Distribution Agreement

Signature: DocuSigned by: Lauren J Hodkinson
A7E8F126F2BD45A...

Lauren J Hodkinson
Name

4/2/2024 | 2:00 PM EDT
Date

Title Nuances for context-dependent transcription factor function

Author Lauren J Hodkinson

Degree Doctor of Philosophy

Program Biological and Biomedical Sciences
Genetics and Molecular Biology

Approved by the Committee

DocuSigned by:
Leila Rieder
1DAA2763F1FE4CF...

Leila Rieder
Advisor

DocuSigned by:
Roger Deal
E48B09A892AD4A7...

Roger Deal
Committee Member

DocuSigned by:
William Kelly
72474D887F3E4C6...

William Kelly
Committee Member

DocuSigned by:
Dorothy Lerit
ABBB6D9F999243A...

Dorothy Lerit
Committee Member

DocuSigned by:
Kenneth Moberg
6E049122B7D6435...

Kenneth Moberg
Committee Member

Committee Member

Accepted by the Laney Graduate School:

Kimberly Jacob Arriola, Ph.D, MPH
Dean, James T. Laney Graduate School

Date

Nuances of context-dependent transcription factor function

By

Lauren J. Hodkinson
B.S., Ithaca College, 2018
B.A., Ithaca College 2018

Advisor: Leila E. Rieder, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in the Graduate Division of Biological and Biomedical Science
Genetics and Molecular Biology
2024

Abstract

Nuances of context-dependent transcription factor function

By Lauren J. Hodkinson

Despite binding similar *cis* elements, transcription factors often perform context-dependent functions at different genomic loci. Furthermore, transcription factors can be involved in large scale coordinated gene events where hundreds of genes need to be targeted and regulated with strict temporal or special requirements. How transcription factors integrate *cis* sequence and genomic context to perform their context-dependent functions is still poorly understood. One example of a context-dependent transcription factor is involved in coordinated gene regulation the *Drosophila* protein Chromatin-Linked Adapter for MSL Proteins (CLAMP), which targets similar GA-rich *cis* elements on the X-chromosome and at the histone locus but recruits very different, locus-specific transcription factors to each of these contexts. We utilized several different techniques to interrogate CLAMP function at the histone locus as a context-dependent transcription factor. First, we focused on understanding how differences in the CLAMP-binding GA-repeat element within the in the promoter of *H3* and *H4* (*H3/H4*p) may impact the overall regulation of histone genes. We found that the *H3/H4*p GA-repeat is variable across the ~100 histone gene arrays however this sequence variation, and subsequent factor targeting, likely does not confer differential expression of the histone genes. Next, we investigated how CLAMP function at the histone genes is impacted by the identity of its *cis* binding elements. Leveraging a powerful transgenic histone array system, we discovered that X-linked CLAMP sequences do not functionally substitute for GA-repeats in the histone gene array. Our results suggest that transcription factors incorporate *cis* sequences and flanking sequence to govern their regulatory function at target loci. Finally, we explored what additional DNA factors may be regulating histone gene expression. Through our undergraduate driven *in silico* screen, we identified 9 novel histone locus regulatory factor candidates that warrant future wet lab studies to interrogate how they may influence histone biogenesis. Combined, these findings broaden our understanding of the nuanced mechanisms of coordinated histone gene regulation.

Acknowledgments

First and foremost, I want to thank the Rieder Lab. I have been fortunate to be surrounded by the most intelligent and encouraging lab members. I would like to recognize each of you for every moment of support and guidance you have given me throughout my PhD journey; (in no specific order) Dr. Casey Schmidt, Dr. Skye Comstra, Tommy O'Haren, Gwyn Puckett, Mellisa Xie, Eric Albanese, Annalise Weber, Greg Kimmerer, Edgar Hsish, Mary Wang, Dabin Cho, John Ali, Henrik Torres, Pamela Diaz-Saldana, Hannah Gilbonio, Nicole Roos, Sisi Falcone, and Amirah Hurst. I also indebted to my mentor Dr. Leila Rieder who I will recognize again, in more detail, at the end of my acknowledgments.

I want to especially thank Dr. Casey Schmidt for her dedication to my own growth as a scientist, as well as a person. Casey, thank you for all of your unwavering support through my science, experiment troubleshooting, paper writing and personal crises. It was a privilege to have you as a post-doctoral mentor and, thinking back on all of our CURE teaching endeavors and constant coffee chats, I'm quite sure I would not have made it through my PhD without you. I apologize for always poking fun at you but know I always do it in jest because of how much I care about you. I didn't just get a post-doc advisor; I got a lifelong friend. Also, to Dr. Skye Comstra, thank you for making me feel the most welcome when I first joined the Rieder Lab. You are majestical (reference to the film *Hunt for the Wilderpeople*). I will always be grateful for becoming instant best friends. I will always cherish yelling about *RuPaul's Drag Race* while simultaneously learning about embryo collection protocols with you.

I want to truly thank each member of my committee for your invaluable feedback, engagement in my project, and encouragement that has led to me success. Dr. Ken Moberg, Dr. Bill Kelly, Dr. Roger Deal and especially Dr. Dorothy Lerit: you all have been an invaluable resource to me for the past five years and I have often told people that I won the lottery for best thesis committee. Beyond my committee at Emory, I want to thank my undergraduate mentor Dr. Susan Witherup, who was monumental in helping me develop my foundation as a young scientist. I would also like to sincerely thank Dr. Te-Wen Lo who is yet another lifechanging mentor I have had in my science career. Te-Wen, you have always been my beam of support and your constant encouragement is what made me confident enough in myself to pursue a PhD in genetics.

I was lucky enough to acquire not one, but two amazing lab families during my PhD experience. Prior to joining the Rieder Lab, I was a member of the Hickman Lab studying host-pathogen yeast genetics. To my core Hickwomen, Dr. Ognenka Avramovska, Dr. Amanda Smith, (soon to be Dr.) Judy Dinh, and Dr. Rema Elmostafa, you have all taught me so much about myself, my confidence, and my abilities as a scientist. From our lab gym days lifting weight I never thought I could, to our midday coffee runs when we were waiting to count our yeast or *C. elegans*, you all are my lifelong lab-family, and I am so grateful to have each of you in my life. Finally, thank you to Dr. Meleah Hickman for taking me on as a student and guiding me through my first years of graduate school. Meleah, you are an inspiration to me not only in my science but in my communication and in my commitment to always staying true to my values. I cannot express how lucky I am to call you my PhD mentor and I couldn’t imagine what my experience at Emory would’ve been like without being in your lab.

I also want to thank my GMB cohort for all the support and comradery we developed through the years. My graduate school experience was made better by each of you. I want to especially thank Dr. Kimberly Diaz-Perez, the first true friend I made in my GMB cohort. Kim, you are one of the most incredible friends I have ever had. Thank you for every study session, movie night, and Atlanta United game we experienced together. You are and will always be one of my closest friends and I am forever grateful to have you in my life. I also need to thank (soon to be Dr.) Samantha Lanjewar from the bottom of my heart. Sam, being roommates for two years through an entire global pandemic was honestly one of the best things that happened to me in graduate school. Thank you for always being willing to take a three mile walk with me to talk about every woe and source of stress we have in our lives. I know our friendship is going to be a lifelong one and I don't think there are words to describe what your support has meant to me. Finally, thank you to the next cohort of GMB students who matriculated in 2019 for adopting me and especially to Emily Hill, Tommy O'Haren, Yemko Pryor, Keenan Wiggins and Jim Rose. You are all amazing individuals and have given me even more lifelong GMB friendships, as well as some extraordinary DnD (Dungeons & Dragons) campaign members. Finally, thank you also to every kind and inspirational member of the Sloan Lab for adopting me into a third lab family.

My time in the GMB program and at Emory introduced me to many other individuals that I now call close friends and who were invaluable in my graduate school experience. I have made countless connections in each of the cohorts that matriculated before and after my own. Although I cannot mention you all by name, I am grateful for the encouragement and advice you all have provided me. To my older and wiser GMB alumni Stephanie Grewenow, MS and Dr. Kari Mattison, I am unimaginably appreciative to you for all of you support and guidance during my time at Emory. You both often say that you forced me into friendship during my first GMB retreat, but I am forever grateful you both wanted to hang out with a brand new, GMB first-year who wanted to make some new friends.

My journey through graduate school was undoubtably changed by meeting my current partner and best friend Emily Hill. Emily, having you as my person through almost the entirety of this wild experience has been a true gift and I am indebted to you for all that you have done for me. Thank you for the hours you have spent listening to me practicing my talks, pontificating about my science, and venting about all of my lab woes. Thank you for always giving me something to smile about, even on my worst days. From every floor picnic to every video game night, you have always been there to bring me out of my science slumps and bring some much-needed light back into my days. You have given so much love, support, and friendship since the moment we met, and I don't think I could ever truly verbalize how meaningful your presence in my life is to me. Also, thank you for giving me the best step-cat in the entire word. Hutch, I want to thank you for always giving me the best kitten cuddles and thank you for always knowing when I need them most.

I owe the sincerest thank you to my two best friends from Ithaca, NY who have given me nothing but encouragement and support throughout my entire PhD. JC Alexander (Jackie) and Dr. Dallas Fonseca, I don't know how I would live without you. You both are my pillars of support. Thank you for both standing by me through all of the trials and tribulations of graduate school. Moving to opposite parts of the country from you both was one of the hardest parts of choosing to come to Emory. I cannot express how grateful I am for all the pandemic zoom calls where you both stayed up until 1:00 am with me practicing for my qual. Our trips to remote cabins in Washington truly saved me and my sanity, and I will always cherish the memories from those escapes. I never could say thank you enough for your friendship and I will always consider you both as my chosen family.

Most importantly, I want to thank my family for being pillars of support not only during my PhD but my entire life. To my parents, I have no words for how truly grateful I am to have you both as my support system. You both have always encouraged me to follow my dreams and have never wavered in your support for my education. I couldn't have asked for a better set of parents. Mom, thank you for always taking my calls and listening to me vent about every challenge I encountered through my PhD. Dad, thank you for always asking questions about my science and wanting to spend hours talking about the latest CRISPR therapies. I would also like to thank my godmother Claudia, who has never stopped encouraging me to reach for the stars. Finally, to my oldest friend and chosen sister Katie Shea, I want to thank you with all my heart for standing by me through (basically) my entire life and I am so grateful that no matter how much time we spend apart, we always pick up right where we left off.

Lastly, I would like to write a special thank you to my advisor Dr. Leila Rieder. The origin story of how we became a mentor-mentee pair started the moment I got to Emory, and I think I only have fate to thank for that. I participated in a two-week course that ran prior to my first day as a graduate student at Emory. There I met Julia Gross, a recent graduate from Brown, who sparked a conversation with me based on my introduction to the group as a "fan of histones." Julia could not wait to tell me how much I would love the brand-new faculty member coming to Emory the following year, Dr. Leila Rieder, and this proved to be a humongous understatement. Leila, I'm confident that there are no words or ways to express how grateful I am to have had you as my PhD mentor. You have given me nothing but unwavering support through some of the most challenging scientific and personal roadblocks I have ever experienced. You have changed my life in so many ways and I have grown so much as a scientist, and as a person, through your mentorship and guidance. I can never thank you enough for taking a chance on an almost third-year graduate student who needed to switch labs. I will always consider you a lifelong mentor, supporter, and friend. I voice this from the deepest part of my heart: you are appreciated by so many and so much, and by no one more than me.

Table of Contents

Chapter 1 – Introduction 19

1.1 Overview 20

1.2 Coordinated gene expression occurs on different scales in the nucleus 21

1.2.1 Coordinating zygotic genome activation 23

1.2.2 Coordinating dosage compensation 25

1.2.3 Coordinated expression of the replication dependent histone genes 27

1.3 The gene clustering and genomic organization facilitate coordinated gene expression 29

1.3.1 Nuclear bodies and three-dimensional genome architecture enable more efficient gene regulation 29

1.3.2 Organizing genes into clusters or repetitive arrays can facilitate coordinated gene expression 31

1.3.3 Combined genomic organizational strategies 32

1.4 Several properties can influence transcription factor targeting and function 33

1.4.1 Cis element sequence 35

1.4.2 Cofactors 36

1.5 Context-dependent transcription factors 38

1.5.1 CLAMP as a context-dependent transcription factor 39

1.5.2 CLAMP function at the histone locus is a model for context-dependent transcription factor function 40

1.6 Goals and major findings 41

1.7 Summary 43

1.8 References 45

Chapter 2 - *Cis* element length variability does not confer differential transcription factor occupancy at the histone locus 55

2.1 Abstract 56

2.2 Introduction 57

2.3 Results 60

2.3.1 The GA-repeat is variable in length across the histone gene arrays. 60

2.3.2 CLAMP, GAF, and Psq all target the GA-repeats in the *H3/H4*p 62

2.3.3 GA-binding factors do not show preference for GA-repeat length at the histone locus.......64
2.4 Discussion.......69
2.5 Conclusions.......72
2.6 Methods.......73
2.7 Acknowledgments.......77
2.8 Supplemental Figures.......78
2.9 References.......80

Chapter 3 Sequence reliance of a *Drosophila* context-dependent transcription factor.........87
3.1 Abstract.......88
3.2 Introduction.......89
3.3 Materials and Methods.......92
3.4 Results and Discussion.......97
3.4.1 CLAMP targets GA-rich *cis* elements at different loci.......97
3.4.2 CLAMP requires GA-rich sequences to bind *in vitro* and *in vivo*.......99
3.4.3 The GA-repeats must reside in the *H3/H4*p for Mxc recruitment *in vivo*.......103
3.4.4 CLAMP binding sequences from different loci do not functionally substitute in the context of the histone array in chromosomal females.......106
3.4.5 X-linked sequences in the context of the histone gene array attract MSL2 in chromosomal males.......110
3.5 Acknowledgements.......114
3.6 Supplemental Figures.......115
3.7 References.......116

Chapter 4 - MSL2 targets histone genes in *Drosophila virilis*.......122
4.1 Abstract.......123
4.2 Introduction.......124
4.3 Methods.......128
4.4 Results.......133
4.4.1 Sequence differences between *D. virilis* histone loci.......133
4.4.2 Both *D. virilis* histone loci are targeted by Mxc and CLAMP.......134

4.4.3 MSL2 targets the major *D. virilis* histone locus 136

4.4.4 MSL2 does not directly interact with histone array sequence 140

4.4.5 CLAMP does not require the GA-rich elements to interact with the *virilis H3/H4* promoter *in vitro* 142

4.4.6 The *D. virilis H3/H4* promoter does not promote Mxc recruitment in *D. melanogaster* ...144

4.5 Discussion 146

4.6 Acknowledgments...152

4.7 Supplemental Figures and Tables 153

4.8 References..164

Chapter 5 - A bioinformatics screen reveals Hox and chromatin remodeling factors at the Drosophila histone locus...174

5.1 Abstract 175

5.2 Introduction...176

5.3 Methods...179

5.3.1 GEO Datasets 179

5.3.2 Bioinformatic Analysis and Data Visualization 181

5.3.3 Criteria for Positive Candidates vs. Negative Candidate..................................182

5.4 Results...183

5.4.1 Validating the bioinformatics pipeline by mapping TATA-associated factors to the histone gene array 183

5.4.2 Candidate DNA-binding factors that did not pass the bioinformatics screen.....................186

5.4.3 Candidates that passed the bioinformatics screen..192

5.4.4 Hox factors localize to the Drosophila histone gene array when overexpressed in cell culture 194

5.4.5 Power and limitations of the screen 197

5.5 Discussion 199

5.6 Conclusions...205

5.7 Acknowledgments and Authors' contributions 206

5.8 References..207

Chapter 6 - A cost-free CURE: Using bioinformatics to identify DNA-binding factors at a specific genomic locus 220

6.1 Abstract 221

6.2 Introduction 222

6.2.1 Intended audience 224

6.2.2 Learning time 225

6.2.3 Prerequisite student knowledge 225

6.2.4 Learning objectives 225

6.3 Procedures 226

6.3.1 Materials 226

6.3.2 Student instruction 228

6.3.3 Faculty Instruction 228

6.3.3.1 Background 228

6.3.3.2 Tutorials 230

6.3.3.3 Work days 231

6.3.3.4 Poster session 232

6.3.3.5 Notes and recommendations 233

6.4 Discussion 237

6.4.1 Field testing 237

6.4.2 Evidence of student learning 238

6.4.3 Possible modifications 241

6.5 Summary 242

6.6 Acknowledgments 243

6.7 References 244

Chapter 7 – Discussion 253

7.1 Overview: Coordinated gene regulation is a heavy burden 254

7.2 The Drosophila melanogaster histone locus is inherently absurd 254

7.3 The impact of cis element length on transcription factor recruitment 257

7.4 A combination of cues can influence transcription factor function 260

7.5 Regulatory mechanisms of histone gene expression vary between species 263

7.6 The HLB is a wild unknown entity....................266
7.7 Histone gene regulation is way more complicated than you think....................268
7.8 The future of studying coordinated histone gene regulation: closing remarks....................272
7.9 References....................274

Appendix A – Supplemental Data of Chapter 5....................279

Figures and Tables

Chapter 1

Figure 1.1 An overview of coordinated gene regulation events in *D. melanogaster*

Figure 1.2 Zygotic genome activation and nuclear cycles in early *D. melanogaster* development

Figure 1.3 The cell cycle and histone gene expression

Figure 1.4 Hallmarks of transcription factor function

Chapter 2

Figure 2.1 The GA-repeat length is variable across the ~100 histone gene arrays

Figure 2.2 GAF, CLAMP, and Psq target the *H3/H4*p GA-repeat

Figure 2.3 The H3/H4 promoter GA-repeat length variability is observed in different datasets

Figure 2.4 GAF and CLAMP target the same length GA-repeats

Table 2.1 ChIP-sequencing datasets

Table 2.2 ChIP-sequencing data used to generate GA-repeat length histograms

Supplemental (S)

SFigure 2.1 The GA-repeat can contain a variety of mismatches and sequence variation

SFigure 2.2 Sample .tsv file output for the GA-repeat counting bioinformatics pipeline

Chapter 3

Figure 3.1 CLAMP binds locations genome-wide while Mxc and MSL3 bind distinct genomic regions

Table 3.1 Sequence of EMSA probes

Figure 3.2 GA-repeats are required for CLAMP binding and function at the transgenic histone gene array

Figure 3.3 GA-repeats must reside in the H3/H4p for proper CLAMP function at the transgenic histone gene array

Figure 3.4 X-linked elements do not functionally substitute GA-repeats in the histone gene array in chromosomal females

Figure 3.5 X-linked elements attract MSL2 in the context of the transgenic histone gene array in chromosomal males

Supplemental (S)

SFigure 3.1 Supershift demonstrating that CLAMP is present in late nuclear extract (LNE)

Chapter 4

Figure 4.1 CLAMP and Mxc target the *D. virilis* histone loci

Figure 4.2 MSL2 targets only one histone locus in *D. virilis*

Figure 4.3 MSL3 does not target the histone locus in *Drosophila*

Figure 4.4 MSL2 does not target histone loci in other *Drosophila* species

Figure 4.5 CLAMP binds the *D. virilis H3/H4* sequence *in vitro* and does not require the GA-rich elements

Figure 4.6 The *D. virilis H3/H4* promoter does not promote HLB formation in *D. melanogaster*

Supplemental (S)

STable 4.1 Summary of MSL protein localization from polytene chromosome immunostaining experiments, related to Figures 4.2 and 4.4

SFigure 4.1 Ectopically expressed MSL2 from either species does not target the *D. melanogaster* histone gene array

STable 4.2 dsDNA EMSA probes, related to Figures 4.5 and S4.7

SFigure 4.2 Anti-GFP ChIP-seq from non-transfected cells gives a peak in the *H2a/H2b* promoter

STable 4.3 *H3/H4* promoter sequences from *D. melanogaster* histone array transgenes., related to Figure 4.6

SFigure 4.3 MSL3 does not target the *D. melanogaster* histone array after MSL2 transfection

SFigure 4.4 H4K16ac is not enriched over the *D. melanogaster* histone array during male *D. melanogaster* embryogenesis

SFigure 4.5 H4K16ac is not enriched over the *D. melanogaster* histone array after MSL2 transfection

SFigure 4.6 CLAMP interacts with the region containing the poorly conserved GA-rich cis elements in the *D. virilis* promoter

SFigure 4.7 Conservation of the *H3/H4* promoter within and between *D. virilis* loci

Chapter 5

Figure 5.1 Known HLB factor CLAMP localizes to the GA-repeat cis element in the *H3/H4* promoter

Table 5.1 DNA Binding Factor Candidates

Figure 5.2 Expected general transcription factors localize to the histone array

Figure 5.3 DNA-binding factors from different categories that did not pass the bioinformatics screen

Figure 5.4 JIL-1, Hr78, and Fs(1)hL localize to the histone gene array

Figure 5.5 Hox factors Ubx, Abd-A, and Abd-B localize to the histone array

Figure 5.6 Ubx localizes to the *H3/H4* promoter in embryos and 3rd instar larva

Figure 5.7 ChIP-seq datasets from different tissues can show different alignment results

Chapter 6

Figure 6.1 Weekly class schedule for the CURE

Figure 6.2 Summary of tools used in Galaxy

Figure 6.3 Example student poster

Figure 6.4 Sample data

Figure 6.5 The primary form of summative assessment for this CURE was the students' posters

Figure 6.6 CURE-related exam questions

Chapter 7

Figure 7.1 A diagram of the *D. melanogaster* histone locus and HLB

Figure 7.2 Model depicting how CLAMP incorporates information from *cis* element sequence and flanking DNA sequence to determine function at the histone array

Figure 7.3 ChIP alignment possibilities

Figure 7.4 Summary of HLB unknowns

Appendix A

Supplemental (S)

SFigure 5.1 Qualitative assessment for scoring candidates as positive or negative

SFigure 5.2 Factors considered negative hits

Chapter 1

Introduction

Chapter 1: Introduction

1.1 Overview

Coordinated gene expression, where genes are expressed in spatial and temporal synchrony, is a crucial but difficult task in the crowded nucleus. To accomplish this feat, transcription factors must traverse the nucleus to find their corresponding *cis* elements. Furthermore, once factors have identified their DNA-binding sites, they impact gene expression on strict temporal and spatial levels. Transcription factors can serve different functions at different loci and may rely on a variety of informational cues from a variety of places to determine how they will function. A large gap in our current understanding of transcription factor function is how the same transcription factor can bind similar looking *cis* DNA elements throughout the genome but function in a completely unique way at these different loci. The nuances of what specific pieces of information or what combination of cues transcription factors incorporate to make sure they function uniquely across the genome is still poorly understood.

In this introduction, I explore several processes of coordinated gene regulation and the role that transcription factors play in ensuring tight regulation of these gene expression events at a variety of scales. I then delve into the intricacies of how the finite number of transcription factors we know of are able to identify and, in many cases, uniquely regulate the thousands of genes within the genome. The examples of coordinated gene expression and transcription factor function I outline span all domains of life, with information from human systems all the way to bacterial systems. Here, I have leveraged the powerful model system *Drosophila melanogaster* to interrogate the nuances of transcription factor function in my dissertation with the aim of gaining a more comprehensive understanding of how transcription factors can function uniquely across the genome.

1.2 Coordinated gene expression occurs on different scales in the nucleus

Coordinated gene regulation, simply, is a phenomenon where a set of genes are orchestrated to be expressed together often under some temporal, spatial, or concentration requirements. Coordinated gene regulation events are extremely broad, varying in scale of orchestration and severity of requirements (Michalak 2008; Nair *et al.* 2022). Large scale coordination can involve hundreds of genes that need to be expressed or repressed at the exact same time, such as during a specific developmental period, or in precise amounts, such as when transcribing the components of a protein complex.

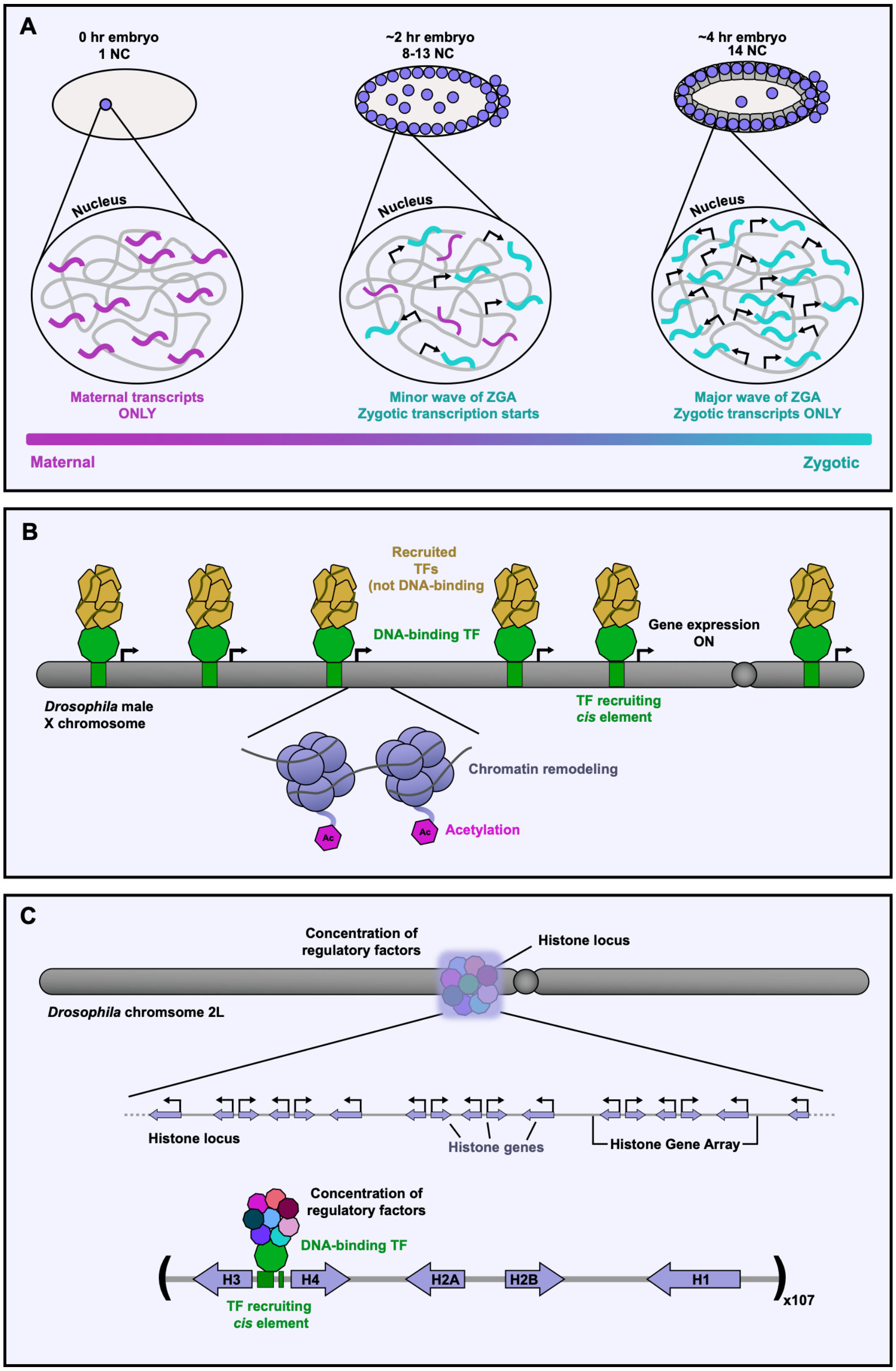


Figure 1.1 An overview of coordinated gene regulation events in *D. melanogaster*. (A) Zygotic genome activation is a large scale coordinated gene regulatory process where, in the first 4 hours of development, maternal transcriptions are consumed/degraded, and the zygotic genome

begins transcribed in two waves. **(B)** Dosage compensation is another coordinated regulatory event where a protein complex is recruited to the X chromosome where it is marked so that transcription can be upregulated to correct for the imbalance of X gene dosage in males. **(C)** Transcription of the replication dependent histone genes is coordinated by a concentration of factors that regulate tandemly repeated histone gene arrays at the histone locus

1.2.1 Coordinating zygotic genome activation

During early animal development, there are several instances of large-scale events where many genes are coordinated to be transcribed around the same time. Zygotic genome activation, or ZGA, is an example of a global coordinated gene regulation event in the early embryo. ZGA occurs in the early stages of all metazoan embryos and occurs in two waves, the minor wave and major wave, which are synchronized with the nuclear cycles of the dividing embryo. ZGA is the second process in the maternal to zygotic transition (MZT) when zygotic transcription is initiated after maternally deposited RNAs and proteins have been degraded or consumed by the developing embryo (Tadros and Lipshitz 2009; Farrell and O'Farrell 2014; Hamm and Harrison 2018).

In *Drosophila,* the early stages of embryo development involve a series of synchronized cell divisions where the embryo is relying on maternally deposited material to support cell functions. The *Drosophila* embryo undergoes 13 nuclear division cycles (nuclear cycles, NC) followed by cellularization after the 14th nuclear division. Therefore, while the degradation of maternally deposited material is occurring over the course of these 14 nuclear cycles, the zygotic genome needs to initiate transcription. The minor wave of ZGA begins around NC 8 and the major wave at NC14, to ensure the developing embryo has access to all mRNAs and proteins it needs (**Figure 1.1A,1.2)** (Hamm and Harrison 2018).

Recent mRNA labeling experiments revealed that 946 genes out of the ~13,600 (Adams 2000, genome sequence) present in the *Drosophila* genome are activated during the minor wave of ZGA and that this number increases to 3588 by the end of the major wave (Kwasnieski *et al.* 2019). Activating ~3600 genes is a large feat for the genome to accomplish, yet ZGA precisely coordinated process occurring over the course of ~4 hours where transcription factors are not only identifying these 3600 gene targets but are initiating their transcription in conjunction loss of maternal material (Harrison and Eisen 2015; Hamm and Harrison 2018).

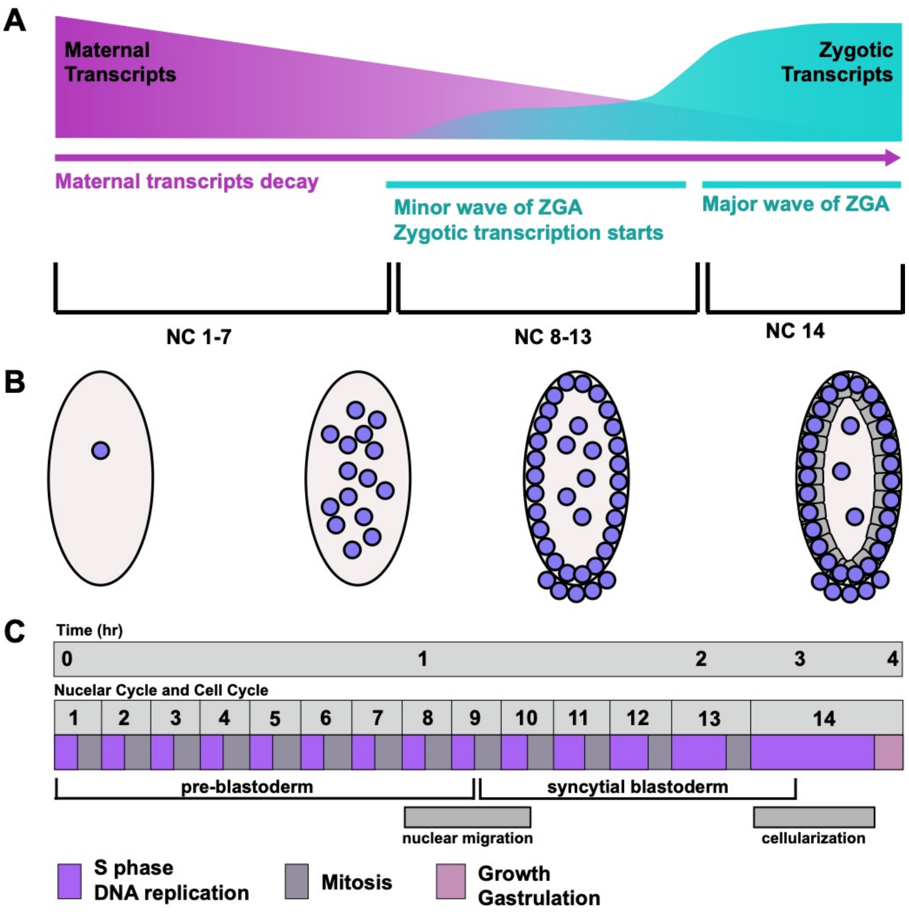


Figure 1.2 Zygotic genome activation and nuclear cycles in early *D. melanogaster* development. **(A)** In the first 14 nuclear cycle divisions of the *Drosophila* embryo, maternal transcripts decay while two waves of zygotic transcription occur to transition from relying on maternal material to relying on zygotic material. **(B)** The nuclei in the syncytial embryo divide in the center of the embryo and subsequently migrate to the periphery to form the blastoderm. Pole cells form around nuclear cycle 11 and cellularization occurs in after the 14th nuclear division. **(C)** The embryo undergoes 14 nuclear divisions, consisting of only DNA replication in S phase then successive mitotic division until full cellularization. The embryo then enters gastrulation where the cell cycle becomes complete and further development continues. This figure was modeled after and recreated from (Tadros and Lipshitz 2009) and (Farrell and O'Farrell 2014).

1.2.2 Coordinating dosage compensation

ZGA is a large-scale example of coordinated gene regulation, however, it is not the only process in the early embryo in which a group of genes needs to be co-regulated. Dosage compensation, the process of correcting the imbalance of sex chromosome gene dosage (Duan and Larschan 2019), is another example coordinated gene regulation .

Dosage compensation is accomplished through distinctive mechanisms in different species but shares the goal of modulating the output of genes linked to the sex chromosomes. In humans, XX individuals have one of their two X chromosomes almost completely silenced, with the exception of a handful of "escaper regions." Human dosage compensation is a well-orchestrated process in which the long non-coding (lncRNA) *Xist* is transcribed from the X chromosomes that is destined to be silenced, and then subsequently coats that X chromosome. Besides the escaper regions which is mechanistically poorly understood, *Xist* simply coats the chromosome it was transcribed from in *cis* and recruits histone modifiers to remove "active" histone marks, replacing them with heterochromatic "inactive" histone marks and DNA methylation (Li *et al.* 2022). Two *Xist* molecules initially concentrate the modifying factors at each of ~50 loci and then complex together with additional factors to form nucleate supramolecular complexes. These complexes subsequently create a spreading gradient of proteins to proximal regions to accomplish full silencing across the entire "inactive" X chromosome (Markaki et al 2021, cell). Organizationally, this "spreading" is made more efficient by the fact that *Xist* and the other factors can spread in *cis* across a single chromosome rather than having to orchestrate these processes by targeting many different loci across the genome.

In contrast to humans, in *Drosophila,* the single male X chromosome is upregulated to compensate for the imbalance of X chromosome gene dosage. A group of five proteins, MSL1, MSL2, MSL3, MOF MLE, and a lncRNA, either *roX1* or *roX2* which are functionally redundant, complex together to form the dosage compensation complex known as the MSL complex (male specific lethal complex, MSLc). MSLc has no formal DNA-binding members but rather targets the X chromosome by associating with the DNA-binding factor CLAMP (chromatin linked adaptor for MSL proteins) (Soruco and Larschan 2014). From the initial CLAMP target sites, MSL complex then spreads across the chromosome assisting in opening chromatin allowing MOF, an acetyltransferase, to deposit H4K16ac to ensure gene expression is upregulated (**Figure 1.1B)**. Similar to humans, because the MSL complex needs to modulate almost every gene on the X chromosome, CLAMP along with the rest of the complex, simply identifies the X chromosome and then is thought to spread in *cis* to the surrounding genes without needing other cues or biases to make sure it is targeting the proper genes (Lucchesi and Kuroda 2015; Ramírez *et al.* 2015). Dosage compensation in both humans and flies exhibits the importance of linear gene clustering and how this organization facilitates effective coordination of hundreds of genes. While humans and *Drosophila* differ in how they achieve dosage compensation, they share the fact that this regulatory event is temporally coordinated to occur in early development, coupled with the challenge of coordinating nearly every gene of an entire chromosome.

1.2.3 Coordinated expression of the replication dependent histone genes

While ZGA and dosage compensation represent large-scale coordinated gene regulation, slightly smaller coordination events involve the co-regulation of gene families or genes that comprise complexes. The histone genes represent both of these, as they often exist as multiples in the genome and make up nucleosomes.

Nucleosomes are critical components of the genome; each nucleosome is comprised of eight positively charged histone proteins and can then associate with a linker histone protein if needed, which function as structural units for negatively charged DNA to wrap around and establish essential genome organization. Many researchers focus on the importance of post-translational modifications to histone tails for the regulation of gene expression. Although this is obviously an exciting area of study, the regulation of histone gene transcription is a hallmark example of coordinated gene regulation. Excitingly, histone gene regulation features several interesting areas of research that are currently understudied or less fully understood than the other examples of coordinated gene regulation mentioned above.

Because of the unique and strict composition requirements of nucleosome structure, histone genes need to be expressed in a dose-dependent manner: the correct concentrations of transcripts and, later, protein need to be synthesized for nucleosome formation to maintain histone homeostasis, where there are neither too few nor too many histones (Chaubal *et al.* 2023). Having even small deviations from the proper concentrations of histone can have severely detrimental consequences (Gunjan and Verreault 2003; McKay *et al.* 2015; Jimeno-González *et al.* 2015; Maya Miles *et al.* 2018; Chari *et al.* 2019). Histone transcripts also have strict processing requirements because they have no introns and, rather than a polyA tail, they form a secondary stem-loop structure that needs to be cleaved for proper nuclear export and before translation (Marzluff *et al.* 2008; Tatomer *et al.* 2016).

Furthermore, histone gene expression itself is coupled to the cell cycle. Histone gene expression is rapidly increases during S phase during genome replication and ceases directly after the completion of S phase, where all remaining histone mRNAs are rapidly degraded in G2 **(Figure 1.3)**. Therefore, the expression of the histone genes not only needs to be coordinated at a strict temporal level based on the cell cycle but also needs to ensure the correct amount of each histone transcript, and subsequent protein, are synthesized to maintain histone homeostasis (Chaubal *et al.* 2023).

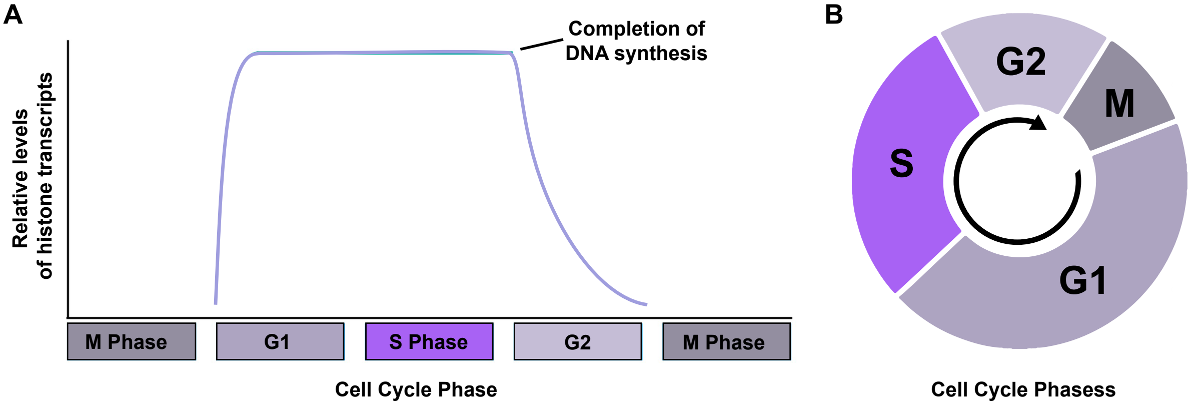


Figure 1.3 The cell cycle and histone gene expression. Expression of the replication dependent histone genes is coupled to the cell cycle. Histone expression is upregulated to reach the highest

relative levels of transcripts during genome duplication in S phase and expression is subsequently reduced directly after the end of S phase. This figure was adapted from (Marzluff *et al.* 2008).

1.3 The gene clustering and genomic organization facilitate coordinated gene expression.

Varying types of organizational strategies can facilitate coordinated gene expression and make transcription more efficient. One genomic organizational tool is clustering genes that need to be expressed at the same time, either into one or several compressed loci or loosely clustered in the same region of a chromosome. Both strict and loose clustering allows for the transcription and regulatory machinery to concentrate in specific genomic locations rather than having to traverse the entire genome to find several separate genomic locations (Tatomer *et al.* 2016). If gene are not clustered linearly, they can also be brought together through three dimensional interactions. Three-dimensional architecture can create microenvironments for transcription hubs, such as nuclear bodies, where, again, all the necessary regulatory factors can be concentrated into a specific region of the nucleus (Carty *et al.* 2017; Ghule *et al.* 2023; Chaubal *et al.* 2023).

1.3.1 Nuclear bodies and three-dimensional genome architecture enable more efficient gene regulation

Although linear or chromosomal clustering provides an obvious spatial benefit for coordinated gene expression, for processes like ZGA where thousands of unlinked or unrelated genes need to be expressed around the same time, this type of organization is impractical. Self-assembling regulatory modules, such as nuclear bodies, and three-dimensional architecture of the genome can allow genes to be organized in the physical space of nucleus and cluster together in the same vicinity to facilitate regulation. Chromosomal territories are established within the nucleus where each chromosomes occupies a particular nuclear compartment, allowing: 1) genes located on the same chromosome are in close 3D space; and 2) genes on different chromosomes that need to be expressed together by the same machinery can be placed in close 3D space (Cremer and Cremer 2001).

Recent studies using high resolution sequencing have established a map of long-range promoter-promoter and promoter-enhancer interactions that allow distant genes to be co-expressed. In a *Drosophila* embryo study*,* researchers termed a “topological operon” as a transcriptional hub consisting of shared pools of transcriptional machinery, including RNA Polymerase II, with outcomes similar to that achieved with bacterial operons (Zhang *et al.* 2022). This “topological operon” incorporates the ability of the eukaryotic genome to organize chromatin to bring distant enhancers and promoters closer together to create these hubs of regulated transcription.

Nuclear bodies are another example of a high order organization strategy where membrane-less organelles allow genes from the same chromosome or different chromosomes to be coregulated (Mao *et al.* 2011). There are a variety of nuclear body structures in the nucleus that allow for concentration of transcription factors and cofactors allowing for aggregation or interaction (Matera *et al.* 2009). The nucleolus, for example concentrates the rDNA genes to allows for a “hub” of transcription and processing rRNAs as well as assembly of the ribosomal subunit all within the same three-dimensional space. The histone locus body is also characterized as a phase separated nuclear body where all of the histone genes and their regulatory factors can be concentrated to facilitate gene expression (Nizami *et al.* 2010; Geisler *et al.* 2023). Again, this three-dimensional organization allows for assembly of both scaffolding factors such as Mxc as well as mRNA processing factors such as FLASH and U7snRNP to concentrate at the histone locus ion *D. melanogaster* or multiple histone loci in humans (White *et al.* 2011; Duronio and Marzluff 2017). Although nuclear bodies were once considered as simple concentrated aggregations of factors, it is now understood that they can play a large role in processes like coordinated gene regulation because of how efficient transcription and processing can be within these subnuclear compartments.

1.3.2 Organizing genes into clusters or repetitive arrays can facilitate coordinated gene expression

On an even smaller scale, groups of linked genes or gene families can be spatially clustered at a single or multiple loci to facilitate regulation. Bacteria, for example, possess a signature example of how genomic organization assists with coordinated gene expression by organizing genes into units known as operons. Operons are groups of coregulated, functionally linked genes that are regulated under the same promoter and usually organized in close proximity. One of the most well-studied operons, the Lac operon, includes the genes of three co-regulated enzymes which share a promoter and are all required to ensure proper transport and metabolism of lactose (Miller and Reznikoff 1978) .

Although eukaryotes do not possess operons exactly like those present in bacteria, they employ similar types of organization such as clustering or arraying genes to facilitate synchronized expression. For example, the Hox genes are essential for body patterning and development in organisms spanning from *Drosophila* all the way up to mammals. Regardless of species, Hox genes are clustered in chromosomal arrays and this organization is distinctive as the order of the genes reflects their spatial activation in the developing embryo. Although each of the Hox genes sports its own unique promoter, being arrayed in close proximity on the chromosome allows for their tight regulation to remain efficient and effective (Pearson *et al.* 2005).

The histone genes also exemplify a set of genes that are organized into clusters and serve as an additional example of genomic organization related to coordinated gene expression. In humans, all of the replication dependent histone genes are loosely clustered together in two genomic loci on chromosome 6 and 1 (Marzluff *et al.* 2002). This clustering allows for the concentrations of factors that regulate histone gene expression to localize efficiently. In *Drosophila melanogaster,* the replication dependent histone genes exist at a single condensed locus on chromosome 2L. In *D. melanogaster,* the histone locus is comprised of approximately 100 tandemly repeated arrays, with each 5 kb array containing the five canonical histone genes (*H3, H4, H2A, H2B,* and *H1*) along with their promoters and regulatory elements (Bongartz and Schloissnig 2018), an extreme example of gene clustering that facilitates coordinated expression.

1.3.3 Combined genomic organizational strategies

Although the above organizational strategies can be stratified into distinct categories, in reality they can be applied in combination to create an ideal regulatory environment for gene groups. For example, ribosomal RNAs, important catalytic components of the protein synthesizing ribosome, are transcribed from the rDNA genes in a strict and highly coordinated manner to ensure proper ribosome biogenesis. The transcription of the rDNA is facilitated by the high copy number of the genes and DNA methylation that allows sets of adjacent gene copies or "clusters" of the genes to be silenced while methylation free clusters are readily transcribed (Hori *et al.* 2023). In humans, the repetitive rDNA clusters are spread across the acrocentric chromosomes and are then three dimensionally organized into the nucleolus, a phase separated hub devoted to rDNA transcription, ribosome assembly, and some mRNA processing to ready transcripts for translation.

The combination of clustering genes together in both two- and three-dimensions is a strategy also applied to the human histone genes. As mentioned above the human replication-dependent histone genes are clustered on two different chromosomes. Surprisingly, only one of the clusters, the major cluster on chromosome 6, includes the *H1* gene. To create the complete nucleosome structure, all of the replication dependent histone genes, including *H1*, need to be expressed together and in the correct stoichiometric amount. While all the histone genes on chromosome 6 are clustered, HiC data have confirmed that long range interactions between loosely clustered histone genes establish a "hub" of histone gene transcription where distal enhancer elements and histone genes that are megabases away are brought together in 3D space (Carty *et al.* 2017; Ghule *et al.* 2023).

1.4 Several properties can influence transcription factor targeting and function

Although a robust nuclear organizational strategy does facilitate efficient coordinated transcription, it does not explain how coordinated gene expression is functionally achieved. Transcription factors are responsible for modulating gene expression and carry out a variety of functions to achieve proper gene regulation. Some transcription factors are defined by their ability to bind specific DNA sequences or *cis* elements, while others are critical players or structural components in larger complexes. Considering the large-scale gene regulatory events where thousands of genes are being modulated like ZGA and dosage compensation, it is clear that the groups of factors that are modulating gene expression need a way to identify their targets so the correct genes are modulated respectively. This targeting is also commonly imperfect; in both humans and *Drosophila* dosage compensation, there are some regions of the X that are not compensated, often referred to as "escaper regions." The factors involved in dosage compensation therefore need to be highly specific to the regions they are targeting or, more interestingly, not targeting. This points to an obvious question: how do transcription factors properly identify their targets?

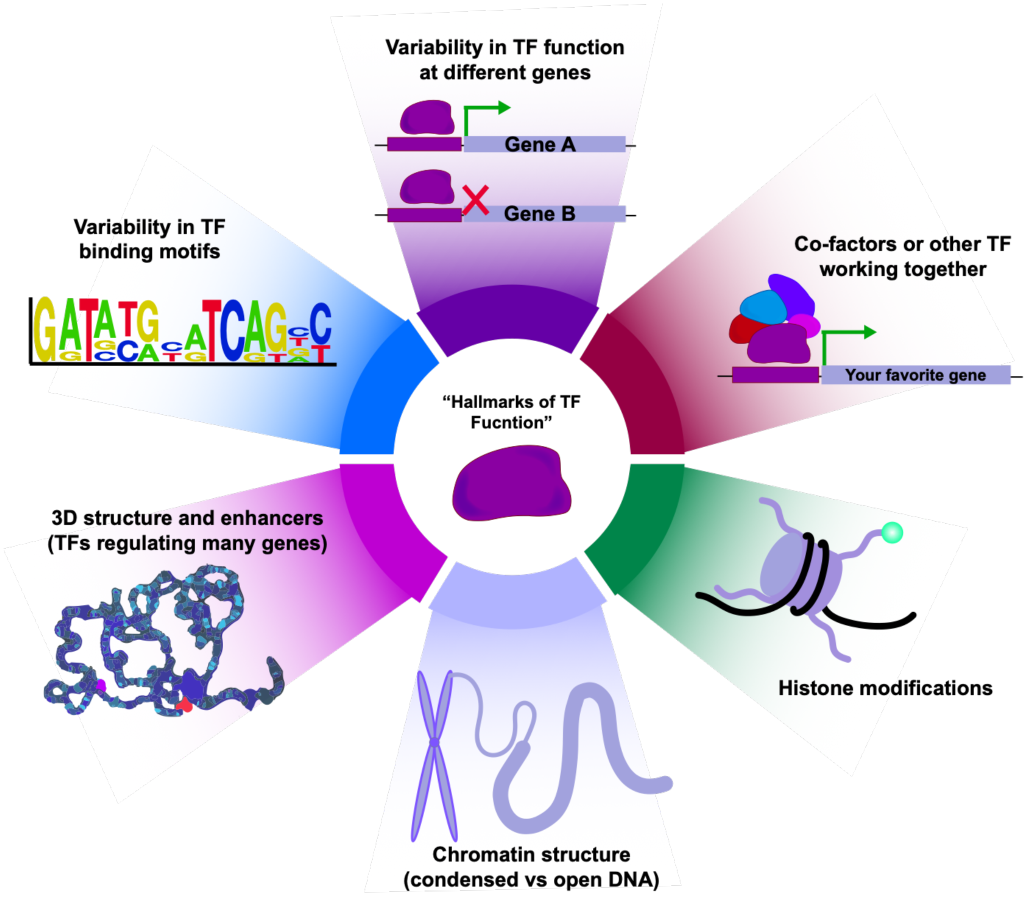

Variability in TF function at different genes
Gene A
Gene B
Variability in TF binding motifs
Co-factors or other TF working together
Your favorite gene
"Hallmarks of TF Fucntion"
3D structure and enhancers (TFs regulating many genes)
Histone modifications
Chromatin structure (condensed vs open DNA)

Figure 1.4 Hallmarks of transcription factor function. Transcription factor function can be impacted by several properties and characteristics including variability in *cis* motif sequence,

variable functions at different genomic loci, cofactors, histone modifications, chromatin structure, and three-dimensional genome interactions.

1.4.1 *Cis* element sequence

Some transcription factors are designated by their ability to physically interact with DNA through specific binding domains and therefore recognize unique *cis* elements called "motifs" within or related to their target loci which give transcription factors some specification for their gene targets. *Cis* elements themselves possess their own characteristics and variability that can influence transcription factor function.

Pioneer factors are one specific category of transcription factors that give some clues as to how transcription factors can identify their targets. Pioneer factors including Zelda (Dufourt *et al.* 2018), GAF (GAGA-factor) (Gaskill *et al.* 2021), and CLAMP (Chromatin Linked Adaptor of MSL protein) bind *cis* elements and have the ability to open the genome by loosening which plays an extremely important role during ZGA (Duan *et al.* 2021). The binding specificity of these and other pioneer factors contributes to them having the ability to accurately target those ~3600 genes that turns on during the waves of ZGA (Kwasnieski *et al.* 2019).

Transcription factors can require incredibly strict binding motifs for proper identification and function. For example, GAF binds a strict GA repeat motif which is required to be at least 5 bp. GAF also preferentially binds smaller GA-repeat *cis* elements in general and will be outcompeted by other GA-repeat binding factors such as CLAMP for longer GA-repeat elements (Kuzu *et al.* 2016; Kaye *et al.* 2018) . Interestingly, binding motifs can also incorporate variability in their sequence to influence transcription factor binding. CLAMP and Pipsqueak (Psq), another GA-repeat binding transcription factor, both differ from GAF in that they can bind more variable GA-rich regions rather than a strict GA-repeat (Lehmann *et al.* 1998; Kaye *et al.* 2018; Gutierrez-Perez *et al.* 2019). Psq is able to withstand an even higher amount of variability in its binding motif. Considering that CLAMP, GAF, and Psq all share binding motifs with GAs, the variability that they can either incorporate or not incorporate into their binding may give some clues to how they are able to retain unique function at some different loci across the genome. I discuss the relationship between these GA-binding factors further in Chapter 2.

Beyond allowing for some uniqueness of otherwise identical DNA targets, the variability of binding motifs can also serve as a variable that causes different functional outcomes for transcription factors based on what specific sequence the given TF binds. Variability in *cis* element sequence has not always been recognized as a characteristic that may influence transcription factor function, but recent studies show that even small sequence differences can have impactful effects on TFs. For example, the pioneer factor Zelda to binds specific *cis* elements called TAGteam sites, which includes a set of binding motifs with a TAG sequence always present (Satija and Bradley 2012). It was previously assumed that if these binding motifs were altered in any way Zelda would be non-functional at those gene targets. However, single nucleotide mutations in the Zelda binding motifs can actually change Zelda's function making it less able to activate its gene target (Harrison *et al.* 2010; Ozdemir *et al.* 2011; Li and Eisen 2018).

1.4.2 Cofactors

Transcription factor function can also be modulated by cofactors that are binding to or near a given target gene. One common assumption of transcription factors outlined by (Zeitlinger 2020) is that transcription factors bind across the genome each with their own impact on the genes they are regulating without incorporating communication or cooperativity. Considering that TF binding motifs exist in incredibly close proximity in the genome, the reality is that transcription factors often work in combination to regulate their target genes and often with strict synchronicity where two or more TF are completely necessary for gene regulation.

An interesting example of transcription factor cooperativity is the regulation of the major histocompatibility complex, MHC, Class II genes. MHC Class II genes, which produce molecules that are critical for cell surface receptors in immune response, are clustered together at a single locus which is composed of three classical class II genes (HLA-DP, -DQ, -DR) and two 'non-classical' class II genes (HLA-DM and -DO) along with their enhancer regulatory elements (Reith *et al.* 2005). The upstream enhancer region includes 4 distinct *cis* elements called "box" domains known as SXY, each of which are bound by a specific TF. The locus is regulated by a "master regulator" called CIITA which does not have DNA binding activity but instead interacts directly with each of the TFs that bind the SXY *cis* elements. Simply mutating one of TFs that bind the SXY or, even a single *cis* element of the SXY itself fully interrupts the regulation of the MHC class II genes, meaning that all of these are required in combination for proper locus regulation (Reith *et al.* 2005). Furthermore, this exemplifies how transcription factors work in conjunction with each other to create a necessary complex for regulation which differs from the model where regulatory events rely on one TF simply binding its target gene and regulating gene expression.

