chromatin accessibility dictates whether or not genes and regulatory elements such as enhancers are available for different steps of transcription. Accessibility is regulated by modifications to the core DNA packaging unit: a nucleosome containing DNA wrapped around histone proteins. These modifications are known as epigenetic marks, and include chemical modifications, such as acetylation, to amino acids in histone protein tails, which change how tightly DNA associates with the histone octamer (Goldberg, Allis, and Bernstein 2007; Sedivy, Banumathy, and Adams 2008). Depending on the modification, this leads to either an increase or decrease in accessibility of DNA to transcriptional machinery, which can regulate different steps of the transcription process. Alterations include: 1) acetylation (Allfrey, Faulkner, and Mirsky 1964), regulated by histone deacetylases and histone acetyl transferases (Brownell and Allis 1995); 2) methylation (Jenuwein et al. 2000), regulated by histone methyltransferases; 3) phosphorylation (Gurley, Walters, and Tobey 1974), regulated by kinases and phosphatases; 4) or ubiquitination (Sun and Allis 2002; H. Wang et al. 2004), deposited by ubiquitin ligases.

In addition to covalent modifications that can loosen or tighten DNA-histone interactions, the spacing of nucleosomes is also critical to gene regulation. For example, in addition to pausing, another role of GAF at promoters is to support a nucleosome-free region to maintain availability of the TSS to the General Transcription Factors and other initiation factors (Fuda et al. 2015). GAF performs this function through recruitment of a complex that is able to move or “remodel” nucleosomes and their position along the DNA. These chromatin-remodeling protein complexes are ATP-dependent and lead to either compaction or opening of chromatin (Kingston and Narlikar 1999). High activation energy is required to reorganize nucleosomes and change them from one conformation

to another. Chromatin-remodeling complexes associate with DNA and nucleosomes to create a low energy intermediate state, lowering the “barrier” to changes (Lorch et al. 1998; Schnitzler, Sif, and Kingston 1998).

Although there is variation among specific complex components, two chromatin remodeling complex families, Switch/sucrose Non-fermentable (SWI/SNF) and ISWI, are conserved between mammals, yeast, and fruit flies. Experiments in yeast studying SWI/SNF activity at the *HO* promoter and the chromatin modifying Snf1p-dependent Spt-Ada-Gcn-5-acetyltransferase (SAGA) complex revealed that, with the exception of pioneer factors, chromatin remodelers act prior to recruitment of TFs and chromatin modifiers that acetylate histones (Cosma, Tanaka, and Nasmyth 1999; Krebs et al. 1999). SWI/SNF complexes can either maintain the original nucleosome components when creating a newly remodeled nucleosome (Lorch et al. 1998; Schnitzler, Sif, and Kingston 1998), or they can physically transfer histones to new template DNA (Lorch, Zhang, and Kornberg 1999; Owen-Hughes and Workman 1996).

The mechanisms of ISWI family remodelers are less conserved than those of SWI/SNF, though one main feature is the ability to promote even spacing of nucleosomes. Nucleosome Remodeling Factor (NURF) complex, the first well-studied ISWI family member, was discovered in *Drosophila* (Toshio Tsukiyama, Becker, and Wu 1994; Toshio Tsukiyama and Wu 1995). GAF and NURF bind to substrate nucleosomes where NURF performs remodeling and GAF then stabilizes the now evenly spaced remodeled nucleosomes. This type of interaction between TFs and chromatin remodeling complexes likely plays a role in broader DNA accessibility regulation. If chromatin is in an intermediate state between open or closed, it may then be up to other activating or repressive factors to promote heterochromatinization versus euchromatinization.

If located in heterochromatin, DNA is so tightly wound around the histone octamer that any underlying regulatory sequences are often hidden, rendering these sequences inaccessible to the necessary regulatory proteins (Figure 1.1A). However a specialized group of factors, referred to as pioneer factors, are capable of binding to packaged DNA to increase accessibility and promote recruitment of other factors (L. A. Cirillo 1998; Lisa Ann Cirillo et al. 2002). These include the Forkhead box A (FoxA) and PU.1 proteins in mammals (Heinz et al. 2010; Barozzi et al. 2014), and Zelda and GAF in *Drosophila* (Foo et al. 2014). After pioneer factors bind, they promote chromatin modifications for transition into an open euchromatin state in which DNA is now accessible to other TFs (Figure 1.1B). Mutations in genes encoding these proteins lead to severe defects, for example *Drosophila zelda* mutants exhibit defective synchronization of mitosis and cellular blastoderm formation (Staudt et al. 2006; Liang et al. 2008), while a null *Foxa2* mutation is lethal in mice due to a lack of a node structure essential for notochord development (Weinstein et al. 1994).

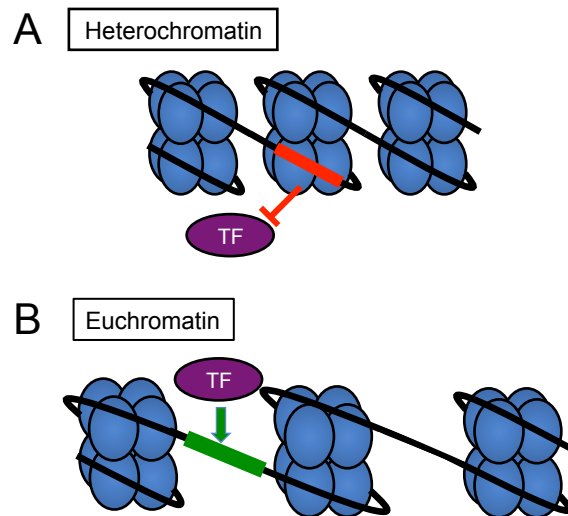


Figure 1.1. Nucleosome spacing dictates accessibility of genes.

A. In heterochromatin, DNA is wrapped tightly around histone octamers to form nucleosomes, rendered inaccessible to most TFs. **B.** Euchromatin is an open DNA conformation, resulting in availability of TF binding sites.

The association of certain histone marks with each other at specific genomic locations has been used to develop multi-state Hidden Markov models that classify the genome into a number of chromatin states (Kharchenko et al. 2011). These states predict accessibility and gene expression. Chromatin states offer more details than the simple classification of the genome into “open”/euchromatin and “closed”/heterochromatin. For example, chromatin states capture positional information relative to genes and specific chromatin marks. Chromatin states have been well characterized in *Drosophila* and are referred to by either a color or number (Kharchenko et al. 2011; Filion et al. 2010). For example, active promoters at TSS are marked by H3K4me3/me2 and H3K9ac and are referred to as “state 1” or the “red state” (Kharchenko et al. 2011). Another kind of active chromatin is enriched specifically on the X-chromosome, and is called the “green state,” or “state 5,” marked by H4K16c and H3K36me3. Alternatively, an example of a repressed chromatin state is the “dark blue state” or “state 7,” which is characterized by pericentromeric heterochromatin and H3K9me2/me3. These states are a useful tool that allows for rapid analysis of new binding patterns relative to existing marks.

Insulators and three-dimensional architecture

In addition to specifically recruiting RNA Pol II and regulating its progression through transcription and performing chromatin regulation, there is another function for TFs in gene regulation. TFs can regulate expression through binding sites known as insulators. Proteins or protein complexes bind to insulator elements and perform at least one of three functions. First, TFs may bind to enhancer-blocking insulators (Kellum and Schedl

1992), which prevent interactions between enhancers and incorrect promoters (Figure 1.2A). Second, TFs may bind to insulators and perform barrier activity (Kellum and Schedl 1991), which prevents silencing heterochromatin marks from spreading into areas of euchromatin where active gene expression is occurring (Figure 1.2B). Finally, TFs bind to insulators and regulate three-dimensional genome organization (Figure 1.2C). In *Drosophila*, this activity has been thoroughly characterized for the *even-skipped* (*eve*) gene locus that forms a ~16 kb loop flanked by the insulator boundary sites *homie* and *nhomie* (Fujioka et al. 2016). This looping activity of chromatin is critical not only for enhancer-promoter interactions, but also for overall genome organization (Dekker and Heard 2015).

Much of what we know about gene regulation has focused on the linear relationship between closely localized elements. However, looping of regulatory domains is an area of increasing interest. In addition to promoter-enhancer looping, larger domains have been identified and are referred to as topologically associated domains (TADs) (Nora et al. 2012; Dixon et al. 2012; Dekker and Heard 2015). Techniques such as Chromosome Conformation Capture (3C) (Dekker et al., 2002), and Circular Chromosome Conformation Capture (4C) (Zhao et al. 2006; Simonis et al. 2006), allow targeted testing of genome-wide interactions that occur between DNA sequences hundreds of kb apart. More recently, the global and high throughput version of Chromosome Conformation Capture, Hi-C (Lieberman-Aiden et al. 2009), has led to the identification of all three-dimensional genome-wide interactions from a single experiment (although final resolution depends on sequencing-depth and fragment size). These technologies have facilitated the identification of regions referred to as TAD boundaries (Nora et al. 2012; Dixon et al. 2012) that separate larger domains from one another.

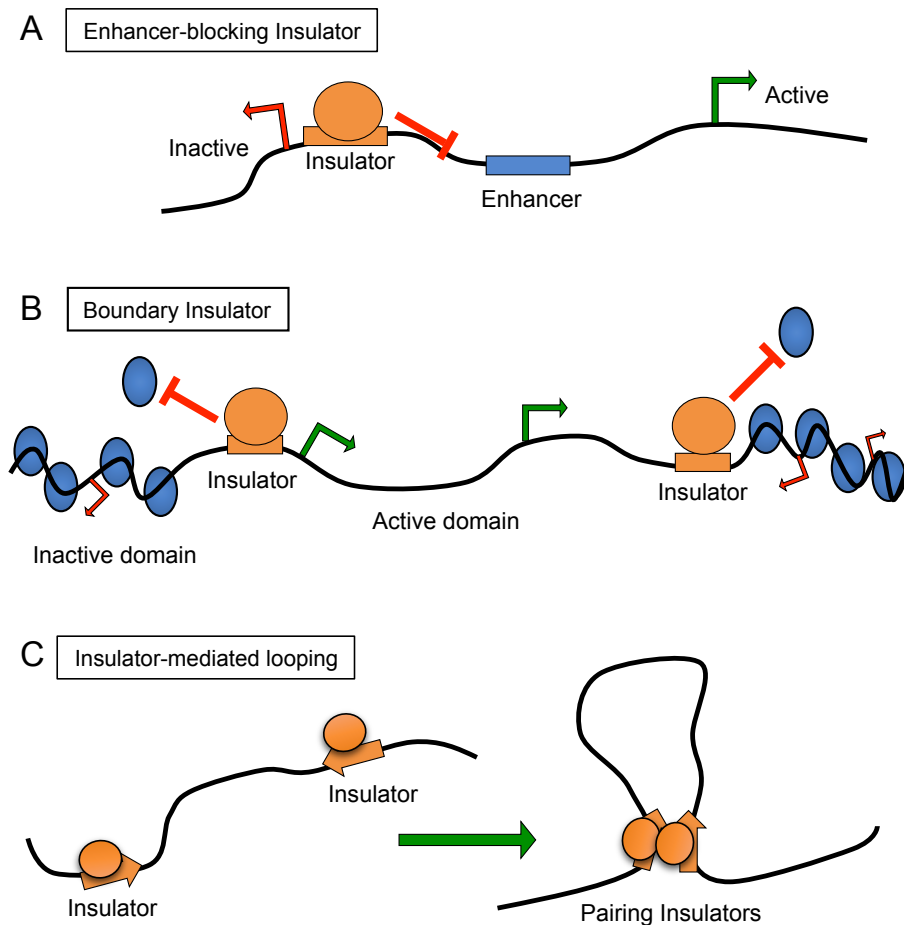


Figure 1.2. Overview of diverse insulator functions.

A. Enhancer-blocking insulators prevent interaction between enhancers and promoters. The gene on the left is inactive (red), as a result of the insulator between the promoter and the enhancer. This does not interfere with the gene on the right (green), which is activated by the enhancer. **B.** Boundary elements are insulators at the border of heterochromatin and euchromatin, preventing the spread of the inactive domain into the active domain. **C.** Insulators regulate three-dimensional organization of DNA through pairing to generate loops. This could incorporate the above categories of insulators through blocking an enhancer from promoters within the loop, and/or insulating an active domain in the loop against surrounding heterochromatin.

Although boundary formation is not yet fully understood, there is much active research in this area. Boundaries are enriched with insulator proteins, such as the conserved CTCF-

binding factor (CTCF) (Lobanenkov et al. 1990; Ghirlando et al. 2012; Dixon et al. 2012). However, *CTCF* RNAi is insufficient for boundary disruption, leading to a model in which the cohesion ring complex drives the formation of TADs through a loop extrusion mechanism (Fudenberg et al. 2016). In this model, the structural cohesion proteins form a ring through which DNA extends into a loop, continuing until a boundary element (such as CTCF) is reached. In *Drosophila*, recent high-resolution (~260 bp) work implicates GA-rich sequences in three-dimensional genome organization as a result of their enrichment at TAD boundaries (Eagen, Aiden, and Kornberg 2017). This suggests a role for the GAF protein as a critical regulator of genome organization, since GAF has also recently been identified in a developmentally critical insulator complex called the Late Boundary Complex (LBC) (Wolle et al. 2015) (see Chapter 4). Other GA-binding proteins are of interest to test for insulator functions, such as the Chromatin-Linked Adaptor for MSL Proteins (CLAMP), described below.

TF binding to target sequences

As discussed above, TFs are versatile regulators of gene expression. They bind nucleosome-bound DNA (pioneer factors) or accessible DNA (other TFs) to directly influence transcription, recruit co-factors (chromatin remodelers), or perform architectural regulation (insulators). TF regulation seems like a straightforward process: bind DNA, recruit co-factor proteins, turn expression on or off, or fine-tune up or down. However, despite all of the progress in the TF field, it is difficult to predict where TFs bind and how they act based only on the identity of the protein and DNA sequence (Siggers and Gordân 2014).

To address the question of TF binding site prediction, one study took an evolutionary approach, testing for optimal binding sites of closely-related TF families using a protein binding microarray (PBM) to measure binding to thousands of sequences (Siggers et al. 2014). This work identified variation within the yeast Msn2 protein family, comparing a shared motif to individual protein-preferred motif sequences, and proposed that small divergences in the amino acid sequence such as an additional modular domain outside of the DNA binding domain, led to binding of a preferred site without altering binding to the shared site (Siggers et al. 2014). In addition to studies aimed at understanding TF binding sites, there have been many studies designed to broadly characterize functional activity resulting from TF binding. For example, a screen of 474 TFs was performed using a UAS-GAL4 targeting luciferase assay to measure activity of TFs at enhancers and categorize them into activators, repressors, and context-specific factors (Stampfel et al. 2015).

A TF and its target sequence are not isolated in the nucleus. There are many other factors present that may interact with the DNA, TF, or both to influence binding interactions. Interactions with other proteins either increases or decreases binding of a factor, and in some cases they can even alter the optimal sequence motif of a TF (Siggers et al. 2011; Slattery et al. 2014). An interesting case has also been described where the *Ciona* body-patterning Orthodenticle homeobox (Otx) gene enhancer evolved sub-optimal binding sites for fibroblast growth factor and the GATA TF as a regulatory mechanism to control tissue specificity (Farley et al. 2015). For some factors, motif orientation dictates the function of binding (e.g. insulator pairing of sites to create looping) (Fujioka et al. 2016), while others bind and perform the same function independent of motif direction. Finally, despite the presence of a motif, a site may not be bound if it is packaged in heterochromatin. Therefore, it is necessary to integrate

multiple types of data such as sequence, genomic location, binding partners, and chromatin state to predict TF binding rather than relying upon sequence alone.

Dosage compensation: a model for TF targeting and formation of a domain of coordinated gene expression

Rarely will a gene or cellular process feature a single type of TF or regulatory activity. An exemplary system is dosage compensation, which integrates several of the above transcription regulatory mechanisms. Dosage compensation evolved to correct an inherent imbalance of gene dosage (Ohno 1967; Lyon 1961). Abnormal gene dosage can be caused by changes in copy number, such as in the cases of chromosomal duplications or loss of chromosomes, referred to as aneuploidy (Gordon, Resio, and Pellman 2012). Aneuploidy is often the result of an abnormal event during meiosis or mitosis. However, a state of natural aneuploidy exists in heterogametic species such as mammals or *Drosophila melanogaster*: XY males have one fewer X-chromosome compared to autosomes and to XX females (C. M. Disteché 2012). This chromosomal gene imbalance is remedied through the essential process of dosage compensation.

During dosage compensation, expression of genes encoded on the X-chromosome(s) is normalized to that of genes encoded on the autosomes (C. M. Disteché 2012). In mammals, dosage compensation is achieved by increasing transcription of most X-linked genes in both sexes, followed by the random inactivation of one of the two female X-chromosomes (Lyon 1961; Nguyen and Disteché 2006). The activation step was initially contested due to difficulty quantifying and comparing chromosome-wide transcription changes from microarray data (Nguyen and Disteché 2006), and an RNA-seq dataset that had the opposite result of the microarray data showing a 1:1 ratio of X-

linked and autosomal gene expression (Xiong et al. 2010). This discrepancy was partially resolved by re-analyzing the RNA-seq dataset, demonstrating that although there is no global up-regulation of X-linked genes, most genes are activated (Deng et al. 2011).

D. melanogaster is an ideal model system in which to study X-chromosome targeting because dosage compensation occurs only in males and, unlike in mammals, there is no additional inactivation step: transcription from the single male X-chromosome is increased approximately twofold, while the female X-chromosomes remain unaffected (Hamada et al. 2005). Interestingly, in *C. elegans*, hermaphroditic XX worms decrease X-linked expression by half to equal that of XO males (Meyer and Casson 1986). Despite the variety of transcription balancing mechanisms, in all of these organisms dosage compensation begins with an X-chromosome identification step. Significant work has been devoted to the inactivation step in mammals, but less is known about the initial targeting step that is conserved across species.

The master regulator of *D. melanogaster* dosage compensation is the conserved Male-Specific Lethal (MSL) complex (J. Belote and Lucchesi 1980). Genes encoding MSL complex are expressed only in males and the complex is targeted to the single male X-chromosome to upregulate transcription of all active genes approximately two-fold (J. Belote and Lucchesi 1980; Larschan et al. 2007a; Larschan et al. 2011). The MSL complex upregulates gene transcription through deposition of the activating H4K16ac chromatin mark, which promotes transcriptional elongation (Larschan et al. 2011). This histone mark is one feature characteristic of the X-chromosome “green,” or “state 5” active chromatin state mentioned above.

The MSL complex is a ribonucleoprotein complex and includes one of the two functionally redundant non-coding RNA components, *RNA on the X 1* and 2 (*roX1* and *roX2*), which are encoded on the X-chromosome and interchangeably associate with the MSL complex (**Figure 1.3**) (Deng et al. 2005; Deng and Meller 2006). While the protein components of MSL complex were identified in genetic screens for male-specific lethality (J. Belote and Lucchesi 1980; J. M. Belote 1983), the *roX* RNAs were identified when mutations led to defects in MSL localization. However, when *roX* genes are placed ectopically on an autosome, MSL targets the new location, spreads to nearby active genes, and still targets the X-chromosome *in trans* (Richard L Kelley et al. 1999; Deng and Meller 2006; Deng and Meller 2009). MSL complex includes two core structural or scaffold-like proteins: Male-Specific Lethal 1 and 2 (MSL1 and MSL2), which have recently been found to dimerize with themselves and each other (Hallacli et al. 2012). The MSL3 complex component contains a chromo-domain, which is characteristic of chromatin remodelers (Paro and Hogness 1991), and allows it to bind to the H3K36me3 chromatin mark associated with active genes (Sural et al. 2008). MSL complex also includes the Maleless (Mle) RNA/DNA helicase (Izzo et al. 2008), and the Males absent On the First (MOF) histone acetyltransferase (Kind et al. 2008).

Many of the MSL complex components are conserved in mammals, although they do not form a complex that regulates dosage compensation. On its own, the MOF human homolog hMOF may play a role in mammalian X-upregulation, though not as part of the assembled MSL complex (Deng et al. 2013a). The mammalian Mle homolog is RNA helicase A, which has broad cellular functions including splicing and transcriptional initiation (Nakajima et al. 1997). Several MSL complex components have been implicated in cancer in mammals, increasing interest in MSL targeting beyond the field of *Drosophila* dosage compensation. For example, the H4K16ac histone modification

deposited by hMOF is depleted in human glioblastoma and other cancers (Pfister et al. 2008; Gelbart et al. 2009). MSL3 is another MSL component commonly mutated in glioblastoma and is also implicated in pediatric gliomas (Leiserson et al. 2013).

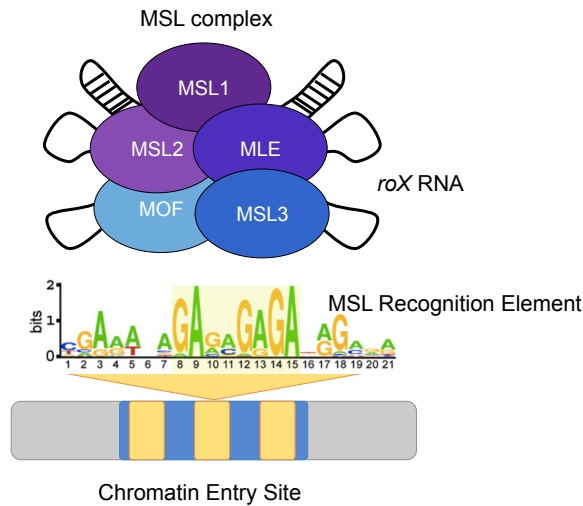


Figure 1.3 The MSL complex localizes to specific sites on the X-chromosome.

MSL complex contains five protein components and one of two interchangeable *roX* RNAs. MSL is targeted to Chromatin Entry Sites (CES), which contain one or more MSL recognition element (MRE) motifs (A. Alekseyenko et al. 2008).

Targeting of the MSL complex in *Drosophila* is proposed to involve a few specific steps. The first step in MSL localization is most likely targeting sequences 3' of the *roX* loci where the *roX* RNAs are thought to be co-transcriptionally incorporated into MSL complex (Kind and Akhtar 2007). Next, MSL complex identifies approximately 150-300 Chromatin Entry Sites (CES), which are 100-bp to ~1.5 kb regions that contain one or more 21-bp GA-rich DNA motifs called MSL Recognition Elements (MREs) (Figure 1.3) (A. Alekseyenko et al. 2008). CES were identified by ChIP-seq experiments as the sites of highest MSL complex occupancy and are also sites that are still targeted by an MSL1/MSL2 sub complex in MSL3 mutant flies (A. Alekseyenko et al. 2008; Straub et al. 2008; Sural et al. 2008). These mutants lacked the ability to perform the final step of

MSL complex targeting: spreading to nearby active genes marked by H3K36me3 (Richard L Kelley et al. 1999), leading to the final binding pattern in which MSL coats active gene bodies along the length of the X-chromosome.

Despite localizing to CES that contain MREs, MSL complex does not directly bind to these motifs with high affinity (Fauth et al. 2010). This suggests that there is an additional co-factor linking the MSL complex to MREs within CES. Given that the targeting of MSL complex features a sequence component, the GA-rich MRE, it was hypothesized that a TF may mediate recruitment of the complex to these sites. The most promising candidate was the well-studied GAF protein, which targets GA-repeat sequences (Farkas et al. 1994). As a protein with many functions, it seemed plausible that it had an additional role as the dosage compensation recruitment factor.

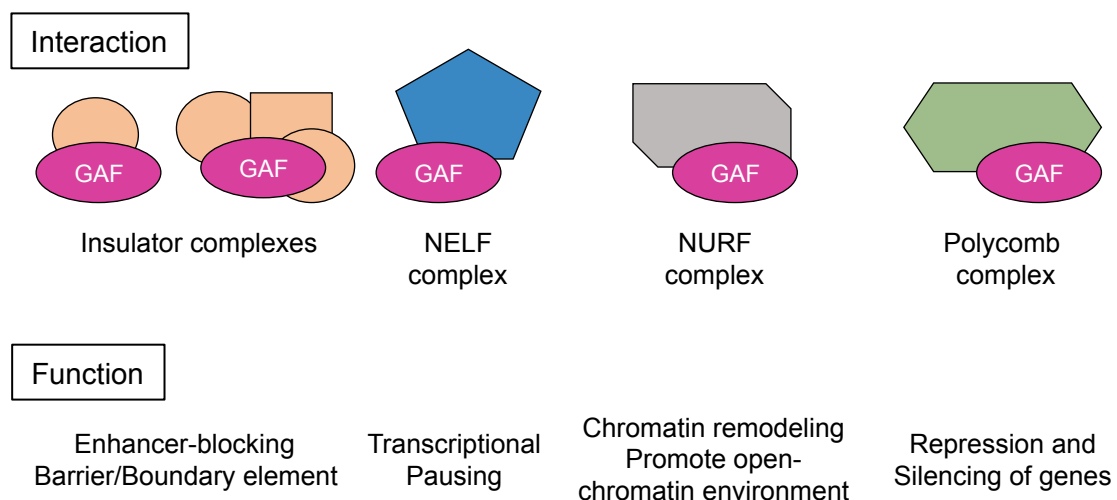


Figure 1.4 GAF interacts with different proteins or protein complexes to perform a broad range of functions.

Overview of the most well-characterized functions of GAF which include its role as an insulator (Ohtsuki and Levine 1998; Wolle et al. 2015), interaction with NELF to promote pausing of RNA Pol II (Lee et al. 2008), Interaction with NURF to regulate chromatin accessibility (Toshio Tsukiyama, Becker, and Wu 1994; Toshio Tsukiyama and Wu 1995), and interaction with the Polycomb repressive complex for gene silencing (Hagstrom, Muller, and Schedl 1997; Strutt, Cavalli, and Paro 1997).

As mentioned above, GAF has a broad range of context-specific roles (Figure 1.4) (Adkins, Hagerman, and Georgel 2006). These include its role in pausing of RNA Pol II during transcription (Lee et al. 2008), and chromatin remodeling through interacting with NURF (Toshio Tsukiyama, Becker, and Wu 1994; Toshio Tsukiyama and Wu 1995). A recent mass-spectrometry approach identified GAF-interacting proteins from chromatin remodeling complexes, including SWI/SNF, ISWI, Polycomb, and Trithorax (Lomaev et al. 2017). GAF has three functional protein domains that mediate these many interactions. An N-terminal Broad-Complex, Tramtrack, Bric a brac/Pox virus and Zinc-finger (BTB/POZ) domain (Zollman et al. 1994; Bardwell and Treisman 1994), a central DNA binding domain containing a C2H2 zinc-finger that binds to the GAGAG consensus sequence (Omichinski and Pedone 1997), and a C-terminal glutamine-rich (Q-rich) domain. This Q-rich domain is the main distinction between the two GAF protein isoforms that are 519 and 581 amino acids in length due to longer polyQ stretches in the 581 protein (Benyajati et al. 1997; Wilkins and Lis 1999).

Many genetic tools were available to test for GAF function in MSL targeting. Specifically, mutant alleles of differing severity in the gene encoding GAF, *trithorax-like (trl)*, were used in combination with each other or with MSL mutants, and male-specific lethality was observed (Greenberg, Yanowitz, and Schedl 2004). Surprisingly, despite this lethal phenotype, when immunofluorescence was used to test MSL localization in *trl* mutants polytene chromosomes, MSL binding was only disrupted at a single site (cytological chromosome location 12DE), but not at the extensive list of other X-linked sites targeted by MSL complex (Greenberg, Yanowitz, and Schedl 2004). Therefore, a new comprehensive RNA interference (RNAi) screen was performed to identify proteins that were required for activity in an MSL-dependent luciferase assay (Larschan et al. 2012). Unlike previous studies, which had fully profiled proteins leading to male-specific lethality

(J. Belote and Lucchesi 1980; J. M. Belote 1983), this screen tested non-sex specific proteins in hopes of identifying a new factor. From this screen, the TF Chromatin-Linked Adaptor for MSL Proteins (CLAMP, previously CG1832) was the most promising candidate. CLAMP has two predicted domain features: a C-terminal C2H2 zinc-finger domain containing six zinc-fingers, which are able to recognize and bind specific DNA sequences, and an N-terminal/central glutamine-rich domain, which is predicted to mediate protein-protein interactions (Michelitsch and Weissman 2000; Atanesyan et al. 2012). Concurrently, an MSL complex ChIP mass spectrometry (ChIP-MS) experiment was performed in *Drosophila* cell culture to identify proteins interacting with chromatin associated MSL complex proteins (C. I. Wang et al. 2013). CLAMP was identified as an interacting factor, though the mechanism of this interaction remains unknown.

Subsequently, the role of CLAMP in MSL recruitment was investigated using RNAi in cell culture and in the fly, as well as using mutant alleles. First, the relationship between MSL and CLAMP was tested via RNAi followed by ChIP-seq to identify sites that were dependent on each other for binding (Soruco et al. 2013). This led to the sorting of CES into three groups based on the effect of *msl2* RNAi on CLAMP binding: those that are dependent on MSL (Group A), those that are partially dependent (Group B) and those that are independent (Group C). At group A and B sites, there is synergistic binding between CLAMP and MSL complex, suggesting a new model for MSL recruitment by CLAMP (Figure 1.5). In this model, CLAMP is present at CES from group B and C (Figure 1.5A), and after MSL complex assembly and recruitment (Figure 1.5B and C), CLAMP and MSL function synergistically to increase each other's occupancy at groups A and B CES. In the last phase of dosage compensation establishment, MSL spreads to nearby active genes to generate its final localization pattern (Figure 1.5D) (Richard L Kelley et al. 1999). To follow up on cell culture studies that led to this model, a protein

null mutant for *clamp* was generated using the CRISPR/Cas9 gene editing system to induce a frame shift, resulting in early termination (J. A. Urban, Doherty, et al. 2017). This mutation, called *clamp2*, results in homozygous lethality in both males and females, indicating that CLAMP is required beyond male dosage compensation.

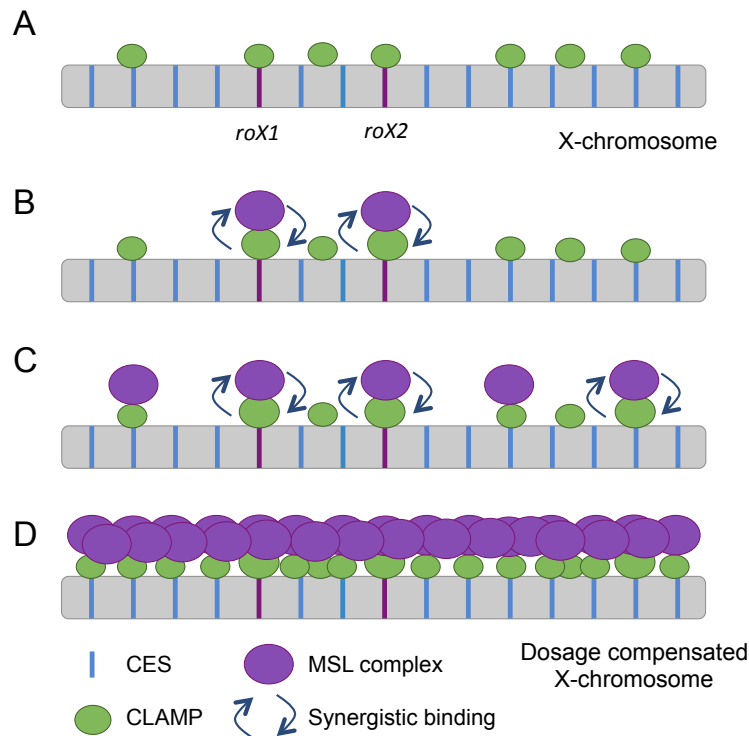


Figure 1.5 Step-wise model for the role of CLAMP in MSL in dosage compensation.

(Adapted from (Soruco et al. 2013)). **A.** CLAMP binds to GA-rich sequences located in CES on the X-chromosome. **B.** MSL complex is assembled at the *roX* RNA loci. MSL and CLAMP function to promote each other's occupancy synergistically. **C.** MSL is recruited by CLAMP to additional CES. At some sites they act in synergy, at others CLAMP binding is independent of MSL interaction. **D.** MSL is recruited to all CES where it co-localizes with CLAMP, and then spreads active genes in a sequence-independent manner, leading to its final binding pattern coating the dosage compensated X-chromosome.

The diverse functions of CLAMP and GAF reveal insight into context-specific transcriptional regulation

Little is understood about how similar TFs perform different context-specific roles and studying the relationship between CLAMP and GAF will advance our understanding of

TF function. GAF and CLAMP share similar binding site motifs (Soruco et al. 2013; Omichinski and Pedone 1997; Lee et al. 2008), yet have different overall binding: CLAMP binds to CES, while GAF only weakly binds CES (Soruco et al. 2013). Although the two proteins have some similarities in domain structure, the central GAF DNA binding domain contains only a single C2H2 zinc finger compared to the six zinc-finger CLAMP DNA binding domain, and the GAF C-terminal Q-rich domain contains polyQ sequences much longer than the Q-rich regions in CLAMP (Benyajati et al. 1997; Wilkins and Lis 1999). Due to the similarity between CLAMP and GAF binding site motifs and their similar domain structure, it seemed likely that CLAMP also has functional roles in non-sex-specific genome-wide transcription regulation, similar to GAF. Preliminary results after *clamp* RNAi from both genome-wide analysis of promoter transcription (Start-seq) and two-dimensional chromatin structure (nucleosome profiles) measured using MNase titrations revealed that CLAMP has at least two context specific roles: CLAMP activates genes and opens chromatin on the male X-chromosome and activates and represses gene expression and closes chromatin on autosomes (J. A. Urban, Kuzu, et al. 2017).

In the three subsequent chapters I investigate the mechanism of differential CLAMP and GAF occupancy in regulation of transcription. In Chapter 2, I thoroughly analyze the CLAMP binding motif using protein binding microarrays and its conservation using ChIP-seq, revealing an enrichment of GA-repeats on the X-chromosome that occurred over evolutionary time leading to enhanced CLAMP occupancy across *Diptera*. In Chapter 3, I investigate the relationship between GAF and CLAMP using competition electrophoretic mobility shift assays and an *in vivo* RNAi approach to test factor relocalization and identify sites of exclusive and interdependent binding. The relative occupancy of GAF and CLAMP differs between CES, where CLAMP frequently

outcompetes GAF, and promoters where both factors often co-occupy similar genomic loci. I determined that after reduction of GAF, MSL complex is mis-localized away from CES, consistent with and potentially explaining previous observations of synthetic male-specific lethal GAF (*Trl*) mutant alleles (Greenberg, Yanowitz, and Schedl 2004).

In Chapter 4, I hypothesize that the common function of CLAMP and GAF in promoting MSL recruitment identified in Chapter 3 may be due to their participation in the same protein complex. CES, which contain many CLAMP binding sites on the X-chromosome, are found at the boundaries between TADs (Ramírez et al. 2015). I therefore asked whether CLAMP may be part of an insulator complex called the Late Boundary Complex that targets GA-rich sequences and is already known to contain GAF (Wolle et al. 2015). I used biochemical approaches to determine that CLAMP is a member of the LBC. The role of insulators in dosage compensation will be an exciting area of future discovery. Overall, my data support a role for the combination of competition for binding to GA-rich sequences, regulation of chromatin accessibility, and co-factor specificity in determining binding patterns and transcriptional regulation by GAF and CLAMP. This result is consistent with the idea introduced above that multiple types of analysis should be integrated to reveal the diverse TF targeting mechanisms in transcription regulation.

Chapter 2 : Expansion of GA Dinucleotide Repeats Increases the Density of CLAMP Binding Sites on the X- Chromosome to Promote *Drosophila* Dosage Compensation

This chapter has been published and is reproduced here with minor formatting changes.

Guray Kuzu*, **Emily G. Kaye***, Jessica Chery, Trevor Siggers, Lin Yang, Jason R. Dobson, Sonia Boor, Jacob Bliss, Wei Liu, Gerwald Jogl, Remo Rohs, Nadia D. Singh, Martha L. Bulyk, Michael Y. Tolstorukov, and Erica Larschan. *PLOS Genetics*. July; 12(7):1-28, 2016. (*Co-first authors)

EL, EGK, GK, RR, MLB, MYT, and GJ conceived and designed the experiments. EGK, JC, TS, SB, JB, and WL performed experiments. GK, TS, LY, NDS, JRD performed data analysis and generated figures. EL, EGK, GK, and MYT wrote the manuscript with input and approval from other co-authors.

Abstract

Dosage compensation is an essential process that equalizes transcript levels of X-linked genes between sexes by forming a domain of coordinated gene expression. Throughout the evolution of *Diptera*, many different X-chromosomes acquired the ability to be dosage compensated. Once each newly evolved X-chromosome is targeted for dosage compensation in XY males, its active genes are upregulated two-fold to equalize gene expression with XX females. In *Drosophila melanogaster*, the CLAMP zinc finger protein links the dosage compensation complex to the X-chromosome. However, the mechanism for X-chromosome identification has remained unknown. Here, we combine biochemical, genomic and evolutionary approaches to reveal that expansion of GA-dinucleotide repeats likely accumulated on the X-chromosome over evolutionary time to increase the density of CLAMP binding sites, thereby driving the evolution of dosage compensation. Overall, we present new insight into how subtle changes in genomic architecture, such as expansions of a simple sequence repeat, promote the evolution of coordinated gene expression.

Author Summary

From stem cells to neurons, regulation of gene dosage is essential in all tissues and species in which it has been studied. Gene dosage must be balanced largely because it is critical to maintain the stoichiometry of components of multi-protein complexes that are encoded at diverse locations throughout the genome. The X-chromosome in many heterogametic species is a natural case where there is a seeming imbalance in gene dosage between sexes. How do cells correct imbalances in gene dosage? The key first step in dosage compensation across species is distinguishing the X-chromosome from the autosomes. Here we report a new mechanism by which newly evolved X-chromosomes become dosage compensated in *Diptera*: the expansion of simple GA dinucleotide repeats increases the density of the conserved CLAMP zinc finger protein that recruits the dosage compensation complex to the X-chromosome. Expansion of dinucleotide repeats can easily occur via slippage of DNA polymerase providing a simple mechanism for nucleating a domain of coordinated gene expression.

Introduction

Changes in primary DNA sequence that occur over evolutionary time alter transcription factor occupancy on DNA (Badis et al. 2009; Paris et al. 2013). Differential occupancy of transcription factors throughout the genome controls the essential gene regulatory programs that define growth and development (Harrison and Eisen 2015). Sex chromosome dosage compensation is a key model system with which to study this essential process because a large number of genes on a single chromosome are co-regulated to form a domain of coordinated gene expression (McElroy, Kang, and Kuroda 2014; Ferrari et al. 2014). However, little is understood about the evolutionary mechanisms that drive the differentiation of the sex chromosomes to ensure the specificity of this new domain.

Recent work demonstrated that the same mechanism of dosage compensation evolved independently across *Dipterans*, although the diverged sex chromosomes are not all derived from the same ancient chromosome (Vicoso and Bachtrog 2015). This conserved mechanism increases the transcript levels of all active genes along the length of the male X-chromosome two-fold to equalize gene expression between males (XY) and females (XX) (Vicoso and Bachtrog 2015; J. M. Belote 1983). Because many different chromosomes evolved the same mechanism independently, we hypothesized that the ability of any *cis*-acting DNA sequences involved in this process to be easily generated is critical. For the past thirty years, it has been known that dinucleotide repeats are enriched on the *Drosophila* X-chromosome (Lowenhaupt, Rich, and Pardue 1989). It was hypothesized that these repeats promote targeting of the dosage compensation machinery to the X-chromosome (Lowenhaupt, Rich, and Pardue 1989), yet the mechanism remained unknown.

In *Drosophila melanogaster*, the best studied *Dipteran* species, dosage compensation is mediated by the MSL (Male-Specific Lethal) complex that deposits H4K16ac, which is likely to increase transcript levels by opening chromatin to allow more rapid progression of RNA Polymerase II through gene bodies (J. M. Belote 1983; Shogren-Knaak et al. 2006; Larschan et al. 2011; Ferrari et al. 2013). The MSL complex includes five protein components and two non-coding RNAs (Lucchesi and Kuroda 2015), and is recruited to X-linked Chromatin Entry Sites (CES) by 21-bp GA-rich DNA sequence elements called MREs (MSL Recognition Elements) (A. Alekseyenko et al. 2008). MREs are 1.8 fold enriched on the X-chromosome compared with autosomes in both *D. melanogaster* and the distantly related *D. miranda*. Moreover, MREs have been acquired on the newly evolved *D. miranda* neo X-chromosome that is only 1 million years old compared with the ancient X-chromosomes that are 30 million years old (A. A. Alekseyenko et al. 2013). The enrichment of MREs on the X-chromosome over evolutionary time likely occurred via a combination of transposon insertions, gene conversion, and errors in DNA replication (Christopher E Ellison and Bachtrog 2015; Q. Zhou et al. 2013; C. E. Ellison and Bachtrog 2013).

Interestingly, the 21-bp MREs are much longer than most transcription factor binding sites that are on average 6 to 8-bp in length (Hume et al. 2015). Not only do MREs have an 8-bp highly conserved core sequence, but they also contain more degenerate flanking sequence outside of the core motif that is required for MSL complex recruitment (A. Alekseyenko et al. 2008). Canonical MSL complex components do not directly interact with the MRE in a sequence specific manner, other than a low affinity interaction between the MSL2 protein and a single cytosine within the MRE (Zheng et al. 2014;

Fauth et al. 2010). Therefore, the mechanism by which the complete 21-bp MRE motif promotes MSL recruitment to the X-chromosome remained poorly understood.

We recently demonstrated that the CLAMP C2H2 zinc finger protein directly recognizes MRE sequences and is required for MSL complex recruitment, thereby providing the first link between MSL complex and the X-chromosome (Soruco et al. 2013). When CLAMP and MSL complex are colocalized, the occupancy of both factors is increased likely due to: 1) the deposition of the H4K16ac mark by MSL complex that opens chromatin to allow CLAMP to identify its binding sites more efficiently (Shogren-Knaak et al. 2006; Soruco et al. 2013) and 2) physical association between CLAMP and MSL complex (C. I. Wang et al. 2013). Moreover, the CLAMP protein occupies MREs on both the X-chromosome and autosomes and is enriched on the X-chromosome even in the absence of MSL complex (Soruco et al. 2013). Therefore, it is key to understand how the binding sites for CLAMP became enriched on the X-chromosome to promote specific X-identification for dosage compensation.

Here, we integrate biochemical, *in vivo*, and evolutionary approaches to provide new insight into how the X-chromosome likely evolved as a domain of coordinated gene regulation. First, CLAMP occupancy increases as the number of GA-repeats increases, a feature that provides a possible mechanism for easily generating new high affinity CLAMP binding sites. Second, the overall density of CLAMP occupancy is highest within the MSL complex CES, and the location of CLAMP binding sites relative to genes differs on the X-chromosome compared to autosomes, increasing X-specificity. Third, CLAMP and its DNA binding sequence are enriched on the X-chromosome across several distantly related species. Therefore, we provide support for a mechanism by which

expansion of simple sequence repeats over evolutionary time increases the density of CLAMP within CES to drive the X-specificity of dosage compensation.

Results

CLAMP directly binds to a long GA-rich motif that is enriched at *in vivo* targets

Previously, we determined that the sequences flanking the 8-bp core of the 21-bp MRE are necessary for MSL recruitment (A. Alekseyenko et al. 2008) (Figure 2.1A: I). It was surprising that such an unusually long motif was required for MSL complex recruitment. Similarly, the CLAMP motif determined by our previous ChIP-seq analysis features an 8-bp core that is part of a longer GA-rich motif that is similar to the MRE (Soruco et al. 2013) (Figure 2.1A: II). Therefore, we hypothesized that bases flanking the 8-bp core contribute to MSL recruitment because they are required for CLAMP binding, which then tethers the MSL complex to the X- chromosome.

To determine the requirements for CLAMP binding directly to DNA, we compared *in vivo* and *in vitro* binding properties. We previously used a universal Protein Binding Microarray (uPBM) that contains all possible combinations of 10-mer sequences and additional flanking bases within a 36-bp variable probe sequence (Soruco et al. 2013). The uPBMs identified only a short 8-bp motif (Figure 2.1A:III) because they were not powered to determine the role of flanking DNA sequences due to insufficient coverage of longer binding sequences (Soruco et al. 2013; Berger and Bulyk 2006). To better recapitulate *in vivo* binding, we therefore designed a custom genomic-context PBM (gcPBM) experiment (Gordân et al. 2013) using sequences derived from CLAMP ChIP-seq binding sites (ChIP+) and control sequences not bound by CLAMP *in vivo* (ChIP-) (Table S2.1). We expressed a GST-CLAMP fusion protein containing the predicted 6

tandem C-terminal zinc fingers of CLAMP for these analyses (a.a. 350-561) using *in vitro* transcription and translation as in our previous studies (Sorucu et al. 2013).

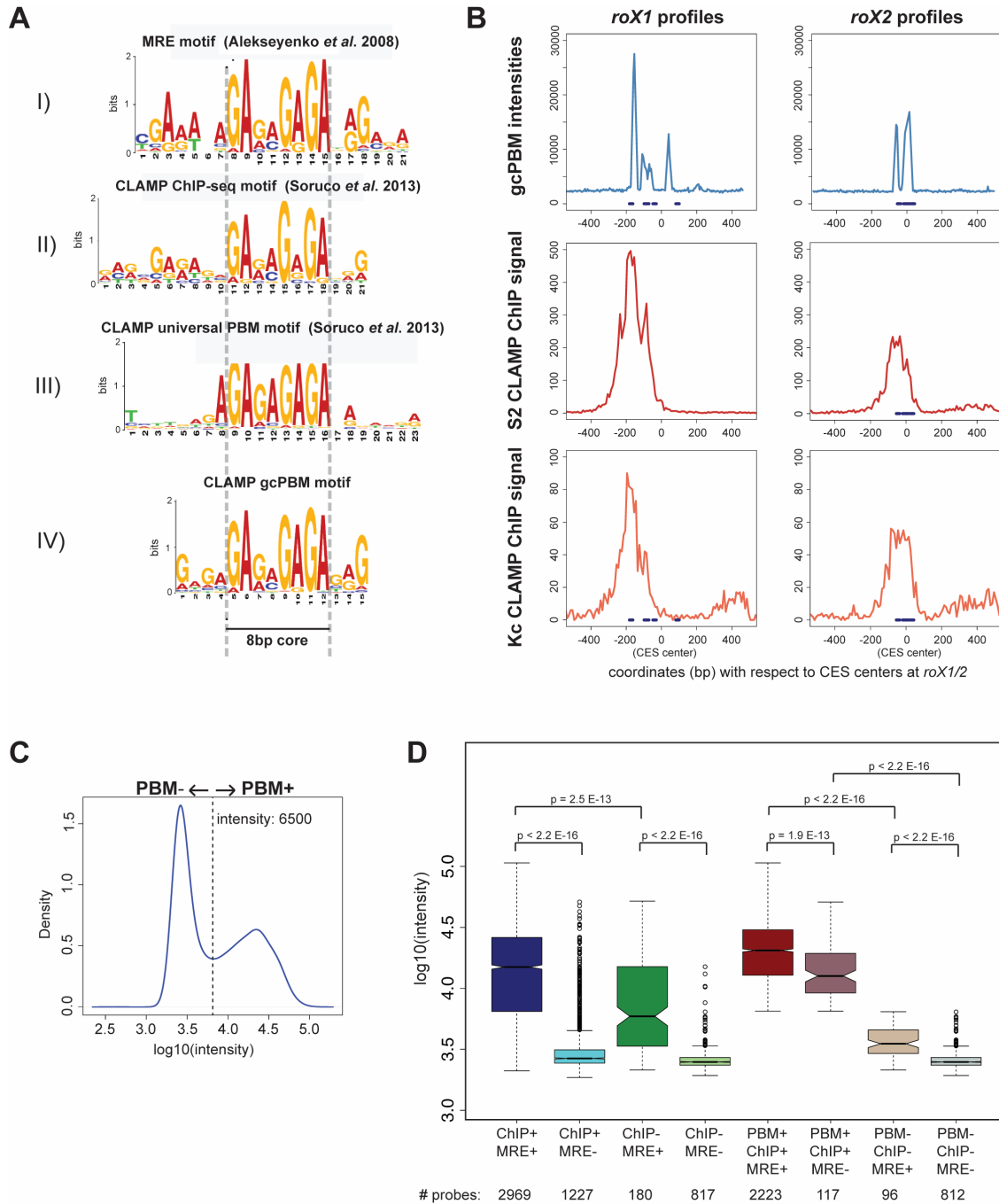


Figure 2.1 CLAMP CLAMP directly binds to a long GA-rich motif that is enriched at *in vivo* targets.

A) The following motifs are shown: I) MRE motif (A. Alekseyenko et al. 2008); II) CLAMP *in vivo* motif derived from ChIP-seq data from S2 cells (Sorucu et al. 2013); III) CLAMP *in vitro* motif derived from the uPBM (Sorucu et al. 2013); and IV) CLAMP *in vitro* motif that we derived from the custom gcPBM. The most conserved part of the motifs, their 8-bp core, is highlighted between

two dashed lines. **B)** gcPBM intensities (intensity from genomic context PBM) (top) and CLAMP ChIP signal (input normalized RPKM: Reads per Kb per Million) for S2 (middle) and Kc (bottom) cells are plotted for 800-bp windows centered at each of two MSL complex CES (*roX1*, left, and *roX2*, right). MRE sequences are shown as blue dashes on each profile. **C)** A histogram of intensities from the gcPBM experiment. The intensity of 6500, between the two peaks of the bimodal distribution, is indicated with a dashed line. Probes with intensities higher than 6500 are designated PBM+ and those that are lower than 6500 are considered PBM-. **D)** Box plots are used to compare the intensities for probes sorted by whether they contain sequences that are bound or unbound from *in vivo* ChIP-seq data (ChIP +/-) and whether or not they conform to the previously characterized MRE motif (MRE +/-). In the right half of the panel, probes were resorted to add the category of bound or unbound on the PBM (PBM +/-). p-values for comparisons between categories are displayed above the relevant boxes, and the number of probes for each category are listed for each group. For each box plot used throughout this study, the collared box indicates that 95% percent confidence interval. If the notches at the center of the box are not overlapping, it indicates that two samples are statistically different from each other.

To validate our gcPBM approach, we included control sequences on the array to compare with *in vivo* ChIP-seq binding profiles. First, we plotted the *in vitro* binding of CLAMP over two tiled regions which surround strong binding sites for MSL complex (*roX1* and *roX2*) (Figure 2.1B: top). We observed specific interaction with most sequences containing previously identified MREs and the gcPBM binding profiles were similar to those from previously generated *in vivo* ChIP-seq data (Figure 2.1B: middle and bottom). In order to extend our analysis beyond two specific locations in the genome, we next defined the CLAMP bound sequences on the gcPBM (PBM+) as those with intensities greater than 6500 based on finding the local minimum in a histogram of all of the intensities (Figure 2.1C). Motifs were generated by MEME analysis from the bound sequences (Figure 2.1A: IV). Using the gcPBM, we captured additional flanking sequences outside the 8-bp core region compared with the uPBM (Figure 2.1A: IV vs. III). Therefore, we demonstrated that the binding of CLAMP derived from custom gcPBMs largely matches previously observed *in vivo* data, validating our approach.

Next, we compared *in vitro* CLAMP binding to sequences that are bound (ChIP+) or not bound (ChIP-) *in vivo*. *In vitro* binding of CLAMP is significantly increased for DNA sequences that are bound *in vivo* (ChIP+ MRE+, dark blue and dark red boxes)

compared to those that are not bound *in vivo* (ChIP- MRE+, dark green and tan boxes) even if both DNA sequences match the MRE that we previously derived from high affinity MSL complex binding sites (A. Alekseyenko et al. 2008) (Figure 2.1D, full set of p-values are reported in Table S2.2). The presence of a sequence that matches the MRE motif further increases CLAMP binding intensity compared to those sequences without an MRE motif even if both are bound *in vivo* (Figure 2.1D: ChIP + MRE+ (dark blue box) vs. ChIP+ MRE- (light blue box)). MREs promote CLAMP binding for both PBM+ and PBM- sites (Figure 2.1D: right). In conclusion, these data suggest that sequences that CLAMP binds more strongly *in vitro* are more likely to be bound *in vivo*.

CLAMP binding requires DNA elements flanking the 8-bp core part of the motif

To determine which parts of the CLAMP motif are most critical for CLAMP binding, we used several different approaches. First, we used MEME to derive motifs from two different classes of sites: 1) PBM+ ChIP+ (**bound** *in vivo* and *in vitro*, Figure 2.2A: I); 2) PBM- ChIP- (**unbound** *in vivo* and *in vitro*, Figure 2.2A: II) (See Figure S2.1A for motifs from additional classes of sites). As expected, we noted overall similarities in the core of the motif because many of our probes contain the short core motif identified from our original uPBMs to allow detailed analysis of the flanking regions (see Methods). We found that DNA elements that are not bound have very little sequence flanking the core 8-bp of the motif compared with bound sites.

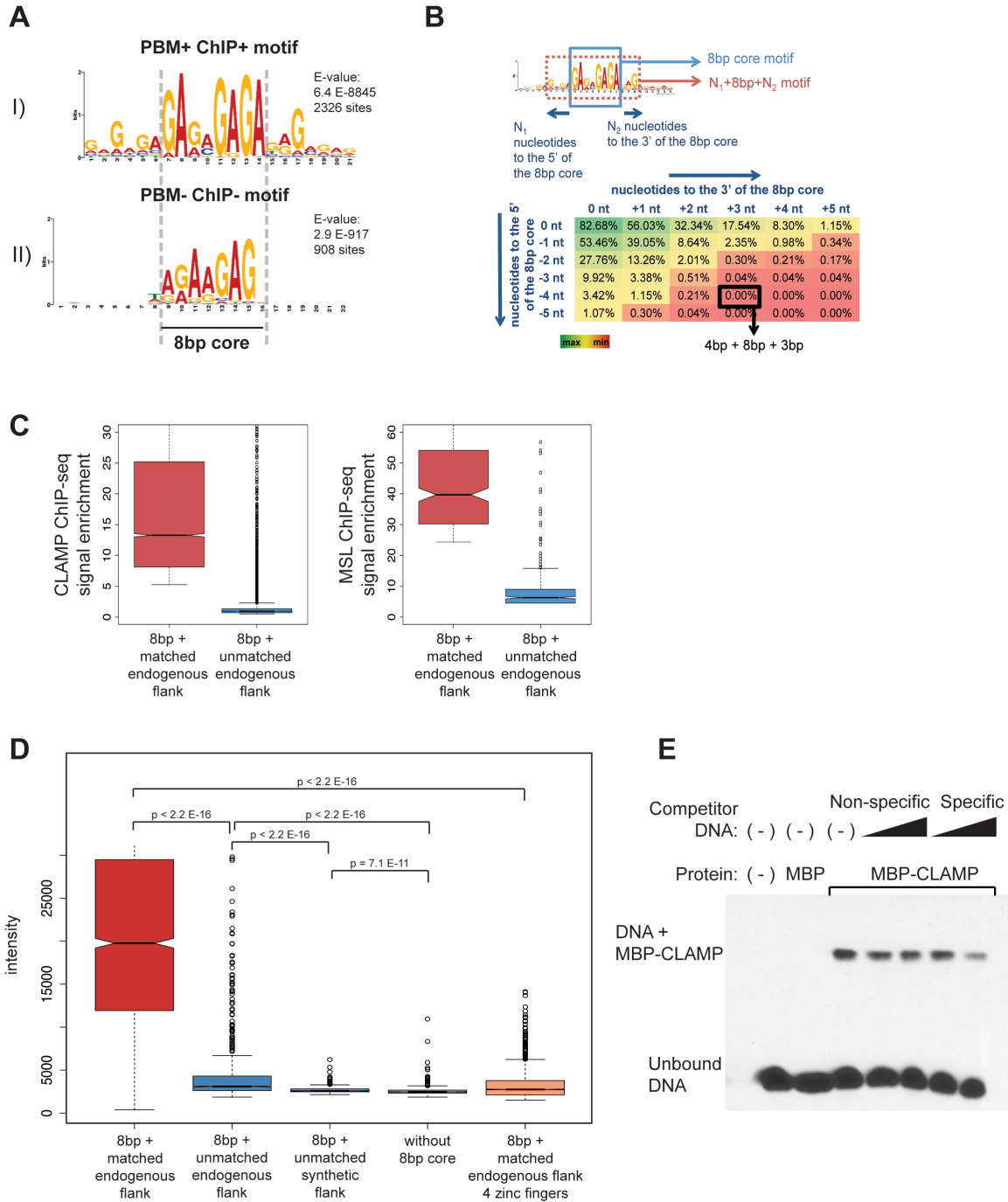


Figure 2.2 A long 15-bp motif increases CLAMP binding to DNA.

A) Motifs obtained from the custom gcPBMs are shown: I) The PBM+ChIP+ motif represents the sequences that CLAMP binds both *in vivo* and *in vitro*. II) The PBM-ChIP- motif represents the sequences that are on the array but not bound by CLAMP *in vivo* or *in vitro*. The most conserved 8-bp core element is indicated by vertical dashed lines. **B)** A representation of the methodology to define the minimal CLAMP-bound motif by scanning both 5' and 3' of the core motif. The table shows the percentage of PBM+ChIP+ (CLAMP binding both *in vitro* and *in vivo*) sequences that overlap with PBM-ChIP- sequences (CLAMP binding neither *in vitro* nor *in vivo*) at the specified motif size. The y-axis shows the nucleotides 5' of the 8-bp core motif and the x-axis shows the nucleotides 3'. The scale ranges from green for the maximal values to red for the minimal values.

Values show the percentage of PBM+ChIP+ sequences shared with PBM-ChIP- sequences for the length window selected. A value of zero overlapping sequences represents complete separation between PBM+ChIP+ and PBM-ChIP- sequences and is obtained at 4-bp 5' and 3-bp 3' of the core motif. **C)** CLAMP and MSL ChIP-seq enrichments are shown for sequences containing the 8-bp core with and without additional flanking sequence matching the motif. Motif hits were found using the FIMO tool ($p < E-4$). All 8-bp core hits were found first and the ones overlapping with the full 15-bp motif were separated as '8bp + matched endogenous flank' and the rest were grouped as '8bp + unmatched endogenous flank'. Since the '8bp + unmatched endogenous flank' group has ~10,000 sites, the top 10,000 enrichments are shown in the CLAMP enrichment plot. Since there are ~300 CES, the top 300 enrichments are shown in the MSL enrichment plot. **D)** Binding intensities are shown for the following classes of probes: 1) probes with the optimal motif (8-bp + matched endogenous flank, red); 2) probes that have matching 8-bp core regions but the endogenous flanks do not match the motif (8-bp + unmatched endogenous flanks, blue); 3) probes that have 8-bp cores with synthetic constant flanking sequences (8-bp + unmatched synthetic flank, cyan); 4) probes that do not have the 8-bp core motif (without 8-bp, brown); 5) Intensities for C-terminal 4 zinc finger GST fusion proteins are shown for probes containing the 15-bp CLAMP motif (15-bp, 4ZF, orange). **E)** CLAMP binds to DNA containing a high affinity, 8 bp + matched flank motif in an electrophoretic mobility shift assay. Biotin-labeled DNA alone (lane 1) and DNA with MBP (lane 2) do not shift, while MBP-CLAMP forms a complex with DNA to shift the signal. This was competed away with specific (high affinity) competitor but not a non-specific competitor that contains the 8-bp core but lacks endogenous flanking sequences (8-bp + unmatched synthetic flank).

Next, we quantitatively determined the optimal size of the CLAMP binding site. We found the minimal number of consensus nucleotides outside of the 8-bp core that were required to separate all bound sequences (PBM+ ChIP+) from all unbound sequences (PBM-ChIP-) (Figure 2.2B, percent overlap of 0%). Using this approach, we identified the minimal key flanking regions as 4-bp 5' and 3-bp 3' of the 8-bp core sequence for a total of a 15-bp motif. After defining the 15-bp optimal CLAMP binding motif, we tested whether the ability of a sequence to match this motif correlates with increased *in vivo* CLAMP and MSL complex ChIP-seq occupancy (Soruco et al. 2013; A. Alekseyenko et al. 2008). We found that the presence of the 15-bp motif (8-bp + matched endogenous flank) greatly increased both CLAMP and MSL complex occupancy compared with the 8-bp core motif alone lacking any flanking sequence that matches the motif (8-bp + unmatched endogenous flank) (Figure 2.2C). Furthermore, we determined that the better a sequence matches the motif (decreased p-value), the higher the *in vivo* occupancy at this motif (Figure S2.1B, Table S2.3).

Next, we measured the role of flanking sequences in a different way by dividing the bound sequences into five different quantiles based on their *in vitro* binding intensities (Figure S2.2A). To quantify these differences, we measured the Euclidean distances between the top quantile of bound motif instances (q1) and the bottom quantile of bound motif instances (q5) for the 8-bp core (blue) and the flanking sequences (red) (Figure S2.2B, Table S2.4). Differences between the strongly and weakly bound motif instances were greater in the flanking region (red) compared to the core (blue), indicating that appropriate flanking bases correlate with enhanced CLAMP interaction with DNA.