This type of cooperativity is also distinct from TF concentrations or bodies that regulate gene expression as such as or the histone locus body (HLB), the concentration of factors that regulates the histone genes, or the nucleolus, where rRNA is made. The HLB is a phase separated body (Hur *et al.* 2020) that includes a large number of different factors, most of which have not been identified as DNA binding factors. In the HLB, CLAMP serves as one of the only known links between the body of factors and the DNA sequence making it a critical factor for identifying the locus itself (Soruco *et al.* 2013; Rieder *et al.* 2017). While the HLB does not necessarily reflect combinations of transcription factors binding DNA and regulating together, the complexes in histone gene regulation do exemplify a set of factors that need to complex and cooperate together to regulate their respective loci.

1.5 Context-dependent transcription factors

Considering all the variables that can influence transcription factor function, one problem still remains; there are only a finite number of transcription factors in the genome therefore each factor cannot simply target only one gene. Because the human system is more complex and of less interest to me, from the perspective of *Drosophila*, there are approximately 16,000 genes that result in protein products and out of those factors, only ~ 700 are predicted to interact with DNA (Hammonds *et al.* 2013; Rhee *et al.* 2014). If there are only 700 transcription factors responsible for regulating the transcription of over 20 times the number of genes, it implies that some transcription factor may only regulate gene however some may regulate dozens. Transcription factors classified as "context-dependent" are able to bind similar *cis* elements across the genome but retain unique regulatory outputs at these different loci. This is in contrast to factors like yeast GAL4 that only binds the upstream activation sequence (UAS) and subsequently upregulates a downstream gene (Brand and Perrimon 1993) or LacI which binds exclusively to the LacO sequence and subsequently downregulates gene expression (Miller and Reznikoff 1978; Chao *et al.* 1980). Context dependent transcription factors may have a variety of functions at different loci ranging from upregulating, downregulating, or even serving as just a cofactor for recruitment.

Since factors can bind similar elements across the genome but perform different functions, a large gap in the field currently is understanding what inputs or cues these context dependent factors use to determine their functions at different loci. If we consider a factor like Zelda as the proxy for our hypothesis, we might suspect that context dependent factors read small changes or variability in their binding motif and adjust their function accordingly. However, context dependent factors may incorporate information from the flanking sequence where the *cis* element they bind resides or even factor in exploit information for the surrounding genomic environment, including chromatin structure or long-range interactions, to determine how to function at different loci.

1.5.1 CLAMP as a context-dependent transcription factor

To understand how transcription factors identify their targets, it is impractical to examine all ~700 *Drosophila* transcription factors wholly. As mentioned throughout this introduction, CLAMP is designated as a content dependent transcription factor for serving critical functions in two major coordinated gene expression events in *Drosophila*: histone gene expression and upregulation of the male X chromosome for dosage compensation. CLAMP is enriched on the male X chromosome where it targets GA-rich regions that often overlap with MREs (MSL recognition elements) (Soruco *et al.* 2013; Villa *et al.* 2016). This targeting allows Male Specific Lethal complex (MSLc) to be recruited accomplish dosage compensation in male flies (Soruco *et al.* 2013; Soruco and Larschan 2014). At the histone locus, CLAMP binds a long GA-repeat cis element in the promoter of *H3* and *H4* (*H3/H4*p) and fosters recruitment of additional factors responsible for histone gene regulation (Rieder *et al.* 2017). CLAMP targets both the histone genes and the X chromosome prior to locus-specific factors such as Mxc and MSLc, neither of which have strong DNA-binding capability. These observations show that, despite binding similar-seeming *cis* elements on the X chromosome and at the histone locus, CLAMP recruits different factors to each location ensuring proper group composition at each genomic location. It is unclear how early transcription factors such as CLAMP integrate information from GA-rich *cis* elements with other genomic context information to perform its locus-specific functions.

1.5.2 CLAMP function at the histone locus is a model for context-dependent transcription factor function

Although CLAMP is a convincing example of a context-dependent gene regulator, the processes CLAMP is involved in present a variety of limitations in trying to study how transcription factors determine their function. As stated previously, CLAMP binds GA-rich regions both on the X chromosome for *Drosophila* male dosage compensation and at the endogenous histone locus for histone gene regulation. Neither the *Drosophila* X-chromosome nor the endogenous histone locus are tractable study systems to explore the nuances of transcription factor function. MSLc coats the entire chromosome, and it is not practical to manipulate each GA-rich MRE in all 150 CES (Alekseyenko *et al.* 2008). Altering a critical number of CES, besides being nearly impossible to execute, would likely cause incomplete dosage compensation and male-specific lethality, while altering just a few CES is unlikely to significantly affect MSL recruitment to the whole chromosome due to complex spreading (Kelley *et al.* 1999; Kageyama *et al.* 2001; Gorchakov *et al.* 2009).

Similarly, the endogenous histone locus is organized in a series of ~100 tandemly repeated 5 kb arrays. This repetitive nature of the histone genes renders them intractable for genetic manipulation as there are CLAMP-binding GA-repeats in the *H3/H4*p of all 100 histone locus arrays. However, the *Drosophila* transgenic histone array provides an excellent genetic system in which to test this hypothesis. Transgenes carrying 1-12 histone gene arrays have been established that allow for genetic manipulation and recapitulate histone locus functionality (Günesdogan *et al.* 2010; Salzler *et al.* 2013; McKay *et al.* 2015; Meers *et al.* 2018) A single histone genes array transgene does not rescue an endogenous histone locus deletion but successfully recruits HLB-specific factors and allow for histone gene expression (Koreski *et al.* 2020). The histone array transgene system is a powerful tool that allows us to perturbate the *cis* elements within a histone gene array without needing to edit all 100 arrays at the endogenous locus.

1.6 Goals and major finding

In **Chapter 2**, I focus on exploring the CLAMP binding GA-repeat within the histone gene arrays in *D. melanogaster*. Interestingly, although the ~107 histone arrays are virtually identical, we discovered there is one feature of the array that varies: the length of the GA-repeat in the *H3/H4*p, ranging from 16 to 35 nucleotides in length. Historically, CLAMP has been characterized for binding MRE motifs on the X chromosome and targets longer GA-rich *cis* elements on the X chromosome with higher affinity than it does shorter repeats (Kuzu *et al.* 2016; Kaye *et al.* 2018). I first sought to describe the impact of GA-repeat variability on transcription factor binding as well as understand if the GA-repeat variability impact differential regulation of the histone gene arrays at the histone locus. I hypothesized that CLAMP would preferentially target arrays that contain longer GA-repeats due to CLAMP's preference for longer GA-repeats on the X chromosome. I confirmed that CLAMP as well as two other GA-repeat binding transcription factors pipsqueak (Psq) and GAGA factor (GAF) target the histone locus based on ChIP-seq analysis. I show that CLAMP and GAF, despite targeting different length GA-repeats across the genome, target all of the GA-repeat lengths within the histone arrays. My results suggested that the GA-repeat itself does not impact differential regulation of the histone arrays and implies there may be additional *cis* elements that recruit cofactors that can impact histone gene regulation.

In **Chapter 3**, I explore how transcription factors integrate information from *cis* element sequence as well as contextual cues to perform their context specific function. I focused on the CLAMP's role at the histone locus and leveraged a single histone array transgenic system to manipulate the CLAMP binding *H3/H4*p GA-repeat. I aimed to integrate how changes in *cis* element sequence and context could impact CLAMP function. I determined that CLAMP gleans information from not only *cis* element sequence but also from flanking sequence to determine its function at the histone locus. My results suggest that transcription factors incorporate more than just information about the *cis* elements they target, and these cues may play into how transcription factors are able to coordinate the expression of different genes.

In **Chapter 4**, my undergraduate mentee and I explored *cis*-element conservation in different *Drosophila* species. The histone gene coding sequences themselves are impressively conserved across species. In the ~ 40 MYa diverged species *Drosophila virilis*, the histone genes are spread across two loci, more like in humans and distinct from the single locus in *Drosophila* melanogaster. Even more interestingly, in *D. virilis*, one of the two histone loci show localization of the X chromosome, dosage compensation MSL2 factor prompting the question of how the regulation of histone gene expression is different between sexes. These results suggest that histone gene regulation may be mechanistically different in other *Drosophila* species and prompts many new questions about what aspects of coordinated histone gene regulation are truly conserves and those that are unique in different systems.

Finally, in **Chapters 5 and 6,** I discuss how we can broaden our understanding of the contextual cues provided to transcription factors by investigating cofactors working in combination to achieve coordinated gene regulation. Despite recent advancements in the field of histone gene expression, we have still yet to create a fully extensive list of factors responsible for regulating the transcription and processing of the histone genes in *Drosophila melanogaster*. To discover novel DNA-binding proteins that target the histone locus, we turned to mining literature for likely candidates and then funneled these into a secondary bioinformatics screen. We established a Course-based Undergraduate Research Experience (CURE) focused on making novel bioinformatics-based research projects that are accessible to undergraduate students. We discovered that the Hox proteins Ubx, Abd-A and Abd-B likely target the histone locus and provided over 40 undergraduate students with hand-on research experience.

1.7 Summary

Coordinated gene regulation, and particularly the coordination of histone gene regulation, is complex. The requirements for genes to be tightly regulated at different times and at the correct levels is an incredibly challenging feat to orchestrate. My work indicates that transcription factors play a large role in coordinated gene regulation and seem to incorporate critical information from the *cis* elements they target, flanking sequences and genomic context, and even cofactors that are targeting the same regions. Further complicating our understanding of these processes, it seems that many coordinated gene events like histone gene regulation, dosage compensation and ZGA have their own unique mechanism and studying each of them is essential for broadening our understanding of all properties that govern coordinated gene regulation.

1.8 References

Alekseyenko A. A., S. Peng, E. Larschan, A. A. Gorchakov, O.-K. Lee, *et al.*, 2008 A sequence motif within chromatin entry sites directs MSL establishment on the Drosophila X chromosome. Cell 134: 599–609. https://doi.org/10.1016/j.cell.2008.06.033

Bongartz P., and S. Schloissnig, 2018 Deep repeat resolution—the assembly of the Drosophila Histone Complex. Nucleic Acids Research 47: e18–e18. https://doi.org/10.1093/nar/gky1194

Brand A. H., and N. Perrimon, 1993 Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. Development 118: 401–415. https://doi.org/10.1242/dev.118.2.401

Carty M., L. Zamparo, M. Sahin, A. González, R. Pelossof, *et al.*, 2017 An integrated model for detecting significant chromatin interactions from high-resolution Hi-C data. Nat Commun 8: 15454. https://doi.org/10.1038/ncomms15454

Chao M. V., H. G. Martinson, and J. D. Gralla, 1980 lac Operator nucleosomes. 2. lac Nucleosomes can change conformation to strengthen binding by lac repressor. Biochemistry 19: 3260–3269. https://doi.org/10.1021/bi00555a025

Chari S., H. Wilky, J. Govindan, and A. A. Amodeo, 2019 Histone concentration regulates the cell cycle and transcription in early development. Development 146: dev177402. https://doi.org/10.1242/dev.177402

Chaubal A., J. M. Waldern, C. Taylor, A. Laederach, W. F. Marzluff, *et al.*, 2023 Coordinated expression of replication-dependent histone genes from multiple loci promotes histone homeostasis in Drosophila. MBoC 34: ar118. https://doi.org/10.1091/mbc.E22-11-0532

Cremer T., and C. Cremer, 2001 Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nat Rev Genet 2: 292–301. https://doi.org/10.1038/35066075

Duan J., and E. N. Larschan, 2019 Dosage Compensation: How to Be Compensated…Or Not? Current Biology 29: R1229–R1231. https://doi.org/10.1016/j.cub.2019.09.065

Duan J., L. Rieder, M. M. Colonnetta, A. Huang, M. Mckenney, *et al.*, 2021 CLAMP and Zelda function together to promote Drosophila zygotic genome activation. eLife.

Dufourt J., A. Trullo, J. Hunter, C. Fernandez, J. Lazaro, *et al.*, 2018 Temporal control of gene expression by the pioneer factor Zelda through transient interactions in hubs. Nat Commun 9: 5194. https://doi.org/10.1038/s41467-018-07613-z

Duronio R. J., and W. F. Marzluff, 2017 Coordinating cell cycle-regulated histone gene expression through assembly and function of the Histone Locus Body. RNA Biol 14: 726–738. https://doi.org/10.1080/15476286.2016.1265198

Farrell J. A., and P. H. O'Farrell, 2014 From egg to gastrula: how the cell cycle is remodeled during the Drosophila mid-blastula transition. Annu Rev Genet 48: 269–294. https://doi.org/10.1146/annurev-genet-111212-133531

Gaskill M. M., T. J. Gibson, E. D. Larson, and M. M. Harrison, 2021 GAF is essential for zygotic genome activation and chromatin accessibility in the early Drosophila embryo, (Y. M. Yamashita, and K. Struhl, Eds.). eLife 10: e66668. https://doi.org/10.7554/eLife.66668

Geisler M. S., J. P. Kemp Jr., and R. J. Duronio, 2023 Histone locus bodies: a paradigm for how nuclear biomolecular condensates control cell cycle regulated gene expression. Nucleus 14: 2293604. https://doi.org/10.1080/19491034.2023.2293604

Ghule P. N., J. R. Boyd, F. Kabala, A. J. Fritz, N. A. Bouffard, *et al.*, 2023 Spatiotemporal higher-order chromatin landscape of human histone gene clusters at histone locus bodies during the cell cycle in breast cancer progression. Gene 872: 147441. https://doi.org/10.1016/j.gene.2023.147441

Gorchakov A. A., A. A. Alekseyenko, P. Kharchenko, P. J. Park, and M. I. Kuroda, 2009 Long-range spreading of dosage compensation in Drosophila captures transcribed autosomal genes inserted on X. Genes Dev 23: 2266–2271. https://doi.org/10.1101/gad.1840409

Günesdogan U., H. Jäckle, and A. Herzig, 2010 A genetic system to assess in vivo the functions of histones and histone modifications in higher eukaryotes. EMBO Rep 11: 772–6. https://doi.org/10.1038/embor.2010.124

Gunjan A., and A. Verreault, 2003 A Rad53 kinase-dependent surveillance mechanism that regulates histone protein levels in S. cerevisiae. Cell 115: 537–549. https://doi.org/10.1016/s0092-8674(03)00896-1

Gutierrez-Perez I., M. J. Rowley, X. Lyu, V. Valadez-Graham, D. M. Vallejo, *et al.*, 2019 Ecdysone-Induced 3D Chromatin Reorganization Involves Active Enhancers Bound by Pipsqueak and Polycomb. Cell Rep 28: 2715-2727.e5. https://doi.org/10.1016/j.celrep.2019.07.096

Hamm D. C., and M. M. Harrison, 2018 Regulatory principles governing the maternal-to-zygotic transition: insights from Drosophila melanogaster. Open Biol 8: 180183. https://doi.org/10.1098/rsob.180183

Hammonds A. S., C. A. Bristow, W. W. Fisher, R. Weiszmann, S. Wu, *et al.*, 2013 Spatial expression of transcription factors in Drosophila embryonic organ development. Genome Biol 14: R140. https://doi.org/10.1186/gb-2013-14-12-r140

Harrison M. M., M. R. Botchan, and T. W. Cline, 2010 Grainyhead and Zelda compete for binding to the promoters of the earliest-expressed Drosophila genes. Developmental Biology 345: 248–255. https://doi.org/10.1016/j.ydbio.2010.06.026

Harrison M. M., and M. B. Eisen, 2015 Transcriptional Activation of the Zygotic Genome in Drosophila. Curr Top Dev Biol 113: 85–112. https://doi.org/10.1016/bs.ctdb.2015.07.028

Hori Y., C. Engel, and T. Kobayashi, 2023 Regulation of ribosomal RNA gene copy number, transcription and nucleolus organization in eukaryotes. Nat Rev Mol Cell Biol 24: 414–429. https://doi.org/10.1038/s41580-022-00573-9

Hur W., J. P. Kemp, M. Tarzia, V. E. Deneke, W. F. Marzluff, *et al.*, 2020 CDK-Regulated Phase Separation Seeded by Histone Genes Ensures Precise Growth and Function of

Histone Locus Bodies. Dev Cell 54: 379-394.e6. https://doi.org/10.1016/j.devcel.2020.06.003

Jimeno-González S., L. Payán-Bravo, A. M. Muñoz-Cabello, M. Guijo, G. Gutierrez, *et al.*, 2015 Defective histone supply causes changes in RNA polymerase II elongation rate and cotranscriptional pre-mRNA splicing. Proceedings of the National Academy of Sciences 112: 14840–14845. https://doi.org/10.1073/pnas.1506760112

Kageyama Y., G. Mengus, G. Gilfillan, H. G. Kennedy, C. Stuckenholz, *et al.*, 2001 Association and spreading of the Drosophila dosage compensation complex from a discrete roX1 chromatin entry site. The EMBO Journal 20: 2236–2245. https://doi.org/10.1093/emboj/20.9.2236

Kaye E. G., M. Booker, J. V. Kurland, A. E. Conicella, N. L. Fawzi, *et al.*, 2018 Differential Occupancy of Two GA-Binding Proteins Promotes Targeting of the Drosophila Dosage Compensation Complex to the Male X Chromosome. Cell Rep 22: 3227–3239. https://doi.org/10.1016/j.celrep.2018.02.098

Kelley R. L., V. H. Meller, P. R. Gordadze, G. Roman, R. L. Davis, *et al.*, 1999 Epigenetic spreading of the Drosophila dosage compensation complex from roX RNA genes into flanking chromatin. Cell 98: 513–22. https://doi.org/10.1016/s0092-8674(00)81979-0

Koreski K. P., L. E. Rieder, L. M. McLain, A. Chaubal, W. F. Marzluff, *et al.*, 2020 Drosophila histone locus body assembly and function involves multiple interactions. Mol Biol Cell 31: 1525–1537. https://doi.org/10.1091/mbc.E20-03-0176

Kuzu G., E. G. Kaye, J. Chery, T. Siggers, L. Yang, *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 Genet 12: e1006120. https://doi.org/10.1371/journal.pgen.1006120

Kwasnieski J. C., T. L. Orr-Weaver, and D. P. Bartel, 2019 Early genome activation in Drosophila is extensive with an initial tendency for aborted transcripts and retained introns. Genome Res. 29: 1188–1197. https://doi.org/10.1101/gr.242164.118

Lehmann M., T. Siegmund, K. G. Lintermann, and G. Korge, 1998 The pipsqueak protein of Drosophila melanogaster binds to GAGA sequences through a novel DNA-binding domain. J Biol Chem 273: 28504–28509. https://doi.org/10.1074/jbc.273.43.28504

Li X.-Y., and M. B. Eisen, 2018 Effects of the maternal factor Zelda on zygotic enhancer activity in the Drosophila embryo. 385070.

Li J., Z. Ming, L. Yang, T. Wang, G. Liu, *et al.*, 2022 Long noncoding RNA XIST: Mechanisms for X chromosome inactivation, roles in sex-biased diseases, and therapeutic opportunities. Genes Dis 9: 1478–1492. https://doi.org/10.1016/j.gendis.2022.04.007

Lucchesi J. C., and M. I. Kuroda, 2015 Dosage Compensation in Drosophila. Cold Spring Harb Perspect Biol 7: a019398. https://doi.org/10.1101/cshperspect.a019398

Mao Y. S., B. Zhang, and D. L. Spector, 2011 Biogenesis and Function of Nuclear Bodies. Trends Genet 27: 295–306. https://doi.org/10.1016/j.tig.2011.05.006

Marzluff W. F., P. Gongidi, K. R. Woods, J. Jin, and L. J. Maltais, 2002 The Human and Mouse Replication-Dependent Histone Genes. Genomics 80: 487–498. https://doi.org/10.1006/geno.2002.6850

Marzluff W. F., E. J. Wagner, and R. J. Duronio, 2008 Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail. Nat Rev Genet 9: 843–854. https://doi.org/10.1038/nrg2438

Matera A. G., M. Izaguire-Sierra, K. Praveen, and T. K. Rajendra, 2009 Nuclear Bodies: Random Aggregates of Sticky Proteins or Crucibles of Macromolecular Assembly? Developmental Cell 17: 639–647. https://doi.org/10.1016/j.devcel.2009.10.017

Maya Miles D., X. Peñate, T. Sanmartín Olmo, F. Jourquin, M. C. Muñoz Centeno, *et al.*, 2018 High levels of histones promote whole-genome-duplications and trigger a Swe1WEE1-dependent phosphorylation of Cdc28CDK1. Elife 7: e35337. https://doi.org/10.7554/eLife.35337

McKay D. J., S. Klusza, T. J. Penke, M. P. Meers, K. P. Curry, *et al.*, 2015 Interrogating the function of metazoan histones using engineered gene clusters. Dev Cell 32: 373–86. https://doi.org/10.1016/j.devcel.2014.12.025

Meers M. P., M. Leatham-Jensen, T. J. R. Penke, D. J. McKay, R. J. Duronio, *et al.*, 2018 An Animal Model for Genetic Analysis of Multi-Gene Families: Cloning and Transgenesis of Large Tandemly Repeated Histone Gene Clusters. Methods Mol Biol 1832: 309–325. https://doi.org/10.1007/978-1-4939-8663-7_17

Michalak P., 2008 Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. Genomics 91: 243–248. https://doi.org/10.1016/j.ygeno.2007.11.002

Miller J. H., and W. S. Reznikoff, 1978 *The Operon*. Cold Spring Harbor Laboratory.

Nair R. R., E. Pataki, and J. E. Gerst, 2022 Transperons: RNA operons as effectors of coordinated gene expression in eukaryotes. Trends in Genetics 38: 1217–1227. https://doi.org/10.1016/j.tig.2022.07.005

Nizami Z., S. Deryusheva, and J. G. Gall, 2010 The Cajal Body and Histone Locus Body. Cold Spring Harb Perspect Biol 2: a000653. https://doi.org/10.1101/cshperspect.a000653

Ozdemir A., K. I. Fisher-Aylor, S. Pepke, M. Samanta, L. Dunipace, *et al.*, 2011 High resolution mapping of Twist to DNA in Drosophila embryos: Efficient functional analysis and evolutionary conservation. Genome Res 21: 566–577. https://doi.org/10.1101/gr.104018.109

Pearson J. C., D. Lemons, and W. McGinnis, 2005 Modulating Hox gene functions during animal body patterning. Nat Rev Genet 6: 893–904. https://doi.org/10.1038/nrg1726

Ramírez F., T. Lingg, S. Toscano, K. C. Lam, P. Georgiev, *et al.*, 2015 High-Affinity Sites Form an Interaction Network to Facilitate Spreading of the MSL Complex across the X Chromosome in Drosophila. Mol Cell 60: 146–162. https://doi.org/10.1016/j.molcel.2015.08.024

Reith W., S. LeibundGut-Landmann, and J.-M. Waldburger, 2005 Regulation of MHC class II gene expression by the class II transactivator. Nat Rev Immunol 5: 793–806. https://doi.org/10.1038/nri1708

Rhee D. Y., D.-Y. Cho, B. Zhai, M. Slattery, L. Ma, *et al.*, 2014 Transcription Factor Networks in Drosophila melanogaster. Cell Rep 8: 2031–2043. https://doi.org/10.1016/j.celrep.2014.08.038

Rieder L. E., K. P. Koreski, K. A. Boltz, G. Kuzu, J. A. Urban, *et al.*, 2017 Histone locus regulation by the Drosophila dosage compensation adaptor protein CLAMP. Genes Dev 31: 1494–1508. https://doi.org/10.1101/gad.300855.117

Salzler H. R., D. C. Tatomer, P. Y. Malek, S. L. McDaniel, A. N. Orlando, *et al.*, 2013 A sequence in the Drosophila H3-H4 Promoter triggers histone locus body assembly and biosynthesis of replication-coupled histone mRNAs. Dev Cell 24: 623–34. https://doi.org/10.1016/j.devcel.2013.02.014

Satija R., and R. K. Bradley, 2012 The TAGteam motif facilitates binding of 21 sequence-specific transcription factors in the Drosophila embryo. Genome Res 22: 656–665. https://doi.org/10.1101/gr.130682.111

Soruco M. M., J. Chery, E. P. Bishop, T. Siggers, M. Y. Tolstorukov, *et al.*, 2013 The CLAMP protein links the MSL complex to the X chromosome during Drosophila dosage compensation. Genes Dev 27: 1551–6. https://doi.org/10.1101/gad.214585.113

Soruco M. M. L., and E. Larschan, 2014 A new player in X identification: the CLAMP protein is a key factor in Drosophila dosage compensation. Chromosome Res 22: 505–515. https://doi.org/10.1007/s10577-014-9438-4

Tadros W., and H. D. Lipshitz, 2009 The maternal-to-zygotic transition: a play in two acts. Development 136: 3033–42. https://doi.org/10.1242/dev.033183

Tatomer D. C., E. Terzo, K. P. Curry, H. Salzler, I. Sabath, *et al.*, 2016 Concentrating pre-mRNA processing factors in the histone locus body facilitates efficient histone mRNA biogenesis. Journal of Cell Biology 213: 557–570. https://doi.org/10.1083/jcb.201504043

Villa R., T. Schauer, P. Smialowski, T. Straub, and P. B. Becker, 2016 PionX sites mark the X chromosome for dosage compensation. Nature 537: 244–248. https://doi.org/10.1038/nature19338

White A. E., B. D. Burch, X. C. Yang, P. Y. Gasdaska, Z. Dominski, *et al.*, 2011 Drosophila histone locus bodies form by hierarchical recruitment of components. J Cell Biol 193: 677–94. https://doi.org/10.1083/jcb.201012077

Zeitlinger J., 2020 Seven myths of how transcription factors read the cis-regulatory code. Current Opinion in Systems Biology 23: 22–31. https://doi.org/10.1016/j.coisb.2020.08.002

Zhang L., Z. Wu, and H. Lu, 2022 Top(ological-operon) secret behind the long-range transcriptional coupling. Sig Transduct Target Ther 7: 1–3. https://doi.org/10.1038/s41392-022-01195-5

Chapter 2

Cis element length variability does not confer differential transcription factor occupancy at the histone locus

2.1 Abstract

Histone genes require precise regulation to maintain histone homeostasis and ensure nucleosome, critical genome packaging units, synthesis. Animal histone genes often have unique clustered genomic organization; however, there is variability of histone gene number and organization as well as differential regulation of the histone genes across species. The *Drosophila melanogaster* histone locus has unique organizational characteristics as it exists as a series of ~100 highly regular, tandemly repeated arrays of the 5 replication-dependent histone genes at a single locus. We hypothesize that the histone genes within arrays across the locus may be differentially regulated based on the fact that *D. melanogaster* are viable with only 12 histone gene arrays. We discovered that the GA-repeat within the *H3/H4* promoter is the only variable sequence across the histone gene arrays. The *H3/H4* promoter GA-repeat is targeted by CLAMP to promote histone gene regulation however we also show two additional GA-bind transcription factors, GAF and Psq may also target the GA-repeat. When we examined CLAMP and GAF targeting further, we determined that neither CLAMP nor GAF showed bias for any GA-repeat lengths. Furthermore, we found that the distribution of GA-repeats targeted by both CLAMP and GAF do not change throughout early development. Together our results suggest that the transcription factors targeting GA-repeat do not themselves impact differential regulation of the histone genes but prompts future studies to interrogate additional *cis* elements or factors that may work together to impact histone gene regulation.

2.2 Introduction

Histone genes need to be strictly regulated so there are neither too few nor too many histones at any given time in the cell. The canonical histone genes, *H3, H4, H2A, H2B,* and *H1*, are replication dependent; their regulation is strictly coupled to the cycle. Histone genes are expressed during S phase to package newly replicated DNA followed by complete halting of the gene expression by the end of G. In part due to the requirement for strict coordinated cell cycle regulation, animal histone genes often have unique clustered genomic organization; however, there is variability of histone gene organization and differential regulation across species.

Histone genes were originally cloned and sequenced from purple and green sea urchins, *S. purpuratus* and *P. miliaris* respectively, in the late 70s from which we learned sea urchin genomes have two clustered sets of histone genes. The first set consists of a tandem repeat of the five canonical histone genes termed the "early histone genes" and a second set of 39 genes that are separated from the early genes, termed the "late histone genes" (Marzluff *et al.* 2006). These two histone gene sets are differentially regulated based on cell type and timing. The early histone genes are only expressed in the egg through the blastula stage whereas the late histone genes are expressed during late embryogenesis and continue expression through adulthood in all somatic cells (Marzluff *et al.* 2006).

The human genome also carries two clusters of histone genes, a major cluster on chromosome 6 and a minor cluster on chromosome 1. All H1 genes are located in the major cluster which is spread across several megabases and contains ~60 histone genes in smaller sub-clusters while the minor cluster only contains around 10-12 histone genes (Seal *et al.* 2022; Ghule *et al.* 2023) . Recent Hi-C data shows there are distinct promoter-promoter interactions between the subclusters of the major histone locus on chromosome 6 which could suggest regulatory mechanisms that are different between the major and minor locus (Carty *et al.* 2017; Ghule *et al.* 2023). Transcription factors that regulate histone genes are shared between these loci however they are differentially regulated to ensure there are correct stoichiometries of H3, H4, H2A, H2B and H1 are made for proper nucleosome structure. The major histone locus also associates with another nuclear body called the Cajal body, x through the cell cycle whereas the minor locus only associates with it during S phase (Ma *et al.* 2000; Shopland *et al.* 2001). From work in human embryonic stem cells, the H4 genes may have distinct regulation patterns between he major and minor loci based on tumor cell type however the patterns of H4 gene expression show only minor differences between loc in embryonic stem cells. This suggests that the overall contribution of histone transcripts from the major and minor loci are similar implying that there are mechanisms of differential regulation that keep this equilibrium despite differences in histone gene copy number between the loci (Becker *et al.* 2007).

Fission yeast are an even more extreme example of how histone genes are differentially regulated. Fission yeast genomes contain three pairs of H3-H4 gene along with just a single pair of H2Aalpha-H2B, and a lone H2A beta gene. A study looking at the three pairs of H3 and H4 genes found that the first and third pair are up-regulated while the second pair is normally downregulated, exhibiting oscillation of expression through the cell cycle (Takayama and Takahashi 2007).

The histone genes in *Drosophila melanogaster* are a unique example of clustered histone gene organization. The *D. melanogaster* genome carries a single repetitive histone locus on chromosome 2L. Based on recent locus assembly from long-read sequencing (Bongartz and Schloissnig 2019), the *D. melanogaster* histone locus includes approximately 107 tandemly repeated histone gene arrays. Each 5kb array includes the five canonical histone genes, *H1, H3,* *H4, H2A* and *H2B* along with their respective *cis* regulatory elements and promoters. *H3* and *H4* share a bi-directional promoter that contains an important GA-repeat *cis* element required for histone gene regulation (Rieder *et al.* 2017). These characteristics of histone gene organization in *D. melanogaster* are somewhat distinct from the above-mentioned yeast, sea urchins, and humans that all have multiple histone gene clusters at different loci.

The studies from yeast, sea urchin, and even humans show that histone genes are differentially regulated based on timing, gene copy number, and number of loci. Despite these different organisms having diverse means of regulating and maintaining histone homeostasis, as a group they set a president for the hypothesis that histone genes in all animals have some level of differential regulatory mechanisms. Now having the knowledge that there are approximately 107 histone gene arrays that comprise the histone locus one large question still remains: are all 107 of the histone gene arrays expressed equally. We have some evidence that would suggest all 107 arrays need to be active from work using a 12-array transgene (Günesdogan *et al.* 2010; Salzler *et al.* 2013; McKay *et al.* 2015). This 12-arrray transgene is sufficient for viability in a genetic background where the endogenous histone locus is deleted suggesting that as few as 12 arrays at the endogenous locus could be active (Salzler *et al.* 2013; Koreski *et al.* 2020). These data confirm that not every array is necessary and therefore we hypothesize that the arrays will be expressed at different times or at the different levels, perhaps due to differences in TF function across arrays.

To explore how the arrays at the endogenous *D. melanogaster* histone locus might be functionally different, we utilized a recent histone locus assembly completed through long-read sequencing (Bongartz and Schloissnig 2019) to search for sequence differences between the histone gene arrays. We discovered that the arrays are nearly identical in sequence, but the GA-repeat in the s *H3/H4* promoter is variable in length ranging from 16-35 nucleotides in length. The *H3/H4* promoter sequence can nucleate recruitment of specific histone regulatory factors, and the GA-repeat is specifically targeted by the transcription factor CLAMP (Salzler *et al.* 2013; Rieder *et al.* 2017; Koreski *et al.* 2020). Further, we recently confirmed that the GA-repeats are critical for histone locus factor recruitment. Therefore, we refined our hypothesis and tested if the length variability of the GA-repeat is responsible for differential transcription factor occupancy. To test this hypothesis, we obtained existing ChIP-seq data from CLAMP and other GA-repeat binding factors GAGA Factor (GAF) and Pipsqueak (Psq) and investigated their differential occupancy over GA-repeat elements. We discovered that all three factors bind the range of GA-repeat lengths and, furthermore, show that CLAMP and GAF are unbiased in the GA-repeat lengths they bind. Our discovery of variable GA-repeats at the histone locus uncovered a previously unknown distinction of the 107 histone gene arrays and may provide a target for future studies on histone array uniqueness and functionality. Furthermore, our observations suggest that the GA-repeat variability likely does not contribute to differential occupancy of transcription factors at histone gene arrays and implies other *cis* elements or cofactors that might contribute to differential histone gene regulation.

2.3 Results

2.3.1 The GA-repeat is variable in length across the histone gene arrays.

Bongartz *et al.* (2019) produced a *de novo* assembly of the *Drosophila melanogaster* repetitive histone gene locus, identifying that the locus contains ~107 histone gene arrays. We aligned the gene arrays and discovered that they are nearly identical in sequence other than length variability of a GA-repeat present in the bidirectional promoter of genes *H3* and *H4*

(**Figure 2.1A,B,C**). The GA-repeat varies in length from 16 base pairs to 35 base pairs (**Figure 2.1C,D**). The most common GA-repeat length is 21 bp (29 of the 107 arrays). We found some clustering of arrays with similar length GA-repeats such as those that have GA-repeats with 29 or 31 bp GA-repeats.

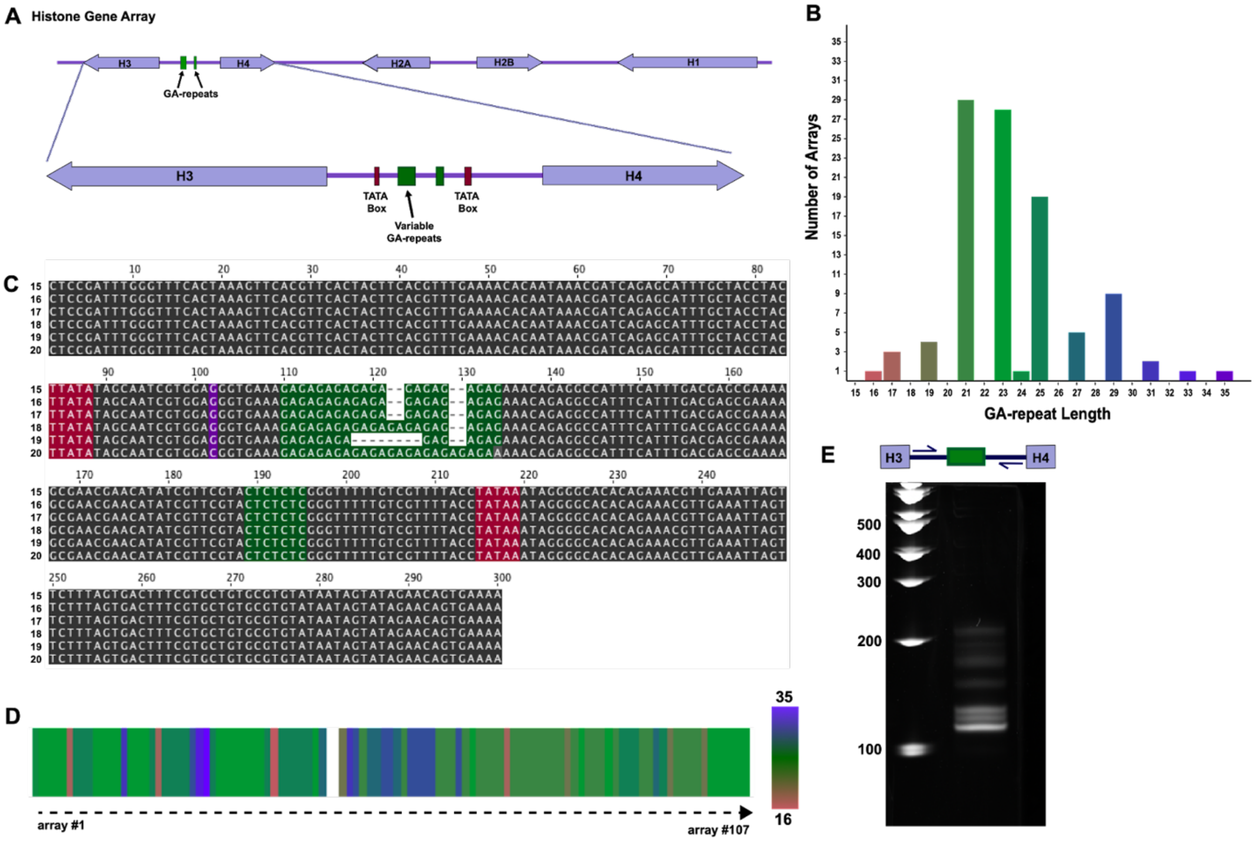


Figure 2.1: The GA-repeat length is variable across the ~100 histone gene arrays. (A) A diagram of a single histone gene array and the GA-repeat element located in the *H3/H4*p. **(B)** We utilized previously assembled histone locus from Bongartz et al. (2019) to compare the sequences of the histone gene arrays. The arrays are virtually identical other than the GA-repeat in the H3H4p, which varies in length. **(C)** We aligned six of the 300 bp *H3/H4*p (arrays 15-20, TATA boxes in maroon). Other than a single SNP (purple), the GA-repeat remains the only

sequence variability. **(D)** A heatmap shows the positions of different GA-repeat lengths across the locus. Each array is represented by one vertical bar. **(E)** We designed primers to amplify the *H3/H4*p of the histone arrays to confirm the variability of the GA-repeat in vivo. Laddering of PCR products in an acrylamide gel confirmed GA-repeat variability.

To confirm the GA-repeat variability *in vivo*, we designed primers to amplify about 115 bp of the endogenous *H3/H4*p region that includes the GA-repeat region. PCR from genomic DNA is predicted to produce amplicons ranging from 110 bp (16 bp GA-repeat) to 129 bp (35 bp GA-repeat). We observed the expected laddering of PCR products on an acrylamide gel, confirming GA-repeat length variability *in vivo* (**Figure 2.1E**). We noticed several amplicons that exceeded the predicted length, possibly due to secondary structure forming due to the GA-repeat.

2.3.2 CLAMP, GAF, and Psq all target the GA-repeats in the *H3/H4*p

Our observations indicate that the most dramatic sequence difference across the histone arrays is the wide variability of the GA-repeat length. We previously demonstrated that this sequence is targeted by the CLAMP transcription factor (Rieder *et al.* 2017) and that the interaction is important for HLB factor recruitment and histone gene expression (Rieder *et al.* 2017; Hodkinson *et al.* 2023). However, the *Drosophila* genome carries two other GA-repeat binding transcription factors: GAF and Psq (Lehmann *et al.* 1998; van Steensel *et al.* 2003). We therefore hypothesized that these other GA-repeat binding factors are also targeting the GA-repeats in the histone gene array. To test our hypothesis, we aligned previously generated ChIP-sequencing data to the histone gene array (Gutierrez-Perez *et al.* 2019; Gaskill *et al.* 2021; Duan *et al.* 2021).

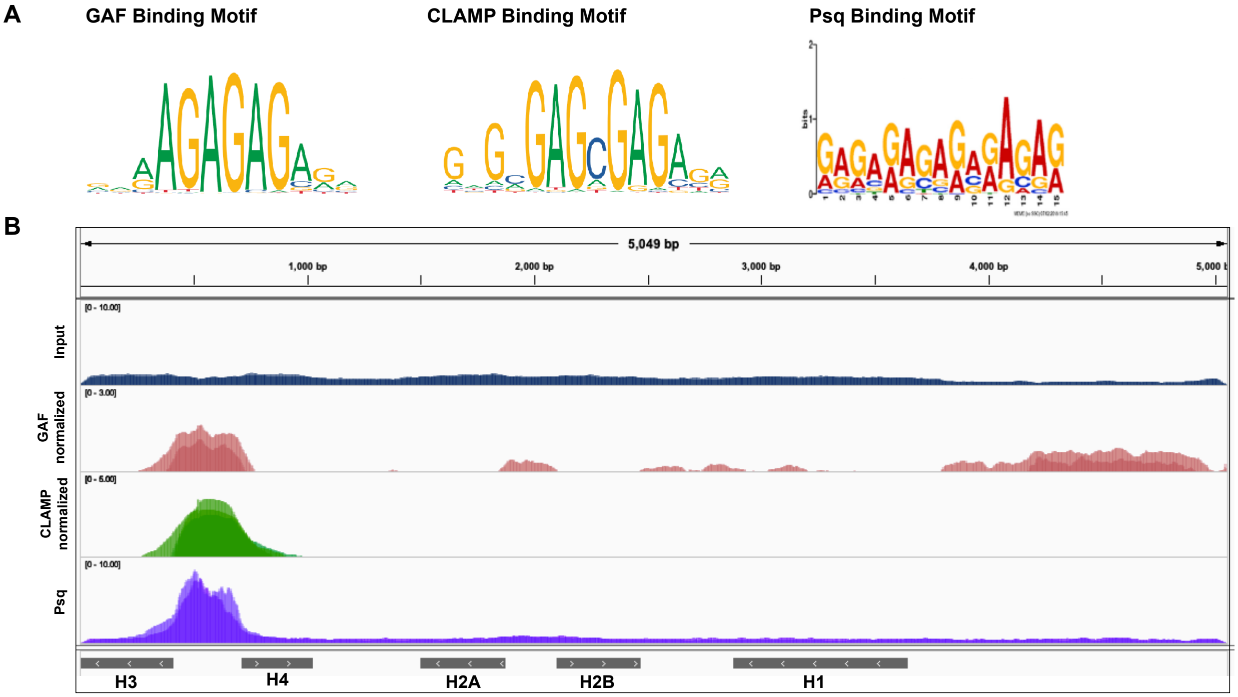


Figure 2.2: GAF, CLAMP, and Psq target the *H3/H4*p GA-repeat. (A) The binding motifs for GAF, CLAMP and Psq all contain GA-repeats. Binding motifs for GAF and CLAMP generated by the open access database JASPAR (Castro-Mondragon et al 2022) and Psq binding motif recreated from Gutierrez-Perez *et al*. 2019. **(B)** We aligned ChIP-seq data for GAF (pink, two replicates overlayed(Gaskill *et al.* 2021)) in 2-3 hr embryos, CLAMP (green, three replicates overlayed (Duan *et al.* 2021)) in 2-4 hr embryos, and Psq (purple, two replicates overlayed (Gutierrez-Perez *et al.* 2019)) in 2-4 hr embryos to the single histone gene array. GAF and CLAMP data were normalized to respective inputs. Psq was not normalized because no inputs were provided. All three factors target the GA-repeat in the *H3/H4*p of the histone gene array. A representative input (blue) from the GAF and CLAMP ChIP-seq data is shown for comparison.

Because of the repetitive nature of the histone locus, aligning sequencing data such as ChIP-seq reads becomes impractical as each read would map to more than one or even all 107 histone gene arrays. Historically, to align sequencing data to the histone gene array, we utilized a condensed or custom version of the histone gene array, similar to the histone gene array outlined in Mckay *et al.* (2015), containing only one copy of each of the canonical histone genes along with their promoters. Using the condensed histone gene array also means there is only one GA-repeat *cis* element, which happens to be 21 bp.

First, we confirmed that CLAMP robustly targets the *H3/H4* promoter GA-repeat **(Figure 2.2).** CLAMP shows a clear peak over this region, as previously observed (Rieder *et al.* 2017; Koreski *et al.* 2020). Previously published GAF ChIP data from 2-3 hr embryos (Gaskill *et al.* 2021) and Psq data from *Drosophila* embryonic stem cells (Kc167 cells, (Gutierrez-Perez *et al.* 2019)) also show a clear peak at the *H3/H4* promoter. Based on these data, CLAMP, GAF and Psq all target the histone locus.

2.3.3 GA-binding factors do not show preference for GA-repeat length at the histone locus

Although the ChIP peaks shown in **Figure 2.2** for all three GA-repeat binding factors imply that they target the histone locus, we cannot deduce which array or arrays they target because we are only looking at the data aligned to a single histone array rather than the entire locus. We next wanted to deduce what arrays each of the three GA-repeat binding factors might target by examining whether they have a bias for certain length GA-repeats. Because the GA-repeats are the only sequence that differs between the 107 histone gene arrays, knowing what length GA-repeats CLAMP, GAF, and Psq target can help us infer what arrays they target. CLAMP shows preference for binding longer GA-repeats on the X chromosome while GAF shows preference for short GA-repeats (Kaye *et al.* 2018). *In vitro*, CLAMP binds DNA probes with long GA-repeats up to 30 nucleotides in length by EMSA where GAF would only shift pieces of DNA with shorter GA-repeats of 8 nucleotides (Kaye *et al.* 2018). Therefore, we hypothesized that these GA-repeat binding factors might target different histone arrays based on their binding preference for GA-repeat length.

We developed a bioinformatics script that selected *H3/H4* promoter sequences from the ChIP-seq dataset by defining two anchor sequences, one upstream (5') and one downstream (3') of the GA-repeat with enough length to ensure specificity to the H3/H4 promoter. The code then extracts the reads that match both anchors, scans to identify the GA-repeat and counts the number of nucleotides that make up the GA-repeat in that given read. We utilized the ChIP input as a positive control, hypothesizing that we would recapitulate the GA-repeat lengths and frequencies we retrieved from the long-read sequencing results.

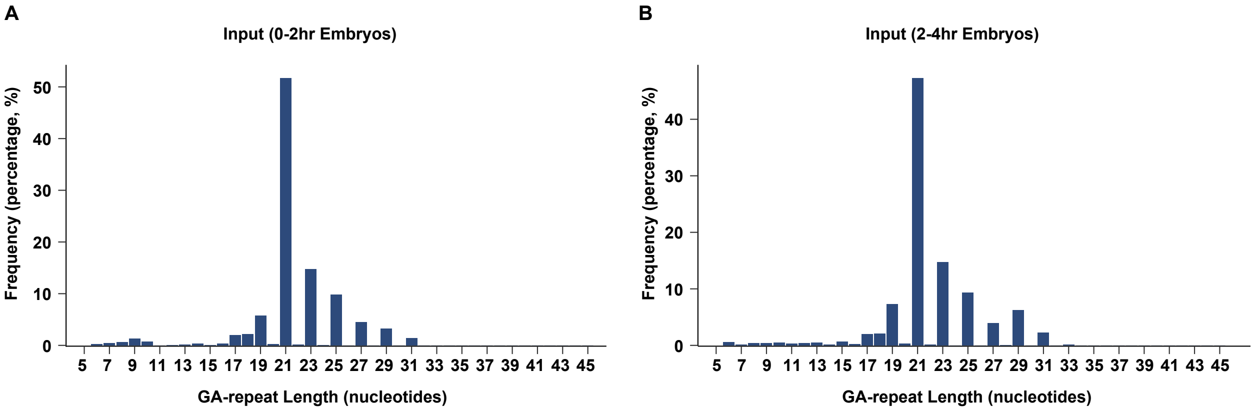


Figure 2.3 The *H3/H4* promoter GA-repeat length variability is observed in different datasets. We designed a bioinformatics code to parse through ChIP-seq datasets, extract reads containing the *H3/H4* promoter GA-repeat and count the number of nucleotides within the

repeat. We extracted reads from input ChIP-seq datasets of **(A)** 0-2hr embryos (three replicates) and **(B)** 2-4hr embryos (three replicates) and created histograms based on the the GA-repeat nucleotide counts. The X-axis shows all GA-repeat lengths, and the Y-axis is the frequency each length was found represented as a percentage of extracted reads which contained that GA-repeat length.

When we generated histograms for GA-repeat length frequency from the input libraries of 0-2hr and 2-4hr embryos, we observed that the distribution of GA-repeat lengths mirrored the distribution we found from the long read Bongartz *et al.* (2019) data. However, we did notice a few differences. We identified some GA-repeats that were shorter than expected due to SNPs in the middle of the GA-repeat (**Supplementary Figure 2.1**). In addition, we noticed some minor differences in the frequencies of element lengths (**Figure 2.1B** vs. **Figure 2.3**). This is likely due to strain genotype, as the Bongartz assembly was obtained from OregonR *Drosophila*, while the ChIP-seq dataset was obtained from *yellow-, white-* animals. Because large, repetitive regions of the genome are subjected to frequent expansion and contraction due to unequal crossing over (Smith 1976), it's also likely that few *Drosophila* strains have exactly the same GA-repeat length distribution. Even individuals within an interbreeding population may have different numbers of arrays and therefore frequencies of GA-repeat lengths. Overall, however, we confirmed that the variability and length distribution of the histone locus GA-repeats is relatively reproducible.

To determine the binding profiles of CLAMP and GAF at the variable GA-repeat, we used an available CLAMP ChIP-seq dataset from Duan *et al.* (2021) and generated a GAF ChIP dataset, both of which include data from 0-2 and 2-4hr *Drosophila* embryos. These are relevant time points for histone gene expression; the early *Drosophila embryo* undergoes 14 nuclear division cycles where every 8-12 minutes, the entire genome is replicated therefore, a large number of histones are rapidly required (Tadros and Lipshitz 2009; Farrell and O'Farrell 2014; Harrison and Eisen 2015). The histone genes are targeted by specific factors as early as nuclear cycle 9 (Terzo *et al.* 2015), and zygotic histone genes are expressed by nuclear cycle 11 (Edgar and Schubiger 1986). CLAMP is maternally deposited and targets the histone locus in the early embryo, prior to detectable histone gene expression (Rieder *et al.* 2017). GAF is not thought to target the zygotic histone locus unless CLAMP is depleted (Rieder *et al.* 2017), although we discovered that it likely does so, at least from some datasets (**Figure 2.2**).

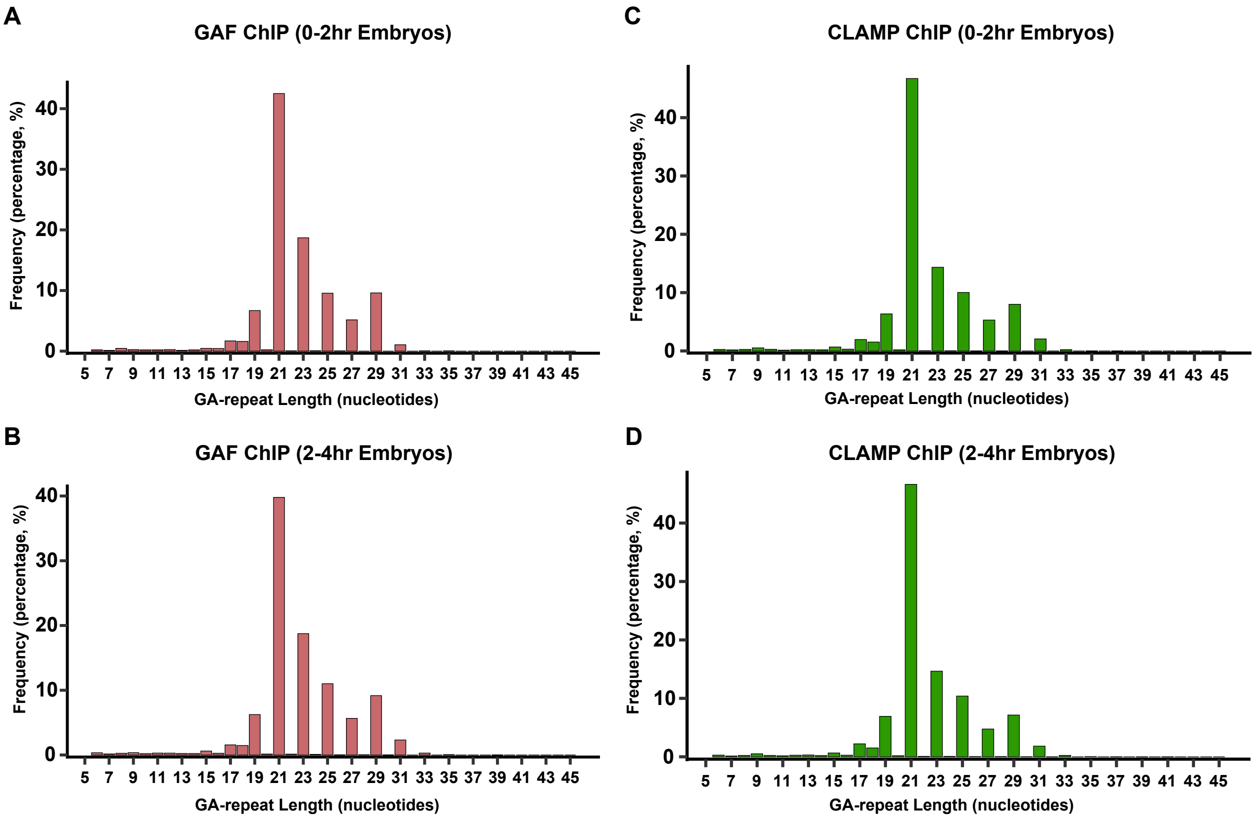


Figure 2.4 GAF and CLAMP target the same length GA-repeats. We extracted reads containing the *H3/H4* promoter GA-repeat from ChIP-seq data for GAF in **(A)** 0-2hr embryos

(three replicates) and **(B)** 2-4hr embryos (three replicates). We also extracted reads contain the *H3/H4* promoter GA-repeat from ChIP-seq data for CLAMP in **(C)** 0-2hr embryos (three replicates) and **(D)** 2-4hr embryos (three replicates). In both histograms, the X-axis shows all GA-repeat lengths, and the Y-axis is the average frequency each length was found represented as a average percentage of extracted reads which contained that GA-repeat length from the three replicates of each data set.

Using these embryonic ChIP-seq datasets, we investigated the frequencies of GA-repeat lengths targeted by CLAMP and GAF. We found that all GA-repeat lengths were bound by CLAMP, which does not show any bias for specific GA-repeat lengths despite preferring long X-linked GA-repeats (Kaye *et al.* 2018). Furthermore, CLAMP seems to target each of the GA-repeat lengths at similar frequencies to their respective counts across the locus (**input, Figure 2.3**). Lastly, we found no difference between the distribution of GA-repeat lengths targeted by CLAMP based on age of embryo, suggesting that developmental timing does not impact CLAMP distribution to the GA-repeats at the histone locus (**Figure 2.4 C,D**).

We next performed a similar analysis for GAF and retrieved similar results. GAF showed no bias for specific GA-repeat lengths in either 0-2 or 2-4 hr embryos (**Figure 2.4 A,B)**. Further, we found that GAF also seems to bind each of the GA-repeat lengths at similar frequencies to their respective counts across the locus (**input, Figure 2.3**) similar to CLAMP (**Figure 2.4 A,B**). Although we identified a previously generated Psq ChIP-seq dataset from 3 hr embryos, we were unable to interrogate GA-repeat binding preference due to the short length (50 bp) of the sequencing reads, which is not sufficient to identify reads that contain both the anchor sequences and the GA-repeat.

2.4 Discussion

The *Drosophila melanogaster* histone locus comprises ~107 virtually identical histone gene arrays and is regulated by a unique nuclear body. It is unknown whether all ~107 histone genes are all targeted by the same transcription factors and produce the same mRNA output, as we are unable to differentiate the histone gene expression of each of the arrays. Some evidence points toward differential expression of genes. Animals carrying 12-array transgenes in the background of an endogenous locus deletion are viable (Günesdogan *et al.* 2010; Salzler *et al.* 2013; Koreski *et al.* 2020) and express histone mRNAs at the same level as the endogenous locus, indicating that 100 are not required for viability. Other species, including other drosophilids have varying numbers of histone genes in differing genomic arrangements, such as the closely related *D. simulans* whose genome carries only 15 histone arrays (Sisi Falcone, unpublished data) or ~40 MYa diverged *D. virilis* whose genome carries two histone loci which combined only contain 32 arrays (Russo *et al.* 1995; Schienman *et al.* 1998). It is difficult to assay how the *D. melanogaster* genes might be differentially expressed, as the histone coding sequences are virtually identical. We therefore sought to uncover mechanisms for differential histone gene regulation by investigating sequence differences between arrays.

Using a recent long-read histone locus assembly (Bongartz et al. 2019), we discovered that the GA-repeat in the *H3/H4* promoter is variable in length across the histone locus, while the rest of the 5 kb arrays are nearly identical in sequence. We previously demonstrated that CLAMP targets the GA-repeat in the *H3/H4* promoter and confirmed that the GA-repeats are important for histone locus factor recruitment (Rieder *et al.* 2017; Hodkinson *et al.* 2023). These observations indicated the importance of the GA-repeat in overall histone gene regulation, so we hypothesized that this sequence variability might be functionally important in recruiting different transcription factors. We found that all GA-binding transcription factors target the element, but that none seems to have a bias for longer or shorter repeats.

GA dinucleotide repeats are fairly common in many genomes and serve a variety of functions. GA-repeat, also called shore tandem repeats (STRs), are commonly found in core promoter sequences to serve as targets for transcription factors or pioneer factors like GAF which displace nucleosomes to ready the gene for transcription (Valipour *et al.* 2013). Recent work looking at a conserved GA-repeat in the core promoter of early human embryonic development genes shows that differences in GA-repeat length at these genes can cause differences in expression levels (Valipour *et al.* 2013). These data confirm that GA-repeat length itself is sufficient to drive differential gene expression and, furthermore, may imply that the length of the GA-repeat in the histone gene arrays may also impact differential expression of the histone gene within array even if this is not due to changes in CLAMP, GAF, or Psq binding.

GA dinucleotide repeats can also serve as insulators. GAF binding at GA-repeats is critical for insulation between genes and unrelated, neighboring enhancer sequences (Lehmann 2004; Gaskill *et al.* 2021). In mice, GA-repeat motifs within the Hox gene clusters are nucleosome-free and, when GAF targets these regions, chromatin boundaries are established to create domains so the Hox genes themselves are insulated from their neighboring regulatory elements (Srivastava *et al.* 2013). Similarly, in *D. melanogaster* GAF localizes to the *Fab-7* boundary element from the Hox genes *Ubx, Abd-A* and *Abd-B*. GAF can target distinct GA-repeats at the *Fab-7* element which can determine its function as an insulator at different developmental time points (Schweinsberg *et al.* 2004). It is possible that the GA-repeat in the histone array acts as an insulator and, although it is located within the *H3/H4* promoter, it may serve multiple functions as the target for binding factors at a subset of arrays and as an insulator for others to modulate the expression if histone genes in different arrays.

Our observations suggest the GA-repeat may not impact differential expression of the histone genes, however here we did not explore if GA-repeat length impacted individual histone gene expression levels. The repetitiveness of the histone locus makes it impossible to assess the expression of individual histone genes because there is no sequence variation to differentiate what histone gene is expressed each gene array. Future experiments could leverage a barcoded 12-array transgene where silent mutations can be made in the histone genes to differential expression from each array. Using this system, we could look at how GA-repeat length impacts histone gene expression rather than just GA-repeat binding. Studies in sea urchins, which have two clusters of histone genes that are differentially regulated to be expressed “early” or “late” in development, show that the specific downregulation of the “early” H2A genes is regulated by an upstream (5’) GA-repeat serving as an insulator (Di Caro *et al.* 2004). This study suggests that the regulation of individual histone genes can be governed by *cis* elements. Furthermore, this data emphasizes that dinucleotide repeats, and specifically GA-repeats, have important functions across species and in many genomic contexts.

GA-repeats are not the only *cis* element that can modulate gene expression and it is likely that there are secondary or several additional *cis* elements that are responsible for regulating different histone arrays or genes (Horton *et al.* 2022). The work here focused specifically on the GA-repeat as it is essential for CLAMP binding and the only variable sequence between the arrays however, additional *cis* elements in the *H3/H4* or *H2A/H2B* promoter could impact differential regulation at the histone locus. Furthermore, it is possible that all the arrays are targeted by GA-repeat binding factors, but additional transcription factors are needed to then activate the arrays and express the histone genes within that array. Here we only consider the impact of the GA-repeat binding factors at the histone array however we know that there is a body of factors that regulate histone gene expression known as the HLB (Duronio and Marzluff 2017), the full composition of which is still unknown. Future studies exploring what other DNA-binding factors target the histone arrays, like the recently published screen from Hodkinson *et al.* 2023, as well as investigating how differential targeting may impact histone gene expression will provide greater understanding of the intricacies involved in histone gene regulation in *D. melanogaster.*

2.5 Conclusions

By utilizing the previously assembled histone locus sequencing data (Bongartz and Schloissnig 2019), we revealed the *H3/H4* promoter GA-repeat *cis* element is the only variable sequence between the ~107 gene arrays. By leveraging previously published ChIP-sequencing datasets we determined that the variability in the GA-repeat does not impact the binding of factors GAF or CLAMP and suggests that these factors alone are not responsible for any differential regulation of the histone. Overall, our results and observations have expanded our understanding of the sequence features of the *D. melanogaster* histone locus and given insight into what may govern histone gene regulation.

2.6 Methods

Promoter Alignment

We obtained the H3/H4 promoter sequences from the Bongertz *et al.* (2019) genome assembly (Figure 2.1) and used reads extracted from input ChIP-sequncing data (Supplemental Figure 2.1). We aligned sequences using T-Coffee Multiple Sequence Alignment (MSA) (Notredame *et al.* 2000) to create a ClustalW output and formatted the shading and features with Jalview (Waterhouse *et al.* 2009).

ChIP-analysis and Data Visualization - IGV plots

We directly imported individual FASTQ datasets into the web-based platform Galaxy (The Galaxy Community 2022) through the NCBI SRA Run Selector by selecting the desired runs and utilizing the computing Galaxy download feature. We retrieved the FASTQ files from SRA using the "Faster Download and Extract Reads in FASTQ format from NCBI SRA" Galaxy command. Because the ~100 histone gene arrays are extremely similar in sequence, we do not utilize the dm6 or dm3 genomes and instead collapse ChIP-seq data onto a single histone array. We used a custom "genome" that includes a single *Drosophila melanogaster* histone array similar to that in Mckay *et al.* (2015), which we directly uploaded to Galaxy using the "upload data" feature, and normalized using the Galaxy command "NormalizeFasta" specifying an 80 bp line length for the output .fasta file. We aligned ChIP reads to the normalized histone gene array using Bowtie2 (Langmead and Salzberg 2012) to create .bam files using the user built-in index and "very sensitive end-to-end" parameter settings. We converted the .bam files to .bigwig files using the "bamCoverage" Galaxy command in which we set the bin size to 1 bp and set the effective genome size to user specified: 5000 bp (approximate size of l histone array). If an input dataset was available, we normalized ChIP datasets to input using the "bamCompare" Galaxy command in which we set the bin size to 1 bp. We visualized the bigwig files using the Integrative Genome Viewer (IGV) (Robinson *et al.* 2011).

Table 2.1 ChIP-sequencing datasets. Specifics for the NCBI GEO datasets used including the GEO Accession number, the SRA Run selector numbers, the developmental time of each sample, and the cited source.

Factor	**GEO Accession #**	**SRA Run Selector**	**Developmental Timepoint**	**Citation**
GAF GAGA Factor (Trl)	GSE152773	**Anti GAF-GFP** 1 -SRR12045586 2 - SRR12045588 **Input** 1- SRR12045585 2 - SRR12045587	2-3hr, stage 5 embryos	(Gaskill *et al.* 2021)
CLAMP Chromatin linked adaptor for MSL proteins	GSE152613	CLAMP antibody 1 - SRR12024931 2 - SRR12024949 3 - SRR12024967 Input 1 - SRR12024933 2 - SRR12024951 3 - SRR12024969	2-4hr embryos	(Duan *et al.* 2021)
Psq Pipsqueak	GSE118047	PsqM (PsqTot) antibody 1- SRR7638403 2 - SRR7638404	Kc167 *Drosophila* embryonic cell line	(Gutierrez-Perez *et al.* 2019)

ChIP-analysis and Bioinformatics Pipeline - GA-repeat Histograms

Our annotated pipeline (code) will be available on GitHub (pending) in the script entitled count_ag_repeats.py. We utilized packages SeqIO from biopython (Cock *et al.* 2009) to parse through fastq files and regex from anaconda or pip for all functional outputs. We also utilized logging from anaconda or pip to create a built-in log for the run as an informational output. We designed the code to first identify, and extract reads that contain the H3H4 promoter sequence by using two short, flanking "anchor" sequences to the left (5') and the right (3') of the GA-repeat (left sequence: TAGCAATCGT right sequence: CATTTCATTTGACGAGC). We used a counting mechanism to ensure that reads with both the left and the right anchor were extracted however there is also an information output for single matches. We then designed the code to scan through the extracted reads until encountering the specified string "AGAGAG" as a seed sequence for the GA-repeat. Once the GA-repeat is identified, we designed the code to count the number of nucleotides within the repeat. Of note, we designed the code to allow for 0 mismatches in the repeat which meaning repeats where two "A" nucleotides or two "G" nucleotides are adjacent to each other will only be counted until that "AA" or "GG" appears. We identified that there are a handful of GA-repeats that contains SNPs causing "AA" or "GG" stretches (**Supplementary Figure 2.1**). However, this feature of the pipeline is changeable to allow for any specified number of mismatches. The script outputs 6 files to a specified path destination. These outputs include a .tsv file with four columns of information; the first column is nucleotide count of the GA-repeat, the second column has the extracted repeat itself, and the third column with the trimmed read where the repeat originated, and the last column has the sequence ID (**Supplementary Figure 2.2**). This file allows confirmation of the GA-repeat nucleotide counts as well as access to the reads the pipeline extracted. The other 5 files are .fastq.gz files that include reads from the script parsing through the entire sequencing file which include: dual_match.fastq.gz containing all the full length reads that had both ancho sequences, left_only.fastq.gz containing reads that only matched the left anchor sequence, no_match.fastq.gz, containing reads that did not have either anchor sequence, right only.fastq.gz containing reads that only matched the right anchor sequence, and strange_match.fastq.gz contain reads with unexpected configurations of the anchor sequences such as forward and reverse complements of these sequences.

ChIP-analysis and Data – GA-repeat Histograms

CLAMP ChIP-seq datasets from Duan et al. 2021 were retrieved from is deposited at NCBI GEO and the accession number is (GSE152598). GAF ChIP data was performed as described in Duan *et al.* 2021 with 10uL of GAF antibody (Fuda *et al.* 2015).

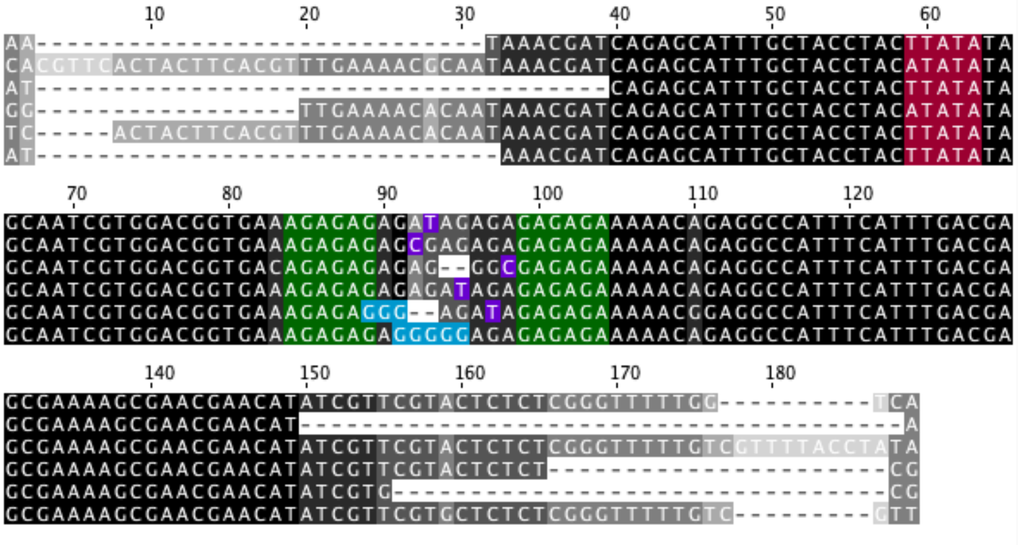
Table 2.2 ChIP-sequencing data used to generate GA-repeat length histograms

Target TF	Developmental Timepoint	GEO Accession #	SRA Run Selector #
Input	0-2hr embryo	GSE152613	Input 1- SRR12024924 2 - SRR12024942 3 - SRR12024960
Input	2-4hr embryo	GSE152613	Input 1 - SRR12024933 2 - SRR12024951 3 - SRR12024969
GAF	0-2hr embryo	pending	GAF antibody pending
GAF	2-4hr embryo	pending	GAF antibody pending
CLAMP	0-2hr embryo	GSE152613	CLAMP antibody 1- SRR12024922 2 - SRR12024940 3 - SRR12024958
CLAMP	2-4hr embryo	GSE152613	CLAMP antibody 1 - SRR12024931 2 - SRR12024949 3 - SRR12024967

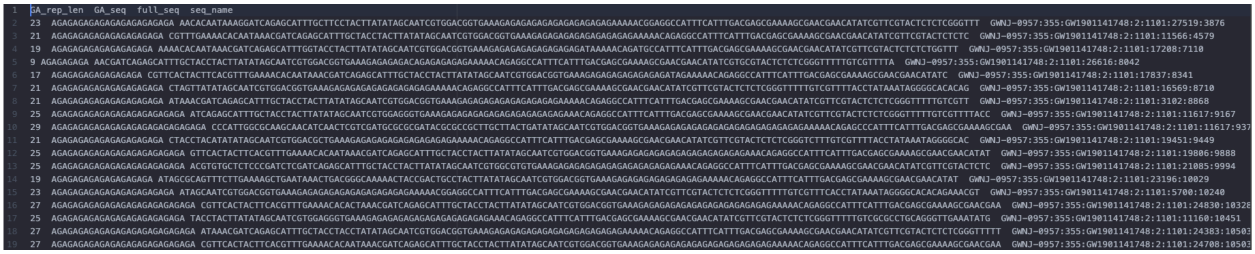
2.7 Acknowledgments

We would like to thank the Rieder Lab for supporting this work. We would specifically like to thank Dr. Casey Schmidt for her invaluable insight into early experimental design and execution. We would also like to thank Dr. David Gorkin who wrote the python pipeline and was an invaluable source of support and bioinformatics knowledge throughout this project. Additionally, we would like to thank the Lis Lab for the GAF antibody. This work was supported by T32GM00008490 and F31HD105452 to LJH; and R00HD092625 and R35GM142724 to LER.

2.8 Supplemental Figures


10
20
30
40
50
60
AA-----------------------------TAAACGATCAGAGCATTTGCTACCTACTTATATA
CACGTTCACTACTTCACGTTTGAAAACGCAATAAACGATCAGAGCATTTGCTACCTACATATATA
AT-------------------------------------CAGAGCATTTGCTACCTACTTATATA
GG-----------------TTGAAAACACAATAAACGATCAGAGCATTTGCTACCTACATATATA
TC-----ACTACTTCACGTTTGAAAACACAATAAACGATCAGAGCATTTGCTACCTACTTATATA
AT------------------------------AAACGATCAGAGCATTTGCTACCTACTTATATA
70
80
90
100
110
120
GCAATCGTGGACGGTGAAAGAGAGAGATAGAGAGAGAGAAAAACAGAGGCCATTTCATTTGACGA
GCAATCGTGGACGGTGAAAGAGAGAGCGAGAGAGAGAGAAAAACAGAGGCCATTTCATTTGACGA
GCAATCGTGGACGGTGACAGAGAGAGAG--GGCGAGAGAAAAACAGAGGCCATTTCATTTGACGA
GCAATCGTGGACGGTGAAAGAGAGAGAGATAGAGAGAGAAAAACAGAGGCCATTTCATTTGACGA
GCAATCGTGGACGGTGAAAGAGAGGG--AGATAGAGAGAAAAACGGAGGCCATTTCATTTGACGA
GCAATCGTGGACGGTGAAAGAGAGAGGGGGAGAGAGAGAAAAACAGAGGCCATTTCATTTGACGA
140
150
160
170
180
GCGAAAAGCGAACGAACATATCGTTCGTACTCTCTCGGGTTTTTGG----------TCA
GCGAAAAGCGAACGAACAT---------------------------------------A
GCGAAAAGCGAACGAACATATCGTTCGTACTCTCTCGGGTTTTTGTCGTTTTACCTATA
GCGAAAAGCGAACGAACATATCGTTCGTACTCTCT----------------------CG
GCGAAAAGCGAACGAACATATCGTG--------------------------------CG
GCGAAAAGCGAACGAACATATCGTTCGTGCTCTCTCGGGTTTTTGTC---------GTT

Supplementary Figure 2.1 The GA-repeat can contain a variety of mismatches and sequence variation. We aligned a representative set of reads extracted by our bioinformatics pipeline where the *H3/H4* promoter GA-repeats contains SNPs or stretches of repeating A or G nucleotides. One of the TATA boxes is labeled in maroon and the GA-repeat is labeled in green. SNPs are shown in purple and stretches of A or G nucleotides are shown in teal. (Note these sequences have been extracted and trimmed by our python script).



```
GA_rep_len  GA_seq  full_seq  seq_name
23  AGAGAGAGAGAGAGAGAGAGAGA AACACAATAAAGGATCAGAGCATTTGCTTCCTACTTATATAGCAATCGTGGACGGTGAAAGAGAGAGAGAGAGAGAGAGAGAGAGAAAAACGGAGGCCATTTCATTTGACGAGCGAAAAGCGAACGAACATATCGTTCGTACTCTCTCGGGTTT  GWNJ-0957:355:GW1901141748:2:1101:27519:3876
21  AGAGAGAGAGAGAGAGAGAGA CGTTTGAAAACACAATAAACGATCAGAGCATTTGCTACCTACTTATATAGCAATCGTGGACGGTGAAAGAGAGAGAGAGAGAGAGAGAGAAAAACAGAGGCCATTTCATTTGACGAGCGAAAAGCGAACGAACATATCGTTCGTACTCTCTC  GWNJ-0957:355:GW1901141748:2:1101:11566:4579
19  AGAGAGAGAGAGAGAGAGA AAAACACAATAAACGATCAGAGCATTTGGTACCTACTTATATAGCAATCGTGGACGGTGAAAGAGAGAGAGAGAGAGAGATAAAAACAGATGCCATTTCATTTGACGAGCGAAAAGCGAACGAACATATCGTTCGTACTCTCTCTCTGGTTT  GWNJ-0957:355:GW1901141748:2:1101:17208:7110
9 AGAGAGAGA AACGATCAGAGCATTTGCTACCTACTTATATAGCAATCGTGGACGGTGAAAGAGAGAGACAGAGAGAGAGAAAAACAGAGGCCATTTCATTTGACGAGCGAAAAGCGAACGAACATATCGTGCGTACTCTCTCGGGTTTTTGTCGTTTTA  GWNJ-0957:355:GW1901141748:2:1101:26616:8042
17  AGAGAGAGAGAGAGAGA CGTTCACTACTTCACGTTTGAAAACACAATAAACGATCAGAGCATTTGCTACCTACTTATATAGCAATCGTGGACGGTGAAAGAGAGAGAGAGAGAGATAGAAAAACAGAGGCCATTTCATTTGACGAGCGAAAAGCGAACGAACATATC  GWNJ-0957:355:GW1901141748:2:1101:17837:8341
21  AGAGAGAGAGAGAGAGAGAGA CTAGTTATATAGCAATCGTGGACGGTGAAAGAGAGAGAGAGAGAGAGAGAAAAACAGAGGCCATTTCATTTGACGAGCGAAAAGCGAACGAACATATCGTTCGTACTCTCTCGGGTTTTTGTCGTTTTACCTATAAATAGGGGCACACAG  GWNJ-0957:355:GW1901141748:2:1101:16569:8710
21  AGAGAGAGAGAGAGAGAGAGA ATAAACGATCAGAGCATTTGCTACCTACTTATATAGCAATCGTGGACGGTGAAAGAGAGAGAGAGAGAGAGAGAAAAACAGAGGCCATTTCATTTGACGAGCGAAAAGCGAACGAACATATCGTTCGTACTCTCTCGGGTTTTTGTCGTT  GWNJ-0957:355:GW1901141748:2:1101:3102:8868
25  AGAGAGAGAGAGAGAGAGAGAGAGA ATCAGAGCATTTGCTACCTACTTATATAGCAATCGTGGAGGGTGAAAGAGAGAGAGAGAGAGAGAGAGAGAAACAGAGGCCATTTCATTTGACGAGCGAAAAGCGAACGAACATATCGTTCGTACTCTCTCGGGTTTTTGTCGTTTTACC  GWNJ-0957:355:GW1901141748:2:1101:11617:9167
29  AGAGAGAGAGAGAGAGAGAGAGAGAGAGA CCCATTGGCGCAAGCAACATCAACTCGTCGATGCGCGCGATACGCGCCGCTTGCTTACTGATATAGCAATCGTGGACGGTGAAAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAAAAACAGAGCCCATTTCATTTGACGAGCGAAAAGCGAA  GWNJ-0957:355:GW1901141748:2:1101:11617:937
21  AGAGAGAGAGAGAGAGAGAGA CTACCTACATATATAGCAATCGTGGACGCTGAAAGAGAGAGAGAGAGAGAGAGAAAAACAGAGGCCATTTCATTTGACGAGCGAAAAGCGAACGAACATATCGTTCGTACTCTCTCGGGTCTTTGTCGTTTTACCTATAAATAGGGGCAC  GWNJ-0957:355:GW1901141748:2:1101:19451:9449
25  AGAGAGAGAGAGAGAGAGAGAGAGA GTTCACTACTTCACGTTTGAAAACACAATAAACGATCAGAGCATTTGCTACCTACTTATATAGCAATCGTGGACGGTGAAAGAGAGAGAGAGAGAGAGAGAGAGAAACAGAGGCCATTTCATTTGACGAGCGAAAAGCGAACGAACATAT  GWNJ-0957:355:GW1901141748:2:1101:19806:9888
25  AGAGAGAGAGAGAGAGAGAGAGAGA ACGTGTGCTCTCCCGATCTCGATCAGAGCATTTGCTACCTACTTATATAGCAATCGTGGCGTGTGAAAGAGAGAGAGAGAGAGAGAGAGAGAGAGAAACAGAGGCCATTTCATTTGACGAGCGAAAAGCGAACGAACATATCGTTCGTACTCTC  GWNJ-0957:355:GW1901141748:2:1101:21085:9994
19  AGAGAGAGAGAGAGAGAGA ATAGCGCAGTTTCTTGAAAAGCTGAATAAACTGACGGGCAAAAACTACCGACTGCCTACTTATATAGCAATCGTGGACGGTGAAAGAGAGAGAGAGAGAGAAAAACAGAGGCCATTTCATTTGACGAGCGAAAAGCGAACGAACATAT  GWNJ-0957:355:GW1901141748:2:1101:23196:10029
23  AGAGAGAGAGAGAGAGAGAGAGA ATAGCAATCGTGGACGGTGAAAGAGAGAGAGAGAGAGAGAGAGAAAAACGGAGGCCATTTCATTTGACGAGCGAAAAGCGAACGAACATATCGTTCGTACTCTCTCGGGTTTTTGTCGTTTCACCTATAAATAGGGGCACACAGAAACGT  GWNJ-0957:355:GW1901141748:2:1101:5700:10240
27  AGAGAGAGAGAGAGAGAGAGAGAGAGA CGTTCACTACTTCACGTTTGAAAACACACTAAACGATCAGAGCATTTGCTACCTACTTATATAGCAATCGTGGACGGTGAAAGAGAGAGAGAGAGAGAGAGAGAGAGAGAAAAACAGAGGCCATTTCATTTGACGAGCGAAAAGCGAACGAA  GWNJ-0957:355:GW1901141748:2:1101:24830:10328
25  AGAGAGAGAGAGAGAGAGAGAGAGA TACCTACTTATATAGCAATCGTGGAGGGTGAAAGAGAGAGAGAGAGAGAGAGAGAGAGAAACAGAGGCCATTTCATTTGACGAGCGAAAAGCGAACGAACATATCGTTCGTACTCTCTCGGGTTTTTGTCGCGCCTGCAGGGTTGAAATATG  GWNJ-0957:355:GW1901141748:2:1101:11160:10451
27  AGAGAGAGAGAGAGAGAGAGAGAGAGA ATAAACGATCAGAGCATTTGCTACCTACCTATATAGCAATCGTGGACGGTGAAAGAGAGAGAGAGAGAGAGAGAGAGAGAAAAACAGAGGCCATTTCATTTGACGAGCGAAAAGCGAACGAACATATCGTTCGTACTCTCTCGGGTTTTT  GWNJ-0957:355:GW1901141748:2:1101:24383:10503
27  AGAGAGAGAGAGAGAGAGAGAGAGAGA CGTTCACTACTTCACGTTTGAAAACACAATAAACGATCAGAGCATTTGCTACCTACTTATATAGCAATCGTGGACGGTGAAAGAGAGAGAGAGAGAGAGAGAGAGAGAGAAAAACAGAGGCCATTTCATTTGACGAGCGAAAAGCGAACGAA  GWNJ-0957:355:GW1901141748:2:1101:24708:10503
```

Supplementary Figure 2.2 Sample .tsv file output for the GA-repeat counting bioinformatics pipeline. Our pipeline parses through sequence.fastq.gz files and extracts reads with the *H3/H4* promoter GA-repeat and subsequently counts the number of nucleotides that makes up the repeat. The main output file for this script is a .tsv file contain three columns. The first column specifies the number of nucleotides that make up the GA-repeat, the second column has the extracted repeat itself, and the third column with the trimmed read where the repeat originated, and the last column has the sequence ID.

2.9 References

Becker K. A., J. L. Stein, J. B. Lian, A. J. van Wijnen, and G. S. Stein, 2007 Establishment of histone gene regulation and cell cycle checkpoint control in human embryonic stem cells. Journal of Cellular Physiology 210: 517–526. https://doi.org/10.1002/jcp.20903

Bongartz P., and S. Schloissnig, 2019 Deep repeat resolution-the assembly of the Drosophila Histone Complex. Nucleic Acids Res 47: e18. https://doi.org/10.1093/nar/gky1194

Carty M., L. Zamparo, M. Sahin, A. González, R. Pelossof, *et al.*, 2017 An integrated model for detecting significant chromatin interactions from high-resolution Hi-C data. Nat Commun 8: 15454. https://doi.org/10.1038/ncomms15454

Cock P. J. A., T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, *et al.*, 2009 Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25: 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Di Caro D., R. Melfi, C. Alessandro, G. Serio, V. Di Caro, *et al.*, 2004 Down-regulation of early sea urchin histone H2A gene relies on cis regulative sequences located in the 5’ and 3’ regions and including the enhancer blocker sns. J Mol Biol 342: 1367–1377. https://doi.org/10.1016/j.jmb.2004.07.101

Duan J., L. Rieder, M. M. Colonnetta, A. Huang, M. Mckenney, *et al.*, 2021 CLAMP and Zelda function together to promote Drosophila zygotic genome activation. eLife.

Duronio R. J., and W. F. Marzluff, 2017 Coordinating cell cycle-regulated histone gene expression through assembly and function of the Histone Locus Body. RNA Biol 14: 726–738. https://doi.org/10.1080/15476286.2016.1265198

Edgar B. A., and G. Schubiger, 1986 Parameters controlling transcriptional activation during early Drosophila development. Cell 44: 871–877. https://doi.org/10.1016/0092-8674(86)90009-7

Farrell J. A., and P. H. O'Farrell, 2014 From egg to gastrula: how the cell cycle is remodeled during the Drosophila mid-blastula transition. Annu Rev Genet 48: 269–294. https://doi.org/10.1146/annurev-genet-111212-133531

Fuda N. J., M. J. Guertin, S. Sharma, C. G. Danko, A. L. Martins, *et al.*, 2015 GAGA Factor Maintains Nucleosome-Free Regions and Has a Role in RNA Polymerase II Recruitment to Promoters. PLOS Genetics 11: e1005108. https://doi.org/10.1371/journal.pgen.1005108

Gaskill M. M., T. J. Gibson, E. D. Larson, and M. M. Harrison, 2021 GAF is essential for zygotic genome activation and chromatin accessibility in the early Drosophila embryo, (Y. M. Yamashita, and K. Struhl, Eds.). eLife 10: e66668. https://doi.org/10.7554/eLife.66668

Ghule P. N., J. R. Boyd, F. Kabala, A. J. Fritz, N. A. Bouffard, *et al.*, 2023 Spatiotemporal higher-order chromatin landscape of human histone gene clusters at histone locus bodies during the cell cycle in breast cancer progression. Gene 872: 147441. https://doi.org/10.1016/j.gene.2023.147441

Günesdogan U., H. Jäckle, and A. Herzig, 2010 A genetic system to assess in vivo the functions of histones and histone modifications in higher eukaryotes. EMBO Rep 11: 772–6. https://doi.org/10.1038/embor.2010.124

Gutierrez-Perez I., M. J. Rowley, X. Lyu, V. Valadez-Graham, D. M. Vallejo, *et al.*, 2019 Ecdysone-Induced 3D Chromatin Reorganization Involves Active Enhancers Bound by Pipsqueak and Polycomb. Cell Rep 28: 2715-2727.e5. https://doi.org/10.1016/j.celrep.2019.07.096

Harrison M. M., and M. B. Eisen, 2015 Transcriptional Activation of the Zygotic Genome in Drosophila. Curr Top Dev Biol 113: 85–112. https://doi.org/10.1016/bs.ctdb.2015.07.028

Hodkinson L. J., J. Gross, C. A. Schmidt, P. P. Diaz-Saldana, T. Aoki, *et al.*, 2023 Sequence reliance of a Drosophila context-dependent transcription factor. 2023.12.07.570650.

Horton C. A., A. M. Alexandari, M. G. B. Hayes, E. Marklund, J. M. Schaepe, *et al.*, 2022 Short tandem repeats bind transcription factors to tune eukaryotic gene expression. 2022.05.24.493321.

Kaye E. G., M. Booker, J. V. Kurland, A. E. Conicella, N. L. Fawzi, *et al.*, 2018 Differential Occupancy of Two GA-Binding Proteins Promotes Targeting of the Drosophila Dosage Compensation Complex to the Male X Chromosome. Cell Rep 22: 3227–3239. https://doi.org/10.1016/j.celrep.2018.02.098

Koreski K. P., L. E. Rieder, L. M. McLain, A. Chaubal, W. F. Marzluff, *et al.*, 2020 Drosophila histone locus body assembly and function involves multiple interactions. Mol Biol Cell 31: 1525–1537. https://doi.org/10.1091/mbc.E20-03-0176

Langmead B., and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357–359. https://doi.org/10.1038/nmeth.1923

Lehmann M., T. Siegmund, K. G. Lintermann, and G. Korge, 1998 The pipsqueak protein of Drosophila melanogaster binds to GAGA sequences through a novel DNA-binding domain. J Biol Chem 273: 28504–28509. https://doi.org/10.1074/jbc.273.43.28504

Lehmann M., 2004 Anything else but GAGA: a nonhistone protein complex reshapes chromatin structure. Trends in Genetics 20: 15–22. https://doi.org/10.1016/j.tig.2003.11.005

Ma T., B. A. Van Tine, Y. Wei, M. D. Garrett, D. Nelson, *et al.*, 2000 Cell cycle-regulated phosphorylation of p220(NPAT) by cyclin E/Cdk2 in Cajal bodies promotes histone gene transcription. Genes Dev 14: 2298–2313. https://doi.org/10.1101/gad.829500

Marzluff W. F., S. Sakallah, and H. Kelkar, 2006 The sea urchin histone gene complement. Developmental Biology 300: 308–320. https://doi.org/10.1016/j.ydbio.2006.08.067

McKay D. J., S. Klusza, T. J. Penke, M. P. Meers, K. P. Curry, *et al.*, 2015 Interrogating the function of metazoan histones using engineered gene clusters. Dev Cell 32: 373–86. https://doi.org/10.1016/j.devcel.2014.12.025

Notredame C., D. G. Higgins, and J. Heringa, 2000 T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302: 205–217. https://doi.org/10.1006/jmbi.2000.4042

Rieder L. E., K. P. Koreski, K. A. Boltz, G. Kuzu, J. A. Urban, *et al.*, 2017 Histone locus regulation by the Drosophila dosage compensation adaptor protein CLAMP. Genes Dev 31: 1494–1508. https://doi.org/10.1101/gad.300855.117

Robinson J. T., H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, *et al.*, 2011 Integrative Genomics Viewer. Nat Biotechnol 29: 24–26. https://doi.org/10.1038/nbt.1754

Russo C. A., N. Takezaki, and M. Nei, 1995 Molecular phylogeny and divergence times of drosophilid species. Mol Biol Evol 12: 391–404. https://doi.org/10.1093/oxfordjournals.molbev.a040214

Salzler H. R., D. C. Tatomer, P. Y. Malek, S. L. McDaniel, A. N. Orlando, *et al.*, 2013 A sequence in the Drosophila H3-H4 Promoter triggers histone locus body assembly and biosynthesis of replication-coupled histone mRNAs. Dev Cell 24: 623–34. https://doi.org/10.1016/j.devcel.2013.02.014

Schienman J. E., E. R. Lozovskaya, and L. D. Strausbaugh, 1998 Drosophila virilis has atypical kinds and arrangements of histone repeats. Chromosoma 107: 529–539. https://doi.org/10.1007/s004120050339

Schweinsberg S., K. Hagstrom, D. Gohl, P. Schedl, R. P. Kumar, *et al.*, 2004 The Enhancer-Blocking Activity of the Fab-7 Boundary From the Drosophila Bithorax Complex Requires GAGA-Factor-Binding Sites. Genetics 168: 1371–1384. https://doi.org/10.1534/genetics.104.029561

Seal R. L., P. Denny, E. A. Bruford, A. K. Gribkova, D. Landsman, *et al.*, 2022 A standardized nomenclature for mammalian histone genes. Epigenetics & Chromatin 15: 1–18. https://doi.org/10.1186/s13072-022-00467-2

Shopland L. S., M. Byron, J. L. Stein, J. B. Lian, G. S. Stein, *et al.*, 2001 Replication-dependent histone gene expression is related to Cajal body (CB) association but does not require sustained CB contact. Mol Biol Cell 12: 565–576. https://doi.org/10.1091/mbc.12.3.565

Smith G. P., 1976 Evolution of Repeated DNA Sequences by Unequal Crossover. Science 191: 528–535. https://doi.org/10.1126/science.1251186

Srivastava S., D. Puri, H. S. Garapati, J. Dhawan, and R. K. Mishra, 2013 Vertebrate GAGA factor associated insulator elements demarcate homeotic genes in the HOX clusters. Epigenetics & Chromatin 6: 8. https://doi.org/10.1186/1756-8935-6-8

Steensel B. van, J. Delrow, and H. J. Bussemaker, 2003 Genomewide analysis of Drosophila GAGA factor target genes reveals context-dependent DNA binding. Proc Natl Acad Sci U S A 100: 2580–2585. https://doi.org/10.1073/pnas.0438000100

Tadros W., and H. D. Lipshitz, 2009 The maternal-to-zygotic transition: a play in two acts. Development 136: 3033–42. https://doi.org/10.1242/dev.033183

Takayama Y., and K. Takahashi, 2007 Differential regulation of repeated histone genes during the fission yeast cell cycle. Nucleic Acids Res 35: 3223–3237. https://doi.org/10.1093/nar/gkm213

Terzo E. A., S. M. Lyons, J. S. Poulton, B. R. S. Temple, W. F. Marzluff, *et al.*, 2015 Distinct self-interaction domains promote Multi Sex Combs accumulation in and formation of the Drosophila histone locus body. Mol Biol Cell 26: 1559–1574. https://doi.org/10.1091/mbc.E14-10-1445

The Galaxy Community, 2022 The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. Nucleic Acids Research 50: W345–W351. https://doi.org/10.1093/nar/gkac247

Valipour E., A. Kowsari, H. Bayat, M. Banan, S. Kazeminasab, *et al.*, 2013 Polymorphic core promoter GA-repeats alter gene expression of the early embryonic developmental genes. Gene 531: 175–179. https://doi.org/10.1016/j.gene.2013.09.032

Waterhouse A. M., J. B. Procter, D. M. A. Martin, M. Clamp, and G. J. Barton, 2009 Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189–1191. https://doi.org/10.1093/bioinformatics/btp033

Chapter 3

Sequence reliance of a *Drosophila* context-dependent transcription factor

Reproduced with permission from

Lauren J. Hodkinson[1], Julia Gross[2], Casey A. Schmidt[3], Pamela P. Diaz-Saldana[3], Tsutomo Aoki[4], Leila E. Rieder[3#] 2023 Sequence reliance of a *Drosophila* context-dependent transcription factor. 2023.12.07.570650. *bioRxiv*

[1] Genetics and Molecular Biology Graduate Program, Emory University, Atlanta, GA 30322 USA

[2] Immunology and Molecular Pathogenesis Graduate Program, Emory University, Atlanta, GA 30322 USA

[3] Department of Biology Emory University, Atlanta, GA 30322 USA

[4] Department of Molecular Biology, Princeton University, Princeton, NJ 08540 USA

[#] Corresponding author: lrieder@emory.edu

3.1 Abstract

Despite binding similar *cis* elements, transcription factors often perform context-dependent functions at different genomic loci. How transcription factors integrate *cis* sequence and genomic context to perform their context-dependent functions is still poorly understood. One example of a context-dependent transcription factor is the *Drosophila* protein Chromatin-Linked Adapter for MSL Proteins (CLAMP), which targets similar GA-rich cis elements on the X-chromosome and at the histone locus but recruits very different, locus-specific transcription factors to each of these contexts. Here we investigate how CLAMP function at the histone genes is impacted by the identity of its *cis* binding elements. CLAMP binds a long GA-repeat element in the promoter of *H3* and *H4* (*H3/H4*p) and recruits histone locus body (HLB) factors needed for histone gene expression. We engineered flies to carry a transgenic histone gene array in which we replaced the *H3/H4*p cis elements with X-linked CLAMP-recruiting GA-rich elements. We discovered that X-linked CLAMP elements do not functionally substitute for GA-repeats in the histone gene array and do not recruit the core HLB factor, indicating that *cis* element sequence is critical. Sufficient X-linked sequence, in the context of the *H3/H4*p results in X-chromosome-specific factor recruitment in males, indicating the importance of local context. Our observations suggest that both sequence and local context dictate CLAMP function.

3.2 Introduction

Coordinated gene expression is a crucial but difficult task in the crowded nucleus. To accomplish this feat, transcription factors (TFs) must first traverse the nucleus to find their corresponding *cis* elements. Furthermore, once factors have identified their DNA-binding sites, they can then impact gene expression on highly constrained temporal and spatial levels. When gene expression programs are misregulated or interrupted by mutations in regulatory elements, it can have catastrophic impacts, causing a variety of disease outcomes including cancer, autoimmunity, and neurological disorders (Lee and Young 2013). To further complicate the hurdle of widespread gene regulation, some TFs are "context dependent"; they bind similar *cis* elements across the genome but retain the ability to perform distinct functions at these different loci (Fry and Farnham 1999). Currently, we still do not fully understand how context-dependent TFs integrate locational information with cues they may receive from cofactors, 3D architecture, and other signals to perform their diverse functions.

CLAMP (Chromatin Linked Adaptor for MSL Proteins) is *Diptera*-specific C2H2 zinc-finger TF that plays a genome-wide role as a pioneer factor (Duan et al. 2021) and targets GA-rich *cis* elements to regulate global gene expression through both chromatin accessibility changes and polymerase pausing (Urban et al. 2017b, 2017a). CLAMP is designated as a context-dependent transcription factor that is critical for two major coordinated gene expression events: histone gene expression and upregulation of the male X chromosome for dosage compensation (Soruco and Larschan 2014). CLAMP is enriched on the male X-chromosome where it binds to GA-rich regions often overlapping with MREs (MSL recognition elements) (Alekseyenko et al. 2008) and recruits the Male Specific Lethal complex (MSLc) to accomplish dosage compensation in male flies (Soruco and Larschan 2014). At the histone gene locus, CLAMP binds a long GA-repeat *cis* element in the promoter of *H3* and *H4* (*H3/H4*p) and fosters recruitment of histone gene locus body (HLB) specific factors including Mxc (Multi Sex combs, Mxc; the *Drosophila* ortholog of the human NPAT; (Terzo et al. 2015)) (Salzler et al. 2013; Rieder et al. 2017). CLAMP targets both the histone genes and the X chromosome prior to locus-specific factors such as Mxc and MSLc, neither of which have strong DNA-binding capability (Villa et al. 2016; Terzo et al. 2015). These observations show that, despite binding similar *cis* elements on the X-chromosome and at the histone gene locus, CLAMP recruits different factors to each location ensuring proper group composition at each genomic location. It is unclear how early transcription factors such as CLAMP integrate information from GA-rich *cis* elements with other genomic contextual information to perform their locus-specific functions.

Drosophila dosage compensation provides some clues as to how TFs like CLAMP integrate sequence and context information. For example, moving X-linked chromosome entry sites (CES; up to ~1500bp which contain one or more GA-rich MREs (Alekseyenko et al. 2008)) to autosomal locations leads to ectopic MSLc recruitment, spreading of the complex into surrounding chromatin, and transcriptional regulation of nearby genes (Gorchakov et al. 2009). A similar phenomenon occurs when the ~300bp *H3/H4*p, which includes a GA-repeat ranging from 16-35 bps (Bongartz and Schloissnig 2018) targeted by CLAMP (Salzler et al. 2013; Rieder et al. 2017). When this segment is placed outside the endogenous histone gene locus on chromosome 2L, HLB-specific factors are recruited to the transgenes resulting in transcription of the adjacent sequences (Salzler et al. 2013). These important experiments show that separation of local contexts (CES, up to ~1500bp; *H3/H3*p, ~300bp) from the larger locus (X-chromosome, histone gene locus on chromosome 2L) still allows for retention of local context function. This retention of function suggests that the larger chromosomal or locus context of these elements is not required for their function in recruiting the factors necessary for coordinated gene expression. Since both regions carry elements that recruit the CLAMP protein (Alekseyenko et al. 2008; Rieder et al. 2017), we hypothesized that the GA *cis* elements themselves are interchangeable and that the flanking local context provides the cues required for CLAMP context-specific function and unique factor recruitment.

Neither the *Drosophila* X-chromosome nor the endogenous histone gene locus are tractable study systems in which to test this hypothesis. MSLc coats the entire chromosome, and it is not practical to manipulate each GA-rich MRE in all 150 CES (Alekseyenko et al. 2008). Altering a critical number of CES, besides being nearly impossible to execute, would likely cause incomplete dosage compensation and male-specific lethality, while altering just a few CES is unlikely to significantly affect MSL recruitment to the whole chromosome due to complex spreading (Kelley et al. 1999; Kageyama et al. 2001; Gorchakov et al. 2009) . Similarly, the endogenous histone gene locus is organized in a series of ~100 tandemly repeated 5 Kb arrays, in which each array contains the five canonical histone genes (*H3*, *H4*, *H2A*, *H2B*, and *H1*).

The repetitive nature of the histone locus renders it intractable for genetic manipulation as it harbors CLAMP-binding GA-repeats in all ~100 *H3/H4*p. However, the *Drosophila* transgenic histone gene array provides an excellent genetic system in which to test our hypothesis. Transgenes carrying 1-12 histone gene arrays have been established that allow for genetic manipulation and recapitulate histone locus functionality (Salzler et al. 2013; McKay et al. 2015; Meers et al. 2018). A single copy histone gene array, while not able to rescue an endogenous histone locus deletion background, successfully recruits HLB-specific factors and drive histone gene expression (Koreski et al. 2020). This becomes a powerful system to perturb the CLAMP-recruiting *cis* elements within the array without editing all ~100 arrays at the endogenous locus.

Leveraging the transgenic system, we confirm that *H3/H4*p GA-repeat CLAMP binding sites are required for Mxc recruitment to the transgene (Rieder et al. 2017). We further demonstrate that transgenes in which we replace the *H3/H4*p GA-repeats with X-linked GA-rich MREs, which are bound by CLAMP in vitro and in their native locations, fail to recruit Mxc despite the larger contextual information of the histone gene array. Finally, we demonstrate that adding back additional X-linked sequence to the transgenic histone gene array results in MSLc recruitment in males. We observe sex-specific differences that suggest a competition between CLAMP-associated factors in males, but not in females. Overall, our observations indicate that *cis* element sequence alone is enough to impact context-dependent TF functions.

3.3 Materials and methods

Transgenics

Transgenes were constructed to include a 5 Kb histone array sequence consisting of the 5 replication-dependent histone genes and their relative promoters (McKay et al. 2015; Meers et al. 2018) in which the H4 and H2A genes are FLAG-tagged (24 bp) at the 5' end to distinguish them from the endogenous histone genes. 500 bp DNA inserts containing the *H3/H4*p changes of interest were ordered from IDT and inserted via Gibson cloning (detailed above). All 1x histone array transgenes were inserted at the VK33 attP site on chromosome 3L (65B2) (Venken et al. 2006) using PhiC3-mediated integration (Groth et al. 2004) by GenetiVision (Houston, TX). Injected chimeric flies were crossed in pairs to a +/+; CyO/If ; TM3 (Sb) / TM6 (Tb) balancer stock. Resulting red-eyed progeny were selected and singly mated back to +/+; CyO/If ; TM3

(Sb) / TM6 (Tb) flies and a homozygous stock was established. Flies were maintained on standard cornmeal/molasses media at 18º C and transferred every 3-4 weeks.

Electrophoretic Mobility Shift Assays:

We performed EMSAs after Aoki et al. 2008 (Aoki et al. 2008) with minor modifications.

DNA probes: We made EMSA probes using PCR from gblocks (IDT) acquired during cloning as templates and the following primers: H3/H4 promoter F1: CACAGCACGAAAGTCACTAAAGAAC, H3/H4 promoter R1: GTTTGAAAACACAATAAACGATCAGAGC. We 5′ end labeled one pmol of probe with γ-32P-ATP (MP Biomedicals) using T4 polynucleotide kinase (New England BioLabs) in a 50 µl total reaction volume at 37°C for 1 hour. We used Sephadex G-50 fine gel (Amersham Biosciences) columns to separate free ATP from labeled probes. We adjusted the volume of the eluted sample to 100 µl using deionized water so that the final concentration of the probe was 10 fmol/µl.

Late embryo nuclear extracts: We prepared embryo extracts from 6-18hr Oregon R embryos collected on apple juice plates and aged 6 hours at room temperature. We performed nuclear extract preparation as in Aoki et al. 2008. We omitted the final dialysis step described in Aoki et al. and completed the extraction with the final concentration of KCl at 360 mM.

Shifts: We performed 20 µl binding reactions consisting of 0.5 µl (5 fmol) of labeled probe in the following buffer: 25 mM Tris-Cl (pH 7.4), 100 mM KCl, 1 mM EDTA, 0.1 mM dithiothreitol, 0.1 mM PMSF, 0.3 mg/ml bovine serum albumin, 10% glycerol, 0.25 mg/ml poly(dI-dC)/poly(dI-dC). We added 1 µl of nuclear extract and incubated samples at room temperature for 30 minutes. We loaded samples onto a 4% acrylamide (mono/bis, 29:1)-0.5× TBE-2.5%

glycerol slab gel. We performed electrophoresis at 4°C, 180 V for 3-4 hours using 0.5× TBE-2.5% glycerol as a running buffer. We dried gels and imaged using a Typhoon 9410 scanner and Image Gauge software.

Supershifts: We preincubated reactions, including 5ug/ul poly(dA-dT)/poly(dA-dT), with OreR late nuclear extract (LNE) and antibodies for 30 min at room temperature before adding hot probe. We used 1 ul rabbit serum and 4 ul anti-CLAMP antibodies.

Probe (length)	**Probe Sequence**
WT (226 bp) GA-repeats bolded	CACAGCACGAAAGTCACTAAAGAACTAATTTCAACGTTTCTGTGTGCCCCTATTTATAGGTAAAACGACAAAAACCC**GAGAGAG**TACGAACGATATGTTCGTTCGCTTTTCGCTCGTCAAATGAAATGGCCTCTGTTTT**TCTCTCTCTCTCTCTCTCTCT**TTCACCGTCCACGATTGCTATATAAGTAGGTAGCAAATGCTCTGATCGTTTATTGTGTTTTCAAAC
GA del (198 bp)	CACAGCACGAAAGTCACTAAAGAACTAATTTCAACGTTTCTGTGTGCCCCTATTTATAGGTAAAACGACAAAAACCC**X**TACGAACGATATGTTCGTTCGCTTTTCGCTCGTCAAATGAAATGGCCTCTGTTTT**X**TTCACCGTCCACGATTGCTATATAAGTAGGTAGCAAATGCTCTGATCGTTTATTGTGTTTTCAAAC
2 MRE (238 bp) MREs bolded	CACAGCACGAAAGTCACTAAAGAACTAATTTCAACGTTTCTGTGTGCCCCTATTTATAGGTAAAACGACAAAAACCC**gatttagagcgagatgacaa**TACGAACGATATGTTCGTTCGCTTTTCGCTCGTCAAATGAAATGGCCTCTGTTTT**ggcgatctctctcgtatacg**TTCACCGTCCACGATTGCTATATAAGTAGGTAGCAAATGCTCTGATCGTTTATTGTGTTTTCAAACG
CES5C2 (232 bp) MREs bolded	CACAGCACGAAAGTCACTAAAGAACTAATTTCAACGTTTCTGTGTGCCCCTATTTATATTTATAGGTAAAACGaaatcacgttc**acacaacttagaaagagatagcgatg**gcgg**tgtgaaagagagcgagatagttgga**agctt**catggaaatgaaagagaggtagttt**ttggaaatgaATTGCTATATAAGTAGGTAGCAAATGCTCTGATCGTTTATTGTGTTTTCAAAC
roX2 (232 bp) MREs bolded	CACAGCACGAAAGTCACTAAAGAACTAATTTCAACGTTTCTGTGTGCCCCTATTTATATTTATAGGaatacagatc**gatttagagcgagatgacaa**tagaga**ggcgatctctctcgtatacg**agtct**tgaaaagaaagagaaggcga**acggtgct**ggcttagagagagatggcaa**tactaattaacTATAAGTAGGTAGCAAATGCTCTGATCGTTTATTGTGTTTTCAAAC

Table 3.1: Sequences of EMSA probes.

Polytene immunofluorescence and microscopy

We performed polytene chromosome squashes from salivary glands of sexed female and male third instar larvae. We passed glands through fix 1 (4% formaldehyde, 1% Triton X-100, in 1× PBS) for 1 min, fix 2 (4% formaldehyde, 50% glacial acetic acid) for 2 min, and 1:2:3 solution (ratio of lactic acid:water:glacial acetic acid) for 5 min prior to squashing and spreading. Slides were washed in 1X PBS, washed in 1% Triton X-100 (diluted in 1X PBS) and blocked for one hour in .5% BSA in 1X PBS. Slides were then incubated with primary antibody diluted in blocking solution (antibody specifics below) overnight at 4º C in a dark humid chamber. Slides were then washed in 1 X PBS and incubated with secondary antibody diluted in blocking solution (antibody specifics below) for two hours. Slides were then mounted using Prolong Diamond anti-fade reagent with DAPI (Thermo Fisher, P36961), and spreads were imaged on a using a Zeiss Scope.A1 equipped with a Zeiss AxioCam using a 40×/0.75 plan neofluar objective using AxioVision software. We used primary antibodies at the following concentrations: rabbit anti-CLAMP (1:1000; Novus/SDIX) (Larschan et al. 2012), guinea pig anti-Mxc (1:5000) (White et al. 2011, gift from the Duronio and Marzluff labs), and rabbit anti-MSL2 (1:150, a gift from the Meller Lab). We used AlexaFluor secondary 488 and 647 antibodies (Thermo Fisher Scientific) at a concentration of 1:1000.

ChIP-seq data

CLAMP ChIP-seq was performed using protocol from Rieder et al. 2017 (Rieder et al. 2017) and data was taken from NCBI GEO (GSE152598, (Duan et al. 2021)). MSL3 ChIP-seq was performed using protocol from Rieder et al, 2017 and data was taken from NCBI GEO

(GSE133637, (Rieder et al. 2019)). Mxc ChIP-seq was performed as in Rieder et al. 2017 using 2 ul guinea pig anti-Mxc antibody (gift from the Duronio/Marzluff laboratories).

Data availability: Mxc ChIP-seq data are deposited on NCBI GEO (GES249797).

Bioinformatic analysis, alignment and visualization

We performed bioinformatics analysis as in Hodkinson *et al.* 2023 (Hodkinson et al. 2023). We directly imported individual FASTQ datasets into the web-based platform Galaxy (The Galaxy Community 2022) through the NCBI SRA Run Selector or using our sequencing files by selecting the desired runs and utilizing the computing Galaxy download feature. We retrieved the FASTQ files from SRA using the "Faster Download and Extract Reads in FASTQ format from NCBI SRA" Galaxy command. Because the ~100 histone gene arrays are extremely similar in sequence (Bongartz and Schloissnig 2018), we do not utilize the dm6 or dm3 genomes and instead can collapse ChIP-seq data onto a single histone array (McKay et al. 2015; Bongartz and Schloissnig 2018; Koreski et al. 2020). We used a custom "genome" that includes a single Drosophila melanogaster histone array similar to that in Mckay et al. 2015, which we directly uploaded to Galaxy using the "upload data" feature, and normalized using the Galaxy command "NormalizeFasta" specifying an 80 bp line length for the output .fasta file. We aligned ChIP reads to the normalized histone gene array using Bowtie2 (Langmead and Salzberg 2012) to create .bam files using the user built-in index and "very sensitive end-to-end" parameter settings. We converted the .bam files to .bigwig files using the "bamCoverage" Galaxy command in which we set the bin size to 1 bp and set the effective genome size to user specified: 5000 bp (approximate size of l histone array). We also mapped relevant input or IgG datasets. If an input dataset was available, we normalized ChIP datasets to input using the "bamCompare" Galaxy command in which we set the bin size to 1 bp. We visualized the bigwig files using the Integrative Genome Viewer (IGV) (Robinson et al. 2011).