To directly test the requirement for flanking sequences to enhance CLAMP binding, we compared the binding of CLAMP to probes containing the optimal 15-bp motif (8-bp + matched endogenous flank) with binding to two additional classes of probes: 1) Probes with the same 8-bp cores as those in the 15-bp motif probes but different flanks derived from endogenous sequences (8-bp core + unmatched endogenous flank); 2) Probes with variable 8-bp cores that all match the original uPBM motif and a single non-endogenous artificial constant flanking sequence (8-bp + unmatched synthetic flank). The probes containing the original 15-bp motifs (8-bp + matched endogenous flank) were bound much more strongly than either class of probes that only contained sequences matched to the 8-bp motif core (Figure 2.2D, p-values for all comparisons are reported as Table S2.5).

Although the gcPBM was performed with a GST-CLAMP fusion as has been done previously (Berger and Bulyk 2006), GST-tags can form homodimers potentially influencing binding profiles (Liew et al. 2008). In order to determine whether the CLAMP motif requirements are specific to GST-tagged CLAMP, we next tested CLAMP binding using a maltose binding protein (MBP) epitope tag by electrophoretic mobility shift assay

(EMSA). We observed that a gcPBM high affinity 15-bp motif-containing sequence was better able to compete for CLAMP compared with a nonspecific gcPBM sequence with constant synthetic flank sequence (Figure 2.2E). Therefore, a long 15-bp motif improves the interaction between CLAMP and DNA independent of the GST-CLAMP fusion.

Next, we hypothesized that the presence of a large DNA binding domain containing six tandem zinc fingers allows CLAMP to recognize a long 15-bp motif. Each DNA-binding zinc finger typically recognizes 3-bp of DNA (Klug 2010), suggesting that five of the six CLAMP zinc fingers would be necessary to recognize a 15-bp motif. To further map the CLAMP DNA interaction domain, we produced constructs with one through five zinc fingers (from both the N-terminal and C-terminal directions). However, the only additional construct that produced soluble protein was the C-terminal four-finger construct (a.a. 412-561). Both the six finger and four finger constructs are able to bind to DNA (Figure 2.2D). However, the six-zinc finger protein bound to probes with the 15-bp motif (red box) with much stronger affinity than the four finger construct (orange box) (Fig 2.2D). This indicates that although the four fingers tested in this construct are sufficient for DNA binding, one of the two deleted fingers is required to significantly increase the affinity of CLAMP binding to probes. It is possible that a specific number of fingers (greater than four) is required, or it may be that one or both of the deleted fingers have the most specificity for the flanking sequence. In summary, we used gcPBMs to rigorously show that a long MRE, such as that previously observed to improve MSL binding (A. Alekseyenko et al. 2008) is required for CLAMP binding to DNA.

CLAMP binds more efficiently to long GA-dinucleotide repeats in vivo and in vitro

Many transcription factors use the biophysical properties of DNA shape to recognize their binding sites in addition to just primary DNA sequence (Abe et al. 2015; Slattery et al. 2014). Therefore, we tested structural features of the MRE and CLAMP motifs such as minor groove width, helix twist, roll, and propeller twist (T. Zhou et al. 2013) for their contribution to CLAMP binding. We found that specific DNA shape features did not influence CLAMP binding at its strong binding sites but there was a modest role for shape at weak binding sites (Figure S2.3). Therefore, shape plays a more significant role in CLAMP binding at its weak binding sites consistent with their lack of a strong consensus motif and similar to the shape requirements observed for other transcription factors (Zentner et al. 2015).

Next, we used a statistical machine learning approach called L2-regularized multiple linear regression (MLR) (Abe et al. 2015; T. Zhou et al. 2013) to determine features that were the best predictors of *in vitro* CLAMP binding based on gcPBM data. We found that the addition of dinucleotides to the model (1mer+2mer) was a better predictor of CLAMP binding than the addition of DNA shape parameters (1mer+shape) ($p < 0.001$) (Figure 2.3A). The addition of trinucleotides (3-mers) did not substantially further increase the ability to predict binding (1.6% percent performance gain). Therefore, we concluded that dinucleotides were the best predictors of CLAMP binding *in vitro*.

Because GA-dinucleotides were the best predictors of CLAMP binding, we examined the binding of CLAMP to PBM probes containing increasing lengths of GA-repeats directly. We found that there is a stepwise increase in binding to longer repeats that begins between 4 repeats (8-bp) and 5 repeats (10-bp) corresponding with extending the motif beyond the 8-bp core of the motif (Figure 2.3B). The increase in binding plateaus at 10 repeats (20-bp) and there are very few probes that have more than 11 repeats within the

36-bp probe sequence so they were binned together (Figure 2.3B, number of probes for each repeat length are reported as Table S2.6).

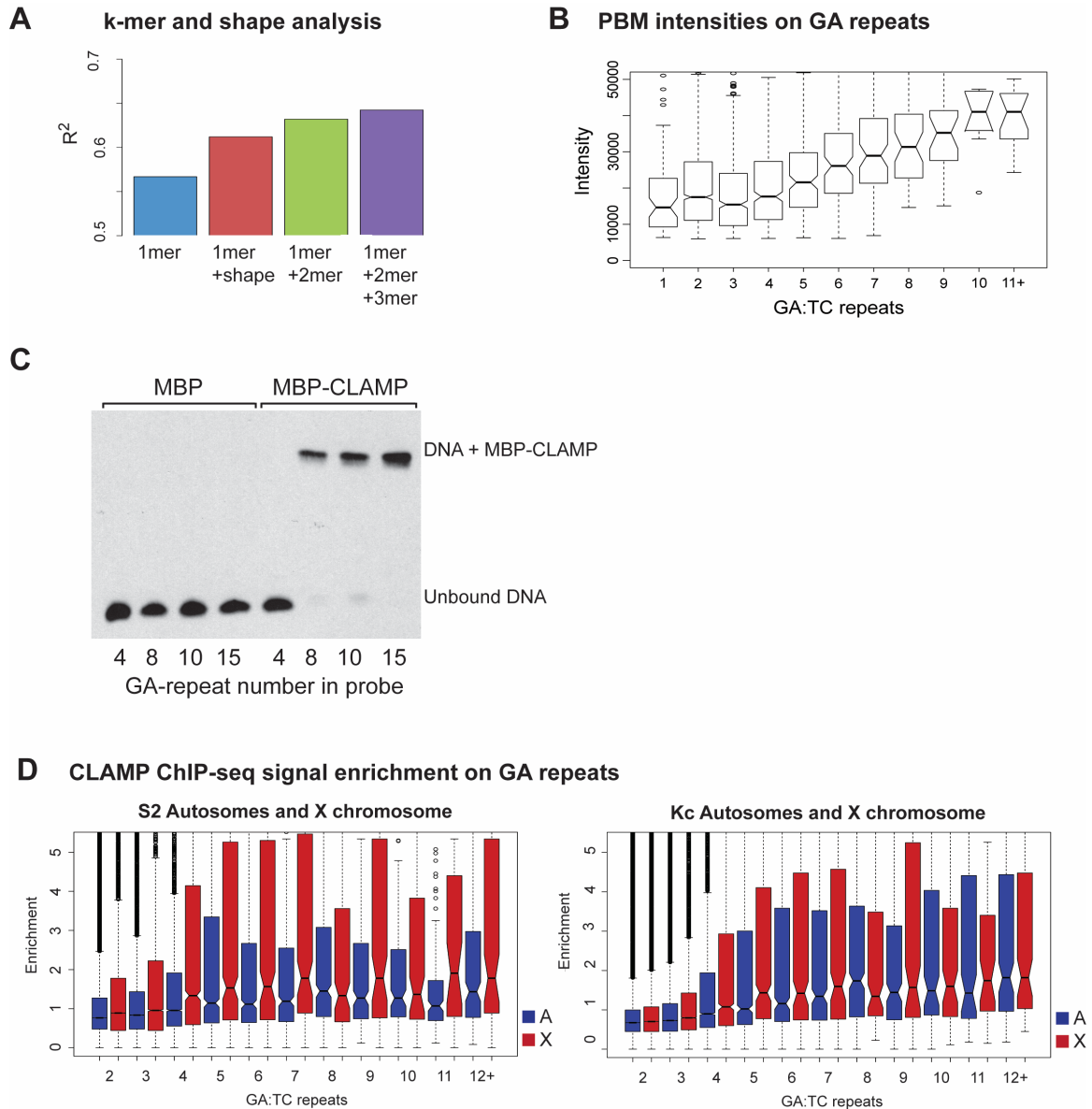


Figure 2.3 Increasing the number of GA-dinucleotide repeats increases CLAMP occupancy.

A) A multiple linear regression to test contribution of sequence length (k-mer) and shape to overall binding. Adding dinucleotide (2mer) features to the sequence-only (1mer) model increases the performance more than adding DNA shape features, indicating the importance of dinucleotides in CLAMP-DNA recognition. Adding trinucleotide (3mer) features further increases the performance marginally. **B)** CLAMP PBM binding for GA-dinucleotide repeats of different lengths is plotted as box plot distributions. The y-axis is the PBM intensity score for each number of GA-repeats, which are plotted along the x-axis, e.g. 1 = GA, 2=GAGA. **C)** An electrophoretic mobility shift assay to test MBP-CLAMP binding to increasing numbers of GA-repeats. The labeled probes contain GA-repeats of 4 (8-bp), 8 (16-bp), 10 (20-bp) and 15 (30-bp) centered

within a 60-bp sequence. The first four lanes are reactions containing MBP control protein with DNA, and the next four are MBP-CLAMP with DNA. **D)** Input-normalized CLAMP ChIP-seq signal enrichments at GA-repeats of different lengths are given for the X-chromosome (red) and autosomes (blue) from male S2 (top) and female Kc cells (bottom). The x-axis shows the number of GA-repeats e.g. 1 = GA, 2=GAGA.

In order to ask whether the signal on the PBM increases with the number of GA-repeats because a single CLAMP protein molecule binds to DNA with greater affinity or multiple CLAMP molecules bind to a single probe, we performed EMSAs with GA-repeat probes. Again, we are using MBP-tagged CLAMP to test if binding increases with GA-length independent of potential dimerization of the GST-tagged CLAMP. The probes contained 4, 8, 10, and 15 GA-repeats (8-bp, 16-bp, 20-bp, 30-bp) embedded within a 60-bp sequence. When MBP-CLAMP was added to each labeled probe, there was an increase in the shifted, protein-bound DNA signal for all probes except the 8-bp probe that corresponds only to the core CLAMP motif (Figure 2.3C). There was a single shifted species for all probes, even the 30-bp GA-repeat which indicates that increased signal is likely due to the greater frequency of a single CLAMP protein binding to DNA not multimerization of the CLAMP DNA binding domain on probes with longer repeats. We cannot fully exclude the possibility that GST-CLAMP dimerizes in the gcPBM experiments, but by testing CLAMP binding with a non-dimerizing tag we observe the same trends in binding indicating it is likely CLAMP rather than the tag predominantly influences binding to longer GA-repeats.

Consistent with the *in vitro* analysis, *in vivo* occupancy of CLAMP from ChIP-seq data in both male (S2) and female (Kc) cells increases as the number of GA-repeats increases on both the X-chromosome (red) and autosomes (blue) (Figure 2.3D, number of repeats analyzed in the genome reported in Table S2.7). We also compared binding on individual autosomal arms for each different GA-repeat length and observed similar

trends (S2 cells: Figure S2.4A and Kc cells: S2.4B). As previously reported, *in vivo* CLAMP binding is modestly increased on the X-chromosome in S2 cells (male) when compared to Kc cells (female) and autosomes due to synergistic interactions between MSL complex and CLAMP (Figure 2.3D) (Soruco et al. 2013). Overall, CLAMP occupancy increases both *in vivo* and *in vitro* as the number of GA-repeats increases on both the X-chromosome and autosomes.

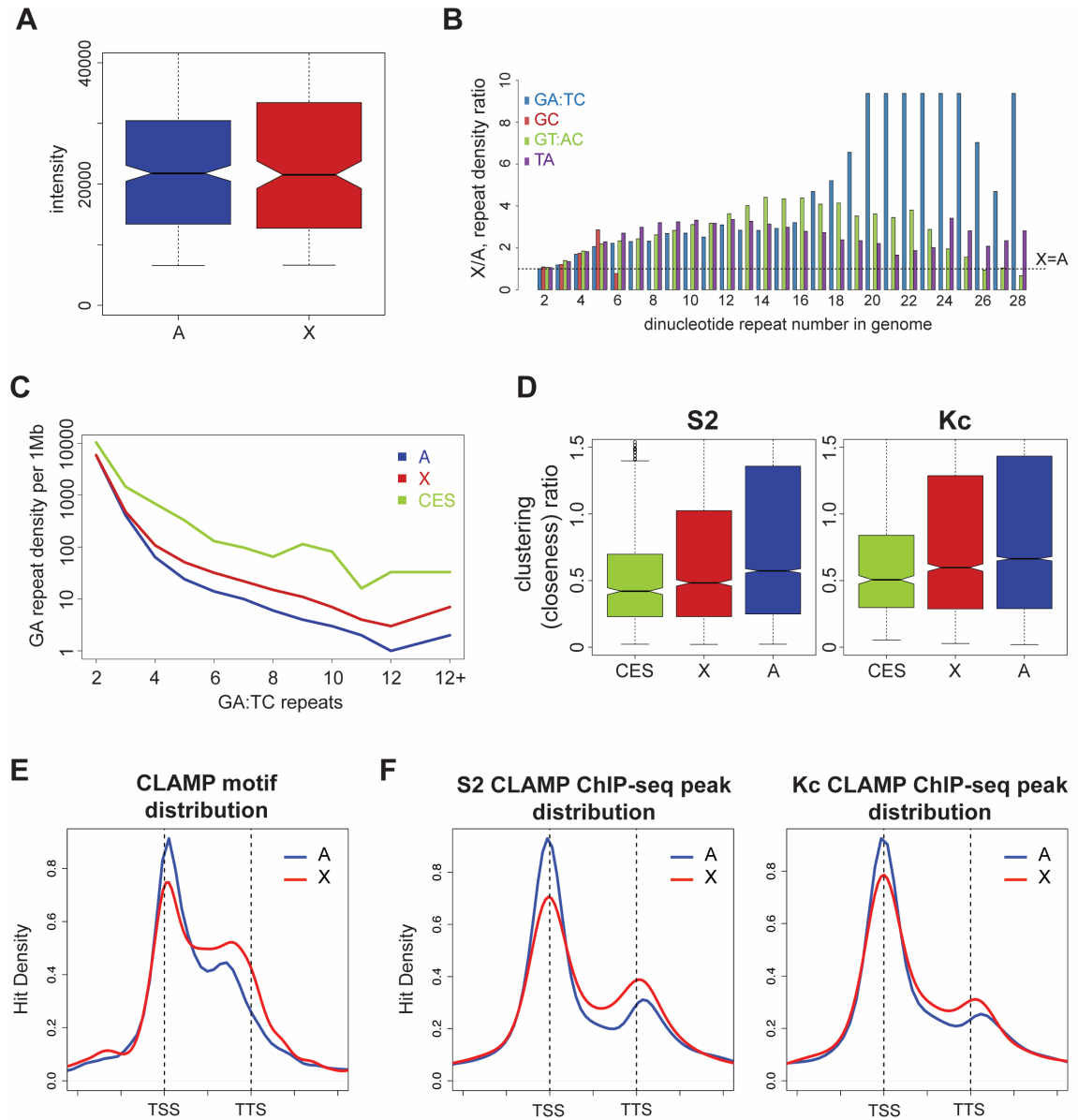


Figure 2.4 The X-chromosome has more GA-repeats than autosomes and the CES have clustering of CLAMP occupancy.

A) Distribution of the PBM intensities of the probes aligned to the X-chromosome (red) versus autosomes (blue). **B)** The density of different dinucleotide repeats on X-chromosome is compared with autosomes for different repeat lengths within the *D. melanogaster* genome. Any value above 1 indicates a higher repeat density on X-chromosome compared with autosomes. **C)** Density of GA repeats (1=GA, 2 = GAGA, etc.) per 1 Mb along Autosomes (A, blue), the X-chromosome (X, red), and X-chromosome CES (CES, green). **D)** Top: Average distance between n neighboring CLAMP ChIP-seq peaks in S2 (left) and Kc (right) cells. Peaks are separated as within CES (green), on X (red), or on autosomes (blue). Bottom: The average distances are replotted for S2 (left) and Kc (right) cells after normalization using distances measured between n random peaks. **E)** A histogram showing the frequency of each 15-bp CLAMP motif at a particular location relative to genes on the X (red) or autosomes (blue). The profiles are shown around gene bodies and the range is from one gene length before the TSS to one gene length after the TTS. The distance from the TSS is normalized to gene length for each gene. **F)** Average gene profiles of CLAMP ChIP-seq peaks on X (red) and autosomes (blue) from S2 cells (male) and Kc cells (female). Normalization was conducted as for E.

The X-chromosome has increased density of GA-repeats and CLAMP occupancy compared to autosomes

While CLAMP is necessary for MSL complex recruitment specifically to the X-chromosome, CLAMP occupies GA-rich binding sites all over the genome (Sorucu et al. 2013). In fact, the average *in vitro* binding to the group of X-linked sequences on the PBM does not differ significantly from the average binding to autosomal sequences (Figure 2.4A). There is a statistically significant increase in CLAMP occupancy at X-linked sites only in the top quantile of binding affinity (p-value <0.01) (Figure S2.5A). These *in vitro* data are consistent with the binding of CLAMP to similar GA-rich motifs throughout the genome based on ChIP-seq analysis (Sorucu et al. 2013). Therefore, we hypothesized that rather than enhanced CLAMP binding to each individual site, an increased density of GA-repeats may occur on the X-chromosome that would generate more CLAMP binding sites and clustering of groups of CLAMP sites in closer proximity to each other.

The CLAMP binding sites we identified from our custom PBMs are 2.2 fold X-enriched, a slightly greater enrichment than the 1.8 fold enrichment of the MRE sequences (A. Alekseyenko et al. 2008). Because CLAMP binds more strongly to sites containing more

GA-repeats (Figure 2.3), we compared the GA-repeat content of the X-chromosome to autosomes. Classical FISH studies conducted before the sequencing of the *Drosophila* genome demonstrated that many types of dinucleotide repeats are increased on the X-chromosome (Lowenhaupt, Rich, and Pardue 1989). However, this previous work did not precisely define the dinucleotide repeat content of the X-chromosome vs. autosomes, so we computationally measured the ratio of dinucleotide repeats of all lengths throughout the genome. We determined that the density of all types of dinucleotide repeats is strongly increased on the X-chromosome vs. autosomes (Figure 2.4B). We also compared X to individual autosomal arms (and observed the same trends (S2.5B, S2.8). Interestingly, long GA-repeats are more highly enriched on the X-chromosome than other types of dinucleotide repeats, consistent with the X-enrichment of long GA-rich CLAMP binding sites (Figure 2.4B). The regions of highest MSL complex occupancy (CES, green) are even more enriched for GA-repeats than other regions of the X-chromosome (red) or autosomes (blue) (Figure 2.4C). Therefore, we hypothesized that increased density of long GA-repeats within CES increases the density of CLAMP binding sites, which then promotes MSL complex targeting to CES.

To test this hypothesis, we directly measured the average distances between *in vivo* CLAMP-occupied sites at CES, other sites on X, and autosomal sites and compared them to the values expected for random site distribution (Figure 2.4D, Figure S2.6, p-values reported as Tables S2.9, S2.10, S2.12, and S2.13). Specifically, we measured the distances between CLAMP ChIP-seq peaks in S2 and Kc cells and calculated average distance of each site to the closest 1-4 neighboring sites (number of peaks used are reported as Table S2.11). The CES have smaller distances between CLAMP sites than all other loci on the X-chromosome or autosomes, including individual sites on

autosomal arms (Figure S2.6A). To account for difference in absolute counts of the CLAMP sites within CES and other regions, we normalized our data by comparing with a random distribution of the same number of sites. We demonstrated that CLAMP sites are relatively closer to each other around the CES compared with all other genomic regions in the presence (S2 cells) and the absence of MSL complex (Kc cells). Therefore, the clustering of CLAMP peaks is increased around CES even when normalized for increased peak number and is independent of MSL complex. Overall, these data support our hypothesis that CES can be distinguished from other genomic locations prior to MSL complex recruitment due to their increased density of CLAMP sites.

Unlike most transcriptional regulators that localize to transcription start sites (TSS), the MSL complex is recruited to active gene bodies and transcription termination sites (TTS) to promote the progression of RNA Polymerase II across genes (Larschan et al. 2011; Ferrari et al. 2013; AA Alekseyenko and Larschan 2006; Gilfillan et al. 2006). Therefore, we hypothesized that increased density of CLAMP peaks and motifs over gene bodies and 3' UTRs on the X-chromosome vs. autosomes would cause CLAMP occupancy to be more broadly distributed on X-linked genes increasing overlap with MSL complex binding sites. To test this hypothesis, we mapped the position of the CLAMP binding motif from our top bound sites to the genome (Figure 2.1A:IV) and generated an average gene profile to measure the frequency of its distribution across genes (Figure 2.4E). Next, we compared the location of CLAMP motifs to a similar plot of CLAMP occupancy from ChIP-seq data on the X-chromosome and autosomes in both male (S2) and female (Kc) cells (Figure 2.4F). We also quantified the motif density and CLAMP occupancy ratios of the gene body (from TSS+250 bp to the TTS) to the 5' end (TSS $\pm$ 250 bp) for the X-chromosome, autosomes, and individual autosome arms (Figures S2.7A and

S2.7B). CLAMP motifs and occupancy are concentrated at the TSS and TTS of genes on both the X-chromosomes and autosomes. However, on the X-chromosome, CLAMP motifs and occupancy levels are more broadly distributed over gene bodies relative to TSS compared with than on autosomes. Overall, the localization of CLAMP on the X-chromosome is more biased towards gene bodies and the TTS and away from the TSS than on autosomes. Therefore, X-linked CLAMP binding sites have a greater chance of co-occupying sites where MSL complex binding would be stabilized by gene body chromatin marks with known roles in dosage compensation such as H3K36me3 (Larschan et al. 2007a) than autosomal sites.

The CLAMP protein and X-enrichment of its GA-repeat binding sites are conserved across species

Recent work demonstrated that dosage compensation across *Diptera* evolved independently as many different chromosome arms transitioned from autosomal to X-chromosome identities (Vicoso and Bachtrog 2015). To determine whether CLAMP could be a key factor involved in the evolution of the X-chromosome across *Diptera*, we first examined its conservation. The CLAMP gene is highly conserved across all *Diptera* that have sufficient sequencing data to find orthologs (Figure 2.5A, Figure S2.8). Moreover, the DNA binding domain (a.a. 350-561) always has six tandem zinc fingers and is very highly conserved across diverse species (Table S2.14) (Weirauch et al. 2014). The glutamine-rich domain is less conserved in terms of specific amino acid position but its glutamine-rich nature is conserved across species (Figure S2.9, Table S2.14) (Sievers et al. 2014).

To determine how conserved the gene encoding CLAMP is compared to all other genes in *Drosophila*, we compared the ratio of non-synonymous to synonymous changes in

CLAMP between *D. melanogaster* and *D. simulans* to the same ratio for all 1:1 orthologs in the *D. melanogaster* species group (Larracunte et al. 2008) (Figure 2.5B). Among *Drosophilids*, CLAMP is more highly conserved than all but 11% of *Drosophila* protein coding sequences. Moreover, CLAMP is much more highly conserved than the MSL complex components for which 1:1 orthologue data were available: MSL1, MSL3 and MLE (maleless). The low level of conservation of MSL components is consistent with previous measurements (Rodriguez et al. 2007). Therefore, it is possible that the dosage compensation machinery could have co-opted the function of the more ancient CLAMP protein independently in multiple species as X-chromosomes evolved.

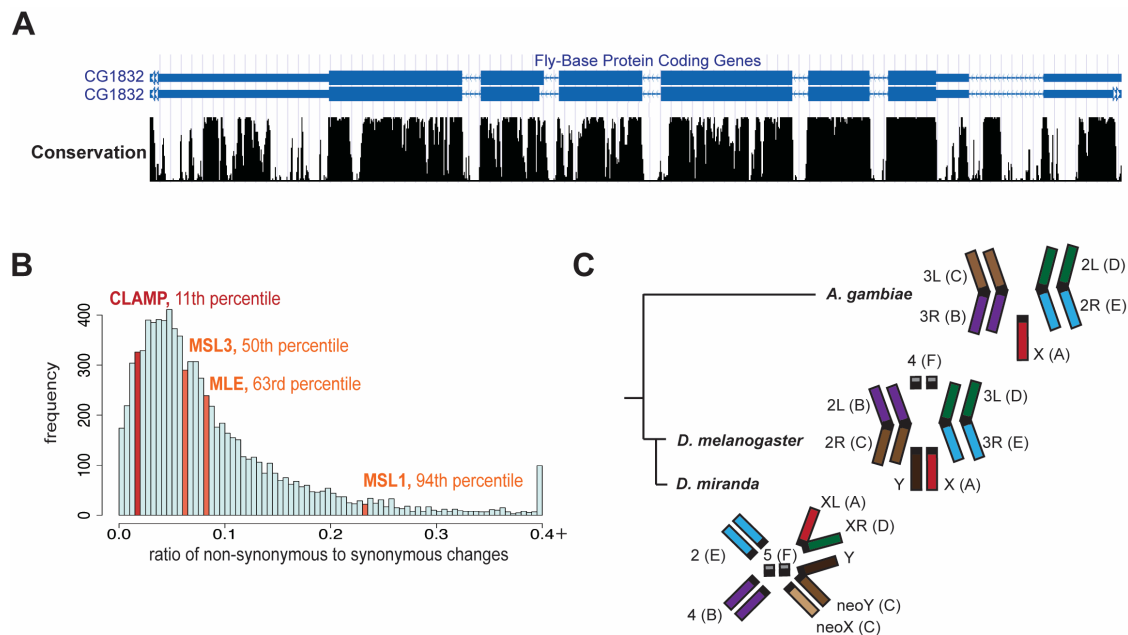


Figure 2.5 CLAMP and enrichment of the CLAMP motif are highly conserved across species.

A) A schematic of the CLAMP (CG1832) gene from *D. melanogaster* is shown in blue. Conservation of the CLAMP gene sequence compared to orthologues in other *Drosophilids* and the mosquito (*A. gambiae*) is shown below (UCSC Genome Browser). **B)** The genomic distribution of the ratio of non-synonymous to synonymous changes is given (dN/dS ratio) for all 8560 genes with 1:1 orthologues between *D. melanogaster* and *D. simulans*. Percentiles are listed for CLAMP and the MSL components for which 1:1 orthologues were present: MSL1, MSL3 and MLE. **C)** Schematic of male *A. gambiae*, *D. melanogaster*, and *D. miranda* chromosomes and corresponding Muller element letters. Colors are matched based on Muller element identity (Vicoso and Bachtrog 2015; Bolshakov et al. 2002).

Based on the strong conservation of CLAMP across species, we hypothesized that the CLAMP motif may have become X-enriched independently on different X-chromosomes in multiple species. To test this hypothesis, we compared the density of CLAMP motifs on the X-chromosome vs. autosomes in several *Diptera* in which appropriate assemblies and sequence information defining the X-chromosome is available. Among *Drosophilids*, we compared *D. melanogaster* with the distantly related *D. miranda* (30 million years apart) (Brown and Bachtrog 2014) that has both ancient X-chromosomes (XL,XR) and a newly evolving X-chromosome (neoX) (Figure 2.5C). Therefore, it is possible to examine the neoX as an intermediate in the evolution of dosage compensation (Artyom a Alekseyenko et al. 2013; Q. Zhou et al. 2013; C. E. Ellison and Bachtrog 2013). As noted previously, the density of CLAMP binding sequences per megabase on the X-chromosome is 2.2-fold enriched compared with autosomes (Figure 2.6A: red, Table S2.15). Within *D. miranda*, the ancient XL and XR chromosomes are more enriched for CLAMP motifs compared with the newly evolving neoX chromosome that has a modest 1.1 fold enrichment (Figure 2.6A: blue, Table S2.15).

To determine the degree of conservation of the X-enrichment of the CLAMP motif outside of *Drosophilids*, we examined the X-enrichment of the CLAMP motifs in *Anopheles gambiae*, the distantly related mosquito (250 million years apart) with a fully assembled genome. For measurements of motifs and repeat density, we found that fully assembled genomes and scaffolds did not correlate well, and therefore we used only fully assembled genomes. Interestingly, we observe the same X-enrichment of CLAMP motifs in *A. gambiae* and *D. melanogaster* (2.2-fold) revealing conservation across 250 million years of evolutionary time (Figure 2.6A: green). Because CLAMP binds more strongly to long GA-repeats (Figure 2.3), we determined the density of GA-repeats of different lengths across species.

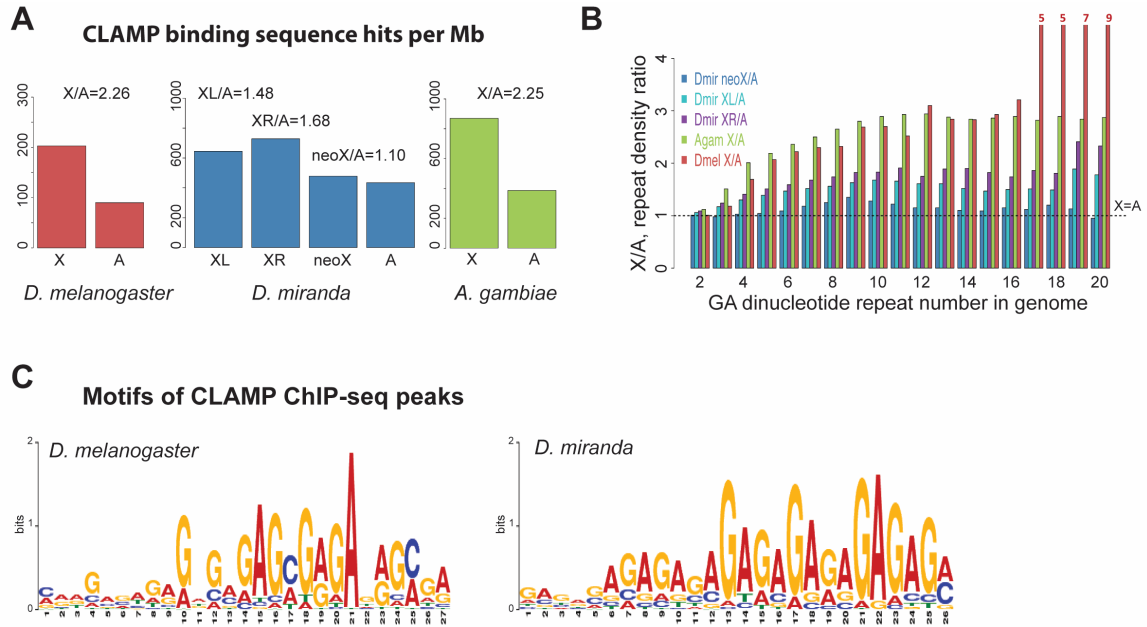


Figure 2.6 CLAMP binds to similar *in vivo* X-enriched binding sites in *D. miranda* and *D. melanogaster*.

A) The ratio of *in vitro* CLAMP binding site density (number of CLAMP binding sequence hits per Mb) for X versus autosomes is plotted for *D. miranda*, *D. melanogaster* and *A. gambiae*. Autosomes of *D. melanogaster* are chromosomes 2L (Muller-B), 2R (Muller-C), 3L (Muller-D), 3R (Muller-E) and 4 (Muller-F); autosomes of *D. miranda* are chromosomes 2 (Muller-E), 4 (Muller-B) and 5 (Muller-F); and autosomes of *A. gambiae* are chromosomes 2L (Muller-D), 2R (Muller-E), 3L (Muller-C), and 3R (Muller-B). **B)** The density ratios of GA-repeats on individual X-chromosome(s) vs. autosomes for different repeat lengths in the *D. melanogaster*, *D. miranda* and *A. gambiae* genomes are plotted. Any value above 1 indicates a higher repeat density on the X-chromosome compared to autosomes. **C)** CLAMP ChIP-seq motifs are shown for *D. melanogaster* and *D. miranda* larval ChIP-seq data. Motifs are found using MEME-ChIP for the peak regions (peak centers ± 100 bp).

We observed elevated GA-repeat density on the X-chromosome in *A. gambiae* and the ancient X-chromosomes in *D. miranda* (XL and XR) (Figure 2.6B). In contrast, the newly evolving neo X-chromosome has very modest enrichment of GA-repeats compared with autosomes (Figure 2.6B). We also analyzed other types of dinucleotide repeats and found that they are also X-enriched that suggests a general expansion of dinucleotide repeats (Figure S2.10 and Table S2.16, *D. miranda*, Table S2.17, *A. gambiae*). However, CLAMP recognizes only a motif containing GA-repeats (Sorucu et al. 2013) and therefore it is unlikely that the expansion of the other repeats would alter CLAMP occupancy. To test that our combined autosome scores were representative of individual

autosomal elements, we compared the density ratios to each individual arm, and in all cases dinucleotide repeats were enriched on X with the exception of Muller Element F that is highly heterochromatic (Figure S2.11, *D. miranda*, and Figure S2.12, *A. gambiae*). Therefore, the X-enrichment of GA and other dinucleotide repeats is conserved across 250 million years of evolution and is increased on ancient X-chromosomes when compared with a newly evolving X-chromosome.

To determine whether CLAMP binds to the same motif across species, we performed ChIP-seq in *D. miranda* larvae and compared our motifs with those from *D. melanogaster* larvae. Overall, the motifs that we derived from *D. miranda* larvae ChIP-seq are similar to those from *D. melanogaster* larvae, *in vivo* motifs from tissue culture cells and the *in vitro* gcPBM motif (Figures 2.6C and 2.1A). When we compared the genomic sequences of each *D. miranda* chromosome to the respective motif identified from *D. miranda*, we found that ancient XL and XR had the strongest motifs with the lowest p-values (Figure S2.13A). Furthermore, when we obtained motifs from the peaks bound at the highest level (top 1000 peaks), we observed even more dramatic elongation of GA-repeats within the *D. miranda* motif (Figure S2.13B). Next, we performed controls to assure the specificity of our motifs by examining the occurrence of the true motifs within 1-kb genomic regions surrounding ChIP-seq peaks vs. scrambled and randomized control genomic regions (Figure S2.13C). As p-values decreased, the specificity of the motifs increased, validating the motifs. Furthermore, the motifs that we obtained were X-enriched and their X-enrichment increased as the p-values decreased (Figure S2.13D, Table S2.15). Overall, the CLAMP protein interacts with a very similar sequence across 30 million years of evolutionary time suggesting that it is an ancient factor that recognizes GA-repeats.

Although the *D. melanogaster* and *D. miranda* CLAMP *in vivo* motifs are very similar (Figure 2.6C), we observed one intriguing difference: The incidence of a cytosine instead of an adenine at position 4 within the GA-rich core occurs more often in *D. melanogaster* (14.3%) than in *D. miranda* (6.0%). While the interaction of CLAMP with DNA is highly sequence-specific (Soruco et al. 2013), the MSL2 component of MSL complex has some specificity for interacting with MREs via its CXC domain when expressed at high levels *in vitro* (Zheng et al. 2014; Fauth et al. 2010). This specificity is conferred by interaction between the MSL2 R543 residue and the cytosine that can occur within some MRE motifs at position 4. Therefore, we examined the *D. miranda* MSL2 CXC domain and determined that it lacks 4 out of 9 cysteine residues that are required to maintain its structure and the key R543 residue that contacts the cytosine specifically (Figure S2.14) (Goujon et al. 2010). Therefore, differences in DNA binding specificity of CLAMP correlate with changes in the MSL2 protein that alter its DNA binding specificity, suggesting potential co-evolution of these two proteins.

In summary, we have integrated biochemical, *in vivo*, genomic, and evolutionary approaches to support a model by which expansion of GA simple sequence repeats on the X-chromosome promotes increased density of CLAMP within CES, which then targets MSL complex. It is likely that increased density of the maternally loaded CLAMP protein at CES functions together with previously identified regulators such as the *roX* RNAs (Koya and Meller 2011) and the H3K36me3 histone mark (Larschan et al. 2007a; Bell et al. 2007) to promote specific recruitment of MSL complex to the X-chromosome.

Discussion

Upon the evolution of heterogametic species, the process of dosage compensation became essential to ensure the appropriate balance of gene expression between males and females and the X and autosomes. Distinguishing the X-chromosome from autosomes is the key step in this process because MSL complex must be targeted to the correct chromosome to ensure the fidelity of dosage compensation. Here, we demonstrate that in several species this process likely involved enriching the evolving X-chromosomes for long GA-repeat binding sites that can be recognized by the highly conserved CLAMP protein that recruits MSL complex.

CLAMP binding sites are not X-specific as the CLAMP protein binds to similar GA-rich sequences all over the genome (Soruco et al. 2013). We propose that a higher density of sites within CES that contain longer GA-repeats evolved to optimize CLAMP binding on X to better target MSL complex for dosage compensation. Then, it is likely that the increased density of CLAMP at CES functions together with other cofactors with known roles in MSL complex recruitment such as H3K36me3 (Larschan et al. 2007a; Bell et al. 2007) and *roX* RNAs (Koya and Meller 2011). Once this initial process of X-chromosome identification occurs, synergistic interactions between maternally loaded CLAMP and the MSL complex (Soruco et al. 2013) increase the X-enrichment of both factors.

Interestingly, the CLAMP motif is much longer than most transcription factor binding sites. It is possible that the length of the CLAMP binding site ensures specificity by reducing the promiscuity of its binding and allowing it to compete with other similar proteins. In addition, recent work on transcriptional regulators in budding yeast has implicated the sequence context of transcription factor binding sites outside of the core

binding site as critical for the recognition process (Levo et al. 2015). Therefore, current approaches to identifying transcription factor binding site motifs have likely underestimated their length due to the approaches used that often allow detection of only short motifs. In the future, it will be important to determine transcription factor recognition motifs using approaches like gcPBM that uses *in vivo* sequences to identify direct binding site motifs.

There are several mechanisms by which the GA-repeat number could have been increased including expansions due to slippage of DNA polymerase. Helitron transposons containing GA-rich sequences have also been implicated in the X-enrichment of these sequences in *D. miranda* (C. E. Ellison and Bachtrog 2013). It is possible that expansions of GA dinucleotides occurred within these transposons after they landed on the X-chromosome. These GA-repeat expansions could have been further propagated by gene conversion events that also occurred during the evolution of dosage compensation (C. E. Ellison and Bachtrog 2013). Finally, long repeat sequences such as the 1.688 elements that produce siRNAs function during dosage compensation via an unknown mechanism (Menon et al. 2014). Therefore, it is possible that GA-repeat elements have been expanded over evolutionary time because of a general role in promoting dosage compensation. To support this hypothesis, a recent report identified GA-rich binding motifs almost identical to those that we characterized as CLAMP binding sites within the strongest MSL complex binding sites in three additional *Drosophila* species (Quinn et al. 2016).

Motifs that contain GA-repeats have been implicated in diverse processes that all involve generating open chromatin regions. GA-repeat containing motifs are highly enriched at sites that promote pausing of RNA Polymerase II and at developmentally regulated

DNase I hypersensitivity sites (Guertin and Lis 2013; Kharchenko et al. 2011). Furthermore, a GA-repeat motif is one of the two motifs that are enriched at genes that are activated first during the maternal to zygotic transition (Chen et al. 2013). The well-studied GAGA factor (GAF) protein also recognizes similar sequences to the CLAMP protein and has been implicated in pausing of RNA Polymerase II and opening of chromatin (T Tsukiyama and Wu 1996). Overall, it is likely that the dosage compensation machinery has evolved to take advantage of targeting GA-repeats that mark open chromatin regions to ensure that it only identifies active genes for further transcriptional upregulation by the MSL complex.

It is possible that GA-rich sequences have roles in dosage compensation outside of *Diptera*. For example, it has been proposed that upregulation of the single active X occurs in mammals and this process is mediated by targeting the conserved MOF histone acetyltransferase component of MSL complex (Deng et al. 2011). Moreover, GA-repeats were found to be significantly enriched within regions of the X-chromosome that escape X-inactivation (X escape regions) (Clemson et al. 2006). There are no strong homologues of CLAMP in mammals but there are several possible functional orthologs such as the ETS family transcription factor GABP1 (GA binding protein-1) (Yang, Mott, and Rosmarin 2007). Furthermore, in *C. elegans*, there is an early upregulation of both X-chromosomes that is also mediated by the MOF histone acetyltransferase (Deng et al. 2011). One of the zinc finger proteins that targets the *C. elegans* dosage compensation machinery is SCC-2 (sister chromatid cohesion - 2) which recognizes a GA-repeat sequence very similar to the CLAMP binding motif (Kranz et al. 2013). Therefore, it is possible that GA-repeats are involved in dosage compensation beyond *Diptera* and this will be an exciting area for future investigation.

Materials and Methods

gcPBM design, experimental methods, and data analysis

Design of custom genomic-context Protein-Binding Microarray (gcPBM). PBM

probes were designed to assay three classes of DNA sequences defined by (1) the presence of an MRE sequence (derived from MSL3-TAP ChIP-seq data (A. Alekseyenko et al. 2008)), and (2) *in vivo* binding determined by CLAMP ChIP-seq experiment (Soruco et al. 2013). CLAMP ChIP-seq peaks (identified at a false-discovery rate of 0.001) were examined for the presence of CLAMP binding sites (Soruco et al. 2013). Matches to two previously identified minimal *in vitro* CLAMP binding motifs from universal PBM (uPBM) experiments (Soruco et al. 2013) were used to define a CLAMP binding site (CLAMP motif, cutoff <7.8; CLAMP secondary motif, cutoff < 6.0). CLAMP ChIP-positive regions, each containing a minimal CLAMP *in vitro* binding site, were then divided based on the presence or absence of an MRE sequence and labeled as ChIP+/MRE+ or ChIP+/MRE-, respectively. Similarity to the MRE motif (A. Alekseyenko et al. 2008) was determined using the FIMO tool (Grant, Bailey, and Noble 2011). Sequences with similarity of less than $p\text{-value}=10^{-4}$ were labeled as MRE+, and sequences with similarity of higher than $p\text{-value}=10^{-3}$ were labeled as MRE-. A third set of probes focused on promoter regions from the fly genome (500 bp up and downstream of transcription start site; fly genome build dm3 from UCSC) that did not overlap with CLAMP ChIP-positive regions (i.e., contain a CLAMP binding site but were not bound *in vivo*). Promoter regions were ranked according to CLAMP binding motif scores (as described), for the top-scoring 250 promoter regions, PBM probes were constructed for all identified CLAMP binding sites, and are referred to as ChIP-/MRE+. We also tiled the genomic regions including the *roX1*, *roX2* and CES5C2 MSL complex binding sites. Our design protocol resulted in 3944 ChIP+/MRE+ probes, 2313 ChIP+/MRE- probes, and for 1282 ChIP-/MRE+ probes. All probes on the PBM are 60-nucleotides (nt) long and

contain a 24-nt primer sequence (positions 37 to 60) and a 36-nt variable region. Probes were constructed with the minimal CLAMP binding site from uPBMs centered at nucleotide position 19, allowing 17-nt of genomic flanking DNA on either side. All probes are included four times on the array in each orientation (i.e., forward sense and reverse complement), resulting in eight replicate measurements per unique binding sequence. See Table S2.1 for complete design details for the gcPBM.

Probes were defined as PBM+/- based on their PBM signal intensities. Probe intensities had a bimodal distribution (Figure 2.1C). We selected a value of 6500, which is between the two peaks of the distribution, as the threshold to define binding *in vitro*: probes with an intensity higher than 6500 were labeled as PBM+ (“bound”) and probes with an intensity lower than 6500 were labeled as PBM- (“not bound”). gcPBM data will publically available upon publication.

***In vitro* protein expression for protein binding microarrays (PBMs).** The pDEST15 GST expressing Gateway vector (Invitrogen) was used to generate the expression plasmid for the C-terminal CLAMP four and six zinc finger constructs. For the six-finger construct (amino acid residues 350-561), the following primers were used for cloning: forward primer 5'-ACTGTTAAACTGAAAATCCGTCCAA-3' and reverse primer 5'-CTATAACCCACCGATAATCACTTCCTGT-3'. For the four-finger construct, (amino acid residues 412-561), the following primers were used for cloning: 5'-GGCGAGAAACCCTATAAATGCCAAAC-3' and reverse primer 5'-CTATAACCCACCGATAATCACTTCCTGT-3'. Cloning was performed using standard Gateway Technology (Invitrogen). Proteins were expressed using the TnT Quick-coupled Transcription and Translation Systems Kit (Promega). *In vitro* transcription and translation reactions were carried out according to standard protocols with the following

adaptation: zinc acetate was added to a final concentration of 50 μ M. Proteins were quantified using Western blotting with a GST dilution series as a standard. The concentrations of the proteins used for PBMs were as follows: 1) C-terminal 4-zinc finger protein (24.3 nM); 2) C-terminal 6-zinc finger protein (13.9 nM).

PBM experiments and initial analysis. PBM experiments were performed using a custom-designed oligonucleotide array (Agilent Technologies, Inc., AMADID #037964, design described above). The array was converted to a double-stranded DNA array by primer extension and used in PBM experiments essentially as described previously (Berger and Bulyk 2006; Berger and Bulyk 2009). Protein samples containing C-terminal 4- and 6-finger fusions of the CLAMP protein to GST (preparation as described above) were incubated in four different chambers of a gcPBM microarray at a final concentration of $\sim$ 200 nM for 1 hour in binding buffer (10 mM Tris-HCl pH 7.4; 0.2 μ g/ μ l bovine serum albumin (BSA) (New England BioLabs #B9001S); 0.3 ng/ μ l salmon testes DNA (Sigma, #D7656); 2% non-fat dry milk (Stop & Shop brand); 1 mM dithiothreitol (DTT); 1 mM ethylenediaminetetraacetic acid (EDTA); 50 μ M zinc acetate dehydrate; 80 mM NaCl). Protein-bound arrays were then washed and incubated with antibody for 20 min (0.05 mg/ml Alexa 488-conjugated anti-GST antibody (Invitrogen, #A11131). Microarray scanning, quantification, and data normalization were performed using GenePix Pro ver. 6 (Axon) and masliner (MicroArray LINEar Regression) software (Dudley et al. 2002) as previously described (Berger and Bulyk 2006; Berger and Bulyk 2009). For all probe sequences, median fluorescence intensities were determined for the eight replicate probes. PBM data are available with the following NCBI GEO accession number: GSE83442.

Motif analysis for both PBM and ChIP-seq data. Motif analysis was performed using the MEME software suite: specifically, MEME (Bailey and Elkan 1994) was used for the motif from PBM data and MEME-ChIP (Machanick and Bailey 2011) was used for the motifs analysis from CLAMP ChIP-seq data. The parameters were set to allow one motif occurrence per input sequence; both forward and reverse sequence orientations were considered; minimal motif length was set to 12-bp; other parameters were set to default values. Motif detection at the ChIP-seq peaks as well as in random and scrambled sequences was performed using the FIMO tool (Grant, Bailey, and Noble 2011) with default parameters. Motif logos were created using Weblogo (Crooks et al. 2004) .

DNA shape analysis

The distribution of the logarithmic signal intensity of the PBM probes indicated two mixed Gaussian distributions, with the Gaussian distribution of higher mean likely representing CLAMP binding. Mixture Gaussian modeling was used to separate out strongly bound probes. First, a two-Gaussian model was fitted to the logarithmic signal intensity of the probes. Then probes that were assigned to the higher Gaussian with confidence of ≥ 0.95 were defined as strongly bound probes. Probes that ambiguously mapped to the *Drosophila* genome were removed. The remaining probes were then fed as input to the MEME suite for motif discovery (Bailey et al. 2009) using default settings. Following this step, the probes were aligned according to the motif generated by MEME.

Using the aligned PBM probes, L2-regularized multiple linear regression (MLR) was used to model the DNA binding specificity of CLAMP. Mononucleotide (1mer), dinucleotide (2mer), and trinucleotide (3mer) features were extracted from the sequences, and the four DNA shape features minor groove width, propeller twist, roll, and helix twist were derived from our DNA shape method (T. Zhou et al. 2013). These

features were encoded for different models as previously described (Tianyin Zhou et al. 2015). These features were then used as the predictor variables and the signal intensity of the probes as the response variable for the L2-regularized MLR. The coefficient of determination R^2 between predicted and experimental probe intensities was calculated as the performance measure of the models using 10-fold cross validation.

Protein expression, DNA labeling, and methods for EMSA

Protein expression from E.coli for EMSAs. The pMAL-c5X Vector containing the MBP (Maltose Binding Tag) was used to express MBP fusion CLAMP 6 zinc finger protein using the Nde1 and Sbf1 restriction enzyme sites. For large-scale *E. coli* expression, the CLAMP cDNA was codon optimized and ordered (Life Technologies). Additionally, for optimal expression amino acids after the zinc fingers that had low-predicted structure (a.a. 528-561) were not include in these constructs. The following primers were used for cloning the 6 zinc finger domain of CLAMP: Forward primer 5'-TACGATCATATGTCCGGTAGCGTTAAACAGAGTGTTACCGT-3' and reverse primer 5'-TACGATCCTGCAGGCTATTACAGACCGCCAATAATAACTTCTT-3'. The pMAL protein fusion and purification system from NEB (Catalog # E8200S) was used.

CLAMP protein was overexpressed in strain *E. coli* Star (Invitrogen) in LB media containing 50 µg/ml ampicillin and 90 µM ZnCl₂. Cultures were grown to mid-log phase at 310 K and expression was induced with 0.4 M IPTG. The cells were then further grown for 18 hours at 293 K and harvested by centrifugation. Cells were lysed by sonication on ice in a buffer containing 20mM Tris pH 7.5, 90 µM ZnCl₂, 1 mM MgCl₂ and 90mM KCl. The soluble lysate was then loaded onto an amylose column. Unbound protein was removed with a wash buffer containing 20 mM Tris pH 8.0, 100 mM NaCl,

and 90 μM ZnCl_2 . MBP-CLAMP was then eluted in the same buffer containing 10 mM Maltose. 5% glycerol was added and protein aliquots were stored at 193 K.

DNA labeling for electrophoretic mobility shift assay. The CLAMP binding sequences were end labeled with Biotin using Biotin 3' End Labeling Kit (Thermo Scientific Pierce Cat#89818) with the following adaptations. Each strand of DNA was Biotin labeled independently then annealed by heating to 95°C for 10 mins then removing from heat to slowly cool to room temperature (20°C). Single strand DNA digestion was performed to remove un-annealed DNA for 30 mins at 37°C with Mung Bean Nuclease 2,000 U/ μL (New England Biolabs Cat#M0250S) followed by Chloroform:isoamyl alcohol (Sigma Aldrich Cat# C0549-1PT) extraction as recommended in the Biotin Labeling Kit protocol. Probe sequences are listed in Table 2.1.

Table 2.1 EMSA probe sequences.

Probe name	Sequence (5' to 3')
High affinity (forward)	CGGGCCAGCTGCTGTCTCGCTCGCACCCGCACCGCT
High affinity (reverse)	AGCGGTGCGGGTGCGAGCGAGACAGCAGCTGGCCCG
Constant flank (forward)	CTACTATAGCAATGGGAGCGAGAAGTATCAGTCAGT
Constant flank (reverse)	ACTGACTGATACTTCTCGCTCCCATTGCTATAGTAG
Four-repeat (forward)	GTATTGTTTATTTATGTAATTATAATGAGAGAGATATTGTTTATTT ATTAATGTATAATT
Four-repeat (reverse)	AATTATACATTAATAAATAAACAATATCTCTCTCATTATAATTACA TAAATAAACAATAC
Eight-repeat (forward)	TGTTTATTTTATGTAATTATAATGAGAGAGAGAGAGAGATATTGTT TATTTATTAATGTAT

Eight-repeat (reverse)	ATACATTAATAAATAAACAATATCTCTCTCTCTCTCATTATAAT TACATAAATAAACA
Ten-repeat (forward)	TTTATTTTATGTAATTATAATGAGAGAGAGAGAGAGAGATATTG TTTATTTATTAATGT
Ten-repeat (reverse)	ACATTAATAAATAAACAATATCTCTCTCTCTCTCTCTCATTATA ATTACATAAATAAA
Fifteen-repeat (forward)	TTATGTAATTATAATGAGAGAGAGAGAGAGAGAGAGAGAGAGA GATATTGTTTATTATT
Fifteen-repeat (reverse)	AATAAATAAACAATATCTCTCTCTCTCTCTCTCTCTCTCTCTCA TTATAATTACATAA

Electrophoretic mobility shift assay. EMSAs were performed using the Chemiluminescent Nucleic Acid Detection Module (Thermo Scientific Cat#89880)/LightShift Chemiluminescent EMSA Kit (Thermo Scientific Cat#20148). DNA binding conditions for CLAMP protein were optimized with 1 µg/µL Poly (dI•dC, Cat#20148E) and Elution Buffer: 5mM Tris-HCL, pH 8.5 (AE Macherey-Nagel Cat#740588.10). Control MBP protein was used from NEB and diluted to equal the molar concentration of CLAMP.

ChIP-seq sample preparation and data analysis

Chromatin preparation. The following protocol is modified from (Larschan et al. 2007a). Chromatin was prepared from 150 sexed third instar larvae ground into a powder with a mortar and pestle while frozen in liquid nitrogen. Larvae were homogenized in a 40 mL Dounce homogenizer in 40 mLs of PBS containing a protease inhibitor cocktail (PI) (pancrease-extract 0.002 mg/ml, thermolysin 0.0005 mg/ml, chymotrypsin 0.002 mg/ml, trypsin 0.02 mg/ml, papain 0.33 mg/ml) (Roche). Next, the extract was crosslinked with 1% formaldehyde. Samples were rotated for 20 min at room temperature followed by quenching with 2 mLs 2.5 M glycine for 5 minutes at room temperature. The extract was chilled on ice and centrifuged for 5 minutes at 1500 g at 4°C. The pellet was resuspended in 40 mLs PBS containing 1x PI and 0.2 mM PMSF followed by

centrifugation for 5 minutes at 1500 g at 4°C. Centrifugation and resuspension were repeated as described above with the following buffers: 1) 10 mLs Wash A (10 mM HEPES pH 7.6, 10 mM EDTA pH 8.0, 0.5 mM EGTA pH 8.0, 0.25% Triton-X 100, 1x PI, 0.2 mM PMSF); 2) 10 mLs Wash B (10 mM HEPES pH 7.6, 10 mM NaCl, 10 mM EDTA pH 8.0, 0.5 mM EGTA pH 8.0, 0.25% Triton-X 100, 1x PI, 0.2 mM PMSF); 3) 10 mLs TE Wash (10 mM Tris-HCL pH 8.0, 1 mM EDTA pH 8.0, 0.01% SDS, 1x PI, 0.2 mM PMSF); 4) 10 mLs TE Wash with SDS (10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0, 1% SDS, 1x PI, 0.2 mM PMSF), and 5) 2 washes with 10 mLs TE Wash. Next, the pellet was resuspended in 2 mLs Pre-RIPA (0.1% SDS, 10 mM Tris-HCL pH 8.0, Larval paste samples were sonicated on ice using a Bioruptor microtip sonicator for 6 30-second cycles at 20% amplitude, with one minute off in between cycles to prevent the sample from overheating. Samples were then cleaned up and run on a gel to validate sonication. To clean up the sonicated chromatin, 25 µL of sonicated sample was added to tubes containing 175 µL elution buffer (1% SDS, 0.1 M Sodium Bicarbonate), and crosslinks were reversed overnight at 65°C. 200 µL of TE buffer was added to each tube (0.01 M Tris, pH 9.0, 1 µM EDTA) along with 8 µL RNase A, and the mixture was incubated at 37°C for 2 hours. 4 µL Proteinase K was added per tube and samples were incubated at 55°C for two hours.

DNA was extracted using a Phenol-Chloroform-Isoamyl Alcohol (25:24:1) extraction using phase-lock tubes. The DNA was then precipitated with 5M NaCl, glycogen and ethanol at -80°C for 1 hour and resuspended in 10 mL water. Precipitate was run on a 1% gel at 120 V for 1 hour and post-stained with Gel Red for 20 minutes. After validating the shearing, 100 µL Triton X-100, 20 µL 10% Sodium Deoxycholate, and 56 µL 5M NaCl were added. Samples were pelleted and the supernatant was saved. 50 µL was set aside as input.

Immunoprecipitation. For immunoprecipitation, 5 μ L CLAMP antibody (Larschan et al. 2012) was added to each sample and rotated at 4°C overnight. 100 μ L Protein A Dynabeads per sample were washed in IP buffer (0.1% SDS, 10 mM Tris-HCl, pH 8.0, 1 mM EDTA, 1% Triton X-100, 0.1% Sodium Deoxycholate, 140 mM NaCl) and added to the IP. The lysate and beads were incubated for 2 hours at 4°C, then the beads were washed in wash buffer (50 mM HEPES-KOH, 0.1 M LiCl, 1 mM EDTA, 1% NP-40, 0.7% Sodium Deoxycholate). Samples were eluted in elution Buffer (1% SDS, 0.1 M Sodium Bicarbonate), and DNA was isolated as described above with a final resuspension into 50 μ L water. DNA libraries were prepared using the Illumina TruSeq Kit, and sequenced on a Illumina HiSeq with 100-bp paired end sequencing.

ChIP-seq data analysis. Reads were aligned to the publically available assemblies for *D. melanogaster* (version dm3) and the *D. miranda* genome (Christopher E Ellison and Bachtrog 2015) allowing only unique alignments. The numbers of aligned reads from each sample are shown below. The positions with anomalously high tag counts were removed for the consideration as potential amplification artifacts (Kharchenko, Tolstorukov, and Park 2008). Tag counts were normalized by the corresponding library sizes and enrichments were calculated as ratios of normalized tag counts in ChIP and input samples. Peaks in the ChIP-Seq tag distribution were called using the SPP software package with default parameters. Overlapping peaks of two replicates were used in the analysis. Data are available with the following NCBI GEO accession number: GSE83435.

Alignment results.

D. melanogaster replicate 1 input: 25208995 reads were processed, 17035648 (67.58%) were aligned.

D. melanogaster replicate 1 ChIP: 10298719 reads were processed, 5882185 (57.12%) were aligned.

D. melanogaster replicate 2 input: 16543983 reads were processed, 12299524 (74.34%) were aligned.

D. melanogaster replicate 2 ChIP: 5639720 reads were processed, 267213 (47.38%) were aligned.

D. miranda replicate 1 input: 9544768 reads were processed, 3234724 (33.89%) were aligned.

D. miranda replicate 1 ChIP: 8751503 reads were processed, 3149806 (35.99%) were aligned.

D. miranda replicate 2 input: 4782344 reads were processed, 2123634 (44.41%) were aligned.

D. miranda replicate 2 ChIP: 15175635 reads were processed, 5512100 (36.32%) were aligned.

Analysis of CLAMP peak clustering

CLAMP binding sites (ChIP-seq peaks from Soruco et al. 2013) were considered in this analysis. We measured the distance of each binding site to its neighbors and calculated average distance up to nth neighbor. We did the same calculation for randomized peaks and compared distance of CLAMP binding sites to the median of 10 randomized sets. We considered binding sites are around CES if they are within the median distance value of chromosome X.

Motif representation and comparison

Nucleotide frequency in the motif is shown in log scale as 'bits'; the maximum value is $\log_2 4 = 2$ bits, where 4 is the number of nucleotide types. The significance of the motif is represented with an *E*-value score calculated by MEME. The *E*-value is an estimate of the expected number of motifs with the given log likelihood ratio (or higher), and with the same width and site count, that one would find in a similarly sized set of random sequences.

To compare motifs, we calculated the Euclidean distance of frequencies which is defined as the-root-mean-square-difference of nucleotide frequencies for each position.

Analysis of motif density within genomes

To determine CLAMP binding site motif density on chromosomes, sequences that match the motif were identified on chromosomes and exact matches were counted as hits. For repeat density analysis, repeats with different lengths were identified on chromosomes and the number of hits were normalized by the length of each chromosome.