3.4 Results and Discussion

3.4.1 CLAMP targets GA-rich *cis* elements at different loci

CLAMP binds genome-wide and is enriched on the X-chromosome as well as on chromosome 2L at the endogenous histone locus (**Figure 3.1A**). We mapped available CLAMP ChIP-seq data from 2-4 hr mixed embryos (Duan et al. 2021) and confirmed that CLAMP targets the GA-repeats in the *H3/H4*p (**Figure 3.1B**). Mxc is specific to the histone locus by immunofluorescence (White et al. 2011) and the mammalian homolog NPAT also only targets histone promoters in cultured human cells (Kaya-Okur et al. 2019). We confirmed that Mxc is specific to the histone genes by performing ChIP-seq from embryo samples and by performing sexed polytene chromosome immunofluorescence using antibodies against Mxc, CLAMP, and MSL3. Mxc targets only the histone locus on chromosome 2L (**Figure 3.1A, D**) and is broadly localized over histone gene promoters, overlapping with CLAMP signal at the *H3/H4p* (**Figure 3.1B**). We also mapped existing MSL3 (Male Specific Lethal 3; a component of MSLc) ChIP-seq data from embryos (Rieder et al. 2019) and confirmed that MSLc is enriched on the X-chromosome (**Figure 3.1A, D**). As expected, MSLc is not enriched at the autosomal histone locus (**Figures 3.1A-B, D**). Both CLAMP and MSL are enriched at CESs, including CES5C2 and the *roX2* (*RNA on X 2*; a lncRNA component of MSLc) CES (**Figure 3.1C).** Although CLAMP is present at both X-linked CES and the GA-repeats within the context of the histone gene array, MSLc and Mxc are locus-specific.

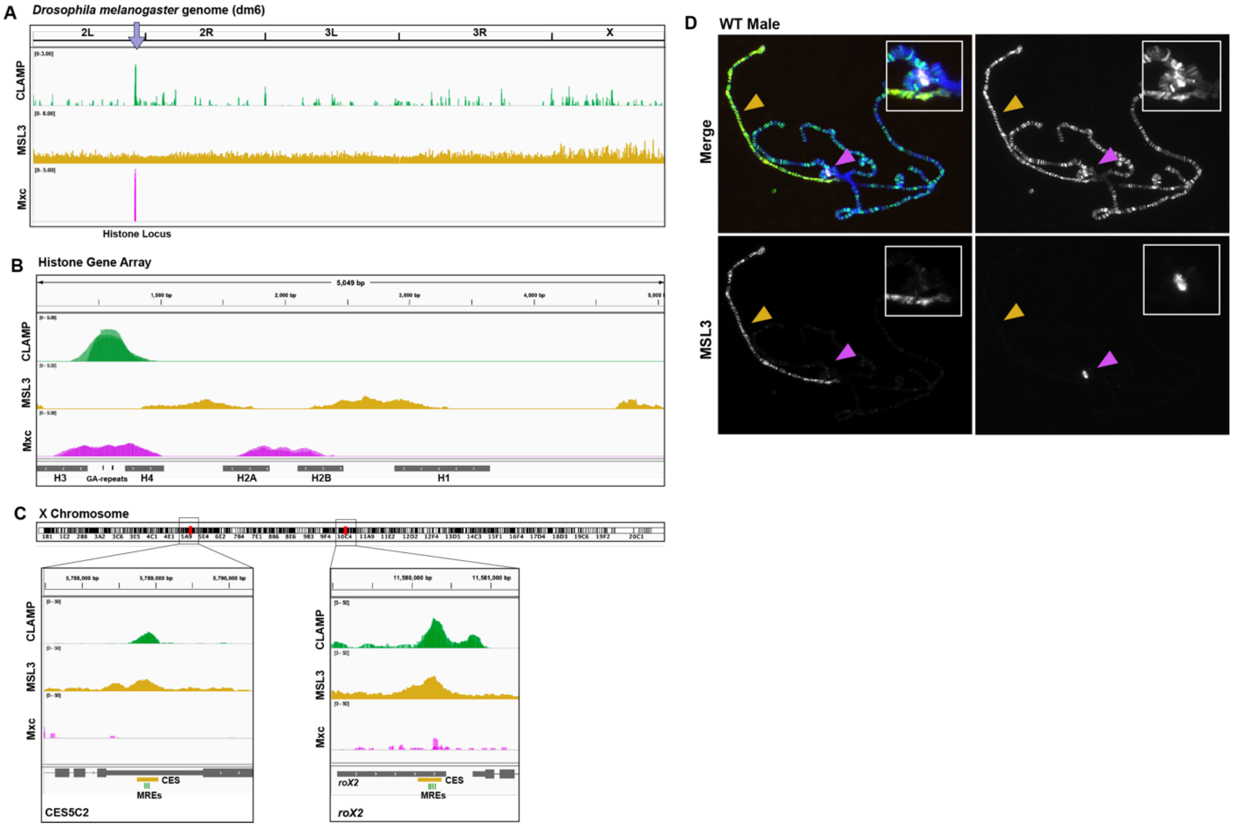


Figure 3.1: CLAMP binds locations genome-wide while Mxc and MSL3 bind distinct genomic regions. **(A)** We mapped ChIP-seq data for CLAMP (green) from 2-4hr mixed sex embryos (Duan *et al.* 2021, GSE152598, three overlaid replicates normalized to respective inputs), MSL3 (yellow) from 2-4hr mixed embryos ((Rieder et al. 2019), GSE133637, normalized to the input), and Mxc (magenta) from 2-4 hr female staged embryos (three overlaid replicates normalized to respective inputs) to the dm6 (*Drosophila melanogaster)* genome. The purple arrow indicates the location of the endogenous histone gene locus. **(B)** CLAMP, MSL3, and Mxc ChIP-seq data mapped to a custom single histone gene array. The dark green bars between the *H3* and *H4* coding sequences mark the CLAMP binding GA-repeats. **(C)** ChIP peaks at two X-chromosome locations, CES5C2 and the *roX2* gene. The yellow bars mark the known chromosome entry sites (CES) and the light green bars mark the GA-rich MSL

recognition elements (MREs) where CLAMP and MSLc colocalize. **(D)** We performed immunofluorescence staining of wildtype chromosomal male third instar larval polytene chromosomes for CLAMP (green), MSL3 (yellow) and Mxc (magenta). DNA is stained with DAPI (blue). CLAMP and Mxc colocalize at the endogenous histone locus (magenta arrow, solid outlined box) and CLAMP and MSL3 colocalize on the X-chromosome (yellow arrow). Mxc and MSL3 do not colocalize.

3.4.2 CLAMP requires GA-rich sequences to bind *in vitro* and *in vivo*

CLAMP is a zinc-finger protein that directly interacts with DNA sequence (Soruco and Larschan 2014). Recombinant full-length CLAMP binds to GA-repeat carrying DNA probes *in vitro* (Duan et al. 2021). We therefore investigated the ability of recombinant CLAMP to interact with different *cis* element and whether the GA-repeats were critical for CLAMP binding *in vitro*. We first designed two biotin-labeled DNA probes based on the sequence from the endogenous *H3/H4*p. The "WT" probe includes the endogenous *H3/H4*p with intact CLAMP-recruiting GA-repeat. The "GA delete" probe includes the same *H3/H4*p sequence except the GA-repeats are removed (**Figure 3.2A**). CLAMP is maternally deposited in the early embryo and is a known member of the Late Boundary Complex which shifts CES probes *in vitro* (Kuzu et al. 2016; Kaye et al. 2017). We therefore also performed radiolabeled EMSAs using embryo extracts, which should include both maternally deposited CLAMP as well as the CLAMP-containing LBC, and our probes. To confirm the embryo extract included CLAMP, we performed shifts with extract, probe, and CLAMP antibody and showed that CLAMP antibody super-shifted with the WT WT probe (**Supplemental Figure 3.1)**. Only the WT probe containing the GA-repeats is shifted with embryo extract; the GA delete probe did not shift (**Figure 3.4B**). We repeated these EMSAs with recombinant full-length CLAMP protein (Duan et al. 2021) with the same result

(**Figure 3.2B**). These observations confirm that the endogenous *H3/H4*p GA-repeats are critical for CLAMP binding *in vitro*.

In vivo, GA-repeat *cis* elements are clearly not sufficient for CLAMP and Mxc recruitment: GA-repeats exist throughout the genome and long repeats are enriched on the X-chromosome (Kuzu et al. 2016) yet Mxc is solely recruited to the histone locus (**Figure 3.1**) (Rieder et al. 2017, 2019). We therefore sought to translate our *in vitro* observations *in vivo*. A transgene carrying twelve wild-type histone gene arrays recruits HLB components, including CLAMP and Mxc, but transgenic arrays lacking the GA-repeats in the *H3/H4*p fail to attract HLB factors when the endogenous locus is present (Rieder et al. 2017; Koreski et al. 2020). To validate these findings and to confirm these observations using a transgene carrying only a single histone gene array, we created two transgenic lines. The "WT" line carries a transgene with a single wild-type copy of the histone gene array, while the "GA deletion" line carries the same transgene lacking the GA-repeat sequences in the *H3/H4*p (**Figure 3.2A**). We then performed sexed polytene chromosome immunofluorescence using an antibody against Mxc and scored for ectopic Mxc.

We observed that both CLAMP and Mxc are recruited to the WT transgene in both chromosomal female and chromosomal male larvae (**Figure 3.2C,D,E**). The endogenous histone locus is visible near the chromocenter and serves as an internal staining control. Approximately 80% of polytenes from female larvae containing the WT transgene and 85% of polytenes from male larvae exhibited ectopic Mxc recruitment (**Figure 3.2C,D)**. In contrast, we rarely observed ectopic Mxc recruitment in larvae carrying the GA delete transgene (**Figure 3.2C,D,E**) and observed a significant difference between the polytene scoring of WT larva and GA-deletion larva (**Figure 3.2F)**. Our *in vivo* observations are therefore in agreement with our *in vitro* EMSA results and establish the transgenic system as a manipulable *in vivo* assay.

Because CLAMP is an integral transcription factor responsible for regulating the expression of essential genes such as the histone genes, it's likely that there are "backup" mechanisms for ensuring CLAMP can locate regions where GA-rich *cis* elements reside and perform its function. Recent work demonstrated that CLAMP is recruited to a transgene carrying twelve histone gene arrays in which the *H3/H4*p is replaced with the *H2A/H2B* promoter, which does not contain GA-repeats. But, the phenomenon where CLAMP and HLB factors are recruited to the transgene only occurs in the background of a endogenous ~100 copy histone locus deletion (Koreski et al. 2020). CLAMP targets the region by immunofluorescence, but does not interact with any sequence by ChIP-seq. This is still surprising, since our observations show that simply deleting the GA-repeats from the *H3/H4*p rendered CLAMP non-functional in the context of the single histone array transgene (**Figure 3.2C,D**). However, all of our observations are by necessity in the background of the endogenous histone locus, as the single histone gene array itself does not support viability (Günesdogan et al. 2010; McKay et al. 2015).

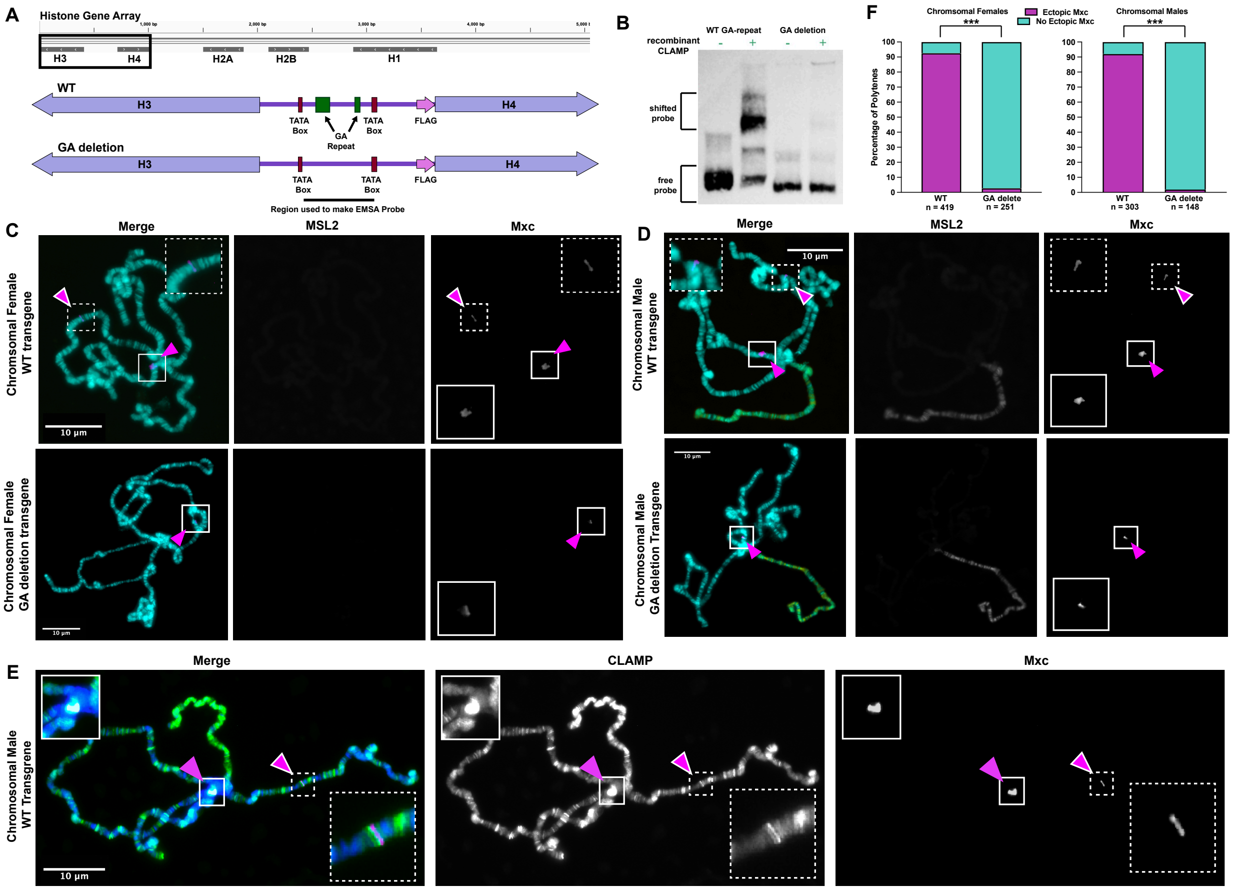


Figure 3.2: GA-repeats are required for CLAMP binding and function at the transgenic histone gene array. (A) We engineered two transgenes carrying a single histone gene array: the "WT" transgene which resembles the endogenous histone arrays and the "GA deletion" transgene in which we deleted the GA-repeats (dark green, labeled bars). **(B)** We performed EMSA (gel shift) assays with recombinant CLAMP and biotinylated probes of the *H3/H4p* sequences from both histone array transgenes (WT: 226 bp, GA deletion: 198 bp). Recombinant CLAMP shifts EMSA probes only when the GA-repeats are present. **(C)** We performed immunofluorescence staining of third instar larval polytene chromosomes in chromosomal females for Mxc (magenta). A chromosomal female carrying the WT transgene (top) shows ectopic Mxc (white outlined magenta arrow, dotted outlined box) and while a chromosomal female carrying the GA deletion transgene (bottom) shows no ectopic Mxc staining. **(D)** A

chromosomal male carrying the WT transgene (top) shows ectopic Mxc while a chromosomal male carrying the GA deletion transgene (bottom) shows no ectopic Mxc staining. Both sexes show Mxc localizing to the endogenous histone locus, which is used as an internal staining control (magenta arrow, solid outlined box). **(E)** We also performed immunofluorescence in chromosomal male and female animals for CLAMP (green) and Mxc (magenta). A chromosomal male shows colocalization of CLAMP (green) and Mxc (magenta) at the endogenous histone locus (magenta arrow, solid outlined box) and at the WT transgene (white outlined magenta arrow, dotted outlined box). Both chromosomal females and chromosomal males show Mxc localizing to the endogenous histone locus, which is used as an internal staining control (magenta arrow, solid outlined box). **(F)** Quantification of ectopic Mxc from polytene scoring shows a significant difference between the percentage of chromosome spreads that have ectopic Mxc between the WT and GA deletion transgenes in both chromosomal males and chromosomal females. n values reflect number of polytenes scored for each respective genotype. *** Chi-squared text, $p < 0.001$

3.4.3 The GA-repeats must reside in the *H3/H4*p for Mxc recruitment *in vivo*

We next sought to determine if simply the local promoter context (~300 bp) affects CLAMP recruitment to target elements. We hypothesized that the GA-repeats could be moved anywhere within the transgenic histone gene array and still attract CLAMP along with other histone locus specific factors, such as Mxc, because the larger context of the array is retained. We engineered two transgenes in which we deleted the GA-repeats from the *H3/H4*p and moved them to one of two locations within the transgenic array: either within the *H2A/H2B* promoter ("*H2A/H2B*") or within the intergenic region between the *H1* and the *H2B* coding sequences

("*H2B/H1*") (**Figure 3.3A**). In both cases, we attempted to avoid disrupting any known or predicted *cis* elements (Crayton et al. 2004). We then performed fluorescent staining on polytene chromosomes from transgenic lines with an antibody against Mxc to assess how the different GA-repeat locations impacted Mxc recruitment compared to controls (**Figure 3.3B,C)**. We rarely observed ectopic Mxc in transgenic lines in which the GA-repeats reside in the *H2A/H2B* promoter or the region between *H1 and H2B* (**Figure 3.3C,D**). Our data showed a significant difference between the percentage of chromosome spreads with ectopic Mxc for both transgenes when compared to our WT transgene and the data more closely resembled that of the GA deletion transgenes in which we completely removed the GA-repeats (**Figure 3.3D)**. Our observations show that neither the *H2A/H2B* or *H2B/H1* transgenes led to ectopic Mxc recruitment, indicating the importance of local *H3/H4*p context for Mxc recruitment. Overall, it is clear that the larger context of the histone gene array is not required for either CLAMP recruitment or specific function but that the local flanking context of the *H3/H4*p is important for CLAMP function at the histone gene array.

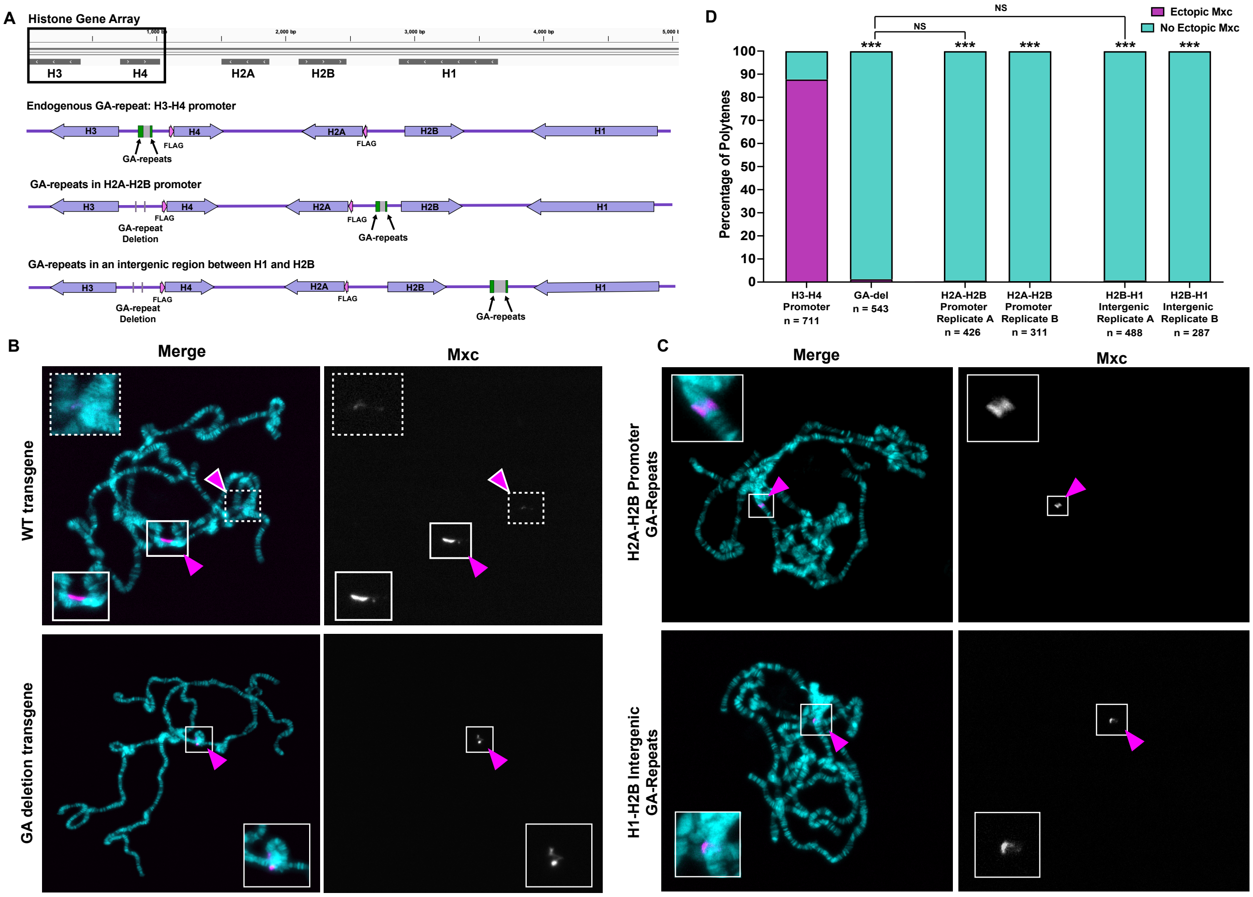


Figure 3.2: GA-repeat*s* must reside in the *H3/H4*p for proper CLAMP function at the transgenic histone gene array. (A) We engineered two histone gene array transgenes *in which* we moved the GA-repeats (green bars) to different regions along the array; one where we placed the GA-repeats in the *H2A/H2B* promoter and one where we placed the GA-repeats in the intergenic region between *H1* and *H2B*. **(B)** We performed immunofluorescence staining *of* third instar larval polytene chromosomes for Mxc (magenta). DNA is stained with DAPI (cyan). Animals carrying the WT transgene (top) show ectopic concentration of Mxc (white outlined magenta arrow, dotted outlined box) while animals carrying the GA deletion transgene (bottom) show no ectopic Mxc staining. **(C)** Animals carrying the transgenes in which the GA-repeats are moved to the *H2A/H2B* promoter (top) or intergenic region between *H1* and *H2B* (bottom) do not show ectopic Mxc. All animals show Mxc localizing to the endogenous histone locus, which

is used as an internal staining control (magenta arrow, solid outlined box). **(D)** Quantification of ectopic Mxc from polytene scoring shows a significant difference in the percentage of chromosome spreads that have ectopic Mxc between WT and the transgenes in which the GA-repeats are moved. Significance above the bars represent the comparison to the WT and the lines with represent direct comparisons between genotypes. n values reflect number of polytenes scored for each respective genotype. Chi-squared test, *** = $p < 0.001$.

Our results were somewhat surprising since other well-studied transcription factors rely exclusively on *cis* element sequence, rather than local context for function. For example, the *Drosophila* pioneer factor Zelda targets “TAGteam” sequences. Zelda competes with another TF, Grainyhead, for binding TAGteam *cis* elements and differences in the motif sequence elicits differential binding and function of Zelda (Harrison et al. 2010; Li and Eisen 2018). Because of this relationship between Zelda and specific *cis* element sequence, specific TAGteam sequences can be placed in combination in transgenes to titrate gene expression output (Li and Eisen 2018). Another example is the early *Drosophila* embryo factor Twist which binds a variety of *cis* element motifs. Twist targets several repetitive *cis* elements as well as E-box motifs, the sequences of which are not interchangeable: each E-box type corresponds to a discrete regulatory role (Ozdemir et al. 2011). However, unlike Zelda and Twist, CLAMP appears to glean significant information from the flanking local context wherein *cis* elements reside.

3.4.4 CLAMP binding sequences from different loci do not functionally substitute in the context of the histone array in chromosomal females

The presence of CLAMP-recruiting elements within the local context of the *H3/H4*p is necessary for ectopic Mxc recruitment. Prior work demonstrated that the *H3/H4*p alone is sufficient to recruit HLB factors and even to initiate histone transcription, indicating that the rest of the array is dispensable at least for these actions (Salzler et al. 2013). However, when CLAMP is tethered to the *H3/H4*p in the absence of the GA-repeats, Mxc and other HLB factors are recruited, but transgenic histone transcription is not initiated (Rieder et al. 2017). Therefore, the specific CLAMP-GA-repeat interaction is critical for full CLAMP function at the histone genes. That being considered, CLAMP may integrate other local information that is unique to the large, repetitive endogenous histone locus. For example, CLAMP-binding sites exist in each of the ~100 histone arrays (Bongartz and Schlossnig 2019), but it is unclear if CLAMP binds to all of these sites. Concentrating HLB factors at the locus likely contributes to important body properties such as phase separation (Hur et al. 2020) and facilitates histone biogenesis (Tatomer et al. 2016). The repetitive nature of the locus also likely leads to unique three-dimensional organization, both within and between loci (Carty et al. 2017; Fritz et al. 2018), which can impact the function of transcription factors such as CLAMP. Since these higher-order organizational aspects are not recapitulated at our transgenic histone gene arrays, we are unable to capture how they contribute to its contest-specific functions.

Given that CLAMP targets many GA-rich elements across the genome and that CLAMP may integrate local genomic context information to perform its functions, we next sought to investigate whether X-linked sequences can functionally substitute for the endogenous GA-repeats in the *H3/H4*p *in vitro*. We designed three hybrid DNA probes based on the sequence of the *H3/H4*p, but in which we replaced parts of the promoter with various amounts of X-linked sequence. In the "2 MRE" probe we replaced the two endogenous CLAMP recruiting GA-repeats with two 21 bp GA-rich MREs from the X-linked *roX2* gene. The "CES5C2" probe replaces the sequence between the *H3* and *H4* TATA boxes with sequence from the X-linked CES5C2 region containing three 21 bp GA-rich MREs. Finally, the "*roX2*" probe replaces the sequence between the *H3* and *H4* TATA boxes with sequence from the X-linked *roX2* CES containing four 21 bp GA-rich MREs (**Figure 3.4A**). We performed radiolabeled EMSAs using embryo extracts and all three hybrid probes robustly shifted (**Figure 3.4B**). CLAMP therefore binds X-linked GA-rich elements in the context of the *H3/H4*p *in vitro,* and this interaction appears similar to CLAMP binding of the endogenous GA-repeats.

To translate our results *in vivo*, we engineered transgenic *Drosophila* lines carrying these hybrid transgenes with the same sequences as our EMSA probes (**Figure 3.4A**) and performed polytene chromosome immunostaining as above. We observed very little ectopic Mxc in all chromosomal female larvae regardless of which hybrid transgene they carried. Animals carrying the 2 MRE transgene were significantly less likely to show ectopic Mxc than the WT control animals, but significantly more likely to show ectopic Mxc compared to animals carrying the CES5C2 and *roX2* transgenes (**Figure 3.4C,D)**. Together these data suggest that replacing just the *H3/H4*p GA-repeats impacts Mxc recruitment, in chromosomal females even though the majority of the *H3/H4*p sequence is preserved and CLAMP binds to this sequence *in vitro* (**Figure 3.4B**). Overall, our data from chromosomal females suggests that *cis* element sequence impacts CLAMP function at the transgenic histone gene array. Furthermore, in combination with our data showing the GA-repeats must reside in the *H3/H4*p for Mxc recruitment to the histone array transgene (**Figure 3)**, our results suggest that CLAMP is using cues from both *cis* element sequence and the local flanking regions of the *cis* element to determine its function at the transgenic histone gene array.

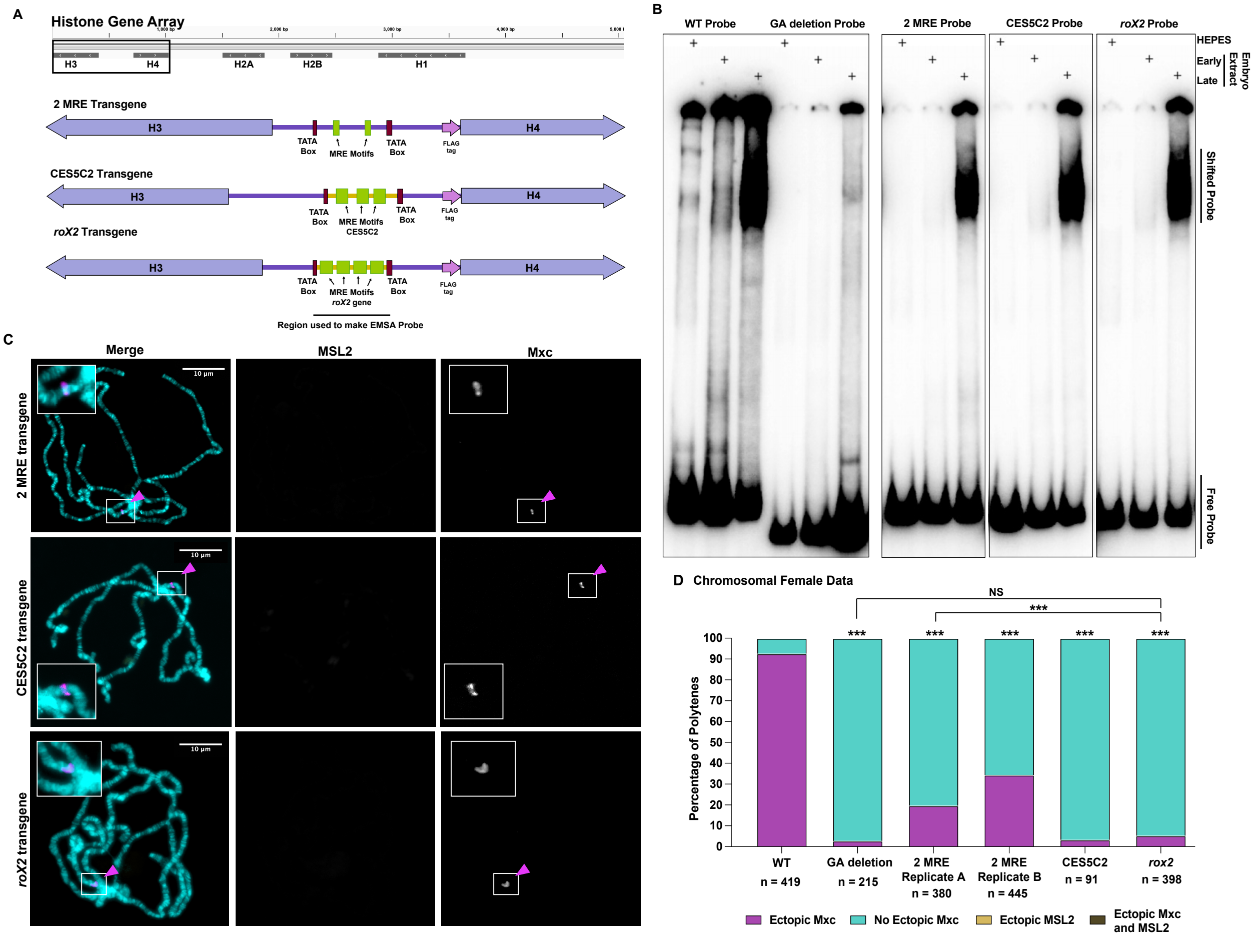


Figure 3.4: X-linked elements do not functionally substitute GA-repeats in the histone gene array in chromosomal females. (A) We engineered histone array transgenes in which we replaced parts of the *H3/H4p* sequence with varying amounts of X-linked sequence; the "2 MRE" transgenes replaces the GA-repeats with X-linked MREs from the *roX2* gene (light green boxes), the "CES5C2" transgene replaces the sequence between the TATA boxes (maroon) with the X-chromosome CES5C2 sequence (yellow line, light green bars indicate MREs), and the "*roX2"* transgene replaces the sequence between the TATA boxes with the CES sequence from the *roX2* gene (yellow line, light green bars indicate MREs). **(B)** We performed EMSA (gel shift) assays with early and late embryo extract (HEPES buffer serves as a control) and radiolabeled probes of the *H3/H4p* sequences of the histone array transgenes in (A) (WT: 226 bp,

GA deletion: 198 bp, 2 MRE: 238 bp, CES5C2: 232 bp, *roX2*: 232 bp). Late embryo extract, containing CLAMP, shifts all EMSA probes of the *H3/H4*p other than the GA deletion probe. **(C)** We performed immunofluorescence staining on third instar larval polytene chromosomes from salivary glands in chromosomal females for Mxc (magenta) and MSL2 (yellow). DNA is stained with DAPI (cyan). Mxc localizes to the endogenous histone locus, which is used as an internal staining control (magenta arrow, solid outlined box). **(D)** Quantification of ectopic Mxc from polytene scoring shows a significant difference in the percentage chromosome spreads that have ectopic Mxc. Significance above the bars represent the comparison to the WT data and the lines with represent direct comparisons between datasets. n values reflect number of polytenes scored for each respective genotype. Chi-squared test, *** = p < 0.001.

3.4.5 X-linked sequences in the context of the histone gene array attract MSL2 in chromosomal males

Because *Drosophila* males undergo dosage compensation, MSLc members such as MSL2 are present in males whereas they are not expressed in females. When CLAMP binds the GA-rich MREs on the X-chromosome, it then recruits MSLc for dosage compensation (Soruco and Larschan 2014; Rieder et al. 2019). We therefore considered that our transgenes might behave differently in chromosomal males compared to females. We performed immunofluorescence staining on male polytene chromosomes with antibodies against Mxc and MSL2. Although MSL2 may have some DNA-binding capability (Villa et al. 2016), it requires CLAMP for efficient X-chromosome targeting (Soruco and Larschan 2014; Rieder et al. 2019).

Similar to our observations in chromosomal females, X-linked sequences do not functionally substitute in the local context of the *H3/H4*p in males: we observed few instances of ectopic Mxc in animals carrying the chimeric transgenes (**Figure 3.5B,C**). Strikingly, ectopic Mxc was significantly rarer in chromosomal males than in females carrying the 2MRE transgene (**Figure 3.5C**). However, in males we observed ectopic, autosomal MSL2: approximately 50% of polytene spreads from animals carrying the 2MRE transgene showed ectopic MSL2, whereas the majority of larvae carrying the CES5C2 or *roX2* transgene showed ectopic MSL2 recruitment (**Figure 3.5B,C)**. These results suggest that X-linked MRE or CES sequences recruit MSLc, even in the context of the transgenic histone gene array. These results confirm that CLAMP utilizes cues from the *cis* element sequence itself as well as the local flanking regions to determine its function at the transgenic histone gene array.

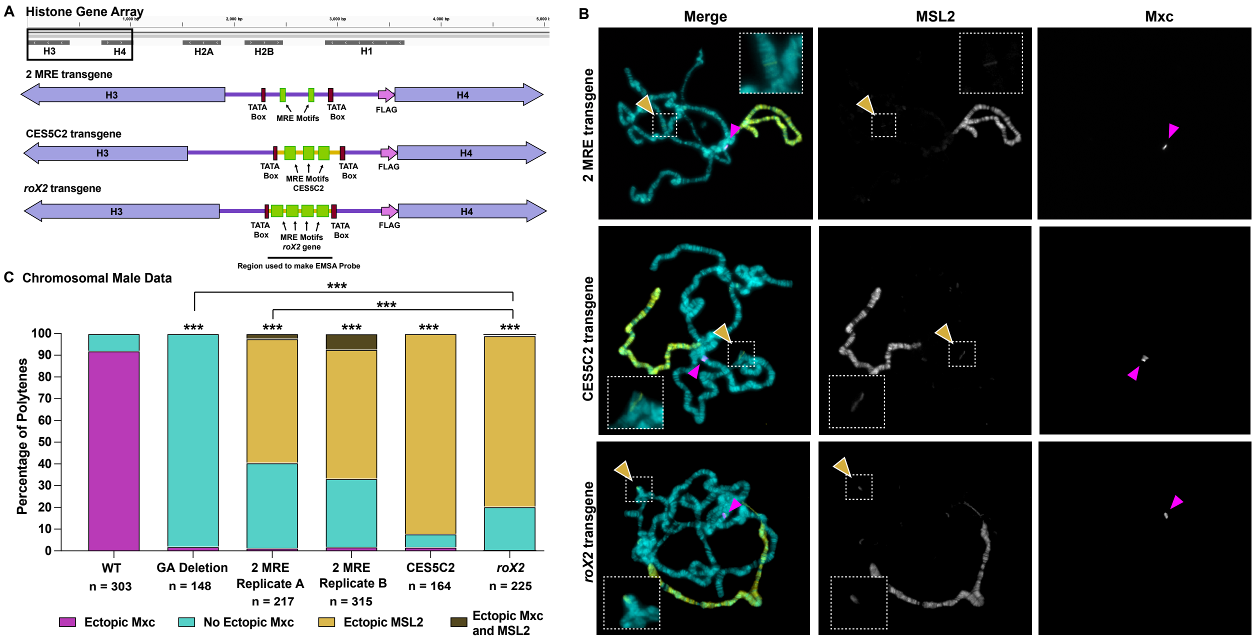


Figure 3.5: X-linked elements attract MSL2 in the context of the transgenic histone gene array in chromosomal males. (A) We engineered single copy histone array transgenes in which

we replaced parts of the *H3/H4p* sequence with varying amounts of X-linked sequence as in Figure 3.4. **(B)** We performed immunofluorescence staining on third instar larval polytene chromosomes in chromosomal males for Mxc (magenta) and MSL2 (yellow). DNA is stained with DAPI (cyan). Mxc localizes to the endogenous histone locus, which is used as an internal staining control (magenta arrow). MSL2 marks the male X-chromosome and also serves as an internal staining control. Chromosomal males containing any of the three histone array transgenes with X-linked sequence show some amount of ectopic MSL2 staining on autosomes (white outlined yellow arrows, dotted white boxes). **(C)** Quantification of ectopic Mxc from polytene scoring shows a significant difference in the percentage chromosome spreads that have ectopic Mxc. Significance above the bars represent the comparison to the WT data and the lines with represent direct comparisons between genotypes. n values reflect number of polytenes scored for each respective genotype. Chi-squared test, *** = $p < 0.001$.

When we replace the majority of the *H3/H4*p with X-linked sequence, CLAMP likely binds this region but performs its X-linked role, even in the context of the transgenic histone gene array, suggesting there may be other information within the local context, such as additional TF recruiting *cis* elements, that impact CLAMP histone locus role. This observation is interesting given that we recently discovered that in *Drosophila virilis,* a species ~40 million years diverged from *D. melanogaster,* MSL2 is recruited to one of their two endogenous histone loci (Xie et al. 2022). The *D. virilis H3/H4*p carries poorly conserved, much shorter GA-repeats than those in *D. melanogaster* (Rieder et al. 2017*),* but CLAMP is still recruited to this sequence *in vitro* and, furthermore, recruits both Mxc and MSL2 to one of the two loci (Xie et al. 2022). This observation suggests that there may be evolutionary differences in CLAMP function and that there are different mechanisms for histone gene regulation, perhaps even within single species (Koreski et al. 2020). In addition, we recently explored other candidates that target the *D. melanogaster* histone gene array (Hodkinson et al. 2023). We identified several DNA-binding factors from this screen, including the Hox factor Ultrabithorax, that may provide CLAMP contextual cues for functioning at the histone locus. Ubx appears to interact specifically with the *H3/H4*p sequence and is therefore positioned close to the CLAMP binding sites. Given that there is likely a "secondary" mechanism to HLB formation that does not involve the CLAMP-GA-repeat interaction (Koreski et al. 2020); see below), other transcription factors emerge as a likely mechanism.

Overall, we show that the context-dependent transcription factor CLAMP incorporates both *cis* element sequence information as well as cues from local flanking context where its *cis* binding elements reside to govern its function. We show that CLAMP *cis* elements are not interchangeable and that, in the context of the histone locus, the local context of the *H3/H4*p provides CLAMP with critical cues to function at the histone genes. Together our findings provide new insights into our understanding of how TFs bind similar *cis* elements in locations across the genome but can preserve their specific regulatory functions at these each of these different loci.

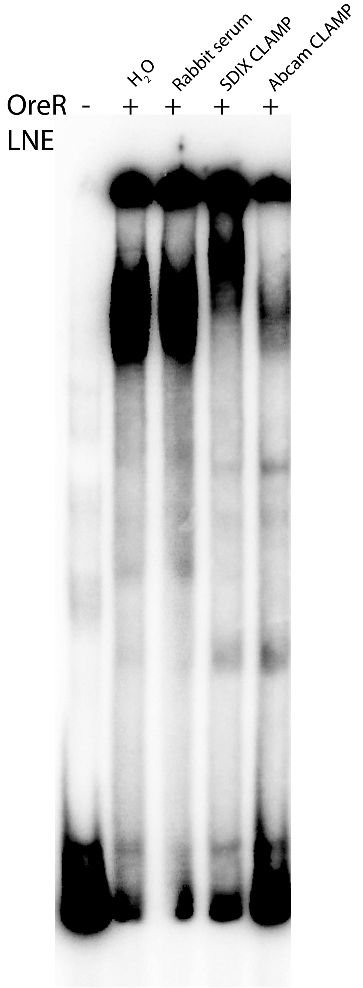
3.5 Acknowledgements

We would like to the members of the Rieder Lab as well as Dr. William Kelly and Dr. Roger Deal for their helpful remarks on the manuscript. We also would like to thank Dr. William Marzluf and Dr. Robert Duronio for Mxc antibody as well as Dr. Victoria Meller for MSL2 antibody. This work was supported by T32GM00008490 and F31HD105452 to LJH, K12GM00068 to CAS and HSC; F32GM140778 to CAS; and R00HD092625 and R35GM142724 to LER.

Author contributions

Conceptualization: JG, CAS, and LER. Methodology: JG, LJH, CAS, PPD, TA, and LER. Validation: LJH, TA, and LER. Formal analysis: LJH, CAS, PPD, and LER. Investigation: LJH, CAS, and TA. Resources: TA and LER. Data curation: LJH and LER. Writing – Original Draft: LJH and LER. Writing – Review and Editing: LJH, JG, CAS, TA, and LER. Visualization: LJH and LER. Supervision: LER. Project Administration: LJH and LER. Funding Acquisition: LJH, CAS, and LER.

3.6 Supplemental Figures



Supplemental Figure 31: Supershift demonstrating that CLAMP is present in late nuclear extract (LNE) from wild type (OregonR) and shifts the *H3H4*p probe. Late nuclear extract shifts the *H3H4*p probe. Rabbit serum (negative control) does not supershift, while two anti-CLAMP antibodies from two different companies bothsupershift.

3.7 References

Alekseyenko AA, Peng S, Larschan E, Gorchakov AA, Lee O-K, Kharchenko P, McGrath SD, Wang CI, Mardis ER, Park PJ, et al. 2008. A sequence motif within chromatin entry sites directs MSL establishment on the Drosophila X chromosome. *Cell* **134**: 599–609.

Aoki T, Schweinsberg S, Manasson J, Schedl P. 2008. A stage-specific factor confers Fab-7 boundary activity during early embryogenesis in Drosophila. *Mol Cell Biol* **28**: 1047–1060.

Bongartz P, Schloissnig S. 2018. Deep repeat resolution—the assembly of the Drosophila Histone Complex. *Nucleic Acids Research* **47**: e18–e18.

Carty M, Zamparo L, Sahin M, González A, Pelossof R, Elemento O, Leslie CS. 2017. An integrated model for detecting significant chromatin interactions from high-resolution Hi-C data. *Nat Commun* **8**: 15454.

Crayton ME, Ladd CE, Sommer M, Hampikian G, Strausbaugh LD. 2004. An organizational model of transcription factor binding sites for a histone promoter in D. melanogaster. *In Silico Biol* **4**: 537–548.

Duan J, Rieder L, Colonnetta MM, Huang A, Mckenney M, Watters S, Girish Deshpande, Jordan W, Fawzi N, Larschan E. 2021. CLAMP and Zelda function together to promote Drosophila zygotic genome activation. *eLife*. https://elifesciences.org/articles/69937 (Accessed January 9, 2023).

Fritz AJ, Ghule PN, Boyd JR, Tye CE, Page NA, Hong D, Shirley DJ, Weinheimer AS, Barutcu AR, Gerrard DL, et al. 2018. Intranuclear and higher-order chromatin organization of the major histone gene cluster in breast cancer. *J Cell Physiol* **233**: 1278–1290.

Fry CJ, Farnham PJ. 1999. Context-dependent Transcriptional Regulation *. *Journal of Biological Chemistry* **274**: 29583–29586.

Gorchakov AA, Alekseyenko AA, Kharchenko P, Park PJ, Kuroda MI. 2009. Long-range spreading of dosage compensation in Drosophila captures transcribed autosomal genes inserted on X. *Genes Dev* **23**: 2266–2271.

Groth AC, Fish M, Nusse R, Calos MP. 2004. Construction of transgenic Drosophila by using the site-specific integrase from phage phiC31. *Genetics* **166**: 1775–82.

Günesdogan U, Jäckle H, Herzig A. 2010. A genetic system to assess in vivo the functions of histones and histone modifications in higher eukaryotes. *EMBO Rep* **11**: 772–6.

Harrison MM, Botchan MR, Cline TW. 2010. Grainyhead and Zelda compete for binding to the promoters of the earliest-expressed Drosophila genes. *Developmental Biology* **345**: 248–255.

Hodkinson LJ, Smith C, Comstra HS, Ajani BA, Albanese EH, Arsalan K, Daisson AP, Forrest KB, Fox EH, Guerette MR, et al. 2023. A bioinformatics screen reveals hox and chromatin remodeling factors at the Drosophila histone locus. *BMC Genom Data* **24**: 54.

Hur W, Kemp JP, Tarzia M, Deneke VE, Marzluff WF, Duronio RJ, Di Talia S. 2020. CDK-Regulated Phase Separation Seeded by Histone Genes Ensures Precise Growth and Function of Histone Locus Bodies. *Dev Cell* **54**: 379-394.e6.

Kageyama Y, Mengus G, Gilfillan G, Kennedy HG, Stuckenholz C, Kelley RL, Becker PB, Kuroda MI. 2001. Association and spreading of the Drosophila dosage compensation complex from a discrete roX1 chromatin entry site. *The EMBO Journal* **20**: 2236–2245.

Kaya-Okur HS, Wu SJ, Codomo CA, Pledger ES, Bryson TD, Henikoff JG, Ahmad K, Henikoff S. 2019. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat Commun* **10**: 1930.

Kaye EG, Kurbidaeva A, Wolle D, Aoki T, Schedl P, Larschan E. 2017. Drosophila Dosage Compensation Loci Associate with a Boundary-Forming Insulator Complex. *Mol Cell Biol* **37**: e00253-17.

Kelley RL, Meller VH, Gordadze PR, Roman G, Davis RL, Kuroda MI. 1999. Epigenetic spreading of the Drosophila dosage compensation complex from roX RNA genes into flanking chromatin. *Cell* **98**: 513–22.

Koreski KP, Rieder LE, McLain LM, Chaubal A, Marzluff WF, Duronio RJ. 2020. Drosophila histone locus body assembly and function involves multiple interactions. *Mol Biol Cell* **31**: 1525–1537.

Kuzu G, Kaye EG, Chery J, Siggers T, Yang L, Dobson JR, Boor S, Bliss J, Liu W, Jogl G, 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 Genet* **12**: e1006120.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.

Lee TI, Young RA. 2013. Transcriptional Regulation and its Misregulation in Disease. *Cell* **152**: 1237–1251.

Li X-Y, Eisen MB. 2018. Effects of the maternal factor Zelda on zygotic enhancer activity in the Drosophila embryo. 385070. https://www.biorxiv.org/content/10.1101/385070v1 (Accessed November 6, 2023).

McKay DJ, Klusza S, Penke TJ, Meers MP, Curry KP, McDaniel SL, Malek PY, Cooper SW, Tatomer DC, Lieb JD, et al. 2015. Interrogating the function of metazoan histones using engineered gene clusters. *Dev Cell* **32**: 373–86.

Meers MP, Leatham-Jensen M, Penke TJR, McKay DJ, Duronio RJ, Matera AG. 2018. An Animal Model for Genetic Analysis of Multi-Gene Families: Cloning and Transgenesis of Large Tandemly Repeated Histone Gene Clusters. *Methods Mol Biol* **1832**: 309–325.

Ozdemir A, Fisher-Aylor KI, Pepke S, Samanta M, Dunipace L, McCue K, Zeng L, Ogawa N, Wold BJ, Stathopoulos A. 2011. High resolution mapping of Twist to DNA in Drosophila embryos: Efficient functional analysis and evolutionary conservation. *Genome Res* **21**: 566–577.

Rieder LE, Jordan WT 3rd, Larschan EN. 2019. Targeting of the Dosage-Compensated Male X-Chromosome during Early Drosophila Development. *Cell Rep* **29**: 4268-4275.e2.

Rieder LE, Koreski KP, Boltz KA, Kuzu G, Urban JA, Bowman SK, Zeidman A, Jordan WT 3rd, Tolstorukov MY, Marzluff WF, et al. 2017. Histone locus regulation by the Drosophila dosage compensation adaptor protein CLAMP. *Genes Dev* **31**: 1494–1508.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative Genomics Viewer. *Nat Biotechnol* **29**: 24–26.

Salzler HR, Tatomer DC, Malek PY, McDaniel SL, Orlando AN, Marzluff WF, Duronio RJ. 2013. A sequence in the Drosophila H3-H4 Promoter triggers histone locus body assembly and biosynthesis of replication-coupled histone mRNAs. *Dev Cell* **24**: 623–34.

Soruco MML, Larschan E. 2014. A new player in X identification: the CLAMP protein is a key factor in Drosophila dosage compensation. *Chromosome Res* **22**: 505–515.

Tatomer DC, Terzo E, Curry KP, Salzler H, Sabath I, Zapotoczny G, McKay DJ, Dominski Z, Marzluff WF, Duronio RJ. 2016. Concentrating pre-mRNA processing factors in the histone locus body facilitates efficient histone mRNA biogenesis. *Journal of Cell Biology* **213**: 557–570.

Terzo EA, Lyons SM, Poulton JS, Temple BRS, Marzluff WF, Duronio RJ. 2015. Distinct self-interaction domains promote Multi Sex Combs accumulation in and formation of the Drosophila histone locus body. *Mol Biol Cell* **26**: 1559–1574.

The Galaxy Community. 2022. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Research* **50**: W345–W351.

Urban J, Kuzu G, Bowman S, Scruggs B, Henriques T, Kingston R, Adelman K, Tolstorukov M, Larschan E. 2017a. Enhanced chromatin accessibility of the dosage compensated Drosophila male X-chromosome requires the CLAMP zinc finger protein. *PLoS One* **12**: e0186855.

Urban JA, Urban JM, Kuzu G, Larschan EN. 2017b. The Drosophila CLAMP protein associates with diverse proteins on chromatin. *PLoS One* **12**: e0189772.

Venken KJT, He Y, Hoskins RA, Bellen HJ. 2006. P[acman]: a BAC transgenic platform for targeted insertion of large DNA fragments in D. melanogaster. *Science* **314**: 1747–1751.

Villa R, Schauer T, Smialowski P, Straub T, Becker PB. 2016. PionX sites mark the X chromosome for dosage compensation. *Nature* **537**: 244–248.

White AE, Burch BD, Yang XC, Gasdaska PY, Dominski Z, Marzluff WF, Duronio RJ. 2011. Drosophila histone locus bodies form by hierarchical recruitment of components. *J Cell Biol* **193**: 677–94.

Xie M, Hodkinson LJ, Comstra HS, Diaz-Saldana PP, Gilbonio HE, Gross JL, Chavez RM, Puckett GL, Aoki T, Schedl P, et al. 2022. MSL2 targets histone genes in Drosophila virilis. 2022.12.14.520423. https://www.biorxiv.org/content/10.1101/2022.12.14.520423v1 (Accessed January 3, 2023).

Chapter 4

MSL2 targets histone genes in *Drosophila virilis*

Reproduced with permission from:

Mellisa Xie*[1], Lauren J. Hodkinson*[1,2], H. Skye Comstra[1], Pamela P. Diaz-Saldana[1], Hannah E. Gilbonio[1], Julia L. Gross[1,3,4], Robert M. Chavez[1,2], Gwyn L. Puckett[1], Tsutomu Aoki[5], Paul Schedl[5], and Leila E. Rieder[1] 2022. MSL2 targets histone genes in Drosophila virilis. 2022.12.14.520423.

*Authors contributed equally

[1] Emory University Department of Biology, Atlanta, GA 30322, USA

[2] Genetics and Molecular Biology graduate program, Emory University, Atlanta, GA 30322, USA

[3] Immunology and Molecular Pathogenesis graduate program, Emory University, Atlanta, GA 30322, USA

[4] Signaling Systems Section, National Institute of Allergy and Infectious Disease, Graduate Partnership Program, Bethesda, MD 20892, USA

[5] Princeton University Department of Molecular Biology, Princeton, NJ 08544, USA

4.1 Abstract

Histone genes are amongst the most evolutionary conserved in eukaryotic genomes, yet *cis*-regulatory mechanisms of histone gene regulation differ considerably amongst species. In *Drosophila melanogaster*, an interaction between GA-rich *cis* elements in the *H3/H4* promoter and the GA-binding transcription factor CLAMP is important for promoting histone gene regulation and factor recruitment to the locus. CLAMP also participates in male dosage compensation by recruiting the Male Specific Lethal Complex (MSLc) to the X-chromosome. We discovered that the male-specific protein of MSLc, MSL2, is recruited to the autosomal major histone locus in *D. virilis* but not to the minor locus or to the single histone locus in other species. While the histone coding sequences are well conserved between species, the critical GA-rich *cis* elements in the *H3/H4* promoter are poorly conserved between *D. melanogaster* and *D. virilis*. We show that CLAMP still targets the two *D. virilis* histone loci *in vivo*. Further, CLAMP interacts with the *D. virilis H3/H4* promoter *in vitro*, even when the poorly-conserved GA-rich *cis* elements are deleted, indicating that the protein interacts differently with the *D. virilis* promoter than it does with the *D. melanogaster* promoter. Since CLAMP and MSL2 directly interact in *D. melanogaster*, we propose that *D. virilis* CLAMP recruits MSL2 to an ectopic autosomal site through interaction with X-like *cis* elements. Further, localization of MSL2 to one of the *D. virilis* histone loci suggests that the loci are regulated differently, and that males and females have different requirements for histone gene regulation.

4.2 Introduction

Histones are critical organizational components of eukaryotic chromatin and are highly conserved. For example, histone H3 is 80% identical at the nucleotide level and 99% identical at the protein level between *Drosophila melanogaster* and humans. Histone levels are carefully controlled during both the cell cycle and development; coordinated expression of histone genes is cell-cycle regulated and peaks during S phase (Marzluff *et al.* 2008). Misregulation of histone genes disrupts the precise cell cycle timing during animal development (Amodeo *et al.* 2015; Chari *et al.* 2019). Unsurprisingly, cell cycle regulatory requirements of the replication-dependent histone genes are similar between species (Mariño-Ramírez *et al.* 2006). Despite similar cell cycle and developmental regulatory requirements between animals, histone gene *cis*-regulatory mechanisms appear to differ between species (Kremer and Hennig 1990; Mariño-Ramírez *et al.* 2006).

Organization of the histone genes within the genome also varies widely between species. Vertebrate genomes tend to have lower histone gene copy number and dispersed distribution, while invertebrate genomes carry high numbers of tandem histone repeats. The human genome has two loose clusters of histone genes interspersed with non-histone genes (Marzluff *et al.* 2002). The *C. elegans* genome carries eleven dispersed clusters of the four core histone genes, and histone loci do not include the histone *H1* gene. The histone locus of *Drosophila melanogaster* is a single locus that carries ~100 copies of a histone gene array that includes all five replication-dependent histone genes (Lifton *et al.* 1978; McKay *et al.* 2015; Bongartz and Schloissnig 2019).

The organization and number of histone genes are not well conserved even within Drosophilidae. *D. hydei* has only about 10 histone array copies, and they are located in the middle of euchromatin on Chromosome 4 (Fitch *et al.* 1990). *D. virilis*, which diverged from *D. melanogaster* ~ 40 MYa (Russo *et al.* 1995) has both regular quartet arrays that include the core histone genes (*H2A, H2B, H3*, and *H4*) and polymorphic quintet arrays that include the histone *H1* gene. The quartet arrays are tandemly distributed and linked to a single "major" locus, while the quintets are distributed between both "major" and "minor" loci (Schienman *et al.* 1998).

The diversity in histone gene organization is striking given similar requirements for histone gene expression across species (Mariño-Ramírez *et al.* 2006). In animals, replication-dependent histone biogenesis is controlled by a suite of factors that target histone genes called the Histone Locus Body (HLB) (Liu *et al.* 2006; Nizami *et al.* 2010; Duronio and Marzluff 2017). The interaction between the scaffolding protein Multi-sex combs (Mxc; *Drosophila* homolog of human nuclear protein of the ataxia telangiectasia mutated locus/NPAT) and the RNA processing factor FLICE-associated huge protein (FLASH) is required for HLB formation in both human and *Drosophila* cells (Yang *et al.* 2014; Kemp *et al.* 2021). However, even these critical HLB proteins are poorly conserved at the sequence level, and there is little indication that Mxc interacts directly with DNA (Terzo *et al.* 2015; Kaya-Okur *et al.* 2019; Kemp *et al.* 2021). Therefore, Mxc and FLASH are unlikely to be the first factors that identify the *Drosophila* zygotic histone genes for unique regulation during development.

Mechanisms of histone gene regulation are well studied in *D. melanogaster*, as the histone genes reside at a single locus (Günesdogan *et al.* 2014; Bongartz and Schloissnig 2019) and histone array transgenes attract HLB factors (Salzler *et al.* 2013). Similar manipulative studies in mammalian systems are comparatively much more difficult (Sankar *et al.* 2022) due to the dispersion of histone genes across two chromosomes and megabases of sequence (Marzluff *et al.* 2002). In *D. melanogaster*, the histone locus is identified by the zinc-finger protein CLAMP, which interacts with long, perfect GA-repeat *cis* elements in the *H3/H4* promoter (Salzler *et al.* 2013; Rieder *et al.* 2017). The CLAMP-histone locus interaction promotes recruitment of Mxc and other HLB proteins (Rieder *et al.* 2017)(L. Hodkinson, observation). Removing the GA-repeat *cis* elements from histone array transgenes abrogates the ability of the transgene to recruit HLB-specific factors (Rieder *et al.* 2017)(L. Hodkinson, observation). At the endogenous locus, CLAMP increases histone locus chromatin accessibility and promotes histone gene expression of all five replication-dependent genes (Rieder *et al.* 2017). While there are likely multiple redundant mechanisms of HLB formation (Koreski *et al.* 2020), CLAMP is the first known factor that directly interacts with histone locus DNA sequence to promote recognition of the histone genes and recruitment of HLB-specific factors.

However, CLAMP is not specific to the histone locus; it is also critical for *Drosophila* dosage compensation (Soruco *et al.* 2013; Soruco and Larschan 2014), which increases X-linked gene expression to equalize gene dosage of males to females and the X-chromosome to the autosomes. CLAMP targets the male X-chromosome at GA-rich sequences (MSL recognition elements; MREs (Alekseyenko *et al.* 2008)), increases male X-chromosome accessibility, and recruits the Male Specific Lethal complex (MSLc) (Kuzu *et al.* 2016; Larschan *et al.* 2017). MSLc spreads across the chromosome and deposits the activating H4K16ac mark to increase male X-linked gene expression (Conrad *et al.* 2012). In addition to its role in histone gene regulation and dosage compensation, CLAMP targets autosomal sites in males and females to increase promoter accessibility and transcriptional elongation (Urban *et al.* 2017). CLAMP is also a component of the Late Boundary Complex, which forms in late-stage *Drosophila* embryos and impacts both dosage compensation (Kaye *et al.* 2017) and insulation within the bithorax Hox gene complex (Wolle *et al.* 2015; Kyrchanova *et al.* 2019). With so many critical functions, it is not surprising that CLAMP is comparatively well conserved amongst Drosophilidae, although it is unique to insects (Kuzu *et al.* 2016).

We previously hypothesized that the well-conserved CLAMP factor provides a bridge between histone locus *cis* elements and poorly conserved locus-specific factors such as Mxc/NPAT (Rieder *et al.* 2017). Similarly, MSLc proteins are poorly conserved (Kuzu *et al.* 2016) and CLAMP may provide a conserved link to newly evolving sex chromosomes (Alekseyenko *et al.* 2013).

Based on the above observations, we were surprised to observe that the male-specific component of MSLc, MSL2, targets one of the two autosomal histone loci in *Drosophila virilis*. MSL2 does not target the single histone locus in other *Drosophila* species and MSL2 is largely confined to the male X-chromosome in *D. melanogaster* (Lucchesi and Kuroda 2015). We report that the critical GA-repeat *cis* element in the *D. melanogaster H3/H4* promoter is almost unrecognizable in *D. virilis*, and more closely resembles GA-rich X-linked MREs. Both *D. melanogaster* and *D. virilis* CLAMP recognize the *D. virilis H3/H4* promoter region *in vivo*. However, *in vitro,* CLAMP does not require the GA-rich element in the *D. virilis* promoter sequence.

Our observations suggest that the two *D. virilis* histone loci are differentially regulated, as previously documented in yeast (Norris and Osley 1987; Cross and Smith 1988) and sea urchin (Marzluff *et al.* 2006). Since MSL2 is only expressed in male *Drosophila*, MSL2 might contribute to sex-specific regulation of the histone genes. Our observations indicate that context-specific transcription factors such as CLAMP may not always completely differentiate between genomic locations, resulting in cross-talk between regions with extremely different regulatory requirements.

4.3 Methods

Drosophila strains

We used the following stocks, maintained on standard cornmeal/molasses food and raised at 18°C: *Drosophila melanogaster* (y[1]w[1118]; +;+;+), *Drosophila virilis* (National *Drosophila* Species Stock Center #15010-1051.88), *Drosophila pseudoobscura* (NDSSC #14011-0121.217), and *Drosophila willistoni* (NDSSC #14030-0811.15).

Cloning and transgenesis

We engineered plasmids that include a 5kb histone array sequence consisting of the 5 replication-dependent histone genes and their relative promoters where the *histone4* and *histone2A* genes are FLAG-tagged (24 bp) at the N-termini (Salzler *et al.* 2013) (original plasmid gift of Drs. Robert Duronio and William Marzluff). We used geneblocks (IDT) carrying the desired changes to alter the sequence of the histone gene array using Gibson cloning. Transgenic sequences are detailed in **Supplemental Table 4.3**. We inserted all 1x histone array transgenes into the genome at the VK33 attP site on chromosome 3L (65B2) (Venken *et al.* 2006) using PhiC3-mediated integration (Genetivision).

Electrophoretic Mobility Shift Assays (EMSAs)

We performed EMSAs after Aoki *et al.* (2008) with minor modifications.

Late embryo nuclear extracts: We prepared embryo extracts from 6-18 hour Oregon R embryos collected on apple juice plates and aged 6 hours at room temperature. We performed nuclear extract preparation as in (Aoki *et al.* 2008). We omitted the final dialysis step described in Aoki *et al.* and completed the extraction with the final concentration of KCl at 360 mM.

DNA probes: We made EMSA probes (sequences in **Supplemental Table 4.2**) using PCR using gblocks (IDT) as templates and the following primers: *D. melanogaster* sequences: H3H4p F1: CACAGCACGAAAGTCACTAAAGAAC, H3H4p R1: GTTTGAAAACACAATAAACGATCAGAGC; *D. virilis* sequences: virilis H3H4p F1: CACCACGAATGTCACTGAGG, virilis H3H4p R1: TGTTAAAAACACAATAATCGTGCGTC. We 5′ end labeled one pmol of probe with γ-32P-ATP (MP Biomedicals) using T4 polynucleotide kinase (New England BioLabs) in a 50 µl total reaction volume at 37°C for 1 hour. We used Sephadex G-50 fine gel (Amersham Biosciences) columns to separate free ATP from labeled probes. We adjusted the volume of the eluted sample to 100 µl using deionized water so that the final concentration of the probe was 10 fmol/µl.

Shifts: We performed 20 µl binding reactions consisting of 0.5 µl (5 fmol) of labeled probe in the following buffer: 25 mM Tris-Cl (pH 7.4), 100 mM KCl, 1 mM EDTA, 0.1 mM dithiothreitol, 0.1 mM PMSF, 0.3 mg/ml bovine serum albumin, 10% glycerol, 0.25 mg/ml poly(dI-dC)/poly(dI-dC). We added 1 µl of nuclear extract and incubated samples at room temperature for 30 minutes. We loaded samples onto a 4% acrylamide (mono/bis, 29:1)-0.5× TBE-2.5% glycerol slab gel. We performed electrophoresis at 4°C, 180 V for 3-4 hours using 0.5× TBE-2.5% glycerol as a running buffer. We dried gels and imaged using a Typhoon 9410 scanner and Image Gauge software.

Recombinant CLAMP EMSAs: We performed recombinant CLAMP EMSAs using full-length recombinant, purified CLAMP protein (61.8 kDa) (Duan *et al.* 2021). We used the LightShift Chemiluminescent EMSA Kit (Thermo Fisher #20148) and performed 20 µl binding reactions with 1 µl CLAMP (1 µM), 1 µl biotinylated probe (0.3 µg/µl) and 1 ug/ul poly(dI-dC)/poly(dI-dC). We incubated reactions for 25 min at room temperature, ran on a 6% nondenaturing polyacrylamide gel and electrophoresed at 100 V for 2 hr in 0.5 X TBE. We performed semi-wet gravity transfer using the TurboBlotter (Cytiva) for 4 hours at room temperature using 20X SSC transfer buffer. We visualized the blotusing the Nucleic Acid Detection Module Kit (Thermo Fisher #89880).

Western blotting

We collected *D. melanogaster* and *D. virilis* embryos on standard grape juice plates for 16 hours and dechorinated on the plate in 100% bleach for 2 minutes. We washed embryos in 1X PBS and then lysed and ground in RIPA buffer + protease inhibitor (Roche #11697498001). We spun samples at 20,000g for 5 minutes and retained the supernatant; this was repeated twice. We diluted the resulting protein lysate in 6X Laemmli sample buffer and ran samples on a 4 - 20% Bolt Bis-Tris gel. We transferred samples to a PVDF membrane, which we blocked for 1 hour in 3% BSA in TBS-T. We incubated the membrane overnight at 4°C with primary antibody at the following concentrations: anti-MSL2^{mel} serum at 1:100 (gift from Dr. Mitzi Kuroda) and anti-β actin at 1:1000 (CST #8457S). We washed the membrane 3x 5 minutes in TBS-T and then incubated it with secondary antibody (LI-COR IRDye® 800CW/680RD Goat anti-Rabbit IgG) at 1:10,000 in TBS-T + 0.01% SDS for 1 hour. We washed the membrane 3x 5 minutes in TBS-T and 1x 5 minutes TBS before imaging (Bio-Rad ChemiDoc).

Immunofluorescence on polytene chromosomes

We performed polytene chromosomes immunostaining from salivary glands dissected from sexed third instar *Drosophila* larvae raised at 18°C on standard cornmeal/molasses food. We passed glands through fix 1 (4% formaldehyde, 1% Triton X-100, in 1× PBS) for 1 min, fix 2 (4% formaldehyde, 50% glacial acetic acid) for 2 min, and 1:2:3 solution (ratio of lactic acid:water:glacial acetic acid) for 5 min prior to squashing and spreading. We washed slides in 1X PBS, then in 1% Triton X-100 (diluted in 1X PBS), and blocked for one hour in .5% BSA diluted in 1X PBS. We then incubated slides with primary antibodies diluted in blocking solution (antibodies specifics below) overnight at 4º C in a dark, humid chamber. We washed slides in 1 X PBS and incubated with secondary antibody diluted in blocking solution (antibody specifics below) for two hours at room temperature. We mounted slides in Prolong Diamond anti-fade reagent with DAPI (ThermoFisher, P36961), and imaged chromosome spreads on a Zeiss Scope.A1 equipped with a Zeiss AxioCam using a 40×/0.75 plan neofluar objective using AxioVision software.

Antibodies

We used primary antibodies at the following concentrations: guinea pig anti-Mxc (1:5000; gift from Drs. Robert Duronio and William Marzluff), rabbit anti-MSL2 (1:150; gift from Dr. Victoria Meller, originally from Dr. Ron Richmond), goat anti-MSL3 serum (1:500; gift from Dr. Erica Larschan, originally from Dr. Mitzi Kuroda). All primary antibodies are raised against the *D. melanogaster* forms of the proteins. We used AlexaFluor secondary antibodies (ThermoFisher Scientific) at a concentration of 1:1000: goat anti-guinea pig AF647 (A-21450), goat anti-rabbit AF488 (A-11008), donkey anti-goat AF488 (A-11055).

Bioinformatics

We used the online platform Galaxy (usegalaxy.org) (Afgan *et al.* 2018) to map ChIP-seq datasets to the histone gene array. We used the following datasets from NCBI GEO: GSE165833 (Villa *et al.* 2021) and GSE133637 (Rieder *et al.* 2019). We mapped reads to a single copy of the histone gene array as in Mckay *et al.* (2015) and normalized when possible to available input samples. We visualized data using Integrative Genomics Viewer (IGV) (Robinson *et al.* 2011). We used a custom R script to combine replicates, when available. The script is deposited at https://github.com/rieder-lab/Omics-Replicate-Merger.

Sequence annotation and alignments

We annotated the *D. virilis* genome assembly (Kim *et al.* 2022) using SnapGene by searching for histone protein conservation with *D. melanogaster*. *D. virilis* genome assembly: ASM798932v2. *D. melanogaster* histone array sequences from (McKay *et al.* 2015). *D. virilis* histone array sequences from UCSC genome browser (http://genome.ucsc.edu) Aug. 2005 (Agencourt prelim/droVir2) release, scaffold_13047, range 1568499-1650698. *D. virilis* histone array sequences from (DDBJ accession no. AB249651) (Nakashima *et al.* 2016). We aligned sequences using Coffee (Erb *et al.* 2012) and SnapGene.

Data availability

Strains and plasmids are available upon request. The authors affirm that all data necessary for confirming the conclusions of the article are present within the article, figures, and tables.

4.4 Results

4.4.1 Sequence differences between *D. virilis* histone loci

Drosophila virilis diverged from *D. melanogaster* ~ 40 million years ago. Instead of one histone locus, as in *D. melanogaster*, it carries two: a major locus on Chromosome II (25F) and a minor locus on Chromosome IV (43C)(Schienman *et al.* 1998). Schienman *et al.* (1998) performed Southern blot analysis and concluded that the major locus includes 25-30 arrays of both a quintet organization (*H2a, H2b, H3, H4*, and *H1*) and a quartet lacking *H1*, while the minor locus includes 6-8 quintet arrays. Kim *et al.* (2022) recently assembled 101 drosophilid genomes using PacBio long-read sequencing. We annotated the *D. virilis* histone loci from the Kim *et al.* assembly and determined that the major locus includes 5 quintet arrays and 27 quartet arrays, as well as many interrupted arrays and gene fragments. The minor locus includes 5 regularly spaced quintet arrays.

The ~100 regularly spaced quintet histone gene arrays in *D. melanogaster* are nearly identical in sequence (Bongartz and Schloissnig 2019). For example, the core *H3* genes vary at only two silent site locations in *D. melanogaster* (G231A, T408C) out of 411 total nt. However, we noticed substantial sequence differences between the same histone genes in *D. virilis*: five *H3* silent locations between the major and minor loci in *D. virilis*. Eight *H3* genes, distributed between the loci, include single nucleotide changes.

Similarity of the linker histone *H1* genes is more variable, compared to core histone genes; the human genome includes H1 subtypes H1.1-H1.5. Similarity of a subtype between species is higher than similarity of subtypes within a species (Di Liegro *et al.* 2018). By comparing the ~100 *D. melanogaster H1* genes, we found only two silent mutations (T312C, C462T) and two coding changes (D82E, V217I). However, we discovered that *D. virilis H1* genes include significant variation: 27 single nucleotide changes and 2 insertion/deletions (of 3 nucleotides each) across 753 nt. Eleven of the single nucleotide changes cause amino acid substitutions and the majority of sequence changes are locus-specific. Our observations suggest that replication-dependent *H1* subtypes across both loci could be present in this species.

4.4.2 Both D. virilis histone loci are targeted by Mxc and CLAMP

The *D. virilis* histone loci differ in sequence, in contrast to the nearly identical sequences of the histone arrays at the single *D. melanogaster* locus. We therefore hypothesized that different factors target the major and minor loci. Mxc is an important HLB scaffolding protein (Hur *et al.* 2020; Kemp *et al.* 2021) that targets the *D. melanogaster* histone genes early during development (White *et al.* 2011). CLAMP is a non-histone-locus specific protein that interacts with GA-repeats in the *H3/H4* promoter, which promotes Mxc recruitment and HLB formation (Rieder *et al.* 2017). We previously observed that CLAMP targets histone loci in both *D. melanogaster* and *D. virilis* by polytene chromosome immunostaining (Rieder *et al.* 2017). We confirmed our previous observation by staining *D. virilis* polytene chromosomes with antibodies against *D. melanogaster* Mxc and CLAMP orthologs and observed that both Mxc and CLAMP target both *D. virilis* loci (**Figure 4.1A-B**).

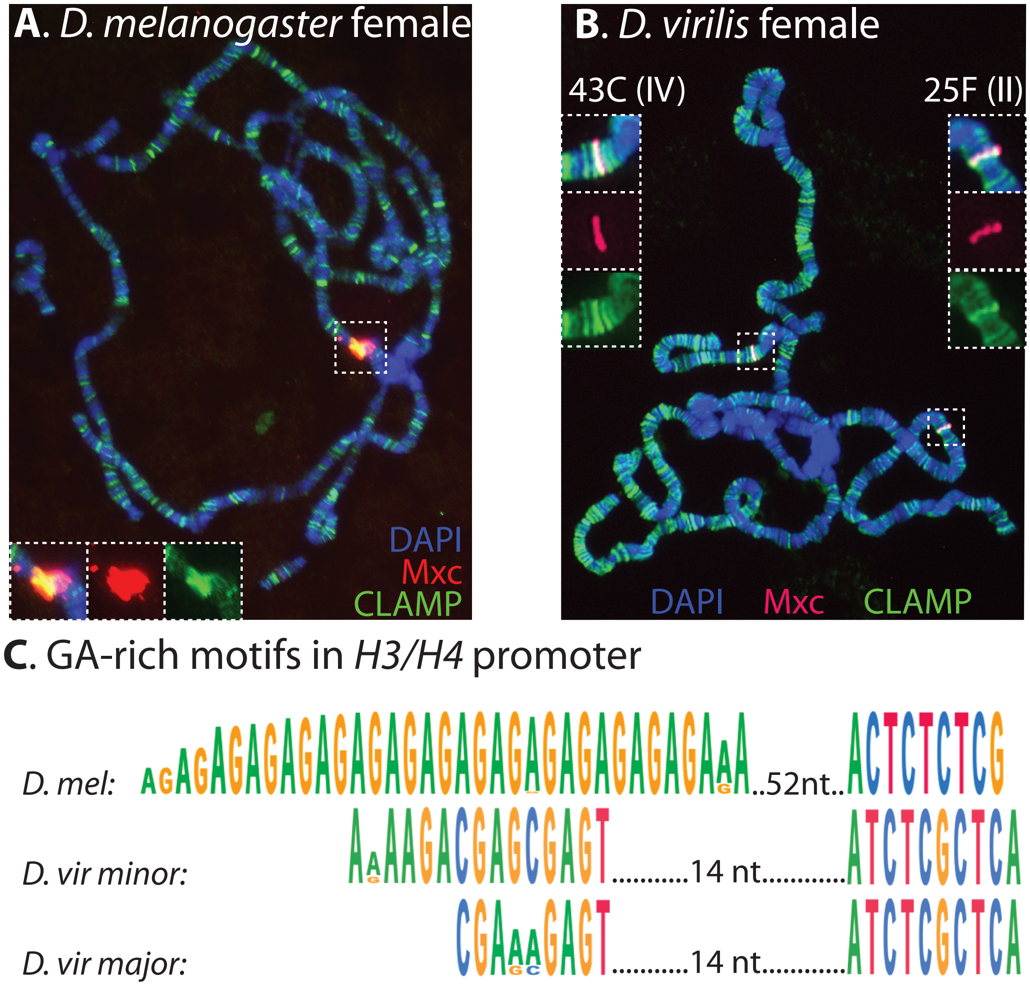

A. D. melanogaster female
DAPI
Mxc
CLAMP
B. D. virilis female
43C (IV)
25F (II)
DAPI
Mxc
CLAMP
C. GA-rich motifs in H3/H4 promoter
D. mel: AGAGAGAGAGAGAGAGAGAGAGAGAGAGAA..52nt..ACTCTCTCG
D. vir minor: AAAAGACGAGCGAGT...........14 nt............ATCTCGCTCA
D. vir major: CGAAAGAGT...........14 nt............ATCTCGCTCA

Figure 4.1: CLAMP and Mxc target the *D. virilis* histone loci. (A) We confirmed previous results (Rieder *et al.* 2017) that CLAMP (green) targets the single histone locus in *D. melanogaster* by staining female larvae polytene chromosome spreads. Multi sex combs (Mxc; pink) is a core HLB protein that specifically targets the histone locus. (**B)** CLAMP (green) also targets the two histone loci (colocalizing with Mxc) in *Drosophila virilis*. (**C)** In *D. melanogaster* (*D. mel*), CLAMP targets two long GA-repeats in the *H3/H4* promoter (Rieder *et al.* 2017). These GA-repeats are conserved but shorter and interrupted in *D. virilis* (*D. vir*).

CLAMP targets GA-rich motifs genome-wide in *D. melanogaster* early during development (Kuzu *et al.* 2016; Rieder *et al.* 2019) and facilitates several essential processes, including dosage compensation (Kuzu *et al.* 2016; Larschan *et al.* 2017; Rieder *et al.* 2019) and transcriptional elongation (Urban *et al.* 2017). The GA-repeats found in the *D. melanogaster H3/H4* promoter are long and unbroken while they are shorter and interrupted in *D. virilis* (**Figure 4.1C**). The *D. virilis H3/H4 cis* elements therefore more closely resemble the interrupted, GA-rich MREs found on *Drosophila melanogaster* X-chromosomes, which are critical for dosage compensation (Alekseyenko *et al.* 2008; Villa *et al.* 2016). We therefore considered that CLAMP might be attracting MSLc to the major *D. virilis* histone locus.

4.4.3 MSL2 targets the major *D. virilis* histone locus

MSLc is confined to the male X-chromosome in *D. melanogaster* (Lucchesi and Kuroda 2015), and this pattern is apparent when staining third instar larval polytene chromosomes (**Figure 4.2A-B**). MSLc does not co-localize with Mxc, which marks the single histone locus on chromosome 2L. We stained male wild-type *D. virilis* polytene chromosomes for Mxc (to mark the histone loci), and MSL2, the male-specific structural component of MSLc (Bashaw and Baker 1995; Kelley *et al.* 1997). We were surprised to observe that MSL2 specifically targets the male *D. virilis* major histone locus on Chromosome II but not the minor histone locus on Chromosome IV (**Figure 4.2C**).

Our anti-MSL2 antibody was raised against *D. melanogaster* MSL2 sequence, and MSLc proteins are not well conserved even in Drosophilidae (Kuzu *et al.* 2016). Strangely, we did not observe male X-chromosome staining in *D. virilis* (**Figure 4.2C; Supplemental Table 4.1**), in contrast to prior work (Marín *et al.* 1996). Therefore, to confirm that our antibody is specific to

MSL2 in both species, we performed western blotting. Our anti-MSL2^{mel} antibody recognized proteins around the correct predicted sizes: ~85 kDa in *D. melanogaster* and ~81 kDa in *D. virilis*, although the *virilis* ortholog may be modified and appears larger than the *melanogaster* ortholog (**Figure 4.2E**). Critically, we do not observe MSL2 staining at 25F in female *D. virilis* (**Figure 4.2D**), further indicating that the anti-MSL2^{mel} antibody is recognizing the *D. virilis* ortholog. These observations indicate that our antibody is specific to MSL2 and recognizes the protein in both species.

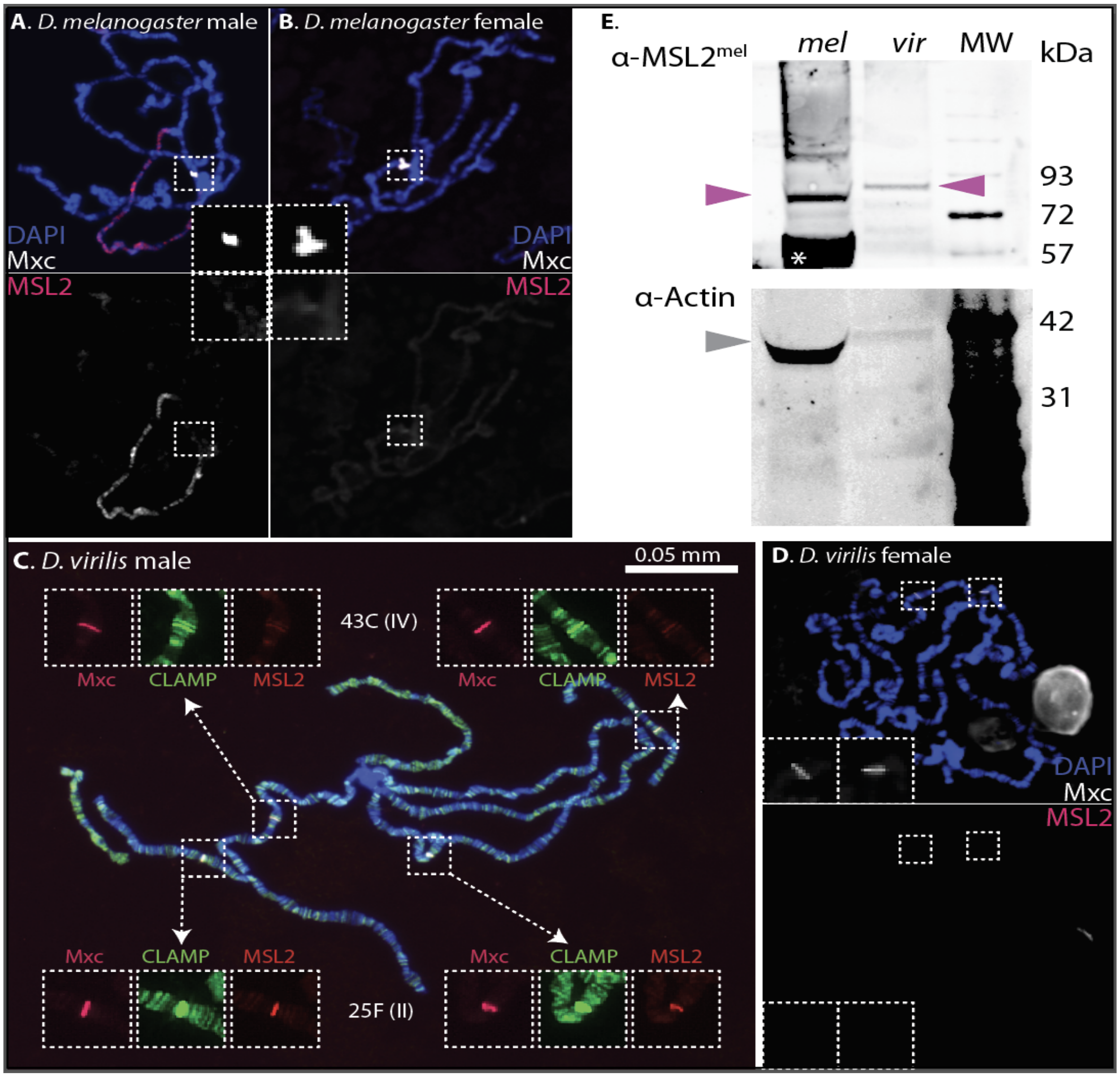


Figure 4.2: MSL2 targets only one histone locus in *D. virilis*. (A) MSL2 (pink) targets the *D. melanogaster* male X-chromosome, but does not overlap with Mxc (white) at the histone locus. (**B)** MSL2 is not present in female *D. melanogaster*. (**C)** Two polytene chromosome spreads from *D. virilis* males show that MSL2 (red) signal overlaps with CLAMP (green) Mxc (pink) at the major *D. virilis* histone locus (25F) but not the minor locus (43C). (**D)** MSL2 is not present in female *D. virilis*. (**E)** The anti-*D. melanogaster* MSL2 antibody detects a ~81-85 kDa protein in both *D. melanogaster* and *D. virilis* embryo extracts, as well as a non-specific protein (*) present in both species. Samples are not concentration-normalized.

MSL2 is the structural component of MSLc and is usually found complexed with other members (Hallacli *et al.* 2012). However, MSL2 has some affinity for DNA sequence in both *D. melanogaster* and *D. virilis* (Villa *et al.* 2016, 2021) and CLAMP and MSL2 directly interact with each other through well-conserved domains (Tikhonova *et al.* 2022b), suggesting that CLAMP might recruit MSL2 outside of the complex. We therefore stained chromosome spreads for another MSLc member, MSL3. We did not observe recruitment of MSL3 to either of the histone loci in male *D. virilis* (**Figure 3**), indicating that MSL2 targets the major histone locus outside of its role in MSLc.

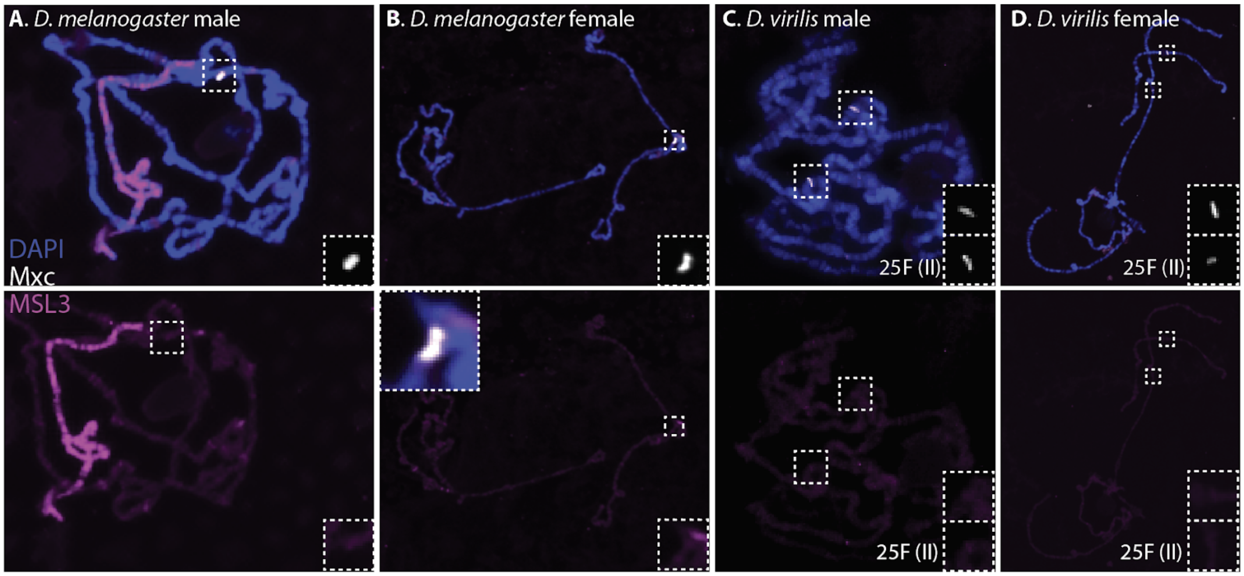


Figure 4.3: MSL3 does not target the histone locus in *Drosophila*. (A) MSL3 (purple) targets the X-chromosome in *D. melanogaster* males, but does not colocalize with Mxc (white). (**B)** MSL3 does not target loci on female *D. melanogaster* polytene chromosomes. (**C)** MSL3 does not target the major histone locus (25F) in *D. virilis* males. (**D)** MSL3 does not target loci on female *D. virilis* chromosomes.

We repeated our polytene experiments in *D. pseudoobscura* and *D. willistoni* (**Figure 4**), which both have a single histone locus. *D. melanogaster* and *D. pseudoobscura* diverged ~25 MYa, while *D. melanogaster* and *D. willistoni* diverged ~35 MYa (Powell 1997). We observed MSL2 specifically on the male X-chromosome in these species, indicating that MSLc targeting the histone genes is specific to *D. virilis* (**Supplemental Table 1**). These data suggest that MSLc specifically targets one of the two histone loci in *D. virilis*, a localization not observed in the other Drosophilids we investigated.

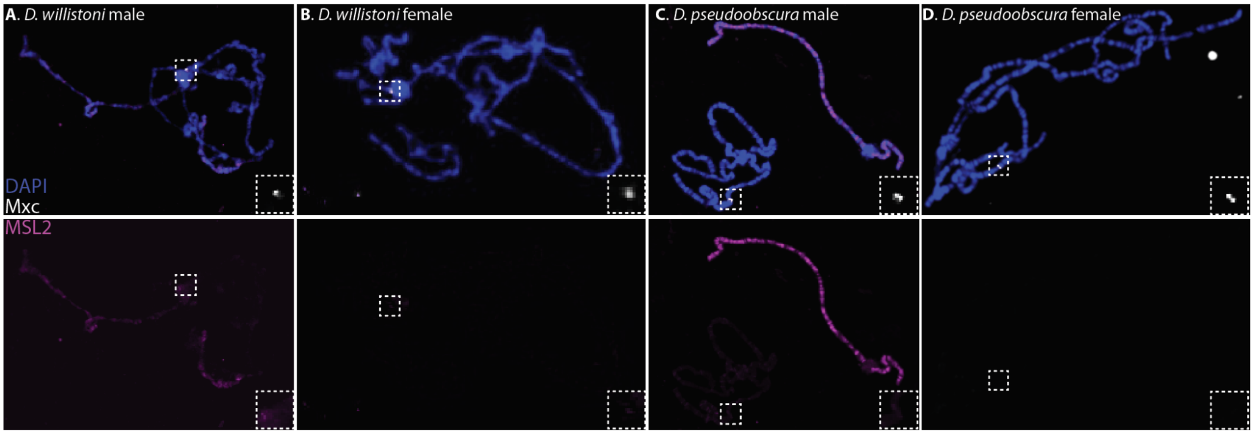


Figure 4.4: MSL2 does not target histone loci in other *Drosophila* species. MSL2 (purple) targets the X-chromosome in *D. willistoni* (**A**) and *D. pseudoobscura* males (**C**) but not females of either species (**B, D**). MSL2 does not co-localize with Mxc (white), which targets the single histone locus in both species.

4.4.4 MSL2 does not directly interact with histone array sequence

CLAMP targets the zygotic *D. melanogaster* histone locus by nuclear cycle 10 (Rieder *et al.* 2017), just prior to detectable Mxc nuclear puncta (White *et al.* 2007) and zygotic histone gene expression in nuclear cycle 11 (Edgar and Schubiger 1986). Similarly, CLAMP targets loci genome-wide, including sites on the male X-chromosome, by nuclear cycle 11 (Rieder *et al.* 2019) and MSLc localizes to the male X-chromosome by nuclear cycle 14 (Rastelli *et al.* 1995). Although MSLc X-chromosome targeting requires CLAMP (Soruco *et al.* 2013), MSL2 has some ability to interact directly with DNA sequence (Villa *et al.* 2016, 2021; Tikhonova *et al.* 2019). Specifically, MSL2^{mel} identifies a subset of X-linked sequences called PionX sites (Villa *et al.* 2016), which often include GA-rich MRE elements. Villa et al. (2021) discovered that the *D. virilis* MSL2 ortholog is able to interact with DNA, but does not show the same specificity for X-linked sequences as the *D. melanogaster* ortholog. We hypothesized that MSL2 might be found specifically at the *H3/H4* promoter within the histone array, since that is where the CLAMP protein targets GA-repeats (**Figure 4.1C**) (Rieder *et al.* 2017), and CLAMP targets X-linked GA-rich MREs prior to MSLc in *D. melanogaster* (Rieder *et al.* 2019).

No studies have examined MSLc localization in *D. virilis* using genomics techniques. However, Villa *et al.* (2021) overexpressed GFP-tagged *D. melanogaster* (MSL2^{mel}-GFP) and *D. virilis* (MSL2^{vir}-GFP) MSL2 in cultured female *D. melanogaster* Kc cells and performed chromatin immunoprecipitation followed by sequencing (ChIP-seq). We mapped these datasets to a single *D. melanogaster* histone gene array (McKay *et al.* 2015). Since the array units in *D. melanogaster* are virtually identical in sequence (Bongartz and Schloissnig 2019), sequencing data is collapsed from ~100 arrays onto a single array. Neither *D. melanogaster* nor *D. virilis* MSL2 targets a sequence in the *D. melanogaster* histone gene array (**Supplemental Figure 4.1**). Importantly, we noticed that the control anti-GFP ChIP-seq dataset from untreated cells gives a sharp peak over the *H2a/H2b* promoter (**Supplemental Figure 4.2**), which is also found in other datasets. This peak in the control indicates that the GFP antibody interacts with sequences at the histone locus and likely elsewhere, confounding conclusions. In addition, we mapped MSL3 ChIP-seq datasets after MSL2 transfection (Villa *et al.* 2021) and did not observe enrichment over the *D. melanogaster* histone gene array (**Supplemental Figure 4.3**).

Finally, MSLc deposits the activating H4K16ac histone mark on the male X-chromosome (Gelbart *et al.* 2009). If MSLc is depositing H4K16ac at the major *D. virilis* histone locus, this post-translational modification could affect histone expression from the locus, specifically in males. In *D. melanogaster*, CLAMP targets sites across the X-chromosome, followed by MSLc and the appearance of H4K16ac by nuclear cycle 14 (Rieder *et al.* 2019). We mapped available H4K16ac ChIP-seq datasets from staged *D. melanogaster* male embryos (Rieder *et al.* 2019) to the histone gene array and did not observe H4K16ac enrichment (**Supplemental Figure 4.4**). However, this finding is not surprising given that we also do not observe MSLc at the *D. melanogaster* histone gene array by polytene chromosome immunostaining (**Figure 4.2A**) or by ChIP-seq (**Supplemental Figure 4.1**). We also mapped H4K16ac ChIP-seq datasets after MSL2^{mel} and MSL2^{vir} transfection (Villa *et al.* 2021) and did not observe H4K16ac enrichment over the *D. melanogaster* histone gene array (**Supplemental Figure 4.5**), which is not surprising since MSL2, but not other MSLc members, is present at the *D. virilis* major locus (**Figures 4.2, 4.3**).

We conclude that MSL2 does not directly interact with histone array sequence in either *D. melanogaster* or *D. virilis* and that other MSLc components are unlikely to be present at the major *D. virilis* histone locus.

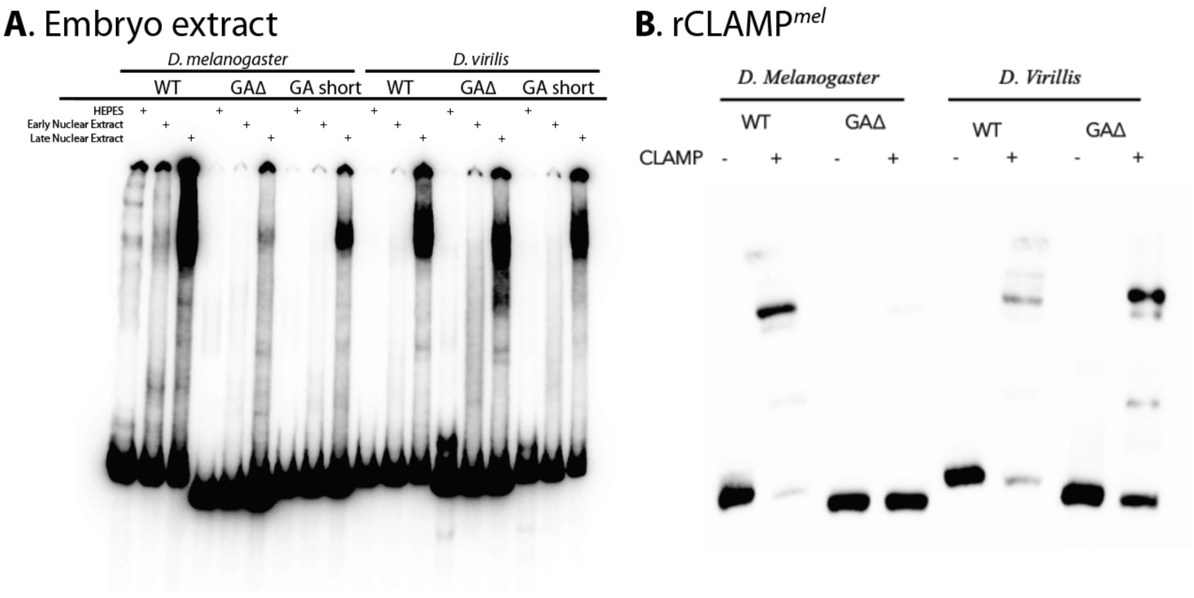
4.4.5 CLAMP does not require the GA-rich elements to interact with the *virilis* H3/H4 promoter *in vitro*

It is difficult to assay CLAMP^{mel} recruitment to single histone array transgenes (L. Hodkinson, observation) since CLAMP is present at loci genome-wide (**Figure 4.1A-B**) (Urban *et al.*). It is also possible that CLAMP is present at the 1xHis^{vir} transgene without interacting directly with DNA sequence, as was previously observed in histone array transgenes lacking GA-repeats (Koreski *et al.* 2020). We therefore turned to an *in vitro* approach to probe the interaction between CLAMP^{mel} and DNA sequence.

CLAMP is a member of the Late Boundary Complex (LBC) (Kaye *et al.* 2017) that forms in late embryogenesis (Wolle *et al.* 2015). We performed electrophoretic mobility shift assays using *D. melanogaster* late embryo extract and DNA probe sequences (**Supplemental Table**

4.2). We found that both *D. melanogaster* embryo extract, as well as recombinant full-length CLAMP^{mel} protein (Kuzu *et al.* 2016), shift both the wild-type *D. melanogaster* and *D. virilis H3/H4* sequences (**Figure 4.5**). The GA-rich *cis* elements in the *D. virilis* sequence are poorly conserved (**Figure 4.1C**) and there are other, short GA-repeats at other locations in the promoter. We therefore performed EMSAs using recombinant CLAMP^{mel} and 60 bp probes that tile the *D. virilis* promoter to confirm that CLAMP is targeting the region that contains the GA-rich *cis* element (**Supplemental Figure 4.6**).

Deleting or shortening the GA-repeats in the *D. melanogaster* sequence compromises the shifting of the *melanogaster* probe. However, even deleting the GA-rich elements entirely from the *D. virilis* probe does not compromise shifting with *D. melanogaster* late embryo extract or recombinant full-length CLAMP (**Figure 4.5A-B**). Our *in vitro* observations suggest that CLAMP^{mel}, and therefore likely CLAMP^{vir}, is directly interacting with the *D. virilis H3/H4* promoter sequence. In addition, CLAMP^{mel}, and therefore likely CLAMP^{vir}, can target non-GA-repeat sequences in the *D. virilis H3/H4*.



A. Embryo extract

B. rCLAMP^{mel}

Figure 4.5: CLAMP binds the *D. virilis H3/H4* sequence *in vitro* and does not require the GA-rich elements. (A) We shifted ^{32}P-labeled dsDNA probes using protein extract from early (0-6hr) and late (6-18hr) *D. melanogaster* embryos. The wild-type (WT) *D. melanogaster H3/H4* sequence is shifted, but the shift is weaker when the GA-repeats are shortened (GA short). *D. melanogaster H3/H4* probe does not shift when the GA-repeats are removed (GAΔ). *D. virilis H3/H4* probes shift with *D. melanogaster* late embryo extract, even when the GA-rich element is removed. (**B)** Recombinant full-length CLAMP^{mel} shifts the wild-type (WT) *D. melanogaster* and *D. virilis H3/H4* probes. Recombinant CLAMP continues to shift the *D. virilis* GAΔ probe. Probe sequences in **Supplemental Table 4.2**.

4.4.6 The *D. virilis H3/H4* promoter does not promote Mxc recruitment in *D. melanogaster*

Since CLAMP^{mel} is able to bind to the *D. virilis H3/H4* promoter sequence, we wondered if the *D. virilis* promoter might promote Mxc recruitment in *D. melanogaster*. It is difficult to manipulate the endogenous *D. melanogaster* histone locus, which includes ~100 nearly-identical histone gene arrays (McKay *et al.* 2015; Bongartz and Schloissnig 2019). However, wild-type *D. melanogaster* histone array transgenes recruit all tested HLB factors and express histone genes similar to the endogenous histone locus (Salzler *et al.* 2013; Rieder *et al.* 2017; Koreski *et al.* 2020) (1xHis^{WT}; **Figure 4.6A**). Histone array transgenes are therefore a powerful tool with which to interrogate DNA sequence contribution to histone locus identification.

As expected, deleting the GA-repeats from the *D. melanogaster H3/H4* promoter leads to failure to recruit the critical HLB scaffolding protein Mxc (1xHis$^{GA\Delta}$; **Figure 4.6B**) (Rieder *et al.* 2017). We recently discovered that replacing the perfect GA-repeat sequence in a histone array transgene with X-linked CLAMP-binding *cis* elements abrogates HLB factor recruitment to the transgene (L. Hodkinson, observation), indicating the critical nature of the *cis* element sequence itself, rather than just the ability to recruit CLAMP.

The *D. virilis* GA-repeats in the *H3/H4* promoter are much shorter than that of *D. melanogaster* (**Figure 4.1C**) and they more closely resemble the X-linked GA-rich MREs involved in male dosage compensation (Alekseyenko *et al.* 2008; Villa *et al.* 2016). Although there are sequence differences between arrays, and even more differences between the major and minor loci, the GA-rich sequences are present in most arrays (**Supplemental Figure 4.7**). We leveraged the transgenic histone array system to determine if the *D. virilis H3/H4* sequence is able to support Mxc recruitment in *D. melanogaster*.

We engineered a histone array transgene that replaces the majority of the *D. melanogaster H3/H4* promoter with a sequence from *D. virilis* (1xHis^{vir}; **Supplemental Table 4.3**). We also engineered a transgene with the *D. melanogaster* promoter sequence but shortened GA-repeats (1xHis^{GA}), using sequences similar to our *in vitro* gel shift assays (**Supplemental Table 4.2**). We discovered that neither transgene is able to recruit Mxc (**Figure 4.6C-D**). Since the *D. virilis* sequence is bound by CLAMP^{mel} *in vitro* (**Figure 4.5B**), our observations suggest that CLAMP targets the 1xHis^{vir} transgene but is unable to recruit Mxc.

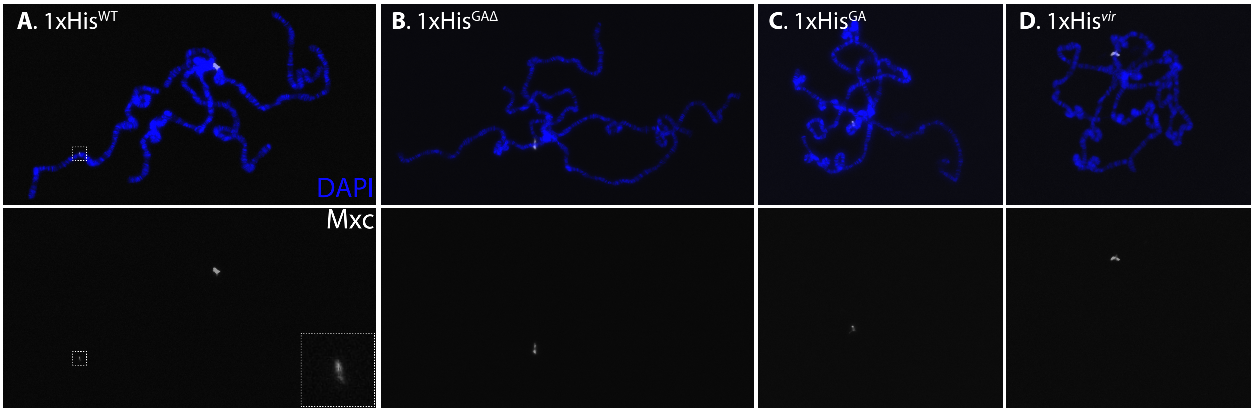


Figure 4.6: The *D. virilis H3/H4* promoter does not promote HLB formation in *D. melanogaster*. We performed polytene immunostaining for Mxc (white; bottom panels) in animals carrying 1x histone array transgenes. (**A)** A wild-type transgene attracts Mxc (inset). (**B)** Deleting the GA-repeats abrogates Mxc recruitment. (**C)** Shortening the GA-repeats also abrogates Mxc recruitment to the transgene. (**D)** Replacing the *H3/H4* promoter with that of *D. virilis* does not support Mxc recruitment. Transgene sequences in **Supplemental Table 4.3.**

4.5 Discussion

Here we show that MSL2 targets the major histone locus in *D. virilis*, but does not target the single histone locus in other *Drosophila* species. We propose that this is due to CLAMP interacting with the *D. virilis H3/H4* promoter sequence in a different manner than it does in *D. melanogaster*. The GA-rich element in the *D. virilis H3/H4* promoter more closely resembles GA-rich X-linked MREs (Alekseyenko *et al.* 2008) than it does the perfect, long GA-repeat of the *D. melanogaster H3/H4* promoter (Rieder *et al.* 2017). We recently discovered that X-linked MREs can drive MSL2 recruitment in the context of a transgenic histone gene array in *D. melanogaster* (L. Hodkinson, observation). However, our *in vitro* observations indicate that this element may even be dispensable for CLAMP interactions in *D. virilis*.

We previously observed that both *D. virilis* histone loci are targeted by CLAMP (Rieder *et al.* 2017). It is curious that we only observe MSL2 at the major locus. There are multiple suggested mechanisms to HLB formation, even in *D. melanogaster*; the GA-repeats are required in transgenic histone gene arrays for localization of CLAMP, Mxc, and other factors in *D. melanogaster*, as long as the transgene is in the background of the endogenous histone locus (Rieder *et al.* 2017). Transgenic arrays lacking the GA-repeats are targeted by Mxc only when the endogenous locus is deleted (Koreski *et al.* 2020). CLAMP is present in transgenic HLBs lacking GA-repeats by polytene chromosome immunostaining, but it does not interact with specific DNA sequences by ChIP-seq (Koreski *et al.* 2020). In addition to a zinc-finger domain, CLAMP contains a disordered prion-like domain (Kaye *et al.* 2018; Tikhonova *et al.* 2022a) that may facilitate dimerization and/or inclusion into the phase-separated HLB (Hur *et al.* 2020) likely through protein-protein interactions (Staller 2022).

This is not the first example of a degenerate *cis* elements facilitating a conserved interaction at the histone locus. In humans, octamer binding transcription factor 1 (Oct-1) controls S-phase H2B expression by targeting an 8 bp "octamer" element in the promoter (Zheng *et al.* 2003). Pdm-1/Nubbin is the Oct-1 counterpart in *Drosophila*, yet only cryptic octamer elements are found in both *H2B* and *H4* promoters and Pdm-1 influences expression of all core histone genes (Lee *et al.* 2010). Although humans and *Drosophila* share similar cell cycle needs for histone expression, histone octamer elements are conserved in vertebrates but not amongst *Drosophila* species. These observations led Lee *et al.* (2010) to suggest that invertebrates have greater tolerance for histone regulatory sequence flexibility, compared to vertebrates.

CLAMP is likely an evolutionary ancient protein that has been co-opted for multiple distinct functions (Kuzu *et al.* 2016; Rieder *et al.* 2017). CLAMP targets locations across the genome during very early embryogenesis. It targets the *D. melanogaster* histone locus around the same time as Mxc (Rieder *et al.* 2017; Kemp *et al.* 2021) and X-linked MREs prior to MSLc (Rieder *et al.* 2019). CLAMP and MSLc synergistically enrich each other's occupancy on the male X-chromosome *in vivo* (Soruco *et al.* 2013; Soruco and Larschan 2014) and *in vitro* (Albig *et al.* 2019). We were surprised that we did not detect MSL2 on the male *D. virilis* X-chromosome, as was previously reported (Marín *et al.* 1996). Despite relatively low protein conservation (Copps *et al.* 1998; Kuzu *et al.* 2016), anti-MSL2^{mel} antibodies have long been used to assay other species. We discovered that MSL2^{vir} appears slightly larger than expected by western blot, which may indicate a post-translational modification on the majority of MSL2^{vir} that interferes with antibody-based detection.

However, MSLc components are not always confined to the X-chromosome. Males absent on the first (MOF) is an MSLc member that is present in other non-sex-specific complexes (Feller *et al.* 2012; Lam *et al.* 2012), while the Maleless helicase (MLE) member of the complex is expressed in both males and females and plays a role in RNA structure and splicing (Reenan *et al.* 2000). MSL2 is a core member of MSLc and is usually only present on the male X-chromosome. However, MSL2 mis-localizes to tandem repeats when the long non-coding *RNA on the X* (*roX*) components of the MSLc are missing in *D. melanogaster* (Figueiredo *et al.* 2014). Compromising the ability of MSL2 to interact with both CLAMP and DNA, via mutation of the MSL2 CLAMP-binding domain (CBD) and CXC domain, respectively, results in loss of complex X-chromosome specificity (Tikhonova *et al.* 2019). Both of these domains are well conserved between *D. melanogaster* and *D. virilis* (Villa *et al.* 2021). CLAMP and MSLc may search for genomic loci together as a “wolf pack” (Staller 2022) and have a higher affinity for sites resembling X-linked MREs. This model is supported by our recent observations that CLAMP targets loci genome-wide, and is followed by MSLc during very early *D. melanogaster* development, prior to MSLc X-chromosome specialization (Rieder *et al.* 2019).