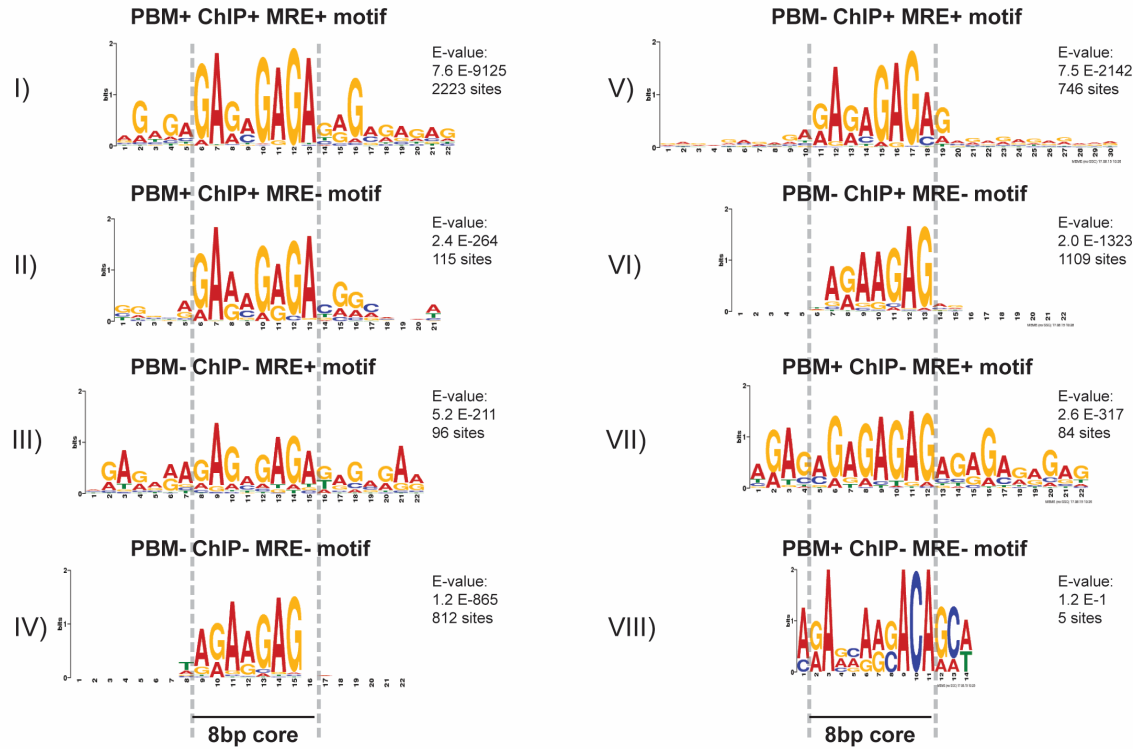
Acknowledgments

We thank Mitzi Kuroda for critical reading of this manuscript and Doris Bachtrog for the *D. miranda* fly line.

Supporting Information

Supplemental Figures

A



B

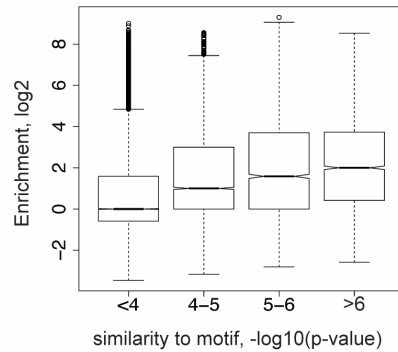


Figure S2.1

A) 'PBM+ChIP+MRE+', 'PBM+ChIP+MRE-', 'PBM-ChIP-MRE+', 'PBM-ChIP-MRE-', 'PBM-ChIP+MRE+', 'PBM-ChIP+MRE-', 'PBM+ChIP-MRE+' and 'PBM+ChIP-MRE-' groups are given with their *E*-values and number of sites. **B)** CLAMP ChIP-seq enrichment is shown for the sequences similar to the custom PBM CLAMP motif found. Similarity of the *in vivo* sequences to the PBM motif was calculated using the FIMO tool.

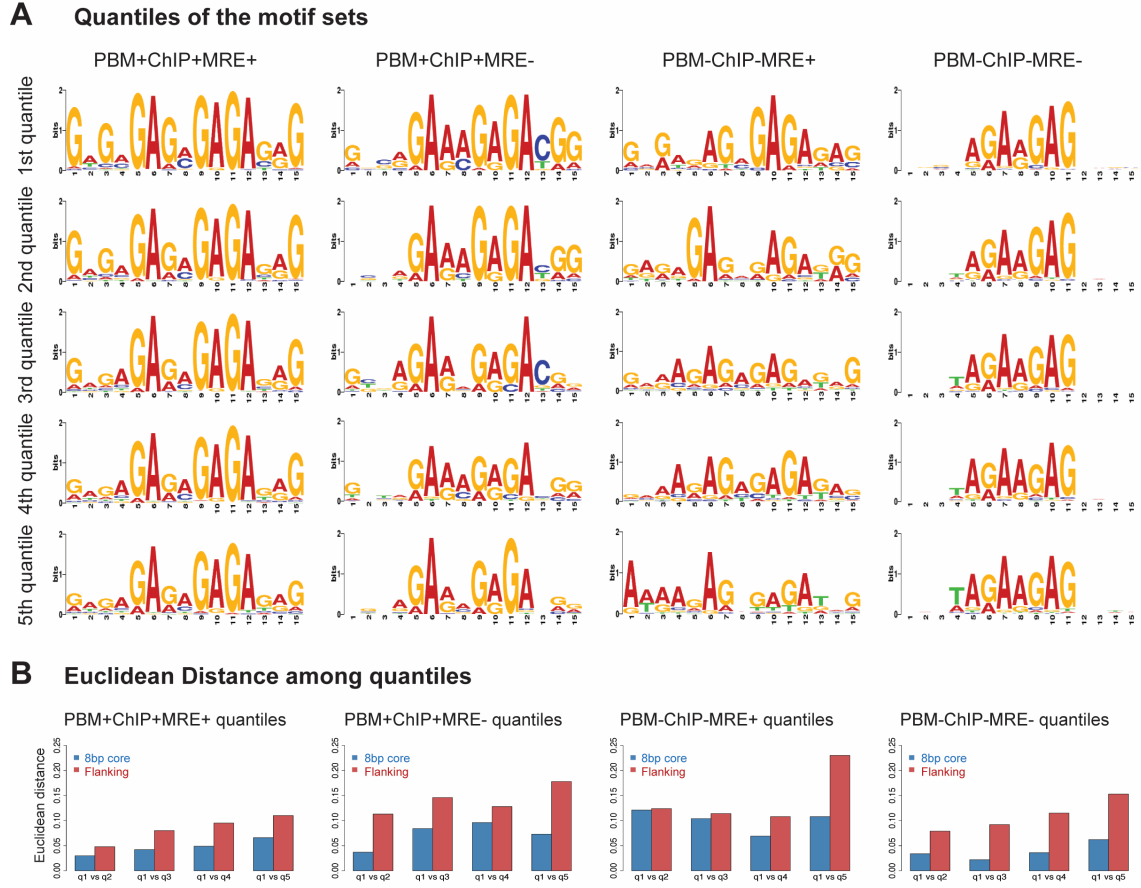


Figure S2.2

A) Quantiles of 'PBM+ChIP+MRE+', 'PBM+ChIP+MRE-', 'PBM-ChIP-MRE+' and 'PBM-ChIP-MRE-' groups are given. Groups are divided into five different quantiles based on binding intensity scores and motifs were obtained from the sequences in each quantile. **B)** Euclidean distances between quantiles are shown for 8-bp core and flanking parts separately. Average Euclidean distances were calculated for each part of the motif.

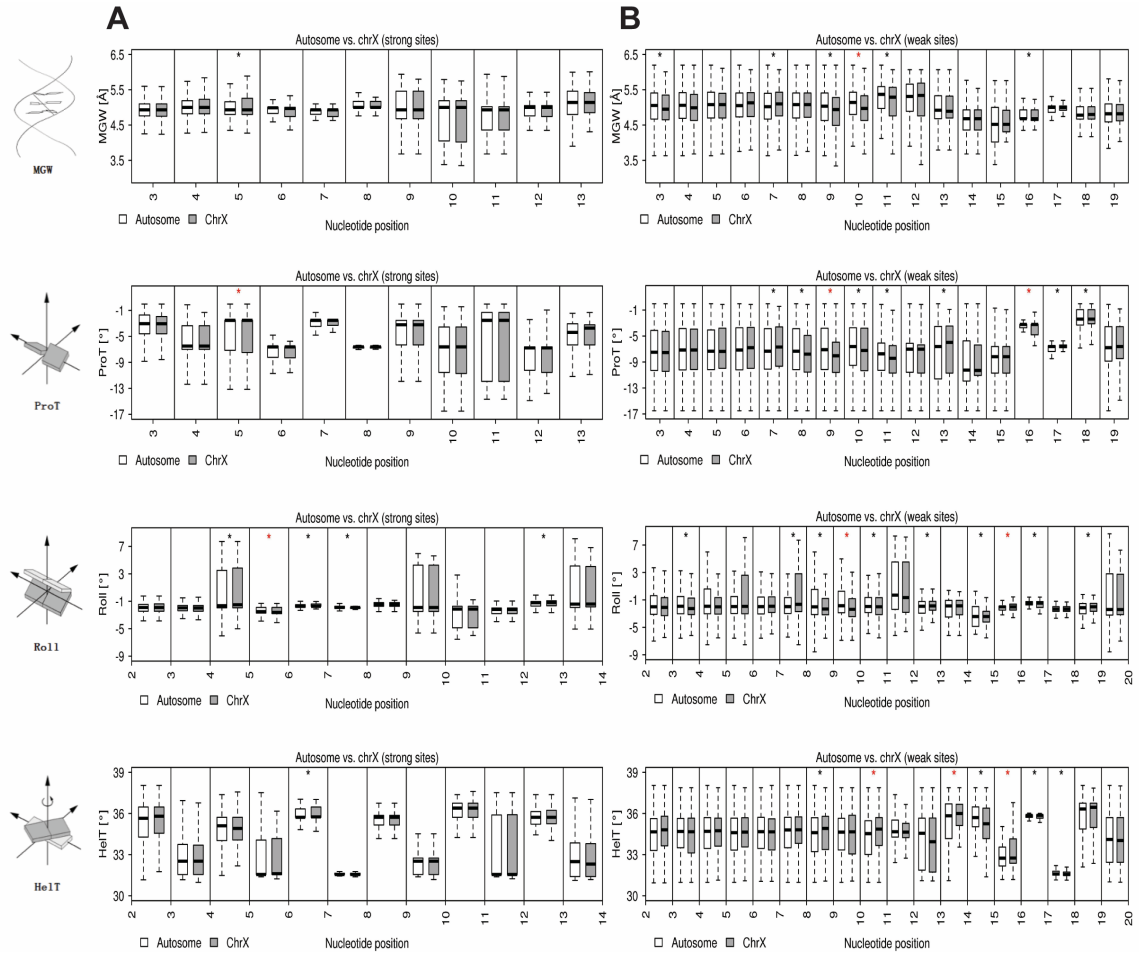


Figure S2.3

Kolmogorov–Smirnov test (KS-test) between autosomes and chromosome X binding site DNA shape profiles in the strong (A) and the weak (B) groups. Black asterisks indicate significant different with p -value < 0.05 , and red asterisks with p -value < 0.001 .

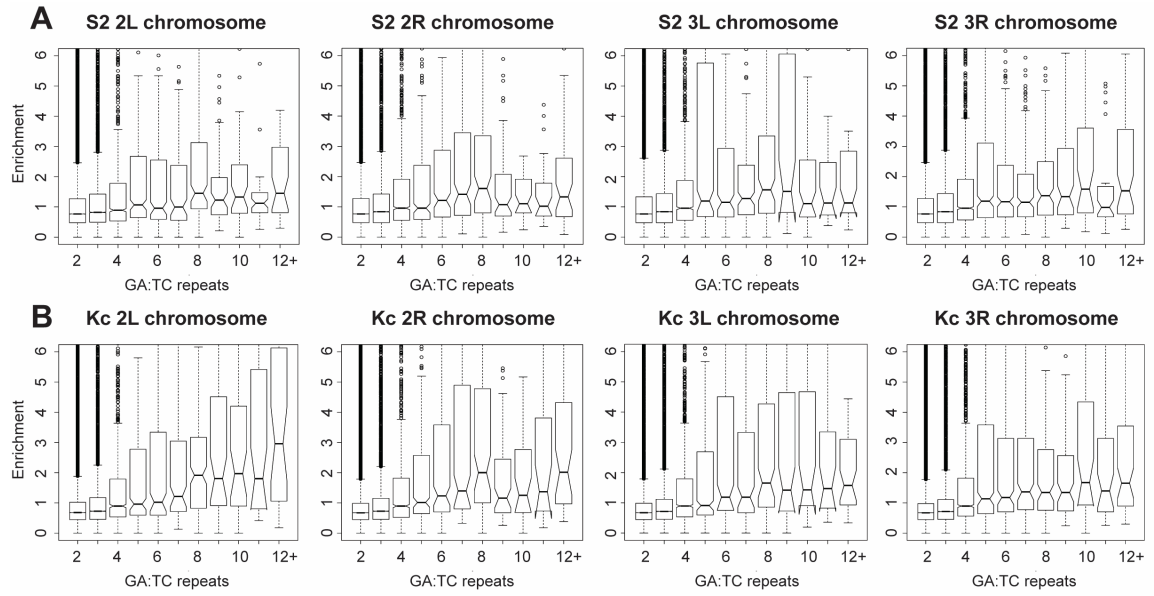


Figure S2.4

CLAMP ChIP-seq enrichment on GA repeats of different lengths is shown for S2 (A) and Kc (B) 2L, 2R, 3L and 3R autosomal arms.

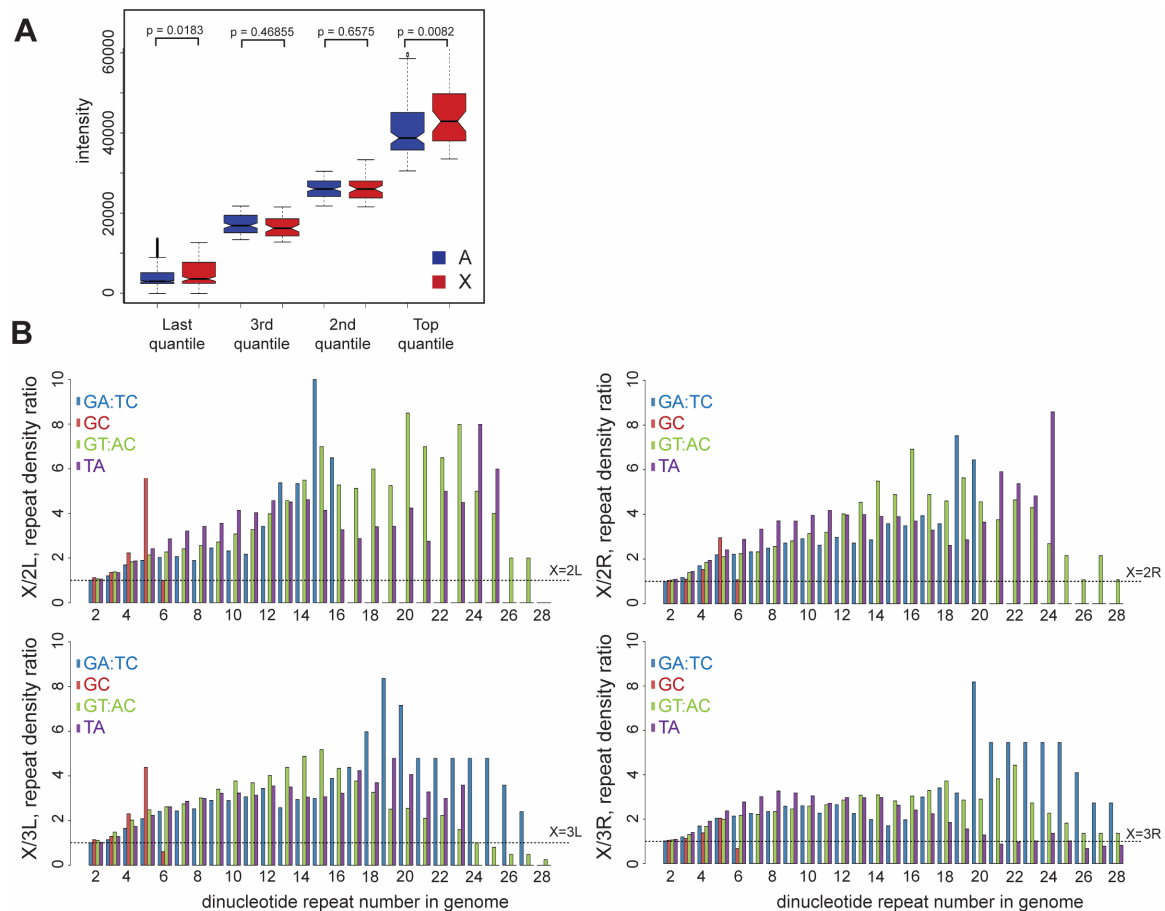


Figure S2.5

A) PBM intensities of probes selected from autosomes and the X-chromosome are plotted for each quantile. p -values calculated via Kolmogorov–Smirnov test are given for each pair. **B)** Density of dinucleotide repeats on the *D. melanogaster* chromosome X is compared with chromosomes 2L, 2R, 3L and 3R individually.

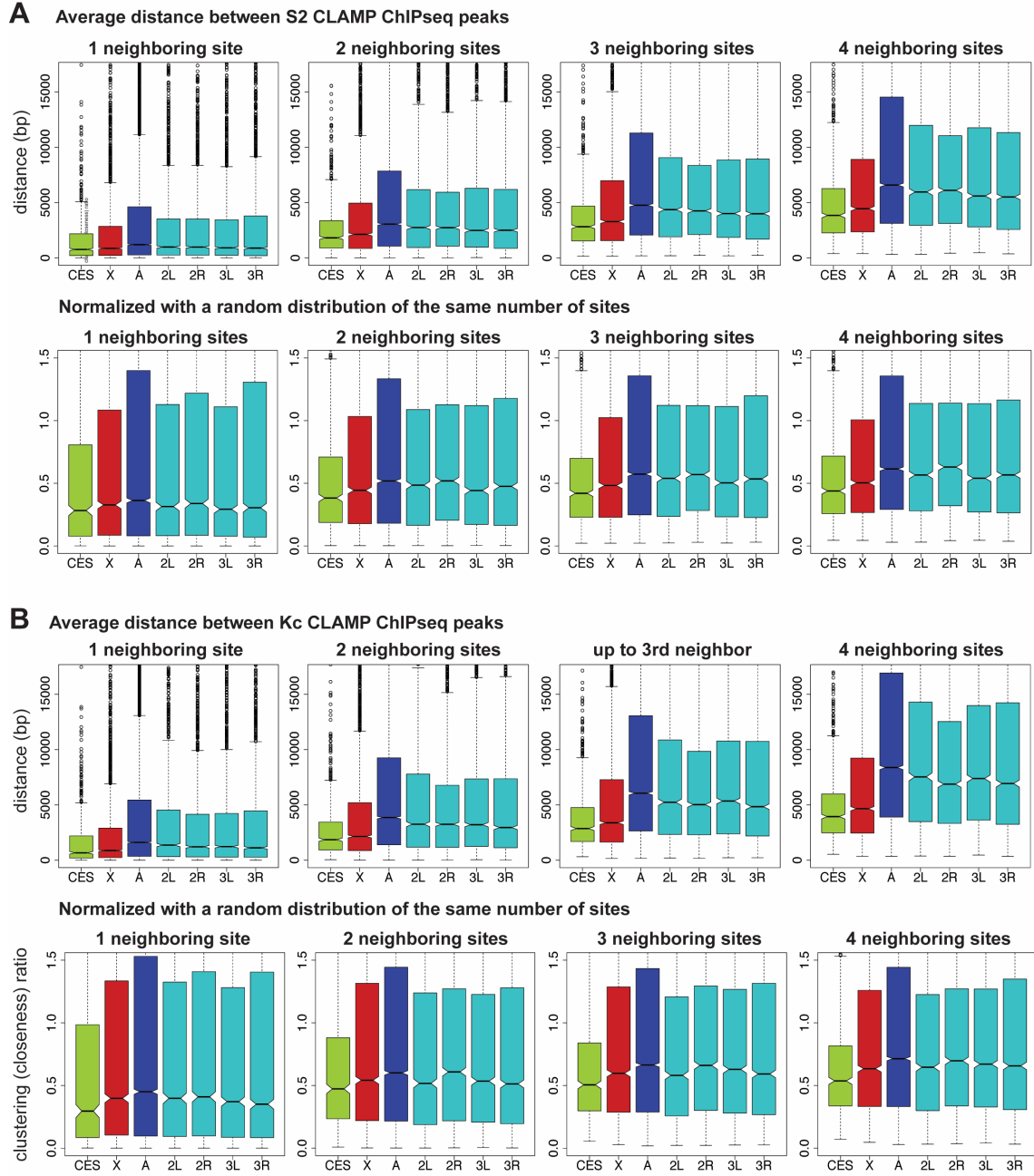


Figure S2.6

A) Average distance between neighboring S2 CLAMP ChIPseq peaks is plotted (top panel). Values are normalized based on a random distribution of peaks (bottom panel). **B)** The same average plots shown for S2 cells in **(A)** are shown for Kc cells.

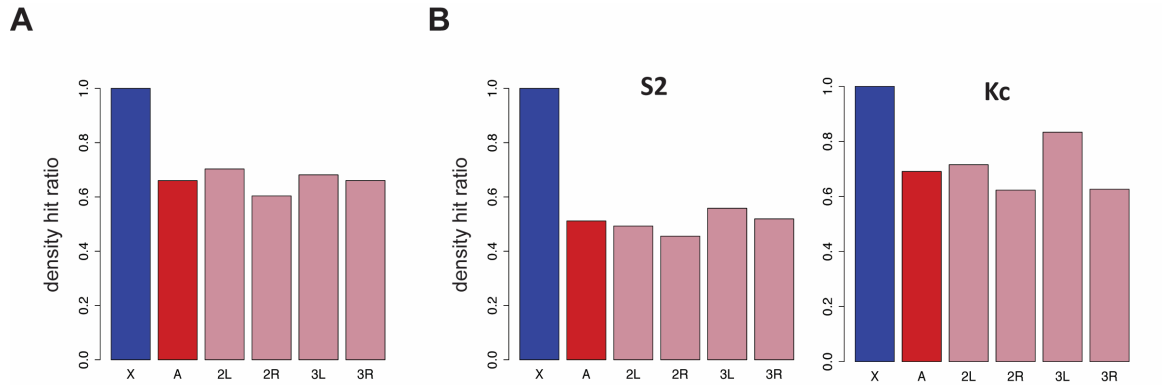


Figure S2.7

A) Ratio of CLAMP motif hit density in the gene body (TSS+250bp-TTS) to 5' end (TSS+/-250-bp) is shown for each chromosomal arm; values are normalized to the X-chromosome value. **B)** The same analysis conducted for part A is shown for S2 and Kc CLAMP ChIP-seq peaks instead of motifs.

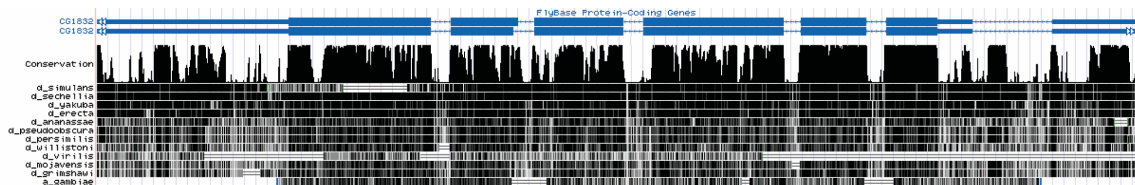


Figure S2.8

A schematic of the CLAMP (CG1832) gene from *D. melanogaster* is shown in blue. Conservation of the CLAMP gene sequence compared to orthologues in other *Drosophilids* and the mosquito (*A. gambiae*) is shown below (UCSC Genome Browser).

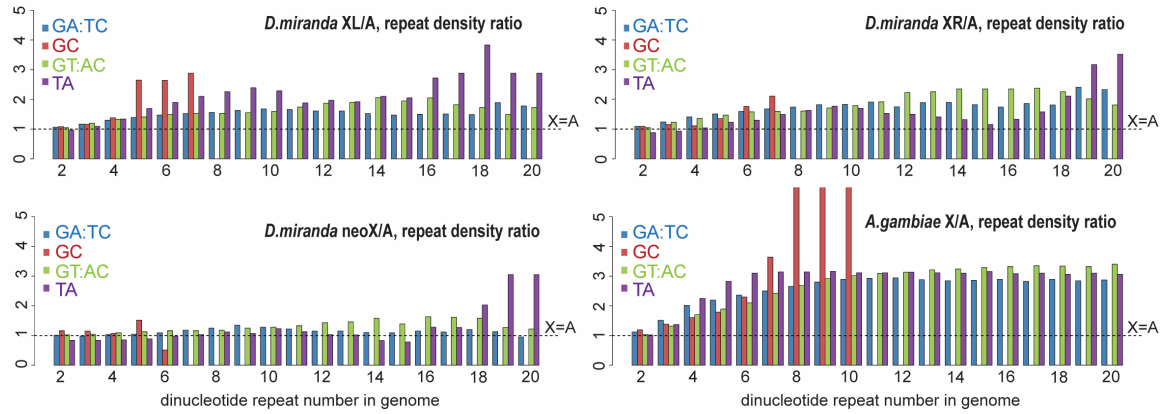


Figure S2.10

Density ratio of neoX (Muller-C), XL (Muller-A), and XR (Muller-D) chromosomes to autosomes in *D. miranda* and X (Muller-A) chromosome to autosome in *A. gambiae* are shown for different types of dinucleotide repeats.



Figure S2.11

Density ratio of XL (Muller-A), XR (Muller-D) and neoX (Muller-C) chromosomes to chromosomes 2 (Muller-B), 4 (Muller-E) and 5 (Muller-F) in *D. miranda* are shown for different repeats.

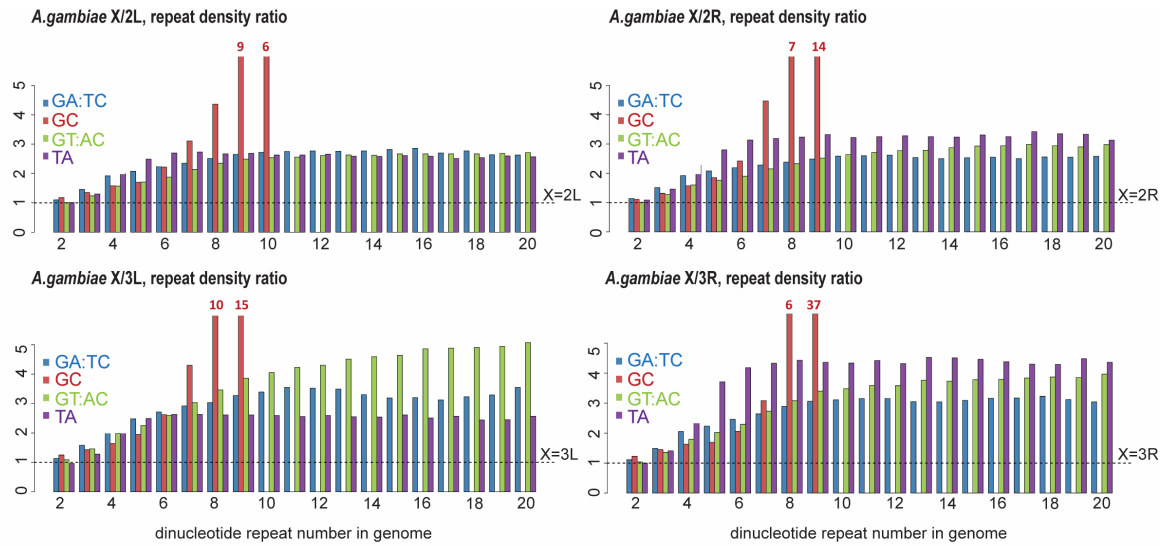


Figure S2.12

Density ratio of X (Muller-A) chromosome to chromosomes 2L (Muller-D), 2R (Muller-E), 3L (Muller-C) and 3R (Muller-B) in *A. gambiae* are shown for different repeats.

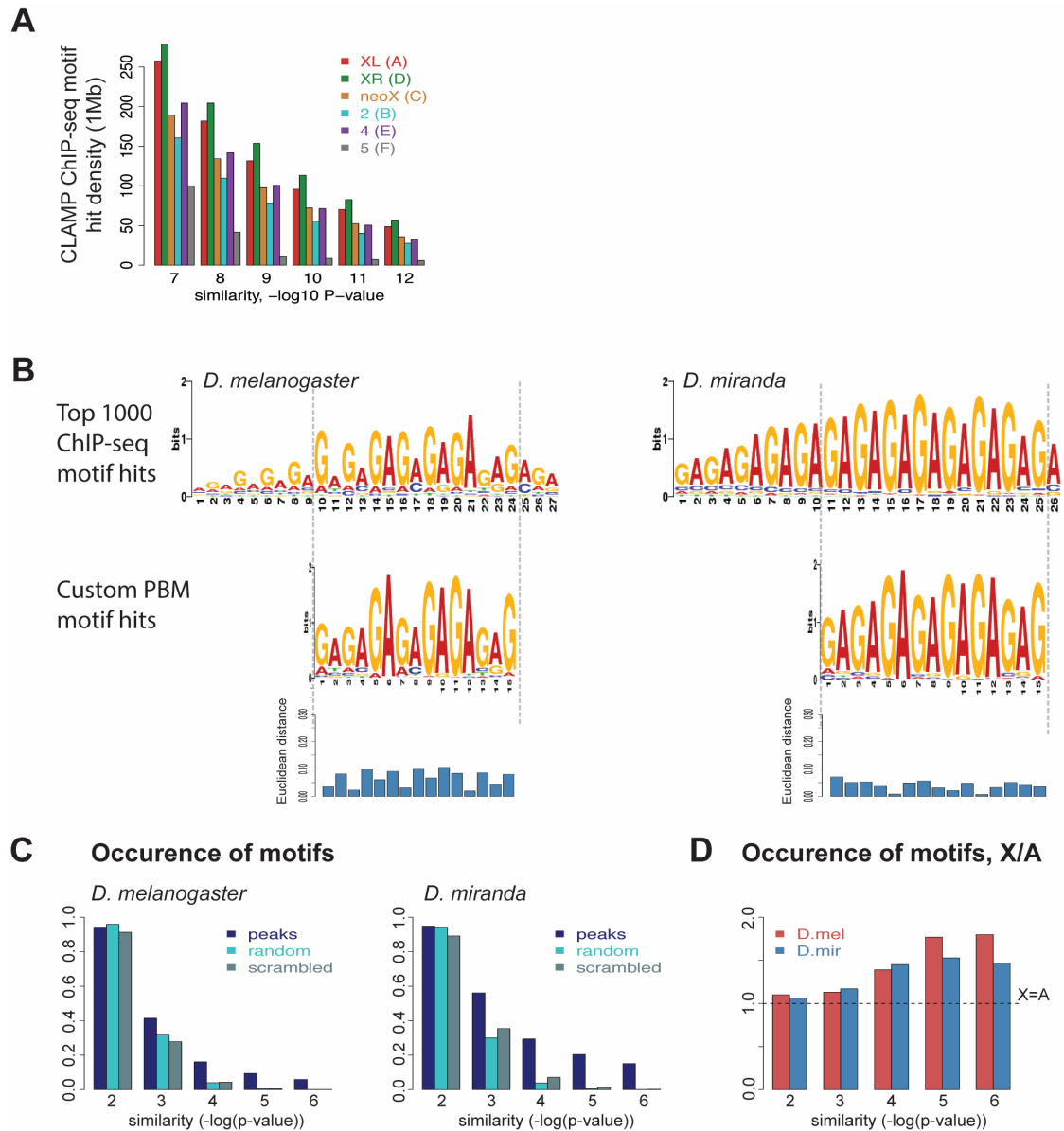


Figure S2.13

A) CLAMP ChIP-seq hit density per 1Mb is shown for each chromosome of *D. miranda*. Similarity of the genomic sequence under each peak to the ChIP-seq motif (Figure 2.6C) was calculated using the FIMO tool. **B)** The motif from the top 1000 ChIP-seq peaks (top panel) was compared with the custom PBM motif. Sequences that include the custom PBM motif are presented as the custom PBM motif hits (middle panel). Euclidean distances between the top and middle motifs are given for each position within the PWM (bottom panel). **C)** As the similarity to an *in vivo* motif increases (lower p-value), the occurrence of the motifs increases in called peaks vs. randomized and scrambled peaks for both *D.melanogaster* and *D. miranda*. **D)** Motif occurrence presented in part C is shown as the ratio of X-chromosome to autosomes.

```

Mel_MSL2_CXC  525  CRCGISGSSNTLTTCRNSRCPCYKSYNSCAGCHCVCC 583
                |.....|          |||||.:.|||.|.|.|.
Mir_MSL2_CXC  364  -----TPAARR-----CYKSGNTCANCRCFGC 385

```

Figure S2.14

Amino acid pairwise alignments using EMBOSS Needle (Goujon et al. 2010) of the *D. melanogaster* and *D. miranda* MSL2 CXC domain. Symbols indicated conserved residues: | = fully conserved, : = similar, . = mismatch.

Supplemental Tables

Table S2.1 The PBM probes. A description of the PBM probe classes and the number of probes in each class.

Probe ID	# of Probes	# Unique Sequences	# of Replicates	Brief Desc.
(A) PROBES EXAMINING SINGLE pMRE/MRE SITES				
mreP_gen_nX_oY_rZ	27216	3402	4	ChIP, +MRE/+pMRE, Genomic flanks
mreN_gen_nX_oY_rZ	10168	1271	4	ChIP, -MRE/+pMRE, Genomic flanks
negP_gen_nX_oX_rX	1664	208	4	(Neg Control) No ChIP, +MRE/+pMRE, Genomic flanks
negN_gen_nX_oX_rX	6576	822	4	(Neg Control) No ChIP, -MRE/+pMRE, Genomic flanks
cons_nX_oY_rZ	2816	352	4	Aggregate probeset of the 8-11bp pMREs with Constant Flanks
(B) PROBES TILING ACROSS DIFFERENT GENOMIC REGIONS				
rox1_n0_pX_oX_rX	1648	206	4	Wide Tiling (1000bp windows), High Affinity Region
5c2_n0_pX_oX_rX	1632	204	4	Wide Tiling (1000bp windows), High Affinity Region
rox2_n0_pX_oX_rX	1648	206	4	Wide Tiling (1000bp windows), High Affinity Region
11d1_n0_pX_oX_rX	1648	206	4	Wide Tiling (1000bp windows), High Affinity Region
15a8_n0_pX_oX_rX	1648	206	4	Wide Tiling (1000bp windows), High Affinity Region
til_nochp_nX_pX_oX_rX	16456	2057	4	Wide Tiling (1000bp windows), No CLAMP Chip, +pMRE
til_chpmre_nX_pX_oX_rX	19568	2446	4	Wide Tiling (1000bp windows) CLAMP Chip, +MRE/pMRE
til_chpno_nX_pX_oX_rX	16424	2053	4	Wide Tiling (1000bp windows) CLAMP Chip, -MRE/-pMRE
ntil_nochp_nX_pX_oX_rX	14104	1763	4	Narrow Tiling (100bp windows), CLAMP Chip, -MRE/-pMRE
ntil_chp_nX_pX_oX_rX	33720	4215	4	Narrow Tiling (100bp windows), CLAMP Chip, +MRE/+pMRE
(C) CONTROL PROBES				
pX_pX_nX_oX_rX	540	90	6	Pairs of Sites
mre1_PO_pX_wX_oX_rX	336	28	6	Position Dependence (PWM1:AGAGAGAGA)
mre2_PO_pX_wX_oX_rX	348	29	6	Position Dependence (PWM2:TAGAAGAG)
extra_nX_oX_rX	570	6	48	

Table S2.2 Kolmogorov–Smirnov test was applied to the PBM intensities of probes.

	ChIP+ MRE+	ChIP+ MRE-	ChIP- MRE+	ChIP- MRE-	PBM+ ChIP+ MRE+	PBM+ ChIP+ MRE-	PBM+ ChIP+ MRE+	PBM+ ChIP+ MRE-
ChIP+ MRE+								
ChIP+ MRE-	< 2.2e-16							
ChIP- MRE+	2.51e-13	< 2.2e-16						
ChIP- MRE-	< 2.2e-16	< 2.2e-16	< 2.2e-16					
PBM+ ChIP+ MRE+	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16				
PBM+ ChIP+ MRE-	1.29e-06	< 2.2e-16	8.31e-10	< 2.2e-16	1.92e-13			
PBM- ChIP- MRE+	< 2.2e-16	2.87e-12	1.73e-10	< 2.2e-16	< 2.2e-16	< 2.2e-16		
PBM- ChIP- MRE-	< 2.2e-16	< 2.2e-16	< 2.2e-16	1	< 2.2e-16	< 2.2e-16	< 2.2e-16	

Categories tested include 'ChIP+MRE+', 'ChIP+MRE-', 'ChIP-MRE+', 'ChIP-MRE-', 'PBM+ChIP+MRE+', 'PBM+ChIP+MRE-', 'PBM-ChIP-MRE+', and 'PBM-ChIP-MRE-', where 'ChIP+' indicates CLAMP binding *in vivo*, 'MRE+' indicates the sequence is with MSL recognition element and 'PBM+' indicates CLAMP binding *in vitro*. Values show the p-values.

Table S2.3 Kolmogorov–Smirnov test was applied to the enrichments on the sequences with different similarities to the motif.

	$-\log_{10}(p) < 4$	$-\log_{10}(p): 4-5$	$-\log_{10}(p): 5-6$	$-\log_{10}(p) > 6$
$-\log_{10}(p) < 4$				
$-\log_{10}(p): 4-5$	$< 2.2e-16$			
$-\log_{10}(p): 5-6$	$< 2.2e-16$	$< 2.2e-16$		
$-\log_{10}(p) > 6$	$< 2.2e-16$	$< 2.2e-16$	$8.4e-4$	

Values show the Kolmogorov–Smirnov test p-values. Similarities to the motif were calculated with the FIMO tool, and categories are $-\log_{10}(p\text{-value})$ is smaller than 4, between 4 and 5, between 5 and 6, and larger than 6.

Table S2.4 Euclidean distance between motifs.

Comparison	8-bp core	Flanking
PBM+ChIP+MRE+ vs PBM+ChIP+MRE-	0.080	0.212
PBM+ChIP+MRE+ vs PBM-ChIP-MRE+	0.091	0.117
PBM+ChIP+MRE+ vs PBM-ChIP-MRE-	0.268	0.264
PBM+ChIP+MRE- vs PBM-ChIP-MRE+	0.119	0.196
PBM+ChIP+MRE- vs PBM-ChIP-MRE-	0.235	0.200
PBM-ChIP-MRE+ vs PBM-ChIP-MRE-	0.247	0.210

The Euclidean distance between all classes of motifs is shown. The Euclidean distance between the 8-bp cores and the flanking sequences are shown separately.

Table S2.5 Kolmogorov-Smirnov test to compare binding to different flanking sequence categories.

Kolmogorov–Smirnov test (p-values)	8bp + matched endogenous flank	8bp + unmatched endogenous flank	8bp + unmatched synthetic flank	without 8bp core part	8bp + matched endogenous flank, 4ZF
8bp + matched endogenous flank					
8bp + unmatched endogenous flank	< 2.2e-16				
8bp + unmatched synthetic flank	< 2.2e-16	< 2.2e-16			
without 8bp core part	< 2.2e-16	< 2.2e-16	7.14e-11		
8bp + matched endogenous flank, 4ZF	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	

Kolmogorov–Smirnov test was applied to the PBM intensities of probes with 8-bp core part and matched endogenous flank, with 8-bp core part and unmatched endogenous flank, with 8-bp core part and unmatched synthetic flank, without 8-bp core part, and with 8-bp core part and matched endogenous flank for 4 zinc finger protein. Values show the p-values.

Table S2.6 The count of probes that contain each number of GA-repeats.

GA repeats	# of probes
1	65
2	1404
3	516
4	365
5	161
6	107
7	79
8	46
9	25
10	10
11+	17

Table S2.7 Number of GA-repeats in *D. melanogaster* autosomes (A) and the X-chromosome (X).

	A	X
GA*2	572439	119767
GA*3	34825	7760
GA*4	3703	1101
GA*5	877	330
GA*6	465	188
GA*7	309	146
GA*8	267	99
GA*9	147	82
GA*10	91	58
GA*11	77	28
GA*12+	106	68

Repeats are not overlapping, i.e. the repeats are assigned by the longest length. For the two replicates used in Figure 2.3D, these numbers are doubled.

Table S2.8 Number of GA repeats in *D. melanogaster* chromosomes and individual chromosomal arms. Repeats with different lengths may be overlapping.

	X	A	2L	2R	3L	3R	4
GA*2	139219	649229	137531	152227	166567	186201	6703
GA*3	11273	44768	9384	10376	11774	12748	486
GA*4	2552	7065	1511	1610	1849	2063	32
GA*5	1191	2696	627	583	684	797	5
GA*6	752	1590	372	366	373	478	1
GA*7	517	1054	249	239	254	312	0
GA*8	352	710	186	152	167	205	0
GA*9	248	432	101	98	102	131	0
GA*10	160	278	69	59	66	84	0
GA*11	100	186	46	41	39	60	0
GA*12	72	109	21	26	25	37	0
GA*13	43	71	8	17	20	26	0
GA*14	32	53	6	12	13	22	0
GA*15	20	32	2	6	8	16	0
GA*16	13	19	2	4	4	9	0
GA*17	11	11	0	3	3	5	0
GA*18	10	9	0	3	2	4	0
GA*19	7	5	0	1	1	3	0
GA*20	6	3	0	1	1	1	0
GA*21	4	2	0	0	1	1	0
GA*22	4	2	0	0	1	1	0
GA*23	4	2	0	0	1	1	0
GA*24	4	2	0	0	1	1	0
GA*25	4	2	0	0	1	1	0
GA*26	3	2	0	0	1	1	0
GA*27	2	2	0	0	1	1	0
GA*28	2	1	0	0	0	1	0

Table S2.9 Kolmogorov–Smirnov test was applied to the average distance values between S2 CLAMP ChIP-seq peaks.

		1st neighbor			up to 2nd neighbor			up to 3rd neighbor			up to 4th neighbor		
		CES	X	A	CES	X	A	CES	X	A	CES	X	A
1st neighbor	CES												
	X	1.64 e-03											
	A	4.54 e-03	< 2.2 e-16										
up to 2nd neighbor	CES	1.62 e-09	< 2.2 e-16	< 2.2 e-16									
	X	1.12 e-11	< 2.2 e-16	< 2.2 e-16	5.72 e-06								
	A	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16							
up to 3rd neighbor	CES	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	4.86 e-11	< 2.2 e-16	< 2.2 e-16						
	X	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	2.86 e-11					
	A	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16				
up to 4th neighbor	CES	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	5.67 e-14	1.37 e-11	< 2.2 e-16			
	X	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	5.86 e-11	8.30 e-13		
	A	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	

Table S2.10 Kolmogorov–Smirnov test was applied to the average distance values between Kc CLAMP ChIP-seq peaks.

		1st neighbor			up to 2nd neighbor			up to 3rd neighbor			up to 4th neighbor		
		CES	X	A	CES	X	A	CES	X	A	CES	X	A
1st neighbor	CES												
	X	1.18 e-01											
	A	9.89 e-06	< 2.2 e-16										
up to 2nd neighbor	CES	1.44 e-15	< 2.2 e-16	< 2.2 e-16									
	X	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	4.26 e-05								
	A	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16							
up to 3rd neighbor	CES	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16						
	X	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	1.08 e-11	< 2.2 e-16	2.75 e-12	1.04 e-11					
	A	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16				
up to 4th neighbor	CES	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	2.33 e-11	5.04 e-12	< 2.2 e-16			
	X	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16		
	A	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	

Table S2.11 Numbers of S2 and KC ChIP-seq peaks used in the average distance calculation are given.

	1st neighbor	up to 2nd neighbor	up to 3rd neighbor	up to 4th neighbor
S2 CES	337	606	864	1055
S2 X	2704	2704	2704	2704
S2 A	11607	11607	11607	11607
Kc CES	248	516	788	1037
Kc X	3220	3220	3220	3220
Kc A	11878	11878	11878	11878

Table S2.12 Kolmogorov–Smirnov test was applied to the normalized distance values between S2 CLAMP ChIP-seq peaks.

		1st neighbor			up to 2nd neighbor			up to 3rd neighbor			up to 4th neighbor		
		CES	X	A	CES	X	A	CES	X	A	CES	X	A
1st neighbor	CES												
	X	2.88 e-03											
	A	1.16 e-02	1.43 e-05										
up to 2nd neighbor	CES	1.84 e-06	4.46 e-14	< 2.2 e-16									
	X	3.73 e-04	< 2.2 e-16	< 2.2 e-16	3.17 e-06								
	A	8.24 e-04	< 2.2 e-16	< 2.2 e-16	1.11 e-15	9.42 e-08							
up to 3rd neighbor	CES	1.51 e-08	< 2.2 e-16	< 2.2 e-16	3.70 e-02	2.83 e-11	< 2.2 e-16						
	X	6.55 e-09	< 2.2 e-16	< 2.2 e-16	5.61 e-07	4.81 e-07	1.05 e-12	9.82 e-12					
	A	4.32 e-09	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	1.98 e-12				
up to 4th neighbor	CES	9.38 e-12	< 2.2 e-16	< 2.2 e-16	3.46 e-05	5.55 e-16	< 2.2 e-16	8.17 e-02	1.48 e-13	< 2.2 e-16			
	X	1.50 e-13	< 2.2 e-16	< 2.2 e-16	7.30 e-07	< 2.2 e-16	< 2.2 e-16	2.74 e-12	9.35 e-04	9.99 e-16	6.10 e-13		
	A	7.49 e-13	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	2.07 e-09	< 2.2 e-16	< 2.2 e-16	

Table S2.13 Kolmogorov–Smirnov test was applied to the normalized distance values between Kc CLAMP ChIP-seq peaks.

		1st neighbor			up to 2nd neighbor			up to 3rd neighbor			up to 4th neighbor		
		CES	X	A	CES	X	A	CES	X	A	CES	X	A
1st neighbor	CES												
	X	1.82 e-01											
	A	1.27 e-01	1.21 e-04										
up to 2nd neighbor	CES	1.36 e-07	4.55 e-15	< 2.2 e-16									
	X	3.01 e-07	< 2.2 e-16	< 2.2 e-16	1.32 e-04								
	A	1.63 e-05	< 2.2 e-16	< 2.2 e-16	1.32 e-07	4.71 e-03							
up to 3rd neighbor	CES	1.17 e-12	< 2.2 e-16	< 2.2 e-16	6.09 e-02	6.93 e-13	< 2.2 e-16						
	X	1.23 e-12	< 2.2 e-16	< 2.2 e-16	2.59 e-04	5.45 e-09	< 2.2 e-16	1.05 e-11					
	A	1.74 e-10	< 2.2 e-16	< 2.2 e-16	9.64 e-08	1.31 e-13	< 2.2 e-16	< 2.2 e-16	1.16 e-05				
up to 4th neighbor	CES	2.83 e-14	< 2.2 e-16	< 2.2 e-16	9.97 e-03	< 2.2 e-16	< 2.2 e-16	4.34 e-01	< 2.2 e-16	< 2.2 e-16			
	X	4.44 e-16	< 2.2 e-16	< 2.2 e-16	8.03 e-05	< 2.2 e-16	< 2.2 e-16	6.49 e-11	1.07 e-04	3.70 e-11	3.33 e-16		
	A	5.37 e-14	< 2.2 e-16	< 2.2 e-16	3.38 e-09	< 2.2 e-16	< 2.2 e-16	< 2.2 e-16	2.43 e-11	1.32 e-08	< 2.2 e-16	9.11 e-09	

Table S2.14 Conservation of CLAMP across species.

Species	FL % Identity	FL % Similarity	DBD % Identity	DBD % Similarity
<i>D. melanogaster</i>	100.0	100.0	100.0	100.0
<i>D. miranda</i>	92.6	95.9	99.4	100.0
<i>A. gambiae</i>	50.5	59.0	89.0	93.9

Conservation of CLAMP relative to *D. melanogaster* was determined for both the full-length protein (FL) and the DNA binding domain (DBD) alone.

Table S2.15 Number of matches to the CLAMP PBM motif per Mb on individual chromosomal arms.

<i>D. melanogaster</i>		<i>D. miranda</i>		<i>A. gambiae</i>	
chr2L	91.6	chr2	387.5	chr2L	409.7
chr2R	88.2	chr4	507.3	chr2R	416.5
chr3L	90.6	chr5	136.0	chr3L	334.1
chr3R	91.4	neoX	478.0	chr3R	372.3
chr4	11.9	chrXL	645.4	chrX	870.0
chrX	202.8	chrXR	729.0		

Table S2.16 Number of GA repeats in *D. miranda* chromosomes. Repeats may be overlapping.

	XL	XR	3	A (2,4,5)	2	4	5
GA*2	92978	130682	83961	252853	130383	117364	5106
GA*3	13816	20057	10967	34014	17372	16151	491
GA*4	4819	7111	3574	10618	5407	5058	153
GA*5	2475	3674	1753	5126	2663	2439	24
GA*6	1637	2416	1147	3207	1645	1542	20
GA*7	1104	1657	809	2083	1049	1023	11
GA*8	749	1138	564	1378	681	689	8
GA*9	519	792	405	917	443	467	7
GA*10	379	562	272	648	301	342	5
GA*11	265	415	183	458	210	243	5
GA*12	193	286	130	344	151	188	5
GA*13	137	218	92	244	102	138	4
GA*14	94	160	64	178	68	107	3
GA*15	66	111	46	129	50	76	3
GA*16	51	81	37	98	38	57	3
GA*17	40	67	28	76	28	46	2
GA*18	29	48	22	56	22	32	2
GA*19	23	40	13	35	11	23	1
GA*20	18	32	9	29	8	21	0
GA*21	16	26	8	24	7	17	0
GA*22	15	24	5	21	7	14	0
GA*23	12	18	3	17	7	10	0
GA*24	10	17	3	15	6	9	0
GA*25	9	15	2	13	5	8	0
GA*26	8	12	2	11	4	7	0
GA*27	6	12	2	7	1	6	0
GA*28	5	9	1	7	1	6	0

Table S2.17 Number of GA repeats in *A. gambiae* chromosomes. Repeats may be overlapping.

	X	A	2L	2R	3L	3R
GA*2	87061	655817	159887	192336	132715	170879
GA*3	14166	79426	19639	23604	15466	20717
GA*4	6356	26736	6686	8331	4955	6764
GA*5	4097	15820	3991	4979	2843	4007
GA*6	3034	10851	2749	3488	1928	2686
GA*7	2273	7687	1954	2514	1341	1878
GA*8	1732	5522	1397	1835	982	1308
GA*9	1343	4048	1024	1360	707	957
GA*10	1063	3110	790	1036	539	745
GA*11	867	2498	637	841	420	600
GA*12	731	2102	534	705	357	506
GA*13	616	1808	451	612	304	441
GA*14	522	1554	381	526	272	375
GA*15	458	1355	329	456	247	323
GA*16	404	1182	286	400	217	279
GA*17	352	1055	264	355	194	242
GA*18	315	921	230	310	168	213
GA*19	270	804	207	267	141	189
GA*20	233	687	179	228	113	167
GA*21	196	594	160	192	102	140
GA*22	168	514	133	171	93	117
GA*23	146	450	116	149	82	103
GA*24	131	399	104	127	75	93
GA*25	117	344	88	110	65	81
GA*26	96	304	76	96	55	77
GA*27	89	258	62	78	52	66
GA*28	74	228	53	68	48	59

Chapter 3 : Differential occupancy of two GA-binding proteins promotes targeting of the *Drosophila* dosage compensation complex to the male X-chromosome

This chapter is currently in preparation for manuscript submission.

Authors of current version are as follows:

Emily G. Kaye*, Matthew Booker*, Alexander Conicella, Nicolas L. Fawzi, Michael Y. Tolstorukov and Erica Larschan. *Co-first authors.

EGK, MB, MT, and EL conceived and designed experiments and data analysis. EGK performed all experiments, with intellectual contribution from AC and NLF for protein expression. MB performed data analysis and generated sequencing analysis figure panels with advisement from MT, EGK, and EL. EGK organized all panels into figures. EGK and EL wrote the chapter with input from MB and AC.

Abstract

Across species, there are transcription factors that bind to similar *cis*-acting sequence elements yet perform different functions. However, little is known about how the relative occupancy of similar transcription factors alters targeting of large transcription complexes. A well-studied model for understanding how transcription complexes are targeted is the *Drosophila* dosage compensation system in which the Male-Specific Lethal (MSL) complex specifically identifies the male X-chromosome to increase transcript levels. We previously demonstrated that the CLAMP zinc-finger protein links MSL complex to the X-chromosome. However, CLAMP is a non-sex specific transcription factor that binds to GA-rich *cis*-elements genome-wide. Another essential, non-sex specific GA-binding factor, GAGA Associated Factor (GAF), also has a role in MSL complex targeting, despite performing many functions throughout the genome. Many of these functions depend on interactions with different co-factors, such as the Negative-Elongation Factor (NELF) for pausing regulation and the Nucleosome Remodeling Factor (NURF) complex for chromatin organization. Therefore, we hypothesized that differential occupancy of CLAMP and other co-factors promotes recruitment of MSL complex specifically to the male X-chromosome. Here, we determine that both CLAMP and GAF promote targeting of MSL complex to its binding sites, known as Chromatin Entry Sites (CES). Interestingly, CLAMP prevents high levels of GAF occupancy at CES, potentially through direct competition for GA-rich motifs. In contrast to CES, GAF is highly enriched at genes that are regulated by CLAMP, GAF or the pausing factor NELF, whose recruitment is modulated by GAF and CLAMP. Our data support a model in which both GA-rich binding factors cycle on and off of their binding sites with different dynamics in a context-specific manner. Therefore, the differential

occupancy relationship between two essential transcription factors promotes the context-specific targeting of a transcription complex.

Introduction

Precise regulation of gene expression is essential for cell viability and requires the activity of transcription complexes. Another layer of regulation is formation of the multi-protein complexes themselves, which can require specific gene dosage to achieve correct stoichiometry (Teichmann and Veitia 2004). Intolerance of incorrect copy-number is exemplified by the disease known as Down syndrome, or trisomy 21, which leads to a range of physical and neurological defects (Rachidi and Lopes 2007). Here, an extra copy of chromosome 21 leads to overexpression of genes including the dual-specificity tyrosine-regulated kinase 1A (*DYRK1A*) gene, which has been linked to developmental disabilities (Guimera et al. 1999; Ferrer et al. 2005). In order to study how gene dosage is regulated, many groups analyze the process of dosage compensation that balances sex chromosome transcript levels in heterogametic species such as mammals or *Drosophila melanogaster*, where XY males have one fewer copy of all genes on the X-chromosome compared to autosomes and to XX females (C. M. Disteché 2012). However, little is known about how the X-chromosome is specifically distinguished from autosomes for dosage compensation in any species.

In mammals, dosage compensation is achieved by identifying and increasing transcription of most X-linked genes in both sexes (Deng et al. 2011; Deng et al. 2013), followed by a second step: the random inactivation of one of the two female X chromosomes (Lyon 1961). Alternately, dosage compensation is accomplished through a single step in *D. melanogaster*, in which transcription is increased two-fold on the single male X-chromosome, while the female X-chromosomes remain unaffected (Hamada et al. 2005). For either mechanism, X-chromosome identification is the critical first step in dosage compensation. Therefore, we use *D. melanogaster* dosage

compensation as a model system to study this first step. Additionally, because it is essential that dosage compensation specifically targets the X-chromosome and not autosomal sites to maintain viability, studies of dosage compensation will also allow us to understand how large transcription complexes are specifically targeted to their binding sites.

The master regulator of dosage compensation in *D. melanogaster* is the Male-Specific Lethal (MSL) complex, a ribonucleoprotein complex assembled only in males (J. Belote and Lucchesi 1980). The MSL complex is specifically targeted to the single male X-chromosome where it upregulates gene transcription two-fold by depositing the histone 4 lysine 16 acetylation (H4K16ac) mark, which promotes transcriptional elongation (Larschan et al. 2011). The MSL complex is enriched on the X-chromosome within 1.5 kb regions at the 3' ends of genes called Chromatin Entry Sites (CES) (A. Alekseyenko et al. 2008; Straub et al. 2008). These CES contain one or more 21-bp GA-rich DNA motifs called MSL Recognition Elements (MREs) (A. Alekseyenko et al. 2008). Localization of MSL complex to MREs is dependent on an additional co-factor, Chromatin-Linked Adaptor for MSL Proteins (CLAMP), which was identified in a comprehensive RNA interference (RNAi) screen to identify proteins required for MSL complex recruitment that were also essential in both males and females and therefore would therefore have been missed by male-specific lethal screens (Larschan et al. 2012). CLAMP binds directly to MRE sequences *in vitro* and *in vivo* (Soruco et al. 2013), and was identified in a chromatin immunoprecipitation (ChIP) mass spectrometry experiment identifying proteins interacting with MSL (C. I. Wang et al. 2013).

Despite having a critical role in the process of dosage compensation, immunostaining and ChIP followed by sequencing (ChIP-seq) analysis reveals that CLAMP is not an X-specific protein and that only 3% of its binding sites overlap with MSL complex (Soruco et al. 2013). CLAMP is also not sex-specific; a decrease in CLAMP protein levels by either RNAi or mutation is lethal to both males and females by the pupal stage of development (Soruco et al. 2013; J. A. Urban, Doherty, et al. 2017). CLAMP is a highly conserved protein across insect species, and it is likely that it has a more ancient, non sex-specific function and was recently co-opted for male-specific dosage compensation due to expansion of GA-repeat binding sequences on the X-chromosome (Kuzu et al. 2016) (Chapter 2). Because the role of transcription factors in gene regulation often depends on their interactions with other proteins, we hypothesized that the presence or absence of a specific combination of co-factors promotes the context-specific function of CLAMP on the male X-chromosome.

We recently determined that CLAMP enhances chromatin accessibility over MSL complex recruitment sites specifically on the X-chromosome, a role previously attributed solely to the function of MSL-complex histone acetylation activity (J. A. Urban, Kuzu, et al. 2017) (under revision). Transcriptional effects of CLAMP reduction were also measured using the nascent RNA sequencing (Start-seq) technique (Nechaev et al. 2010), which identified CLAMP as an activator of transcription on the male X-chromosome, while on autosomes it both activates and represses gene expression (J. A. Urban, Kuzu, et al. 2017) (under revision). Proteomic analysis identified NELF as a non-sex specific interactor that co-localizes with CLAMP at promoters and whose occupancy levels are reduced by CLAMP (J. A. Urban, Urban, and Larschan 2017) (submitted). Another protein previously found to interact with MSL, NURF301 (C. I. Wang et al.

2013), was of interest due to its role in chromatin remodeling and previous findings that mutations in *NURF301* led to compaction of the male X-chromosome (Bai et al. 2007). ChIP-qPCR was performed for NURF301 after *clamp* RNAi, and revealed a specific interaction at CES on the male X-chromosome (J. A. Urban, Kuzu, et al. 2017) (under revision). Therefore, CLAMP functions to promote the occupancy of different co-factors at different genomic locations.

Previous investigation of CLAMP binding sites throughout the genome determined that CLAMP occupies long, clustered GA-repeats motifs that are enriched on the X-chromosome, especially within CES (Kuzu et al. 2016) (Chapter 2). CLAMP, however, is not the only protein capable of binding to GA-rich sequences. In fact, prior to identification of CLAMP, a different well-studied transcription factor encoded by the *trithorax-like (trl)* gene was known to bind directly to GA-rich sequences, earning the name GAGA Factor (GAF) (Farkas et al. 1994). Functional analysis determined that hypomorphic mutations that decrease GAF function have male-specific lethal phenotypes (Greenberg, Yanowitz, and Schedl 2004). However, GAF did not appear to be essential for recruiting MSL complex to the X-chromosome at the resolution of polytene chromosome staining, because only a single MSL complex polytene band was lost after depletion of GAF (Greenberg, Yanowitz, and Schedl 2004). Furthermore, CLAMP but not GAF was identified by ChIP-mass spectrometry for proteins that physically associate with MSL complex on chromatin (C. I. Wang et al. 2013).

There are several lines of evidence that suggested the relationship between these two GA-binding transcription factors may regulate MSL complex targeting and function of the NELF complex: 1) We recently determined that GAF and CLAMP are components of the

same insulator protein complex that physically interacts with CES (Kaye et al. 2017) (Chapter 4); 2) GAF and CLAMP can both recruit NURF301, a chromatin-remodeling enzyme shown to promote chromatin accessibility at CES (J. A. Urban, Kuzu, et al. 2017) (under revision) (Lomaev et al. 2017; Toshio Tsukiyama, Becker, and Wu 1994; Bai et al. 2007); 3) CLAMP and GAF are the only two known direct GA-binding transcription factors in *Drosophila* that both regulate chromatin (Fuda et al. 2015; J. A. Urban, Kuzu, et al. 2017) (under revision); 4) CES, which are the optimal MSL complex recruitment sites are enriched for GA-binding motifs (Kuzu et al. 2016) (Chapter 2); 5) CLAMP null mutant males die earlier during development than females (J. A. Urban, Doherty, et al. 2017), and GAF hypomorphic alleles confer male-specific lethal phenotypes (Greenberg, Yanowitz, and Schedl 2004); 6) GAF promotes targeting of NELF to paused promoters (Lee et al. 2008), and CLAMP reduces targeting of the NELF complex (J. A. Urban, Urban, and Larschan 2017) (submitted). Therefore, we hypothesized that the relative occupancy of CLAMP and GAF at different genomic loci promotes accurate targeting of MSL complex at CES and NELF at promoters.