While MSL2 is usually found complexed with other MSLc members (Hallacli *et al.* 2012), it retains some DNA binding ability in both *D. melanogaster (Villa et al. 2016)* and *D. virilis* (Villa *et al.* 2021). Further, White *et al.* (2011) identified MSL2, although not other MSLc proteins, in a proteomics screen for phosphorylated Mxc in cultured male *D. melanogaster* S2 cells. Mxc is phosphorylated by CyclinE/Cdk2, which activates *histone* gene expression and HLB phase separation (Wei *et al.* 2003; Hur *et al.* 2020). These observations provide evidence for the presence of MSL2 at histone loci, even sometimes in *D. melanogaster*.

Yet several lines of evidence argue against the presence of MSL2 at *D. melanogaster* histone genes. We do not observe MSL2 targeting the histone locus in *D. melanogaster in vivo* by polytene chromosome immunostaining or MSLc member ChIP-seq from multiple tissues. *D. melanogaster* polytene chromosome proximity ligation assays indicate that CLAMP only interacts with MSLc on the male X-chromosome (Lindehell *et al.* 2015). The above observations include both polytene chromosome immunostaining and analysis of sequencing datasets from cultured cells. Although the histone loci of polytene chromosomes appear to recruit all known HLB factors (Salzler *et al.* 2013; Rieder *et al.* 2017; Koreski *et al.* 2020), salivary gland nuclei undergo endoreplication without cell division and therefore might have unusual histone biogenesis requirements (Andreyeva *et al.* 2017). Cultured cells are asynchronous, and MSL2 might target the *D. melanogaster* histone locus at a specific cell cycle time point or in a subset of cell types, confounding results. Therefore, it is unlikely that the tissues examined here represent the histone regulatory needs of all tissues across developmental time.

It is curious that MSL2 targets only the major *D. virilis* histone locus on Chromosome II, which includes ~32 arrays of both the quintet and quartet organization (Schienman *et al.* 1998). The minor locus on Chromosome IV, which includes only ~5 quintet arrays, is targeted by CLAMP and Mxc but not MSL2. These observations indicate that the two histone loci may be differentially regulated in male *D. virilis*. There is no direct evidence that the ~100 nearly identical histone arrays in *D. melanogaster* are differentially regulated, however 100 copies are not required for viability; 12 transgenic arrays rescue viability when the endogenous locus is deleted (Günesdogan *et al.* 2014; McKay *et al.* 2015; Zhang *et al.* 2019).

It is not uncommon for histone genes to experience differential regulation. For example, the sea urchin genome carries three sets of histone genes: 2000 "early" ɑ-histone genes are expressed in early embryogenesis, 35 "late" histone genes are expressed in somatic cells, and three histone genes are testes-specific (Marzluff *et al.* 2006). *Saccharomyces cerevisiae* mutants lacking one *H2a/H2b* unit (*TRT1*) cannot undergo mitosis, while those lacking the other unit (*TRT2*) have no dramatic phenotypes, indicating differential histone biogenesis from the two loci (Norris and Osley 1987; Cross and Smith 1988). The two histone loci in *D. virilis* may fulfill different developmental or cell cycle needs for histone production.

MSL2 is only expressed in XY *Drosophila*; *msl2* translation is repressed by Sex Lethal in XX individuals (Bashaw and Baker 1995; Kelley *et al.* 1997). The presence of MSL2 at the major histone locus in male, but not female, *D. virilis* indicates that histone genes might be differentially regulated between males and females. Although there is little current evidence of differential histone gene regulation between males and females, differences in sex chromosome size and chromatin content, and the existence of dosage compensation suggest that XX and XY individuals are likely to have different histone requirements. In addition, the requirement for maternal deposition of histone proteins (Horard and Loppin 2015) and the presence of maternal-effect *histone* gene-specific transcription factors such as *abnormal oocyte* (Berloco *et al.* 2001) indicate that males and females likely regulate histone loci differently, since female nurse cells must produce large amounts of histone mRNAs and proteins for egg deposition (Horard and Loppin 2015).

CLAMP participates in both male dosage compensation and histone biogenesis and therefore crosstalk between these gene regulatory networks (Friedlander *et al.* 2016) could occur in males but not in females. Factors are often shared between membraneless compartments (also called nuclear bodies). For example, Coilin is shared between Cajal and histone locus bodies at different developmental time points in *Drosophila* and *Xenopus* (Liu *et al.* 2006; Nizami *et al.* 2010). Nucleolin, fibrillarin, and other nucleolus factors are found in the Cajal body (Trinkle-Mulcahy and Sleeman 2017).

Overall, our results indicate that the two histone loci in *D. virilis* may be differentially regulated in males and females. The recruitment of MSL2 to the major *D. virilis* histone locus may be due to differential interactions between local DNA sequence and the CLAMP factor. Finally, CLAMP is shared between several compartments, which may lead to cross talk between gene regulatory networks.

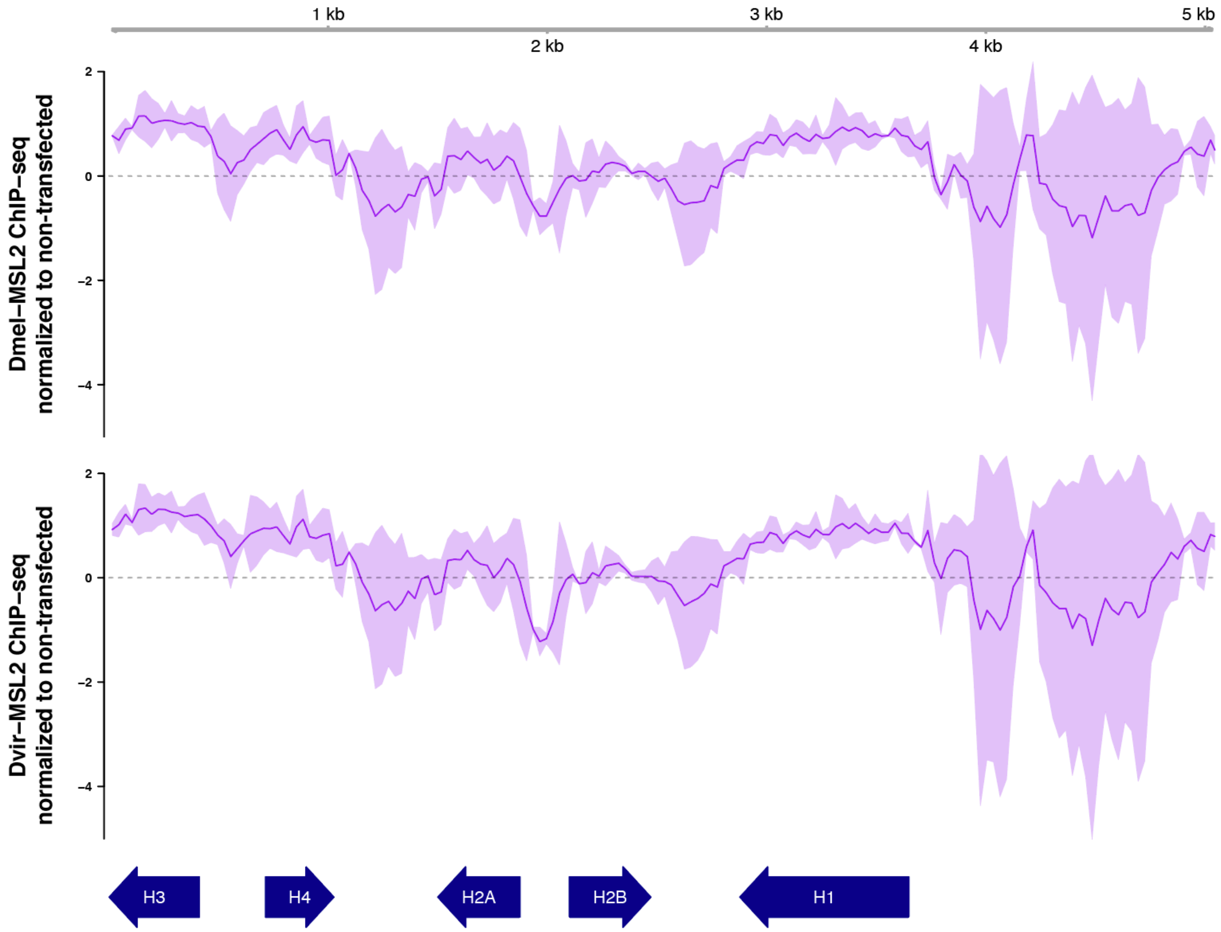
4.6 Acknowledgments

We thank John Ali for generating the original Galaxy pipeline. We thank Dr. Bob Duronio and Dr. William Marzluff for the anti-Mxc antibody, Dr. Victoria Meller (originally Dr. Ron Richmond) for the affinity purified anti-MSL2 antibody, Dr. Mitzi Kuroda for the anti-MSL2 serum, and Dr. Erica Larschan (originally Dr. Mitzi Kuroda) for the anti-MSL3 serum. We thank fellow group members Dr. Casey Schmidt, Tommy O'Haren, Greg Kimmerer, Eric Albanese, and Edgar Hsieh, who all supported this project. This work was supported by an Emory Summer Undergraduate Research Experience (SURE) Fellowship to MX, grants T32GM008490 and F31HD105452 to LJH, and grants K99HD092625, R00HD092625, and R35GM142724 (including S1) to LER.

4.7 Supplemental Figures and Tables:

Supplemental Table 4.1: Summary of MSL protein localization from polytene chromosome immunostaining experiments, related to Figures 4.2 and 4.4

Species	M/F	MSL2 at histone locus/loci?	MSL2 detected on X?	MSL3 at histone locus/loci?	MSL3 detected on X?
D. melanogaster	M	No	**Yes**	No	**Yes**
	F	No	No	No	No
D. virilis	M	**Yes**	No	**Yes**	No
	F	No	No	No	No
D. pseudoobscura	M	No	**Yes**	Not tested	Not tested
	F	No	No	Not tested	Not tested
D. willistoni	M	No	**Yes**	Not tested	Not tested
	F	No	No	Not tested	Not tested

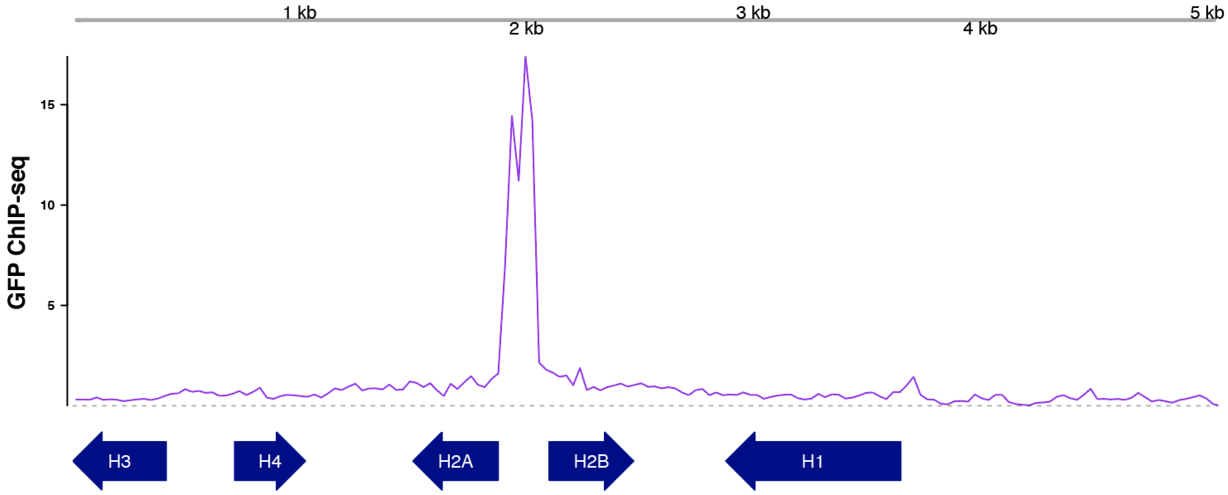


Supplemental Figure 4.1: Ectopically expressed MSL2 from either species does not target the *D. melanogaster* histone gene array. Villa *et al.* (2021) overexpressed GFP-tagged MSL2 from *D. melanogaster* and *D. virilis* in female *D. melanogaster* cell culture. They performed GFP ChIP-seq and we mapped their datasets to the *D. melanogaster* histone gene array, normalizing to the non-transfected control. Neither *D. melanogaster* MSL2 (top) nor *D. virilis* MSL2 (bottom) target a specific DNA sequence in the *D. melanogaster* histone gene array. The light purple indicates spread between two biological replicates, while the dark purple line indicates the average between replicates.

Supplemental Table 4.2: dsDNA EMSA probes, related to Figures 4.5 and S4.7

Probe (length)	Probe sequence
Wild-type *Dmel* **GA-repeats in bold** (226 bp; **Fig 6**)	CACAGCACGAAAGTCACTAAAGAACTAATTTCAACGTTTCTG TGTGCCCCTATTTATAGGTAAAACGACAAAAACCC**GAGAGAG** TACGAACGATATGTTCGTTCGCTTTTCGCTCGTCAAATGAAAT GGCCTCTGTTTT**TCTCTCTCTCTCTCTCTCTCT**TTCACCGTCC ACGATTGCTATATAAGTAGGTAGCAAATGCTCTGATCGTTTAT TGTGTTTTCAAAC
Dmel GA **GA-repeats in bold** (211 bp; **Fig 6**)	CACAGCACGAAAGTCACTAAAGAACTAATTTCAACGTTTCTG TGTGCCCCTATTTATAGGTAAAACGACAAAAACCCT**AGAGA**C TACGAACGATATGTTCGTTCGCTTTTCGCTCGTCAAATGAAAT GGCCTCTGTTTT**TCTCT**TTCACCGTCCACGATTGCTATATAAG TAGGTAGCAAATGCTCTGATCGTTTATTGTGTTTTCAAAC
Dmel GAΔ **Locations of deletions (X)** (198 bp; **Fig 6**)	CACAGCACGAAAGTCACTAAAGAACTAATTTCAACGTTTCTG TGTGCCCCTATTTATAGGTAAAACGACAAAAACCC**X**TACGAA CGATATGTTCGTTCGCTTTTCGCTCGTCAAATGAAATGGCCTC TGTTTT**X**TTCACCGTCCACGATTGCTATATAAGTAGGTAGCAA ATGCTCTGATCGTTTATTGTGTTTTCAAAC
Wild-type *Dvir* **GA-repeats in bold** (235 bp; **Fig 6**)	CACCACGAATGTCACTGAGGTACTAATGCTAGCTCTTCGGGC AGCGCTTATATTTATACCAAAAACCAAAA**AGACGAG**CGAGT GAAAACATATTTCCATCTCGCTCACATACTACCCTTGTAACAT ATTCGACAAAACAGCGAACAGCGAATATATCGT**TCTCT**TTCT AACTTATCACTCATTTTCTATATAAGCGATACACAAACGAGAC GCACGATTATTGTGTTTTTAACA
Dvir GA **GA-repeats in bold** (231 bp; **Fig 6**)	CACCACGAATGTCACTGAGGTACTAATGCTAGCTCTTCGGGC AGCGCTTATATTTATACCAAAAACTCAAAAAGAC**AGAGA**TGA AAACATATTTCCAT**TCTCT**ACATACTACCCTTGTAACATATTC GACAAAACAGCGAACAGCGAATATATCGTTCTCTTTCTAACTT ATCACTCATTTTCTATATAAGCGATACACAAACGAGACGCAC GATTATTGTGTTTTTAACA
Dvir GAΔ **Locations of deletions (X)** (221 bp; **Fig 6**)	CACCACGAATGTCACTGAGGTACTAATGCTAGCTCTTCGGGC AGCGCTTATATTTATACCAAAAACTCAAAAAGAC**X**TGAAAAC ATATTTCCAT**X**ACATACTACCCTTGTAACATATTCGACAAAAC AGCGAACAGCGAATATATCGTTCTCTTTCTAACTTATCACTCA TTTTCTATATAAGCGATACACAAACGAGACGCACGATTATTGT GTTTTTAACA
Wild-type *Dmel* **GA-repeats in bold**	TGAAATGGCCTCTGTTTT**TCTCTCTCTCTCTCTCTCTCTCT**TT CACCGTCCACGATTGCT

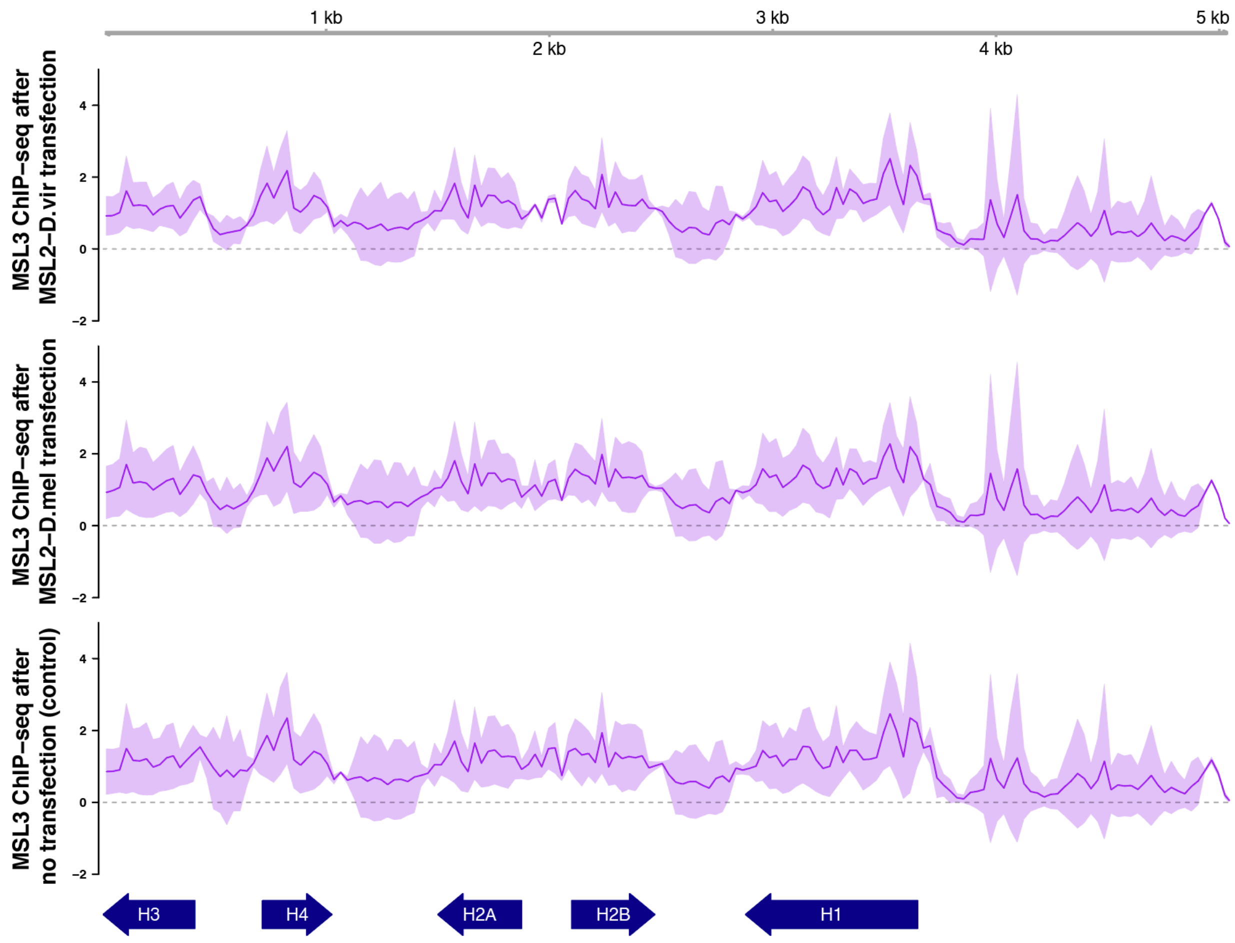
(60 bp; **Fig S7**)	
Dvir 1 **GA-repeats in bold** (60 bp; **Fig S7**)	CACCACGAATGTCACTGAGGTACTAATGCTAG**CTCT**TCGGGCAGCGCTTATATTTATACC
Dvir 2 **GA-rich elements in bold** (60 bp; **Fig S7**)	TACCAAAAACTCAAAAA**GACGAG**CGAGTGAAAACATATTTCCAT**CTCGCTC**ACATACTAC
Dvir 3 **GA-repeats in bold** (60 bp; **Fig S7**)	CATACTACCCTTGTAACATATTCGACAAAACAGCGAACAGCGAATATATCGT**TCTCT**TTC
Dvir 4 **GA-repeats in bold** (60 bp; **Fig S7**)	TATCGT**TCTCT**TTCTAACTTATCACTCATTTTCTATATAAGCGATACACAAAC**GAGA**CGC
Dvir 5 **GA-repeats in bold** (60 bp; **Fig S7**)	TCACTCATTTTCTATATAAGCGATACACAAAC**GAGA**CGCACGATTATTGTGTTTTTAACA



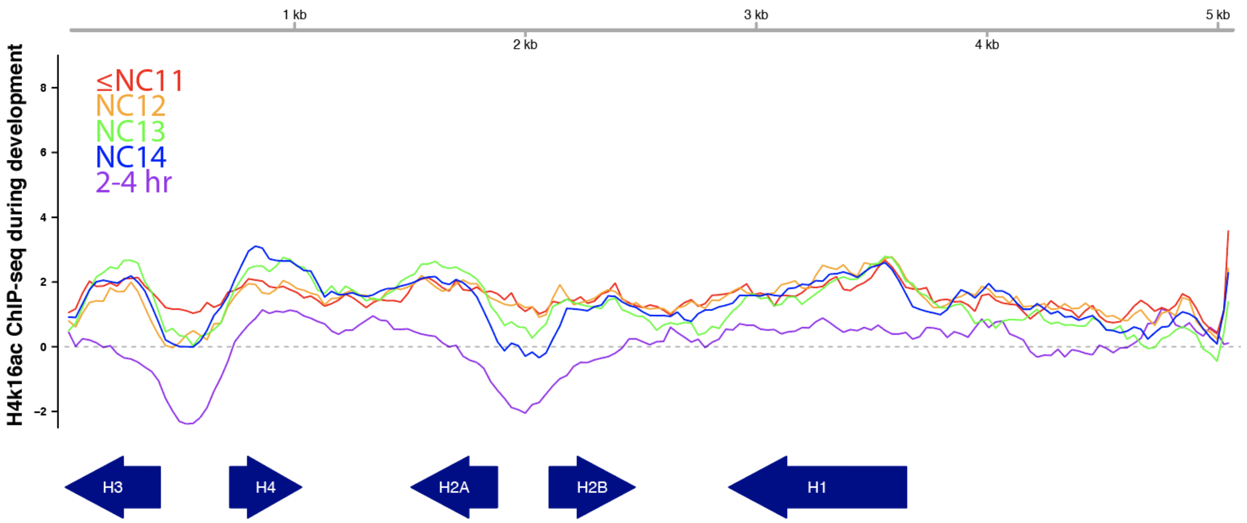
Supplemental Figure 4.2: Anti-GFP ChIP-seq from non-transfected cells gives a peak in the *H2a/H2b* promoter. Villa *et al.* (2021) overexpressed GFP-tagged MSL2 from *D. melanogaster* and *D. virilis* in female *D. melanogaster* cell culture. They performed GFP ChIP-seq in control, non-transfected cells and we mapped their datasets to the *D. melanogaster* histone gene array. We discovered a sharp peak in the *H2a/H2b* promoter, which is present in all anti-GFP ChIP-seq datasets.

Supplemental Table 4.3: *H3/H4* promoter sequences from *D. melanogaster* histone array transgenes, related to Figure 4.6

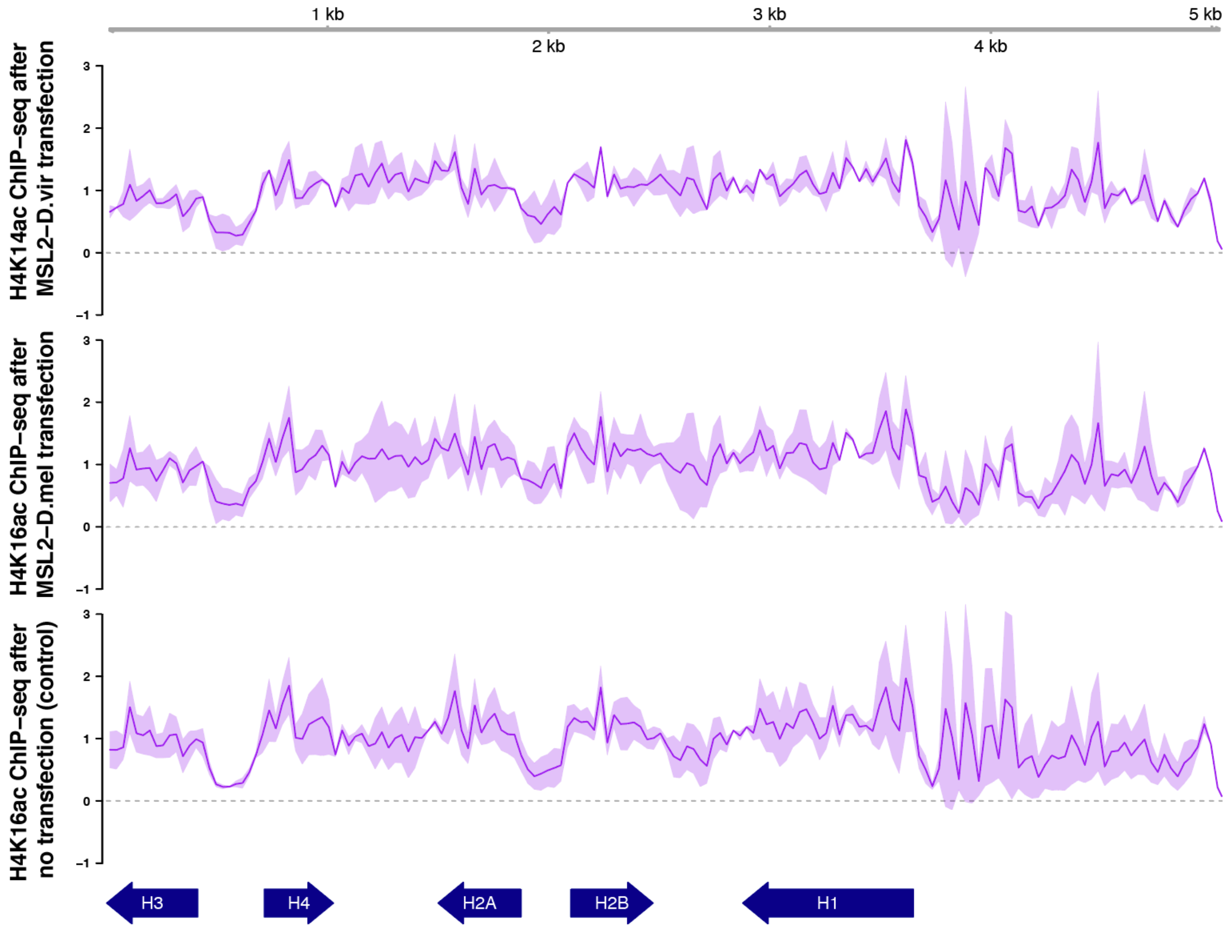
Transgene	Promoter sequence (between *H4* and *H3* start codons)
1xHis^{WT} (Hodkinson et al.) **GA-repeats in bold** (239 bp)	TTTTCACTGTTCTATACTATTATACACGCACAGCACGAAAGTC ACTAAAGAACTAATTTCAACGTTTCTGTGTGCCCCTATTTATA GGTAAAACGACAAAAACCC**GAGAGAG**TACGAACGATATGTT CGTTCGCTTTTCGCTCGTCAAATGAAATGGCCTCTGTTTT**TCT** **CTCTCTCTCTCTCTCTCT**TTCACCGTCCACGATTGCTATATA AGTAGGTAGCAAATGCTCTGATCGTTT
1xHis$^{GA\Delta}$ (Hodkinson et al.) **Locations of deletions (X)** (211 bp)	TTTTCACTGTTCTATACTATTATACACGCACAGCACGAAAGTC ACTAAAGAACTAATTTCAACGTTTCTGTGTGCCCCTATTTATA GGTAAAACGACAAAAACCC**X**TACGAACGATATGTTCGTTCGC TTTTCGCTCGTCAAATGAAATGGCCTCTGTTTT**X**TTCACCGTC CACGATTGCTATATAAGTAGGTAGCAAATGCTCTGATCGTTT
1xHis^{GA} **GA-repeats in bold** (222 bp)	TTTTCACTGTTCTATACTATTATACACGCACAGCACGAAAGTC ACTAAAGAACTAATTTCAACGTTTCTGTGTGCCCCTATTTATA GGTAAAACGACAAAAACCCT**AGAGA**TACGAACGATATGTTC GTTCGCTTTTCGCTCGTCAAATGAAATGGCCTCTGTTTT**TCTC** **T**TTCACCGTCCACGATTGCTATATAAGTAGGTAGCAAATGCTC TGATCGTTT
1xHis^{vir} **GA-rich elements in bold** (297 bp)	TTTTCACTTTATATTTTTTTTTAACTTAACACCACGAATGTCAC TGAGGTACTAATGCTAGCTCTTCGGGCAGCGCTTATATTTATA CCAAAAACTCAAAA**AGACGAGCGAG**TGAAAACATATTTCCA **TCTCGCTC**ACATACTACCCTTGTAACATATTCGACAAAACAG CGAACAGCGAATATATCGTTCTCTTTCTAACTTATCACTCATT TTCTATATAAGCGATACACAAACGAGACGCACGATTATTGTG TTTTTAACAGTGACAGTGTGAAGTTGGAATTGTGAAAGAAAG



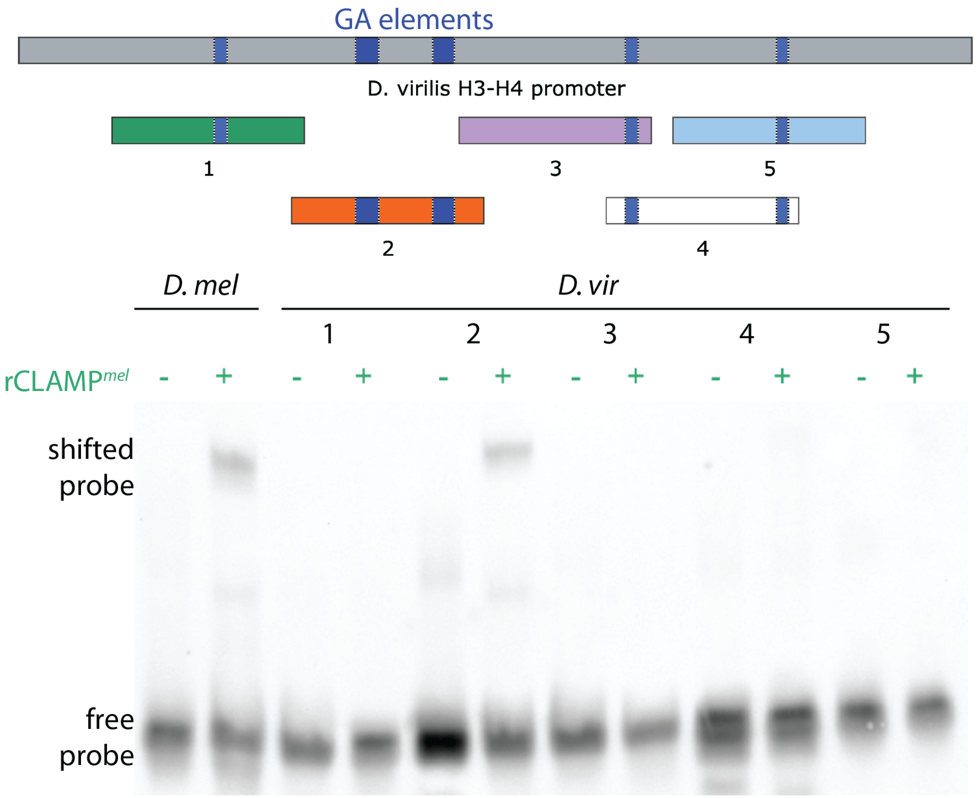
Supplemental Figure 4.3: MSL3 does not target the *D. melanogaster* histone array after MSL2 transfection. Villa *et al*. (2021) overexpressed MSL2 from *D. melanogaster* and *D. virilis* in female *D. melanogaster* cell culture. They performed MSL3 ChIP-seq and we mapped their datasets to the *D. melanogaster* histone gene array. MSL3 does not target the *D. melanogaster* histone gene array after transfection from either *D. virilis* MSL2 (top) nor *D. melanogaster* MSL2 (middle). These datasets look similar to MSL3 ChIP-seq from the non-transfected control (bottom). The light purple indicates spread between two biological replicates, while the dark purple line indicates the average between replicates.



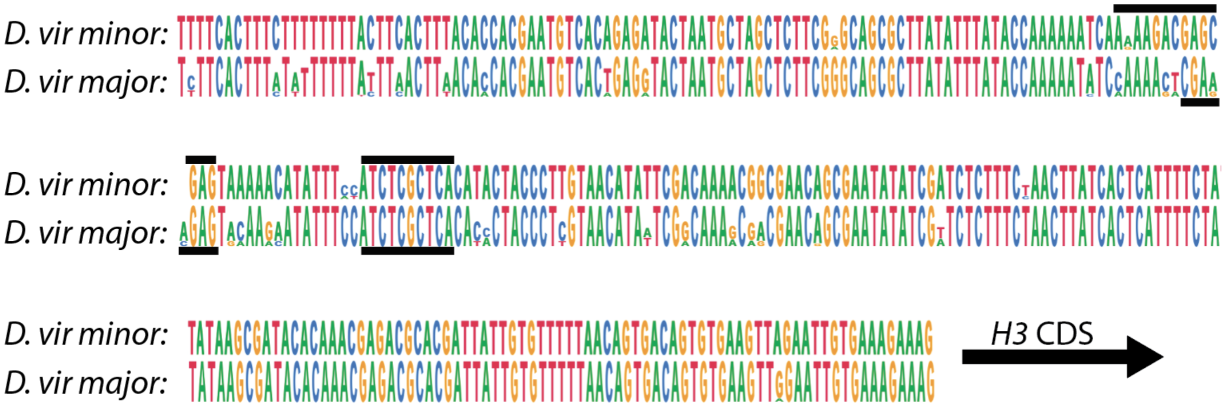
Supplemental Figure 4.4: H4K16ac is not enriched over the *D. melanogaster* histone array during male *D. melanogaster* embryogenesis. Rieder *et al*. (2019) performed male embryo H4K16ac ChIP-seq over a tight developmental time course (nuclear cycles = NC). We mapped single replicate datasets from this study to the *D. melanogaster* histone gene array, normalized to input samples. We observe no enrichment of H4K16ac over the *D. melanogaster* histone gene array.



Supplemental Figure 5: H4K16ac is not enriched over the *D. melanogaster* histone array after MSL2 transfection. Villa *et al.* (2021) overexpressed MSL2 from *D. melanogaster* and *D. virilis* in female *D. melanogaster* cell culture. They performed H4K16ac ChIP-seq and we mapped their datasets to the *D. melanogaster* histone gene array. H4K16ac is not enriched over the *D. melanogaster* histone gene array after transfection from either *D. virilis* MSL2 (top) nor *D. melanogaster* MSL2 (middle). These datasets look similar to H4K16ac ChIP-seq from the non-transfected control (bottom). The light purple indicates spread between two biological replicates, while the dark purple line indicates the average between replicates.



Supplemental Figure 6: CLAMP interacts with the region containing the poorly conserved GA-rich *cis* elements in the *D. virilis* promoter. The *D. virilis* promoter carries poorly conserved GA-rich elements (blue) as well as shorter GA-repeats (4 bp each; blue). We segmented the promoter into 60 bp probes (**Supplemental Table 2**) and performed EMSAs with recombinant CLAMP^{mel}. We confirmed that only probe 2 (orange), which carries the projected GA-rich *cis* elements, shifts with recombinant CLAMP.



Supplemental Figure 4.7: Conservation of the H3/H4 promoter within and between *D. virilis* loci. We aligned sequences in SnapGene using Coffee (Erb *et al.* 2012). The minor (top) *D. virilis* locus includes five intact *H3/H4* promoters, while the major (bottom) includes 30 similar promoters. Conservation within loci is indicated by nucleotide height. GA-rich sequences shown in **Figure 4.1** are indicated by black bars. *H3* CDS represents the coding sequence, beginning with the start codon, of the *H3* gene.

4.8 References

Afgan, E., D. Baker, B. Batut, M. van den Beek, D. Bouvier *et al.*, 2018 The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 46: W537–W544.

Albig, C., E. Tikhonova, S. Krause, O. Maksimenko, C. Regnard *et al.*, 2019 Factor cooperation for chromosome discrimination in Drosophila. Nucleic Acids Res. 47: 1706–1724.

Alekseyenko, A. A., C. E. Ellison, A. A. Gorchakov, Q. Zhou, V. B. Kaiser *et al.*, 2013 Conservation and de novo acquisition of dosage compensation on newly evolved sex chromosomes in Drosophila. Genes Dev. 27: 853–858.

Alekseyenko, A. A., S. Peng, E. Larschan, A. A. Gorchakov, O.-K. Lee *et al.*, 2008 A sequence motif within chromatin entry sites directs MSL establishment on the Drosophila X chromosome. Cell 134: 599–609.

Amodeo, A. A., D. Jukam, A. F. Straight, and J. M. Skotheim, 2015 Histone titration against the genome sets the DNA-to-cytoplasm threshold for the Xenopus midblastula transition. Proc. Natl. Acad. Sci. U. S. A. 112: E1086–95.

Andreyeva, E. N., T. J. Bernardo, T. D. Kolesnikova, X. Lu, L. A. Yarinich *et al.*, 2017 Regulatory functions and chromatin loading dynamics of linker histone H1 during endoreplication in Drosophila. Genes Dev. 31: 603–616.

Aoki, T., S. Schweinsberg, J. Manasson, and P. Schedl, 2008 A stage-specific factor confers Fab-7 boundary activity during early embryogenesis in Drosophila. Mol. Cell. Biol. 28: 1047–1060.

Bashaw, G. J., and B. S. Baker, 1995 The msl-2 dosage compensation gene of Drosophila encodes a putative DNA-binding protein whose expression is sex specifically regulated by

Sex-lethal. Development 121: 3245–3258.

Berloco, M., L. Fanti, A. Breiling, V. Orlando, and S. Pimpinelli, 2001 The maternal effect gene, abnormal oocyte (abo), of Drosophila melanogaster encodes a specific negative regulator of histones. Proc. Natl. Acad. Sci. U. S. A. 98: 12126–12131.

Bone, J. R., and M. I. Kuroda, 1996 Dosage compensation regulatory proteins and the evolution of sex chromosomes in Drosophila. Genetics 144: 705–713.

Bongartz, P., and S. Schloissnig, 2019 Deep repeat resolution—the assembly of the Drosophila Histone Complex. Nucleic Acids Res. 47: e18–e18.

Chari, S., H. Wilky, J. Govindan, and A. A. Amodeo, 2019 Histone concentration regulates the cell cycle and transcription in early development. Development 146.:

Conrad, T., F. M. G. Cavalli, H. Holz, E. Hallacli, J. Kind *et al.*, 2012 The MOF chromobarrel domain controls genome-wide H4K16 acetylation and spreading of the MSL complex. Dev. Cell 22: 610–624.

Copps, K., R. Richman, L. M. Lyman, K. A. Chang, J. Rampersad-Ammons *et al.*, 1998 Complex formation by the Drosophila MSL proteins: role of the MSL2 RING finger in protein complex assembly. EMBO J. 17: 5409–5417.

Cross, S. L., and M. M. Smith, 1988 Comparison of the structure and cell cycle expression of mRNAs encoded by two histone H3-H4 loci in Saccharomyces cerevisiae. Mol. Cell. Biol. 8: 945–954.

Di Liegro, C. M., G. Schiera, and I. Di Liegro, 2018 H1.0 Linker Histone as an Epigenetic Regulator of Cell Proliferation and Differentiation. Genes 9.:

Duan, J., L. Rieder, M. M. Colonnetta, A. Huang, M. Mckenney *et al.*, 2021 Author response: CLAMP and Zelda function together to promote Drosophila zygotic genome activation.

Duronio, R. J., and W. F. Marzluff, 2017 Coordinating cell cycle-regulated histone gene expression through assembly and function of the Histone Locus Body. RNA Biol. 14: 726–738.

Edgar, B. A., and G. Schubiger, 1986 Parameters controlling transcriptional activation during early Drosophila development. Cell 44: 871–877.

Erb, I., J. R. González-Vallinas, G. Bussotti, E. Blanco, E. Eyras *et al.*, 2012 Use of ChIP-Seq data for the design of a multiple promoter-alignment method. Nucleic Acids Res. 40: e52.

Feller, C., M. Prestel, H. Hartmann, T. Straub, J. Söding *et al.*, 2012 The MOF-containing NSL complex associates globally with housekeeping genes, but activates only a defined subset. Nucleic Acids Res. 40: 1509–1522.

Figueiredo, M. L. A., M. Kim, P. Philip, A. Allgardsson, P. Stenberg *et al.*, 2014 Non-coding roX RNAs prevent the binding of the MSL-complex to heterochromatic regions. PLoS Genet. 10: e1004865.

Fitch, D. H., L. D. Strausbaugh, and V. Barrett, 1990 On the origins of tandemly repeated genes: does histone gene copy number in Drosophila reflect chromosomal location? Chromosoma 99: 118–124.

Friedlander, T., R. Prizak, C. C. Guet, N. H. Barton, and G. Tkačik, 2016 Intrinsic limits to gene regulation by global crosstalk. Nat. Commun. 7: 12307.

Gelbart, M. E., E. Larschan, S. Peng, P. J. Park, and M. I. Kuroda, 2009 Drosophila MSL complex globally acetylates H4K16 on the male X chromosome for dosage compensation. Nat. Struct. Mol. Biol. 16: 825–832.

Günesdogan, U., H. Jäckle, and A. Herzig, 2014 Histone supply regulates S phase timing and cell cycle progression. Elife 3: e02443.

Hallacli, E., M. Lipp, P. Georgiev, C. Spielman, S. Cusack *et al.*, 2012 Msl1-mediated dimerization of the dosage compensation complex is essential for male X-chromosome regulation in Drosophila. Mol. Cell 48: 587–600.

Horard, B., and B. Loppin, 2015 Histone storage and deposition in the early Drosophila embryo. Chromosoma 124: 163–175.

Hur, W., J. P. Kemp Jr, M. Tarzia, V. E. Deneke, W. F. Marzluff *et al.*, 2020 CDK-Regulated Phase Separation Seeded by Histone Genes Ensures Precise Growth and Function of Histone Locus Bodies. Dev. Cell 54: 379–394.e6.

Kaya-Okur, H. S., S. J. Wu, C. A. Codomo, E. S. Pledger, T. D. Bryson *et al.*, 2019 CUT&Tag for efficient epigenomic profiling of small samples and single cells. Nat. Commun. 10: 1930.

Kaye, E. G., M. Booker, J. V. Kurland, A. E. Conicella, N. L. Fawzi *et al.*, 2018 Differential Occupancy of Two GA-Binding Proteins Promotes Targeting of the Drosophila Dosage Compensation Complex to the Male X Chromosome. Cell Rep. 22: 3227–3239.

Kaye, E. G., A. Kurbidaeva, D. Wolle, T. Aoki, P. Schedl *et al.*, 2017 Drosophila Dosage Compensation Loci Associate with a Boundary-Forming Insulator Complex. Mol. Cell. Biol. 37.:

Kelley, R. L., J. Wang, L. Bell, and M. I. Kuroda, 1997 Sex lethal controls dosage compensation in Drosophila by a non-splicing mechanism. Nature 387: 195–199.

Kemp, J. P., Jr, X.-C. Yang, Z. Dominski, W. F. Marzluff, and R. J. Duronio, 2021 Superresolution light microscopy of the Drosophila histone locus body reveals a core-shell organization associated with expression of replication-dependent histone genes. Mol. Biol. Cell 32: 942–955.

Kim, B. Y., J. R. Wang, D. E. Miller, O. Barmina, E. Delaney *et al.*, 2022 Correction: Highly contiguous assemblies of 101 drosophilid genomes. Elife 11.:

Koreski, K. P., L. E. Rieder, A. Chubal, L. M. McLain, W. F. Marzluff *et al.*, 2020 Drosophila Histone Locus Body assembly and function involves multiple interactions. Molecular Biology of the Cell 2020.03.16.994483.

Kremer, H., and W. Hennig, 1990 Isolation and characterization of a Drosophila hydei histone DNA repeat unit. Nucleic Acids Res. 18: 1573–1580.

Kuzu, G., E. G. Kaye, J. Chery, T. Siggers, L. Yang *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 Genet. 12: e1006120.

Kyrchanova, O., M. Sabirov, V. Mogila, A. Kurbidaeva, N. Postika *et al.*, 2019 Complete reconstitution of bypass and blocking functions in a minimal artificial Fab-7 insulator from Drosophila bithorax complex. Proc. Natl. Acad. Sci. U. S. A. 116: 13462–13467.

Lam, K. C., F. Mühlpfordt, J. M. Vaquerizas, S. J. Raja, H. Holz *et al.*, 2012 The NSL complex regulates housekeeping genes in Drosophila. PLoS Genet. 8: e1002736.

Larschan, E., J. Urban, and G. Kuzu, 2017 Chromatin accessibility of the dosage compensated Drosophila male X-chromosome is established by a context-specific role for the CLAMP zinc finger protein. The FASEB Journal 31: 593–510.

Lee, M.-C., L.-L. Toh, L.-P. Yaw, and Y. Luo, 2010 Drosophila octamer elements and Pdm-1 dictate the coordinated transcription of core histone genes. J. Biol. Chem. 285: 9041–9053.

Lifton, R. P., M. L. Goldberg, R. W. Karp, and D. S. Hogness, 1978 The organization of the histone genes in Drosophila melanogaster: functional and evolutionary implications. Cold Spring Harb. Symp. Quant. Biol. 42 Pt 2: 1047–1051.

Lindehell, H., M. Kim, and J. Larsson, 2015 Proximity ligation assays of protein and RNA interactions in the male-specific lethal complex on Drosophila melanogaster polytene chromosomes. Chromosoma 124: 385–395.

Liu, J.-L., C. Murphy, M. Buszczak, S. Clatterbuck, R. Goodman *et al.*, 2006 The Drosophila melanogaster Cajal body. J. Cell Biol. 172: 875–884.

Lucchesi, J. C., and M. I. Kuroda, 2015 Dosage compensation in Drosophila. Cold Spring Harb. Perspect. Biol. 7.:

Marín, I., A. Franke, G. J. Bashaw, and B. S. Baker, 1996 The dosage compensation system of Drosophila is co-opted by newly evolved X chromosomes. Nature 383: 160–163.

Mariño-Ramírez, L., I. K. Jordan, and D. Landsman, 2006 Multiple independent evolutionary solutions to core histone gene regulation. Genome Biol. 7: R122.

Marzluff, W. F., P. Gongidi, K. R. Woods, J. Jin, and L. J. Maltais, 2002 The human and mouse replication-dependent histone genes. Genomics 80: 487–498.

Marzluff, W. F., S. Sakallah, and H. Kelkar, 2006 The sea urchin histone gene complement. Dev. Biol. 300: 308–320.

Marzluff, W. F., E. J. Wagner, and R. J. Duronio, 2008 Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail. Nat. Rev. Genet. 9: 843–854.

McKay, D. J., S. Klusza, T. J. R. Penke, M. P. Meers, K. P. Curry *et al.*, 2015 Interrogating the function of metazoan histones using engineered gene clusters. Dev. Cell 32: 373–386.

Nakashima, Y., A. Higashiyama, A. Ushimaru, N. Nagoda, and Y. Matsuo, 2016 Evolution of GC content in the histone gene repeating units from Drosophila lutescens, D. takahashii and D. pseudoobscura. Genes Genet. Syst. 91: 27–36.

Nizami, Z. F., S. Deryusheva, and J. G. Gall, 2010 Cajal bodies and histone locus bodies in

Drosophila and Xenopus. Cold Spring Harb. Symp. Quant. Biol. 75: 313–320.

Norris, D., and M. A. Osley, 1987 The two gene pairs encoding H2A and H2B play different roles in the Saccharomyces cerevisiae life cycle. Mol. Cell. Biol. 7: 3473–3481.

Powell, J. R., 1997 *Progress and Prospects in Evolutionary Biology: The Drosophila Model*. Oxford University Press.

Rastelli, L., R. Richman, and M. I. Kuroda, 1995 The dosage compensation regulators MLE, MSL-1 and MSL-2 are interdependent since early embryogenesis in Drosophila. Mech. Dev. 53: 223–233.

Rebeiz, M., N. Jikomes, V. A. Kassner, and S. B. Carroll, 2011 Evolutionary origin of a novel gene expression pattern through co-option of the latent activities of existing regulatory sequences. Proc. Natl. Acad. Sci. U. S. A. 108: 10036–10043.

Reenan, R. A., C. J. Hanrahan, and B. Ganetzky, 2000 The mle(napts) RNA helicase mutation in drosophila results in a splicing catastrophe of the para Na+ channel transcript in a region of RNA editing. Neuron 25: 139–149.

Rieder, L. E., W. T. Jordan, and E. N. Larschan, 2019 Targeting of the dosage-compensated male X-chromosome during early Drosophila development. Cell Rep. 29(13):4268-4275.

Rieder, L. E., K. P. Koreski, K. A. Boltz, G. Kuzu, J. A. Urban *et al.*, 2017 Histone locus regulation by the Drosophila dosage compensation adaptor protein CLAMP. Genes Dev. 31: 1494–1508.

Robinson, J. T., H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander *et al.*, 2011 Integrative genomics viewer. Nat. Biotechnol. 29: 24–26.

Russo, C. A., N. Takezaki, and M. Nei, 1995 Molecular phylogeny and divergence times of drosophilid species. Mol. Biol. Evol. 12: 391–404.

Salzler, H. R., D. C. Tatomer, P. Y. Malek, S. L. McDaniel, A. N. Orlando *et al.*, 2013 A sequence in the Drosophila H3-H4 Promoter triggers histone locus body assembly and biosynthesis of replication-coupled histone mRNAs. Dev. Cell 24: 623–634.

Sankar, A., F. Mohammad, A. K. Sundaramurthy, H. Wang, M. Lerdrup *et al.*, 2022 Histone editing elucidates the functional roles of H3K27 methylation and acetylation in mammals. Nat. Genet. 54: 754–760.

Schienman, J. E., E. R. Lozovskaya, and L. D. Strausbaugh, 1998 Drosophila virilis has atypical kinds and arrangements of histone repeats. Chromosoma 107: 529–539.

Soruco, M. M. L., J. Chery, E. P. Bishop, T. Siggers, M. Y. Tolstorukov *et al.*, 2013 The CLAMP protein links the MSL complex to the X chromosome during Drosophila dosage compensation. Genes & Development 27: 1551–1556.

Soruco, M. M. L., and E. Larschan, 2014 A new player in X identification: the CLAMP protein is a key factor in Drosophila dosage compensation. Chromosome Res. 22: 505–515.

Staller, M. V., 2022 Transcription factors perform a 2-step search of the nucleus. Genetics iyac111.

Terzo, E. A., S. M. Lyons, J. S. Poulton, B. R. S. Temple, W. F. Marzluff *et al.*, 2015 Distinct self-interaction domains promote Multi Sex Combs accumulation in and formation of the Drosophila histone locus body. Mol. Biol. Cell 26: 1559–1574.

Tikhonova, E., A. Fedotova, A. Bonchuk, V. Mogila, E. N. Larschan *et al.*, 2019 The simultaneous interaction of MSL2 with CLAMP and DNA provides redundancy in the initiation of dosage compensation in Drosophila males. Development 146.:

Tikhonova, E., S. Mariasina, O. Arkova, O. Maksimenko, P. Georgiev *et al.*, 2022a Dimerization Activity of a Disordered N-Terminal Domain from Drosophila CLAMP Protein. Int. J. Mol.

Sci. 23.:

Tikhonova, E., S. Mariasina, S. Efimov, V. Polshakov, O. Maksimenko *et al.*, 2022b Structural basis for interaction between CLAMP and MSL2 proteins involved in the specific recruitment of the dosage compensation complex in Drosophila. Nucleic Acids Res.

Trinkle-Mulcahy, L., and J. E. Sleeman, 2017 The Cajal body and the nucleolus: "In a relationship" or "It's complicated"? RNA Biol. 14: 739–751.

Urban, J. A., C. A. Doherty, W. T. Jordan, J. E. Bliss, J. Feng *et al.* Drosophila CLAMP is an essential protein with sex-specific roles in males and females.

Urban, J. A., J. M. Urban, G. Kuzu, and E. N. Larschan, 2017 The Drosophila CLAMP protein associates with diverse proteins on chromatin. PLoS One 12: e0189772.

Venken, K. J. T., Y. He, R. A. Hoskins, and H. J. Bellen, 2006 P[acman]: a BAC transgenic platform for targeted insertion of large DNA fragments in D. melanogaster. Science 314: 1747–1751.

Villa, R., P. K. A. Jagtap, A. W. Thomae, A. Campos Sparr, I. Forné *et al.*, 2021 Divergent evolution toward sex chromosome-specific gene regulation in Drosophila. Genes Dev. 35: 1055–1070.

Villa, R., T. Schauer, P. Smialowski, T. Straub, and P. B. Becker, 2016 PionX sites mark the X chromosome for dosage compensation. Nature 537: 244–248.

Wei, Y., J. Jin, and J. W. Harper, 2003 The cyclin E/Cdk2 substrate and Cajal body component p220(NPAT) activates histone transcription through a novel LisH-like domain. Mol. Cell. Biol. 23: 3669–3680.

White, A. E., B. D. Burch, X.-C. Yang, P. Y. Gasdaska, Z. Dominski *et al.*, 2011 Drosophila histone locus bodies form by hierarchical recruitment of components. J. Cell Biol. 193:

677–694.

White, A. E., M. E. Leslie, B. R. Calvi, W. F. Marzluff, and R. J. Duronio, 2007 Developmental and cell cycle regulation of the Drosophila histone locus body. Mol. Biol. Cell 18: 2491–2502.

Wolle, D., F. Cleard, T. Aoki, G. Deshpande, P. Schedl *et al.*, 2015 Functional Requirements for Fab-7 Boundary Activity in the Bithorax Complex. Mol. Cell. Biol. 35: 3739–3752.

Yang, X.-C., I. Sabath, L. Kunduru, A. J. van Wijnen, W. F. Marzluff *et al.*, 2014 A conserved interaction that is essential for the biogenesis of histone locus bodies. J. Biol. Chem. 289: 33767–33782.

Zheng, L., R. G. Roeder, and Y. Luo, 2003 S phase activation of the histone H2B promoter by OCA-S, a coactivator complex that contains GAPDH as a key component. Cell 114: 255–266.