Using a combination of genetic, genomic and biochemical approaches, we determined that differential occupancy of CLAMP and GAF regulates MSL complex recruitment. GAF and CLAMP directly compete for binding sites *in vitro* and specific X-enriched GA-rich *cis*-elements favor CLAMP binding. At CES, CLAMP prevents GAF occupancy and GAF enrichment increases in the absence of CLAMP. Surprisingly, despite not directly interacting with MSL complex or localizing to most CES, GAF promotes accurate targeting of MSL binding to CES and reduces MSL binding to non-target sites. In contrast, at promoters of genes that are regulated by CLAMP, GAF, or NELF, CLAMP only partially excludes GAF occupancy. GAF remains enriched at these genomic

locations in the presence of CLAMP, where both CLAMP and GAF are known to modulate NELF occupancy in opposite ways (Lee et al. 2008) (J. A. Urban, Urban, and Larschan 2017) (in prep). Therefore, differential occupancy of CLAMP and GAF at different genomic locations regulates the targeting of different large transcription complexes. Overall, we have provided new insight into how differential occupancy of similar transcription factors promotes accurate targeting of transcription complexes.

Results

MSL complex binding sites have high CLAMP occupancy and low GAF occupancy

CLAMP and GAF both have roles in chromatin regulation. At its highest occupancy sites, CLAMP promotes the open chromatin environment on the male X-chromosome and alters positioning of the nucleosome-depleted region at promoters of genes (J. A. Urban, Kuzu, et al. 2017)(In revision). Similarly, GAF can maintain nucleosome-free promoter regions to activate gene transcription and recruit NELF to paused promoters (Fuda et al. 2015). We therefore hypothesized GAF and CLAMP may functionally cooperate through the regulation of chromatin accessibility.

To understand the functional relationship between CLAMP and GAF on chromatin *in vivo*, we first examined genetic interactions between *clamp* and *trl* mutants using a position effect variegation assay (PEV) with the white-mottled 4 (w^{m4}) eye color reporter system (Judd 1955; Gerasimova et al. 1995). In this assay, a section of the X-chromosome including the *white* gene, which encodes red eye pigment, is inverted such that the *white* gene is adjacent to a heterochromatin boundary. Spreading of heterochromatin can be inferred from a change in the level of mosaicism of the eye color from red (open chromatin) to white (closed chromatin) as the *white* gene becomes

inactivated (Eissenberg 1989). We used flies heterozygous for mutant alleles due to lethality of homozygous mutants. The *clamp*² allele is a protein null mutation introduced via CRISPR/Cas9 editing that results in homozygous lethality in both males and females (J. A. Urban, Doherty, et al. 2017). The GAF mutant allele, *Trl*^{13C} is a hypomorphic allele resulting from a p-element insertion and has previously been characterized to have enhancer of variegation E(var) activity in PEV assays, referring to promotion of the spread of heterochromatin in the mutant state (Katokhin et al. 2001).

We measured eye pigment using absorbance of homogenized heads from male mutant flies and identified that both *clamp*² and *Trl*^{13C} single mutants are enhancers of PEV or E(var) mutants based on their whiter eye color relative to a control suppressor of variegation, or Su(var), mutant allele of *modulo* (*Mod*^{L8}) (Graba et al. 1994). Both *clamp*² and *Trl*^{13C} single mutants were significantly different from *modulo* (*mod*^{L8}) mutant flies which have high red pigment (Figure 3.1A, p-values reported as Table S3.1). Double heterozygous mutant flies (*clamp*²/+;*Trl*^{13C}/+) have a stronger E(var) phenotype compared to single mutants (*clamp*²/+ versus *clamp*²/+;*Trl*^{13C}/+ p-value = .045, *Trl*^{13C}/+ versus *clamp*²/+;*Trl*^{13C}/+ p-value = 0.0024) consistent with a synthetic genetic interaction between the *clamp*² and *Trl*^{13C} alleles (Figure 3.1A). Our interpretation of these results is that CLAMP and GAF function as activators that open chromatin because their mutants cause a spreading of heterochromatin. Therefore, the CLAMP and GAF proteins may be able to cooperate to promote the euchromatic state on the X-chromosome. We hypothesized that one impact of this function may involve increasing X-chromosome chromatin accessibility to promote MSL complex targeting.

Next, we examined the localization pattern of CLAMP and GAF relative to MSL complex on wild type male polytene chromosomes (Figure 3.1B). We found that MSL complex co-

localizes with CLAMP, as previously reported (Soruco et al. 2013). In contrast, we observed less co-localization between GAF and MSL complex. Together, these data qualitatively demonstrate that CLAMP and GAF exhibit different occupancy patterns compared to each other and to MSL complex.

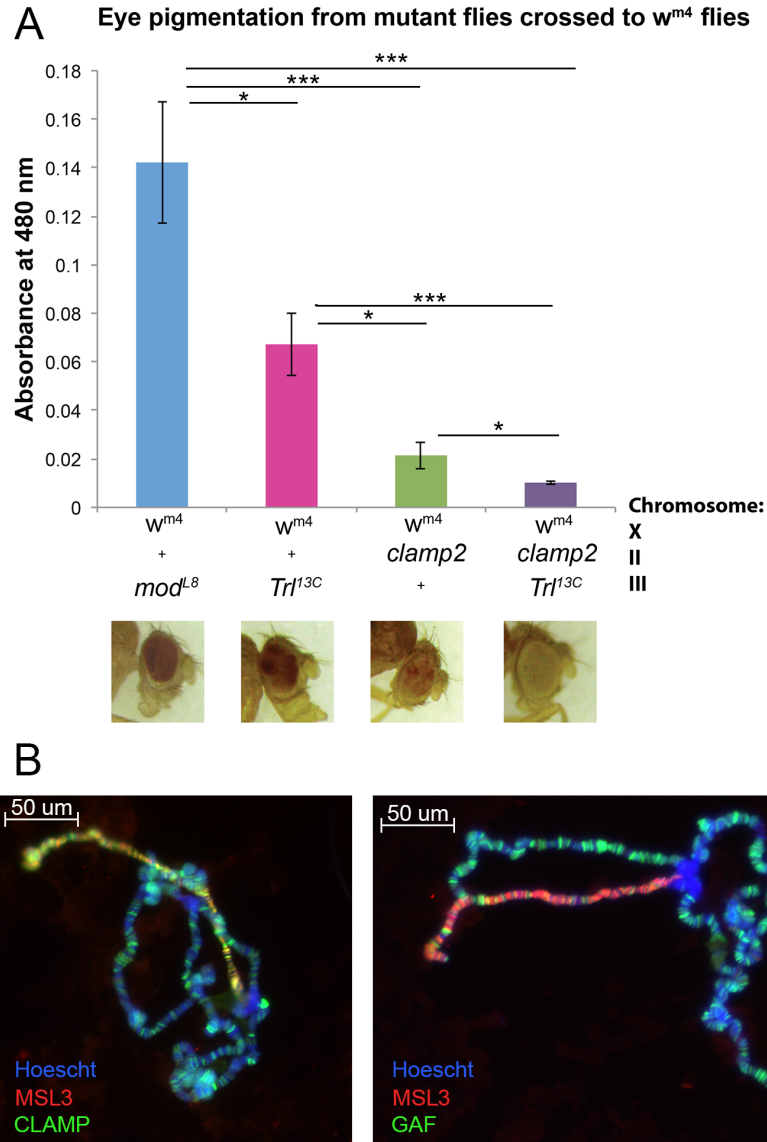


Figure 3.1 GAF and CLAMP open chromatin, a feature of the dosage compensated male X-chromosome, but only CLAMP co-localizes with MSL.

A. Eye pigmentation assay for PEV of mod^{L8} , Trl^{13C} , $clamp^2$ and $Trl^{13C}; clamp^2$ mutants. Absorbance measured at 480 nm plotted with standard error of 4 replicates, $n=5$ male fly heads per replicate. Representative image of eye color included below the appropriate genotype. t-test p-values: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. **B.** Immunostaining of wild type male *Drosophila* salivary gland polytene chromosomes. DNA stained blue, MSL3 in red, and CLAMP (left) or GAF (right) in green. Overlap in localization with MSL observed as yellow.

To map the occupancy of CLAMP and GAF relative to each other and to MSL complex at higher resolution than possible with polytene chromosomes, we performed ChIP-seq on all three factors in male *Drosophila* S2 cells. Four replicates were performed for each immunoprecipitation condition, and we plotted replicability versus enrichment score to assess data quality (Figure S3.1A – S3.1C). Replicates were then removed from future analysis based on poor replicability when compared to remaining samples, e.g. CLAMP replicate three (Figure S3.1A). Peaks were called using MACS2 and filtered for an enrichment value where peaks replicating > 0.8, resulting in cutoffs of fold enrichment of 3 or higher for GAF and CLAMP and 1.8 for MSL (Zhang et al. 2008). Note that due to the known nature of MSL binding along active genes rather than at sharp peaks (A. Alekseyenko et al. 2008), the MACS2 broad peak algorithm function was used to call MSL3 peaks.

We first quantified GAF and CLAMP overlap for all peaks (Figure 3.2A). GAF had more total peaks called compared to CLAMP (5818 versus 3754), as well as more unique peaks with 61% versus 32% unique CLAMP peaks. We further analyzed peak overlap by separating peaks on the X-chromosome versus autosomes (Figure S3.2A and S3.2B). On the X-chromosome, CLAMP had a lower percentage of peaks overlapping with GAF (56%) compared to autosomes (68%). Next, we measured overlap of each factor with MSL on the X-chromosome (Figure 3.2B and C). Consistent with polytene staining, we found that CLAMP and MSL complex occupancy on the X-chromosome overlaps at 21% of CLAMP peaks (218 shared peaks), which is four times greater than the fraction of GAF peaks that are co-occupied by MSL complex (5% of GAF peaks: 54 shared peaks).

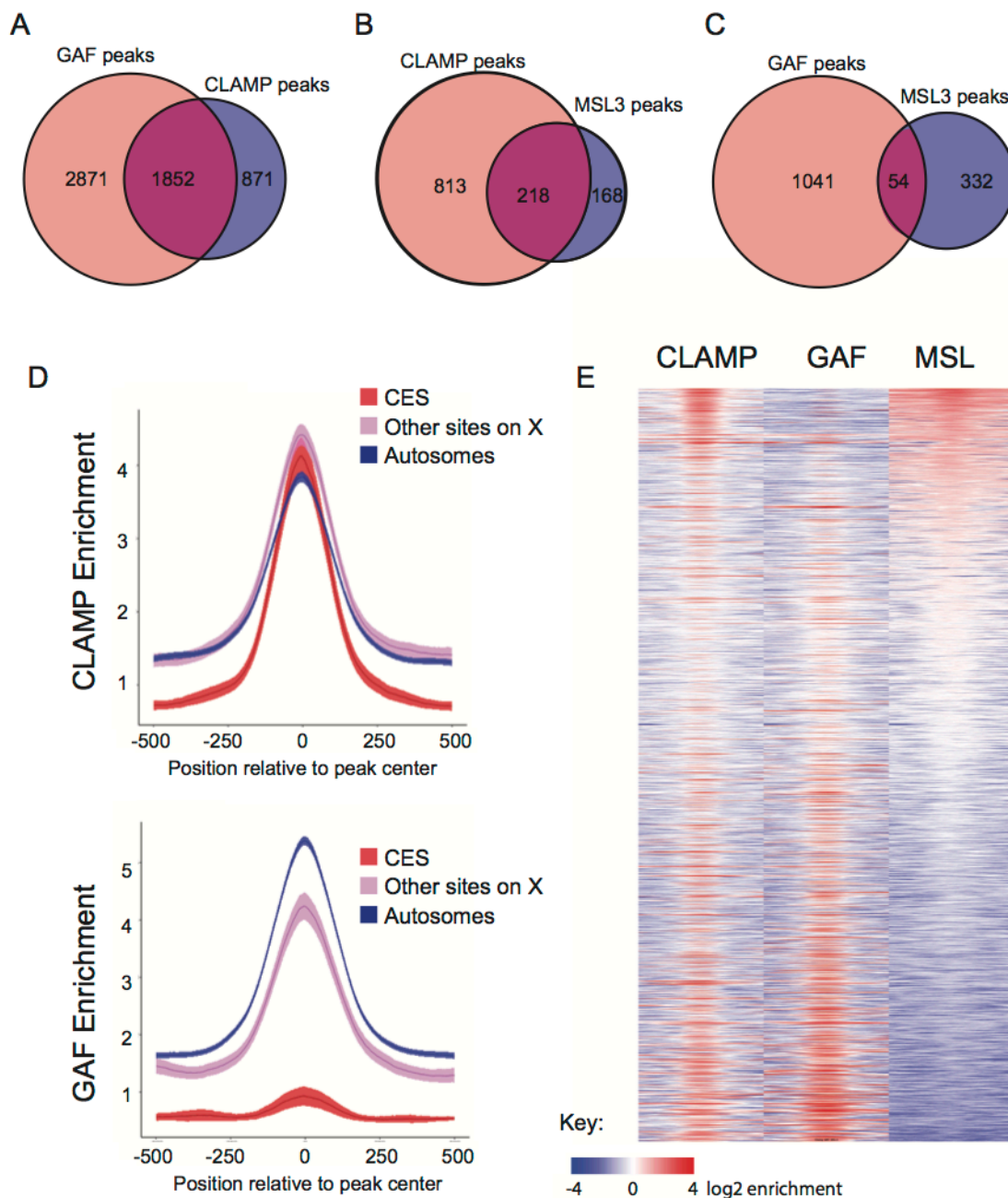


Figure 3.2 ChIP-seq determines CLAMP overlaps better with MSL than GAF.

A. Venn diagram showing number of GAF (light red) and CLAMP (blue) peaks and peaks where both GAF and CLAMP are present (purple). **B.** Venn diagrams showing number of CLAMP (light red) and MSL3 (blue) peaks and peaks where both MSL3 and CLAMP are present (purple) on the X-chromosome. **C.** Venn diagram showing number of GAF (light red) and MSL3 (blue) peaks and peaks where both MSL3 and GAF are present (purple) on the X-chromosome. **D.** Fold enrichment average gene profiles of CLAMP (top) and GAF (bottom) from ChIP-seq data. Averages were plotted for peaks on autosomes (blue), CES (red), and other sites on the X-chromosome (pink) with the 95% confidence interval represented by shading around the line. **E.**

Heat map of 1-kb regions centered on CLAMP and/or GAF peaks on the X-chromosome. White regions indicate background enrichment while red and blue represent above and below background enrichment, respectively (see key). Sites are rank-ordered by MSL3 enrichment.

Peak counting analysis does not account for peak magnitude. Therefore, we next plotted enrichment profiles over peaks at CES compared to other genomic binding sites on X or on autosomes (Figure 3.2D, line shading indicates 95% confidence intervals). CLAMP binds to all sites with equal enrichment (top panel), with a slight preference for X-chromosome sites compared to autosomes. This is consistent with the previous finding that the number of total sites and the density and clustering of these sites enhance CLAMP occupancy of the X-chromosome rather than higher occupancy of individual sites (Kuzu et al. 2016) (Chapter 2). In contrast, GAF shows more differences between categories (bottom panel), where CES are significantly less occupied compared to other sites on X and sites on autosomes. To define how GAF and CLAMP binding occupancy compares to MSL complex occupancy, we next rank-ordered all sites on the X-chromosome that are occupied by CLAMP, GAF or MSL complex by MSL complex occupancy levels (Figure 3.2E). The most enriched MSL sites (top of plot) were depleted for GAF, indicating an inverse relationship between GAF and MSL occupancy levels. We formally tested for an inverse correlation using a scatter plot of GAF occupancy vs MSL occupancy on X, and obtained a Pearson's correlation coefficient of r^2 statistic of -0.39 (Figure S3.2C). Interestingly, CLAMP and GAF co-occupy sites at the bottom of the MSL complex occupancy rank-ordered heat map suggesting that their co-occupancy is inversely correlated with MSL complex occupancy (Figure 3.2E).

dosage compensation (State 5) states. Consistent with the co-localization of CLAMP and MSL complex, CLAMP is more enriched for the dosage compensation chromatin state (State 5) than GAF (Figure S3.3A and S3.3B). Interestingly, we also found that sites enriched for the Polycomb repressor complex (State 6) are more enriched for CLAMP than GAF. This analysis confirms a more specific association with dosage compensation for CLAMP versus GAF, and suggests that Polycomb complex repressive sites are an interesting area for future comparisons between CLAMP and GAF.

CLAMP and GAF compete for the same binding sites *in vitro*

We have previously shown that CLAMP binds to sites on the X-chromosome or autosomes equally well, as long as the DNA contains specific long, GA-rich *cis*-elements (Kuzu et al. 2016) (Chapter 2). We characterized additional features of the motif to understand the mechanism of a potential X-chromosome specific function. On the X-chromosome, CLAMP binding sites occur approximately two-fold more frequently than on autosomes and are biased more towards gene bodies where MSL complex is recruited. Also, CES contain enhanced clustering of CLAMP motifs compared to other genomic locations. Therefore, we asked whether sites that contain the X-enriched CLAMP motif have a higher affinity for CLAMP than GAF, which may promote the specific CLAMP enrichment at CES *in vivo* (Figure 3.2D and E).

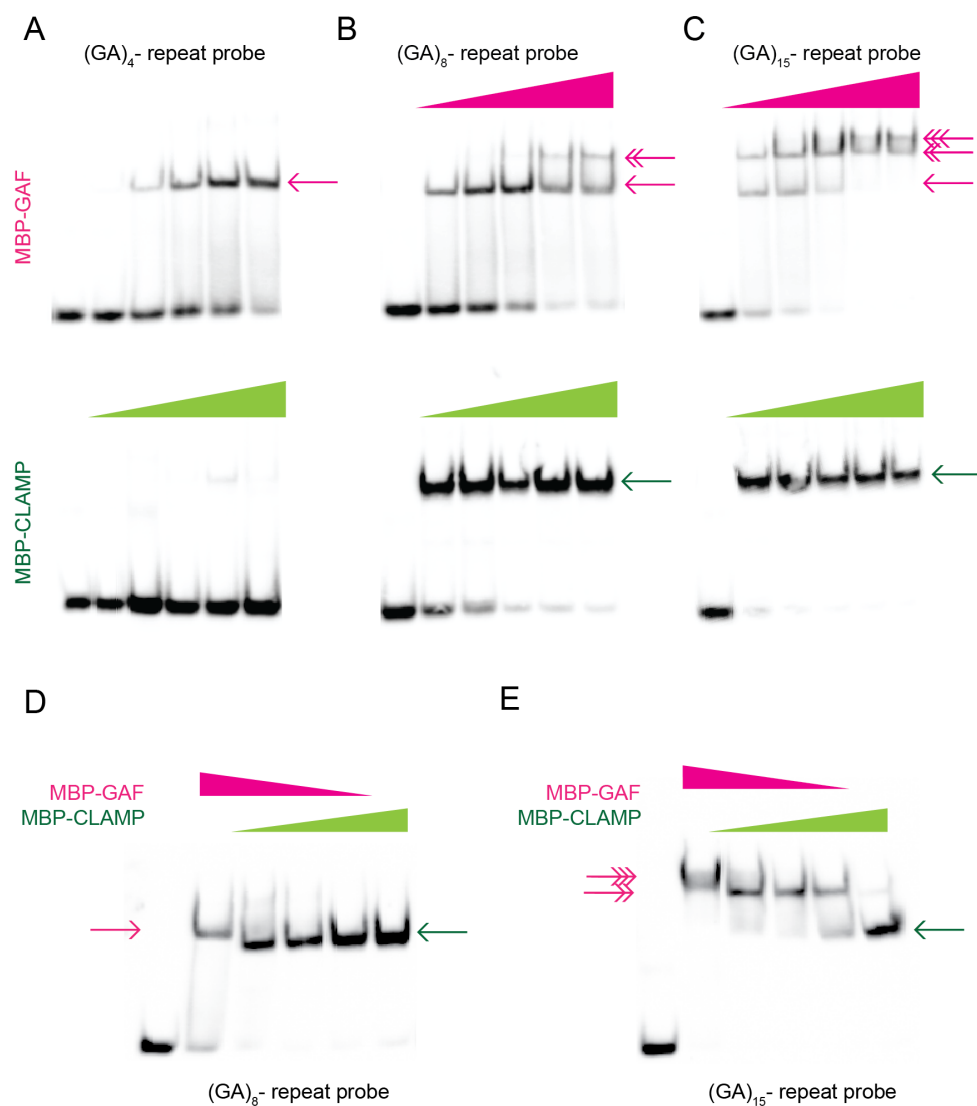


Figure 3.4 GAF and CLAMP compete for GA-repeat sequences *in vitro*.

A. EMSA using a biotin-labeled DNA probe containing an 8-bp (GA)₄-repeat. GAF (top) and CLAMP (bottom) DNA binding domain MBP-fusion proteins were added individually. Final concentration of protein in the reaction is listed above each lane. Shifts of protein-bound DNA indicated by arrows. **B.** EMSA as in A above, now using DNA containing a 16-bp (GA)₈-repeat. DNA shifted by multiple proteins (i.e. two GAF proteins) indicated by double or triple arrowheads. **C.** EMSA as in A and B, now using DNA containing a 30-bp (GA)₁₅-repeat. **D.** Competition EMSA using the 16-bp GA₈-repeat probe. The first lane is a DNA-only control, the second lane is GAF-only, lanes 3-5 are competition with both proteins in the reaction, and the final lane is CLAMP-only. Shifts of protein-bound DNA indicated by arrows. **E.** Competition EMSA as in D, now using the 30-bp GA₁₅-repeat. Shifts of protein-bound DNA indicated by arrows, and as above DNA shifted by multiple proteins indicated by double or triple arrowheads.

Although GAF and CLAMP both contain C2H2 zinc-finger DNA binding domains, CLAMP has six tandem zinc fingers, compared with GAF that has only a single zinc finger (Figure 3.3A). We were therefore not surprised to find different length motifs when we compared the *in vivo* CLAMP binding motif with the *in vivo* GAF binding motif derived from our ChIP-seq data, using MEME for motif identification (Machanick and Bailey 2011) (Figure 3.3B). Consistent with our previous reports (Kuzu et al. 2016) (Chapter 2), we found that the CLAMP 21-bp binding motif includes a core 8-bp region (starting at position 5) with a characteristic high enrichment of the first two “GA” nucleotides within the repeat followed by two less conserved “GA” nucleotides, and finally four additional “GAGA” nucleotides.

We previously demonstrated that both the core and flanking sequences are required for CLAMP occupancy (Kuzu et al. 2016) (Chapter 2). In contrast, the GAF binding motif is shorter but requires more highly conserved stretches of GA-repeats and the minimal motif contains five highly conserved contiguous GA-base pairs, “GAGAG,” consistent with a previously described GAF-DNA crystal structure (Omichinski and Pedone 1997), that are not always present within the CLAMP binding motif (Kuzu et al. 2016) (Chapter 2). We next determined the motifs from sites bound to either CLAMP or GAF, excluding overlapping sites (Figure 3.3C and D). While both are still GA-rich, GAF has two short contiguous GAGAG 5-bp consensus sequences at positions 1-5 and 9-13 within the motif logo (Figure 3.3D). In contrast, sites that bind only to CLAMP rarely have the 5-bp GAF consensus (Figure 3.3C). Instead, CLAMP tolerates breaks that are less conserved, especially at the 3rd and 4th positions within the core 8-bp region. Therefore, it is possible that the less conserved bases within the middle of the CLAMP consensus sequence function to allow CLAMP occupancy and prevent GAF occupancy.

In order to test our hypothesis that CLAMP and GAF directly compete with each other, we measured their binding individually and in competition with each other using an electrophoretic mobility shift assay (EMSA, or “gel-shift”). We expressed and purified MBP fusions of both proteins containing either one of their DNA binding domains (Figure 3.3A). We found that CLAMP did not bind to a probe containing an 8-bp (GA)₄-repeat (Figure 3.4A), but bound well to probes that contained 16 or 30 bp of GA-repeat sequence (Figure 3.4B and C), consistent with previous results using these same DNA sequences but with only a single concentration of protein (Kuzu et al. 2016) (Chapter 2). In contrast, GAF can bind *in vitro* to probes containing an 8bp (GA)₄-repeat, in addition to longer sequences. Interestingly, as GAF protein concentration was increased, additional shifted species were observed for the probes containing longer GA-stretches, suggesting the opportunity for multiple GAF-DNA binding domain fusion proteins to bind (Figure 3.4B and C).

We next performed competition EMSAs to test whether both GA-binding proteins compete for the same probes, and whether changing the relative concentrations of proteins or GA-repeat length can alter protein binding properties. CLAMP outcompetes GAF for a 16-bp (GA)₈-repeat probe, and GAF was unable to compete even at a 3:1 stoichiometric ratio of (Figure 3.4D). In contrast, we observe GAF binding signal more than CLAMP for the 30 bp (GA)₁₅-repeat, at a shift corresponding to multiple GAF proteins, indicating GAF is able to compete with CLAMP likely as a multimer (Figure 3.4E). Overall, these results demonstrate that CLAMP and GAF can compete with each other for binding to GA-repeat containing DNA sequences, and that their relative occupancy can be altered by changing the number of continuous GA-repeats.

CLAMP and GAF alter the other's occupancy to promote specific targeting of MSL complex to CES

Despite similar binding motifs, CLAMP and GAF have both shared and unique binding sites (Figure 3.2A and Figure S3.2A and S3.2B). Since *in vitro* experiments suggest that they can compete for the same binding sites (Figure 3.3D and E), we hypothesized that competition influences their *in vivo* binding patterns and that this affects MSL complex recruitment. Therefore, we measured the occupancy relationship between CLAMP, GAF and MSL complex by performing RNAi targeting *clamp* and *trl*, followed by ChIP-seq for each factor in male S2 cells. We confirmed efficient knockdown efficiency for *clamp* and *trl* by Western blotting (Figure S3.4), in agreement with previous studies using these RNAi targets (Sorucu et al. 2013; Fuda et al. 2015).

We next examined average gene profile enrichment of all peaks for CLAMP and GAF under *gfp* (control, blue), *trl* (pink) or *clamp* (green) RNAi conditions (Figure 3.5A). As expected, CLAMP occupancy at most peaks was reduced after *clamp* RNAi and GAF occupancy at most peaks was reduced after *trl* RNAi. We next examined GAF localization after *clamp* RNAi and noted that, on average, GAF occupancy increased, consistent with *in vitro* competition between the two factors. CLAMP occupancy also increased after *trl* RNAi, although more modestly, on average. Next, we plotted average gene profiles centered at CES peaks for all three factors, to determine the relationships between CLAMP and GAF at CES and their impact on MSL binding (Figure 3.5B). We observed similar trends for RNAi effects on GAF and CLAMP at CES, compared to their profiles at all sites in the genome (Figure 3.5A), although GAF enrichment is lower at CES when compared with its other binding sites (Figure 3.2D). Despite weak GAF binding to CES, average MSL enrichment significantly decreased after *trl* RNAi, although not as completely as after *clamp* RNAi.

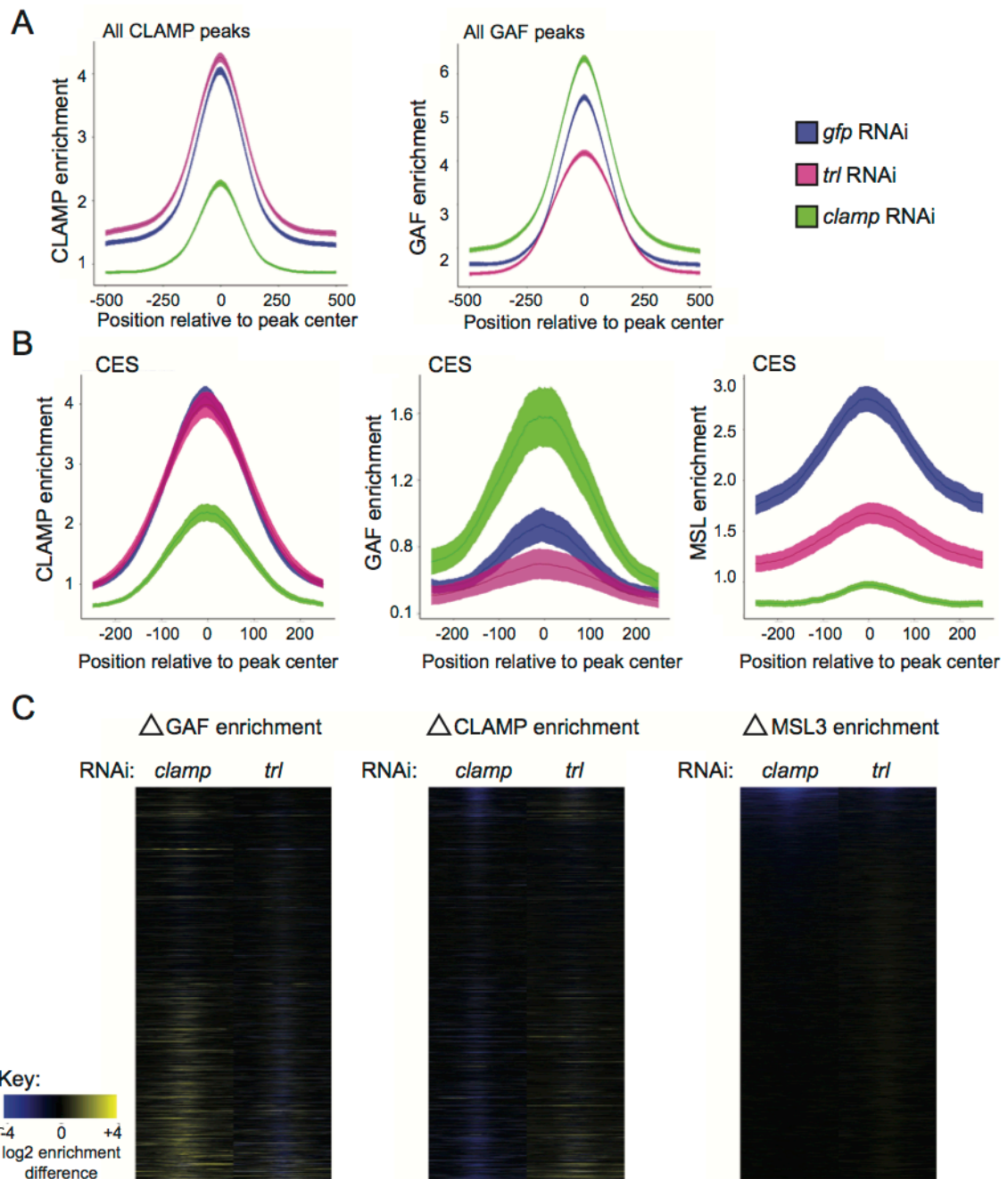


Figure 3.5 Effects of RNAi on GAF, CLAMP and MSL localization.

A. Average enrichment of CLAMP (left) and GAF (right) at CLAMP or GAF peaks centered on peak summit under *gfp* (blue), *trl* (pink), and *clamp* (green) RNAi conditions. Lines indicate mean enrichment with shaded areas representing 95% confidence intervals. **B.** Average enrichment of CLAMP (left), GAF (middle), and MSL3 (right) at CES under *gfp* (blue), *trl* (pink), and *clamp* (green), RNAi conditions. Lines indicate mean enrichment with shaded areas representing 95% confidence intervals. **C.** Heat map of the change in GAF enrichment (left), CLAMP enrichment (middle) and MSL3 enrichment (right) of the 1000-bp regions on the X-chromosome ordered as depicted in 2E under *clamp* RNAi (left column) and *trl* RNAi (right column). Black regions indicate

background enrichment while yellow and blue represent increases and decreases in enrichment, respectively (see key).

In order to define the different types of occupancy changes that occur after RNAi treatments, we generated heat maps of the difference in CLAMP, GAF or MSL occupancy between control and *clamp* or *trl* RNAi treatment at all sites on the X-chromosome that were bound by any factor after any RNAi treatment (Figure 3.5C). We ordered these by MSL occupancy under wild type conditions (as in Figure 3.2E), and the full set of sites is inclusive of 1-kb regions centered over peaks bound by any factor under any condition. As expected from the average profile plots, *trl* RNAi decreased GAF signal and *clamp* RNAi decreased CLAMP signal (Figure 3.5C, columns 2 and 3). In contrast, *trl* RNAi causes both increases and decreases in CLAMP occupancy throughout the genome (Figure 3.5C, column 4), perhaps accounting for the weak overall difference observed in the average gene profiles analysis (Figure 3.5A and B). Similarly, *clamp* RNAi causes diverse changes on GAF occupancy (Figure 3.5C column 1), but increases GAF occupancy at more sites, leading to the overall increase observed in the average gene profiles.

We next considered the changes in MSL complex occupancy after RNAi treatment. We observed that *trl* RNAi results in a decrease in MSL complex occupancy at the highest MSL complex binding sites (Figure 3.5C). In contrast, *trl* RNAi treatment causes MSL complex to modestly increase at sites where MSL complex does not normally bind (Figure 3.5C and Figure S3.5). Therefore, even though GAF is not stably bound at most MSL complex binding sites, it promotes the specificity of MSL complex recruitment to these sites and reduces low levels of aberrant localization to inappropriate targets. Aberrant localization of MSL complex appears in the center of the heat map, indicating

the increases are centered on CLAMP or GAF binding sites. These sites were previously identified to be some of the most accessible regions in the genome (Fuda et al. 2015; J. A. Urban, Kuzu, et al. 2017) (in revision). It is thus possible MSL complex is recognizing open chromatin regions after *trl* RNAi, similar to previous analysis of weak MSL complex binding sites to open chromatin regions on autosomes (Artyom A. Alekseyenko et al. 2012).

CLAMP reduces GAF occupancy at genes where CLAMP, GAF, and NELF regulate transcription

In order to determine how the competition between GAF and CLAMP correlates with gene expression changes, we examined factor recruitment at genes that are either regulated by GAF or CLAMP, compared with unregulated genes. We were also interested in investigating the relationship between CLAMP and GAF at NELF-regulated genes due to the known physical association and functional interaction between NELF and both GA-repeat binding proteins (Lomaev et al. 2017; Lee et al. 2008; J. A. Urban, Urban, and Larschan 2017) (submitted). Previous work indicates that GAF promotes NELF recruitment while CLAMP reduces NELF recruitment at paused genes (Lee et al. 2008; J. A. Urban, Urban, and Larschan 2017) (submitted). We used three available data sets that tested for transcriptional changes in male S2 *Drosophila* cells after RNAi treatments with our factors of interest. These data include *clamp* RNAi nascent transcript sequencing (Start-seq) (J. A. Urban, Kuzu, et al. 2017) (in revision) and *trl* or *nelf* RNAi global run-on sequencing (GRO-seq) (Core et al. 2012; Fuda et al. 2015).

First, we measured the ChIP-seq enrichment of GAF at promoters of genes regulated by CLAMP, GAF, or NELF where transcriptional changes after RNAi were greater than one standard deviation, compared to unregulated genes. We determined that GAF is more enriched at genes regulated by CLAMP, GAF or NELF, compared with unregulated genes from all three data sets (Figure 3.6, A-C). When considering the direction of the transcriptional regulation, GAF is more enriched at CLAMP-repressed than CLAMP-activated genes (Figure 3.6A), yet GAF enrichment is highest for positively regulated genes in the NELF and GAF-regulated categories (Figure 3.6B and C). Note that NELF is an activating factor due to the critical function of pausing in transcription, so there are very few genes in the NELF-repressed category that increase after *nelf* RNAi (Core et al. 2012).

We next plotted CLAMP ChIP-seq enrichment at genes regulated by CLAMP, GAF, or NELF and compared this to enrichment at unregulated genes. As expected, CLAMP was enriched at genes where it regulated transcription (Figure 3.6D), although there are still sites where it binds but does not alter transcription. CLAMP binding was enriched at NELF-positively regulated genes (Figure 3.6E), and although CLAMP was enriched at both categories of GAF-regulated genes, enrichment was slightly higher at GAF-repressed genes (Figure 3.6F).

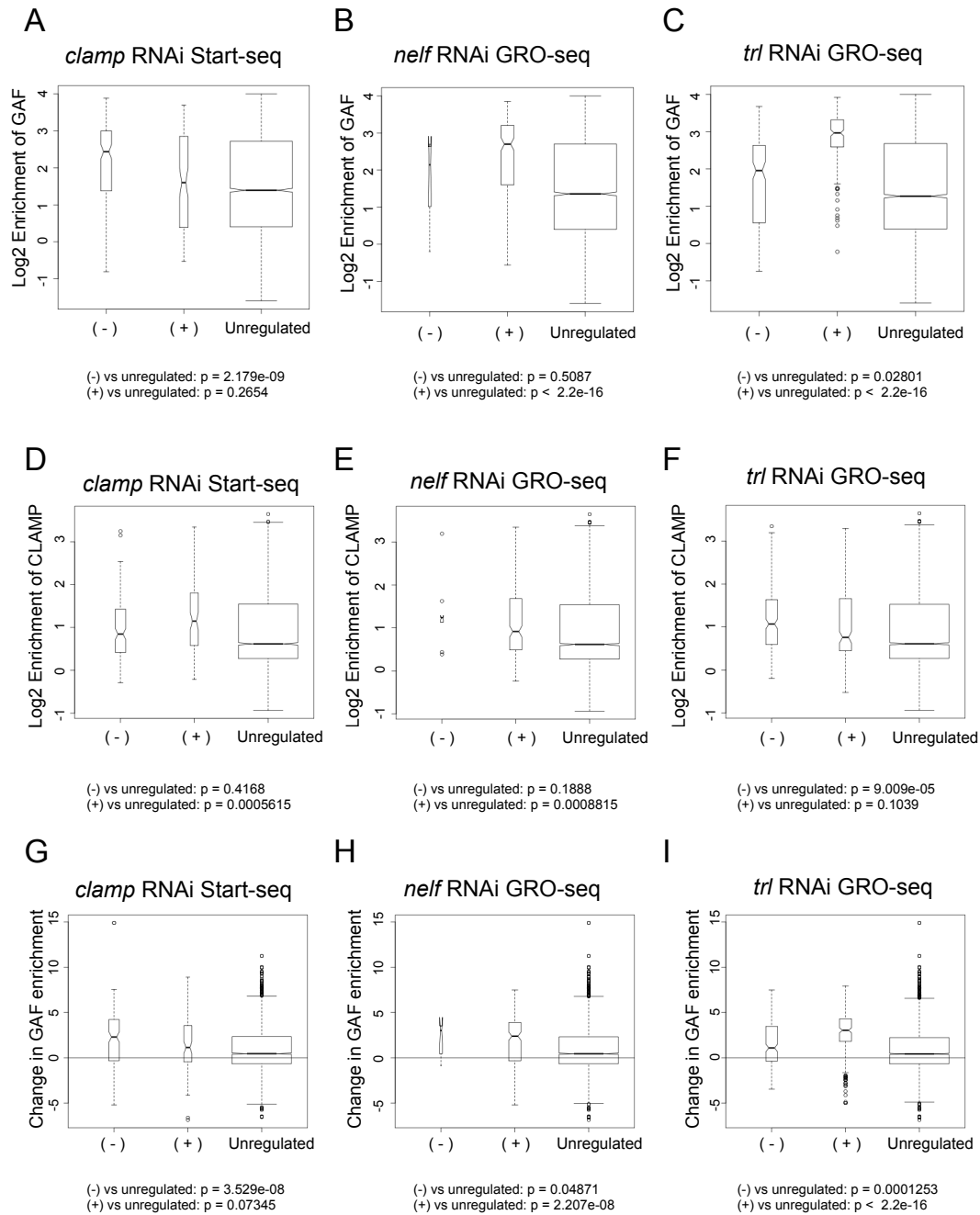


Figure 3.6 Enrichment of GAF and CLAMP at genes sorted by transcriptional changes due to *clamp*, *nelf*, and *trl* RNAi treatments.

GAF enrichment (A-C), CLAMP enrichment (D-F), or change in GAF enrichment after *clamp* RNAi (G-I) were plotted for genes identified as positively regulated (+), negatively regulated (-), or unregulated by either CLAMP (A, D, G), NELF (B, E, H), or GAF (C, F, I). Collared box plots indicating confidence intervals were generated using the R package "boxplot" and represent (+/-) 1.58 times the inter-quartile range divided by the square root of the sample size. p-values generated from a paired Student's t-test.

Due to our interest in the relationship between GAF and CLAMP, we investigated the effects of *clamp* RNAi on GAF enrichment. We observed that GAF occupancy increased at CLAMP-regulated genes, suggesting that the presence of CLAMP reduces GAF binding at sites where CLAMP regulates transcription (Figure 3.6G). Interestingly, CLAMP more strongly reduces GAF recruitment at CLAMP-negatively regulated genes than at CLAMP-activated or unregulated genes. In contrast, the change after *clamp* RNAi was greater at genes positively regulated by GAF or NELF than those that are negatively regulated (Figure 3.6H and I). Since GAF is known to recruit NELF, the activity of CLAMP decreasing GAF occupancy at NELF-regulated genes may then lead to the decreases in NELF occupancy observed at paused genes after *clamp* RNAi (J. A. Urban, Urban, and Larschan 2017) (submitted).

Together, these data suggest that one of the mechanisms by which CLAMP regulates transcription, reducing GAF occupancy, is similar to its function at CES. However, GAF remains enriched at regulated promoters but not at CES, likely due to the clustered strong CLAMP binding sites that are present at CES but not at most promoters. Therefore, the rapidly evolving MSL complex (Rodriguez et al. 2007; Kuzu et al. 2016) (Chapter 2), may have co-opted an ancient, non sex-specific function of CLAMP: regulating gene expression by reducing GAF occupancy as CLAMP binding sites became enriched on the X-chromosome through transposon hopping (C. E. Ellison and Bachtrog 2013).

Discussion

Here we analyzed the relationship between CLAMP and GAF as a model for understanding how differential binding of similar transcription factors promotes the

recruitment of large transcription complexes. By combining *in vivo* and *in vitro* approaches, we determined that CLAMP and GAF compete with each other *in vitro* and can exhibit either synergy or competition *in vivo*. CLAMP binding sites, containing long GA-rich motifs, are highly enriched and clustered at CES. At these locations, CLAMP recruits MSL, while stable GAF occupancy is excluded. Despite not specifically binding to these sites, a reduction of GAF leads to loss of specificity in MSL recruitment, which potentially explains the male-specific lethal genetic interactions between GAF and MSL from early studies of GAF in dosage compensation.

There are several possible models for the role of GAF in MSL targeting. One is that despite not capturing a stable interaction with ChIP, GAF may still bind CES transiently to alter CLAMP binding. Recent work tested a facilitated diffusion model for competition between transcription factors, which may be relevant for the GAF and CLAMP interactions. In this model, transcription factors in solution can facilitate the binding and release of transcription factors bound to DNA (Kamar et al. 2017). After reduction of GAF and the removal of facilitated diffusion, CLAMP localization is disrupted and MSL complex no longer identifies CES. It will be interesting to investigate the new MSL sites to identify features that may recruit MSL in the absence of the sequence specific targeting to CES.

Both GAF and CLAMP functionally interact with the NURF chromatin-remodeling complex (Toshio Tsukiyama, Becker, and Wu 1994; Toshio Tsukiyama and Wu 1995; J. A. Urban, Kuzu, et al. 2017) (in revision). Therefore, it is possible that by cycling on and off of chromatin both factors are able to maintain the high level of chromatin accessibility required for targeting MSL complex specifically to the X-chromosome. At CES with the highest level of chromatin accessibility, *clamp* RNAi does not cause a decrease in

accessibility (J. A. Urban, Kuzu, et al. 2017) (in revision), suggesting redundancy with a similar factor such as GAF. In parallel work, we have also investigated the relationship between GAF and CLAMP at another specific genomic location, the histone locus. At the histone locus, which contains 100 tandem copies of the histone gene array, CLAMP also excludes GAF such that GAF only binds after reduction of CLAMP (Rieder, Koreski, Boltz, Kuzu, Urban, Bowman, Zeidman, et al. 2017). An intriguing interpretation is that CLAMP is specialized for domains of co-regulated gene expression due to its ability to open chromatin at a distance up to 7 kb from its peak of occupancy (J. A. Urban, Kuzu, et al. 2017) (in revision) compared with GAF, which acts more locally such as at specific promoter sites (Fuda et al. 2015).

The different binding specificities of GAF and CLAMP also provide insight into how CLAMP is specifically enriched at CES. CLAMP has more tolerance for variation in base pairs in the middle of its GA-rich motif and does not require five contiguous GAGAG nucleotides. In contrast, GAF binds to the “GAGAG” pentamer as well as longer GA-repeat stretches. It is possible that there is more variability in GA-rich motifs with CES because they recently inserted onto the X-chromosome through transposon-hopping (C. E. Ellison and Bachtrof 2013; Joshi and Meller 2017). The MRE motif enriched at MSL complex binding sites is not a continuous GA-repeat, which is consistent with increased CLAMP occupancy outcompeting GAF. Furthermore, GAF has a single zinc-finger, while CLAMP has six tandem fingers, so their DNA-binding kinetics are likely to be different.

Based on an available crystal structure of the six zinc-finger protein Aart (Segal et al. 2006), proteins with six tandem zinc-fingers may wrap around the entire double helix. Multiple fingers may allow CLAMP to be flexible and tolerate more changes in its *cis*-element binding sites. Interestingly, some MREs have a C at the 3rd or 4th position with

the core GA-rich element, which is thought to promote MSL2 interaction (Fauth et al. 2010). Therefore, it is possible that after GAF and CLAMP open chromatin, CLAMP associates with MSL complex to increase the local concentration of MSL complex such that it can physically interact with DNA. CLAMP and GAF are maternally loaded into the early embryo before zygotic genome activation (Harrison and Eisen 2015), so CLAMP and GAF are potentially bound to chromatin before the formation and recruitment of MSL complex, supporting this order of action.

In addition to their functions at CES, CLAMP and GAF also compete at many promoters genome-wide, which could modulate the recruitment of large transcription complexes such as the NELF complex. In fact, we observed that CLAMP reduces GAF occupancy at NELF-regulated genes but not at unregulated genes. CLAMP and GAF have opposite effects on recruitment of NELF at selected target genes (J. A. Urban, Urban, and Larschan 2017) (Submitted), therefore one of the mechanisms by which these factors regulate transcription is through modulating the occupancy of NELF, which regulates RNA Pol II pausing.

It is possible that the more ancient role for CLAMP across *Diptera* could be to regulate transcription by modulating GAF occupancy. The recent enrichment for non-exact GA-repeat containing *cis*-elements on the X-chromosome then promoted CLAMP over GAF occupancy at CES (Kuzu et al. 2016) (Chapter 2). MSL complex, which is less conserved than CLAMP and may be evolving more rapidly, then acquired the ability to physically associate with CLAMP and not GAF (C. I. Wang et al. 2013; Lomaev et al. 2017). Overall, we have provided new insight into how competition between two similar transcription factors promotes a specialized domain of coordinate gene regulation.

Materials and Methods

Fly stocks:

All flies were maintained in a 25°C incubator. Flies for PEV assay include the following stocks: white mottled 4 flies ($ln(1)w^{m4};++$ Bloomington stock center #807), modulo mutant flies ($w^{+};+;mod^{L8}/TM3$, Sb Bloomington stock center #38432), CLAMP mutant flies ($w^{1118};clamp^2/cyoGFP;+$ (J. A. Urban, Doherty, et al. 2017), GAF mutant flies ($w^{1118};+; Trl^{13C}/TM6B$, Tb, Sb Bloomington stock center #58473), and the double-heterozygous CLAMP and GAF mutants were generated from the above stocks to a final genotype of: $w^{1118}; clamp^2/cyoGFP; trl^{13C}/TM6B$, Tb. Flies for polytene spreads are a control GFP RNAi line ($w^{1118}; P\{UAS-GFP.dsRNA.R\}142$ Bloomington stock center #9330).

Position Effect Variegation assay:

Female white mottled flies were crossed to male mutant flies, larvae and adult flies were sorted to select for male heterozygous mutants, retention of mutation indicated by loss of balancer allele. After two days, images of eye colors were taken of flies on a CO₂ pad using an OptixCam Summit D3K2-5 camera mounted on an Olympus SZX12 scope with an Amscope LED 144 ring light adaptor. Each replicate was collected as a group of five flies, flies were put into 2.0 mL microcentrifuge tubes and flash frozen in liquid nitrogen. Four replicates were collected for a total of 20 flies. Frozen flies were decapitated by agitation, or razor blade if needed. Five heads were added to a fresh tube containing a steel bead and 250 uL pigment assay buffer (3% HCl in ethanol). Fly heads were then homogenized for 1 min at a frequency of 30 per second using a bead mill mixer (Retsch MM300). Heads were incubated at 50°C for 10 minutes, then centrifuged at max speed for 10 minutes at 22°C. ~200 uL of supernatant was transferred to a fresh tube. 150 uL was then transferred to a clear 96-well plate, absorbance at 480 nm was measured on a

plate reader. A pigment assay buffer-only well absorbance value was subtracted from sample-well readings.

Immunostaining of Polytenes:

Polytene squashes were prepared from male third instar larvae as in (W. Cai et al. 2010). For immunostaining, first blocking was performed in 0.5% Bovine Serum Albumin in Phosphate buffered saline (1X PBS) for 1 hr. Primary antibodies were incubated overnight at 4°C. We used a rabbit anti-CLAMP antibody (SDIX, (Soruco et al. 2013)), at a 1:1000 dilution, a rabbit anti-GAF antibody (gift of G. Cavalli) at 1:1000, a goat anti-MSL3 serum (gift of M. Kuroda) at a dilution of 1:500. Slides were then washed in 1X PBS three times for 15 min each. Slides were incubated in secondary antibodies for 2 hr at 22°C in a dark humid chamber. Secondary antibodies were donkey anti-rabbit Alexafluor 488 and donkey anti-goat Alexafluor 594 (Thermofisher Scientific). Slides were counterstained with Hoescht for 10 sec, then washed three times in 1X PBS for 10 min. Slides were mounted with Prolong Gold mounting medium and imaged on a Zeiss Axio Imager M1 Epifluorescence upright microscope using AxioVision version 4.8.2 software.

Cloning for GAF DNA binding domain protein:

The GAF DNA binding domain and surrounding sequence (AA 244-407) were ordered as *E. coli* codon optimized cDNA (Genewiz gene synthesis, P2313-1/C78696). Restriction sites were added on through PCR with the following forward primer (NdeI): 5'-cttcagggtcatatgAGTGGTAGTGTGCAGCAG-3' and reverse primer (XhoI): 5'-gtgggtgtgctcgagCTATGCACCATCACTACCAAC-3'. Restriction site cloning for NdeI and XhoI was then used to insert the GAF DNA binding domain into the pET-THMT

expression vector (Peti and Page 2007) containing a His-MBP fusion tag sequence. The vector was transformed into Bioline BL21 (DE3) cells for *E. coli* expression.

Expression and purification of MBP-GAF DNA binding domain:

Cells were grown in LB incubated at 37°C with shaking until reaching an OD of ~0.6-0.9. Expression was then induced for 4 hours by adding IPTG to a final concentration of 1 mM and incubating at 37°C with shaking. Cells were pelleted at 6,000 rpm and frozen in liquid nitrogen for storage at -80°C. Pellet of cells expressing MBP-GAF was resuspended in buffer A (20 mM Tris-Cl, 500 mM NaCl, 10 mM imidazole, 1 mM DTT, pH 8.0, 0.1 mM ZnCl). Cells were mechanically lysed with an EmulsiFlex C3 (Avestin) and centrifuged at 20,000 rpm for 1 hr. Lysate was then filtered through a 0.22 µm filter and run on a 5 mL HisTrap HP column (GE Healthcare). For column elution, buffer B was used (20 mM Tris-Cl, 500 mM NaCl, 500 mM imidazole, 1 mM DTT, pH 8.0, 0.1 mM ZnCl). Fractions were collected and analyzed by SDS-PAGE. Fractions containing protein were combined and further purified by size-exclusion chromatography (SEC) using a Superdex 200 (GE Healthcare) equilibrated with SEC buffer (20 mM Tris-Cl, 1 mM DTT, 150 mM NaCl, pH 8.0, 0.1 mM ZnCl). 1 mL fractions were collected and analyzed via SDS-PAGE. Fractions containing MBP-GAF were combined and concentrated to 40 µM. Aliquots were flash frozen in liquid nitrogen and stored at -80°C until use.

Electrophoretic Mobility Shift Assay:

The LightShift Chemiluminescent EMSA Kit (Thermo Scientific) was used according to manufacturer protocols and as previously reported (Kuzu et al. 2016) (Chapter 2). Biotin-labeled DNA probes containing (GA)_n-repeats were the same as used previously (Kuzu et al. 2016) (Chapter 2), and were added to a final concentration of 0.6 nM in each

sample reaction. The first protein sample was the MBP-CLAMP DNA binding domain from the same protein preparation as in previously published gel-shift experiments (Kuzu et al. 2016) (Chapter 2). The other protein was the MBP-GAF DNA binding domain prepared above. Images were captured on an Azure c600 imaging system.

Chromatin Immunoprecipitation:

S2 cells used for chromatin preparation were treated with previously validated RNAi targets (*gfp* (Hamada et al. 2005), *clamp* (Larschan et al. 2012; Soruco et al. 2013), *trl* (Fuda et al. 2015)), for 6 days as described in previous work (Kaye et al. 2017) (Chapter 4). Knockdown was validated via western blotting with the Western Breeze kit (Invitrogen). Antibodies used for detection were rabbit anti-CLAMP (SDIX, (Soruco et al. 2013) at 1:1000, and rabbit anti-GAF (Gift of G. Cavalli) at 1:5000. As a loading control, actin was detected with mouse anti-actin (Sigma) at a 1:50,000 dilution. Chromatin was prepared as described in (Kaye et al. 2017) (Chapter 4). Immunoprecipitation was performed in (Kaye et al., 2017) (Chapter 4) with 2 mL IPs of CLAMP (2uL/mL SDIX), GAF (10 uL/mL Gift of J. Lis), or MSL3 serum (.4 uL / mL, Gift of M. Kuroda). Sequencing libraries were prepared using the NEBNext Ultra II DNA library preparation kit according to user manual protocol. For sequencing, NEBNext index adaptors 1-12 were used, with each replicate set as its own lane.

Bioinformatics:

Sequencing reads were mapped against FlyBase release 5.22 reference assembly with bowtie release 0.12.9 with the --solexa-quals and --best options. Reads with more than one match were not aligned. Alignment statistics are reported in Table S3.2.

Peaks were called using MACS2 version 2.1.1 with p cutoff of 0.05, -c input-seq sample bam file, and -t ChIP-seq sample bam file (for GAF and CLAMP-IPs). To define the peaks used in the heat map and venn diagrams, GAF and CLAMP peaks with a negative $\log_{10} \geq 1$ q-score (equivalent to an FDR ≤ 0.1) were included. If the summits of the peaks were within 100-bp, the peaks were merged into a single peak. IP Fold Enrichment tracks (for GAF, CLAMP, and MSL3) were created with MACS2 bdgcmp and the -m FE option and -c input-seq sample bam file, and -t ChIP-seq sample bam file. Mean enrichments across replicates were calculated.

The resulting peaks were considered CLAMP peaks if the maximum fold enrichment was ≥ 3 , GAF peaks if the maximum fold enrichment was ≥ 3 , and MSL3 peaks if the fold enrichment was ≥ 1.8 . Values for cut offs were selected from Figure S3.1, where replicates became consistent with each other (Fraction replicating > 0.8).

Enrichment profiles were generated across individual regions of interest by averaging scores at each base pair across experimental replicates from MACS2-generated enrichment bedGraph files. Next, these scores were used to generate a mean score at each base pair across genomic regions of the given type (i.e., CES or CLAMP peaks) and a 95% standard error confidence interval for that mean. Heat maps were generated across 1000 base pair genomic regions centered on GAF and/or CLAMP peak summits using mean enrichment across experimental replicates at any given position. Difference maps were generated by subtracting the *gfp* treated sample IP enrichment scores from the *trl* or *clamp* treated sample IP enrichment score.

Transcription data was available from (Core et al. 2012; Fuda et al. 2015; J. A. Urban, Kuzu, et al. 2017) (in revision), and separated into categories as follows: “positively

regulated” genes decreased by greater than 1 SD after reduction (RNAi treatment) of factor of interest (e.g. CLAMP in the Start-seq data set, GAF or NELF for GRO-seq data sets), “negatively regulated” genes increased by greater than 1 SD after reduction (RNAi treatment) of factor of interest, and finally “unchanged” genes that remained within + or – 1 SD after reduction of factor of interest. Box plots were generated in R package “boxplot” and include notches that represent (+/-) 1.58 times the inter-quartile range divided by the square root of the sample size. p-values generated from a paired Student's t-test.

Supporting Information

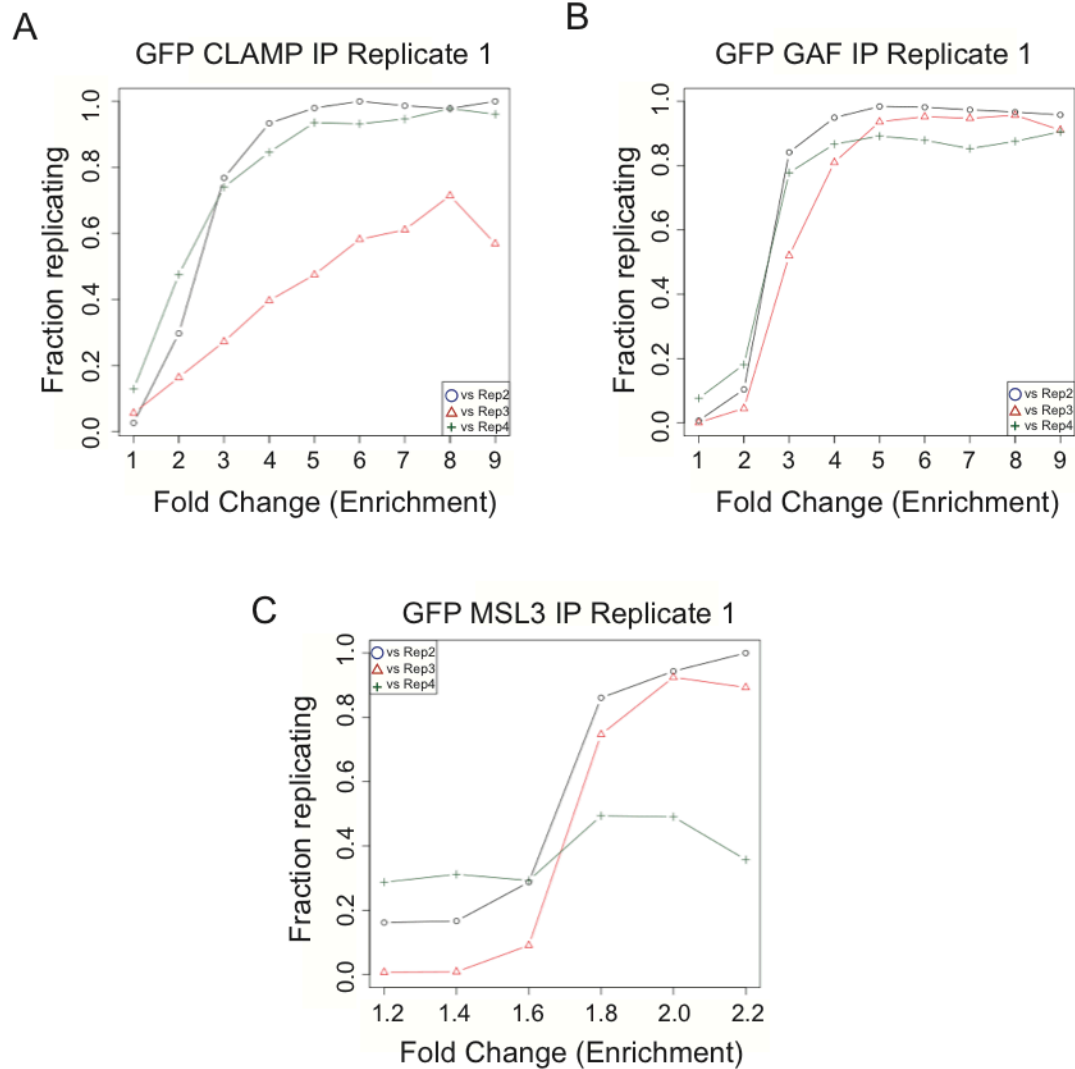


Figure S3.1

Plots of replicate 1 versus replicates 2,3, and 4 to compare the fraction of peaks replicating versus enrichment level for CLAMP (A), GAF (B), and MSL3 (C) ChIP-seq under control (*gfp*) RNAi. From these data, peaks were called for enrichment scores matching a fraction replicating of 0.8 or higher.

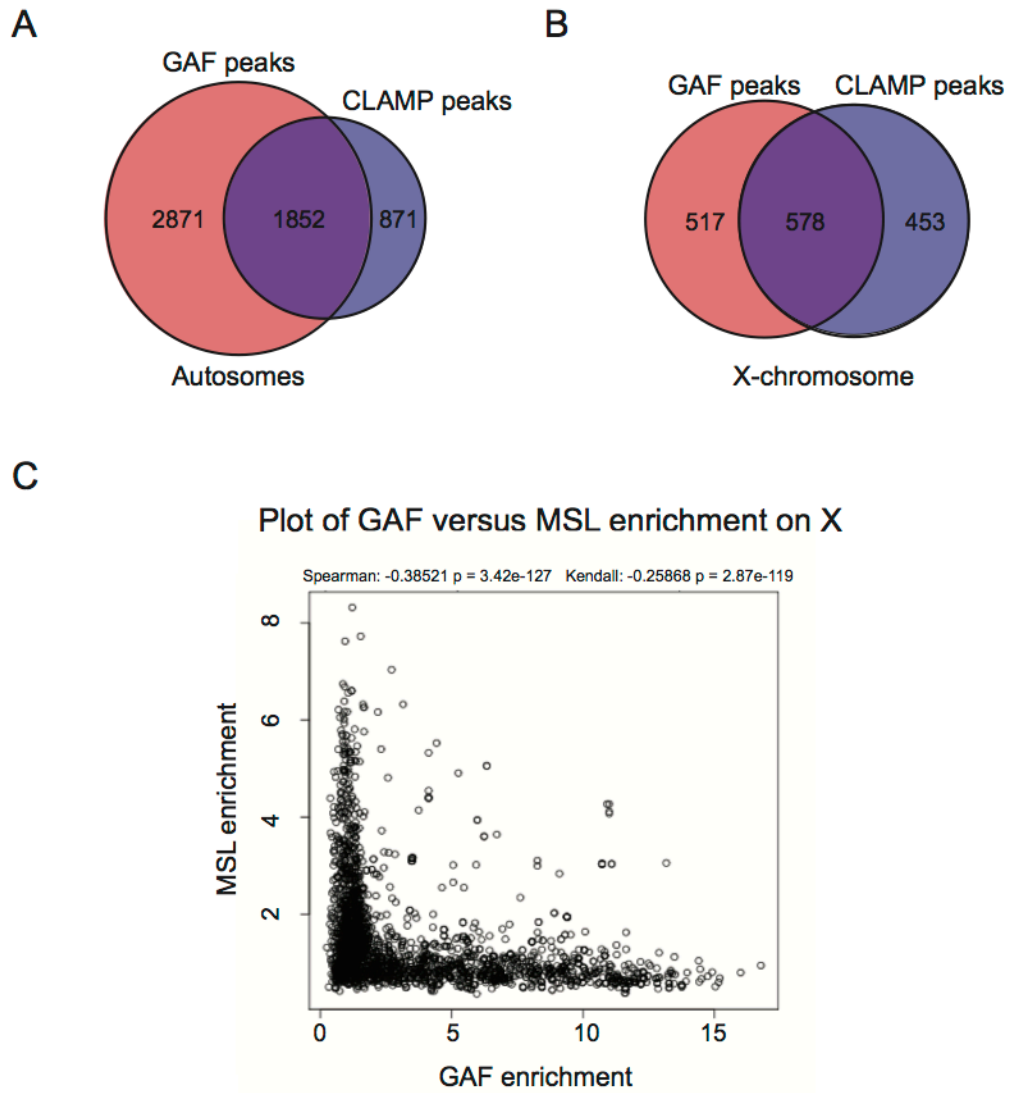
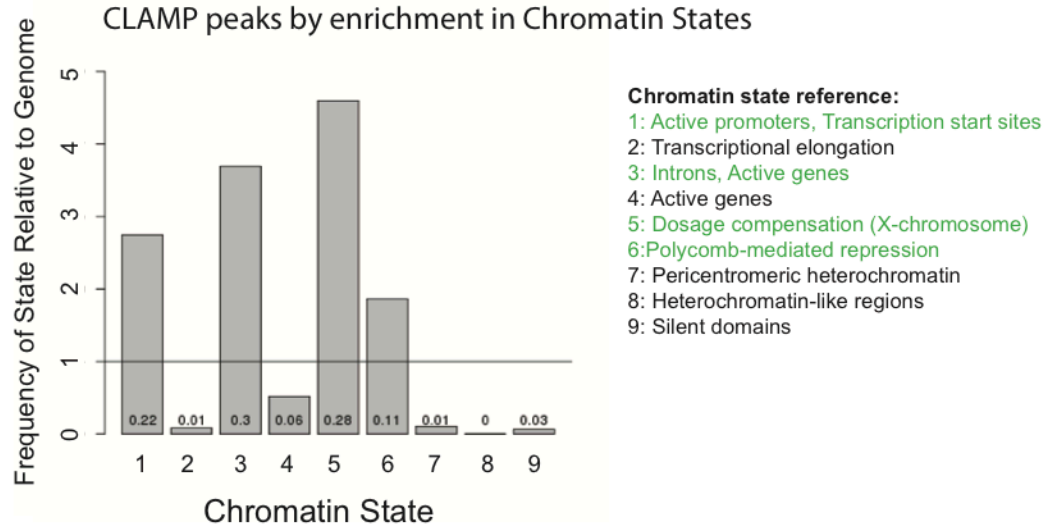


Figure S3.2

A and B. Comparing overlap of binding sites between GAF and CLAMP. Venn diagram showing number of unique GAF (red) and CLAMP (blue) peaks or overlapping peaks (purple). Peaks are separated based by location on autosomes (**A**) or the X-chromosome (**B**). **C.** Scatter plot of maximum GAF enrichment scores (x-axis) and maximum MSL3 enrichment scores (y-axis) in the regions depicted in panel 2E. MSL3 and GAF exhibit a negative correlation with a Pearson's correlation coefficient of -0.39 with a significance of $p = 3.43 \times 10^{-127}$ using a t-distribution.

A



B

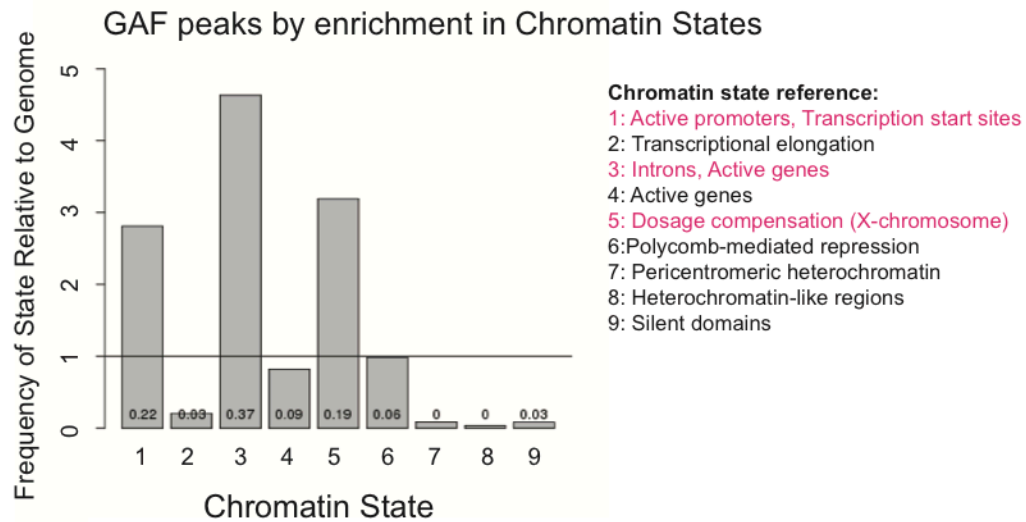


Figure S3.3

Chromatin state enrichment for CLAMP **(A)** and GAF **(B)**. Enrichment presented as frequency compared to abundance of that state in the genome. States are listed to the right summarized from (Kharchenko et al. 2011), with enriched states highlighted in green (CLAMP) or pink (GAF).

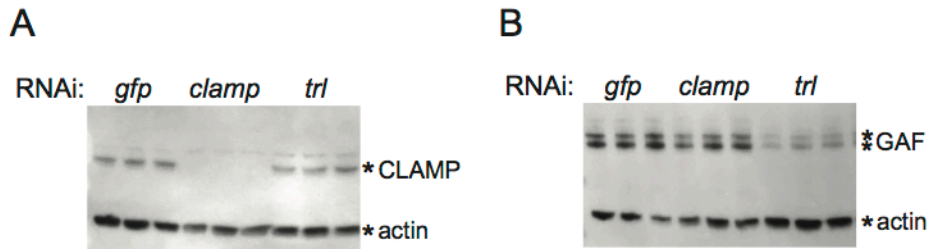


Figure S3.4

Western blotting for CLAMP (A) and GAF (B). Protein samples prepared from S2 cells treated with RNAi targeting *gfp*, *clamp*, or *trl*. Three replicates were loaded on each gel, and protein was detected with either rabbit anti-CLAMP or rabbit anti-GAF antibody, as well as a mouse anti-actin loading control. Specific bands detected are indicated by an asterisks. Note that this antibody detects both isoforms of GAF.

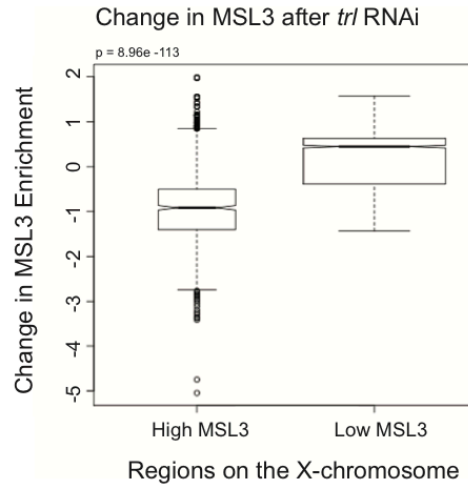


Figure S3.5

Boxplot of the difference in MSL enrichment after *trl* RNAi compared to control *gfp* RNAi. Peaks are separated into two categories based on high wild type enrichment (greater than 1.8) or low MSL3 enrichment (less than 1.8).

Table S3.1 P-values from student's t-test of PEV data assuming unequal variance.

	<i>w^{m4};+; Mod^{L8}</i>	<i>Trl^{13C}</i>	<i>clamp²</i>	<i>clamp²; Trl^{13C}</i>
<i>Trl^{13C}</i>	1.92 e -2	-	-	-
<i>clamp²</i>	1.84 e -3	1.10 e -2	-	-
<i>clamp²; Trl^{13C}</i>	9.72 e -4	2.39 e -3	4.52 e -2	-

All alleles are heterozygous over wild type. - = redundant comparison.

Table S3.2 Aligned reads from each S2 ChIP-seq sample.

Sample	Reads	Alignments
GFP Input R1	19218421	13,630,217
GFP CIP R1	25437957	17,703,752
GFP GIP R1	19411300	14,265,721
GFP MIP R1	25376516	18,146,111
GAF Input R1	23397826	17,282,427
GAF CIP R1	17869683	12,269,533
GAF GIP R1	27868552	19,730,344
GAF MIP R1	29437758	21,835,078
CLAMP Input R1	6525567	1,651,091
CLAMP CIP R2	29502913	19,919,643
CLAMP GIP R2	22527570	15,683,734
CLAMP MIP R2	22696988	15,898,513
GFP Input R2	4468053	1,500,751
GFP CIP R2	27442936	18,934,987
GFP GIP R2	25924790	14,426,424
GFP MIP R2	18009319	13,458,480
GAF Input R2	8780963	5,904,853
GAF CIP R2	30860817	21,463,377
GAF GIP R2	16204495	12,548,896
GAF MIP R2	29461218	23,146,958
CLAMP Input R2	2033630	732,493
CLAMP CIP R3	12835320	9,734,810
CLAMP GIP R3	28263295	19,607,145
CLAMP MIP R3	26991232	18,885,028
GFP Input R3	6702486	2,087,432
GFP CIP R3	24256110	18,641,930
GFP GIP R3	28602646	19,344,367
GFP MIP R3	26288689	18,174,554
GAF Input R3	4636303	1,357,768
GAF CIP R3	34376624	23,327,186
GAF GIP R3	10792374	4,826,453

GAF MIP R3	24357846	17,369,589
CLAMP Input R3	17892526	12,574,408
CLAMP CIP R4	31833990	21,434,420
GFP Input R4	15415451	6,228,407
GFP CIP R4	28739336	23,000,396
GFP GIP R4	20056467	6,898,070
GFP MIP R4	3029004	1,075,554
GAF Input R4	17571339	11,792,293
GAF CIP R4	17136070	7,422,095
GAF GIP R4	9056573	2,736,732
GAF MIP R4	35271326	23,766,572
CLAMP Input R4	33401306	24,691,475
CLAMP CIP R1	21356281	14,590,584
CLAMP GIP R1	18987942	13,255,475
CLAMP MIP R1	19585467	13,347,245

RNAi treatment, followed by input or IP antibody: “CIP” = CLAMP IP, “GIP” = GAF IP, “MIP” = MSL3 IP. Replicates are indicated as Replicate 1 = R1, 2 = R2, 3 = R3, and 4=R4.

Chapter 4 : *Drosophila* dosage compensation loci associate with a boundary forming insulator complex.

This chapter is a manuscript that has been accepted by MCB. It is reproduced here with minor formatting changes.

Emily G. Kaye*, Amina Kurbidaeva*, Daniel Wolle, Tsutomu Aoki, Paul Schedl, and Erica Larschan. *Accepted, MCB July 2017.* *Co-first authors.

Conceptualization, E.G.K., A.K., D.W., P.S., and E.N.L.; Investigation, E.G.K., A.K., D.W., and T.A.; Writing- Original Draft, P.S., and E.N.L; Writing-Review & Editing: E.G.K., A.K., P.S., and E.N.L.; Funding Acquisition, P.S. and E.N.L.; Supervision: P.S. and E.N.L.

Abstract

CES (chromatin entry sites) are 100-1,500 bp elements that recruit MSL (Male Specific Lethal) complexes to the X-chromosome to upregulate expression of X-linked genes in male flies. CES contain one or more ~20 bp GA rich sequences called MREs (MSL recognition elements) that are critical for dosage compensation. Recent studies indicate that CES also correspond to boundaries of X-chromosomal TADs (topologically associated domains). Here we show that a ~1,000 kDa complex called the Late Boundary Complex (LBC), which is required for the functioning of the Bithorax complex boundary *Fab-7*, interacts specifically with a special class of CES that contain multiple MREs. Mutations in the MRE sequences of three of these CES that disrupt function *in vivo* abrogate interactions with the LBC. Moreover, reducing the levels of two LBC components compromises MSL recruitment. Finally, we show that several of the CES that are physically linked to each other *in vivo* are LBC interactors.

Keywords

Insulators, boundaries, TADs, dosage compensation, *Drosophila*, Male Specific Lethal complexes, chromatin entry sites, MSL Recognition Elements, CLAMP, GAGA Factor

Introduction

The three-dimensional organization of eukaryotic chromosomes plays a central role in diverse processes ranging from gene regulation to recombination and repair (Hnisz et al. 2016; Chetverina et al. 2014). Chromatin conformation capture (CCC) experiments have shown that a key organizing principle is the subdivision of chromosomes into a series of looped domains called TADs (topologically associated domains) (Dixon, Gorkin, and Ren 2016; Dekker and Heard 2015).

Though the scale of the loops is smaller, a similar subdivision of chromosomes into looped domains is present in *Drosophila* and GA-rich *cis*-elements are present at the base of most loops (Eagen, Aiden, and Kornberg 2017). In both mammals and *Drosophila*, the end-points of the loops are delimited by special architectural elements called boundaries or insulators. Boundaries in flies generate loops by pairing with each other. For example, pairing interactions between the two boundaries, *homie* and *nhomie*, that flank the *even-skipped* (*eve*) gene generates a ~16 kb loop that encompasses *eve* and all of its regulatory elements (Fujioka et al. 2016). In addition to determining chromosome architecture, pairing interactions between boundaries have regulatory consequences. When interposed between enhancers/silencers and a target gene, boundary factors block regulatory interactions (Kellum and Schedl 1992) or insulate regions of the genome against chromosomal position effects (Kellum and Schedl 1991).

In addition to insulation, pairing interactions between boundaries can also facilitate long distance gene regulation. In flies, this phenomenon was first observed for the *su(Hw)* and *Mcp* boundaries carried by transgene reporters inserted at distant sites (Vazquez and Schedl 1994; Sigrist and Pirrotta 1997; Muller et al. 1999). Pairing has subsequently

been documented for many other boundaries in transgene boundary bypass experiments and in experiments in which boundaries, reporters, and enhancers are placed at distant locations using site-specific insertion techniques (Muravyova et al. 2001; Y. Cai et al. 2010; Fujioka et al. 2016). For example, pairing interactions between the *eve homie* and *nhomie* boundaries can bring enhancers and reporters together over distances ranging from 140 kb to 2 Mb.

Another context in *Drosophila* in which communication over large chromosomal distances is thought to be critically important is dosage compensation where a coordinated domain of active gene expression is established across the entire X-chromosome (J. Belote and Lucchesi 1980; Gelbart and Kuroda 2009; Keller and Akhtar 2015). In *Drosophila*, expression of X-linked genes in males is upregulated to equalize dosage with the two X-chromosomes in females. Dosage compensation depends upon a MSL (Male Specific Lethal) dosage compensation complex that contains five proteins (MSL1, MSL2, MSL3, MOF and MLE) and one of two non-coding RNAs *roX1* or *roX2* (RNA on the X) (Meller and Rattner 2002; Gelbart and Kuroda 2009). MSL complexes associate with actively transcribed genes on the male X-chromosome, acetylating histone H4 lysine 16 (Smith, Allis, and Lucchesi 2001) and increasing the amount of active RNA Polymerase II over gene bodies (Larschan et al. 2011).

At least two steps are involved in targeting the dosage compensation machinery to active X-linked genes. First, MSL complexes are recruited to the X-chromosome by special 100-1,500 bp *cis* acting elements called Chromatin Entry Sites (CES) or High Affinity sites (HAS) (R L Kelley et al. 1999; Kageyama et al. 2001; A. Alekseyenko et al. 2008). Next, MSL complexes spread from CES to the bodies of actively transcribed genes (A A Alekseyenko et al. 2006; Gorchakov 2009). Most CES contain one or more

smaller *cis* elements called MSL recognition elements (MREs) (A. Alekseyenko et al. 2008). MREs are ~21-bp GA-rich sequences that include sequences that are similar to the recognition motif for the well-studied *Trithorax-like* GAGA factor (GAF), GAGAG (Adkins, Hagerman, and Georgel 2006). Two of the loci most enriched for MSL complex are the CES within the *roX1* and *roX2* encoding genes (Meller and Kuroda 2002; Oh, Park, and Kuroda 2003; R L Kelley et al. 1999) that both contain multiple MRE sequences.

Recent experiments reported by Ramirez et al. (2015) have shown that many of the TAD boundaries on the X-chromosome correspond to CES. As boundary elements, CES could provide hubs for the recruitment of MSL complex and promote their bi-directional spread from the base of each loop towards active genes located within the loop. Consistent with an underlying structural role in organizing the topology of the X-chromosome, the three-dimensional organization of the X-chromosome is the same in male and female cells as is the co-localization of CES with loop boundaries. Moreover, depleting MSL2 or MSL3 in male cells does not induce major alterations in the looping pattern. The only clear difference in topological organization of male and female X-chromosomes is in the frequency of long distance interactions between *roX1* and *roX2* and sites elsewhere on the X-chromosome, which is consistent with their male-specific transcriptional activity.

Other observations are consistent with the idea that CES are pre-existing scaffolds for MSL recruitment and likely have properties in common with architectural elements elsewhere in the genome. Thus far, two zinc finger DNA binding proteins have been implicated in dosage compensation and/or CES function. One of the zinc finger proteins is Chromatin-linked adaptor for MSL proteins (CLAMP). CLAMP is present at all CES

where it directly binds to MRE elements and is required for MSL complex recruitment (Larschan et al. 2012; Soruco et al. 2013; Kuzu et al. 2016). The other zinc finger protein is GAF. Like CLAMP, GAF also recognizes GA-rich motifs including MREs. However, the connection between GAF and CES is so far only indirect. *Trans*-heterozygous combinations of strong and weak mutations in the GAF gene, *Trithorax-like (Trl)*, have enhanced lethal effects in males compared to in females (Greenberg, Yanowitz, and Schedl 2004). Moreover, the male-specific lethal effects of *Trl* are related to the dosage compensation machinery because they can be enhanced when the male flies are also heterozygous for mutations in one of the MSL complex encoding genes. While these findings link CLAMP and perhaps also GAF to CES, both proteins have functions beyond their involvement in dosage compensation. GAF and CLAMP are essential proteins in both males and females (J. Urban et al. 2016) and recognize several thousand GA-rich sites on the X and autosomes (Soruco et al. 2013).

Of potential relevance to the link between CES and TAD boundaries are the architectural functions of the GAF protein. GAF was first implicated as a boundary factor in transgene assays (Ohtsuki and Levine 1998; Belozarov et al. 2003). More recently, GAF was found to be a component of a very large (>780 kDa) protein complex, called the Late Boundary Complex (LBC) that is critical for the functioning of a *Drosophila* Bithorax complex boundary, *Fab-7* (Wolle et al. 2015; S. E. Schweinsberg and Schedl 2004). Though the identity of all of the different components of the LBC are not yet known, antibody “supershift” electrophoretic mobility shift (EMSA) experiments indicate that it contains three proteins, GAF, Mod(mdg4) and E(y)2. While GAF and Mod(mdg4) are found associated with the *Fab-7* boundary element using chromatin immunoprecipitation (ChIP) experiments, these are not the only proteins that localize to

this boundary. In fact, one of the other DNA binding proteins found at *Fab-7 in vivo* is CLAMP.

Here, we show that CES share properties with previously characterized autosomal chromatin boundaries. We demonstrate that the *Fab-7* boundary factor LBC binds specifically to several X-linked CES that contain multiple MREs, including *roX1*, *roX2* and CES 5C2. Furthermore, LBC binding to these CES is abrogated by mutations in the MREs that disrupt CES activity *in vivo*. Connecting the architectural functions of CES to the LBC, we show that the LBC binds to CES that are known to interact with each other *in vivo*. We present biochemical evidence that CLAMP is a component of the LBC and use RNAi depletion experiments to show that CLAMP and GAF bind interdependently *in vivo*. Moreover, we show that CLAMP and GAF are both important for the recruitment of MSL complexes to *roX1*, *roX2* and CES 5C2. Overall, we provide a novel link between boundary forming protein complexes and dosage compensation.

Results

CLAMP binds to the *Fab-7* boundary *in vivo* and *in vitro*

The *Fab-7* boundary spans a DNA segment that includes three chromatin specific nuclease hypersensitive sequences, HS*, HS1 and HS2 (Figure 4.1B) (J. Zhou et al. 1996; Rodin et al. 2007; Hagstrom, Muller, and Schedl 1997). While all of these sequences contribute to boundary activity, replacement experiments indicate that HS1 alone is sufficient for full function in an otherwise wild type background (Wolle et al. 2015). Molecular and biochemical experiments have shown that one of the factors responsible for HS1 boundary activity is the LBC (Wolle et al. 2015; S. E. Schweinsberg and Schedl 2004). The LBC binds to the distal half of HS1, dHS1. This 240 bp sequence

has 4 GAGAG motifs: GAGA3, GAGA4, GAGA5, and GAGA6 (Figure 4.1B). Probes spanning all four GAGAG motifs, pairs of GAGAG motifs (GAGA3+4, GAGA5+6) or single GAGAG motifs (GAGA3, GAGA4, GAGA5 but not GAGA6) are shifted by the LBC (Wolle et al. 2015). Deletion and mutational analysis of the GAGA3 and GAGA4 probes showed that the minimal length required for LBC binding is 65 bp and the GAGA motifs are essential. Consistent with these *in vitro* findings, mutations in the GAGA3 and GAGA4 motifs disrupt *Fab-7* boundary function *in vivo* (S. Schweinsberg et al. 2004). Not surprisingly given the importance of the GAGA motifs, EMSA shifts generated by the LBC were found to be supershifted by GAF antibodies. Antibodies against Mod(mdg4) and E(y)2 also supershifted the LBC shift; however, antibodies against other known boundary proteins, such as Beaf, Zw5 and dCTCF had no effect. For this reason, we sought other potential LBC associated proteins.

One plausible candidate is CLAMP. Like GAF, CLAMP recognizes GA-rich motifs and ChIP-seq experiments reveal that many of the sequences bound by GAF and CLAMP on the X-chromosome and autosomes closely overlap (Soruco et al. 2013). In fact, one of the sites occupied by both CLAMP and GAF is *Fab-7* (Figure 4.1A, Figure S4.1). Aside from *Fab-7*, CLAMP localizes to other sequences within the Bithorax complex locus. Some of the CLAMP bound loci are shared with GAF, while others are unique. Included among the common sites are two boundaries, *Fab-8* and *AB-I*, from the *Abd-B* region of the complex which both have LBC recognition sequences (Wolle et al. 2015). Like *Fab-7*, *Fab-8* and *AB-I* include GAGA motifs.

To determine if the presence of both CLAMP and GAF in the LBC has functional relevance, we asked whether CLAMP and GAF associate with the *Fab-7* boundary interdependently. For this purpose, we used RNAi to reduce levels of either CLAMP or

GAF in tissue culture cells, and then monitored their occupancy at *Fab-7* *in vivo* by ChIP-qPCR. GAF RNAi reduces CLAMP occupancy at *Fab-7* (Figure 4.1C) and conversely, although to a lesser extent, CLAMP RNAi reduces GAF occupancy at *Fab-7* (Figure 4.1D). These data suggest that the CLAMP and GAF associations with *Fab-7* are interdependent.

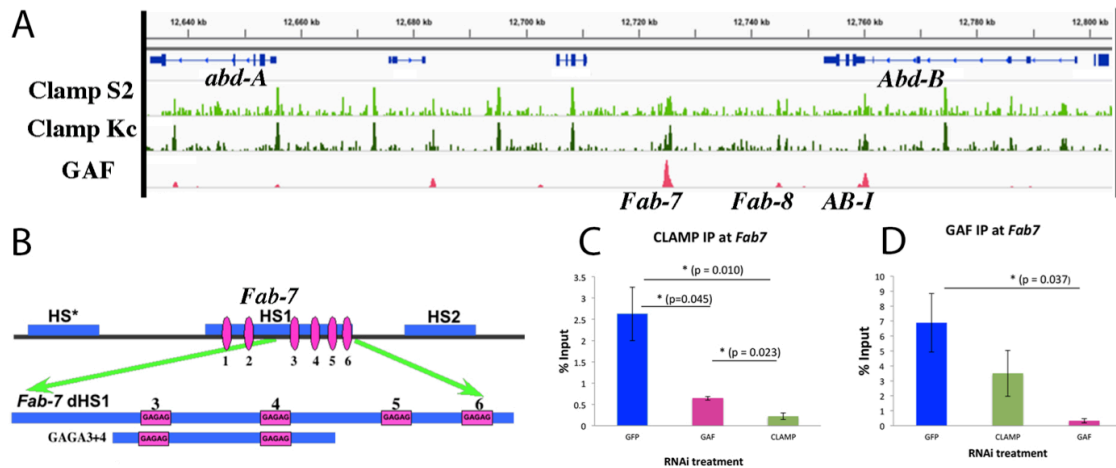


Figure 4.1 CLAMP colocalizes with several boundaries in BX-C.

A) ChIP localization profiles for CLAMP in S2 and Kc cell lines (green) (Sorucu et al. 2013) and GAF (pink) (Fuda et al. 2015), in a 160 Kb region including the *Fab-7*, *Fab-8* and *AB-I* boundaries. **B)** Top: Chromatin organization of the *Fab-7* boundary (1.2 kb) showing nucleosome hypersensitive regions. Distal 240 bp of *Fab-7* HS1 (dHS1) contains 4 GAGAG motifs (pink). Below: The 130 bp GAGA3+4 probe used for EMSA. **C)** and **D)** ChIP followed by qPCR following CLAMP or GAF knockdowns. Error bars represent standard error obtained from 4 (GFP) or 3 (GAF and CLAMP) independent experiments. P-values from t-tests are displayed for significance * $p < 0.05$ or ** $p < 0.01$ differences between sample means.

One plausible explanation for this relationship is that CLAMP is a component of the LBC. To explore this possibility we used antibody supershift EMSA experiments with late stage 6-22 hr embryo nuclear extracts. Figure 4.2A and B show EMSA supershift experiments with the *Fab-7* probe, GAGA3+4, spanning GAGA motifs 3 and 4, and two different CLAMP antibodies. As positive controls we included antibodies against the other known LBC components (E(y)2, GAF, Mod(mdg4)) in these EMSA experiments. The CLAMP antibody used in Figure 4.2A (CLAMP-A) supershifts the GAGA3+4 LBC

shift, while the CLAMP antibody in Figure 4.2B (CLAMP-B) inhibits LBC binding to the probe. For CLAMP-B, we found that the extent of inhibition varies depending upon the ratio of antibody to extract and the duration of time that the antibody is preincubated with the extract (not shown). For the Mod(mdg4) antibody, the inclusion of a preincubation step alters the relative balance between supershifting and inhibition (Figure 4.2A and B). The LBC shift generated by the *Fab-7* GAGA5+6 probe is also supershifted by CLAMP-A antibodies (Figure S4.2). In summary, CLAMP binds to *Fab-7* both *in vivo* and *in vitro*.

CLAMP is an integral LBC component

Our previous gel filtration experiments showed that the LBC exists as a very large pre-assembled complex in nuclear extracts (Wolle et al. 2015). However, we did not determine whether the proteins found associated with the LBC shift in nuclear extracts by antibody supershift experiments co-fractionate with the complex on gel filtration columns. In fact, since the pattern and yield of LBC shifts can differ from one probe to the next (Figure 4.2B, Figure 4.3), it is possible that some factors might not be stably associated with the LBC. To further define the relationship between CLAMP and the other components of the LBC, we size fractionated nuclear extracts on a Superose 6 gel filtration column and assayed odd numbered fractions for DNA binding activity using the GAGA3+4 probe. Figure 4.3A shows EMSA assays of early fractions from the gel filtration column. The peak fraction for the 670 kDa marker is 51 (arrowhead), while the peak fractions for the LBC shift are between 33 and 41. This would be consistent with a molecular weight of approximately 1,000 kDa, which is larger than our previous estimates for the size of the LBC. The mobility of the two shifts generated by the size fractionated LBC (see arrows in Figure 4.3A) is slightly more rapid than the LBC shift in nuclear extracts. This finding indicates that the LBC GAGA3+4 shift detected in nuclear extracts is likely generated by a core complex of integral LBC proteins plus proteins that

are not stably associated with the complex and/or associate with the complex only when the LBC binds DNA.

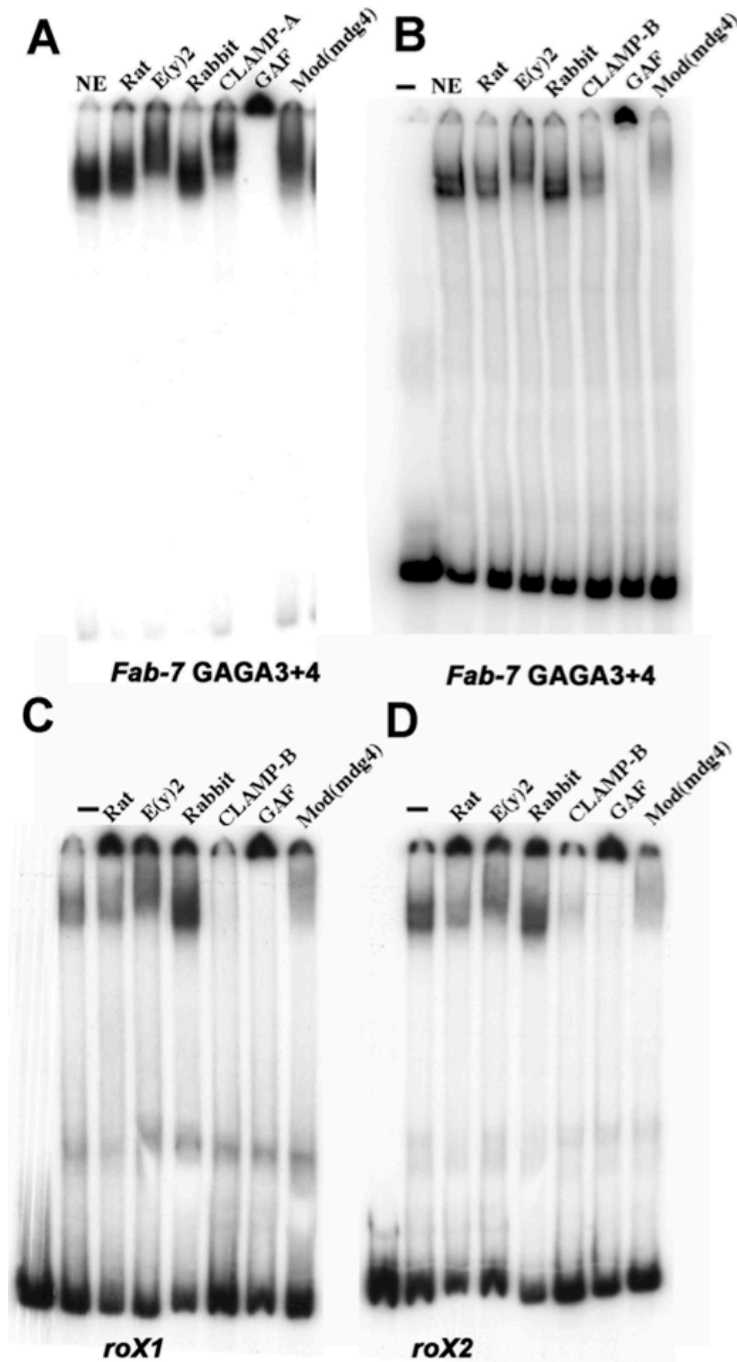


Figure 4.2 LBC binds to *Fab-7 GAGA3+4*, *roX1* and *roX2*.

Probes *Fab-7 GAGA3+4* (A, B), *roX1* (C) or *roX2* (D) were incubated with 6-22 hr nuclear extracts, with or without either a control preimmune rat or rabbit serum, a rat polyclonal E(y)2 antibody, two different rabbit polyclonal CLAMP antibodies (A, B), rabbit polyclonal GAF antibody, or a rabbit polyclonal Mod(mdg4) antibody. (A, B) The nuclear extract, antibody and labeled probe were added together and the reaction mix was then incubated for 30 min prior to

gel electrophoresis. (**C**, **D**) the nuclear extract was first preincubated with the indicated antibody for 30 min, labeled probe was then added and the reaction mix incubated for an additional 30 min prior to gel electrophoresis.

To determine whether CLAMP, GAF and Mod(mdg4) are components of the core LBC complex, two fractions, 35 and 36, that contain the LBC, and a control fraction, 65, distant from the LBC peak, were used for supershift experiments (Figure 4.3B). Like the experiment in Figure 4.2B, we preincubated the antibody with the fractions prior to the addition of labeled probe. As found previously, the addition of serum alone to the peak fractions enhances LBC binding. For the GAF antibody, we observe a supershift like that seen in nuclear extracts. For CLAMP and Mod(mdg4), the antibodies completely block LBC binding. The inhibition of binding seen with the gel filtration fractions likely reflects differences in the relative amount of CLAMP and Mod(mdg4) in extracts compared to that in the column fractions containing LBC activity. Consistent with the inhibition of binding, CLAMP is present in column fractions containing LBC binding activity (see Figure S4.3A).

The LBC binds to the *roX1* and *roX2* CES

The *in vivo* localization of CLAMP to the *Fab-7* boundary and its presence within the LBC raises the possibility that CES on the X-chromosome might utilize this same boundary complex. To test this hypothesis we first focused on two canonical CES, *roX1* and *roX2*. As shown in Figure 4.4B, the *roX1* CES has three MREs, while the *roX2* CES has four. The central *roX1* MRE has two GAGAG motifs, while the two flanking MREs each have a single GAGAG motif. For *roX2* two of the four MREs have a single GAGAG motif. Both CES have an additional GAGAG motif (*) that is not classified as an MRE because the flanking sequences do not match the 21-bp MRE consensus sequence.

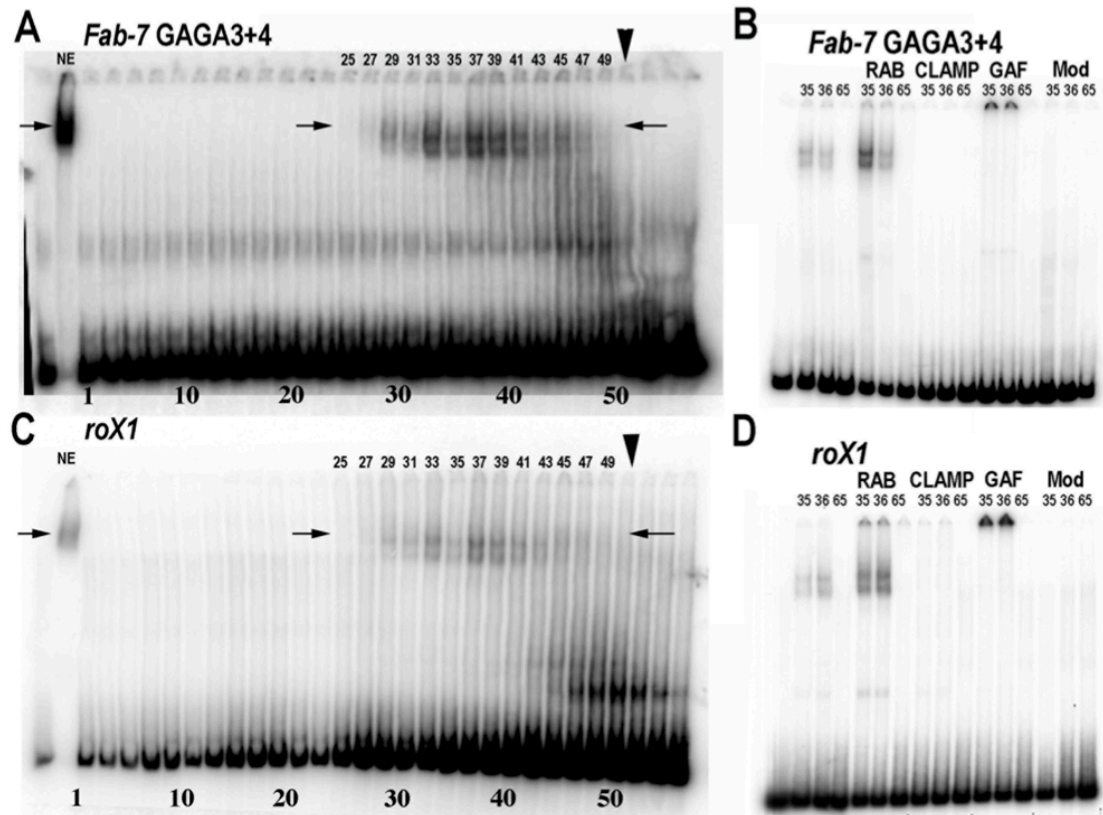


Figure 4.3 The LBC is a >1000 kDa complex that contains CLAMP, GAF, Mod(mdg4).

Nuclear extracts were fractionated on a Superose 6 column and then used for EMSAs with *Fab-7* GAGA3+4 (**A**, **B**) or *roX1* (**C**, **D**) probes. Panels **A** and **C**: 10 μ l of odd numbered column fractions (1-57) were incubated with GAGA3+4 (**A**) or *roX1* (**C**). The peak fraction for the 670 kDa marker used as a standard is indicated by the arrowhead. Arrows indicate center position of LBC shift in nuclear extracts. Panels **B** and **D**: EMSAs using 10 μ l of indicated column fractions preincubated with no antibody, rabbit serum, CLAMP-B, GAF or Mod(mdg4) antibodies for 30 min. The labeled probe was then added, and the incubation continued for 30 min prior to gel electrophoresis.

We first examined the ChIP profiles of CLAMP, GAF and Mod(mdg4) in the regions around the *roX1* and *roX2* genes. Within the resolution of available genome-wide data sets, GAF and Mod(mdg4) colocalize with CLAMP at the CES located within the *roX1* and *roX2* genes *in vivo* (Figure 4.4A). Since all three of the known DNA binding components of the LBC (CLAMP, GAF and Mod(mdg4)) associate with the *roX1* and *roX2* CES *in vivo*, we next tested for LBC binding *in vitro*. We generated 200 bp probes spanning the core *roX1* and *roX2* CES DNA sequences. We also subdivided each CES into three smaller ~100 bp overlapping probes (Figure 4.4B). Figure 4.4C shows EMSAs

with these CES probes and, as a positive control, the 134 bp *Fab-7* GAGA3+4 probe. For the two full-length *roX1* and *roX2* CES probes, we observe two major slowly migrating bands whose mobility closely match those seen for the GAGA3+4 LBC shift. For the smaller sub-fragments, two of the *roX1* and two of the *roX2* probes generate shifts that co-migrate with the GAGA3+4 LBC shifts. The relative yield and precise pattern of shifted bands differ with these smaller probes. For *roX1*, the highest yields are for probe B, which has MRE2 plus part of MRE3, and probe C, which has MRE2 plus MRE3. For *roX2* the highest binding efficiency is seen for probe A, which has three MREs. By contrast, the LBC binds poorly to probes with only a single MRE (*roX1*-A and *roX2*-C). Similar variations in the relative yield and pattern of bands generated by the LBC was also noted when smaller *Fab-7* probes were used (Cleard et al. 2017; Wolle et al. 2015). In addition to the LBC shift, the smaller probes give a collection of faster migrating shifts.

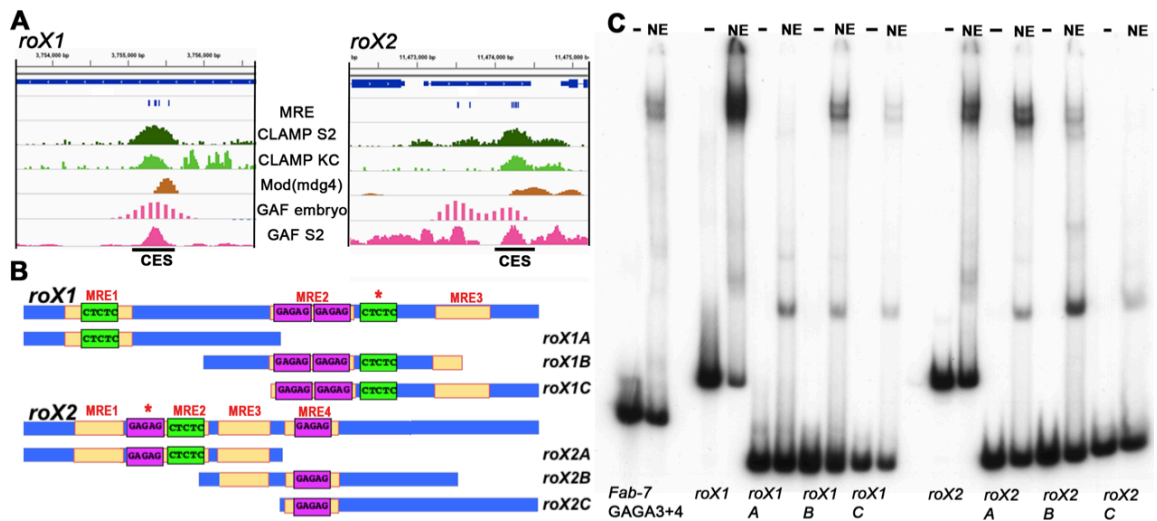


Figure 4.4 The LBC can bind to *roX1* and *roX2* CES.

(A) ChIP profiles of *roX1* and *roX2* for CLAMP (Sorucu et al. 2013), GAF (Fuda et al. 2015), and Mod(mdg4) (ChIP-chip) (modENCODE) in S2 and Kc cells and in embryos (ChIP-chip). (B) A schematic drawing of the sequence organization of the CES *roX1* and *roX2* (200 bp) and the probes used for EMSAs. MRE motifs: yellow; GAF binding sites in forward orientation: purple;

GAF binding sites in reverse orientation: green. The labeling of the GA-rich elements represents the exact sequences that are present within each element. **(C)** EMSAs of full length and truncated *roX1* and *roX2* probes. Probes: *Fab-7* GAGA3+4, *roX1*, *roX2* and *roX1/roX2* subfragments incubated with (NE) or without (-) late embryo nuclear extracts. Note: The input of labeled *roX1C* is not equivalent to that of the other *roX1* probes. For this reason, the yield of the LBC *roX1C* shift appears reduced relative to other probes.

We used several different experimental approaches to determine whether the LBC binds to *roX1* and *roX2*. First, if the complex binding to the *roX1* and *roX2* CES is the same as the complex that binds to *Fab-7* GAGA3+4, then binding should be sensitive to cross-competition. Figure S4.4A shows that LBC binding to the 200 bp *roX1* probe is competed not only by itself, but also by excess cold *roX2* and *Fab-7* GAGA3+4 DNA. Similar results are observed when the labeled probes are *roX2* or *Fab-7* GAGA3+4 (Figure S4.4B and S4.4C); LBC binding to these two probes is competed by excess cold *roX1*, *roX2* and *Fab-7* GAGA3+4. Second, if the factor binding to the *roX1* and *roX2* probes in nuclear extracts corresponds to the LBC, then it should contain the four known protein components: GAF, Mod(mdg4), CLAMP and E(y)2. Figure 4.2C and D show that antibodies against these four LBC proteins either supershift the *roX1* and *roX2* LBC shifts or inhibit LBC binding to DNA (Figure 4.2C, D). In each case, the effects are similar to that observed with the *Fab-7* probe GAGA3+4 (Figure 4.2A, B). Third, we incubated the 200 bp *roX1* probes with fractions from the Superose 6 gel filtration column. Peak fractions for the *roX1* LBC shift (33-41: Figure 4.3C) correspond to the peak fractions for the *Fab-7* GAGA3+4 LBC shift (Figure 4.3A). Finally, as was observed for *Fab-7* GAGA3+4, supershift experiments using gel filtration fractions containing LBC activity show that CLAMP, GAF and Mod(mdg4) are core components of the complex that generates both the *roX1* (Figure 4.3D) and the *roX2* LBC shift (Figure S4.3B). Taken together, these findings demonstrate that the LBC binds to two key CES, *roX1* and *roX2*.

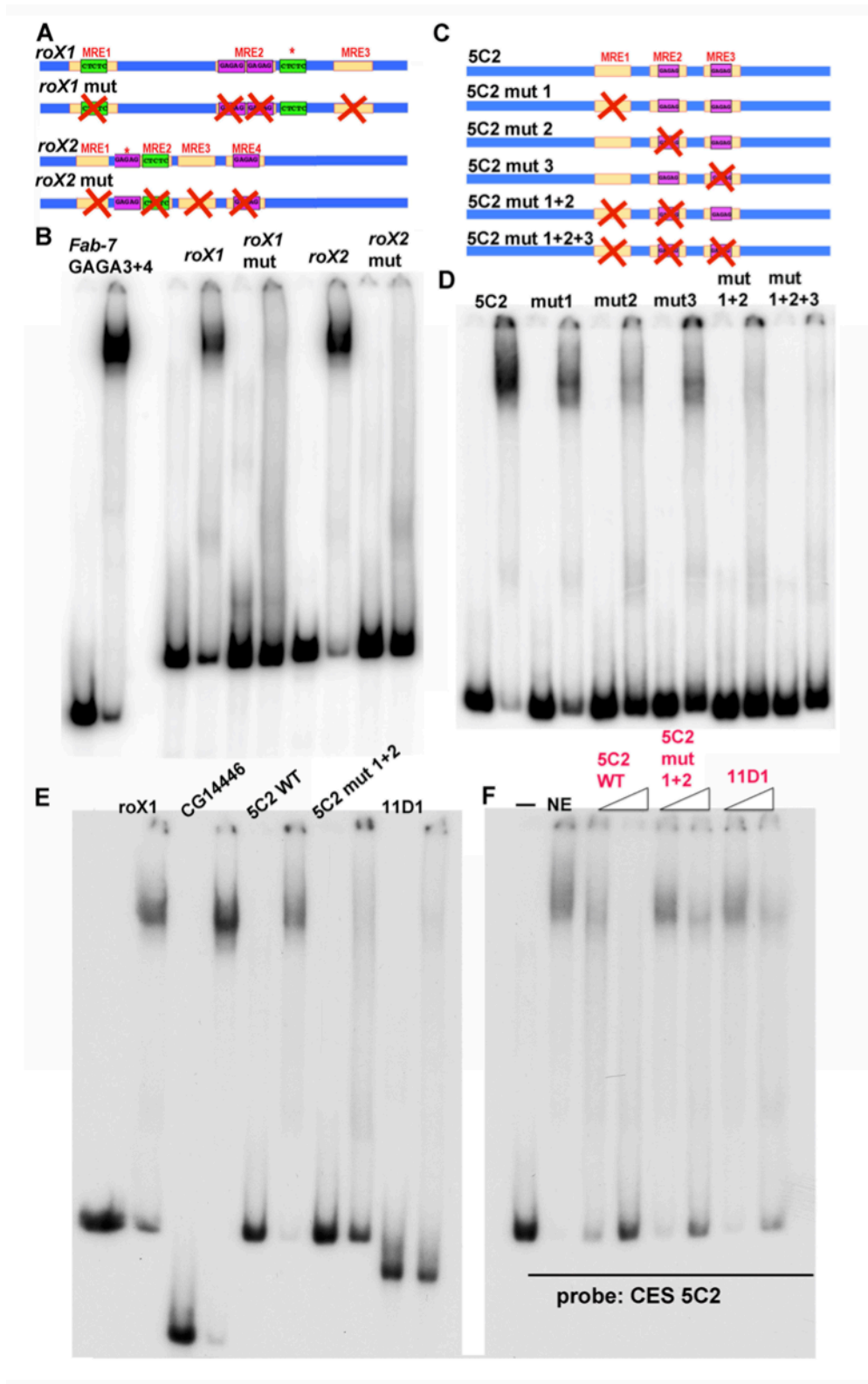


Figure 4.5 LBC binding is MRE dependent.

A schematic drawing of the wild type (top) and mutated versions of the 200 bp *roX1/roX2* (A) sequences used for EMSAs. (B) Mutations in the *roX1* and *roX2* MREs abrogate

LBC binding. (C) Schematic drawing of the wild type and mutant versions of the 5C2 probes. (D) Single and combination mutations in three 5C2 MREs differentially affect LBC binding. (E) Comparison of LBC binding to CES 11D1 with other LBC substrates as indicated. (F) Excess unlabeled WT 5C2, 5C2 mut 1+2 and CES 11D1 DNAs were used to compete for LBC binding to labeled CES 5C2. Fold excess of cold competitor added left to right: CES 5C2: 12.5x and 100x; CES 5C2 mut1+2: 25x and 200x; CES 11D1: 25x and 200x.

LBC binding to CES is MRE dependent

Previous studies have shown that the MRE sequences in the *roX1* and *roX2* CES are required to recruit MSL complex to ectopic autosomal sites. If the LBC plays an important role in the functioning of these two CES, then mutations in the MREs that disrupt MSL recruitment *in vivo* should reduce or eliminate the LBC shift in nuclear extracts. Figure 4.5B shows that this prediction is correct: MRE mutations, which abrogate CES activity *in vivo*, eliminate LBC binding *in vitro*. To confirm the effects of the MRE mutations, we used wild type or mutant *roX1* and *roX2* CES DNAs as competitors for LBC binding to the *Fab-7* GAGA3+4 probe. Figure S4.4D shows that unlabeled wild type *roX1* and *roX2* compete for LBC binding to GAGA3+4, whereas the corresponding mutant DNAs fail to compete effectively.

Other than *roX1* and *roX2*, the most thoroughly characterized multi-MRE CES is CES 5C2 (A. Alekseyenko et al. 2008). As indicated in Figure 4.5C, CES 5C2 contains three 21 bp MREs. Two of these MREs (MRE2 and 3) have a GAGAG motif, while the third, MRE1, has a GA-rich sequence, but no GAGAG motif. Like *roX1* and *roX2*, a shift that co-migrates with the *Fab-7* GAGA3+4 LBC shift is observed with a probe spanning the three MREs in CES 5C2 (Figure 4.5D). We used competition and supershift experiments to demonstrate that the CES 5C2 shift is generated by the LBC. Figure S4.5A shows that excess unlabeled 5C2 DNA competes for LBC binding to labeled GAGA3+4 and *roX1* probes. Figure S4.6A shows that as is observed for *roX1*, *roX2* and *Fab-7*

GAGA3+4, the 5C2 shift is generated by a complex that contains CLAMP, GAF and Mod(mdg4).

Previous studies have shown that mutations in one or a combination of the 5C2 MREs abrogate CES function *in vivo* to different extents (A. Alekseyenko et al. 2008). Two different assays were used to measure CES function. One was a luciferase assay in S2 tissue culture cells while the other was a ChIP assay of MSL2 binding to the 5C2 sequence inserted into an autosomal locus in fly larvae. Mutations in a single 5C2 MRE weakened CES function in one or both of these assays, but do not completely eliminate activity. In contrast, the combination mutants (mut1+2 and mut1+2+3) essentially eliminate both the stimulatory activity of CES 5C2 on transcription and Msl2 recruitment.

The effects of the different MRE mutations on the binding of the LBC to the CES 5C2 probe in nuclear extracts parallel those observed in the two *in vivo* assays (Figure 4.5D). LBC binding is reduced in the three single mutations, 5C2 mut1, 5C2 mut2 and 5C2 mut3, while binding is largely or completely lost in the double (5C2 mut1+2) or triple (5C2 mut1+2+3) mutants, respectively. Consistent with the elimination of LBC binding, excess unlabelled 5C mut1+2+3 DNA fails to compete for LBC binding to either the *roX1* or *Fab-7* GAGA3+4 probes (Figure S4.5A). It is interesting to note that though MRE1 does not contain a precise GAGAG motif, mutation of the GA-rich core weakens LBC binding just like mutations in MRE 2 and 3.

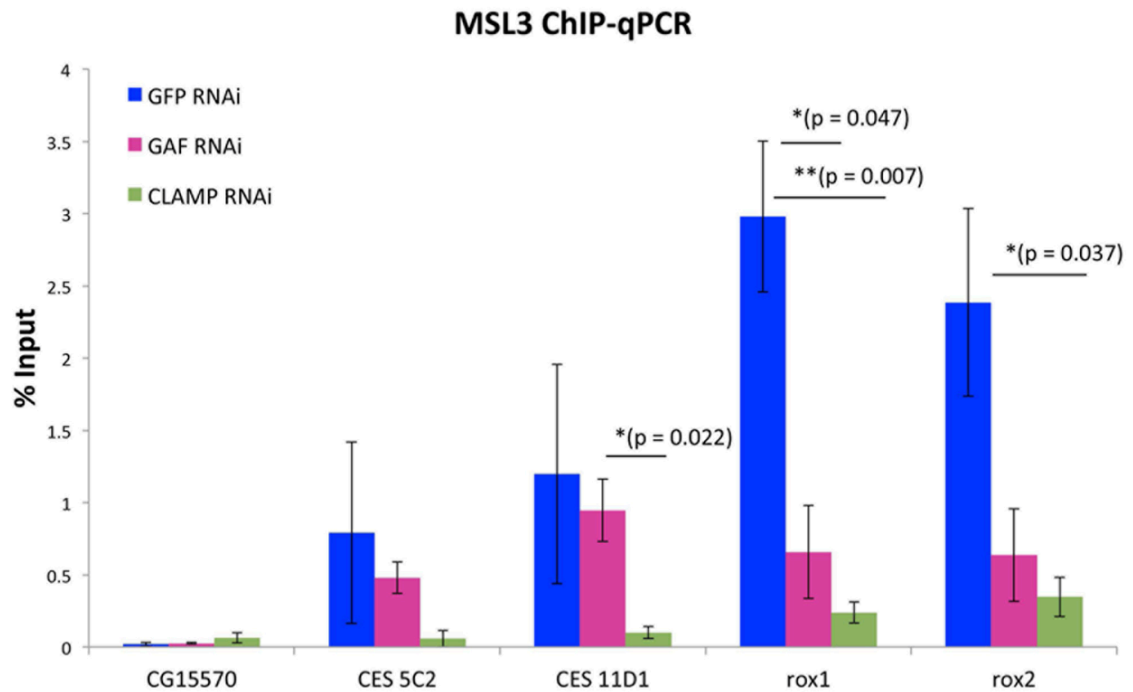


Figure 4.6 GAF and CLAMP contribute to MSL complex recruitment.

ChIP-qPCR was conducted using an antibody specific for the MSL3 protein after RNAi treatment targeting GAF and CLAMP (or GFP as a control). CG15570 was used as a negative control target. The CES 5C2, 11D1, *roX1*, and *roX2* were tested for a change in MSL3 binding. P-values from t-tests are displayed for significance *p<0.05 or **p<0.01 differences between sample means.

Knockdowns of CLAMP and GAF reduce MSL3 binding to the *roX1*, *roX2* and 5C2 CES

The experiments in the previous sections indicate that mutations in *roX1*, *roX2* and 5C2, which weaken or disrupt CES function *in vivo* compromise or eliminate LBC binding *in vitro*. To further test the connection between LBC and the recruitment of the dosage compensation machinery to these three CES, we used RNAi to reduce levels of CLAMP and GAF in male S2 tissue culture cells and measured MSL3 occupancy *in vivo*. As previously documented (Sorucu et al. 2013), we found that MSL3 association with *roX1*, *roX2* and CES 5C2 is CLAMP dependent (Figure 4.6). Unlike the *roX* loci, it is important

to note that the changes in occupancy at CES 5C2 are not statistically significant due to high variability within the control sample specifically at this locus. This is consistent with lower occupancy of CLAMP and GAF at CES 5C2 compared to the other two CES. As would be expected if MSL recruitment depends upon the LBC, MSL association with the *roX1* and *roX2* CES is also reduced when GAF levels are reduced.

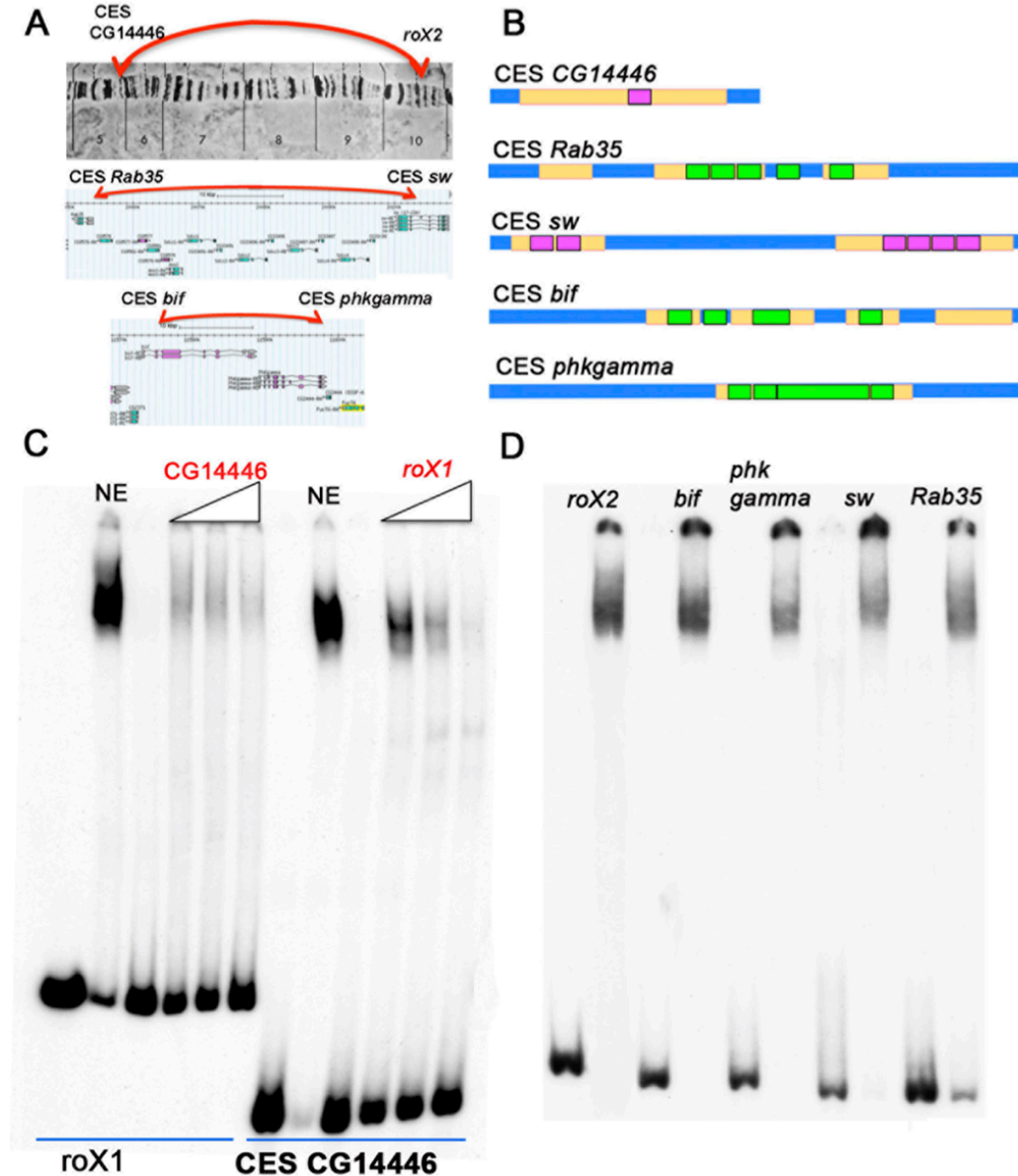


Figure 4.7 The LBC binds to topologically linked CES.