Chapter 5

A bioinformatics screen reveals Hox and chromatin remodeling factors at the *Drosophila* histone locus

Reproduced with permission by:

Lauren J. Hodkinson[1]*, Connor Smith[2]*, H. Skye Comstra[2], Bukola A. Ajani[2], Eric H. Albanese[2], Kawsar Arsalan[2], Alvaro Perez Daisson[2], Katherine B. Forrest[2], Elijah H. Fox[2], Matthew R. Guerette[2], Samia Khan[2], Madeleine P. Koenig[2], Shivani Lam[2], Ava S. Lewandowski[2], Lauren J. Mahoney[2], Nasserallah Manai[2], JonCarlo Miglay[2], Blake A. Miller[2], Olivia Milloway[2], Nhi Ngo[2], Vu D. Ngo[2], Nicole F. Oey[2], Tanya A. Punjani[2], HaoMin SiMa[2], Hollis Zeng[2], **Casey A. Schmidt[2]*, Leila E. Rieder[2]*** 2023 A bioinformatics screen reveals hox and chromatin remodeling factors at the *Drosophila* histone locus. BMC Genomic Data 24: 54. https://doi.org/10.1186/s12863-023-01147-0

*Authors contributed equally

5.1 Abstract

Cells orchestrate histone biogenesis with strict temporal and quantitative control. To efficiently regulate histone biogenesis, the repetitive *Drosophila melanogaster* replication-dependent histone genes are arrayed and clustered at a single locus. Regulatory factors concentrate in a nuclear body known as the histone locus body (HLB), which forms around the locus. Historically, HLB factors are largely discovered by chance, and few are known to interact directly with DNA. It is therefore unclear how the histone genes are specifically targeted for unique and coordinated regulation. To expand the list of known HLB factors, we performed a candidate-based screen by mapping 30 publicly available ChIP datasets of 27 unique factors to the *Drosophila* histone gene array. We identified novel transcription factor candidates, including the *Drosophila* Hox proteins Ultrabithorax (Ubx), Abdominal-A (Abd-A) and Abdominal-B (Abd-B), suggesting a new pathway for these factors in influencing body plan morphogenesis. Additionally, we identified six other factors that target the histone gene array: JIL-1, hormone-like receptor 78 (Hr78), the long isoform of female sterile homeotic (1) (fs(1)h) as well as the general transcription factors TBP associated factor 1 (TAF-1), Transcription Factor IIB (TFIIB), and Transcription Factor IIF (TFIIF). Our foundational screen provides several candidates for future studies into factors that may influence histone biogenesis. Further, our study emphasizes the powerful reservoir of publicly available datasets, which can be mined as a primary screening technique.

5.2 Introduction

Cells rely on strict temporal and quantitative orchestration of gene expression. One way the nucleus accomplishes coordinated gene regulation is through the establishment of nuclear bodies (NBs), membraneless concentrations of proteins and RNAs. The NB micro-environment facilitates processes such as efficient gene expression through transcription and RNA-processing (Matera *et al.* 2009; Tatomer *et al.* 2016; Arias Escayola and Neugebauer 2018).

The histone locus body (HLB) is a conserved NB that regulates histone gene expression and forms at the loci of the replication-dependent histone genes (Duronio and Marzluff 2017) in many different organisms, including humans and *Drosophila*. The HLB is characterized by a set of factors that collectively regulate the uniquely organized histone genes. The *Drosophila melanogaster* histone locus is a cluster of ~100 tandemly repeated arrays, in which each 5 Kb array includes the 5 canonical histone genes along with their respective promoters and regulatory elements (McKay *et al.* 2015; Duronio and Marzluff 2017; Bongartz and Schloissnig 2018). Each array contains two TATA-box containing promoters, one for *H3* and *H4* and one *H2A* and *H2B* (**Figure 5.1A)**. Additionally, the *H1* gene has its own unique promoter that lacks a TATA-box. The promoters contains some known transcription factor motifs (Crayton *et al.* 2004; Isogai *et al.* 2007; Rieder *et al.* 2017), but overall little is known about how the locus is transcriptionally controlled. The clustered, repetitive organization of the locus allows for precise HLB formation at a single genomic location and highly coordinated histone biogenesis linked to S-phase of the cell cycle (Marzluff *et al.* 2002; White *et al.* 2011).

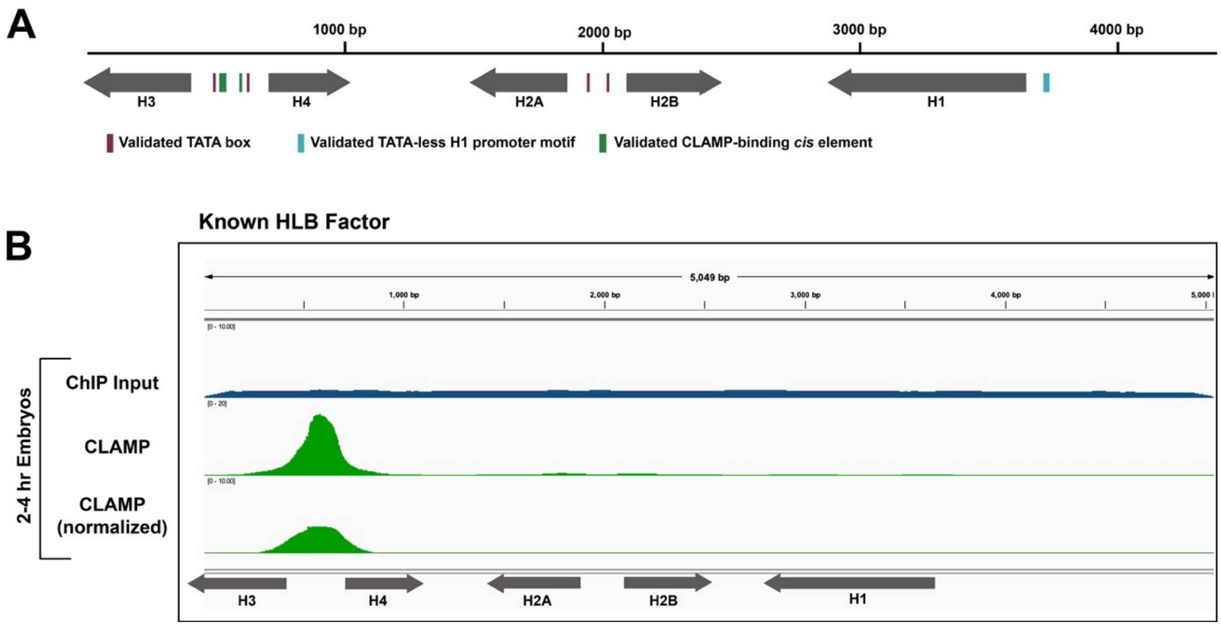


Figure 5.1: Known HLB factor CLAMP localizes to the GA-repeat *cis* element in the *H3/H4* promoter. (A) A diagram detailing the validated *cis* elements in the histone array including the TATA-box elements (teal boxes), the TATA-less motif (maroon box), and the CLAMP binding GA-repeat elements (green boxes). (B) We mapped the known ChIP-seq data for the known HLB factor CLAMP (green) from 2-4 hr embryos (Duan *et al.* 2021).The ChIP signal was normalized to its respective ChIP input signal (blue).

The *Drosophila* HLB is a well-characterized NB that includes several known components that play a role in both the cell cycle regulation of histone gene transcription and the unique processing of histone mRNA transcripts. Several proteins are involved in the initiation and regulation of histone gene transcription including Chromatin Linked Adaptor for MSL proteins (CLAMP; (Rieder *et al.* 2017)), Multi Sex combs (Mxc (White *et al.* 2011; Yang *et al.* 2014); the *Drosophila* ortholog of human Nuclear Protein mapped to the Ataxia-Telangiectasia locus (NPAT; (Terzo *et al.* 2015)), FLICE-associated huge protein (FLASH (Tatomer *et al.* 2016)) and Muscle wasted (Mute (Bulchand *et al.* 2010)). Histone mRNA processing is distinct from that of other mRNAs because histone pre-mRNAs lack polyA tails and introns (Duronio and Marzluff 2017). Several known factors are involved in histone mRNA processing and target the histone gene locus including, the U7snRNP (Godfrey *et al.* 2009), Stem Loop Binding Protein (SLBP(Jaeger *et al.* 2006)), and Lsm11 (Duronio and Marzluff 2017).

Other than CLAMP, the above-mentioned factors target the histone locus but do not interact directly with DNA sequence. Since CLAMP is found at locations genome wide, it is currently unclear how non-DNA binding factors identify and target the histone locus. The presence of histone mRNA is likely to play a role (Shevtsov and Dundr 2011) as are the presence of *cis* elements within the histone gene array (Salzler *et al.* 2013; Rieder *et al.* 2017). One critical interaction involves CLAMP recognizing GA-repeat sequences within the *H3/H4* promoter (Rieder *et al.* 2017) (**Figure 5.1)**. Although the presence of CLAMP is critical for the localization of HLB-specific factors such as Mxc (Rieder *et al.* 2017), the interaction between CLAMP and GA-repeat is not strictly necessary for HLB formation (Koreski *et al.* 2020) and CLAMP is not sufficient for HLB formation (Rieder *et al.* 2017). Therefore, it is likely that other DNA-interacting proteins participate in defining the histone locus. We still lack a comprehensive list of factors associated with histone biogenesis and therefore our model of the mechanisms of histone gene regulation remains incomplete.

Historically, novel HLB factors are often discovered by chance through immunofluorescence such as CLAMP (Rieder *et al.* 2017), Myc (Daneshvar *et al.* 2011), Mute (Bulchand *et al.* 2010), and Abnormal oocyte (Berloco *et al.* 2001). To discover novel DNA-binding proteins that target the histone locus, we first screened the literature for likely candidates and then funneled these into a secondary bioinformatics screen. We leveraged publicly available *Drosophila* ChIP-seq datasets and knowledge of histone gene regulation to curate and analyze a list of candidate DNA-binding factors. We used a bioinformatics pipeline on Galaxy (Afgan *et al.* 2016; The Galaxy Community 2022) to map candidate ChIP-seq data to a single copy of the histone gene array. The ~100 histone gene arrays are nearly identical in sequence (Bongartz and Schloissnig 2018) and we can collapse -omics data from the entire locus onto a single array (McKay *et al.* 2015; Rieder *et al.* 2017; Koreski *et al.* 2020). Supervised undergraduate students conducted much of the initial screen as part of a course-based undergraduate research experience (CURE; (Schmidt *et al.* 2022), demonstrating the simplicity and versatility of the pipeline design. Using our qualitative analysis criteria (**Supplemental Figure 5.1 in Appendix A)**, we discovered several DNA-interacting proteins that pass our initial bioinformatics screen. Our novel candidates that target the histone gene array include development transcription factors such as Hox factors, which may provide a mechanistic link between segment identity and cell division.

Future wet lab studies are required to confirm the presence of these candidates at the histone locus, determine any tissue and temporal specificity, and describe the precise roles of candidates in HLB formation and histone biogenesis. As a whole, our screen establishes mining of existing -omics data as a tool to identify new candidate HLB factors. Although we are limited by the factors, tissues, treatments, and timepoints interrogated by the dataset generators, our pipeline is an inexpensive and rapid tool to screen candidate factors for future wet-lab studies.

5.3 Methods

5.3.1 GEO Datasets

All datasets were downloaded from the NCBI SRA Run Selector through the Gene Expression Omnibus (GEO). See **Table 1** for Accession numbers and references.

Table 5.1: DNA-binding factor candidate datasets

Candidate	GEO Accession #	SRA Run Selector #	Paper citation
Abd-A **Abdominal-A**	GSE69796	**anti-GFP ChIP DNA from Kc167 cells expressing AbdA-GFP** **1**- SRR2060648 **2** -SRR2060649 **Input** **1** - SRR2060652 **2** - SRR2060653	(Beh *et al.* 2016)
Abd-B **Abdominal-B**	GSE69796	**anti-GFP ChIP DNA from Kc167 cells expressing AbdB-GFP** **1**- SRR2060650 **2** -SRR2060651 **Input** **1** - SRR2060652 **2** - SRR2060653	(Beh *et al.* 2016)
ANTP **Antennapedia**	GSE125604	**anti-GFP (Invitrogen) from ANTP-GFP genotype** **1** - SRR8483063 **Input** **1** - SRR8483064	(Kribelbauer *et al.* 2020)
CP190 **Centrosomal protein 190kD**	GSE118699	**CP190 rabbit (Pai et al 2004)** **1** - SRR7706256 **2** - SRR7706258 **Input** **1** - SRR7706251 **2** - SRR7706252	(Bag *et al.* 2019)
CTCF	GSE175402	CTCF 1 - SRR14631231 2 - SRR14631232 Input 1 - SRR14631233 2 - SRR14631234	(Kyrchanova *et al.* 2021)
Exd **Extradenticle**	GSE125604	**anti-V5 (Invitrogen) on exd-V5 transgene genotype** **1** - SRR8483055 **Input** **1** - SRR8483056	(Kribelbauer *et al.* 2020)
Fs(1)h **Female sterile (1) homeotic**	GSE42086	**Female late embryo-derived cell line, ChIP of Fs(1)h long isoform** **1**- SRR611533 **Female late embryo-derived cell line, ChIP of both isoforms of Fs(1)h** **1** - SRR611535 **Input** **1** - SRR611537	(Kellner *et al.* 2013)
Gcn5	GSE83408	**Gcn5 rabbit polyclonal antibody (5 ug/IP)** **1** - SRR3671294 **2** - SRR3671295 **3** - SRR3671298 **Input** **1** - SRR3671296 **2** - SRR3671297 **3** - SRR3671299	(Ali *et al.* 2017)
Hr78 **Hormone-receptor-like 78**	GSE50370	**Hr78-GFP_8-16_embryonic_ChIP-seq_ChIP** **1** - SRR1198798 **2** - SRR1198799 **Input** **1** - SRR1198796 **2** - SRR1198797	(THE MODENCODE CONSORTIUM *et al.* 2010)
Hnf4 **Hepatocyte nuclear factor 4**	GSE73675	**rat anti-dHNF4 3600** **1** - SRR2548371 **2** - SRR2548372 **3** - SRR2548373 **4** - SRR2548374 **Inputs** **1** - SRR2548367 **2** - SRR2548368 **3** - SRR2548369 **4** - SRR2548370	(Barry and Thummel 2016)
HTH **Homothorax**	GSE125604	**anti-Hth (gp52, N-terminal)** **1** - SRR8483065 **Input** **1** - SRR8483066	(Kribelbauer *et al.* 2020)
JIL-1	GSE54438	**JIL-1 monoclonal antibody 5C9** **1** - SRR1145605 **2** - SRR1145606 **Input** **1** - SRR1145612 **2** - SRR1145613	(Cai *et al.* 2014)
M1BP **Motif 1 Binding Protein**	GSE97841	**M1BP_Antibody** **1** - SRR10759878 **Input** **1** - SRR10759877	(Baumann and Gilmour 2017)
MSL-1 **Male-specific Lethal 1**	GSE37864	**polyclonal rabbit MSL1, crude serum** **1** - SRR495366 **2** - SRR495367 **Input** **1** - SRR495378 **2** - SRR495380	(Straub *et al.* 2013a)
Ndf/CG4747 **Nucleosome-destabilizing factor**	GSE42025	**PAP antibody (Sigma P1291)** **1** - SRR611192 **2** - SRR611194 **3** - SRR611196 **4** - SRR611198 **Input** **1** - SRR611193 **2** - SRR611195 **3** - SRR611197 **4** - SRR611199	(Wang *et al.* 2013)
Nej (S2 cells) **Nejire**	GSE72666	**anti-CBP, custom-made antibodies** **1**- SRR2232434 **Input** **1** - SRR2232432	(Doiguchi *et al.* 2016)
Nej (Embryos) **Nejire**	GSE68983	**Nej** **1** - SRR4044401	(Koenecke *et al.* 2016)

		Input 1 - SRR2031906	
Opa **Odd Paired**	GSE140722	**In-house anti-Opa antibody** 1 - SRR10502454 2 - SRR10502455 3 - SRR10502458 4 - SRR10502459 **Input** 1 - SRR10502456 2 - SRR10502457 3 - SRR10502460 4 - SRR10502461	(Koromila *et al.* 2020)
Pan **Pangolin**	GSE50340	**Pan** 1 - SRR1198824 2 - SRR1198825 **Input** 1 - SRR1198822 2 - SRR1198823	(THE MODENCODE CONSORTIUM *et al.* 2010)
Pnt **Pointed**	GSE114092	**Pnt** 1 - SRR7126165 **Input** 1 - SRR7126164	(Webber *et al.* 2018)
Psc **Posterior sex combs**	GSE38166	**Psc Mitotic S2** 1 - SRR 500149 2 - SRR 500150 **Psc Control S2** 1 - SRR500151 2 - SRR500152 **Psc Mitotic S2 Input** 1 - SRR 500153 2 - SRR 500154 **Psc Control S2 Input** 1 - SRR 500155 2 - SRR 500156	(Follmer *et al.* 2012)
Scm **Sex comb on midleg**	GSE66183	**BioTAP-N-Scm** 1 - SRR1813233 2 - SRR1813243 3 - SRR1813245 **Input** 1 - SRR1813234 2 - SRR1813244 3 - SRR1813246	(Kang *et al.* 2015)
su(z)12 **suppressor of zeste 12**	GSE36039	**Su(z)12 ChIP** 1 - SRR363407 2 - SRR363408 **Input** 1 - SRR363409 2 - SRR363410	(Herz *et al.* 2012)
TAF1 **TBP-Associated Factor 1**	GSE97841	**TAF1 Antibody** 1 - SRR5452843 2 - SRR5452844 **Inputs** 1 - SRR5452847 2 - SRR5452848	(Baumann and Gilmour 2017)
TFIIB **Transcription Factor II B**	GSE120152	**anti-TFIIB rabbit polyclonal, custom** 1 - SRR7874066 2 - SRR7874067 **Inputs** 1 - SRR7874069 2 - SR7874070	(Ramalingam *et al.* 2021)
TFIIF **Transcription Factor II F**	GSE120152	**anti-TFIIF rabbit polyclonal, custom** 1 - SRR7874068 **Inputs** 1 - SRR7874069	(Ramalingam *et al.* 2021)
TRF2 **TBP protein-related factor 2**	GSE97841	**TRF2 Antibody** 1 - SRR5452845 2 - SRR5452846 **Inputs** 1 - SRR5452847 2 - SRR5452848	(Baumann and Gilmour 2017)
Ubx (Kc cells) **Ultrabithorax**	GSE69796	**anti-GFP ChIP DNA from Kc167 cells expressing Ubx-GFP** 1 - SRR2060646 2 - SRR2060647 **Inputs:** 1 - SRR2060652 2 - SRR2060653	(Beh *et al.* 2016)
Ubx (embryos) **Ultrabithorax**	GSE64284	**Anti-V5 ChIP, Ubx-V5** 1 - SRR1721317 2 - SRR1721321 **Inputs** 1 - SRR1721316 2 - SRR1721320	(Shlyueva *et al.* 2016)
Ubx (larva) **Ultrabithorax**	GSE184454	**Anti-FLAG monoconal, 3xFLAG-Ubx** 1 - SRR15972582 2 - SRR15972584 **Inputs** 1- SRR15972583 2 - SRR15972585	(Feng *et al.* 2022)

5.3.2 Bioinformatic Analysis and Data Visualization

We directly imported individual FASTQ datasets into the web-based platform Galaxy (Afgan *et al.* 2016; The Galaxy Community 2022) through the NCBI SRA Run Selector by selecting the desired runs and utilizing the computing Galaxy download feature. We retrieved the FASTQ files from SRA using the "Faster Download and Extract Reads in FASTQ format from NCBI SRA" Galaxy command. Because the ~100 histone gene arrays are extremely similar in sequence (Bongartz and Schloissnig 2018), we do not utilize the dm6 or dm3 genomes and instead can collapse ChIP-seq data onto a single histone array (McKay *et al.* 2015; Bongartz and Schloissnig 2018; Koreski *et al.* 2020). We used a custom "genome" that includes a single *Drosophila melanogaster* histone array similar to that in Mckay *et al.* 2015, which we directly uploaded to Galaxy using the "upload data" feature, and normalized using the Galaxy command "NormalizeFasta" specifying an 80 bp line length for the output.fasta file. We aligned ChIP reads to the normalized histone gene array using Bowtie2 (Langmead and Salzberg 2012) to create .bam files using the user built-in index and "very sensitive end-to-end" parameter settings. We converted the .bam files to .bigwig files using the "bamCoverage" Galaxy command in which we set the bin size to 1 bp and set the effective genome size to user specified: 5000 bp (approximate size of l histone array). We also mapped relevant input or IgG datasets. If an input dataset was available, we normalized ChIP datasets to input using the "bamCompare" Galaxy command in which we set the bin size to 1 bp. We visualized the bigwig files using the Integrative Genome Viewer (IGV) (Robinson *et al.* 2011).

5.3.3 Criteria for Positive Candidates vs. Negative Candidate

Because we focused our analysis on a single 5 Kb sequence and condensed data from ~100 identical histone arrays onto a single array, we were unable to use quantitative peak calling programs. We instead utilized the following qualitative criteria to determine positive and negative candidates (**Supplemental Figure 5.1 in Appendix A)**. We only considered the candidate as positive if a peak emerged in the ChIP data that was not present in the input. We considered the following false positives: 1) obvious overrepresentation of gene bodies (e.g. Su(z)12, **Supplemental Figure 5.2 in Appendix A**), 2) underrepresentation of intergenic regions

(CP190 input, **Figure 5.3C**) and 3) if the input coverage and ChIP coverage peaks looked identical, candidate was also considered a negative hit (e.g MSL1, **Figure 5.3B**). Datasets with the above-mentioned characteristics cause peaks to emerge in the normalized data that do not represent the binding of the factor but rather a bias in the amplification of the ChIP library or alignment. We also checked spot length (read length) and considered peaks over the GA-repeat *cis* element in the *H3/H4* promoter found in datasets with read lengths ≤50 bp false positive peaks (e.g Psc, **Supplemental Figure 5.2 in Appendix A**).

5.4 Results

5.4.1 Validating the bioinformatics pipeline by mapping TATA-associated factors to the histone gene array

We first sought to validate our bioinformatics pipeline through analysis of known histone locus proteins and associated factors. Isogai *et al.* (2007) used immunofluorescence and cell culture ChIP-qPCR assays to demonstrate that the TATA binding protein (TBP)/TFIID complex selectively binds to the *H3/H4* promoter and the *H2A/H2B* promoter, but TBP-related factor 2 (TRF2) targets the promoter of the TATA-less *H1* promoter. We identified a publicly available TRF2 ChIP-exo dataset from Baumann *et al.* (2017) for TRF2 and used our pipeline to map the data to the histone gene array. ChIP-exo is similar to ChIP-seq but identifies a more complete set of binding locations for a factor with higher resolution than standard ChIP-seq (Rhee and Pugh 2012). We validated that TRF2 localizes to the H1 promoter (**Figure 5.2A**). Because we were unable to normalize to an input dataset, we compared the TRF2 alignment to an IgG control. The localization of TRF2 to the TATA-less *H1* promoter is consistent with Isogai *et al*. (2007) and is consistent with where a TBP-related factor (TRF) would be expected to bind as they are known to target TATA-less promoters (Wang *et al.* 2013). Baumann *et al.* 2017 demonstrated that Motif 1 binding protein (M1BP) interacts with TRF2 but that this interaction is mostly restricted to the ribosomal protein genes (Baumann and Gilmour 2017). We mapped ChIP-exo data for M1BP and observed that it did not localize to the *H1* promoter under our qualitative criteria (**Supplemental Figure 5.1)** as we saw with TRF2 nor to any other part of the histone array (**Figure 5.2A**), further validating our pipeline.

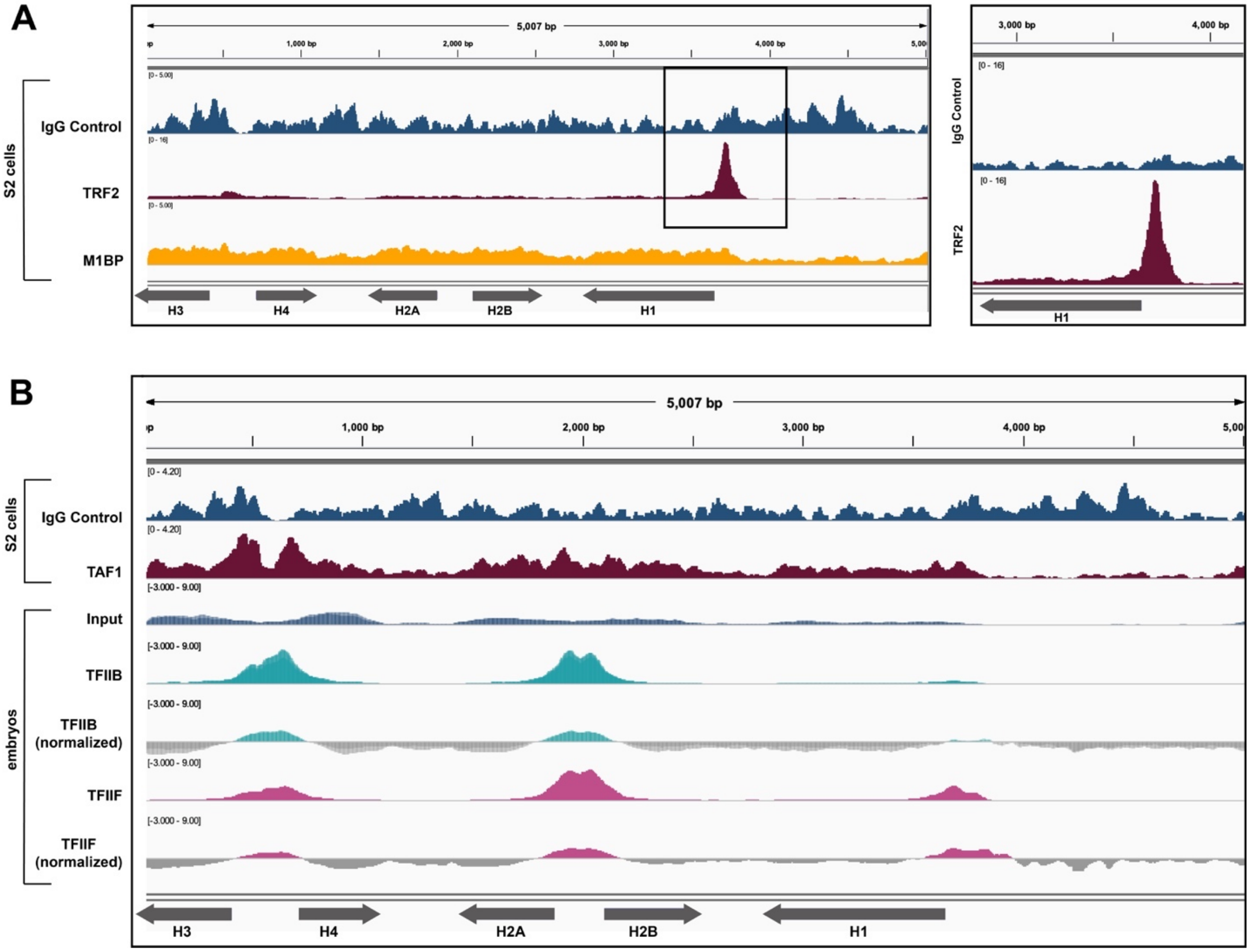


Figure 5.2: Expected general transcription factors localize to the histone array. (A) We mapped ChIP-exo data for TRF2 (maroon,*(Baumann and Gilmour 2017)*) from S2 cells to the histone gene array which recapitulates results from Isogia et al. 2007 showing localization

specifically to the *H1* promoter validating our bioinformatics pipeline. We also mapped ChIP-exo data for M1BP (yellow, (Baumann and Gilmour 2017)) which did not localize to the histone gene arr*ay,* further validating our pipeline. *We* compared ChIP-exo data to an IgG control (blue, *(Baumann and Gilmour 2017)* did not provide input sample). (B) We aligned ChIP-exo data for TAF-1 (maroon, (Baumann and Gilmour 2017)) from S2 cells to the histone gene array and compared to a corresponding IgG control. We aligned ChIP-seq datasets for TFIIB (teal, two replicates overlayed, (Ramalingam *et al.* 2021)) and TFIIF (pink, one replicate, (Ramalingam *et al.* 2021)) from OregonR mixed population embryos to the histone gene array and normalized to the provided input (blue). TFIIB shows localization to the *H3/H4* promoter and the *H2A/H2B* promoter and TFIIF shows localization to both core promoters and the *H1* promoter confirming that our bioinformatics pipeline can be used to identify novel factors that localize to the histone gene array.

Novel general transcription factors that target the histone locus

To expand the list of general transcription factors that target the histone locus, we mapped an additional ChIP-exo dataset from Baumann *et al.* 2017 for TAF1 (TBP associated factor 1). TAF1 is a member of the Transcription Factor IID (TFIID) complex which Isogai *et al.* (2007) also suggested localized to the same regions of the histone gene array as TBP. When we mapped the TAF1 ChIP-exo data we observed that TAF1 targets the TATA-box regions of the *H3*/*H4* promoter and, less robustly, to the TATA-box regions of the *H2A*/*H2B* promoter (**Figure 5.2B,** elements annotated in **Figure 5.1A**). Again, we compared this alignment to an IgG control because we were unable to normalize to an input, but because TAF1 associates with TBP which binds to AT-rich (TATA box) regions (Baumann and Gilmour 2017), the localization of TAF1 to the TATA-box regions of the core histone genes is expected.

To test the ability of our pipeline to identify novel factors that localize to the histone gene array, we investigated the relationships of additional general transcription factors to the histone array. We identified ChIP-seq datasets for both TFIIB and TFIIF. Both TFIIB and TFIIF are associated with TBP (Ramalingam *et al.* 2021) and therefore we would expect them to localize to the *H3/H4* and *H2A/H2B* promoters, similar to TBP (Isogai *et al.* 2007). We observed both TFIIB and TFIIF localized to the *H3/H4* and *H2A/H2B* promoters while, surprisingly, TFIIF localized to the *H1* promoter (**Figure 5.2B**).

5.4.2 Candidate DNA-binding factors that did not pass the bioinformatics screen

After verifying our bioinformatics pipeline, we curated a list of candidate DNA-binding factors (**Table 1**, **Supplemental Table 5.1 in Appendix A**) that we hypothesized would target the histone gene array. To create this candidate list, we prioritized factors that meet at least one of the following criteria: 1) DNA-binding factors with a relationship to a validated HLB factor; 2) DNA-binding factors involved in dosage compensation because CLAMP, a non-sex specific dosage compensation factor, targets the histone locus (Rieder *et al.* 2017; Koreski *et al.* 2020); 3) chromatin remodeling or histone-interacting factors since the epigenetic landscape of the histone locus is largely undefined; 4) early developmental transcription factors since histone gene regulation is critical during early development and synchronized cell division (Chari *et al.* 2019). We also utilized the online platform STRING (Szklarczyk *et al.* 2019) that provides the known and inferred interactomes of a given protein to identify candidates that met the above criteria. Out of the 27 candidates, we rejected 19 as likely not targeting the histone gene array based on our qualitative analysis of the datasets we investigated (**Supplemental Figure 5.1 in Appendix A)**.

HLB factor-associated candidates:

We investigated the DNA-binding factor Sex comb on midleg (Scm), because of its suspected interaction with the known HLB factor Multi-sex combs (Mxc; (White *et al.* 2011; Yang *et al.* 2014). Based on STRING, Scm is predicted to interact with Mxc, as determined by a genetic interference assay in which a double Mxc/Scm mutant resulted in enhanced mutant sex combs phenotypes (Docquier *et al.* 1996; Saget *et al.* 1998). Despite possible interaction with Mxc, neither Scm ChIP-seq data from S2 cells nor 12-24 hr embryos gave meaningful signal over the histone gene array (**Figure 5.3A**). This result was surprising because the human ortholog of Mxc associates exclusively with the histone promoters (Kaya-Okur *et al.* 2019) and Mxc is only found at the histone locus (Terzo *et al.* 2015).

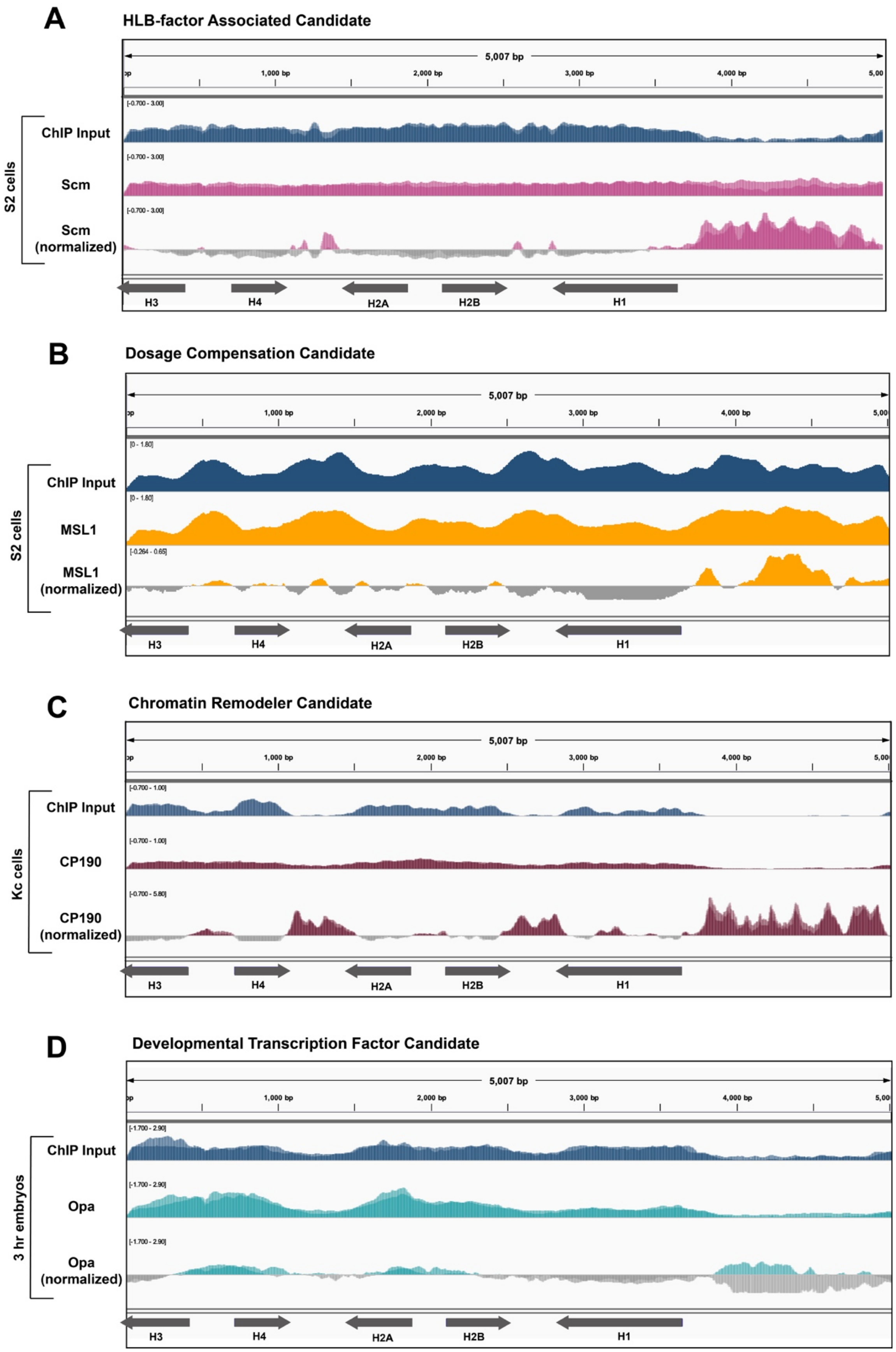

A
HLB-factor Associated Candidate
5,007 bp
bp
1,000 bp
2,000 bp
3,000 bp
4,000 bp
5,00
[-0.700 - 3.00]
S2 cells
ChIP Input
Scm
Scm (normalized)
H3
H4
H2A
H2B
H1
B
Dosage Compensation Candidate
5,007 bp
[0 - 1.80]
[-0.264 - 0.65]
S2 cells
ChIP Input
MSL1
MSL1 (normalized)
H3
H4
H2A
H2B
H1
C
Chromatin Remodeler Candidate
5,007 bp
[-0.700 - 1.00]
[-0.700 - 5.80]
Kc cells
ChIP Input
CP190
CP190 (normalized)
H3
H4
H2A
H2B
H1
D
Developmental Transcription Factor Candidate
5,007 bp
[-1.700 - 2.90]
3 hr embryos
ChIP Input
Opa
Opa (normalized)
H3
H4
H2A
H2B
H1

Figure 5.3: DNA-binding factors from different categories that did not pass the bioinformatics screen. We aligned ChIP-seq datasets for (A) Scm (pink, two replicates overlayed, (Kang *et al.* 2015)) from S2 cells, (B) MSL1 (yellow, one replicate, (Straub *et al.* 2013b)) from S2 cells, (C) CP190 (maroon, two replicates overlayed, (Bag *et al.* 2019)) from Kc cells, and (D) Opa (teal, two replicates overlayed, (Koromila *et al.* 2020)) from 3 hr mixed population embryos to the histone array. We normalized *e*ach ChIP signal to its respective ChIP input signal (blue).

<u>Dosage compensation candidates</u>:

The HLB factor CLAMP targets the *H3/H4* promoter (**Figure 5.1B**) and regulates histone gene expression (Rieder *et al.* 2017), but also plays additional roles in *Drosophila* male dosage compensation: it binds to GA-rich elements along the male X-chromosome and recruits the Male Specific Lethal complex (MSLc). Further, MSL2, the male specific component of MSLc, also emerged from a cell-based HLB factor screen (White *et al.* 2011) and we recently discovered that MSL2 targets one histone gene locus in *Drosophila virilis* (Xie *et al.* 2022b). We therefore hypothesized that dosage compensation factors target the histone gene array along with CLAMP. We chose the following DNA-binding factors for our candidate screen because of their relationship to dosage compensation: MSL1, a protein that scaffolds MSLc (Larschan *et al.* 2006; Straub *et al.* 2013a), and nucleosome destabilizing factor (Ndf, CG4747), a putative H3K36me3-binding protein that is important for MSLc localization (Wang *et al.* 2013). When we mapped ChIP-seq datasets from these factors, we found that neither gave meaningful signal over the histone gene array (MSL1: **Figure 5.3B**, Ndf/CG4747: **Supplemental Figure 5.2 in Appendix A**). This is not surprising as we previously determined that MSL2 does not target the histone locus in *Drosophila melanogaster* by polytene chromosome immunofluorescence (Xie *et al.* 2022b).

Chromatin remodeling candidates:

One of the lesser-studied characteristics of the histone locus is the regional chromatin environment. The endogenous histone locus is located on chromosome 2L, proximal to pericentric heterochromatin. Despite this proximity, histone expression rapidly increases at the start of G1 in preparation for DNA synthesis during S phase, and quickly ceases upon G2 (Duronio and Marzluff 2017) indicating that chromatin remodeling is likely critical in precisely controlling histone gene expression. We therefore hypothesized that chromatin remodeling factors target the histone locus. We chose the following candidates because of their association with chromatin or role in chromatin remodeling: centrosomal 190 kDa protein (CP190), an insulator protein that impacts enhancer protein interactions and stops the spread of heterochromatin (Bag *et al.* 2019); Gcn5, a lysine acetyltransferase critical for oogenesis and morphogenesis (Ali *et al.* 2017); CCCTC-binding factor (CTCF), a genome architectural protein (Kyrchanova *et al.* 2021); Posterior sex combs (Psc), a polycomb-group gene (Follmer *et al.* 2012); and Suppressor 12 of zeste (su(z)12), a subunit of polycomb repressive complex 2 (Herz *et al.* 2012) (CP190: **Figure 5.3C,** all others: **Supplemental Figure 5.2 in Appendix A**).

After identifying relevant ChIP-seq datasets (**Table 5.1**), we used our analysis pipeline to map data to the histone gene array. We observed that none of the above chromatin remodeling candidates gave meaningful signal over the histone gene array (CP190: **Figure 5.3C**, all others: **Supplemental Figure 5.2 in Appendix A**). We were especially surprised that CP190 did not target the histone array. CP190 binds promoter regions, aids enhancer-promoter interactions, and halts the spreading of heterochromatin. Because the histone locus is proximal to pericentric heterochromatin, we hypothesized the presence of CP190 could explain how centromeric heterochromatin does not expand into the histone locus. In addition, CP190 is a member of the Late Boundary Complex (LBC) (Wolle *et al.* 2015), which also contains the CLAMP protein (Kaye *et al.* 2018). We discovered that the LBC binds to the *H3/H4* promoter region *in vitro* (Xie *et al.* 2022b). We were therefore surprised that CP190 does not appear to target the histone gene array, based on the ChIP-seq datasets we analyzed. These data underscore the requirement for visualizing both ChIP and input datasets, rather than just the final normalized trace: although CP190 ChIP-seq data does not show enrichment over the histone gene array, bias in the input dataset lead to misleading peaks in the normalized data (**Figure 5.3C**, **Supplemental Figure 5.2 in Appendix A**).

Developmental transcription factor candidates:

Zygotic histone biogenesis is critical for the constantly dividing embryo; increased histone expression can lengthen the cell cycle whereas decreased histone levels can shorten the cell cycle (Amodeo *et al.* 2015; Chari *et al.* 2019). Histone biogenesis is tightly coupled to DNA replication, and excess histones are buffered so as not to interfere with zygotic chromatin (Li *et al.* 2012, 2014; Stephenson *et al.* 2021). We therefore hypothesized that early embryonic transcription factors target the histone locus. We chose the following DNA-binding factors based on their roles in the early embryo: Odd paired (Opa), a pair ruled gene that contributes to morphogenesis (Koromila *et al.* 2020); Motif 1 binding protein (M1BP), a transcription pausing factor that interacts with the Hox proteins (Baumann and Gilmour 2017; Bag *et al.* 2021); Hepatocyte nuclear factor 4 (Hnf4), a general developmental transcription factor (Barry and Thummel 2016), Pangolin (Pan), a component of the Wingless signaling pathway (Ravindranath and Cadigan 2014); and Pointed (Pnt), a factor the regulates cell proliferation and differentiation during development (Webber *et al.* 2018; Vivekanand 2018). When we mapped appropriate ChIP-seq datasets from these factors, none gave meaningful signal over the histone array (Opa: **Figure 5.3D**, M1BP: **Figure 5.2A,** all others: **Supplemental Figure 5.2 in Appendix A**).

5.4.3 Candidates that passed the bioinformatics screen

We found several factors that exhibited distinct, meaningful localization patterns to the histone gene array and therefore warrant further investigation (**Figure 5.4**). First, we used our bioinformatics pipeline to map a ChIP-seq dataset for the kinase JIL-1, which is responsible for phosphorylating serine 10 on histone 3 (Cai *et al.* 2014; Albig *et al.* 2019). We observed JIL-1 localizing to the histone gene array, specifically to the *H2A/H2B* promoter (**Figure 5.3A**). We observed an additional sharp peak at the *H3/H4* promoter, but this peak is likely an artifact of short read lengths from the dataset and overlaps with a perfect, long GA-repeat sequence in the *H3/H4* promoter (**Supplemental Figure 5.1 in Appendix A)**. JIL-1 is a DNA-binding factor that associates with the Maleless helicase and MSL1, two members of MSLc (Albig *et al.* 2019). In addition to CLAMP performing a role in histone biogenesis, it also plays a role in dosage compensation and associates with MSLc (Larschan *et al.* 2012).

We also observed hormone-like receptor 78 (Hr78) localize to the *H3/H4* promoter (**Figure 5.4B**). Finally, we mapped two isoforms of female sterile (1) homeotic (fs(1)h; the *Drosophila* homolog of BRD4). The long and short isoforms of fs(1)h have distinct binding profiles but both are assumed have a role in chromatin architecture (Kellner *et al.* 2013). We observed that the long isoform, but not the short isoform, localizes to both the *H2A/H2B* and the *H3/H4* promoters (**Figure 5.4C**). Interestingly, Kellner *et al.* (2013) inferred that the fs(1)h long isoform has a unique role in chromatin remodeling by interacting with specific insulator proteins, including CP190, which did not pass our screen (**Figure 5.3C**).

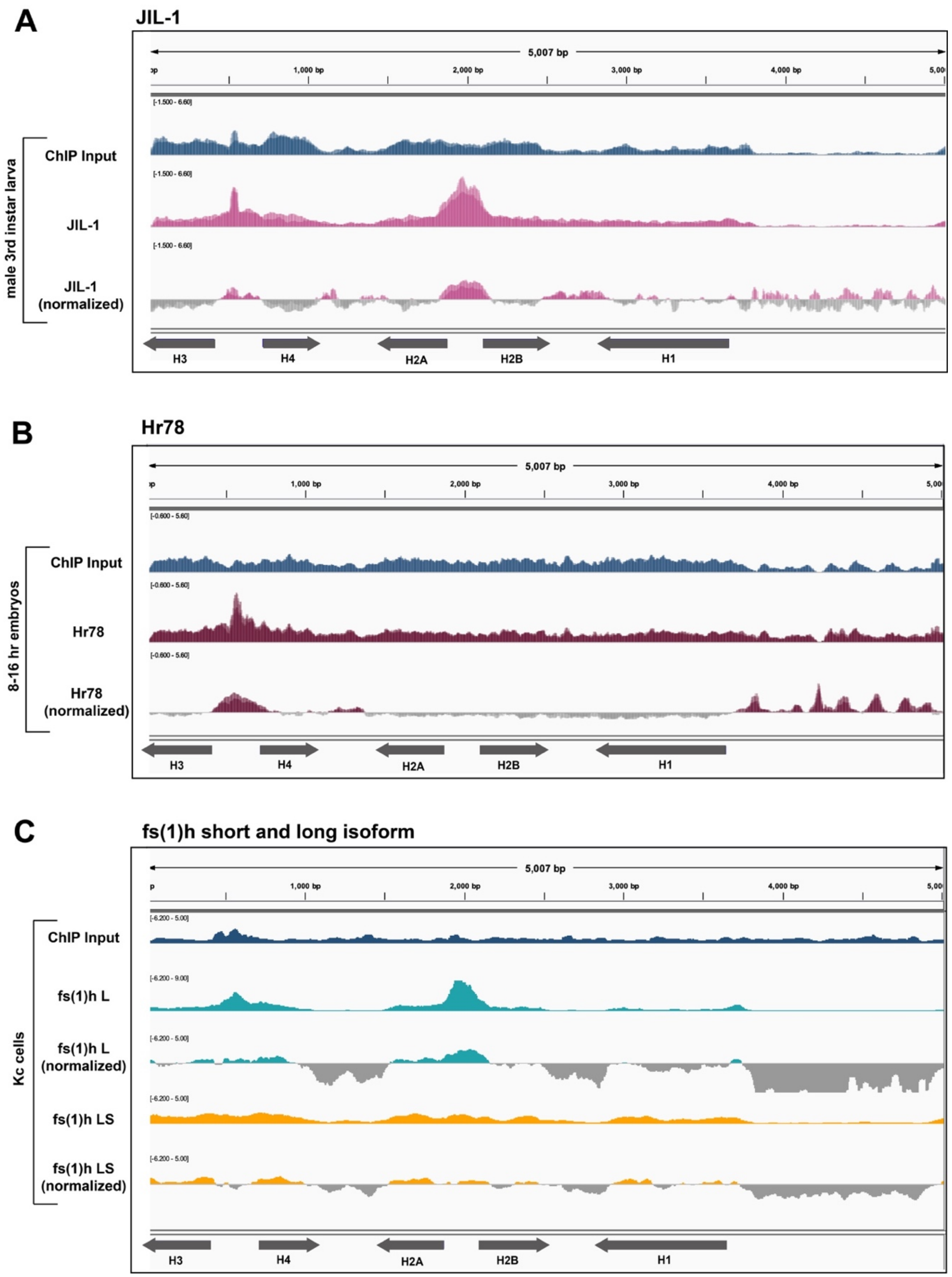


Figure 5.4: JIL-1, Hr78, and Fs(1)hL localize to the histone gene array. We mapped ChIP datasets for (A) JIL-1 (pink, two replicates overlayed, (56)) from male third instar larva, (B) Hr78 (maroon, two replicates overlayed, (72)) from 8-16 hr mixed population embryos and (C) the long((L, teal) and short (S, yellow) isoform of fs(1)h from Kc cells (59) to the histone gene array. We normalized each ChIP-seq dataset to its respective input (blue).

5.4.4 Hox factors localize to the *Drosophila* histone gene array when overexpressed in cell culture

Hox factors are critical for developmental processes like morphogenesis in which cells are constantly dividing and therefore require a near constant supply of histones (Duronio and Marzluff 2017). Histone biogenesis is critical within the first few hours of *Drosophila* development (Amodeo *et al.* 2015; Chari *et al.* 2019). We therefore investigated histone array localization patterns of transcription factors that are critical during early development, including Hox proteins. We identified a publicly available dataset (**Table 1**) in which Beh *et al.* (2016) individually expressed the three Bithorax complex Hox proteins, Ultrabithorax (Ubx), Abdominal-A (Abd-A) and Abdominal-B (Abd-B), in Kc167 cells and performed ChIP-seq. We used our analysis pipeline to map the Ubx, Abd-A, and Abd-B ChIP-seq datasets to the histone gene array and observed striking localization to the *H3/H4* promoter (**Figure 5.5**). We conclude that when overexpressed in cultured cells, Ubx, Abd-A, and Abd-B all target the histone gene array by ChIP-seq.

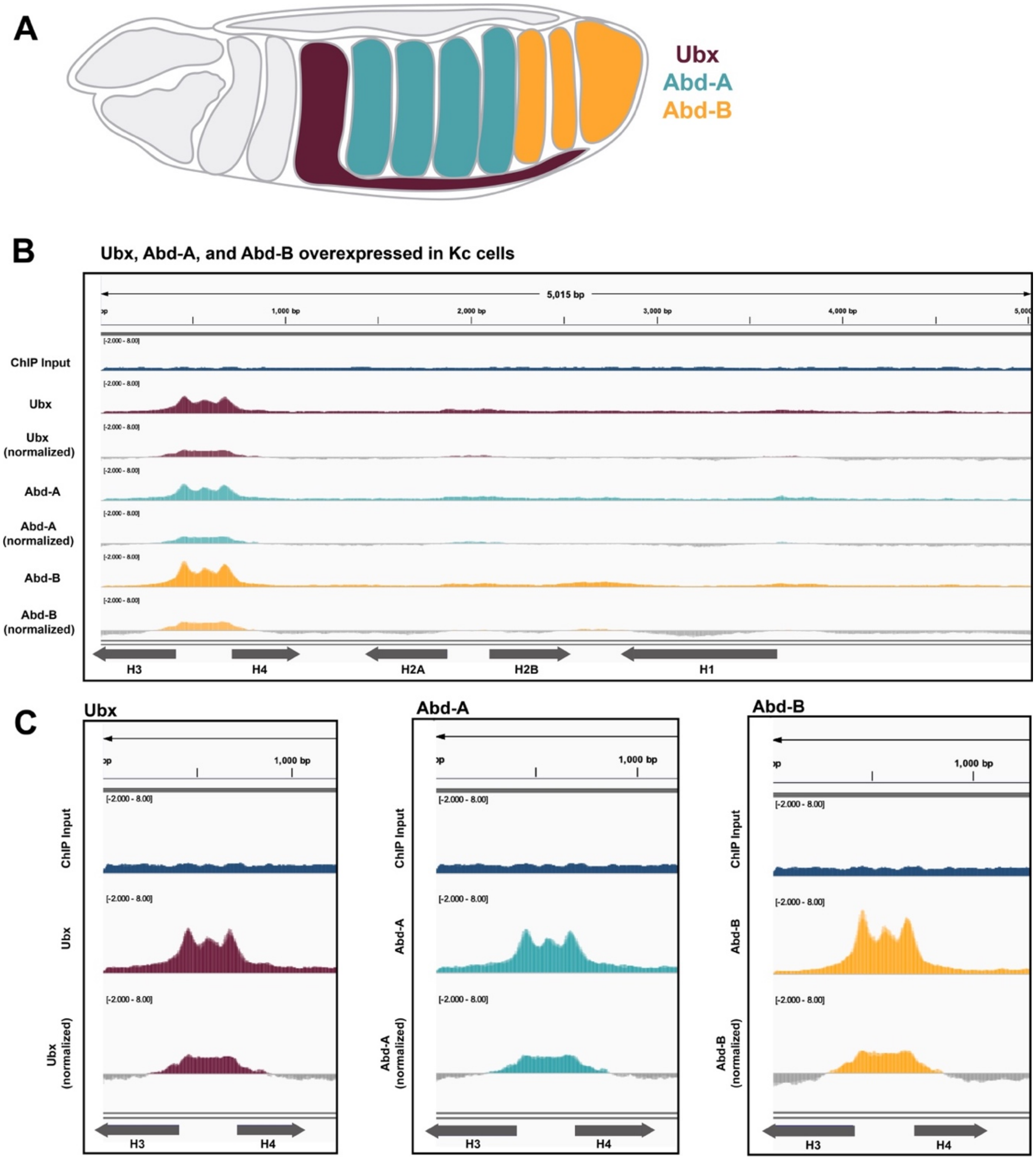


Figure 5.5: Hox factors Ubx, Abd-A, and Abd-B localize to the histone array. (A) Diagram of relative tissue expression patterns for Ubx (maroon), Abd-A (teal) and Abd-B (yellow). (B) We aligned ChIP-seq datasets from Kc cells expressing Ubx (marron, two replicates overlayed, (62)), Abd-A (teal, two replicates overlayed, (62)), and Abd-B (yellow, two replicates overlayed, (62)) to the histone gene array. We normalized each ChIP-seq dataset to the provided input (blue,

two replicates overlayed, (62)). (C) Enlarged signal from (B) of Ubx (maroon), Abd-A (teal), and Abd-B (yellow) over the *H3/H4* promoter.

Because our Hox factor observation (**Figure 5.5**), could be an artifact of overexpression in cultured cells, we identified two additional Ubx ChIP-seq datasets from 0-16 hr embryos and third instar larval imaginal discs (**Table 1**). We used our pipeline to map these data to the histone gene array and observed that Ubx targets the *H3/H4* promoter and, to a lesser extent, the *H2A/H2B* promoter (**Figure 5.6**). We conclude that Ubx targets the histone gene array at various developmental stages and in various tissues, and is therefore a promising candidate for future wet-lab research designed to validate these bioinformatic observations.

To further investigate the relationship between Hox factors and the histone locus, we identified three additional datasets for Hox proteins and Hox cofactors. There are two Hox gene complexes in *Drosophila:* the Bithorax complex (which includes Ubx, Abd-A, and Abd-B) and the Antennapedia complex. We first mapped ChIP-seq data for Antennapedia (Antp) (Kribelbauer *et al.* 2020) but did not observe robust localization to the histone gene array (**Supplementary Figure 5.2 in Appendix A**). We next mapped ChIP-seq datasets for the Hox cofactors extradenticle (Exd) and Homothorax (Hth) (Kribelbauer *et al.* 2020). Exd and Hth associate with the hexapeptide motif in Hox proteins and form heterodimers to impact Hox binding specificity to their gene targets (Rezsohazy *et al.* 2015; Beh *et al.* 2016). We observed that neither Exd nor Hth gave meaningful ChIP signal over the histone gene array (**Supplementary Figure 5.2 in Appendix A**).