(A) Schematic of long-range interactions between three pairs of CES in a polytene X chromosome (top: CG14446 $\leftrightarrow$ roX2) and in GBrowser (bottom: *rab35* $\leftrightarrow$ *sw*; *bif* $\leftrightarrow$ *phkgamma*). (B) A schematic of the 5 CES probes that were used for EMSAs.

MRE motifs, yellow; GAF motifs in forward orientation, purple; GAF motifs in reverse orientation, green. **(C)** Cross-competition experiment with increasing amounts of cold CES CG14446 or *roX1* probes showing that the same complex binds to the different CES. Cold competitor was added in 5x, 10x and 50x excess. **(D)** LBC binds to physically interacting CESs.

CES 11D1: an example of a CES with only a single MRE

The three CES we tested for LBC binding thus far (*roX1*, *roX2* and 5C2) fall into the subclass that has multiple MREs; however, most of the CES identified by ChIP experiments have only a single MRE sequence (A. Alekseyenko et al. 2008). Since the 5C2 mutant, 5C2 mut1+2, gave only a barely detectable LBC shift, we wondered whether CES with a single MRE are bound by the LBC. The only CES with a single MRE that has been studied in detail is CES 11D1. It differs from 5C2, and also *roX1* and *roX2*, in that a small 150 bp fragment spanning the 11D1 MRE has only minimal CES function. Instead, a larger 500 bp element is needed for full CES activity, and in this context the MRE is required (A. Alekseyenko et al. 2008). Based on these *in vivo* results and the EMSA experiments described above, a plausible expectation is that the 150 bp 11D1 fragment would be a poor substrate for the LBC. This is the case. In the experiment in Figure 4.5E we compared the shift generated by the 11D1 probe with that observed for three multi-MRE probes. Two are the 200 bp *roX1* and 5C2 sequences discussed above, while the third is a CES at CG14446 (see further below). The CES CG14446 is only 100 bp and was included in this experiment to control for the possible effects of probe length. By comparison to these three positive controls, the 11D1 probe generates only a very weakly labeled LBC shift. In fact, the yield of the LBC shifted DNA is not much different from that observed for the MRE 5C2 mutant, 5C2 mut1+2, which retains only a single MRE. To confirm these findings, we used excess cold 5C2, 5C2 mut1+2 and 11D1 DNA to compete for LBC binding to labeled 5C2 (Figure 4.5F). Although a 200-fold excess of unlabeled 11D1 DNA reduces LBC binding to 5C2, a significant fraction (~50%) of the 5C2 probe is still shifted by the LBC. A similar result is observed

when the 5C2 mut1+2 DNA is used as the cold competitor. By comparison to these two single MRE DNAs, a 12.5-fold excess of the wild type 5C2 DNA is sufficient to give an equivalent if not a slightly greater degree of competition. Therefore, CES with a single MRE like 11D1 and 5C2 mut1+2 bind to the LBC much more poorly than CES that have multiple MRE sequences.

Given the low affinity of the LBC for CES 11D1, we wondered whether CLAMP and GAF are important for recruiting MSL complexes to this sequence *in vivo*. Figure 4.6 shows that CES 11D1 differs from the multi-MRE CES in that knocking down CLAMP compromises MSL accumulation, while knocking down GAF does not. A plausible interpretation of this finding is that the LBC is not critical for the CES function of CES11D1.

The LBC binds to topologically linked CES

The CCC experiments performed by (Ramírez et al. 2015) show that paired CES often define the endpoints of chromatin looped domains on the X-chromosome of both males and females and thus are likely to have architectural functions equivalent those of classical boundary elements or insulators. Although the basis for pairing specificity is not understood, pairing in transgene assays has been shown to be mediated by protein-protein self-interactions. For example, boundary bypass experiments indicate that boundary elements consisting of multimerized dCTCF sites pair with multimerized dCTCF sites, but not don't pair with multimerized sites for Su(Hw) (Kyrchanova et al. 2008). Thus, a plausible expectation is that the LBC will bind to both partners in at least a subset of the paired CES elements on the X-chromosome.

To test this hypothesis, we selected three of the interacting CES pairs Ramirez et al. identified. These CES pairs were chosen because ChIP experiments indicate that they are bound by both CLAMP and GAF *in vivo* and because they have multiple GA-rich MRE sequences. The first interacting pair is *roX2* and the CES associated with CG14446. These two CES are quite distant from each other, approximately 5 Mb apart. The 100-bp CG14446 CES has a large 72 bp region that contains multiple overlapping MREs, but only one GAGAG motif (see Figure 4.7). The second CES pair corresponds to the CES associated with *rab35a* and *sw*. These two CES are much closer, approximately 50 kb apart. The 200 bp CES *rab35a* has three MREs that have a total of 6 GAGAG motifs (CTCTC). The CES *sw* has two long GA-rich MREs (36 bp and 68 bp) and a total of six GAGAG motifs. The third pair are the CES associated with *bif* and *phkgamma*. They are separated from each other by about 15 kb. CES *bif* has four MREs, of which three have a GAGAG motif (CTCTC). CES *phkgamma* has a single long MRE-containing locus containing five GAGAG motifs (CTCTC) (see Figure 4.7).

Figure 4.7 shows that these physically-linked CES generate an LBC-like shift when incubated with 6-22 hr nuclear extracts. We used several experimental approaches to confirm that the observed shifts correspond to the LBC. In the case of the CES that pairs with *roX2*, CG14446, antibody supershift experiments indicate that as expected the shift is generated by a complex that contains the same proteins as the LBC: GAF, Mod(mdg4), E(y)2 and CLAMP (Figure S4.6B). Precisely the same result was obtained for the *rab35a* CES, which physically pairs with *sw* CES (Figure 4.7, Figure S4.6C). The identification of the LBC is further confirmed by assays of fractions from Superose 6 gel filtration column. Figure S4.7 shows that the elution profile of the LBC shift for the CG14446 CES probe is the same as that observed for *Fab-7* and *roX1* probes. We also

used cross-competition experiments to demonstrate that the LBC complex binds to the paired CES probes (Figure S4.6B).

Discussion

Boundary elements (or insulators) function to subdivide chromosomes in multicellular eukaryotes into a series of topologically independent looped domains. The importance of these architectural elements in the regulation of gene expression was first established in experiments dating back nearly thirty years (Kellum and Schedl 1992; Kellum and Schedl 1991). These experiments showed that boundaries have blocking activity that can restrict the action of enhancers and silencers to the loops in which they reside. Subsequent work revealed yet another function of these architectural elements, namely promoting long distance regulatory interactions (H. N. Cai and Shen 2001; Muravyova et al. 2001). One context in which long distance regulation is especially important is X-chromosome dosage compensation. In male flies, MSL complexes recruited to the X-chromosome by CES spread to nearby genes and upregulate their transcription. Recent experiments have drawn a connection between the recruitment and spreading of the MSL complexes from CES and an architectural function in which the CES define boundaries of looped domains (TADs) along the X-chromosome in both sexes (Ramírez et al. 2015). However, it is not clear from these experiments whether CES share other properties with classical fly boundary elements. Additionally, the factors responsible for the architectural activities of the CES were not identified. Here we show that a canonical Bithorax complex boundary factor, the LBC, is deployed by a subset of CES for their regulatory and architectural functions.

LBC and boundaries in BX-C

A combination of genetic and biochemical experiments have implicated the LBC in the functioning of three BX-C boundaries: *Fab-7*, *Fab-8*, and *AB-I* (S. Schweinsberg et al. 2004; Wolle et al. 2015; Kyrchanova et al. 2008; Cleard et al. 2017). The LBC has several unusual properties: 1) the relatively long ~65 bp DNA minimal binding length; 2) the diversity of the sequences recognition element; and 3) the size and composition of the LBC.

The gel filtration experiments reported here suggest that the LBC is significantly larger than our previous estimate of 670 kDa. These experiments also indicate that the complex detected after gel filtration differs from that observed in nuclear extracts in that it generates a more rapidly migrating shift. The change in mobility does not seem to be probe specific and is observed with all three of the DNAs we tested (*Fab-7* GAGA3+4, *roX1* and CES CG14446). This finding would suggest that some proteins dissociate from the LBC during gel filtration, or, alternatively, only interact with the LBC after it binds to DNA. Included in the group of proteins that remain included in the LBC during gel filtration are the two DNA binding proteins that we detected previously in nuclear extracts, GAF and Mod(mdg4). In the studies reported here we have identified a third LBC associated DNA binding protein, CLAMP. Like these two other proteins, CLAMP co-fractionates with LBC DNA binding activity, and thus appears to be an integral component of the LBC. Strengthening the connection between CLAMP and the LBC, ChIP experiments show that CLAMP interacts with three BX-C boundaries *in vivo* (*Fab-7*, *Fab-8* and *AB-I*) that are recognized by the LBC in nuclear extracts. Additionally, CLAMP and GAF binding to *Fab-7* in tissue culture cells is interdependent suggesting that they are part of the same protein complex that is destabilized in the absence of either component.

LBC and CES

A key question posed by the discovery that CES are architectural elements that link distant X-chromosome sequences to each other (Ramírez et al. 2015) is their relationship to chromatin boundaries. Our finding that CLAMP is an integral component of the BX-C boundary factor LBC provides a plausible link between CES and boundary elements. Previous experiments have shown that CLAMP binds to CES and that this association is critical for the recruitment of MSL complexes to the elements (Soruco et al. 2013). However, these experiments did not illuminate how CLAMP interacts with CES. Here, we show that in nuclear extracts CLAMP binds to nine multi-MRE CES, not independently, but rather as a component of the LBC chromatin boundary factor.

A number of lines of evidence argue that LBC binding to *roX1*, *roX2*, and 5C2 *in vitro* is relevant to their ability to recruit MSL complexes *in vivo*. The first is the requirement for both GAF and CLAMP in recruitment of MSL complex to several key CES. With the important caveat that there could be indirect effects caused by gradually depleting GAF and CLAMP, knockdowns of both proteins diminish MSL recruitment to these CES *in vivo*. The second line of evidence is the close match between the sequence recognition properties of the LBC and the requirements for CES function *in vivo*. A feature common to most CES is the presence of one or more GA-rich MREs in a DNA segment of between ~100-250 bp. The MRE sequences have been shown to be required for the CES functions of *roX1*, *roX2*, CES 5C2 and CES 11D1. Of these four CES, the relationship between CES function *in vivo* and MREs has been analyzed in most detail for CES 5C2. Alekseyenko et al. (2008) showed that mutations in individual MREs within a 150 bp CES 5C2 element reduced but did not eliminate CES activity, while CES activity was disrupted by mutations in two or all of the three MREs. Consistent with the *in vivo* effects on CES activity, we found that mutations in individual 5C2 MREs weaken

LBC binding, whereas binding is disrupted when all three of the MREs are mutant. Importantly, the failure of LBC to bind to CES 5C2 mut1+2+3 in nuclear extracts also recapitulates that the effects of the triple mutant on CLAMP occupancy *in vivo* (Soruco et al. 2013).

In the case of *roX1* and *roX2*, we found that mutations in all of their MREs eliminate the LBC shift in nuclear extracts. Though we did not test the effects of mutations in individual MREs, experiments with smaller 100-bp sub-fragments indicate that LBC binding to *roX1* and *roX2* probes requires the presence of at least two MREs. This conclusion is reinforced by our experiments with CES 11D1. This CES differs from *roX1*, *roX2* and CES 5C2 in that it has only a single MRE and no GAGAG motifs. Unlike other CES with multiple MREs, the 11D1 element is a poor substrate for the LBC.

The observation that the LBC binds preferentially to CES that have multiple MREs, which span a relatively large 100-200 bp sequence, is also consistent with other *in vivo* experiments on these elements. For example, (Gilfillan et al. 2007) identified a ~500 bp CES element DBF-DHS. Within this element, they found that a 40 bp GA rich sequence DBF 12-L15 is essential for recruiting the MSL complex in flies. Though required for the functioning of the full length CES, this 40 bp sequence is not sufficient on its own. On the other hand, when multimerized, the 40 bp element is able to recruit MSL complexes. In the case of *roX2*, (Park et al. 2002) found that a 134 bp fragment spanning the four *roX2* MREs 1-4 functions as a CES at ectopic sites on autosomes. As can be seen in Fig. 4C, the 100 bp *roX2A* probe which contains the first three *roX2* MREs binds the LBC almost as well as the larger 200 bp *roX2* probe.

While we have not tested the architectural functions of CES in the studies presented here, previous work on *Fab-7* and *Fab-8* (Kyrchanova et al. 2015; S. Schweinsberg et al. 2004; Wolle et al. 2015) demonstrates that the LBC sequences in both of these boundaries have classical insulator activity and can mediate long distance pairing interactions. Consequently, a reasonable expectation is that the LBC will have similar architectural functions when associated with other sequences in the genome, including CES. Consistent with this notion, we have shown that three of the pairs of interacting CES identified as “TAD boundaries” in tissue culture cells by Ramirez et al 2015 are recognized by the LBC in nuclear extracts. These interacting pairs are CES *roX2* $\leftrightarrow$ CES *CG14446*, CES *rab35* $\leftrightarrow$ CES *sw*, and CES *bif1* $\leftrightarrow$ CES *phkgamma*.

Boundaries versus CES

Although our findings clearly implicate the LBC in CES function, many questions remain:

Is LBC association a common feature of CES? We tested only a small subset of sequences that have been classified as CES. With the exception of CES 11D1 all of them have multiple, closely clustered MREs and the LBC binds to these elements with an affinity at least equivalent to that observed for *Fab-7* probes. For three of the multi-MRE CES (*roX1*, *roX2* and CES 5C2) there is *in vivo* evidence that they function as CES, and can recruit MSL complexes to ectopic sites. The five remaining sequences were classified as CES on the basis of ChIP experiments and the presence of MREs. Consequently, we do not know whether they can function as CES *in vivo*. However, if they do, it seems likely that both MSL recruitment and LBC binding will depend on the clustered MREs as is the case for *roX1*, *roX2*, and CES 5C2.

While other multi-MRE CES should also have similar properties to those studied here, nearly two thirds of the CES that have been identified in ChIP experiments have a single

MRE (A. Alekseyenko et al. 2008). The only member of this group that we tested is CES 11D1 and its interaction with the LBC clearly differs from CES containing multiple MREs. Whereas all of the multi-MRE CES we tested bind strongly to the LBC, the LBC binds only weakly to a 150 bp probe containing the single CES 11D1 MRE. The limited affinity for the LBC could explain why CES 11D1 requires a 500 bp fragment for CES activity, while smaller 100-200 bp fragments spanning the clustered MREs are sufficient for the CES activity of *roX1*, *roX2* and CES 5C2. One question is whether other single MRE CES have properties similar to CES 11D1? Since probes derived from the *Fab-7* dHS1 sequence are bound with relatively high affinity by the LBC yet do not have any MRE consensus motifs, it is possible that some of the single MRE CES will have other motifs that promote LBC association.

On the other hand, there are reasons to think that the distinct properties of CES 11D1 may be typical of many single MRE CES. In ChIP experiments, Soruco et al. (2013) found that CLAMP association with single-MRE CES is dependent on synergistic interactions with MSL. By contrast, CLAMP associates with multi-MRE sequences independent of the dosage compensation machinery. Furthermore, nucleosome occupancy of single MRE CES differs between the two sexes; it is typically lower in male cells and higher in female (Soruco et al. 2013). In contrast, occupancy of multi-MRE CES is similar in both sexes suggesting that these are constitutive binding sites. These differences, taken together with the limited affinity for the LBC and the persistence of MSL association after GAF RNAi treatment support a model in which single and multi-MRE CES play distinct roles in dosage compensation (Soruco et al. 2013). In this model, the LBC would bind to multi-MRE CES and they would provide a platform for recruiting MSL complexes. Chromatin modifications induced by MSL association with multi-MRE CES along the X-chromosome would then open up the single MRE CES, potentially

permitting the binding of CLAMP and presumably yet other factors. The presence of these proteins would setup a positive feedback loop by facilitating the recruitment of MSL complexes.

What distinguishes CES from autosomal boundaries like Fab-7? Though our experiments implicate the LBC in the functioning of multi-MRE CES, how it promotes the recruitment of MSL complexes to these CES but not to *Fab-7* is unclear. One possibility is that the LBC variants that bind to CES are distinct from those that interact with canonical boundary elements. While this may seem farfetched, until we know more about the composition of the LBC it cannot be excluded. The reason for this is the presence of Mod(mdg4) in the LBC. There are 31 predicted Mod(mdg4) isoforms (Dorn and Krauss 2003). All share a common N-terminal Mod(mdg4) BTB domain which mediates the assembly of Mod(mdg4) octomers (Bonchuk et al. 2011), while they have different C-terminal domains. Of these 31 isoforms, 27 have unique FLYWCH DNA binding domains. If Mod(mdg4) is incorporated into the LBC as an octomer, there could be many different versions of the complex with potentially different DNA binding and protein-protein interaction properties.

However, a more likely alternative is that the LBC provides a platform for other factors that are actually responsible for recruiting/assembling MSL complexes. For example, (Villa et al. 2016) have shown that MSL2 has sequence/shape specific DNA binding activity, and that $MSL-2 \leftrightarrow DNA$ interactions are responsible for mediating the recruitment of the MSL dosage compensation machinery to a specific class of CES called PionX sites when they are present at high local concentrations. According to their model, MSL2 interacts specifically with an extension of the MRE sequence that is located 5' to the GA-rich core. Moreover, in contrast to the known requirements for CES

function *in vivo*, the interaction between MSL2 and PionX sequences *in vitro* appears to be largely independent of the GA-rich motifs. This model would potentially fit with our findings, which show that LBC binding to X-chromosome CES requires GA-rich sequences. In principle, $LBC \leftarrow \rightarrow GA_n\text{-MRE}$ interactions could provide a mechanism for presenting the 5' PionX sequences in a configuration suitable for MSL2 interaction with adjacent sequences. Clearly, further studies will be needed to assess whether this model is correct. Overall, we have provided new mechanistic insights into how sequences on the X-chromosome function to establish chromatin boundaries that are associated with an active domain of coordinate gene regulation.

Acknowledgements

The authors would like to thank Anton Golovnin, Pavel Georgiev, Elissa Lei, Victor Corces, Carl Wu, Mitzi Kuroda, John Lis, and David Gilmour for the gift of antibodies. This work was supported by grants from NIH (R01GM098461-1), American Cancer Society Research Scholar (123682-RSG-13-040-01-DMC), and a Pew Biomedical Scholars program grant awarded to E.N.L and to P.S. (GM043432). A.K. and P.S. would also like to acknowledge support from grants to the Gene Biology Institute by the Russian Federation Ministry of Education and Science (RF 14.B25.31.0022) and the Russian Scientific Foundation (16-14-10346). We have no conflicts of interest to disclose.

Materials and Methods

Nuclear extracts. Nuclear extracts from 6- to 22-h embryos were prepared as described previously (Aoki et al., 2008) with small modifications. 0- to 12-h embryos from Oregon R were collected from apple juice plates and aged 10 h at room temperature. The extraction was completed with the final concentration of KCl as 360 mM.

Probes for EMSA. Probes were obtained by PCR, purified on agarose–1XTris-acetate-EDTA (TAE) gel followed by phenol/chloroform extraction. Probe sequences, as well as primer sequences used for PCR are listed in Table S4.1.

EMSA. Electrophoretic mobility shift assays were performed using ³²P labeled DNA probes under conditions described previously (Wolle et al., 2015). See Supplemental Experimental Procedures for details.

Gel filtration. Fractionation of the nuclear extracts derived from 6- to 22-h embryos was performed by size exclusion chromatography using Superose 6 10/330 GL column (GE Healthcare). Molecular mass markers ranging from 1,350 to 670,000 Da (Bio-Rad) were used as gel filtration standards.

Western Blotting. Either 0.1 µl of late stage nuclear extract (about 2 µg of total protein) or 2 µl of the column fractionated material was run on 10% SDS-PAGE, transferred to a PVDF membrane, probed with anti-CLAMP B antibody (1:250, new antibody generated using Abcam custom services) and developed with Amersham ECL Prime Western Blotting Detection Reagent (GE Healthcare) following the manufacturer's protocol.

Generation of dsRNA for RNAi treatment. Templates for dsRNA to target the CLAMP gene (*CG1832*), the GAF gene (*Trl*), and the control GFP (*gfp*) sequence were generated by PCR using the following template sequences: CLAMP RNAi template, DRSC03718, was designed as described in detail at www.flyrnai.org (Larschan et al 2012). *gfp* dsRNA was used as a control, designed as previously described (Hamada 2005). The GAF (*Trl*) dsRNA target sequence used was also previously designed (Fuda et al. 2015). The CLAMP, GAF, and GFP constructs were synthesized using the T7 megascript kit (Ambion), including a DNase I treatment for 15 min at 37°C following dsRNA synthesis. dsRNA was then purified using the RNeasy kit (Qiagen). Sequences are reported as Supplemental Table S2.

RNAi treatment of SL2 cells. RNAi treatment was adapted from previous studies (Larschan et al. 2012; Soruco et al. 2013). Briefly, RNAi was set up in T150 flasks as follows: 135 µg dsRNA was diluted into 3 mL UltraPure water (Invitrogen) with 6 mL SL2

cells at a concentration of 8×10^5 cells/ml in Schneider's medium (Gibco) for 45 min at 25°C. Next, 21 mL Schneider's medium with 10% heat inactivated FBS was added and RNAi treated cells were grown for 6 days at 25°C until collection for chromatin preparation.

Chromatin preparation. This experiment is based on previous work (Larschan et al. 2007b), modified for T150 cell culture flasks, in which cross-linking and quenching were performed. Samples were then washed three times, pelleted, and resuspended in lysis buffer for sonication. See supplemental experimental procedures for specific protocol and buffer details.

Immunoprecipitation (IP). Either 5 μ L CLAMP antibody (SDIX)(Larschan et al. 2012), 10 μ L GAF antibody (gift of John Lis) or 4 μ L MSL3 serum (Gift of Mitzi Kuroda) was added to each 1 mL IP sample and rotated at 4°C overnight. At the same time, 50 μ L Protein G Dynabeads per sample were incubated in bead blocking buffer (10 mM PBS, 0.1% Triton X-100, and 3% BSA) with rotation at 4°C overnight. The lysate and beads were then incubated together for 2 hours at 4°C. Samples were then washed, eluted from beads, and a DNA clean-up was performed. See supplemental experimental details for wash steps, buffers, and DNA clean-up information.

Quantitative Real-Time PCR. All target primer sequences used were previously designed (A. Alekseyenko et al. 2008; Wolle et al. 2015). Briefly, the average relative quantitation (CT value) of two replicate reactions was determined for all samples. Platinum SYBR master mix w/ROX (Invitrogen) was used. Three independent chromatin preparations were conducted for each experiment and standard deviations are reported. Dilutions of input to 1% were conducted to assure that samples were in the linear range

for quantitative PCR. To assess the overall ChIP enrichment, sample CTs were normalized with respect to the 1% input values, and then the averages and standard deviations of the biological replicates were reported for each primer target.

Supporting Information

Supplemental Methods

EMSA. Purified DNA probes (1 picomole) were 5' end labeled with [γ - ^{32}P]ATP (MP Biomedicals/ Perkin Elmer) using T4 polynucleotide kinase (New England Bio-Labs) in a 50 μl total reaction volume at 37°C for 1 h. Samples were run through columns packed with Sephadex G-50 fine gel (Amersham Biosciences) to separate free ATP from the labeled probes. The volume of the sample eluted from the column was adjusted to 100 μl using deionized water so that the final concentration of the probe was 10 fmol/ μl .

Binding reactions were performed in a 20 μl volume using the conditions described previously (Wolle et al., 2015) except for the concentration of the non-specific competitor poly(dA-dT):poly(dA-dT) in the binding reaction. The final concentration of poly(dA-dT):poly(dA-dT) was varied between 0.1 and 0.25 mg/ml depending on the DNA probe used. 1 μl of nuclear extract (corresponding to about 20 g of protein) or an equal volume of 360 mM nuclear extraction buffer (for negative control) was used. For the experiments with fractionated nuclear extracts, 10 μl of the indicated gel filtration fraction was used instead of the nuclear extract. In such cases, the addition of KCl into the reaction mixture was omitted. In some reactions, unlabeled competitor DNA was included so that the final concentration of the competitor would be in 5- to 100-fold excess. The reaction mixtures containing the ^{32}P - labeled DNA probes were incubated for 30 min at room temperature.

For supershift experiments, pre-immune rat or rabbit serum or antibodies against different proteins were pre-incubated in the reaction mixtures described above with the nuclear extract or gel column fractions for 30 min at room temperature to allow the protein-antibody association, followed by an incubation with ^{32}P - labeled DNA probes

for 30 min at room temperature. Either 1 μ l of rabbit polyclonal anti-CLAMP antibody A (SDIX, Larschan et al 2012), 4 μ l of rabbit polyclonal anti-CLAMP antibody B (Abcam), 1 μ l of rabbit polyclonal antibodies against GAF and E(y)2 or 1 μ l of rat polyclonal against Mod(mdg4) (generous gift from Anton Golovnin and Pavel Georgiev, Elissa Lei [Mod(mdg4)], Carl Wu (GAF), and David Gilmour (GAF)) was used.

Binding reactions were electrophoresed using the conditions described previously (Wolle et al., 2015). The gels were run at 180 V for 3 to 4 h at 4°C, dried, and imaged using a Typhoon 9410 scanner and Image Gauge software or X-ray film.

Chromatin Preparation. Cells were crosslinked with 1% formaldehyde in T150 flasks. Samples were mixed (110 rpm) for 10 min at room temperature (~23°C) followed by quenching with 0.960 mLs 2.5 M glycine for 5 min at room temperature (110 rpm). The cells were then transferred to a 50 mL centrifuge tube and centrifuged for 5 min at 1500 g at 4°C. The pellet was resuspended in 30 mLs PBS containing 1x Protease Inhibitor cocktail (pancrease-extract 0.002 mg/ml, thermolysin 0.0005 mg/ml, chymotrypsin 0.002 mg/ml, trypsin 0.02 mg/ml, papain 0.33 mg/ml) (Roche) and 0.2 mM PMSF followed by centrifugation for 5 minutes at 1500 g at 4°C. Centrifugation and resuspension were repeated as described above with the following buffers: 1) 6 mLs Wash A (10 mM HEPES pH 7.6, 10 mM EDTA pH 8.0, 0.5 mM EGTA pH 8.0, 0.25% Triton-X 100, 1x PI, and 0.2 mM PMSF) and 2) 6 mLs Wash B (10 mM HEPES pH 7.6, 10 mM NaCl, 10 mM EDTA pH 8.0, 0.5 mM EGTA pH 8.0, 0.25% Triton-X 100, 1x PI, 0.2 mM PMSF). In between each wash, cells were rotated at 4°C for 5 min. After Wash B, cells were pelleted and resuspended in Lysis buffer (0.1% SDS, 10 mM EDTA pH 8.0, and 50 mM Tris-HCL pH 8.0) at a concentration of 1.0×10^8 cells/mL.

Chromatin samples were sonicated on ice using a Bioruptor water bath sonicator for six

5 min cycles with 30 sec on 30 sec off on the high setting. A one min incubation on ice in between 5 min cycles was used to prevent the samples from overheating. Chromatin was diluted with 9 volumes Dilution Buffer (0.01% SDS, 1.2 mM EDTA pH 8.0, 16.7 mM Tris-HCL pH 8.0, 1.1% Triton X-100, and 167 mM NaCl) and filtered on a PolyPrep column (BioRad). Next, 200 μ L was taken as 20% "Input" DNA sample and the remaining chromatin was aliquoted into 1 mL volumes for immunoprecipitation and flash frozen in liquid nitrogen and stored at -80°C until used.

For clean-up of input DNA samples, crosslinks were reversed overnight at 65°C. 1 μ L RNase A was added to the inputs and incubated at 37°C for 30 min. 3 μ L Proteinase K was added per tube and samples were incubated at 42°C for 2.5 h. DNA was extracted using a Phenol-Chloroform-Isoamyl Alcohol (25:24:1) extraction using phase-lock tubes. The DNA was then precipitated with 5M NaCl, glycogen and ethanol at -80 for 1 h and resuspended in 50 μ L UltraPure water. DNA was run on a 2% gel at 130 V for 1 h and post-stained with Gel Red for 20 min.

Immunoprecipitation (IP). Incubations of the chromatin with antibodies and magnetic beads were performed as described in the Experimental procedures. The following washes were performed with 1 mL buffer, with 3 min rotating at 4°C followed by collection on a magnetic bead rack: 2X 1 mL Wash Buffer I (20 mM Tris-HCL pH 8.0, 0.1% SDS, 1% Triton X-100, 2 mM EDTA pH 8.0, and 150 mM NaCl), 2X x 1 mL Wash Buffer II (20 mM Tris-HCL pH 8.0, 0.1% SDS, 1% Triton X-100, 2 mM EDTA pH 8.0, and 500 mM NaCl), 1 mL TE wash buffer (10 mM Tris-HCL pH 8.0, 1 mM EDTA pH 8.0), and then eluted in Elution buffer (50 mM Tris-HCL pH 8.0, 10 mM EDTA pH 8.0, and 1% SDS) for 10 min in a 65°C with vortexing every 2 min. Samples were centrifuged at max speed for 1 min, beads collected with a magnetic rack, and samples transferred to new

tubes for DNA clean-up as described above (chromatin preparation supplemental experimental procedures) with a final resuspension in 50 μ L UltraPure water.

Supplemental Figures

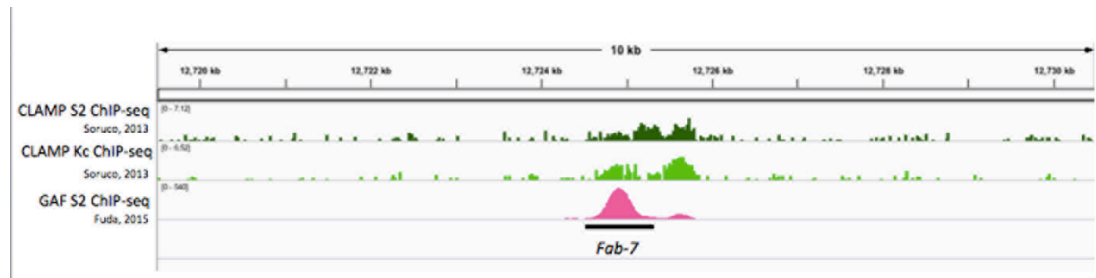


Figure S4.1 GAF and CLAMP bind to the *Fab-7* boundary *in vivo*.

Profiles for a 10 kb region spanning *Fab-7* are shown from genome-wide experiments on the localization of CLAMP (ChIP-seq: (Soruco et al. 2013)), GAF (Fuda et al. 2015).

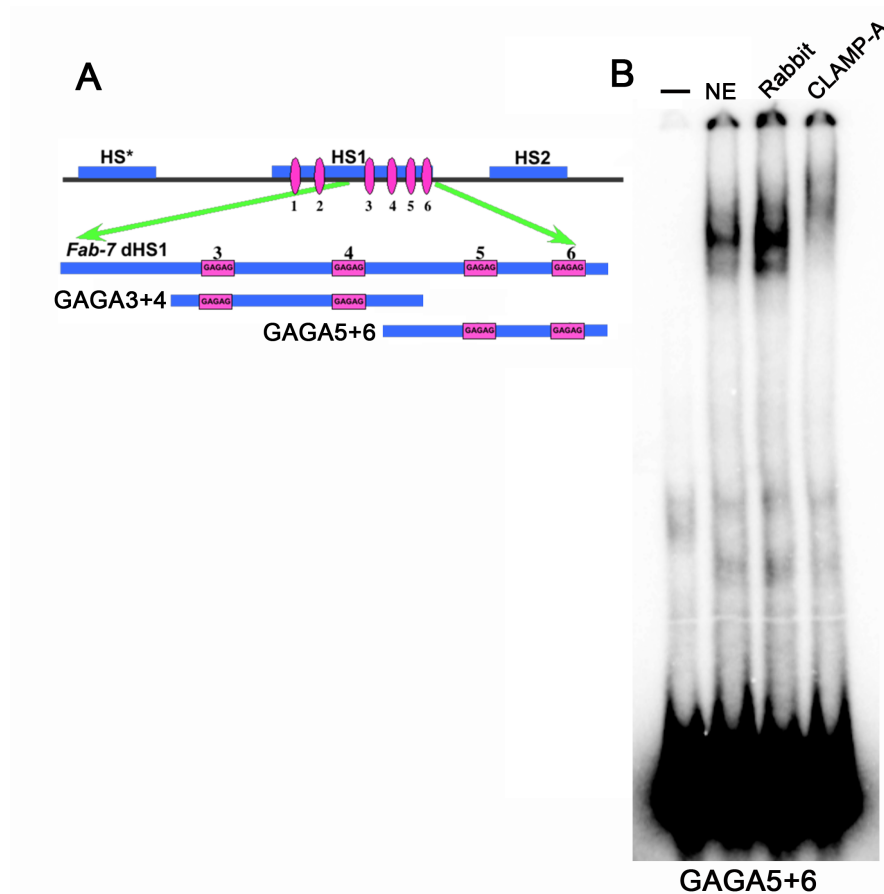


Figure S4.2 CLAMP is part of LBC binding to *Fab-7* GAGA5+6 sequence.

(A) A schematic drawing of the *Fab-7* boundary with six GAGAG sites (pink) and probes (GAGA3+4 and GAGA5+6) used for EMSAs. (B) Probe GAGA5+6 was incubated with late nuclear extracts, without and with either a control preimmune rabbit serum (Rabbit) (lane 3) or the rabbit polyclonal CLAMP A antibody (lane 4).

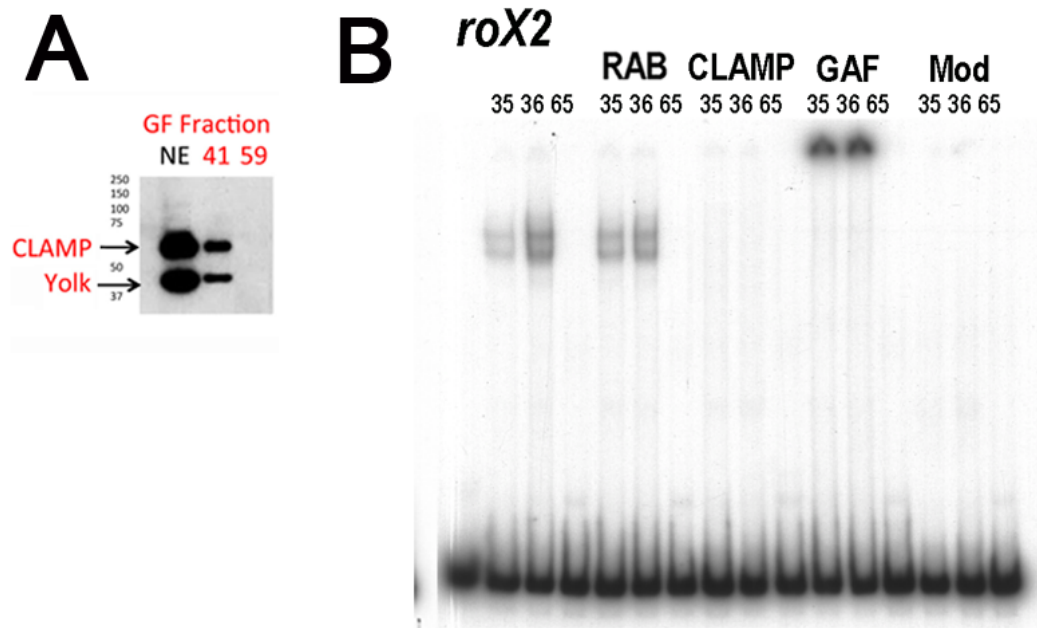


Figure S4.3 CLAMP, GAF and Mod(mdg4) co-fractionate with the LBC.

(A) Western blotting shows that CLAMP protein is present in a peak gel filtration fractions for LBC activity but not in a fraction that did not have LBC activity (as assayed by EMSAs). A non-specific band at 45kDa corresponds to yolk protein present in nuclear extract in excessive amounts. (B) CLAMP, GAF and Mod(mdg4) are integral components of the LBC. 10 ul of Superose 6 fractions 33, 36 and 65 were preincubated with no antibody, rabbit serum, CLAMP-B, GAF or Mod(mdg4) antibodies for 30 min. The labeled *roX2* probe was then added, and the incubation continued for 30 min prior to gel electrophoresis.

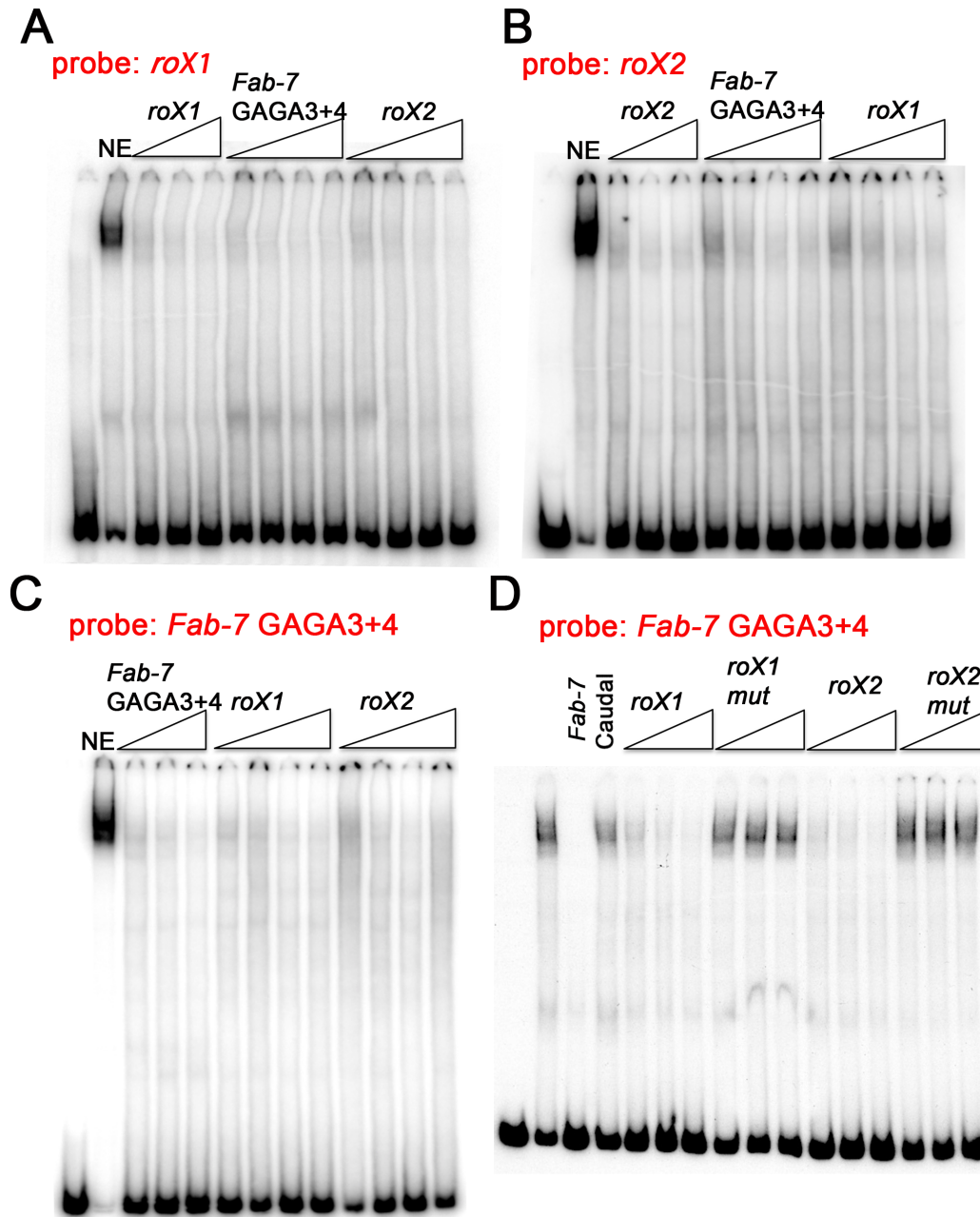


Figure S4.4 Competition experiments with wild type and mutant *roX1* and *roX2* DNAs.

Cross-competition experiments with labeled (A) *roX1*, (B) *roX2*, (C and D) *Fab-7* GAGA3+4 probes. Labeled probes (as indicated in red) were incubated with late nuclear extracts in the absence (NE) or presence of increasing amounts of unlabeled cold competitor as indicated above the lanes. For self-competition (first set of 3 in A, B and C) a 25x, 50x or 100x excess of cold competitor was added to the reaction mixtures. For cross competition (next two sets of four) 25x, 50x, 75x and 100x excess cold competitor as indicated was added to the reaction mix. (D) Mutated *roX1* and *roX2* probes fail to compete for LBC binding to the *Fab-7* GAGA3+4 probe. A 100 bp Caudal DNA competitor fails to compete for LBC binding and is used as a negative control. In each

set of three 25x, 50x and 100x excess of cold competitor as indicated was added to the reaction mix.

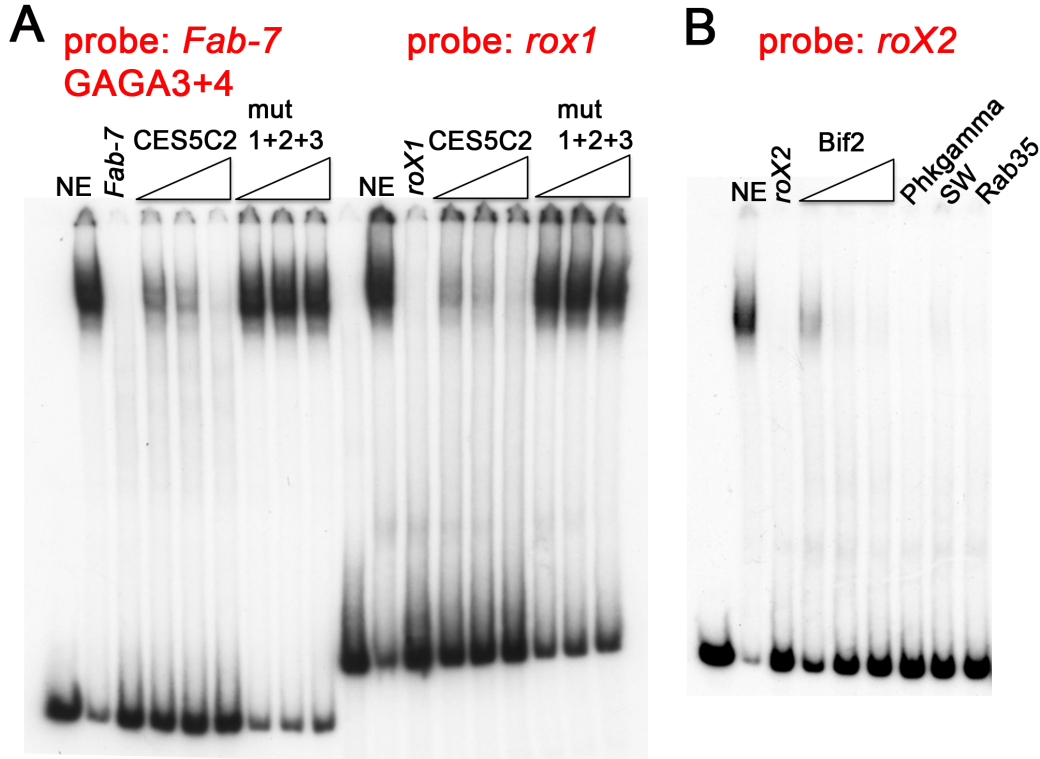


Figure S4.5 Cross competition of CES 5C2, CES 5C2 mut 1+2+3 and interacting CES.

(A) For the cross competition experiments the labeled probes were either *Fab-7* GAGA3+4 or CES 5C2 as indicated above in red. A 12.5x, 25x or 100x excess of the wild type or mutant CES 5C2 cold competitor was added. (B) Labeled *roX2* probe was competed with excess unlabeled DNA corresponding to the four interacting CES as indicated. For Bif2, a 12.5x, 25x, or 100x excess of the unlabeled DNA was added to the reaction mix. For the other unlabeled DNAs (*roX2*, *phkgamma*, *sw* and *Rab35* 100x excess was added.

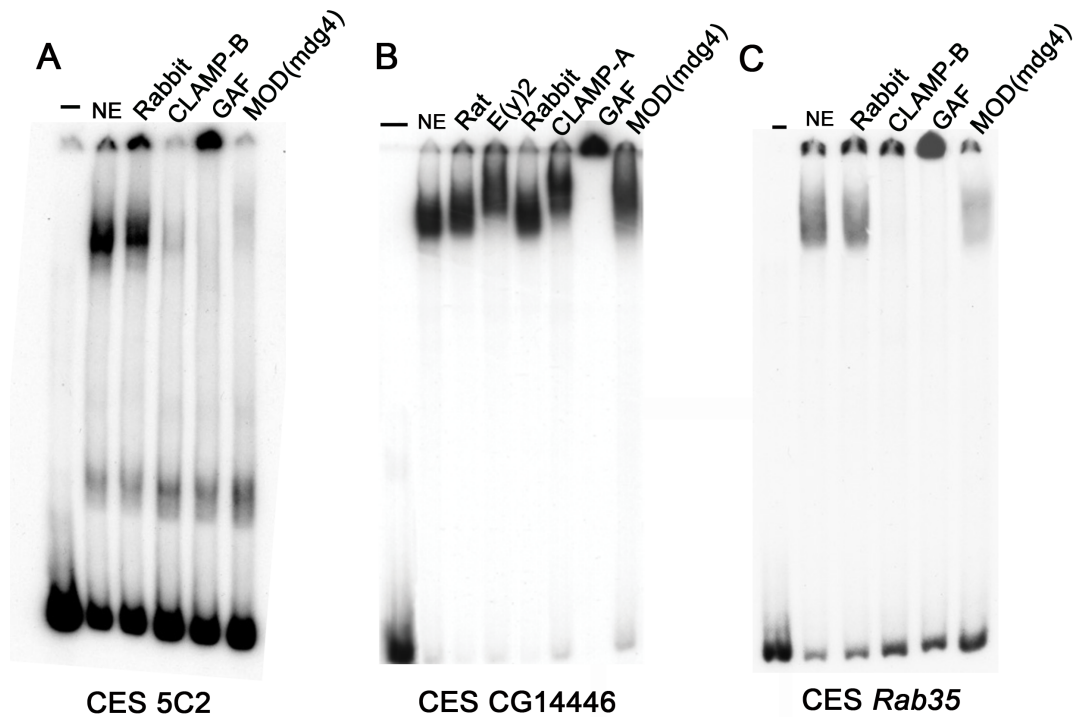


Figure S4.6 Supershift experiments demonstrating that the LBC binds to three additional CES.

(A) CES 5C2, (B) CG14446 and (C) Rab35. Probes were incubated with late nuclear extracts, without and with either a control preimmune rat or rabbit serum, a rat polyclonal E(y)2 antibody, two different rabbit polyclonal CLAMP antibodies (A and B), rabbit polyclonal GAF antibody or a rabbit polyclonal Mod(mdg4) antibody. A) CES 5C2, B) CES CG14446 and C) CES *Rab35*. For the experiments in panels A and C, the nuclear extract was preincubated with the indicated antibody for 30 min prior to the addition of the labeled probe, and then the reaction mix was incubated for an additional 30 min prior to gel electrophoresis. For panel B, the nuclear extract, antibody and labeled probe were added together and the reaction mix was then incubated for 30 min prior to gel electrophoresis.

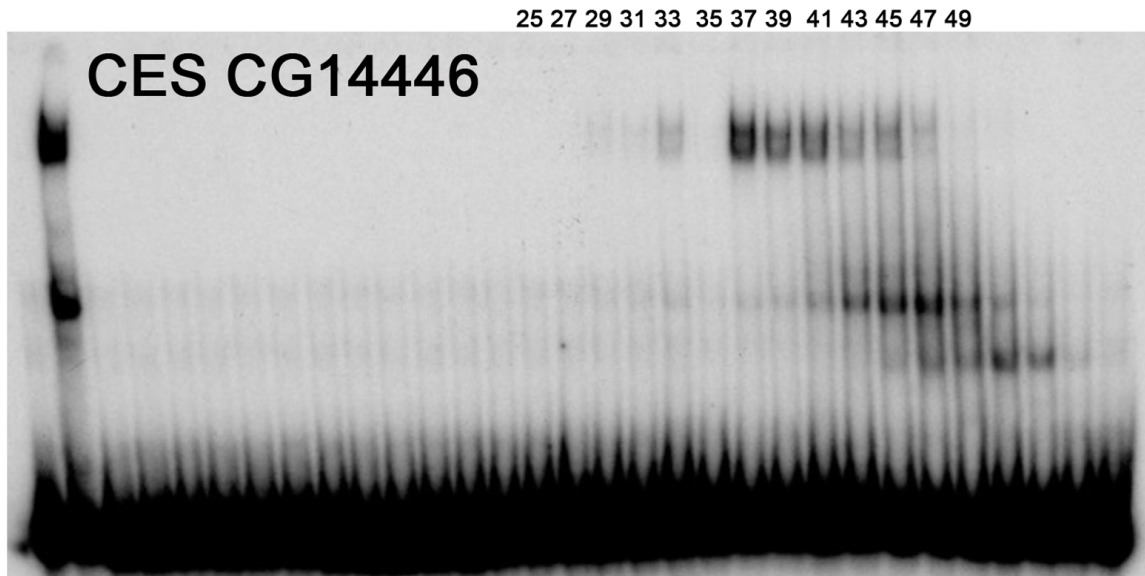


Figure S4.7 LBC shift of CES CG14446 by Superose 6 column fractions.

Nuclear extracts were size separated on a Superose 6 column and then the fractions used for EMSA of the CES CG14446 probe. 10 ul of odd numbered column fractions (1-57) were incubated with the CES CG14446 probe. Note that the peak fractions are the same as observed with the *Fab-7* and *roX1* probes. Note: as in experiments with GAGA3+4 and *roX1*, the activity of fraction 35 is reduced compared to its neighbors. The reduction is, however, more pronounced for CG14446, probably because the fractions had been stored for a longer period of time when they were used for this EMSA experiment.

Supplemental Tables

Table S4.1 DNA sequences used for the EMSA experiments.

Probe	Sequence	Primer sequences (5' to 3')	
Fab-7 GAGA3+ 4	CAAAGAGCGACACGTGAACAG GTGCAGTAGTAAATATAAGCAA AGAGAGTTGGAAAGAGTATTG GCTAAGAGCGTCCGCTCACTA ACACATAGATAAATTAAGAGAG ACGTGATAAGAGAACCGCACG CACACC	F	CAAAGAGCGACAC GTGAACAG
		R	GGTGTGCGTGCGG TTCTC
Fab-7 GAGA5+ 6	CGTGATAAGAGAACCGCACGC ACACCACCGCAAAATCCAATTG GAAGAGAGCGACTGCTTGAGT GTATTGGTTAGCAAGAGAGCG GCTAAGGTTTGATGGTTTGATT GGAATTCAGTTGCCGTTTCG	F	CGTGATAAGAGAA CCGCACGC
		R	CGAACGGCAACTG AATTCCAATC
roX1	TTATAGGGGTTATATGGCTTTA CCGCTCTCTTTCGGGACTTGCA GTCCGCCCTATGGCCGTCTCA CTCACCCGCTTGCGATGCAAA CGCCTGGCCAGCCCGATAGAG AGTGATGGAGAGAGGGAGCCG TCTCTCCTTGTCTACAGTATCA TTGTCTCGCTCGCATCCGCGTA CCGCCACATGCATCCGATAGA GGTATGC	F	TTATAGGGGTTATA TGGCTTTACCG
		R	GCATACCTCTATCG GATGC
roX1 mut	TTATAGGGGTTATATGGCTTTA CCTTTTTTTTTTCGGGACTTGCA GTCCGCCCTATGGCCGTCTCA CTCACCCGCTTGCGATGCAAA CGCCTGGCCAGCCCGATAAAA ATTTTTTTTTTTTTGGAGCCGTC TCTCCTTGTCTACAGTATCATT GTTTTTTTTTTCATCCGCGTACC GCCACATGCATCCGATAGAGG TATGC		
roX1A	TTATAGGGGTTATATGGCTTTA CCGCTCTCTTTCGGGACTTGCA GTCCGCCCTATGGCCGTCTCA CTCACCCGCTTGCGATGCAAA CGCCTGGCCAGCCC	F	TTATAGGGGTTATA TGGCTTTACCG
		R	GGGCTGGCCAGGC GT
roX1B	GCTTGCGATGCAAACGCCTGG CCAGCCCGATAGAGAGTGATG GAGAGAGGGAGCCGTCTCTCC	F	GCTTGCGATGCAA ACGCC
		R	GCGAGCGAGACAA

	TTGTCTACAGTATCATTGTCTC GCTCGC		TGATACTGTAG
roX1C	GATAGAGAGTGATGGAGAGAG GGAGCCGTCTCTCCTTGTCTAC AGTATCATTGTCTCGCTCGCAT CCGCGTACCGCCACATGCATC CGATAGAGGTATGC	F	GATGGAGAGAGGG AGCC
		R	GCATACCTCTATCG GATGC
roX2	ACAATTGCCAAATAATACAGAT CGATTTAGAGCGAGATGACAAT AGAGAGGGCGATCTCTCTCGTAT ACGAGTCTTGAAAAGAAAGAGA AGGCGAACGGTGCTGGCTTAG AGAGAGATGGCAATACTAATTA ACTGCAAATACATTTCCGCCAT TTTGTGGCGCTAAAAGTAACG GAAATTCGAGATGCTTTTAGGG CTG	F	ACAATTGCCAAATA ATACAGATCG
		R	CAGCCCTAAAAGC ATCTCG
roX2 mut	ACAATTGCCAAATAATACAGAT CGATTTATTTTTTTTTGACAATA GAGAGATTTATTTTTTTTTGACA AAGTCTTGAAAATTTTTTTAGG CGAACGGTGCTGGCTTATTTTT TTTTGGCAATACTAATTAAGT CAAATACATTTCCGCCATTTTG TTGGCGCTAAAAGTAACGGAAA TTCGAGATGCTTTTAGGGCTG		
roX2A	ACAATTGCCAAATAATACAGAT CGATTTAGAGCGAGATGACAAT AGAGAGGGCGATCTCTCTCGTAT ACGAGTCTTGAAAAGAAAGAGA AGGCGAACGGTG	F	ACAATTGCCAAATA ATACAGATCG
		R	CACCGTTGCGCTT CTCTTT
roX2B	CGAGTCTTGAAAAGAAAGAGAA GGCGAACGGTGCTGGCTTAGA GAGAGATGGCAATACTAATTA CTGCAAATACATTTCCGCCATT TTGTTGGCGCTAA	F	CGAGTCTTGAAA GAAAGAGAAGG
		R	TTAGCGCCAACAA AATGG
roX2C	CTGGCTTAGAGAGAGATGGCA ATACTAATTAAGTCAAATACAT TTCCGCCATTTTGTGGCGCTA AAAGTAACGGAAATTCGAGATG CTTTTAGGGCTG	F	CTGGCTTAGAGAG AGATGGCAATA
		R	CAGCCCTAAAAGC ATCTCG
CG14446	GTGTGTGCGTGCGCCAGAGAA AGTGAGATAGAGAGAGAGAGA GAGCGAGCGAGAGCGAAAGTG AGGGAGACAGCGACAGAGAGA	F	GTGTGTGCGTGCG CC
		R	CCTCTTTCTCTTAA CTTCTCTCTG

	AGTTAAGAGAAAGAGG		
5C2	TTAGAGCAGAATGTATTTTAAAT ATCAATGTTTCGATGTAGAAAT TGAATGGTTTAAATCACGTTCA CACAACTTAGAAAAGAGATAGCG ATGGCGGTGTGAAAGAGAGCG AGATAGTTGGAAGCTTCATGGA AATGAAAGAGAGGTAGTTTTTG GAAATGAAAGTTGTACTAGAAA TAAGTATTTTATGTATATAGAAT A	F	TTAGAGCAGAATGT ATTTTAAATATCAA TG
		R	TATTCTATATACAT AAAATACTTATTTT TAGTAC
5C2 mut1	TTAGAGCAGAATGTATTTTAAAT ATCAATGTTTCGATGTAGAAAT TGAATGGTTTAAATCACGTTCA CACAACTTATTTTTTTTTAGCGA TGGCGGTGTGAAAGAGAGCGA GATAGTTGGAAGCTTCATGGAA ATGAAAGAGAGGTAGTTTTTG AAATGAAAGTTGTACTAGAAAT AAGTATTTTATGTATATAGAATA		
5C2 mut2	TTAGAGCAGAATGTATTTTAAAT ATCAATGTTTCGATGTAGAAAT TGAATGGTTTAAATCACGTTCA CACAACTTAGAAAAGAGATAGCG ATGGCGGTGTGAAAGATTTTTT TTAGTTGGAAGCTTCATGGAA ATGAAAGAGAGGTAGTTTTTG AAATGAAAGTTGTACTAGAAAT AAGTATTTTATGTATATAGAATA		
5C2 mut3	TTAGAGCAGAATGTATTTTAAAT ATCAATGTTTCGATGTAGAAAT TGAATGGTTTAAATCACGTTCA CACAACTTAGAAAAGAGATAGCG ATGGCGGTGTGAAAGAGAGCG AGATAGTTGGAAGCTTCATGGA AATGATTTTTTTTTAGTTTTTG AAATGAAAGTTGTACTAGAAAT AAGTATTTTATGTATATAGAATA		
5C2 mut1+2	TTAGAGCAGAATGTATTTTAAAT ATCAATGTTTCGATGTAGAAAT TGAATGGTTTAAATCACGTTCA CACAACTTATTTTTTTTTAGCGA TGGCGGTGTGAAAGATTTTTT TTAGTTGGAAGCTTCATGGAAA TGAAAGAGAGGTAGTTTTTGGA AATGAAAGTTGTACTAGAAATA AGTATTTTATGTATATAGAATA		

5C2 mut1+2+ 3	TTAGAGCAGAATGTATTTTAAAT ATCAATGTTTCGATGTAGAAAT TGAATGGTTTAAATCACGTTCA CACAACCTATTTTTTTTTAGCGA TGGCGGTGTGAAAGATTTTTTT TTAGTTGGAAGCTTCATGGAAA TGATTTTTTTTTAGTTTTTGAA ATGAAAGTTGTACTAGAAATAA GTATTTTATGTATATAGAATA			
bif	CGGTTGAGTTTTCTCTTGCCT TCTTCTCGTTGGCAAACCCGTT TTGTTTGTGCTTTGTGCCGTTT TCCTCTCCCACTCCCTCTCCCT CTCTCTCTCTCTCTCTCTCTCT CCCTCTCAGCTGCCATGTTTCT CTCTATGTCCATATCCATATCT CCATATCTCAATGGCTCCATGG CTCCATGG CTCTCTTCGCCCTCTT	F	CGGTTGAGTTTTCC TCTTGCC	
		R	AAGAGGGCGAAGA GAGCC	
phk gamma	CGTGTGCACCTCTAACCCAAAC ACGAAGCCAAAAGAAAGAGAT CGATCCGCATTGCCGAGTATC CAGATGGTTTAAATCCGTATCCT TTGCCCTCTCTCTCTCTCCCTC TCTCTCTCTCTGTCTCTCTCTT CTGATGTTTAGCTGCGCAAACA AAGTCGCTTCAACGCGCGCAA AAAGTTCCAATTCGCCATACTC GTGA	F	CGTGTGCACCTCT AACCCA	
		R	TCACGAGTATGGC GAATTGGA	
sw	GTGTAAGGGGGAAAGCGAGAG AGAGGGAGAGAAGGAGCATTG CAGCACATTGGTTCCTCTCGCC TGTCTCGTCTTCTTTTCGAGTG TGGCTGTGTGCGTGCGTGCA CACATCGGGGGAAAGGGATAG ACAGAGGCAGAGAAAGTGAGT GAGAGGGAGAGAGAGAGAGAG AGAGAAAGGAAAGATGTGGCG GCAGGTGCA	F	GTGTAAGGGGGAA AGCGAGA	
		R	TGCACCTGCCGCC AC	
rab35	TGTTTTCTTTTTTCTGCTGCCT GCTCCACACTCACACACACA GGAGCACACTCCCTGCGCTTA CTCTCTCTCTCTCTCTTTTCG CACTCTCTCGGAAGCGCCCTC TCTTTCGCCCCAAGCTACGACC CTCTCGCTCGCACTGGTGAG CCGGCTGAATATCTGAGATAAG CTGTTGTTCTTTCTGGCTGCTG GTCTG	F	TGTTTTCTTTTTTTC TGCTGCCT	
		R	CAGACCAGCAGCC AGAAA	

CES 11D1	AGCCTGAAGCCTGTGGACTCC	F	AGCCTGAAGCCTG
	GCTGATGCCGTGGAAACGGAC		TGG
	AGCGTAAGTTGCATTGGCATTG	R	TATCCATAGATGCA
	CCATCCATCCATCCATCTCGCT		CATCAGCG
	CATATTCGGGTTTTTTATTGCA		
	GGCCAGCTACACGCCGAGCCT		
	CTTGTCGCTGATGTGCATCTAT		
	GGATA		

Table S4.1 Sequences (5' - 3') used to generate dsRNA.

GFP	F	TAATACGACTCACTATAGGGAGAGGTGAGCAAGGGCGAGGAGCT
	R	TAATACGACTCACTATAGGGAGATCTTGAAGTTCACCTTGATGCCG
CLAMP	F	TAATACGACTCACTATAGGGGGGCGGCGATTCTATTG
	R	TAATACGACTCACTATAGGGCATGGGAGTGCTGCCCC
GAF	F	GAATTAATACGACTCACTATAGGGATGGTTATGTTGGCTGGCGTCAA
	R	GAATTAATACGACTCACTATAGGGATCTTTACGCGTGGTTTGCGT

Table S4.2 Sequences (5' – 3') used for ChIP-qPCR targets.

Fab-7	F	TAAGCCAACTGGTTTCCAACCTCT
	R	TTGCCCAGGGTAAGTAACGGTAT
roX1	F	ATGCGAGCGAGACAATGATACT
	R	GACTTGCACTCCGCCCTATG
roX2	F	ACGGTGCTGGCTTAGAGAGA
	R	AGTTCTGGTCACCCTGGAAA
CES 5C2	F	ATCAATGTTTCGATGTAGA
	R	CTTCCAACCTATCTCGCTCT
CES 11D1	F	GTGGAAACGGACAGCGTAA
	R	CACATCAGCGACAAGAGGC

Chapter 5 : Discussion of the Function of GAF and CLAMP in Transcriptional Regulation

I wrote this discussion chapter and created the included figures. I received input through comments and edits from Dr. Erica Larschan, Dr. Jennifer Urban, and Dr. Leila Rieder.

Discussion

Gene regulation requires precise targeting of transcriptional complexes. The transcription factors that recruit these complexes therefore must also be carefully regulated. An intricate network of protein interactions orchestrates gene expression, where proper levels of gene expression are reliant on precise targeting of transcription factors. In this work I have defined the dynamic interactions between two similar transcription factors and explored their context-specific functions in the genome. The competition mechanism leading to specific targeting of the MSL complex to the X-chromosome and fine tuning of promoter regulation may be broadly applicable to other groups of transcription factors that share similar binding sites.

Different binding specificities for similar GA-binding proteins

DNA sequence alone is not sufficient to predict the binding pattern of a transcription factor, but it is often the first step in evaluating binding. My goal was to understand the contribution of DNA sequence to CLAMP and GAF localization. At first glance, CLAMP and GAF would both be predicted to bind to most GA-rich *cis*-elements because they have very similar binding motifs. Both motifs are GA-rich, and GAF and CLAMP do overlap in occupancy at thousands of locations (Chapter 3). What is surprising is that these proteins do not overlap in occupancy at all sites in the genome. For example, unlike CLAMP, GAF binds very weakly to CES (Chapter 3; (Soruco et al. 2013), the initial MSL complex recruitment sites (A. Alekseyenko et al. 2008; Straub et al. 2008).

We can evaluate the structure of the proteins themselves for clues about binding differences. First, CLAMP has more zinc-fingers compared to GAF (6 fingers in the CLAMP DNA binding domain versus 1 finger in the GAF DNA binding domain). While GAF can oligomerize to extend its DNA sequence motif (Wilkins and Lis 1999; Bonchuk et al. 2011), the modularity of multiple zinc-fingers within one protein likely allows CLAMP to tolerate breaks within the GA-repeat sequences. Therefore, rather than adding specificity, the additional CLAMP zinc-fingers could increase the diversity of sites that are bound and allow the protein to tolerate more deviation from a strict GA-repeat.

In Chapter 2 I defined the *in vitro* CLAMP binding motif through protein binding microarrays (PBMs) and the *in vivo* motif through ChIP-seq (Kuzu et al. 2016). I identified the same GA-rich motif in each case, and specifically determined the minimal 15 bp GA-rich sequence required for CLAMP binding that often has more sequence flexibility within the core binding site. In fact, the 8 bp core of the CLAMP motif is consistently interrupted by two poorly conserved bases (here indicated by lower case). These two bases are found in between more strongly conserved bases resulting in a “GAgAGAGa” motif. In Chapter 3 I performed GAF ChIP-seq and identified the GAF binding motif from binding sites that do not recruit CLAMP, the main feature was a “GAGAG” pentamer that appeared twice in the motif. Therefore, the five contiguous GAGAG bases that are required for

GAF occupancy are not always present in CLAMP binding sites, which likely explains their differential occupancy patterns.

Due to lack of protein stability of most CLAMP zinc finger truncations *in vitro*, I did not directly test whether additional zinc-fingers lead to more stable CLAMP-DNA interactions. After *trl* RNAi, CLAMP did not appear to change in binding to the same extent that GAF changed in response to *clamp* RNAi. It is tempting to hypothesize that perhaps this is due to the larger DNA binding domain of CLAMP, and once it is bound is it more likely to persist at a site even if new sites become available. The stability of CLAMP bound to DNA could be further increased by zinc-fingers wrapping around the DNA more than GAF with its single-zinc finger (Segal et al. 2006; Omichinski and Pedone 1997). Structural approaches, specifically those including the full-length versions of the proteins rather than only the DNA binding domains, will be needed to test this idea. Future experiments could determine if the N-terminal zinc-finger present within CLAMP's glutamine-rich domain contributes to sequence-specific binding, binding stability, or protein-protein interactions (see Appendix).

I further explored the dynamics of binding between GAF and CLAMP in Chapter 4, where I characterized GAF and CLAMP as components of the LBC. It is clear that these proteins can exist as part of the complex as well as independently, but requires further evaluation of LBC binding sites. Although additional sites can be selected for testing based on sites of overlap in ChIP-seq profiles, a caveat from

these data sets is that they are averaged from a population of cells so all of the binding sites identified are unlikely to be bound simultaneously in all cells. Therefore, it may appear as though a binding site is shared between GAF and CLAMP and therefore a promising LBC site, but instead it is actually two separate sub-populations where either GAF or CLAMP are bound at a time. This question could be resolved by performing sequential GAF and CLAMP ChIPs followed by sequencing. The resulting DNA would thus be enriched for sequences interacting simultaneously with both GAF and CLAMP, and therefore more likely to be the LBC rather than the individual proteins. This approach will be valuable not only for identifying LBC sites, but also for future analysis interpreting the RNAi ChIP-seq results presented in Chapter 3. This would include comparing shared versus unique CLAMP and GAF binding sites and reclassifying them based on co-occupancy versus a competition model where the factors may trade-off in binding to the same location.

GAF and CLAMP regulate gene expression through chromatin

Based on the high conservation of CLAMP protein across *Diptera* and its requirement for both male and female viability, I propose that the more ancient function of CLAMP is non sex-specific, and that it was later co-opted for male-specific dosage compensation as a result of the enrichment of its binding sites on the X-chromosome. In Chapter 2 I used the *D. miranda* neo-X chromosome, which is evolutionarily young at one million years old, as an example of an actively evolving X-chromosome (Ellison Bachtrog; Alekseyenko 2013), and

identified that the neo-X is beginning to become enriched for GA repeats. Together, my results favor a model in which the *D. melanogaster* X-chromosome evolved to contain long GA-rich stretches, which are increased in both number and density on the X-chromosome, to drive increased occupancy of CLAMP and subsequent recruitment of MSL complex (Figure 5.1).

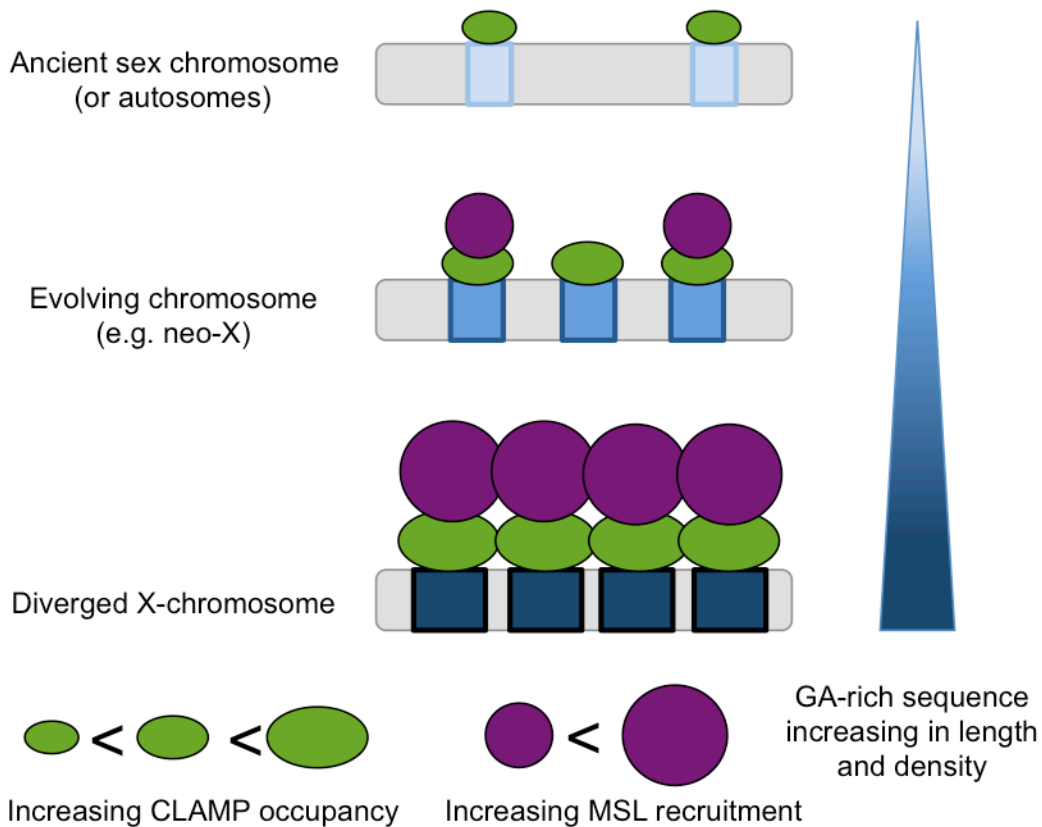


Figure 5.1 The X-chromosome evolved longer, more numerous GA-rich sequences compared to autosomes.

Model of ancient chromosome evolving to acquire more GA-rich sequences for optimal CLAMP and MSL recruitment. As an autosomal chromosome evolves into an X-chromosome, GA-rich sequences expand such as on the *D. miranda* Neo-X chromosome. The diverged X-chromosome has the most long GA-rich sequences, which are also more densely clustered than GA-rich sequences on autosomes.

Increased presence of CLAMP on the X-chromosome presents the opportunity to gain insight into the general transcriptional regulation function that likely predated a role in MSL recruitment. I speculate that the activity of CLAMP on the male X may be an enhanced version of its general function, either due to increased overall amount of CLAMP on the X-chromosome, or due to specific co-factor interactions on the X. It is known that CLAMP plays a major role in promoting the open chromatin environment that is present on the male X-chromosome (J. A. Urban, Kuzu, et al. 2017) (in revision). Based on its DNA binding capabilities, is it possible that CLAMP has a pioneer factor-like function in which it could bind to nucleosome-packaged DNA. This strong chromatin accessibility promoting effect may then be the combination of a global function as a pioneer factor and of additional X-specific co-factor interactions. Co-factors could include MSL complex as well as the NURF chromatin remodeling complex. Follow-up experiments such as EMSAs directly testing the ability of CLAMP to bind to nucleosome-packaged DNA will provide mechanistic insight into the role of CLAMP in chromatin regulation.

Another intriguing explanation for how CLAMP promotes the X-specific open chromatin environment is through competition with GAF for binding partners rather than for DNA sequence. If CLAMP can bind similarly on autosomes and the X-chromosome, perhaps GAF outcompetes CLAMP to recruit NURF at non-CES sites, whereas CLAMP has no competition at CES, allowing it to better perform its chromatin-opening function. In Chapter 4 I address an additional role

for both proteins in context specific chromatin regulation dynamics: their recent characterization as part of the LBC insulator complex. Perhaps GAF and CLAMP have antagonistic functions on autosomes, aside from when they are part of the LBC performing architectural organization of the genome. In Chapter 3 I identified a shared chromatin-opening activity observed through genetic interactions in the PEV assay, in which both factors act as enhancers of variegation leading to whiter eye color, which may be due to both proteins functioning together in the LBC. However, given that this assay requires a transgene on the X-chromosome, it would be interesting to test a similar reporter system located on an autosome.

Differential interaction between CLAMP and GAF promotes context-specific functions of GA-repeat *cis*-elements

GA-rich *cis*-elements have diverse functions throughout the *Drosophila* genome from dosage compensation (A. Alekseyenko et al. 2008) and RNA Pol II pausing (Fuda et al. 2015; Lee et al. 2008) to insulator activity (Ohtsuki and Levine 1998; Wolle et al. 2015), and Polycomb recruitment (Hagstrom, Muller, and Schedl 1997; Strutt, Cavalli, and Paro 1997). Therefore, it is important to determine how these similar motifs are modulated by differential occupancy of GA-binding proteins. For example, CLAMP functions during dosage compensation to open chromatin by recruiting the NURF301 protein to CES (J. A. Urban, Kuzu, et al. 2017) (in revision), whereas it modulates NELF recruitment at promoters (J. A. Urban, Urban, and Larschan 2017) (submitted).

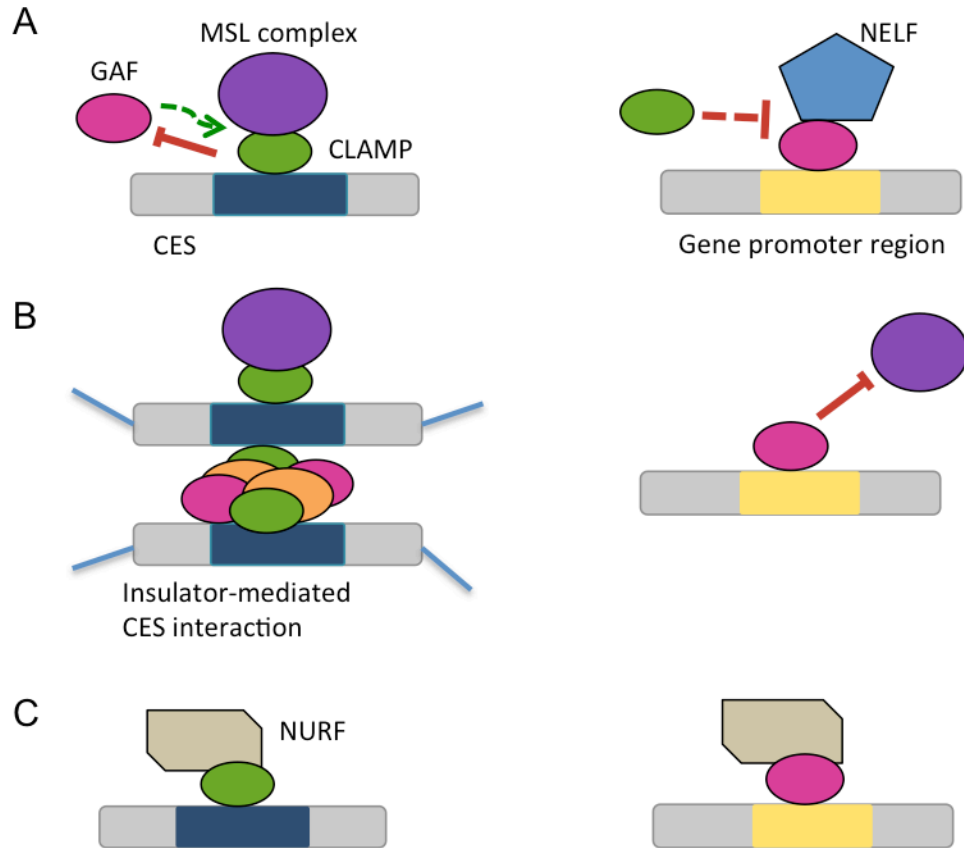


Figure 5.2 Overview of the function of CLAMP and GAF interactions on the X-chromosome.

A) At CES (dark blue), CLAMP (green) outcompetes GAF (pink) for binding, similar to its function at promoters (yellow) where it decreases GAF and NELF (blue) recruitment. Dotted green arrow indicates GAF promotes recruitment MSL complex but is not interacting directly with MSL. Dotted red inhibitor symbol indicates CLAMP negatively regulates but does not completely exclude GAF and NELF from promoters. **B)** Proposed model for GAF requirement in MSL localization. Without directly binding to CES, GAF may facilitate looping of CES through its function with CLAMP in an insulator complex (orange). Alternatively, or in addition to this, GAF localizing to other genomic loci such as promoters prevents off-target MSL-localization. **C)** GAF and CLAMP both open chromatin through NURF (light brown) recruitment, CLAMP specifically at CES, GAF primarily at promoters genome-wide although it may transiently perform this function at CES.

My work has demonstrated that differential competition between CLAMP and GAF promotes the different regulatory function of specific classes of binding sites (Figure 5.2). In Chapter 3 I determined that CLAMP almost completely outcompetes GAF at CES involved in dosage compensation, leading to an

inverse relationship between GAF and MSL occupancy. However, I demonstrated that GAF is also required to promote MSL complex recruitment to these sites, possibly through a mechanism where CLAMP and GAF cycle on and off the chromatin and recruit the NURF complex to CES, maintaining an open chromatin environment where MSL complex can interact with CLAMP.

Furthermore, in Chapter 3 I showed that CLAMP modestly out competes GAF at genes regulated by GAF, CLAMP and NELF, especially at genes negatively regulated by CLAMP. Coupled with prior data that indicates NELF occupancy is positively regulated by GAF at several paused genes (Lee et al. 2008) and negatively regulated by CLAMP (J. A. Urban, Urban, and Larschan 2017) (submitted), my work indicates that CLAMP prevents elevated levels of GAF and NELF occupancy at paused promoters. Co-localizing a negative and a positive regulator of NELF recruitment at the same promoters provides a potential mechanism to assure that NELF occupancy is precisely regulated for accurate rate of release of RNA Pol II from promoters into gene bodies.

As part of Chapter 3 I determined that CLAMP is more enriched than GAF at Polycomb repressive complex target sites. The polycomb response element (PRE) is also GA-rich. However, the preference for CLAMP binding versus GAF was unexpected given the previous literature on GAF's role in Polycomb regulation (Strutt, Cavalli, and Paro 1997; Hagstrom, Muller, and Schedl 1997). Therefore, it is possible that CLAMP and GAF also cycle on and off of PREs to

mediate repression. In the future, it will be of interest to compare the dynamics of how CLAMP and GAF interact with PREs, compared other locations in the genome. Despite “State 6” Polycomb silenced chromatin not being statistically enriched for GAF in our GAF ChIP-seq data (Chapter 3), this does not necessarily indicate GAF is not present at Polycomb sites at all. In fact, we found that GAF, despite not being enriched at CES, influences MSL-recruitment to CES. Therefore, it is possible that GAF may still regulate Polycomb recruitment in a similar indirect manner. This has implications for higher order organization of the genome because recent high resolution Hi-C determined that GA-rich PRE elements are enriched at boundaries between TADs (Eagen, Aiden, and Kornberg 2017).

The specific three-dimensional organization of the X-chromosome likely promotes targeting by MSL complex.

Recent work identifying GA-rich PRE sites at boundaries is not the first link between CLAMP and/or GAF binding sites and three-dimensional genome organization. CES are enriched at TAD boundaries, which are established independently of MSL complex binding (Ramírez et al. 2015). Since TAD boundaries have insulator-like qualities, in Chapter 4 I functionally tested the role of the LBC complex at a subset of CES, including *roX1*, *roX2*, and CES5C2. I demonstrated a link between sites interacting in three-dimensional space and LBC binding, although as mentioned above, the full LBC binding profile remains unknown.

In addition to the finding that CLAMP is part of the LBC, evidence that the three-dimensional organization of the X-chromosome forms independently of MSL complex (Ramírez et al. 2015) and that CLAMP can alter chromatin accessibility over 15 kb regions surrounding its binding site independently of MSL binding (J. A. Urban, Kuzu, et al. 2017) (in revision) adds to our interest in CLAMP as a regulator of long-range genome interactions. The role of CLAMP as an insulator (Chapter 4) has already launched new areas of investigation, and follow-up Hi-C experiments before and after RNAi reduction of *clamp* are already underway (William Jordan, unpublished). In addition to this structural insulator function, it will be of interest to test the role of CLAMP for other insulator properties introduced in Chapter 1, such as enhancer-blocking. When physically located in between the enhancer and the promoter, enhancer-blocking insulators prevent inappropriate enhancer-promoter interaction (Kellum and Schedl 1992). CLAMP can repress expression of some genes while activating others, and so it is possible that enhancer-blocking activity could be part of this diversified regulatory capacity. Experiments designed to characterize CLAMP insulator activity *in vivo* will help interpret existing transcriptional data sets.

Summary

My results have defined the evolution of the GA-rich CLAMP binding site (Chapter 2), determined the functional binding relationship between GAF and CLAMP through competition and relocalization experiments (Chapter 3), and discovered a new insulator function for CLAMP as part of the GAF-containing

LBC (Chapter 4). Further analysis of recently completed sequencing work (described in Chapter 3) may lead to additional insights on the relationship between the GAF and CLAMP GA-binding proteins. In addition, current research projects are building on these results through collaboration on a structure-function analysis of CLAMP (see Appendix), as well as a Hi-C experiment mentioned above. Overall these experiments provide a solid foundation for understanding transcription factor binding and the role of these two GA-targeting proteins in dosage compensation, and have informed future studies that will yield even more exciting steps forward.

A. Appendix: Tools for mapping CLAMP and MSL interaction domains

I wrote the text and generated the figures for this Appendix. In a rotation project conceptualized by MCB alumni Dr. Jessica Chery, I performed CXC-MSL2 cloning and pull-down assays. I designed the CLAMP truncation constructs, which were cloned by Larschan Laboratory alumni Sonia Boor during her undergraduate work in the laboratory under my mentorship.

Abstract:

The Male-Specific Lethal (MSL) complex in *Drosophila melanogaster* is necessary for the increase in transcription levels of genes on the single male X-chromosome during dosage compensation. The MSL complex localizes to MSL recognition elements (MREs), which are motifs clustered on the X-chromosome. Recently, the glutamine-rich zinc-finger protein Chromatin-Linked Adaptor for MSL Proteins (CLAMP) was discovered and demonstrated to bind the same MRE sequences as the MSL complex. Although CLAMP was identified as a required component for MSL targeting to the X-chromosome, the question of which protein-protein interactions are required for this recruitment remains open. There are no putative sequence-specific DNA binding domains in the MSL complex proteins. However, the CXC zinc-finger domain of the MSL protein MSL2 was found to have some low-affinity DNA binding activity. Here, the CXC domain of MSL2 was cloned with a GST-tag and used in pull-down binding assays with MBP-tagged CLAMP. Additional CLAMP truncation constructs were generated that will be valuable tools for future *in vivo* experiments. Mapping the interaction between CLAMP and the MSL complex will be an exciting step forward toward answering the question of how these proteins specifically interact to coordinate the upregulated gene expression on the male X-chromosome.

Introduction:

Dosage compensation is the essential biological process by which heterogametic species normalize gene expression on the sex chromosomes to that of the autosomes (Ohno 1967; C. Disteche 2012). X-chromosome identification is the first step in this multi-step process of dosage compensation. After initial X-identification in mammals, dosage compensation occurs by increasing transcription of most X-linked genes followed by the inactivation of one of the two female X-chromosomes (Lyon 1961; Deng et al. 2011). Alternately, in *Drosophila melanogaster* transcription is increased on the single male X-chromosome after the X-identification step (Hamada et al. 2005). Although mechanisms of dosage compensation differ among species, the initial step of X-chromosome identification is conserved and remains poorly understood.

Drosophila melanogaster is an ideal model system to study X-identification because it lacks the extra inactivation step, and is easily manipulated due to the array of genetic tools available. In *Drosophila*, the master regulator of dosage compensation is the Male Specific Lethal (MSL) complex that is expressed only in males, and is targeted to the single X-chromosome (J. Belote and Lucchesi 1980). MSL complex is comprised of two core structural proteins, Male-Specific Lethal 1 and 2 (MSL1 and MSL2), as well as three other proteins (MSL3, maleless (MLE), and Males absent On the First (MOF)) and two non-coding RNAs (Sural et al. 2008; Izzo et al. 2008; Meller and Rattner 2002). Initial MSL localization to the X-chromosome is associated with DNA motifs that are enriched for “GA” dinucleotides (A. Alekseyenko et al. 2008). The sites where MSL localizes initially prior to spreading to active genes are referred to as Chromatin Entry Sites (CES) (A. Alekseyenko et al. 2008). Despite sequence-specific targeting of MSL, these DNA sequences are not unique to the X-chromosome. However, they are more enriched on X compared to autosomes as well as more densely clustered potentially providing optimal

targets (Kuzu et al. 2016) (Chapter 2). Direct binding between MSL and these sites has been tested in electrophoretic mobility shift assays (EMSAs) and only weak interactions with a single “C-G” DNA base pair were identified using the CXC domain of MSL2 (Fauth et al. 2010). A CXC domain features 9 cysteine amino acids that coordinate three zinc ions (Hauser et al. 2000). More recently, a more high-throughput protein binding sequencing assay was performed, which again identified weak interaction with an extended GA-rich motif using the CXC domain and the neighboring Proline/Basic-rich C-terminal region of the MSL2 protein (Villa et al. 2012). Given the poor binding specificity of MSL2 that is unlikely to be sufficient for MSL complex recruitment to its target sites that contain full 21-bp MRE sequences, it was logical to consider testing additional proteins not in the MSL complex but found to be involved in MSL localization.

Specifically, the ideal candidate had been identified in an MSL activity-dependent RNAi screen (Larschan et al. 2012) and in a cross-linking mass spectrometry approach identifying MSL-interacting proteins (C. I. Wang et al. 2013). This protein, previously called CG1832, was further characterized and named Chromatin-Linked Adaptor for MSL Proteins (CLAMP) (Soruco et al. 2013). CLAMP is required for MSL recruitment to CES but also binds directly to MRE-like sequences genome-wide. A CLAMP protein-null mutant generated using the CRISPR/Cas9 gene editing system is lethal in both males and females (J. A. Urban, Doherty, et al. 2017). The requirement of CLAMP for female viability was not surprising given that CLAMP is not an X-specific protein (Soruco et al. 2013). However, the role of CLAMP genome-wide does complicate the search for an MSL-recruitment mechanism which is X-chromosome specific. Although some context-specific functions of CLAMP have been investigated, such as the role of CLAMP at the histone locus (Rieder, Koreski, Boltz, Kuzu, Urban, Bowman, Zeidman, et al. 2017),

specifically mapping the domains required for CLAMP and MSL interactions will allow us to understand this X-specific interaction.

Results and Discussion:

The mechanism of MSL complex targeting remains poorly understood. The core MSL complex proteins, MSL1 and MSL2, can form a subcomplex in the absence of the other complex components (A. Alekseyenko et al. 2008; Hallacli et al. 2012). Together, MSL1 and MSL2 are sufficient for targeting the complex to CLAMP-bound CES on the X-chromosome (A. Alekseyenko et al. 2008; Sural et al. 2008; Richard L Kelley et al. 1999; Straub et al. 2008). In the absence of CLAMP, however, MSL is absent from the X-chromosome when measured via fluorescent signal on male *Drosophila* salivary gland polytene chromosomes and by ChIP-seq (Soruco et al. 2013) (Chapter 3). Therefore, MSL1 and MSL2 are promising candidates for CLAMP interaction and will be tested separately from the other complex components.

Given the low-affinity DNA interaction demonstrated by the CXC domain of MSL2, we hypothesized that this domain may interact with CLAMP to identify correct binding sites. CLAMP has a predicted DNA binding C-terminal domain containing six tandem C2H2 zinc-fingers (Soruco et al. 2013; Kuzu et al. 2016; J. A. Urban, Doherty, et al. 2017) (Chapters 2 and 3). Zinc-fingers have been shown to have protein-protein interaction potential (Leon and Roth 2000). Therefore, the zinc-finger domain of CLAMP was first tested for interaction with the MSL2 CXC zinc-finger domain. A Maltose Binding Protein (MBP) tagged version of the CLAMP DNA binding domain was already available (Amino acids 350-561) (Kuzu et al. 2016) (Chapter 2). Therefore, a different tag was used for the MSL2-CXC domain, the Glutathione-S-Transferase (GST) tag.

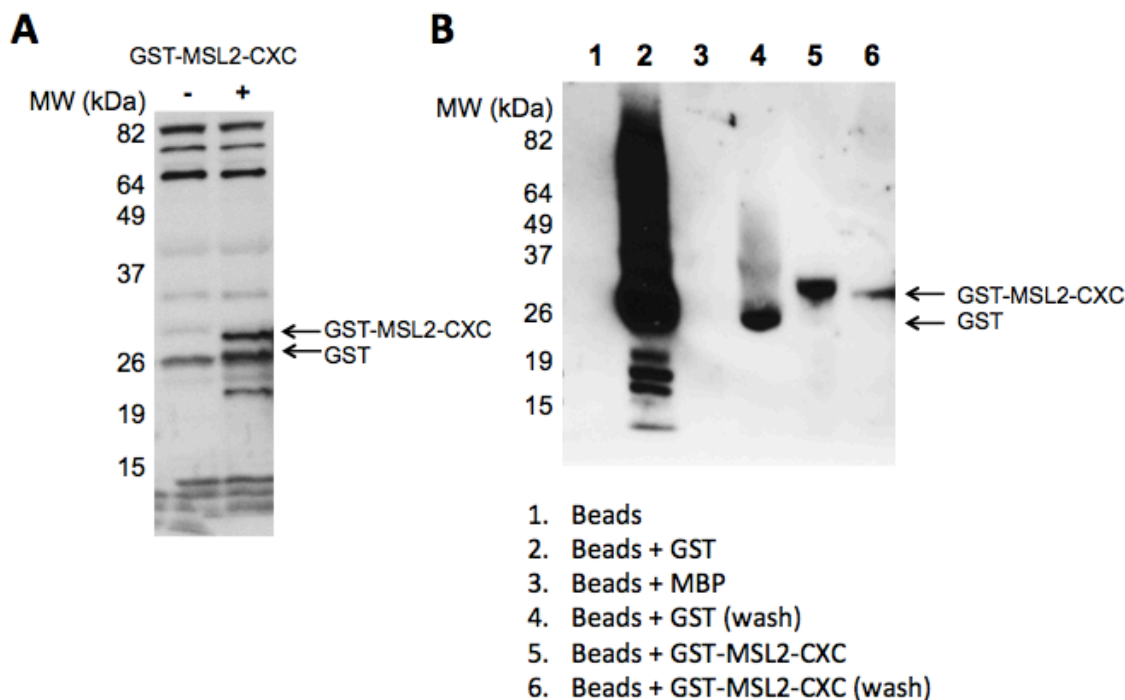


Figure A.1 Expression and binding of GST-MSL2-CXC fusion protein to glutathione beads.
A. Expression of either empty pDEST15 vector (-) control, or expression GST-MSL2-CXC construct (+) in pDEST15. **B.** Testing for proper binding to glutathione beads using anti-GST western blotting of protein samples. Control maltose binding protein was run in lane 3 and did not bind to beads (tested for detection in anti-MBP western, not shown).

It was necessary to test expression of GST-MSL2-CXC prior to performing interaction assays. Successful expression was indicated by the presence of a 29 kDa protein detected using western blotting in an expression reaction (+) compared to an empty vector control (-) lacking the MSL2-CXC DNA insert (Figure A.1A). After confirming expression of the fusion GST-MSL2-CXC protein, a control experiment was performed using glutathione beads with either GST alone or GST-MSL2-CXC followed by anti-GST western blotting (Figure A.1B). This confirmed that the experimental conditions selected would allow for successful GST-glutathione bead interactions. Glutathione beads were then used to bind GST-MSL2-CXC and any interacting proteins present after incubation

with the CLAMP DNA binding domain (MBP-CLAMP) fusion protein. Western blotting for GST (left) and MBP (right) of the resulting protein sample was performed to test for interaction with MBP-CLAMP (Figure A.2). The presence of a 64 kDa band in the eluted protein lane of the anti-MBP western indicates that it was pulled-down by interacting to glutathione bead-bound GST-MSL2-CXC protein.

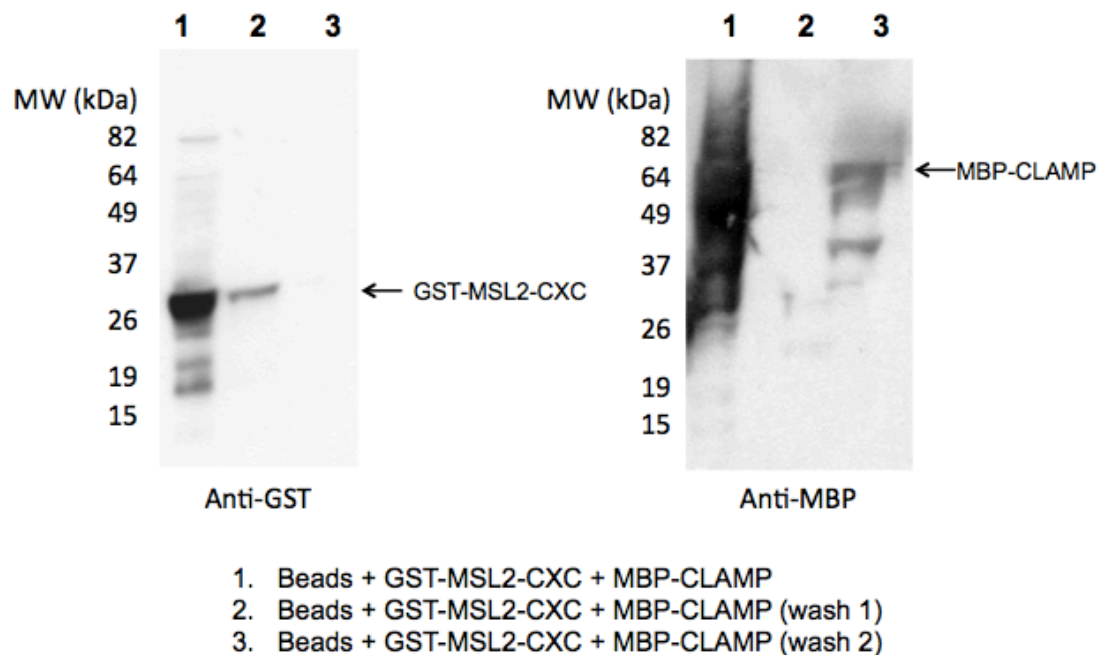


Figure A.2 GST-MSL2-CXC interacts with MBP-CLAMP in a GST pull-down assay.

Left: anti-GST western blot testing for GST-tagged protein in samples containing GST-MSL2-CXC and MBP-CLAMP (lane 1), and washes collected during the experiment (lanes 2 and 3). **Right:** anti-MBP western blot testing for MBP-tagged protein in the same set of samples.

Therefore, there is a putative interaction between MSL2-CXC domain and the CLAMP DNA binding domain. There are a few caveats to consider for follow up studies. The GST-MSL2-CXC protein was expressed using an *in vitro* transcription and translation expression kit, and was not purified prior to incubation with the glutathione beads. Therefore, it is possible that other GST proteins (or even GST itself) were mediating the interaction with the MBP-CLAMP DNA binding domain fusion protein. Ideally, this experiment would be repeated swapping the tags using a GST-CLAMP and MBP-MSL2-

CXC domain (or other combinations of epitope tags). Additional controls would include testing for the MBP and GST protein tags alone in combination with the tagged CXC or CLAMP proteins to make sure the result is not an artifact of the selected tags.

Additional future directions should include other CLAMP domains as well as mutation of specific sub-domains (such as individual zinc-fingers) to identify the specific interaction site. Recent work has indicated that there is variation in the binding relationship between MSL complex and CLAMP at CES (Soruco et al. 2013). Specifically, the CES have been divided into three groups: MSL-dependent sites, MSL-independent sites, and partially MSL-dependent sites that result in synergistic binding between CLAMP and MSL complex. It would be exciting to test interactions at these different groups to see if different parts of CLAMP or MSL interact at different groups of sites.

To determine if the interaction between CLAMP and MSL complex differs at different CES, it will be best to map their interaction domains *in vivo* using tagged CLAMP and MSL2 mutant constructs in flies that lack the endogenous proteins. Chromatin immunoprecipitation of mutant proteins followed by quantitative polymerase chain reaction (ChIP-qPCR) amplification of binding sites will allow us to determine whether CLAMP and MSL still co-localize at different classes of CES. Recent unpublished work has been completed in the laboratory of Dr. Pavel Georgiev to replace the endogenous CLAMP locus with an attP genomic integration site. This will allow the opportunity to insert CLAMP truncation constructs into the genome to replace endogenous CLAMP. The CLAMP truncation-mutant constructs have also been created, with carefully selected features such as an N-terminal Halo tag. The versatility of this tag allows it to be used in biochemical assays, such as chromatin immunoprecipitation and mass

spectrometry as well as microscopy and live imaging (Los and Wood 2007). A full-length control construct was made to ensure that the Halo tag does not alter the localization of CLAMP, and that any results observed are due to the truncations rather than the tag. The other constructs include amino acids 1-287, 288-561, 154-561, and 1-153 fused to 288-561 (Figure A.3). These were designed to test requirement for the N-terminal glutamine-rich protein regions since *in vitro* studies focused on the DNA binding domain. Other than a predicted zinc-finger at AAs 127-153, the N-terminal domain has no predicted structural features. It is however glutamine-rich, which is a known protein-protein interaction feature (Atanesyan et al. 2012). The collaborating laboratory will make smaller mutations in the zinc-finger domain rather than truncations that may disrupt folding of this domain.

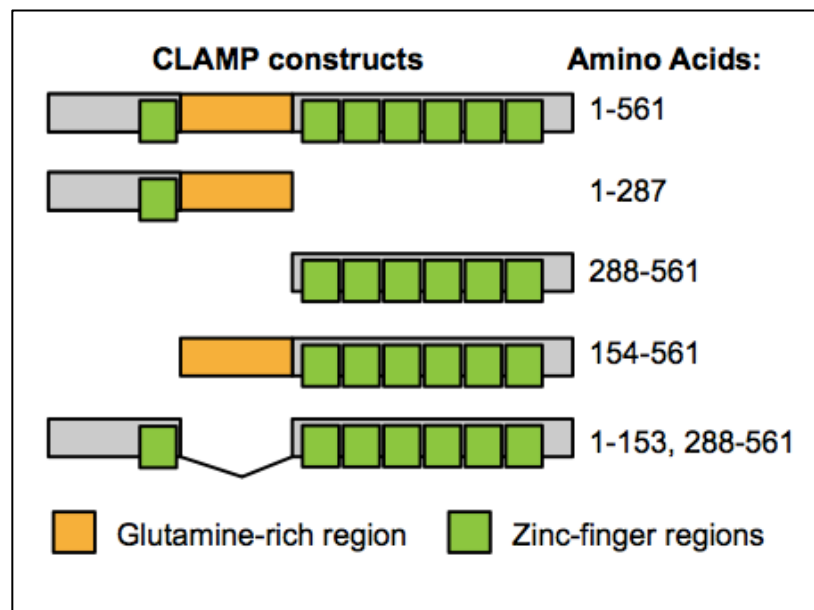


Figure A.3 CLAMP truncation construct design.

Five constructs were generated of the CLAMP protein. The glutamine-rich region is highlighted yellow, while the C-terminal six zinc-fingers and N-terminal zinc finger regions are in green. The corresponding amino acids are listed to the right of each construct. Not shown: N-terminal Halo Tag, C-terminal 3'UTR.

With these new tools to complement the previous work it will be possible to not only test for MSL-CLAMP interactions, but also to investigate the non-sex-specific function of CLAMP and its role in females. CLAMP is a transcription factor with several clear context specific roles including functions on the X-chromosome and histone locus (Soruco et al. 2013; Kuzu et al. 2016; Rieder, Koreski, Boltz, Kuzu, Urban, Bowman, Zeideman, et al. 2017; J. A. Urban, Kuzu, et al. 2017) (Chapters 2-4; Urban in revision). It will be exciting to look for phenotypes caused by truncations that will provide clues about each of the context-specific functions. Recent work includes a CLAMP mass-spectrometry experiment to identify interacting proteins (J. A. Urban, Urban, and Larschan 2017) (submitted). This has led to several candidates whose recruitment could be investigated using the CLAMP truncations including NELF, NURF, and the insulator protein Alan Shepard (J. A. Urban, Urban, and Larschan 2017) (submitted). Possible future approaches would include immunostaining, co-IP experiments, and ChIP experiments. The use of the Halo tag will also allow mass spec of the truncations to be performed using an anti-Halo antibody, since some constructs lack the N-terminal CLAMP antibody epitope that would have prevented antibody based approaches without an additional tag. Therefore, these new constructs provide the opportunity to use a broad range of methods to investigate the interaction between MSL and CLAMP. Overall, these experiments will increase our knowledge of dosage compensation, as well as how a transcription factor can perform multiple context-specific functions in the cell.

Materials and Methods:

MSL2-CXC domain cloning:

The gateway cloning system was used to clone MSL2-CXC into a GST-containing destination vector. The 111-bp MSL2-CXC domain was codon optimized and ordered using Genewiz custom gene synthesis, ID: P0943-1/A75919. The following primers were used to add attB primers for cloning into the pDONR221 gateway vector (Invitrogen):

MSL2-CXC Forward primer (5' to 3'):

GGGGACAAGTTTGTACAAAAAAGCAGGCTCCTGCCGCTGCGGCATCAGCG

MSL2-CXC Reverse primer (5' to 3')

GGGGACCACTTTGTACAAGAAAGCTGGGTCCTAGCAGCACACGCAATGGCAGCC

The gateway cloning kit was used to perform the BP Clonase II reaction for cloning the PCR product into pDONR221, and then the LR Clonase II enzyme was used to clone into the destination vector, pDEST15.

GST-MSL2-CXC expression:

The Pure Express *in vitro* transcription and translation kit (New England Biolabs) was used to express GST-MSL2-CXC. Manufacturer protocol was altered by adding 1.25 uL 1mM Zinc Acetate to 25 uL reactions, which were then expressed for 2 hours at 37°C.

MSL2-CXC GST pull down experiments:

The binding/wash buffer used was the same buffer used for previous MBP-CLAMP expression (Kuzu et al. 2016) (Chapter 2). This buffer was 20 mM Tris pH 7.5, 10 uM ZnCl₂, 100 mM NaCl, 5 mM Citrate, and 5 mM B-mercaptoethanol. The pull-down interaction experiments were performed at 4°C or on ice. Beads were pre-bound to GST-tagged protein or control GST with 50 uL glutathione-agarose beads (BD Biosciences) in

protein lo-bind tubes (Eppendorf) for 2 hrs at 4°C on a rotator. Beads were then washed 4X at 4°C with 1 mL cold wash buffer. Beads were collected between washes by spinning for 1 min max speed. Supernatants were saved after the first and last wash for analysis. GST-bound beads were then incubated with MBP-CLAMP. Again, they were rotated for 2 hrs at 4°C, washed 4X. Samples were eluted by adding 1X Laemmli buffer to beads, boiled 5 min at 95°C, centrifuged 2 min at max speed and the supernatant was collected for anti-GST and anti-MBP westerns.

Western blotting:

Western blotting was performed with the Western Breeze kit (Invitrogen), following manufacturer protocols. Samples were run on pre-cast Tris-SDS-glycine 8-16% gels (Invitrogen). Blots were incubated overnight at 4°C in primary antibody. Rabbit anti-GST antibody (Sigma) was used at a 1:1000 dilution, rabbit anti-MBP serum (NEB) was used at 1:1000.

New truncation cloning

Primers were designed to amplify truncations of the CLAMP protein out of cDNA prepared from cell culture. Full-length CLAMP, 154-561 and 287-561 with the CLAMP 3' UTR were all amplified by standard PCR, with a SgfI site on the 5' end and a PmeI site on the 3' end of the construct. PCR products and pFN21A Halo Tag CMV Flexi Vector (Promega) were digested using the provided enzyme bend (PmeI and SgfI) and 5X Flexi buffer (Promega) for 30 minutes at 37°C and heat inactivated for 20 minutes at 65°C. Constructs were ligated in 2X Flexi Ligase Buffer and 1 uL T4 DNA ligase as described in the provided technical manual.

Primers for the constructs were as follows, all 5' to 3' orientation:

1-561 Forward primer: TAAACGCGATCGCCGAAGACCTTACCAAAAACATTATCT

1-561 Reverse primer: TAATGTTTAAACGTATTTTATTCAAATATTTATTTGATT

154-561 Forward primer: TAATGCGATCGCCCAGCAGCAACAACAGCAACATAACG

154-561 Reverse primer: TAATGTTTAAACGTATTTTATTCAAATATTTATTTGATT

288-561 Forward primer: TAATGCGATCGCCGCACAGACTGGCGGTACAACGCCAA

288-561 Reverse primer: TAATGTTTAAACGTATTTTATTCAAATATTTATTTGATT

1-287 Forward primer: GAGCGATAACGCGATCGCCGAAGACCTTA

1-287 Reverse primer: TGCATTTAACTATCTTAAGCCACAATTTGC

3'UTR Forward primer: GATACTTTAAATGCAGAGGTGACAGGC

3'UTR Reverse primer: CCGAATTCGTTTAAACGTATTTTATTCAAATA

Fusion Forward primer: GAAGACCTTACCAAAAACATTATCT

Fusion Reverse primer: TTAAGCCACAATTTGCTGAAGTTGTGCTT

1-153 Forward primer: CGATAACGCGATCGCCGAAGACCTTACCAA

1-153 Reverse primer: TTGTACCGCCAGTCTGTGCCCCGTCTGTATGCAT

288-561 Forward primer: GCACAGACTGGCGGTACAAC

288-561 Reverse primer: GTATTTTATTCAAATATTTATTTGAATTTATGCATT

Fusion Forward primer: GAAGACCTTACCAAAAACATTATCT

Fusion Reverse primer: CCGTCTCTATGCATCCGCT

Two constructs, 1-287 and 1-153:288-561 were generated using the In-Fusion cloning system (Takeda Clontech). Ligations were performed as directed in the manual with 200 ng of purified PCR fragments and 50 ng linearized pFN21A vector. Ligation reactions were transformed into competent cells using the established laboratory transformation protocol, and colonies were screened using restriction digest fragment analysis after preparing mini-prep DNA samples with the Zyppy kit (Zymo research).

To transfer to the final pValium20 *Drosophila* expression vector, all constructs were digested using the EcoRI restriction enzyme (NEB) in buffer 2.1 at 37°C for 4 hours. Samples were heat inactivated at 65°C for 20 min. Ligation, transformation and colony screening was performed as described above.

References

- Abe, Namiko, Iris Dror, Lin Yang, Matthew Slattery, Tianyin Zhou, Harmen J. Bussemaker, Remo Rohs, and Richard S. Mann. 2015. "Deconvolving the Recognition of DNA Shape from Sequence." *Cell* 161 (2): 307–18. doi:10.1016/j.cell.2015.02.008.
- Adelman, Karen, and John T. Lis. 2012. "Promoter-Proximal Pausing of RNA Polymerase II: Emerging Roles in Metazoans." *Nature Reviews Genetics* 13 (10). Nature Publishing Group: 720–31. doi:10.1038/nrg3293.
- Adkins, Nicholas L, Thomas A Hagerman, and Philippe Georgel. 2006. "GAGA Protein: A Multi-Faceted Transcription Factor." *Biochemistry and Cell Biology = Biochimie et Biologie Cellulaire* 84 (4): 559–67. doi:10.1139/o06-062.
- Alekseyenko, A., Shouyong Peng, Erica Larschan, Andrey A. Gorchakov, Ok-Kyung Kyung Lee, Peter Kharchenko, Sean D. McGrath, et al. 2008. "A Sequence Motif within Chromatin Entry Sites Directs MSL Establishment on the Drosophila X Chromosome." *Cell* 134 (4): 599–609. doi:10.1016/j.cell.2008.06.033.
- Alekseyenko, A A, E Larschan, W R Lai, P J Park, and M I Kuroda. 2006. "High-Resolution ChIP-Chip Analysis Reveals That the Drosophila MSL Complex Selectively Identifies Active Genes on the Male X Chromosome." *Genes Dev* 20 (7): 848–57.
- Alekseyenko, AA, and Erica Larschan. 2006. "High-Resolution ChIP–chip Analysis Reveals That the Drosophila MSL Complex Selectively Identifies Active Genes on the Male X Chromosome." *Genes & ...*, 848–57. doi:10.1101/gad.1400206.is.
- Alekseyenko, Artyom A., Joshua W K Ho, Shouyong Peng, Marnie Gelbart, Michael Y. Tolstorukov, Annette Plachetka, Peter V. Kharchenko, et al. 2012. "Sequence-Specific Targeting of Dosage Compensation in Drosophila Favors an Active

- Chromatin Context." *PLoS Genetics* 8 (4). doi:10.1371/journal.pgen.1002646.
- Alekseyenko, Artyom a, Christopher E Ellison, Andrey a Gorchakov, Qi Zhou, Vera B Kaiser, Nick Toda, Zaak Walton, et al. 2013. "Conservation and de Novo Acquisition of Dosage Compensation on Newly Evolved Sex Chromosomes in *Drosophila*." *Genes & Development* 27 (8): 853–58. doi:10.1101/gad.215426.113.
- Alén, Claudia, Nicholas A Kent, Hannah S Jones, Justin O’Sullivan, Agustín Aranda, and Nicholas J Proudfoot. 2002. "A Role for Chromatin Remodeling in Transcriptional Termination by RNA Polymerase II." *Molecular Cell* 10 (6): 1441–52. doi:10.1016/S1097-2765(02)00778-5.
- Alfrey, V. G., R. Faulkner, and A. E. Mirsky. 1964. "ACETYLATION AND METHYLATION OF HISTONES AND THEIR POSSIBLE ROLE IN THE REGULATION OF RNA SYNTHESIS." *Proceedings of the National Academy of Sciences* 51 (5): 786–94. doi:10.1073/pnas.51.5.786.
- Arnold, Cosmas D., Daniel Gerlach, Christoph Stelzer, Łukasz M. Boryń, Martina Rath, and Alexander Stark. 2013. "Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-Seq." *Science* 339 (6123).
- Atanesyan, Lilit, Viola Günther, Bernhard Dichtl, Oleg Georgiev, and Walter Schaffner. 2012. "Polyglutamine Tracts as Modulators of Transcriptional Activation from Yeast to Mammals." *Biological Chemistry* 393 (1–2): 63–70. doi:10.1515/BC-2011-252.
- Badis, Gwenaél, Michael F. Berger, Anthony A. Philippakis, Shaheynoor Talukder, Andrew R. Gehrke, Savina A. Jaeger, Esther T. Chan, et al. 2009. "Diversity and Complexity in DNA Recognition by Transcription Factors." *Science (New York, N.Y.)* 324 (5935): 1720–23. doi:10.1126/science.1162327.
- Bai, X, E Larschan, S Y Kwon, P Badenhorst, and M I Kuroda. 2007. "Regional Control of Chromatin Organization by Noncoding roX RNAs and the NURF Remodeling Complex in *Drosophila Melanogaster*." *Genetics* 176 (3): 1491–99.

doi:10.1534/genetics.107.071571.

- Bailey, T. L., M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble. 2009. "MEME SUITE: Tools for Motif Discovery and Searching." *Nucleic Acids Research* 37 (Web Server): W202–8. doi:10.1093/nar/gkp335.
- Bailey, T L, and C Elkan. 1994. "Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Biopolymers." *Proceedings. International Conference on Intelligent Systems for Molecular Biology* 2: 28–36.
- Banerji, Julian, Sandro Rusconi, and Walter Schaffner. 1981. "Expression of a β -Globin Gene Is Enhanced by Remote SV40 DNA Sequences." *Cell* 27 (1): 299–308.
- Bardwell, V J, and R Treisman. 1994. "The POZ Domain: A Conserved Protein-Protein Interaction Motif." *Genes & Development* 8 (14). Cold Spring Harbor Laboratory Press: 1664–77. doi:10.1101/GAD.8.14.1664.
- Barozzi, Iros, Marta Simonatto, Silvia Bonifacio, Lin Yang, Remo Rohs, Serena Ghisletti, and Gioacchino Natoli. 2014. "Coregulation of Transcription Factor Binding and Nucleosome Occupancy through DNA Features of Mammalian Enhancers." *Molecular Cell* 54 (5): 844–57. doi:10.1016/j.molcel.2014.04.006.
- Bell, Oliver, Christiane Wirbelauer, Marc Hild, Annette N D Scharf, Michaela Schwaiger, David M MacAlpine, Frédéric Zilbermann, et al. 2007. "Localized H3K36 Methylation States Define Histone H4K16 Acetylation during Transcriptional Elongation in Drosophila." *The EMBO Journal* 26 (24): 4974–84. doi:10.1038/sj.emboj.7601926.
- Belote, J M. 1983. "Male-Specific Lethal Mutations of DROSOPHILA MELANOGASTER . II. Parameters of Gene Action during Male Development." *Genetics* 105 (4): 881–96.
- Belote, JM, and JC Lucchesi. 1980. "Male-Specific Lethal Mutations of Drosophila

- Melanogaster.” *Genetics*, 165–86.
- Belozerov, Vladimir E., Parimal Majumder, Ping Shen, and Haini N. Cai. 2003. “A Novel Boundary Element May Facilitate Independent Gene Regulation in the Antennapedia Complex of *Drosophila*.” *EMBO Journal* 22 (12): 3113–21. doi:10.1093/emboj/cdg297.
- Benyajati, C, L Mueller, N Xu, M Pappano, J Gao, M Mosammamarast, D Conklin, H Granok, C Craig, and S Elgin. 1997. “Multiple Isoforms of GAGA Factor, a Critical Component of Chromatin Structure.” *Nucleic Acids Research* 25 (16): 3345–53.
- Berdasco, María, and Manel Esteller. 2010. “Aberrant Epigenetic Landscape in Cancer: How Cellular Identity Goes Awry.” *Developmental Cell* 19 (5): 698–711. doi:10.1016/j.devcel.2010.10.005.
- Berger, Michael F., and Martha L. Bulyk. 2006. “Protein Binding Microarrays (PBMs) for Rapid, High-Throughput Characterization of the Sequence Specificities of DNA Binding Proteins.” In *Gene Mapping, Discovery, and Expression*, 338:245–60. New Jersey: Humana Press. doi:10.1385/1-59745-097-9:245.
- Berger, Michael F, and Martha L Bulyk. 2009. “Universal Protein-Binding Microarrays for the Comprehensive Characterization of the DNA-Binding Specificities of Transcription Factors.” *Nature Protocols* 4 (3): 393–411. doi:10.1038/nprot.2008.195.
- Blau, J, H Xiao, S McCracken, P O’Hare, J Greenblatt, and D Bentley. 1996. “Three Functional Classes of Transcriptional Activation Domain.” *Molecular and Cellular Biology* 16 (5): 2044–55.
- Bolshakov, V. N., Pantelis Topalis, Claudia Blass, Elena Kokoza, Alessandra della Torre, Fotis C Kafatos, and Christos Louis. 2002. “A Comparative Genomic Analysis of Two Distant Diptera, the Fruit Fly, *Drosophila Melanogaster*, and the Malaria Mosquito, *Anopheles Gambiae*.” *Genome Research* 12 (1): 57–66.

doi:10.1101/gr.196101.

Bonchuk, Artem, Stepan Denisov, Pavel Georgiev, and Oksana Maksimenko. 2011.

“Drosophila BTB/POZ Domains Of ‘ttk Group’ can Form Multimers and Selectively Interact with Each Other.” *Journal of Molecular Biology* 412 (3): 423–36.

doi:10.1016/j.jmb.2011.07.052.

Brown, E. J., and D. Bachtrog. 2014. “The Chromatin Landscape of Drosophila:

Comparisons between Species, Sexes, and Chromosomes.” *Genome Research* 24 (7): 1125–37. doi:10.1101/gr.172155.114.

Brownell, J E, and C D Allis. 1995. “An Activity Gel Assay Detects a Single, Catalytically

Active Histone Acetyltransferase Subunit in Tetrahymena Macronuclei.”

Proceedings of the National Academy of Sciences of the United States of America 92 (14): 6364–68.

Cai, H N, and P Shen. 2001. “Effects of Cis Arrangement of Chromatin Insulators on

Enhancer-Blocking Activity.” *Science (New York, N.Y.)* 291 (5503): 493–95.

doi:10.1126/science.291.5503.493.

Cai, Weili, Ye Jin, Jack Girton, Jorgen Johansen, and Kristen M Johansen. 2010.

“Preparation of Drosophila Polytene Chromosome Squashes for Antibody Labeling.” *Journal of Visualized Experiments : JoVE*, no. 36(February). MyJoVE Corporation. doi:10.3791/1748.

Cai, Yong, Jingji Jin, Selene K. Swanson, Michael D. Cole, Seung Hyuk Choi, Laurence

Florens, Michael P. Washburn, Joan W. Conaway, and Ronald C. Conaway. 2010.

“Subunit Composition and Substrate Specificity of a MOF-Containing Histone Acetyltransferase Distinct from the Male-Specific Lethal (MSL) Complex.” *Journal of Biological Chemistry* 285 (7): 4268–72. doi:10.1074/jbc.C109.087981.

Chen, Kai, Jeff Johnston, Wanqing Shao, Samuel Meier, Cynthia Staber, and Julia

Zeitlinger. 2013. “A Global Change in RNA Polymerase II Pausing during the

- Drosophila Midblastula Transition." *eLife* 2 (January): e00861.
doi:10.7554/eLife.00861.
- Chetverina, Darya, Tsutomu Aoki, Maksim Erokhin, Pavel Georgiev, and Paul Schedl. 2014. "Making Connections: Insulators Organize Eukaryotic Chromosomes into Independent Cis-Regulatory Networks." *BioEssays* 36 (2): 163–72.
doi:10.1002/bies.201300125.
- Cirillo, L. A. 1998. "Binding of the Winged-Helix Transcription Factor HNF3 to a Linker Histone Site on the Nucleosome." *The EMBO Journal* 17 (1): 244–54.
doi:10.1093/emboj/17.1.244.
- Cirillo, Lisa Ann, Frank Robert Lin, Isabel Cuesta, Dara Friedman, Michal Jarnik, and Kenneth S Zaret. 2002. "Opening of Compacted Chromatin by Early Developmental Transcription Factors HNF3 (FoxA) and GATA-4." *Molecular Cell* 9 (2): 279–89.
doi:10.1016/S1097-2765(02)00459-8.
- Cleard, F, Wolle D, Taverner AM, Aoki T, Deshpande G, Andolfatto P, Karch F, and Schedl P. 2017. "Different Evolutionary Strategies To Conserve Chromatin Boundary Function in the Bithorax Complex." *Genetics* 205 (2): 589.
- Clemson, C. M., L. L. Hall, M. Byron, J. McNeil, and J. B. Lawrence. 2006. "The X Chromosome Is Organized into a Gene-Rich Outer Rim and an Internal Core Containing Silenced Nongenic Sequences." *Proceedings of the National Academy of Sciences* 103 (20): 7688–93. doi:10.1073/pnas.0601069103.
- Conaway, Ronald C., and Joan Weliky Conaway. 1990. "Transcription Initiated by RNA Polymerase II and Purified Transcription Factors from Liver." *J.Biol.Chem.* 265 (13): 7559–63.
- Core, Leighton J, Joshua J Waterfall, Daniel A Gilchrist, David C Fargo, Hojoong Kwak, Karen Adelman, and John T Lis. 2012. "Defining the Status of RNA Polymerase at Promoters." *Cell Reports* 2 (4). NIH Public Access: 1025–35.

- doi:10.1016/j.celrep.2012.08.034.
- Cosma, M P, T Tanaka, and K Nasmyth. 1999. "Ordered Recruitment of Transcription and Chromatin Remodeling Factors to a Cell Cycle- and Developmentally Regulated Promoter." *Cell* 97 (3): 299–311.
- Crooks, G. E., Gary Hon, John-Marc Chandonia, and Steven E Brenner. 2004. "WebLogo: A Sequence Logo Generator." *Genome Research* 14 (6): 1188–90. doi:10.1101/gr.849004.
- Dekker, Job, and Edith Heard. 2015. "Structural and Functional Diversity of Topologically Associating Domains." *FEBS Letters* 589 (20PartA): 2877–84. doi:10.1016/j.febslet.2015.08.044.
- Dekker, Job, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. n.d. "Capturing Chromosome Conformation." *Science* 295. American Association for the Advancement of Science: 1306–11. doi:10.2307/3075925.
- Deng, Xinxian, Joel B. Berletch, Wenxiu Ma, Di Kim Nguyen, Joseph B. Hiatt, William S. Noble, Jay Shendure, and Christine M. Disteche. 2013a. "Mammalian X Upregulation Is Associated with Enhanced Transcription Initiation, RNA Half-Life, and MOF-Mediated H4K16 Acetylation." *Developmental Cell* 25 (1): 55–68. doi:10.1016/j.devcel.2013.01.028.
- Deng, Xinxian, Joel B Berletch, Wenxiu Ma, Di Kim Nguyen, Joseph B Hiatt, William S Noble, Jay Shendure, and Christine M Disteche. 2013b. "Mammalian X Upregulation Is Associated with Enhanced Transcription Initiation, RNA Half-Life, and MOF-Mediated H4K16 Acetylation." *Developmental Cell* 25 (1): 55–68. doi:10.1016/j.devcel.2013.01.028.
- Deng, Xinxian, Joseph B Hiatt, Di Kim Nguyen, Sevinc Ercan, David Sturgill, LaDeana W Hillier, Felix Schlesinger, et al. 2011. "Evidence for Compensatory Upregulation of Expressed X-Linked Genes in Mammals, Caenorhabditis Elegans and Drosophila

- Melanogaster.” *Nature Genetics* 43 (12): 1179–85. doi:10.1038/ng.948.
- Deng, Xinxian, and Victoria H Meller. 2006. “roX RNAs Are Required for Increased Expression of X-Linked Genes in *Drosophila Melanogaster* Males.” *Genetics* 174 (4): 1859–66. doi:10.1534/genetics.106.064568.
- . 2009. “Molecularly Severe roX1 Mutations Contribute to Dosage Compensation in *Drosophila*.” *Genesis (New York, N.Y. : 2000)* 47 (1): 49–54. doi:10.1002/dvg.20463.
- Deng, Xinxian, Barbara P Rattner, Skye Souter, and Victoria H Meller. 2005. “The Severity of roX1 Mutations Is Predicted by MSL Localization on the X Chromosome.” *Mechanisms of Development* 122 (10): 1094–1105. doi:10.1016/j.mod.2005.06.004.
- Disteche, Christine M. 2012. “Dosage Compensation of the Sex Chromosomes.” *Annual Review of Genetics* 46 (January): 537–60. doi:10.1146/annurev-genet-110711-155454.
- Disteche, CM. 2012. “Dosage Compensation of the Sex Chromosomes.” *Annual Review of Genetics*, 537–60. doi:10.1146/annurev-genet-110711-155454.Dosage.
- Dixon, Jesse R., David U. Gorkin, and Bing Ren. 2016. “Chromatin Domains: The Unit of Chromosome Organization.” *Molecular Cell*. doi:10.1016/j.molcel.2016.05.018.
- Dixon, Jesse R., Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu, and Bing Ren. 2012. “Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions.” *Nature* 485 (7398): 376–80. doi:10.1038/nature11082.
- Dorn, Rainer, and Veiko Krauss. 2003. “The Modifier of mdg4 Locus in *Drosophila*: Functional Complexity Is Resolved by Trans Splicing.” *Genetica*. doi:10.1023/A:1022983810016.
- Dudley, A. M., J. Aach, M. A. Steffen, and G. M. Church. 2002. “Measuring Absolute

- Expression with Microarrays with a Calibrated Reference Sample and an Extended Signal Intensity Range.” *Proceedings of the National Academy of Sciences* 99 (11): 7554–59. doi:10.1073/pnas.112683499.
- Eagen, Kyle P., Erez Lieberman Aiden, and Roger D. Kornberg. 2017. “Polycomb-Mediated Chromatin Loops Revealed by a Sub-Kilobase Resolution Chromatin Interaction Map.” *bioRxiv*, 99804. doi:10.1101/099804.
- Eissenberg, Joel C. 1989. “Position Effect Variegation in *Drosophila*: Towards a Genetics of Chromatin Assembly.” *BioEssays* 11 (1): 14–17. doi:10.1002/bies.950110105.
- Ellison, C. E., and D. Bachtrög. 2013. “Dosage Compensation via Transposable Element Mediated Rewiring of a Regulatory Network.” *Science* 342 (6160): 846–50. doi:10.1126/science.1239552.
- Ellison, Christopher E, and Doris Bachtrög. 2015. “Non-Allelic Gene Conversion Enables Rapid Evolutionary Change at Multiple Regulatory Sites Encoded by Transposable Elements.” *eLife* 4 (February). doi:10.7554/eLife.05899.
- Farkas, G, J Gausz, M Galloni, and G Reuter. 1994. “The Trithorax-like Gene Encodes the *Drosophila* GAGA Factor.”
- Farkas, G, B A Leibovitch, and S C Elgin. 2000. “Chromatin Organization and Transcriptional Control of Gene Expression in *Drosophila*.” *Gene* 253 (2): 117–36. doi:10.1016/S0378-1119(00)00240-7.
- Farley, E. K., K. M. Olson, W. Zhang, A. J. Brandt, D. S. Rokhsar, and M. S. Levine. 2015. “Suboptimization of Developmental Enhancers.” *Science* 350 (6258): 325–28. doi:10.1126/science.aac6948.
- Fauth, Torsten, Felix Müller-Planitz, Cornelia König, Tobias Straub, and Peter B Becker. 2010. “The DNA Binding CXC Domain of MSL2 Is Required for Faithful Targeting the Dosage Compensation Complex to the X Chromosome.” *Nucleic Acids Research* 38 (10): 3209–21. doi:10.1093/nar/gkq026.

- Ferrari, Francesco, Artyom A Alekseyenko, Peter J Park, and Mitzi I Kuroda. 2014. "Transcriptional Control of a Whole Chromosome: Emerging Models for Dosage Compensation." *Nature Structural & Molecular Biology* 21 (2): 118–25. doi:10.1038/nsmb.2763.
- Ferrari, Francesco, Annette Plachetka, Artyom A. Alekseyenko, Youngsook L. Jung, Fatih Ozsolak, Peter V. Kharchenko, Peter J. Park, and Mitzi I. Kuroda. 2013. "'Jump Start and Gain' Model for Dosage Compensation in Drosophila Based on Direct Sequencing of Nascent Transcripts." *Cell Reports* 5 (3): 629–36. doi:10.1016/j.celrep.2013.09.037.
- Ferrer, Isidro, Marta Barrachina, Berta Puig, M. Martínez de Lagrán, Eulalia Martí, Jesús Avila, and Mara Dierssen. 2005. "Constitutive Dyrk1A Is Abnormally Expressed in Alzheimer Disease, Down Syndrome, Pick Disease, and Related Transgenic Models." *Neurobiology of Disease* 20 (2): 392–400. doi:10.1016/j.nbd.2005.03.020.
- Filion, Guillaume J., Joke G. van Bommel, Ulrich Braunschweig, Wendy Talhout, Jop Kind, Lucas D. Ward, Wim Brugman, et al. 2010. "Systematic Protein Location Mapping Reveals Five Principal Chromatin Types in Drosophila Cells." *Cell* 143 (2): 212–24. doi:10.1016/j.cell.2010.09.009.
- Fitzgerald, Molly, and Thomas Shenk. 1978. "The Sequence 5'-AAUAAA-3' Forms Part of the Recognition Site for Polyadenylation of SV40 mRNAs." *Cell* 24 (April): 251–60. doi:https://doi.org/10.1016/0092-8674(81)90521-3.
- Flanagan, Peter M., Raymond J. Kelleher, Michael H. Sayre, Herbert Tschochner, and Roger D. Kornberg. 1991. "A Mediator Required for Activation of RNA Polymerase II Transcription in Vitro." *Nature* 350 (6317): 436–38. doi:10.1038/350436a0.
- Foo, Sun Melody, Yujia Sun, Bomyi Lim, Ruta Ziukaite, Kevin O'Brien, Chung Yi Nien, Nikolai Kirov, Stanislav Y. Shvartsman, and Christine A. Rushlow. 2014. "Zelda Potentiates Morphogen Activity by Increasing Chromatin Accessibility." *Current*

- Biology* 24 (12): 1341–46. doi:10.1016/j.cub.2014.04.032.
- Fuda, Nicholas J., Michael J. Guertin, Sumeet Sharma, Charles G. Danko, André L. Martins, Adam Siepel, and John T. Lis. 2015. “GAGA Factor Maintains Nucleosome-Free Regions and Has a Role in RNA Polymerase II Recruitment to Promoters.” *PLoS Genetics* 11 (3): 1–22. doi:10.1371/journal.pgen.1005108.
- Fudenberg, Geoffrey, Maxim Imakaev, Carolyn Lu, Anton Goloborodko, Nezar Abdennur, and Leonid A Mirny Correspondence. 2016. “Formation of Chromosomal Domains by Loop Extrusion.” *CellReports* 15: 2038–49. doi:10.1016/j.celrep.2016.04.085.
- Fujioka, Miki, Hemlata Mistry, Paul Schedl, James B. Jaynes, and MA Crosby. 2016. “Determinants of Chromosome Architecture: Insulator Pairing in Cis and in Trans.” Edited by Bas van Steensel. *PLOS Genetics* 12 (2). Public Library of Science: e1005889. doi:10.1371/journal.pgen.1005889.
- Gelbart, Marnie E, and Mitzi I Kuroda. 2009. “Drosophila Dosage Compensation: A Complex Voyage to the X Chromosome.” *Development (Cambridge, England)* 136 (9): 1399–1410. doi:10.1242/dev.029645.
- Gelbart, Marnie E, Erica Larschan, Shouyong Peng, Peter J Park, and Mitzi I Kuroda. 2009. “Drosophila MSL Complex Globally Acetylates H4K16 on the Male X Chromosome for Dosage Compensation.” *Nature Structural & Molecular Biology* 16 (8). Nature Publishing Group: 825–32. doi:10.1038/nsmb.1644.
- Gerasimova, Tatiana I., David A. Gdula, Denis V. Gerasimov, Olga Simonova, and Victor G. Corces. 1995. “A Drosophila Protein That Imparts Directionality on a Chromatin Insulator Is an Enhancer of Position-Effect Variegation.” *Cell* 82 (4): 587–97. doi:10.1016/0092-8674(95)90031-4.
- Ghirlando, Rodolfo, Keith Giles, Humaira Gowher, Tiaojiang Xiao, Zhixiong Xu, Hongjie Yao, and Gary Felsenfeld. 2012. “Chromatin Domains, Insulators, and the

- Regulation of Gene Expression.” *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*. Elsevier B.V. doi:10.1016/j.bbagr.2012.01.016.
- Gilfillan, Gregor D., Cornelia König, Ina K. Dahlsveen, Nicky Prakoura, Tobias Straub, Rosemarie Lamm, Torsten Fauth, and Peter B. Becker. 2007. “Cumulative Contributions of Weak DNA Determinants to Targeting the Drosophila Dosage Compensation Complex.” *Nucleic Acids Research* 35 (11): 3561–72. doi:10.1093/nar/gkm282.
- Gilfillan, Gregor D, Tobias Straub, Elzo De Wit, Frauke Greil, Rosemarie Lamm, Bas Van Steensel, and Peter B Becker. 2006. “Chromosome-Wide Gene-Specific Targeting of the Drosophila Dosage Compensation Complex.” *Genes and Development* 20 (7): 858–70. doi:10.1101/gad.1399406.
- Goldberg, Aaron D, C David Allis, and Emily Bernstein. 2007. “Epigenetics: A Landscape Takes Shape.” *Cell* 128 (4): 635–38. doi:10.1016/j.cell.2007.02.006.
- Gorchakov, AA. 2009. “Long-Range Spreading of Dosage Compensation in Drosophila Captures Transcribed Autosomal Genes Inserted on X.” *Genes & ...*, 2266–71. doi:10.1101/gad.1840409.How.
- Gordân, Raluca, Ning Shen, Iris Dror, Tianyin Zhou, John Horton, Remo Rohs, and Martha L. Bulyk. 2013. “Genomic Regions Flanking E-Box Binding Sites Influence DNA Binding Specificity of bHLH Transcription Factors through DNA Shape.” *Cell Reports* 3 (4): 1093–1104. doi:10.1016/j.celrep.2013.03.014.
- Gordon, David J, Benjamin Resio, and David Pellman. 2012. “Causes and Consequences of Aneuploidy in Cancer.” *Nature Reviews. Genetics* 13 (3). Nature Publishing Group: 189–203. doi:10.1038/nrg3123.
- Goujon, M., H. McWilliam, W. Li, F. Valentin, S. Squizzato, J. Paern, and R. Lopez. 2010. “A New Bioinformatics Analysis Tools Framework at EMBL-EBI.” *Nucleic Acids Research* 38 (Web Server): W695–99. doi:10.1093/nar/gkq313.

- Graba, Yacine, Patrick Laurenti, Laurent Perrin, Denise Aragnol, and Jacques Pradel. 1994. "The Modifier of Variegation modulo Gene Acts Downstream of Dorsoventral and HOM-C Genes and Is Required for Morphogenesis in *Drosophila*." *Developmental Biology* 166 (2): 704–15. doi:10.1006/dbio.1994.1349.
- Grant, C. E., T. L. Bailey, and W. S. Noble. 2011. "FIMO: Scanning for Occurrences of a given Motif." *Bioinformatics* 27 (7): 1017–18. doi:10.1093/bioinformatics/btr064.
- Greenberg, Anthony J, Judith L Yanowitz, and Paul Schedl. 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 (1): 279–89. doi:10.1534/genetics.166.1.279.
- Guertin, Michael J, and John T Lis. 2013. "Mechanisms by Which Transcription Factors Gain Access to Target Sequence Elements in Chromatin." *Current Opinion in Genetics and Development*. doi:10.1016/j.gde.2012.11.008.
- Guimera, Jordi, Caty Casas, Xavier Estivill, and Melanie Pritchard. 1999. "HumanMinibrainHomologue (MNBH/DYRK1): Characterization, Alternative Splicing, Differential Tissue Expression, and Overexpression in Down Syndrome." *Genomics* 57 (3): 407–18. doi:10.1006/geno.1999.5775.
- Gurley, L R, R A Walters, and R A Tobey. 1974. "Cell Cycle-Specific Changes in Histone Phosphorylation Associated with Cell Proliferation and Chromosome Condensation." *The Journal of Cell Biology* 60 (2). The Rockefeller University Press: 356–64.
- Hagstrom, K, M Muller, and P Schedl. 1997. "A Polycomb and GAGA Dependent Silencer Adjoins the Fab-7 Boundary in the *Drosophila* Bithorax Complex." *Genetics* 146 (4): 1365–80.
- Hallacli, Erinc, Michael Lipp, Plamen Georgiev, Clare Spielman, Stephen Cusack, Asifa Akhtar, and Jan Kadlec. 2012. "Msl1-Mediated Dimerization of the Dosage

- Compensation Complex Is Essential for Male X-Chromosome Regulation in *Drosophila*." *Molecular Cell* 48 (4). Elsevier Inc.: 587–600.
doi:10.1016/j.molcel.2012.09.014.
- Hamada, Fumika N. 2005. "Global Regulation of X Chromosomal Genes by the MSL Complex in *Drosophila Melanogaster*." *Genes & Development* 19 (19): 2289–94.
doi:10.1101/gad.1343705.
- Hamada, Fumika N, Peter J Park, Polina R Gordadze, and Mitzi I Kuroda. 2005. "Global Regulation of X Chromosomal Genes by the MSL Complex in *Drosophila Melanogaster*." *Genes and Development* 19 (19): 2289–94.
doi:10.1101/gad.1343705.
- Harrison, Melissa M., and Michael B. Eisen. 2015. "Transcriptional Activation of the Zygotic Genome in *Drosophila*." In *Current Topics in Developmental Biology*, 113:85–112. doi:10.1016/bs.ctdb.2015.07.028.
- Hauser, B A, J Q He, S O Park, and C S Gasser. 2000. "TSO1 Is a Novel Protein That Modulates Cytokinesis and Cell Expansion in *Arabidopsis*." *Development (Cambridge, England)* 127 (10): 2219–26.
- Heinz, Sven, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, and Christopher K. Glass. 2010. "Simple Combinations of Lineage-Determining Transcription Factors Prime Cis-Regulatory Elements Required for Macrophage and B Cell Identities." *Molecular Cell* 38 (4): 576–89. doi:10.1016/j.molcel.2010.05.004.
- Henriques, Telmo, Daniel A. Gilchrist, Sergei Nechaev, Michael Bern, Ginger W. Muse, Adam Burkholder, David C. Fargo, et al. 2013. "Stable Pausing by RNA Polymerase II Provides an Opportunity to Target and Integrate Regulatory Signals." *Molecular Cell* 52 (4). Elsevier: 517–28. doi:10.1016/j.molcel.2013.10.001.
- Hnisz, D., A. S. Weintraub, D. S. Day, A.-L. Valton, R. O. Bak, C. H. Li, J. Goldmann, et

- al. 2016. “Activation of Proto-Oncogenes by Disruption of Chromosome Neighborhoods.” *Science* 351 (6280): 1454–58. doi:10.1126/science.aad9024.
- Hume, M. A., L. A. Barrera, S. S. Gisselbrecht, and M. L. Bulyk. 2015. “UniPROBE, Update 2015: New Tools and Content for the Online Database of Protein-Binding Microarray Data on Protein-DNA Interactions.” *Nucleic Acids Research* 43 (D1): D117–22. doi:10.1093/nar/gku1045.
- Izzo, Annalisa, Catherine Regnard, Violette Morales, Elisabeth Kremmer, and Peter B Becker. 2008. “Structure-Function Analysis of the RNA Helicase Maleless.” *Nucleic Acids Research* 36 (3): 950–62. doi:10.1093/nar/gkm1108.
- Jenuwein, Thomas, Stephen Rea, Frank Eisenhaber, Dónal O’Carroll, Brian D. Strahl, Zu-Wen Sun, Manfred Schmid, et al. 2000. “Regulation of Chromatin Structure by Site-Specific Histone H3 Methyltransferases.” *Nature* 406 (6796): 593–99. doi:10.1038/35020506.
- Jonkers, Iris, Hojoong Kwak, and John T Lis. 2014. “Genome-Wide Dynamics of Pol II Elongation and Its Interplay with Promoter Proximal Pausing, Chromatin, and Exons.” *eLife* 2014 (3). doi:10.7554/eLife.02407.
- Jonkers, Iris, and John T. Lis. 2015. “Getting up to Speed with Transcription Elongation by RNA Polymerase II.” *Nature Reviews Molecular Cell Biology* 16 (3). Nature Publishing Group: 167–77. doi:10.1038/nrm3953.
- Joshi, Sonal S., and Victoria H. Meller. 2017. “Satellite Repeats Identify X Chromatin for Dosage Compensation in *Drosophila Melanogaster* Males.” *Current Biology* 27 (10): 1393–1402.e2. doi:10.1016/j.cub.2017.03.078.
- Judd, B H. 1955. “Direct Proof of a Variegated-Type Position Effect at the White Locus in *Drosophila Melanogaster*.” *Genetics* 40 (5): 739–44.
- Kageyama, Y., G. Mengus, G. Gilfillan, H. G. Kennedy, C. Stuckenholtz, R. L. Kelley, P. B. Becker, and M. I. Kuroda. 2001. “Association and Spreading of the *Drosophila*

- Dosage Compensation Complex from a Discrete roX1 Chromatin Entry Site.”
EMBO Journal 20 (9): 2236–45. doi:10.1093/emboj/20.9.2236.
- Kamar, Ramsey I, Edward J Banigan, Aykut Erbas, Rebecca D Giuntoli, Monica Olvera de la Cruz, Reid C Johnson, and John F Marko. 2017. “Facilitated Dissociation of Transcription Factors from Single DNA Binding Sites.” *Proceedings of the National Academy of Sciences* 114 (16): E3251–57. doi:10.1073/pnas.1701884114.
- Katokhin, A V, A V Pindyurin, E V Fedorova, and E M Baricheva. 2001. “Molecular Genetic Analysis of the *Drosophila Melanogaster* Trithorax-like Gene Coding for the GAGA Transcription Factor” 37 (4): 368–74.
- Kaye, Emily G, Amina Kurbidaeva, Daniel Wolle, Tsutomu Aoki, Paul Schedl, and Erica N. Larschan. 2017. “*Drosophila* Dosage Compensation Loci Associate with a Boundary Forming Insulator Complex.” *Molecular and Cellular Biology*, August. American Society for Microbiology, MCB.00253-17. doi:10.1128/MCB.00253-17.
- Kelleher, Raymond J., Peter M. Flanagan, and Roger D. Kornberg. 1990. “A Novel Mediator between Activator Proteins and the RNA Polymerase II Transcription Apparatus.” *Cell* 61 (7): 1209–15. doi:10.1016/0092-8674(90)90685-8.
- Keller, Claudia Isabelle, and Asifa Akhtar. 2015. “The MSL Complex: Juggling RNA-Protein Interactions for Dosage Compensation and beyond.” *Current Opinion in Genetics and Development*. doi:10.1016/j.gde.2015.03.007.
- Kelley, R L, V H Meller, P R Gordadze, G Roman, R L Davis, and M I Kuroda. 1999. “Epigenetic Spreading of the *Drosophila* Dosage Compensation Complex from roX RNA Genes into Flanking Chromatin.” *Cell* 98 (4): 513–22.
- Kelley, Richard L, Victoria H Meller, Polina R Gordadze, Gregg Roman, Ronald L Davis, and Mitzi I Kuroda. 1999. “Epigenetic Spreading of the *Drosophila* Dosage Compensation Complex from roX RNA Genes into Flanking Chromatin.” *Cell* 98 (4): 513–22. doi:10.1016/S0092-8674(00)81979-0.

- Kellum, R, and P Schedl. 1991. "A Position-Effect Assay for Boundaries of Higher Order Chromosomal Domains." *Cell* 64 (5): 941–50.
- . 1992. "A Group of Scs Elements Function as Domain Boundaries in an Enhancer-Blocking Assay." *Molecular and Cellular Biology* 12 (5): 2424–31. doi:10.1128/MCB.12.5.2424.Updated.
- Kharchenko, Peter V, Artyom A Alekseyenko, Yuri B Schwartz, Aki Minoda, Nicole C Riddle, Jason Ernst, Peter J Sabo, et al. 2011. "Comprehensive Analysis of the Chromatin Landscape in *Drosophila Melanogaster*." *Nature* 471 (7339). Nature Publishing Group: 480–85. doi:10.1038/nature09725.
- Kharchenko, Peter V, Michael Y Tolstorukov, and Peter J Park. 2008. "Design and Analysis of ChIP-Seq Experiments for DNA-Binding Proteins." *Nature Biotechnology* 26 (12): 1351–59. doi:10.1038/nbt.1508.
- Kind, Jop, and Asifa Akhtar. 2007. "Cotranscriptional Recruitment of the Dosage Compensation Complex to X-Linked Target Genes." *Genes & Development* 21 (16): 2030–40. doi:10.1101/gad.430807.
- Kind, Jop, Juan M Vaquerizas, Philipp Gebhardt, Marc Gentzel, Nicholas M Luscombe, Paul Bertone, and Asifa Akhtar. 2008. "Genome-Wide Analysis Reveals MOF as a Key Regulator of Dosage Compensation and Gene Expression in *Drosophila*." *Cell* 133 (5): 813–28. doi:10.1016/j.cell.2008.04.036.
- Kingston, Robert E, and Geeta J Narlikar. 1999. "ATP-Dependent Remodeling and Acetylation as Regulators of Chromatin Fluidity." *Genes & Development* 13 (18): 2339–52.
- Klug, Aaron. 2010. "The Discovery of Zinc Fingers and Their Applications in Gene Regulation and Genome Manipulation." *Annual Review of Biochemistry* 79 (December 2009): 213–31. doi:10.1146/annurev-biochem-010909-095056.
- Koya, S. Kiran, and Victoria H. Meller. 2011. "roX RNAs and Genome Regulation in

- Drosophila Melanogaster*.” In *Progress in Molecular and Subcellular Biology*, 51:147–60. doi:10.1007/978-3-642-16502-3_7.
- Kranz, Anna-Lena, Chen-Yu Jiao, Lara Winterkorn, Sarah Albritton, Maxwell Kramer, and Sevinç Ercan. 2013. “Genome-Wide Analysis of Condensin Binding in *Caenorhabditis Elegans*.” *Genome Biology* 14 (10): R112. doi:10.1186/gb-2013-14-10-r112.
- Krebs, J E, M H Kuo, C D Allis, and C L Peterson. 1999. “Cell Cycle-Regulated Histone Acetylation Required for Expression of the Yeast HO Gene.” *Genes & Development* 13 (11). Cold Spring Harbor Laboratory Press: 1412–21.
- Krumm, A, L B Hickey, and M Groudine. 1995. “Promoter-Proximal Pausing of RNA Polymerase II Defines a General Rate-Limiting Step after Transcription Initiation.” *Genes & Development* 9 (5): 559–72. doi:10.1101/gad.9.5.559.
- Kuzu, Guray, Emily G. Kaye, Jessica Chery, Trevor Siggers, Lin Yang, Jason R. Dobson, Sonia Boor, et al. 2016. “Expansion of GA Dinucleotide Repeats Increases the Density of CLAMP Binding Sites on the X-Chromosome to Promote *Drosophila* Dosage Compensation.” *PLoS Genetics* 12 (7): 1–28. doi:10.1371/journal.pgen.1006120.
- Kwak, Hojoong, and John T. Lis. 2013. “Control of Transcriptional Elongation.” *Annual Review of Genetics* 47 (1). Annual Reviews : 483–508. doi:10.1146/annurev-genet-110711-155440.
- Kyrchanova, Olga, Darya Chetverina, Oksana Maksimenko, Andrey Kullyev, and Pavel Georgiev. 2008. “Orientation-Dependent Interaction between *Drosophila* Insulators Is a Property of This Class of Regulatory Elements.” *Nucleic Acids Research* 36 (22): 7019–28. doi:10.1093/nar/gkn781.
- Kyrchanova, Olga, Vladic Mogila, Daniel Wolle, Jose Paolo Magbanua, Robert White, Pavel Georgiev, and Paul Schedl. 2015. “The Boundary Paradox in the Bithorax

- Complex.” *Mechanisms of Development* 138: 122–32.
doi:10.1016/j.mod.2015.07.002.
- Larracuenta, Amanda M., Timothy B. Sackton, Anthony J. Greenberg, Alex Wong, Nadia D. Singh, David Sturgill, Yu Zhang, Brian Oliver, and Andrew G. Clark. 2008. “Evolution of Protein-Coding Genes in *Drosophila*.” *Trends in Genetics* 24 (3): 114–23. doi:10.1016/j.tig.2007.12.001.
- Larschan, Erica, Artyom A Alekseyenko, Andrey A Gortchakov, Shouyong Peng, Bing Li, Pok Yang, Jerry L Workman, Peter J Park, and Mitzi I Kuroda. 2007a. “MSL Complex Is Attracted to Genes Marked by H3K36 Trimethylation Using a Sequence-Independent Mechanism.” *Molecular Cell* 28 (1): 121–33.
doi:10.1016/j.molcel.2007.08.011.
- Larschan, Erica, Artyom A Alekseyenko, Andrey A Gortchakov, Shouyong Peng, Bing Li, Pok Yang, Jerry L Workman, Peter J Park, and Mitzi I Kuroda. 2007b. “MSL Complex Is Attracted to Genes Marked by H3K36 Trimethylation Using a Sequence-Independent Mechanism.” *Molecular Cell* 28 (1): 121–33.
doi:10.1016/j.molcel.2007.08.011.
- Larschan, Erica, Eric P Bishop, Peter V Kharchenko, Leighton J Core, John T Lis, Peter J Park, and Mitzi I Kuroda. 2011. “X Chromosome Dosage Compensation via Enhanced Transcriptional Elongation in *Drosophila*.” *Nature* 471 (7336). Nature Publishing Group: 115–18. doi:10.1038/nature09757.
- Larschan, Erica, Marcela M L Soruco, Ok-Kyung Lee, Shouyong Peng, Eric Bishop, Jessica Chery, Karen Goebel, Jessica Feng, Peter J Park, and Mitzi I Kuroda. 2012. “Identification of Chromatin-Associated Regulators of MSL Complex Targeting in *Drosophila* Dosage Compensation.” *PLoS Genetics* 8 (7): e1002830.
doi:10.1371/journal.pgen.1002830.
- Lee, Chanhyo, X. Li, Aaron Hechmer, Michael Eisen, Mark D Biggin, Bryan J Venters,

- Cizhong Jiang, Jian Li, B Franklin Pugh, and David S Gilmour. 2008. "NELF and GAGA Factor Are Linked to Promoter-Proximal Pausing at Many Genes in *Drosophila*." *Molecular and Cellular Biology* 28 (10): 3290–3300.
doi:10.1128/MCB.02224-07.
- Leiserson, Mark D M, Dima Blokh, Roded Sharan, and Benjamin J Raphael. 2013. "Simultaneous Identification of Multiple Driver Pathways in Cancer." *PLoS Computational Biology* 9 (5): e1003054. doi:10.1371/journal.pcbi.1003054.
- Leon, O, and M Roth. 2000. "Zinc Fingers: DNA Binding and Protein-Protein Interactions." *Biological Research* 33 (1): 21–30.
- Levo, Michal, Einat Zalckvar, Eilon Sharon, Ana Carolina Dantas Machado, Yael Kalma, Maya Lotam-Pompan, Adina Weinberger, Zohar Yakhini, Remo Rohs, and Eran Segal. 2015. "Unraveling Determinants of Transcription Factor Binding Outside the Core Binding Site." *Genome Research* 25 (7): 1018–29.
doi:10.1101/gr.185033.114.
- Li, Jian, Yingyun Liu, HoSung Rhee, Saikat Kumar B Ghosh, Lu Bai, B. Franklin Pugh, and David S. Gilmour. 2013. "Kinetic Competition between Elongation Rate and Binding of NELF Controls Promoter-Proximal Pausing." *Molecular Cell* 50 (5). Elsevier Inc.: 711–22. doi:10.1016/j.molcel.2013.05.016.
- Liang, Hsiao-Lan, Chung-Yi Nien, Hsiao-Yun Liu, Mark M Metzstein, Nikolai Kirov, and Christine Rushlow. 2008. "The Zinc-Finger Protein Zelda Is a Key Activator of the Early Zygotic Genome in *Drosophila*." *Nature* 456 (7220): 400–403.
doi:10.1038/nature07388.
- Lieberman-Aiden, Erez, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. 2009. "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome." *Science* 326 (5950).

- Liew, Chu Kong, Roland Gamsjaeger, Robyn E. Mansfield, and Joel P. Mackay. 2008. "NMR Spectroscopy as a Tool for the Rapid Assessment of the Conformation of GST-Fusion Proteins." *Protein Science* 17 (9): 1630–35. doi:10.1110/ps.034983.108.
- Lobanenkov, V V, R H Nicolas, V V Adler, H Paterson, E M Klenova, A V Polotskaja, and G H Goodwin. 1990. "A Novel Sequence-Specific DNA Binding Protein Which Interacts with Three Regularly Spaced Direct Repeats of the CCCTC-Motif in the 5'-flanking Sequence of the Chicken c-Myc Gene." *Oncogene* 5 (12): 1743–53.
- Lomaev, Dmitry, Anna Mikhailova, Maksim Erokhin, Alexander V. Shaposhnikov, James J. Moresco, Tatiana Blokhina, Daniel Wolle, et al. 2017. "The GAGA Factor Regulatory Network: Identification of GAGA Factor Associated Proteins." *PLoS ONE* 12 (3): 1–20. doi:10.1371/journal.pone.0173602.
- Lorch, Y, B R Cairns, M Zhang, and R D Kornberg. 1998. "Activated RSC-Nucleosome Complex and Persistently Altered Form of the Nucleosome." *Cell* 94 (1): 29–34.
- Lorch, Y, M Zhang, and R D Kornberg. 1999. "Histone Octamer Transfer by a Chromatin-Remodeling Complex." *Cell* 96 (3): 389–92.
- Los, Georgyi V, and Keith Wood. 2007. "The HaloTag: A Novel Technology for Cell Imaging and Protein Analysis." *Methods in Molecular Biology (Clifton, N.J.)* 356: 195–208.
- Lowenhaupt, K, A Rich, and M L Pardue. 1989. "Nonrandom Distribution of Long Mono- and Dinucleotide Repeats in Drosophila Chromosomes: Correlations with Dosage Compensation, Heterochromatin, and Recombination." *Molecular and Cellular Biology* 9 (3): 1173–82.
- Lucchesi, John C, and Mitzi I Kuroda. 2015. "Dosage Compensation in Drosophila." *Cold Spring Harbor Perspectives in Biology* 7 (5): a019398. doi:10.1101/cshperspect.a019398.

- Lue, N F, and R D Kornberg. 1987. "Accurate Initiation at RNA Polymerase II Promoters in Extracts from *Saccharomyces Cerevisiae*." *Proceedings of the National Academy of Sciences of the United States of America* 84 (24): 8839–43.
- Lyon, Mary F. 1961. "Gene Action in the X-Chromosome of the Mouse (*Mus Musculus* L.)." *Nature* 190 (4773): 372–73. doi:10.1038/190372a0.
- Machanick, P., and T. L. Bailey. 2011. "MEME-ChIP: Motif Analysis of Large DNA Datasets." *Bioinformatics* 27 (12): 1696–97. doi:10.1093/bioinformatics/btr189.
- Matsui, T, J Segall, P A Weil, and R G Roeder. 1980. "Multiple Factors Required for Accurate Initiation of Transcription by Purified RNA Polymerase II." *The Journal of Biological Chemistry* 255 (24): 11992–96.
- McElroy, Kyle A, Hyuckjoon Kang, and Mitzi I Kuroda. 2014. "Are We There yet? Initial Targeting of the Male-Specific Lethal and Polycomb Group Chromatin Complexes in *Drosophila*." *Open Biology* 4 (140006). doi:10.1098/rsob.140006.
- Meller, V H, and M I Kuroda. 2002. "Sex and the Single Chromosome." *Adv Genet* 46: 1–24.
- Meller, V H, and B P Rattner. 2002. "The roX Genes Encode Redundant Male-Specific Lethal Transcripts Required for Targeting of the MSL Complex." *Embo J* 21 (5): 1084–91.
- Menon, Debashish U., Cristian Coarfa, Weimin Xiao, Preethi H. Gunaratne, and Victoria H. Meller. 2014. "siRNAs from an X-Linked Satellite Repeat Promote X-Chromosome Recognition in *Drosophila Melanogaster*." *Proceedings of the National Academy of Sciences* 111 (46): 16460–65. doi:10.1073/pnas.1410534111.
- Meyer, B J, and L P Casson. 1986. "Caenorhabditis Elegans Compensates for the Difference in X Chromosome Dosage between the Sexes by Regulating Transcript Levels." *Cell* 47 (6): 871–81.
- Michelitsch, M D, and J S Weissman. 2000. "A Census of Glutamine/asparagine-Rich

- Regions: Implications for Their Conserved Function and the Prediction of Novel Prions.” *Proceedings of the National Academy of Sciences of the United States of America* 97 (22): 11910–15. doi:10.1073/pnas.97.22.11910.
- Muller, Martin, Kirsten Hagstrom, Henrik Gyrkovics, Vincenzo Pirrotta, and Paul Schedl. 1999. “The Mcp Element from the *Drosophila Melanogaster Bithorax* Complex Mediates Long-Distance Regulatory Interactions.” *Genetics* 153 (3): 1333–56.
- Muravyova, E, A Golovnin, E Gracheva, A Parshikov, T Belenkaya, V Pirrotta, and P Georgiev. 2001. “Loss of Insulator Activity by Paired Su(Hw) Chromatin Insulators.” *Science* 291 (5503): 495–98. doi:10.1126/science.291.5503.495.
- Nakajima, Toshihiro, Chiharu Uchida, Stephen F. Anderson, Chee-Gun Lee, Jerard Hurwitz, Jeffrey D. Parvin, and Marc Montminy. 1997. “RNA Helicase A Mediates Association of CBP with RNA Polymerase II.” *Cell* 90 (6): 1107–12. doi:10.1016/S0092-8674(00)80376-1.
- Nechaev, Sergei, David C Fargo, Gilberto dos Santos, Liwen Liu, Yuan Gao, and Karen Adelman. 2010. “Global Analysis of Short RNAs Reveals Widespread Promoter-Proximal Stalling and Arrest of Pol II in *Drosophila*.” *Science (New York, N.Y.)* 327 (5963). NIH Public Access: 335–38. doi:10.1126/science.1181421.
- Nguyen, Di Kim, and Christine M Disteché. 2006. “Dosage Compensation of the Active X Chromosome in Mammals.” *Nature Genetics* 38 (1): 47–53. doi:10.1038/ng1705.
- Nora, Elphège P., Bryan R. Lajoie, Edda G. Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, et al. 2012. “Spatial Partitioning of the Regulatory Landscape of the X-Inactivation Centre.” *Nature* 485 (7398): 381–85. doi:10.1038/nature11049.
- Oh, H, Y Park, and M I Kuroda. 2003. “Local Spreading of MSL Complexes from roX Genes on the *Drosophila* X Chromosome.” *Genes Dev* 17 (11): 1334–39.

- Ohno, Susumu. 1967. "The Basic Difference in Constitution between the Mammalian X and the Drosophila X." In *Sex Chromosomes and Sex-Linked Genes*, 1:82–91. Monographs on Endocrinology. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-88178-7.
- Ohtsuki, Sumio, and Michael Levine. 1998. "GAGA Mediates the Enhancer Blocking Activity of the Eve Promoter in the Drosophila Embryo." *Genes and Development* 12 (21): 3325–30. doi:10.1101/gad.12.21.3325.
- Omichinski, JG, and P Pedone. 1997. "The Solution Structure of a Specific GAGA Factor-DNA Complex Reveals a Modular Binding Mode." *Nature Structural & Molecular Biology* 4 (2).
- Owen-Hughes, T, and J L Workman. 1996. "Remodeling the Chromatin Structure of a Nucleosome Array by Transcription Factor-Targeted Trans-Displacement of Histones." *The EMBO Journal* 15 (17). European Molecular Biology Organization: 4702–12.
- Paris, Mathilde, Tommy Kaplan, Xiao Yong Li, Jacqueline E. Villalta, Susan E. Lott, and Michael B. Eisen. 2013. "Extensive Divergence of Transcription Factor Binding in Drosophila Embryos with Highly Conserved Gene Expression." Edited by Patricia Wittkopp. *PLoS Genetics* 9 (9): e1003748. doi:10.1371/journal.pgen.1003748.
- Park, Y, R L Kelley, H Oh, M I Kuroda, and V H Meller. 2002. "Extent of Chromatin Spreading Determined by roX RNA Recruitment of MSL Proteins." *Science* 298 (5598): 1620–23.
- Paro, R, and D S Hogness. 1991. "The Polycomb Protein Shares a Homologous Domain with a Heterochromatin-Associated Protein of Drosophila." *Proceedings of the National Academy of Sciences of the United States of America* 88 (1): 263–67.
- Patikoglou, G A, J L Kim, L Sun, S H Yang, T Kodadek, and S K Burley. 1999. "TATA Element Recognition by the TATA Box-Binding Protein Has Been Conserved

- throughout Evolution.” *Genes & Development* 13 (24). Cold Spring Harbor Laboratory Press: 3217–30.
- Peterlin, B Matija, and David H Price. 2006. “Controlling the Elongation Phase of Transcription with P-TEFb.” *Molecular Cell* 23 (3): 297–305.
doi:10.1016/j.molcel.2006.06.014.
- Peti, Wolfgang, and Rebecca Page. 2007. “Strategies to Maximize Heterologous Protein Expression in Escherichia Coli with Minimal Cost.” *Protein Expression and Purification* 51 (1): 1–10. doi:10.1016/j.pep.2006.06.024.
- Pfister, Stefan, Stephen Rea, Mikko Taipale, Frank Mendrzyk, Beate Straub, Carina Ittrich, Olaf Thuerigen, Hans Peter Sinn, Asifa Akhtar, and Peter Lichter. 2008. “The Histone Acetyltransferase hMOF Is Frequently Downregulated in Primary Breast Carcinoma and Medulloblastoma and Constitutes a Biomarker for Clinical Outcome in Medulloblastoma.” *International Journal of Cancer. Journal International Du Cancer* 122 (6): 1207–13. doi:10.1002/ijc.23283.
- Proudfoot, N J, and G G Brownlee. 1974. “Nucleotide Sequence Adjacent to Polyadenylic Acid in Globulin Messenger RNA.” *FEBS Letters* 38 (2): 179–83.
- Quinn, Jeffrey J, Qiangfeng C Zhang, Plamen Georgiev, Ibrahim A Ilik, Asifa Akhtar, and Howard Y Chang. 2016. “Rapid Evolutionary Turnover Underlies Conserved lncRNA-Genome Interactions.” *Genes & Development* 30 (2): 191–207.
doi:10.1101/gad.272187.115.
- Rachidi, Mohammed, and Carmela Lopes. 2007. “Mental Retardation in Down Syndrome: From Gene Dosage Imbalance to Molecular and Cellular Mechanisms.” *Neuroscience Research* 59 (4): 349–69. doi:10.1016/j.neures.2007.08.007.
- Rahl, Peter B, Charles Y Lin, Amy C Seila, Ryan A Flynn, Scott McCuine, Christopher B Burge, Phillip A Sharp, and Richard A Young. 2010. “C-Myc Regulates Transcriptional Pause Release.” *Cell* 141 (3): 432–45.

doi:10.1016/j.cell.2010.03.030.

Ramírez, Fidel, Thomas Lingg, Sarah Toscano, Kin Chung Chung Lam, Plamen

Georgiev, Ho-Ryun Ryun Chung, Bryan R. Lajoie, et al. 2015. “High-Affinity Sites Form an Interaction Network to Facilitate Spreading of the MSL Complex across the X Chromosome in *Drosophila*.” *Molecular Cell* 60 (1): 146–62.

doi:10.1016/j.molcel.2015.08.024.

Rieder, Leila E., Kaitlin P. Koreski, Kara A. Boltz, Guray Kuzu, Jennifer A. Urban, Sarah K. Bowman, A. Zeideman, et al. 2017. “Histone Locus Regulation by the *Drosophila* Dosage Compensation Adaptor Protein CLAMP.”

Rieder, Leila E, Kaitlin P Koreski, Kara A Boltz, Guray Kuzu, Jennifer A Urban, Sarah Bowman, Anna Zeidman, et al. 2017. “Histone Locus Regulation by the *Drosophila* Dosage Compensation Adaptor Protein CLAMP.” *Genes & Development*, August. Cold Spring Harbor Laboratory Press. doi:10.1101/gad.300855.117.

Rodin, Sergey, Olga Kyrchanova, Ekaterina Pomerantseva, Alexander Parshikov, and Pavel Georgiev. 2007. “New Properties of *Drosophila* Fab-7 Insulator.” *Genetics* 177 (1): 113–21. doi:10.1534/genetics.107.075887.

Rodriguez, M. A., D. Vermaak, J. J. Bayes, and H. S. Malik. 2007. “Species-Specific Positive Selection of the Male-Specific Lethal Complex That Participates in Dosage Compensation in *Drosophila*.” *Proceedings of the National Academy of Sciences* 104 (39): 15412–17. doi:10.1073/pnas.0707445104.

Rougvié, A E, and J T Lis. 1988. “The RNA Polymerase II Molecule at the 5' End of the Uninduced hsp70 Gene of *D. Melanogaster* Is Transcriptionally Engaged.” *Cell* 54 (6): 795–804.

Schnitzler, G, S Sif, and R E Kingston. 1998. “Human SWI/SNF Interconverts a Nucleosome between Its Base State and a Stable Remodeled State.” *Cell* 94 (1): 17–27.

- Schweinsberg, Susan E, and Paul Schedl. 2004. "Developmental Modulation of Fab-7 Boundary Function." *Development* 131 (19): 4743–49. doi:10.1242/dev.01343.
- Schweinsberg, Susan, Kirsten Hagstrom, Daryl Gohl, Paul Schedl, Ram P Kumar, Rakesh Mishra, and Francois Karch. 2004. "The Enhancer-Blocking Activity of the Fab-7 Boundary From the Drosophila Bithorax Complex Requires GAGA-Factor-Binding Sites" 1384 (November): 1371–84. doi:10.1534/genetics.104.029561.
- Sedivy, John M, Gowrishankar Banumathy, and Peter D Adams. 2008. "Aging by Epigenetics--a Consequence of Chromatin Damage?" *Experimental Cell Research* 314 (9): 1909–17. doi:10.1016/j.yexcr.2008.02.023.
- Segal, David J., Justin W. Crotty, Mital S. Bhakta, Carlos F. Barbas, and Nancy C. Horton. 2006. "Structure of Aart, a Designed Six-Finger Zinc Finger Peptide, Bound to DNA." *Journal of Molecular Biology* 363 (2): 405–21. doi:10.1016/j.jmb.2006.08.016.
- Shogren-Knaak, M., Haruhiko Ishii, Jian-Min Sun, Michael J Pazin, James R Davie, and Craig L Peterson. 2006. "Histone H4-K16 Acetylation Controls Chromatin Structure and Protein Interactions." *Science* 311 (5762): 844–47. doi:10.1126/science.1124000.
- Sievers, F., A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, et al. 2014. "Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega." *Molecular Systems Biology* 7 (1): 539–539. doi:10.1038/msb.2011.75.
- Siggers, Trevor, Michael H Duyzend, Jessica Reddy, Sidra Khan, and Martha L Bulyk. 2011. "Non-DNA-Binding Cofactors Enhance DNA-Binding Specificity of a Transcriptional Regulatory Complex." *Molecular Systems Biology* 7 (555). Nature Publishing Group: 555. doi:10.1038/msb.2011.89.
- Siggers, Trevor, and Raluca Gordân. 2014. "Protein-DNA Binding: Complexities and

- Multi-Protein Codes." *Nucleic Acids Research* 42 (4): 2099–2111.
doi:10.1093/nar/gkt1112.
- Siggers, Trevor, Jessica Reddy, Brian Barron, and Martha L. Bulyk. 2014.
"Diversification of Transcription Factor Paralogs via Noncanonical Modularity in
C2H2 Zinc Finger DNA Binding." *Molecular Cell* 55 (4): 640–48.
doi:10.1016/j.molcel.2014.06.019.
- Sigrist, Christian J A, and Vincenzo Pirrotta. 1997. "Chromatin Insulator Elements Block
the Silencing of a Target Gene by the Drosophila Polycomb Response Element
(PRE) but Allow Trans Interactions between PREs on Different Chromosomes."
Genetics 147 (1): 209–21.
- Simonis, Marieke, Petra Klous, Erik Splinter, Yuri Moshkin, Rob Willemsen, Elzo de Wit,
Bas van Steensel, and Wouter de Laat. 2006. "Nuclear Organization of Active and
Inactive Chromatin Domains Uncovered by Chromosome Conformation Capture–
on-Chip (4C)." *Nature Genetics* 38 (11): 1348–54. doi:10.1038/ng1896.
- Slattery, Matthew, Tianyin Zhou, Lin Yang, Ana Carolina Dantas Machado, Raluca
Gordân, and Remo Rohs. 2014. "Absence of a Simple Code: How Transcription
Factors Read the Genome." *Trends in Biochemical Sciences* 39 (9): 381–99.
doi:10.1016/j.tibs.2014.07.002.
- Smith, E R, C D Allis, and J C Lucchesi. 2001. "Linking Global Histone Acetylation to the
Transcription Enhancement of X-Chromosomal Genes in Drosophila Males." *J Biol
Chem* 276 (34): 31483–86.
- Soruco, Marcela M L, Jessica Chery, Eric P Bishop, Trevor Siggers, Michael Y
Tolstorukov, Alexander R Leydon, Arthur U Sugden, et al. 2013. "The CLAMP
Protein Links the MSL Complex to the X Chromosome during Drosophila Dosage
Compensation." *Genes & Development* 27 (14): 1551–56.
doi:10.1101/gad.214585.113.

- Stampfel, Gerald, Tomáš Kazmar, Olga Frank, Sebastian Wienerroither, Franziska Reiter, and Alexander Stark. 2015. "Transcriptional Regulators Form Diverse Groups with Context-Dependent Regulatory Functions." *Nature*. doi:10.1038/nature15545.
- Staudt, Nicole, Sonja Fellert, Ho-Ryun Chung, Herbert Jäckle, and Gerd Vorbrüggen. 2006. "Mutations of the Drosophila Zinc Finger-Encoding Gene *Vielfältig* Impair Mitotic Cell Divisions and Cause Improper Chromosome Segregation." *Molecular Biology of the Cell* 17 (5): 2356–65. doi:10.1091/mbc.E05-11-1056.
- Straub, Tobias, Charlotte Grimaud, Gregor D Gilfillan, Angelika Mitterweger, and Peter B Becker. 2008. "The Chromosomal High-Affinity Binding Sites for the Drosophila Dosage Compensation Complex." *PLoS Genetics* 4 (12): e1000302. doi:10.1371/journal.pgen.1000302.
- Strutt, H., G Cavalli, and R Paro. 1997. "Co-Localization of Polycomb Protein and GAGA Factor on Regulatory Elements Responsible for the Maintenance of Homeotic Gene Expression." *The EMBO Journal* 16 (12): 3621–32. doi:10.1093/emboj/16.12.3621.
- Sun, Zu-Wen, and C. David Allis. 2002. "Ubiquitination of Histone H2B Regulates H3 Methylation and Gene Silencing in Yeast." *Nature* 418 (6893). Nature Publishing Group: 104–8. doi:10.1038/nature00883.
- Sural, Tuba H, Shouyong Peng, Bing Li, Jerry L Workman, Peter J Park, and Mitzi I Kuroda. 2008. "The MSL3 Chromodomain Directs a Key Targeting Step for Dosage Compensation of the Drosophila Melanogaster X Chromosome." *Nature Structural & Molecular Biology* 15 (12): 1318–25. doi:10.1038/nsmb.1520.
- Teichmann, Sarah Amalia, and Reiner Albert Veitia. 2004. "Genes Encoding Subunits of Stable Complexes Are Clustered on the Yeast Chromosomes: An Interpretation from a Dosage Balance Perspective." *Genetics*. Genetics Society of America. doi:10.1534/genetics.103.024505.

- Tsukiyama, T, and C Wu. 1996. "Purification of GAGA Factor of *Drosophila* and Its Role in Nucleosome Disruption." *Methods in Enzymology* 274: 291–99.
- Tsukiyama, Toshio, Peter B. Becker, and Carl Wu. 1994. "ATP-Dependent Nucleosome Disruption at a Heat-Shock Promoter Mediated by Binding of GAGA Transcription Factor." *Nature* 367 (6463): 525–32. doi:10.1038/367525a0.
- Tsukiyama, Toshio, and Carl Wu. 1995. "Purification and Properties of an ATP-Dependent Nucleosome Remodeling Factor." *Cell* 83 (6): 1011–20. doi:10.1016/0092-8674(95)90216-3.
- Urban, J, C Doherty, L Rieder, W Jordan, M Tsiarli, and E Larschan. 2016. "The Essential *Drosophila* CLAMP Protein Differentially Regulates Non-Coding roX RNAs in Male and Females." *Chromosome Res.*
- Urban, Jennifer A., Caroline A. Doherty, William T. Jordan, Jacob E. Bliss, Jessica Feng, Marcela M. Soruco, Leila E. Rieder, Maria A. Tsiarli, and Erica N. Larschan. 2017. "The Essential *Drosophila* CLAMP Protein Differentially Regulates Non-Coding roX RNAs in Male and Females." *Chromosome Research* 25 (2). Springer Netherlands: 101–13. doi:10.1007/s10577-016-9541-9.
- Urban, Jennifer A., John Urban, and Erica N. Larschan. 2017. "The *Drosophila* CLAMP Protein Associates with Diverse Sex-Specific and Non-Sex Specific Proteins on Chromatin."
- Urban, Jennifer A, Guray Kuzu, Sarah K Bowman, Benjamin Scruggs, Telmo Henriques, Robert E Kingston, Karen Adelman, Michael Y. Tolstorukov, and Erica N. Larschan. 2017. "Enhanced Chromatin Accessibility of the Dosage Compensated *Drosophila* Male X-Chromosome Requires the CLAMP Zinc-Finger Protein."
- Vazquez, J, and P Schedl. 1994. "Sequences Required for Enhancer Blocking Activity of Scs Are Located within Two Nuclease-Hypersensitive Regions." *The EMBO Journal* 13 (24): 5984–93.

- Vicoso, Beatriz, and Doris Bachtrog. 2015. "Numerous Transitions of Sex Chromosomes in Diptera." *PLoS Biology* 13 (4): 1–22. doi:10.1371/journal.pbio.1002078.
- Villa, Raffaella, Ignasi Forné, Marisa Müller, Axel Imhof, Tobias Straub, and Peter B Becker. 2012. "MSL2 Combines Sensor and Effector Functions in Homeostatic Control of the Drosophila Dosage Compensation Machinery." *Molecular Cell* 48 (4): 647–54. doi:10.1016/j.molcel.2012.09.012.
- Villa, Raffaella, Tamas Schauer, Pawel Smialowski, Tobias Straub, and Peter B Becker. 2016. "PionX Sites Mark the X Chromosome for Dosage Compensation." *Nature* 537 (7619). Nature Publishing Group: 244–48. doi:10.1038/nature19338.
- Villar, Diego, Camille Berthelot, Paul Flicek, and Duncan T Odom Correspondence. 2015. "Enhancer Evolution across 20 Mammalian Species." *Cell* 160: 554–66. doi:10.1016/j.cell.2015.01.006.
- Wang, Charlotte I, Artyom a Alekseyenko, Gary LeRoy, Andrew E H Elia, Andrey a Gorchakov, Laura-Mae P Britton, Stephen J Elledge, Peter V Kharchenko, Benjamin a Garcia, and Mitzi I Kuroda. 2013. "Chromatin Proteins Captured by ChIP-Mass Spectrometry Are Linked to Dosage Compensation in Drosophila." *Nature Structural & Molecular Biology* 20 (2). Nature Publishing Group: 202–9. doi:10.1038/nsmb.2477.
- Wang, Hengbin, Liangjun Wang, Hediye Erdjument-Bromage, Miguel Vidal, Paul Tempst, Richard S Jones, and Yi Zhang. 2004. "Role of Histone H2A Ubiquitination in Polycomb Silencing." *Nature* 431 (7010): 873–78. doi:10.1038/nature02985.
- Weinstein, D C, A Ruiz i Altaba, W S Chen, P Hoodless, V R Prezioso, T M Jessell, and J E Darnell. 1994. "The Winged-Helix Transcription Factor HNF-3 Beta Is Required for Notochord Development in the Mouse Embryo." *Cell* 78 (4): 575–88.
- Weirauch, Matthew T., Ally Yang, Mihai Albu, Atina G. Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S. Najafabadi, et al. 2014. "Determination and

- Inference of Eukaryotic Transcription Factor Sequence Specificity." *Cell* 158 (6): 1431–43. doi:10.1016/j.cell.2014.08.009.
- Wilkins, R C, and J T Lis. 1999. "DNA Distortion and Multimerization: Novel Functions of the Glutamine-Rich Domain of GAGA Factor." *Journal of Molecular Biology* 285 (2): 515–25. doi:10.1006/jmbi.1998.2356.
- Wolle, Daniel, Fabienne Cleard, Tsutomu Aoki, Girish Deshpande, Paul Schedl, and Francois Karch. 2015. "Functional Requirements for Fab-7 Boundary Activity in the Bithorax Complex." *Molecular and Cellular Biology* 35 (21): 3739–52. doi:10.1128/MCB.00456-15.Address.
- Xiong, Yuanyan, Xiaoshu Chen, Zhidong Chen, Xunzhang Wang, Suhua Shi, Xueqin Wang, Jianzhi Zhang, and Xionglei He. 2010. "RNA Sequencing Shows No Dosage Compensation of the Active X-Chromosome." *Nature Genetics* 42 (12): 1043–47. doi:10.1038/ng.711.
- Yanez-Cuna, J. O., H. Q. Dinh, E. Z. Kvon, D. Shlyueva, and A. Stark. 2012. "Uncovering Cis-Regulatory Sequence Requirements for Context-Specific Transcription Factor Binding." *Genome Research* 22 (10): 2018–30. doi:10.1101/gr.132811.111.
- Yang, Zhong-Fa, Stephanie Mott, and Alan G. Rosmarin. 2007. "The Ets Transcription Factor GABP Is Required for Cell-Cycle Progression." *Nature Cell Biology* 9 (3): 339–46. doi:10.1038/ncb1548.
- Zaret, K S, and F Sherman. 1982. "DNA Sequence Required for Efficient Transcription Termination in Yeast." *Cell* 28 (3): 563–73.
- Zentner, Gabriel E., Sivakanthan Kasinathan, Beibei Xin, Remo Rohs, and Steven Henikoff. 2015. "ChEC-Seq Kinetics Discriminates Transcription Factor Binding Sites by DNA Sequence and Shape in Vivo." *Nature Communications* 6 (October): 8733. doi:10.1038/ncomms9733.

- Zhang, Yong, Tao Liu, Clifford A Meyer, Jérôme Eeckhoutte, David S Johnson, Bradley E Bernstein, Chad Nussbaum, et al. 2008. "Model-Based Analysis of ChIP-Seq (MACS)." *Genome Biology* 9 (9): R137. doi:10.1186/gb-2008-9-9-r137.
- Zhao, Zhihu, Gholamreza Tavoosidana, Mikael Sjölander, Anita Göndör, Piero Mariano, Sha Wang, Chandrasekhar Kanduri, et al. 2006. "Circular Chromosome Conformation Capture (4C) Uncovers Extensive Networks of Epigenetically Regulated Intra- and Interchromosomal Interactions." *Nature Genetics* 38 (11): 1341–47. doi:10.1038/ng1891.
- Zheng, Sanduo, Raffaella Villa, Jia Wang, Sanduo Zheng, Raffaella Villa, Jia Wang, Yingang Feng, Jinfeng Wang, and Peter B Becker. 2014. "Structural Basis of X Chromosome DNA Recognition by the MSL2 CXC Domain during *Drosophila* Dosage Compensation." *Genes and Development* 28 (23): 2652–62. doi:10.1101/gad.250936.114.
- Zhou, Jumin, Scott Barolo, Paul Szymanski, and Michael Levine. 1996. "The Fab-7 Element of the Bithorax Complex Attenuates Enhancer-Promoter Interactions in the *Drosophila* Embryo." *Genes and Development* 10 (24): 3195–3201. doi:10.1101/gad.10.24.3195.
- Zhou, Qi, Christopher E. Ellison, Vera B. Kaiser, Artyom A. Alekseyenko, Andrey A. Gorchakov, and Doris Bachtrog. 2013. "The Epigenome of Evolving *Drosophila* Neo-Sex Chromosomes: Dosage Compensation and Heterochromatin Formation." Edited by Peter B. Becker. *PLoS Biology* 11 (11): e1001711. doi:10.1371/journal.pbio.1001711.
- Zhou, T., L. Yang, Y. Lu, I. Dror, A. C. Dantas Machado, T. Ghane, R. Di Felice, and R. Rohs. 2013. "DNAshape: A Method for the High-Throughput Prediction of DNA Structural Features on a Genomic Scale." *Nucleic Acids Research* 41 (W1): W56–62. doi:10.1093/nar/gkt437.

- Zhou, Tianyin, Ning Shen, Lin Yang, Namiko Abe, John Horton, Richard S. Mann, Harmen J. Bussemaker, Raluca Gordân, and Remo Rohs. 2015. "Quantitative Modeling of Transcription Factor Binding Specificities Using DNA Shape." *Proceedings of the National Academy of Sciences* 112 (15): 4654–59. doi:10.1073/pnas.1422023112.
- Zollman, Susan, Dorothea Godtt, Gilbert G Priveo, Jean-Louis Couderc, Frank A Laski, and Dan L Lindsley. 1994. "The BTB Domain, Found Primarily in Zinc Finger Proteins, Defines an Evolutionarily Conserved Family That Includes Several Developmentally Regulated Genes in *Drosophila* (Bric A Brac/tramtrack/transcriptional Regulation/amlino-Terminal Domain/enbryogenesis)." *Developmental Biology* 91: 10717–21.