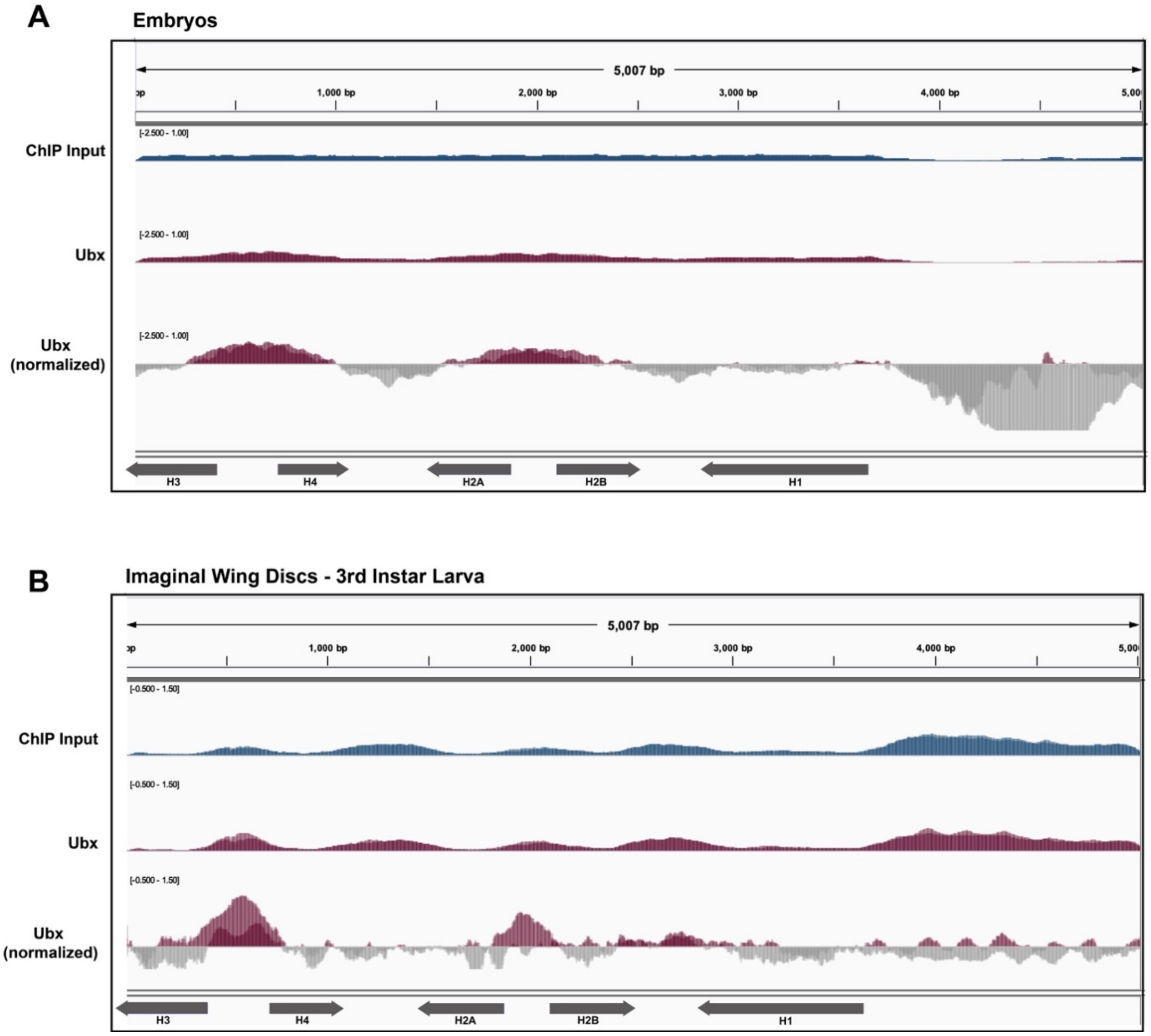


Figure 5.6: Ubx localizes to the H3/H4 promoter in embryos and 3rd instar larva. We mapped Ubx ChIP-seq datasets from (A) mixed population embryos (maroon, top panel, two replicates overlayed, (76)) and (B) imaginal wing discs in third instar larva (maroon, bottom panel, two replicates overlayed, (77)) to the histone gene array. We normalized ChIP-seq datasets to the provided inputs (blue, two replicates overlayed). Signal from the *H3/H4* promoter is enlarged in the panels on the right.

5.4.5 Power and limitations of the screen

The range of results from our candidate screen demonstrates both the power and limitations of our bioinformatics pipeline. In total, we analyzed datasets for 27 different DNA-binding factors and produced 9 candidates that warrant further wet lab investigation. Despite the power of this screen, we are limited by the availability of public datasets. Characteristics of these datasets, such as quality of reads, read length, and inclusions of controls such as inputs are based on the original experimental design and research. Furthermore, we are also restricted by the tissues or genotypes investigated in the original study, limiting the scope of our investigation.

For example, we analyzed several datasets for Nejire (Nej; homolog of mammalian CREB-binding protein (CBP) and Pointed (Pnt). A previous screen in S2 cells identified Nej and Pnt as potential HLB factors (White *et al.* 2011). We mapped a Pnt ChIP-seq dataset from Stage 11 embryos (**Table 1**) and observed that Pnt does not give meaningful signal over the histone gene array (**Figure 5.7**, bottom). Additionally, we investigated two Nej ChIP-seq datasets in which we obtained disparate results. The Nej ChIP-seq dataset from S2 cells did not yield meaningful signal over the histone gene array (**Figure 5.7**, center). In contrast, we investigated a Nej ChIP-seq dataset from early *Drosophila* embryos and observed robust localization to the *H3/H4* promoter, *H2A/H2B* promoter and, to a lesser extent, the *H1* promoter (**Figure 5.7**, top). From these observations, we conclude that Nej likely targets the histone gene array in embryos and would therefore be a strong candidate for future wet-lab studies to validate this observation. Our Pnt and Nej observations demonstrate how our screening approach is powerful but limited by data availability and experimental variables.

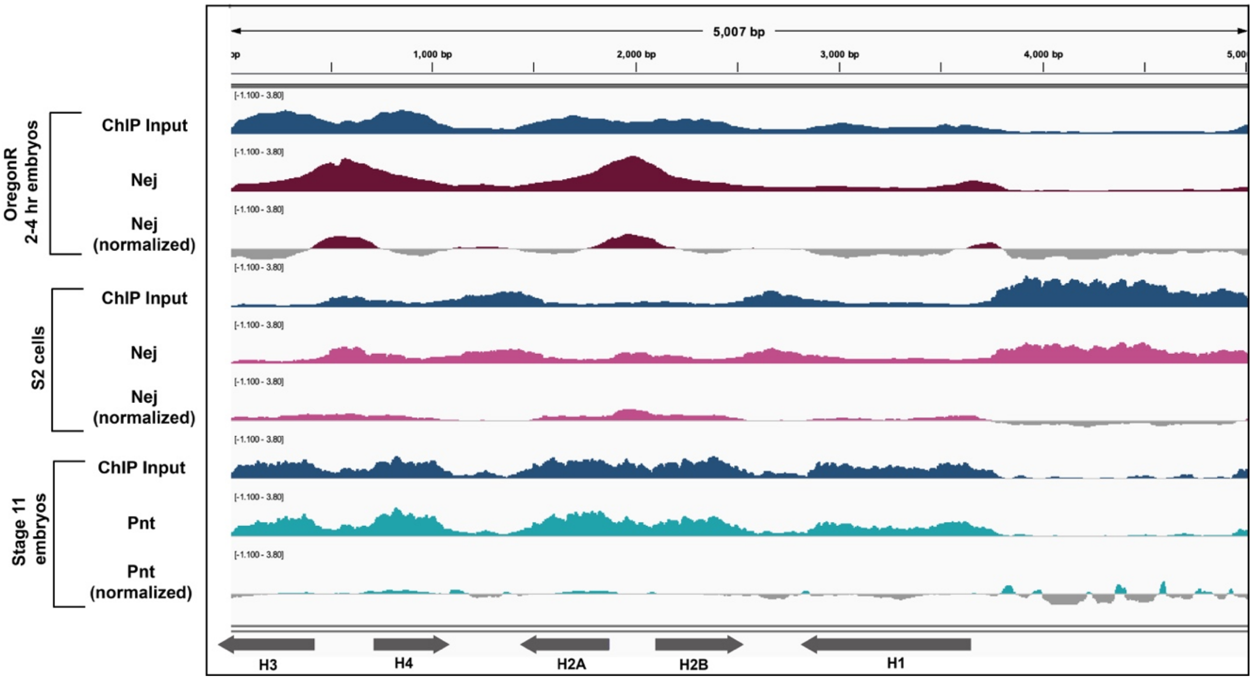


Figure 5.7: ChIP-seq datasets from different tissues can show different alignment results. We mapped two different ChIP-seq datasets for Nejire (Nej) were aligned to the histone gene array. ChIP data from 2-4 hr embryos (maroon, one replicate, (74)), showed localization to the *H3/H4* promoter and the *H2A/H2B* promoter, while ChIP-seq data from S2 cells (pink, one replicate, (73)) showed no localization to the histone gene array. We also aligned ChIP-seq data for Pnt from stage 11 embryos (54) to the histone gene array. We normalized the ChIP-seq signals to their respective input signals (blue).

5.5 Discussion

To broaden our understanding of factors that impact histone biogenesis in *Drosophila melanogaster,* we conducted a candidate-based bioinformatics screen for DNA-binding factors that localize to histone gene array. Although many HLB factors are known, it is likely that there are many other factors critical for histone biogenesis that have yet to be identified, since several have been discovered by chance in the past few years including CLAMP (Rieder *et al.* 2017), Winged-Eye (WGE; (Ozawa *et al.* 2016), and Myc (Daneshvar *et al.* 2011). To begin to close this gap in knowledge, we chose 27 factors based on their roles in chromatin remodeling, dosage compensation, development, and interaction with known HLB factors, hypothesizing that these represent strong candidates for novel HLB factors. As our screen is limited by availability of relevant datasets, it will likely produce both false positives and negatives. Additionally, because we used a targeted screening approach by investigating factors with relevant functions and at relevant developmental timepoints to histone gene expression, we expected more positive hits than we would find using completely unbiased screen. Given our starting pool of 27 factors, we were pleased to produce 9 candidates of potential HLB factors. We envision that the final 9 candidates that passed our qualitative analysis will be investigated through future wet lab experiments (Salzler *et al.* 2013; Rieder *et al.* 2017; Xie *et al.* 2022a).

We validated our bioinformatics pipeline by investigating TRF2, a general transcription factor known to target the histone genes (Isogai *et al.* 2007) and confirmed that TRF2 binds to the TATA-less *H1* promoter. Isogai *et al.* (2007) determined that TBP, another general transcription factor, targets the TATA-containing *H3/H4* and *H2A/H2B* promoters. We expanded this observation by investigating TBP-associated factors TAF1, TFIID, and TFIIF. We discovered that all of these general transcription factors target the histone gene array, further validating our pipeline.

We also discovered that the localization of some factors such as Nej to the histone gene array is tissue specific. Nej emerged from a proteomic screen for factors involved in HLB activation in cultured cells (White *et al.* 2011). However, Nej ChIP-seq from cultured cells did not give meaningful signal over the histone gene array, whereas embryo ChIP-seq showed Nej at histone promoters. These observations denote limitations of our screening technique: we are hindered by the availability and quality of datasets for candidate proteins in specific tissues, genotypes, and conditions. Despite the constraints of data availability, we identified 9 out of 27 candidates that give meaningful signal over the histone gene array based on our qualitative analysis criteria and warrant future wet lab study.

We initially identified several categories of candidate factors, some of which produced positive hit whereas some did not. For example, Scm, which may interact with the confirmed HLB scaffolding factor, Mxc (Docquier *et al.* 1996; Saget *et al.* 1998; Kemp *et al.* 2021) did not show meaningful signal over the histone gene array and therefore we determined that it does not likely target the histone genes.

We also investigated factors involved in dosage compensation, including MSL1, Ndf/CG4747, and JIL-1, because the HLB factor CLAMP plays a key role in male X-chromosome activation. MSL2, another member of the MSLc, was identified in an unbiased proteomics-based HLB candidate screen in cultured cells (White *et al.* 2011), and we recently discovered that MSLc targets one of the two histone loci in *Drosophila virilis* in salivary gland polytene chromosomes (Xie *et al.* 2022b). Although neither MSL1 nor Ndf localized to the histone gene array, JIL-1 robustly localized to the histone gene array.

Of note, the ChIP-seq datasets for MSL1 were produced from S2 cells, the Ndf datasets were from both male and female larvae, and the JIL-1 dataset came from specifically male third instar larva. MSL1 and Ndf may target the histone gene array in other tissues or only in embryos, representing potential false negatives in our bioinformatics screen. However, JIL-1 is a more generalized kinase that is responsible for phosphorylating serine 10 on histone 3 across the genome, not just on the male X-chromosome (Regnard *et al.* 2011; Cai *et al.* 2014; Albig *et al.*

2019). JIL-1 may therefore be present at the histone locus independent of its role in dosage compensation by contributing to the epigenetic landscape of the histone locus. Taken together, our results indicate that dosage compensation and histone gene expression are likely distinct regulatory events, and the majority of factors are not shared between these processes in *Drosophila melanogaster*.

One of the lesser studied characteristics of the histone locus is the local chromatin environment and how epigenetic marks influences histone gene expression. We chose CP190, Gcn5, Psc, Pangolin, and su(z)12 as chromatin remodeling candidates that might target the histone genes but, after mapping relevant datasets, none of these candidate chromatin remodelers target the histone gene array. We did, however, discover that the long isoform of fs(1)h (fs(1)hL) robustly localizes to the histone gene array. Fs(1)hL has a unique role in chromatin remodeling that differs the short fs(1)h isoform, as it associates with insulator proteins, including CP190 (Kellner *et al.* 2013). Since the histone locus is situated near heterochromatin, it is possible that insulators prevent spreading of heterochromatin into the histone locus. CP190 was also a strong candidate for histone locus-association. CLAMP and CP190 share binding profiles at many promoters and each is important for the other's localization (Bag *et al.* 2019). However, when we mapped a CP190 ChIP-seq dataset from female embryos, we did not observe histone array localization. Based on these observations, we conclude that fs(1)hL is a strong candidate for future wet lab studies. Fs(1)hL and CLAMP may interact with CP190 at the histone locus, in specific tissues or at precise developmental timepoints that were not captured in the datasets we investigated.

Finally, we explored several developmental transcription factors because histone biogenesis is critical in the first few hours of *Drosophila* development during rapid zygotic cell divisions. We chose Opa, M1BP, and Hnf4 as candidates. Despite their roles in early development and patterning, these factors did not target the histone gene array. However, we identified Nej (CREB-binding protein; CBP) as a candidate that targets the histone gene array, specifically in *Drosophila* embryos but not in S2 cells. Nej was previously identified as an HLB candidate through a cell-based proteomics screen (White *et al.* 2011). Nej is a histone acetyltransferase, but it has roles in cell proliferation and developmental patterning. Nej could influence the chromatin environment of the histone locus during key times in development or in tissues that are constantly dividing where histone proteins would be needed. Because of the roles Nej plays in general developmental processes, it is a strong candidate for future wet lab studies.

We were surprised to discover that the Hox proteins Ubx, Abd-A and Abd-B, all localize to the histone array when overexpressed in Kc cells. Specifically, these factors all target the *H3/H4* promoter. This ~300 bp promoter is unique within the 5 Kb histone gene array; it is the minimal sequence required for Mxc localization and HLB formation (Salzler *et al.* 2013) and contains critical GA-repeat *cis* elements targeted by CLAMP (Rieder *et al.* 2017). The CLAMP-GA-repeat interaction promotes recruitment of histone-locus specific transcription factors (Rieder *et al.* 2017; Koreski *et al.* 2020). To confirm that our observations are not a byproduct of overexpression, we also investigated independent Ubx ChIP-seq datasets prepared from early embryos (0-16 hrs) and from third instar larval imaginal wing discs. These datasets confirm that Ubx targets the histone gene array, although the distribution across the array varies between tissues. Ubx, as well as Abd-A and Abd-B, are all highly active in the early embryo when histone proteins are needed to organize newly synthesized DNA. Therefore Ubx, Abd-A, and Abd-B could provide a spatial and temporal link between histone biogenesis, cell division, and morphogenesis in the embryo.

With 9 out of 27 hits from our screen emerging as strong candidates for future studies, our screen has proven to be a powerful tool to identify strong candidates for DNA-binding factors that target this histone gene array. However, our screen also demonstrates the limitations of using publicly available data. Although we curated a list of candidates that were based on known characteristics of histone biogenesis, we were limited by several aspects of these datasets, such as quality of reads, read length, and inclusions of proper controls such as inputs. Controls are specifically important to our pipeline because relative peaks at a given location do not always represent true localization. Our negative hits show a range of different negative signals displayed in **Figure 5.3**. In some cases, we saw clear enrichment for open chromatin regions, over promoters and/or gene bodies, but did not characterize these factors as hits based on our qualitative analysis criteria. These regions can be overrepresented in the ChIP sequencing experiment as a whole and, therefore, do not reflect where the DNA-binding factor is truly localizing. This phenomenon is best demonstrated when looking at inputs that also show enrichment over open chromatin or gene bodies as shown in our **Supplemental Figure 5.2 in Appendix A**. Inputs between datasets can be highly variable and, because they are used in the normalization process, can bias the final visualization.

The HLB was discovered by Liu and Gall only seventeen years ago (Liu *et al.* 2006). Since then, novel HLB factors have largely been discovered one at a time by chance. Proteomic screens identified several new candidates but failed to identify known factors, including CLAMP (White *et al.* 2011). A comprehensive inventory of HLB factors is necessary to establish a thorough mechanism of histone biogenesis. Histone regulation is especially critical in the early animal embryo: excess histones drive extra, asynchronous mitotic cycles, whereas depletion of maternal histones lengthens cell division in *Drosophila* embryos (Chari *et al.* 2019). The timing of important early developmental events such as the mid-blastula transition is influenced by histone to DNA ratios (Amodeo *et al.* 2015). Histone levels also affect pre-mRNA splicing in human cells (Jimeno-González *et al.* 2015), and *H1* isoform loss-of-function mutations are associated with B cell lymphomas (Yusufova *et al.* 2021). Factors that influence histone biogenesis likely contribute to these developmental and disease phenotypes.

5.6 Conclusions

Here we present a candidate-based screen for novel histone locus-associating factors. Our screen was largely driven by the undergraduate student coauthors in two stages: first, we identified strong candidates based on their established or inferred roles, second, we identified and mapped relevant ChIP-seq datasets to the histone gene array. A similar recent bioinformatic screen searched through thousands of datasets and hundreds of hematopoietic transcription factors for those associated with the repetitive mammalian rDNA array. This analysis identified numerous candidate transcription factors but required intensive computational pairwise comparisons and thresholding (Antony *et al.* 2022). Another recent screen searched through 1200 chromatin proteins and post-translational modifications to identify those associated with repetitive human centromeres (Corless *et al.* 2023). We instead chose an informed, narrow list of initial candidates and identified 9 out of 27 that we will prioritize for future wet lab studies. Our results not only identify factors that may be involved in histone biogenesis, but also demonstrate the power of a candidate-based bioinformatics screen driven by students.

5.7 Acknowledgments and Authors' contributions

This work was supported by T32GM00008490 and F31HD105452 to LJH, 3R35GM142724-01S2 to CS, K12GM00068 to CAS and HSC; F32GM140778 to CAS; and R00HD092625 and R35GM142724 to LER. We would like to thank the Emory University students who piloted the project during its earliest stages: Mary Wang, Greg Kimmerer, Mellisa Xie, Dabin Cho, Henrik Torres, Yono Bulis, Edgar Hsieh, Shaariq Khan, Andre Mijacika, Sean Parker, Rohan Ramdeholl, Annalise Weber, and Kelly Yoon. We thank all the Rieder Lab members for their helpful contributions to project development.

Conceptualization, LER, HSC, CAS, and LJH. Data curation, LJH, CS, EHA, BAA, KA, APD, KBF, EHF, MRG, SK, MPK, SL, ASL, LJM, NM, JM, BAM, OM, NN, VDN, NFO, TAP, HS, and HZ; Formal Analysis, LJH and CS; Funding Acquisition, LER, CAS, HSC, LJH, and CS; Investigation, LJH, CS, EHA, BAA, KA, APD, KBF, EHF, MRG, SK, MPK, SL, ASL, LJM, NM, JM, BAM, OM, NN, VDN, NFO, TAP, HS, and HZ; Methodology, HSC; Project Administration, LJH, CAS, and LER; Resources, LJH, CAS, HSC, and LER; Software, HSC; Supervision, LJH, CAS, HSC and LER; Validation, CS and LJH; Visualization, LJH and CS, Writing – Original Draft, LJH and LER.; Writing – Review & Editing, LJH, CS, CAS, and LER.

5.8 References

Afgan E., D. Baker, M. van den Beek, D. Blankenberg, D. Bouvier, *et al.*, 2016 The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res 44: W3–W10. https://doi.org/10.1093/nar/gkw343

Albig C., C. Wang, G. P. Dann, F. Wojcik, T. Schauer, *et al.*, 2019 JASPer controls interphase histone H3S10 phosphorylation by chromosomal kinase JIL-1 in Drosophila. Nat Commun 10: 5343. https://doi.org/10.1038/s41467-019-13174-6

Ali T., M. Krüger, S. Bhuju, M. Jarek, M. Bartkuhn, *et al.*, 2017 Chromatin binding of Gcn5 in Drosophila is largely mediated by CP190. Nucleic Acids Res 45: 2384–2395. https://doi.org/10.1093/nar/gkw1178

Amodeo A. A., D. Jukam, A. F. Straight, and J. M. Skotheim, 2015 Histone titration against the genome sets the DNA-to-cytoplasm threshold for the Xenopus midblastula transition. Proceedings of the National Academy of Sciences 112: E1086–E1095. https://doi.org/10.1073/pnas.1413990112

Antony C., S. S. George, J. Blum, P. Somers, C. L. Thorsheim, *et al.*, 2022 Control of ribosomal RNA synthesis by hematopoietic transcription factors. Mol Cell 82: 3826-3839.e9. https://doi.org/10.1016/j.molcel.2022.08.027

Arias Escayola D., and K. M. Neugebauer, 2018 Dynamics and Function of Nuclear Bodies during Embryogenesis. Biochemistry 57: 2462–2469. https://doi.org/10.1021/acs.biochem.7b01262

Bag I., R. K. Dale, C. Palmer, and E. P. Lei, 2019 The zinc-finger protein CLAMP promotes gypsy chromatin insulator function in Drosophila. Journal of Cell Science 132: jcs226092. https://doi.org/10.1242/jcs.226092

Bag I., S. Chen, L. F. Rosin, Y. Chen, C.-Y. Liu, *et al.*, 2021 M1BP cooperates with CP190 to activate transcription at TAD borders and promote chromatin insulator activity. Nat Commun 12: 4170. https://doi.org/10.1038/s41467-021-24407-y

Barry W. E., and C. S. Thummel, 2016 The Drosophila HNF4 nuclear receptor promotes glucose-stimulated insulin secretion and mitochondrial function in adults. Elife 5: e11183. https://doi.org/10.7554/eLife.11183

Baumann D. G., and D. S. Gilmour, 2017 A sequence-specific core promoter-binding transcription factor recruits TRF2 to coordinately transcribe ribosomal protein genes. Nucleic Acids Res 45: 10481–10491. https://doi.org/10.1093/nar/gkx676

Beh C. Y., S. El-Sharnouby, A. Chatzipli, S. Russell, S. W. Choo, *et al.*, 2016 Roles of cofactors and chromatin accessibility in Hox protein target specificity. Epigenetics Chromatin 9: 1. https://doi.org/10.1186/s13072-015-0049-x

Berloco M., L. Fanti, A. Breiling, V. Orlando, and S. Pimpinelli, 2001 The maternal effect gene, abnormal oocyte (abo), of Drosophila melanogaster encodes a specific negative regulator of histones. Proc Natl Acad Sci U S A 98: 12126–12131. https://doi.org/10.1073/pnas.211428798

Bongartz P., and S. Schloissnig, 2018 Deep repeat resolution—the assembly of the Drosophila Histone Complex. Nucleic Acids Research 47: e18–e18. https://doi.org/10.1093/nar/gky1194

Bulchand S., S. D. Menon, S. E. George, and W. Chia, 2010 Muscle wasted: a novel component of the Drosophila histone locus body required for muscle integrity. Journal of Cell Science 123: 2697–2707. https://doi.org/10.1242/jcs.063172

Cai W., C. Wang, Y. Li, C. Yao, L. Shen, *et al.*, 2014 Genome-wide analysis of regulation of gene expression and H3K9me2 distribution by JIL-1 kinase mediated histone H3S10 phosphorylation in Drosophila. Nucleic Acids Res 42: 5456–5467. https://doi.org/10.1093/nar/gku173

Chari S., H. Wilky, J. Govindan, and A. A. Amodeo, 2019 Histone concentration regulates the cell cycle and transcription in early development. Development 146: dev177402. https://doi.org/10.1242/dev.177402

Corless S., N. Pratap-Singh, N. S. Benabdallah, J. Böhm, A. M. Simon, *et al.*, 2023 The bromodomain inhibitor JQ1 is a molecular glue targeting centromeres. 2023.03.15.532673.

Crayton M. E., C. E. Ladd, M. Sommer, G. Hampikian, and L. D. Strausbaugh, 2004 An organizational model of transcription factor binding sites for a histone promoter in D. melanogaster. In Silico Biol 4: 537–548.

Daneshvar K., A. Khan, and J. M. Goodliffe, 2011 Myc Localizes to Histone Locus Bodies during Replication in Drosophila. PLOS ONE 6: e23928. https://doi.org/10.1371/journal.pone.0023928

Docquier F., O. Saget, F. Forquignon, N. B. Randsholt, and P. Santamaria, 1996 The multi sex combs gene of Drosophila melanogaster is required for proliferation of the germline. Rouxs Arch Dev Biol 205: 203–214. https://doi.org/10.1007/BF00365798

Doiguchi M., T. Nakagawa, Y. Imamura, M. Yoneda, M. Higashi, *et al.*, 2016 SMARCAD1 is an ATP-dependent stimulator of nucleosomal H2A acetylation via CBP, resulting in transcriptional regulation. Sci Rep 6: 20179. https://doi.org/10.1038/srep20179

Duan J., L. Rieder, M. M. Colonnetta, A. Huang, M. Mckenney, *et al.*, 2021 CLAMP and Zelda function together to promote Drosophila zygotic genome activation. eLife.

Duronio R. J., and W. F. Marzluff, 2017 Coordinating cell cycle-regulated histone gene expression through assembly and function of the Histone Locus Body. RNA Biol 14: 726–738. https://doi.org/10.1080/15476286.2016.1265198

Feng S., C. Rastogi, R. Loker, W. J. Glassford, H. Tomas Rube, *et al.*, 2022 Transcription factor paralogs orchestrate alternative gene regulatory networks by context-dependent cooperation with multiple cofactors. Nat Commun 13: 3808. https://doi.org/10.1038/s41467-022-31501-2

Follmer N. E., A. H. Wani, and N. J. Francis, 2012 A polycomb group protein is retained at specific sites on chromatin in mitosis. PLoS Genet 8: e1003135. https://doi.org/10.1371/journal.pgen.1003135

Godfrey A. C., A. E. White, D. C. Tatomer, W. F. Marzluff, and R. J. Duronio, 2009 The Drosophila U7 snRNP proteins Lsm10 and Lsm11 are required for histone pre-mRNA processing and play an essential role in development. RNA 15: 1661–1672. https://doi.org/10.1261/rna.1518009

Herz H.-M., M. Mohan, A. S. Garrett, C. Miller, D. Casto, *et al.*, 2012 Polycomb repressive complex 2-dependent and -independent functions of Jarid2 in transcriptional regulation in Drosophila. Mol Cell Biol 32: 1683–1693. https://doi.org/10.1128/MCB.06503-11

Isogai Y., S. Keles, M. Prestel, A. Hochheimer, and R. Tjian, 2007 Transcription of histone gene cluster by differential core-promoter factors. Genes Dev. 21: 2936–2949. https://doi.org/10.1101/gad.1608807

Jaeger S., F. Martin, J. Rudinger-Thirion, R. Giegé, and G. Eriani, 2006 Binding of human SLBP on the 3'-UTR of histone precursor H4-12 mRNA induces structural rearrangements that enable U7 snRNA anchoring. Nucleic Acids Res 34: 4987–4995. https://doi.org/10.1093/nar/gkl666

Jimeno-González S., L. Payán-Bravo, A. M. Muñoz-Cabello, M. Guijo, G. Gutierrez, *et al.*, 2015 Defective histone supply causes changes in RNA polymerase II elongation rate and cotranscriptional pre-mRNA splicing. Proceedings of the National Academy of Sciences 112: 14840–14845. https://doi.org/10.1073/pnas.1506760112

Kang H., K. A. McElroy, Y. L. Jung, A. A. Alekseyenko, B. M. Zee, *et al.*, 2015 Sex comb on midleg (Scm) is a functional link between PcG-repressive complexes in Drosophila. Genes Dev. 29: 1136–1150. https://doi.org/10.1101/gad.260562.115

Kaya-Okur H. S., S. J. Wu, C. A. Codomo, E. S. Pledger, T. D. Bryson, *et al.*, 2019 CUT&Tag for efficient epigenomic profiling of small samples and single cells. Nat Commun 10: 1930. https://doi.org/10.1038/s41467-019-09982-5

Kaye E. G., M. Booker, J. V. Kurland, A. E. Conicella, N. L. Fawzi, *et al.*, 2018 Differential Occupancy of Two GA-Binding Proteins Promotes Targeting of the Drosophila Dosage Compensation Complex to the Male X Chromosome. Cell Rep 22: 3227–3239. https://doi.org/10.1016/j.celrep.2018.02.098

Kellner W. A., K. Van Bortle, L. Li, E. Ramos, N. Takenaka, *et al.*, 2013 Distinct isoforms of the Drosophila Brd4 homologue are present at enhancers, promoters and insulator sites. Nucleic Acids Res 41: 9274–9283. https://doi.org/10.1093/nar/gkt722

Kemp J. P., X.-C. Yang, Z. Dominski, W. F. Marzluff, and R. J. Duronio, 2021 Superresolution light microscopy of the Drosophila histone locus body reveals a core–shell organization associated with expression of replication–dependent histone genes. MBoC 32: 942–955. https://doi.org/10.1091/mbc.E20-10-0645

Koenecke N., J. Johnston, B. Gaertner, M. Natarajan, and J. Zeitlinger, 2016 Genome-wide identification of Drosophila dorso-ventral enhancers by differential histone acetylation analysis. Genome Biology 17: 196. https://doi.org/10.1186/s13059-016-1057-2

Koreski K. P., L. E. Rieder, L. M. McLain, W. F. Marzluff, and R. J. Duronio, 2020 Drosophila Histone Locus Body assembly and function involves multiple interactions. bioRxiv 2020.03.16.994483. https://doi.org/10.1101/2020.03.16.994483

Koromila T., F. Gao, Y. Iwasaki, P. He, L. Pachter, *et al.*, 2020 Odd-paired is a pioneer-like factor that coordinates with Zelda to control gene expression in embryos, (K. Struhl, O. Hobert, and E. Clark, Eds.). eLife 9: e59610. https://doi.org/10.7554/eLife.59610

Kribelbauer J. F., R. E. Loker, S. Feng, C. Rastogi, N. Abe, *et al.*, 2020 Context-Dependent Gene Regulation by Homeodomain Transcription Factor Complexes Revealed by Shape-Readout Deficient Proteins. Molecular Cell 78: 152-167.e11. https://doi.org/10.1016/j.molcel.2020.01.027

Kyrchanova O., N. Klimenko, N. Postika, A. Bonchuk, N. Zolotarev, *et al.*, 2021 Drosophila architectural protein CTCF is not essential for fly survival and is able to function independently of CP190. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms 1864: 194733. https://doi.org/10.1016/j.bbagrm.2021.194733

Langmead B., and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357–359. https://doi.org/10.1038/nmeth.1923

Larschan E., A. A. Alekseyenko, W. R. Lai, P. J. Park, and M. I. Kuroda, 2006 MSL complex associates with clusters of actively transcribed genes along the Drosophila male X chromosome. Cold Spring Harb Symp Quant Biol 71: 385–94. https://doi.org/10.1101/sqb.2006.71.026

Larschan E., M. M. Soruco, O. K. Lee, S. Peng, E. Bishop, *et al.*, 2012 Identification of chromatin-associated regulators of MSL complex targeting in Drosophila dosage compensation. PLoS Genet 8: e1002830. https://doi.org/10.1371/journal.pgen.1002830

Li Z., K. Thiel, P. J. Thul, M. Beller, R. P. Kühnlein, *et al.*, 2012 Lipid droplets control the maternal histone supply of Drosophila embryos. Curr Biol 22: 2104–2113. https://doi.org/10.1016/j.cub.2012.09.018

Li Z., M. R. Johnson, Z. Ke, L. Chen, and M. A. Welte, 2014 Drosophila lipid droplets buffer the H2Av supply to protect early embryonic development. Curr Biol 24: 1485–1491. https://doi.org/10.1016/j.cub.2014.05.022

Liu J.-L., C. Murphy, M. Buszczak, S. Clatterbuck, R. Goodman, *et al.*, 2006 The Drosophila melanogaster Cajal body. J Cell Biol 172: 875–884. https://doi.org/10.1083/jcb.200511038

Marzluff W. F., P. Gongidi, K. R. Woods, J. Jin, and L. J. Maltais, 2002 The human and mouse replication-dependent histone genes. Genomics 80: 487–98.

Matera A. G., M. Izaguire-Sierra, K. Praveen, and T. K. Rajendra, 2009 Nuclear Bodies: Random Aggregates of Sticky Proteins or Crucibles of Macromolecular Assembly? Developmental Cell 17: 639–647. https://doi.org/10.1016/j.devcel.2009.10.017

McKay D. J., S. Klusza, T. J. Penke, M. P. Meers, K. P. Curry, *et al.*, 2015 Interrogating the function of metazoan histones using engineered gene clusters. Dev Cell 32: 373–86. https://doi.org/10.1016/j.devcel.2014.12.025

Ozawa N., H. Furuhashi, K. Masuko, E. Numao, T. Makino, *et al.*, 2016 Organ identity specification factor WGE localizes to the histone locus body and regulates histone expression to ensure genomic stability in Drosophila. Genes to Cells 21: 442–456. https://doi.org/10.1111/gtc.12354

Ramalingam V., M. Natarajan, J. Johnston, and J. Zeitlinger, 2021 TATA and paused promoters active in differentiated tissues have distinct expression characteristics. Molecular Systems Biology 17: e9866. https://doi.org/10.15252/msb.20209866

Ravindranath A., and K. M. Cadigan, 2014 Structure-Function Analysis of the C-clamp of TCF/Pangolin in Wnt/ß-catenin Signaling. PLOS ONE 9: e86180. https://doi.org/10.1371/journal.pone.0086180

Regnard C., T. Straub, A. Mitterweger, I. K. Dahlsveen, V. Fabian, *et al.*, 2011 Global Analysis of the Relationship between JIL-1 Kinase and Transcription. PLOS Genetics 7: e1001327. https://doi.org/10.1371/journal.pgen.1001327

Rezsohazy R., A. J. Saurin, C. Maurel-Zaffran, and Y. Graba, 2015 Cellular and molecular insights into Hox protein action. Development 142: 1212–1227. https://doi.org/10.1242/dev.109785

Rhee H. S., and B. F. Pugh, 2012 ChIP-exo Method for Identifying Genomic Location of DNA-Binding Proteins with Near-Single-Nucleotide Accuracy. Current Protocols in Molecular Biology 100: 21.24.1-21.24.14. https://doi.org/10.1002/0471142727.mb2124s100

Rieder L. E., K. P. Koreski, K. A. Boltz, G. Kuzu, J. A. Urban, *et al.*, 2017 Histone locus regulation by the Drosophila dosage compensation adaptor protein CLAMP. Genes Dev 31: 1494–1508. https://doi.org/10.1101/gad.300855.117

Robinson J. T., H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, *et al.*, 2011 Integrative Genomics Viewer. Nat Biotechnol 29: 24–26. https://doi.org/10.1038/nbt.1754

Saget O., F. Forquignon, P. Santamaria, and N. B. Randsholt, 1998 Needs and targets for the multi sex combs gene product in Drosophila melanogaster. Genetics 149: 1823–1838.

Salzler H. R., D. C. Tatomer, P. Y. Malek, S. L. McDaniel, A. N. Orlando, *et al.*, 2013 A sequence in the Drosophila H3-H4 Promoter triggers histone locus body assembly and biosynthesis of replication-coupled histone mRNAs. Dev Cell 24: 623–34. https://doi.org/10.1016/j.devcel.2013.02.014

Schmidt C. A., L. J. Hodkinson, H. S. Comstra, and L. E. Rieder, 2022 A cost-free CURE: Using bioinformatics to identify DNA-binding factors at a specific genomic locus. 2022.10.21.513244.

Shevtsov S. P., and M. Dundr, 2011 Nucleation of nuclear bodies by RNA. Nat Cell Biol 13: 167–173. https://doi.org/10.1038/ncb2157

Shlyueva D., A. C. A. Meireles-Filho, M. Pagani, and A. Stark, 2016 Genome-Wide Ultrabithorax Binding Analysis Reveals Highly Targeted Genomic Loci at Developmental Regulators and a Potential Connection to Polycomb-Mediated Regulation. PLoS One 11: e0161997. https://doi.org/10.1371/journal.pone.0161997

Stephenson R. A., J. M. Thomalla, L. Chen, P. Kolkhof, R. P. White, *et al.*, 2021 Sequestration to lipid droplets promotes histone availability by preventing turnover of excess histones. Development 148: dev199381. https://doi.org/10.1242/dev.199381

Straub T., A. Zabel, G. D. Gilfillan, C. Feller, and P. B. Becker, 2013a Different chromatin interfaces of the *Drosophila* dosage compensation complex revealed by high-shear ChIP-seq. Genome Res. 23: 473–485. https://doi.org/10.1101/gr.146407.112

Straub T., A. Zabel, G. D. Gilfillan, C. Feller, and P. B. Becker, 2013b Different chromatin interfaces of the Drosophila dosage compensation complex revealed by high-shear ChIP-seq. Genome Res 23: 473–485. https://doi.org/10.1101/gr.146407.112

Szklarczyk D., A. L. Gable, D. Lyon, A. Junge, S. Wyder, *et al.*, 2019 STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 47: D607–D613. https://doi.org/10.1093/nar/gky1131

Tatomer D. C., E. Terzo, K. P. Curry, H. Salzler, I. Sabath, *et al.*, 2016 Concentrating pre-mRNA processing factors in the histone locus body facilitates efficient histone mRNA biogenesis. Journal of Cell Biology 213: 557–570. https://doi.org/10.1083/jcb.201504043

Terzo E. A., S. M. Lyons, J. S. Poulton, B. R. S. Temple, W. F. Marzluff, *et al.*, 2015 Distinct self-interaction domains promote Multi Sex Combs accumulation in and formation of the Drosophila histone locus body. Mol Biol Cell 26: 1559–1574. https://doi.org/10.1091/mbc.E14-10-1445

The Galaxy Community, 2022 The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. Nucleic Acids Research 50: W345–W351. https://doi.org/10.1093/nar/gkac247

THE MODENCODE CONSORTIUM, S. Roy, J. Ernst, P. V. Kharchenko, P. Kheradpour, *et al.*, 2010 Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE. Science 330: 1787–1797. https://doi.org/10.1126/science.1198374

Vivekanand P., 2018 Lessons from Drosophila Pointed, an ETS family transcription factor and key nuclear effector of the RTK signaling pathway. Genesis 56: e23257. https://doi.org/10.1002/dvg.23257

Wang C. I., A. A. Alekseyenko, G. LeRoy, A. E. H. Elia, A. A. Gorchakov, *et al.*, 2013 Chromatin proteins captured by ChIP-mass spectrometry are linked to dosage compensation in Drosophila. Nat Struct Mol Biol 20: 202–209. https://doi.org/10.1038/nsmb.2477

Webber J. L., J. Zhang, A. Massey, N. Sanchez-Luege, and I. Rebay, 2018 Collaborative repressive action of the antagonistic ETS transcription factors Pointed and Yan fine-tunes gene expression to confer robustness in Drosophila. Development 145: dev165985. https://doi.org/10.1242/dev.165985

White A. E., B. D. Burch, X. C. Yang, P. Y. Gasdaska, Z. Dominski, *et al.*, 2011 Drosophila histone locus bodies form by hierarchical recruitment of components. J Cell Biol 193: 677–94. https://doi.org/10.1083/jcb.201012077

Wolle D., F. Cleard, T. Aoki, G. Deshpande, P. Schedl, *et al.*, 2015 Functional Requirements for Fab-7 Boundary Activity in the Bithorax Complex. Molecular and Cellular Biology 35: 3739–3752. https://doi.org/10.1128/MCB.00456-15

Xie M., S. Comstra, C. Schmidt, L. Hodkinson, and L. E. Rieder, 2022a Max is likely not at the Drosophila histone locus. 2022.09.11.507040.

Xie M., L. J. Hodkinson, H. S. Comstra, P. P. Diaz-Saldana, H. E. Gilbonio, *et al.*, 2022b MSL2 targets histone genes in Drosophila virilis. 2022.12.14.520423.

Yang X., I. Sabath, L. Kunduru, A. J. van Wijnen, W. F. Marzluff, *et al.*, 2014 A conserved interaction that is essential for the biogenesis of histone locus bodies. J Biol Chem 289: 33767–33782. https://doi.org/10.1074/jbc.M114.616466

Yusufova N., A. Kloetgen, M. Teater, A. Osunsade, J. M. Camarillo, *et al.*, 2021 Histone H1 loss drives lymphoma by disrupting 3D chromatin architecture. Nature 589: 299–305. https://doi.org/10.1038/s41586-020-3017-y

Chapter 6

A cost-free CURE: Using bioinformatics to identify DNA-binding factors at a specific genomic locus

Reproduced with permission from:

Casey A. Schmidt (Ph.D.)[1], Lauren J. Hodkinson (B.S. and B.A.)[2], H. Skye Comstra (Ph.D.)[1], Samia Khan (B.S.)[1], Henrik Torres (high school diploma)[3], and Leila E. Rieder (Ph.D.)[1,2]

[1]Department of Biology, Emory University, Atlanta, GA, USA

[2]Graduate Program in Genetics and Molecular Biology, Emory University, Atlanta, GA, USA

[3]Choate Rosemary Hall, Wallingford, CT, USA

6.1 Abstract

Research experiences provide diverse benefits for undergraduates. Many academic institutions have adopted course-based undergraduate research experiences (CUREs) to improve student access to research opportunities. However, potential instructors of a CURE might still face financial or practical hurdles that prevent implementation. Bioinformatics research offers an alternative that is free, safe, compatible with remote learning, and may be more accessible for students with disabilities. Here, we describe a bioinformatics CURE that leverages publicly available datasets to discover novel proteins that target an instructor-determined genomic locus of interest. We use the free, user-friendly bioinformatics platform Galaxy to map ChIP-seq datasets to a genome, which removes the computing burden from students. Both faculty and students directly benefit from this CURE, as faculty can perform candidate screens and publish CURE results. Students gain not only basic bioinformatics knowledge, but also transferable skills, including scientific communication, database navigation, and primary literature experience. The CURE is flexible and can be expanded to analyze different types of high-throughput data or to investigate different genomic loci in any species.

6.2 Introduction

Undergraduate research experiences are invaluable to students. Documented benefits include retention in STEM (1), increased confidence in research abilities (2), and inclusion of underrepresented populations (3). Yet many students struggle to find a space in laboratories already at capacity. Course-based undergraduate research experiences, or CUREs, can remedy this problem, as they offer students authentic research experiences within the context of a classroom (4). Not only do CUREs involve many more undergraduates in research than the traditional "apprentice" model, but they also allow faculty (especially those with high teaching responsibilities) to make research progress. For example, the instructor of a CURE course can perform a screen (5, 6), follow up on an interesting result from their lab (7), or increase the rigor and reproducibility of a research project through replication by different lab groups or sections.

Despite these clear benefits, there are often limitations to running bench-based CUREs. For example, large schools with high enrollment might face space and time constraints. In addition, the materials required to perform wet-lab experiments may be expensive and time-consuming to prepare for large classes. Overall, these and other limitations can be prohibitive to implementing wet-laboratory CUREs (8).

Bioinformatics CUREs can skirt these hurdles. Because laboratory space is not necessary, the class can be held in a computer lab, a classroom (if the students have access to personal laptops), or completely virtually. There are no costly reagents to purchase or biohazard concerns. Bioinformatics research can offer students with disabilities a less physically demanding alternative to bench-based experiments. It is also compatible with remote or asynchronous teaching, which became necessary during the early COVID-19 pandemic (9, 10).

Although bioinformatics research is typically performed on expensive computing clusters, we instead use Galaxy (11), which is a free, user-friendly platform that integrates many widely-used bioinformatics tools. All memory-intensive computing is performed on Galaxy's servers, allowing students to simply set up commands, execute, and log off; neither sophisticated programming knowledge nor computing power is needed. Bioinformatics research is easily integrated into students' busy schedules, and each activity can typically be completed in less time than a traditional 3-hour wet laboratory. Students participating in bioinformatics CUREs report high sense of achievement and high levels of satisfaction with their projects (12). Furthermore, students can publish their discoveries, which fosters a sense of belonging to the scientific community (13, 14).

Here, we document a successful CURE that applies bioinformatics tools to discover candidate DNA-binding factors that interact with a genomic locus. Specifically, we investigated the *Drosophila melanogaster* histone gene array, which encodes the replication-dependent histones. Because histones undergo non-canonical mRNA processing and exhibit cell cycle-dependent expression, they require a unique suite of transcription and processing factors (15). Although many of these factors are known, the complete inventory of histone gene expression regulators remains incomplete.

In this CURE, students utilize a hypothesis-based candidate approach to identify existing high-throughput datasets (specifically, ChIP-seq or similar techniques). By mapping the reads from a ChIP-seq experiment to the *Drosophila* histone gene array and critically examining the alignment data, students determine if a transcription factor likely targets the locus, suggesting that it may contribute to histone biogenesis. Our approach functions as a primary screen to identify candidate regulatory proteins and provides opportunities for wet-lab follow-up undergraduate research projects (for example, co-immunostaining for the candidate and a positive control to validate bioinformatics findings) (16).

We piloted our CURE remotely with students who were confined at home during the early COVID-19 pandemic. We then transitioned to an in-person experience during a 50-minute weekly "discussion" period attached to a sophomore-level genetics course. Over the course of a semester, each student chose at least one protein to investigate, identified appropriate datasets, mapped datasets to the *Drosophila* histone gene array using Galaxy, and produced alignment figures. The semester culminated in a poster session, during which the students presented their findings to members of the Biology Department (i.e., faculty, staff, and students).

The CURE presented here is beneficial to all parties involved: not only did the students obtain valuable research experience and transferable skills, but they also identified new candidate factors to further investigate in our wet laboratory, allowing us to make research progress (14, 16). There are thousands of ChIP-seq datasets across multiple repositories that are available for analysis. Future students could examine other types of high-throughput datasets, such as ATAC-seq, to further probe the chromatin landscape of the histone gene locus. The bioinformatics analysis presented here can be extended to any annotated locus of interest in any organism. These seemingly endless possibilities support the sustainable implementation and adaptation of this CURE.

6.2.1 Intended audience

We implemented this CURE in a 200-level genetics course that contained 25 sophomores, juniors, and seniors, most of whom were biology majors. Previously, we piloted the CURE virtually with smaller groups of volunteer college students of similar demographics. We also sponsored a remote high school student, indicating that students with a wide range of experience levels can perform the research with appropriate training.

6.2.2. Learning time

The course had two 75-minute lecture periods and one 50-minute "discussion" period per week over a 14-week semester. Traditionally, the discussion period for this course was used for worksheets, activities, and/or literature discussions. Instead, we implemented the CURE during this time over the entire semester, which accounted for 20% of their overall course grade.

6.2.3 Prerequisite student knowledge

We covered all of the background information on conceptual topics, such as transcription factors and ChIP-seq, in the lecture portion of the class (see Appendix A for ChIP-seq resources). Therefore, the only prerequisites for the CURE were the course prerequisites (freshmen-level introductory courses for biology majors). In addition, students did not need prior bioinformatics or computer science experience; all required skills were taught in the training modules.

6.2.4 Learning objectives

Our overall goal was to provide students with an authentic bioinformatics research experience. Upon completion of this CURE, students will be able to:

1. Search peer-reviewed literature to identify candidate proteins that target the *Drosophila* histone gene locus.
2. Form a hypothesis about the candidate protein based on background literature.

3. Identify appropriate datasets (e.g., ChIP-seq or CUT&RUN) through literature or database searches.
4. Map datasets to the *Drosophila* histone gene locus using bioinformatics tools in Galaxy.
5. Visualize data by producing alignment figures in Integrative Genomics Viewer (IGV) software.
6. Synthesize data and conclude if the candidate targets the *Drosophila* histone gene locus.
7. Propose at least two follow-up experiments related to the candidate protein based on ChIP-seq outcome.
8. Present findings to a wider audience (i.e., peers and department) at an in-person poster session.

6.3 Procedure

6.3.1 Materials

The following materials are required for this CURE:

- Computer and Internet access
- Galaxy account (free web-based platform, www.usegalaxy.org)
- Integrative Genomics Viewer software (free downloadable software, https://software.broadinstitute.org/software/igv/)
- Learning management software such as Canvas, or cloud storage program such as Google Drive or OneDrive to house files and course materials
- Customizable form software, such as Google forms, to assess weekly student progress. Alternatively, students could use a software such as Benchling, OneNote, or Google Docs as a lab notebook, and allow instructors access to monitor progress

- Poster making software, such as PowerPoint, Google Slides, or BioRender
- Poster printing facility or online poster platform such as SpatialChat
- Optional: video production software such as Zoom, if the instructor is generating pre-recorded tutorials or the CURE is conducted remotely

Week	Category	Topic	Assignment
1	Background	Introductions - histone gene expression, high-throughput dataset repositories	
2		Discuss review paper (15)	Read paper (due before class)
3		Discuss research paper (17)	Read paper (due before class)
4		Histone gene expression knowns & unknowns; how to select a candidate	Fill out Google spreadsheet with your candidate
5	Tutorials	Tutorial - finding data (NCBI GEO) and Galaxy introduction	Google form with screenshot
6		Tutorial - Galaxy commands	Google form with screenshot
7		Tutorial - Galaxy outputs, IGV	Google form with screenshot
8	Work days	Work session 1	Google form with screenshot
9		Work session 2	Google form with screenshot
10		Work session 3	Google form with screenshot
11		Work session 4	Google form with screenshot
12		Poster tutorial (work session 5)	
13		Poster making session (work session 6)	Poster draft
14	Poster session	Poster session	Fill out 3 peer review forms during the poster session

Figure 6.1: Weekly class schedule for the CURE. We divided the 14-week semester into 4 categories (background, tutorials, work days, and poster session). We assessed student participation through activity logs (Google forms).

6.3.2 Student instructions

Students received the schedule (**Figure 6.1**) at the beginning of the semester, which we divided into four general categories:

1. Background (weeks 1-4), during which students read and discussed review (15) and research (17) articles
2. Tutorials (weeks 5-7), during which students learned how to use Galaxy and IGV through instructor-led in-person tutorials
3. Work days (weeks 8-13), during which students independently carried out their bioinformatics analyses and created their poster, under in-person supervision from the instructor
4. The poster session (week 14), during which students presented their work

For the background sessions, we assigned small groups a figure from the review and research papers to annotate using a presentation template (Appendix 2). During the tutorial and work day sessions, students completed a Google form at the end of class describing their efforts and progress that day (Appendix 3). At the poster session, each student presented their poster and filled out three peer review forms (Appendix 4).

6.3.3 Faculty instructions

6.3.3.1 Background

Our class met twice weekly for the lecture portion (75-minute periods) and once weekly for the bioinformatics CURE portion (two sections of a 50-minute period) over 14 weeks (see **Figure 6.1** for the schedule). During lectures, we followed a “molecules first” rather than “Mendel first” approach (18) to introduce CURE-relevant concepts earlier. For example, concepts covered in the first weeks included transcription, transcriptional regulation, and epigenetics. Lecture topics also paid special attention to high-throughput procedures, such as ChIP-seq (see Appendix A for ChIP-seq teaching resources). Students learned how wet-lab scientists generate sequencing data, how to identify appropriate experimental controls, and the types of research questions that these techniques address. This approach synchronized the lecture and discussion sessions, and provided students with the required background knowledge for the CURE.

During the discussion period, we spent the first four weeks introducing students to *Drosophila* histone gene expression through literature discussions. Students read and discussed both a review article (15) and a research article that used a bioinformatics approach similar to that introduced in the CURE (17). For each paper, we assigned small groups a figure to annotate during class and submit to the instructor (see Appendix 2), which served as their graded assessment for the week.

In the fourth week, we shifted to candidate protein selection. Students gathered additional information on histone gene expression and DNA-binding proteins from PubMed and FlyBase (19). We gave several guiding criteria for finding a candidate factor, such as: (A) proteins that interact with known histone regulators, using protein interaction databases such as STRING (https://string-db.org/) (20); (B) transcription factors that act in the early *Drosophila* embryo, which requires rapid histone biosynthesis (21); (C) DNA-binding factors implicated in cell cycle progression, as histone expression is linked to S-phase (15); and (D) dosage compensation factors, because a prominent histone gene regulator is also involved in dosage compensation (17). Students worked independently while the instructor circulated the classroom for individual ad hoc check-ins. Although the instructors provided guidance, candidate selection was ultimately student-driven. At the end of this class period, students recorded their chosen protein on a class-wide Google spreadsheet, which served as their assessment for the week.

6.3.3.2 Tutorials

We followed background and brainstorming sessions with three weeks of bioinformatics tutorials, during which we led students through analysis and visualization of example data using Galaxy (11) and IGV (22). Pre-recorded tutorials were also posted on our learning management site (Canvas) for students to reference outside class, and contained the same information as what was presented in class. In the tutorials, we used ChIP-seq data from the background primary research article (17) to ensure that their results matched the published figures. See **Figure 6.2** for an overview of the tools we used in Galaxy, and Appendix 5 for the Galaxy workflow tutorial. Due to computing demands on the Galaxy servers, some tools can take several hours to complete. During any downtime, students continued their background research on candidate proteins in preparation for designing their poster. We consulted with each student individually during class time to provide guidance, and students could also come to office hours for additional help.

Galaxy tool	**Description**	**Input**	**Output**
Faster Download and Extract Reads in FASTQ	Extracts sequencing reads (.fastq files) from an SRA import folder	SRA import folder	.fastq file(s)

FastQC	Quality control of the sequencing reads	.fastq file(s)	(1) “webpage” readout (2) “raw data” readout
Bowtie2	Aligns sequencing reads to genome (either built-in genome or user-provided genome)	(1) .fastq file(s) (2) Normalized .fasta genome file (only if using a user-provided genome)	.bam file (ChIP-seq reads mapped to user-specified genome)
bamCoverage	Converts .bam files to .bigwig files, which are better for visualization	.bam file	.bigwig file
bamCompare	Normalizes experimental conditions to input (if available)	(1) Input .bam file (2) Experimental .bam file	.bigwig file

Figure 6.2: Summary of tools used in Galaxy. Each tool can be found by using the search function in Galaxy (see Appendix for Galaxy tutorial).

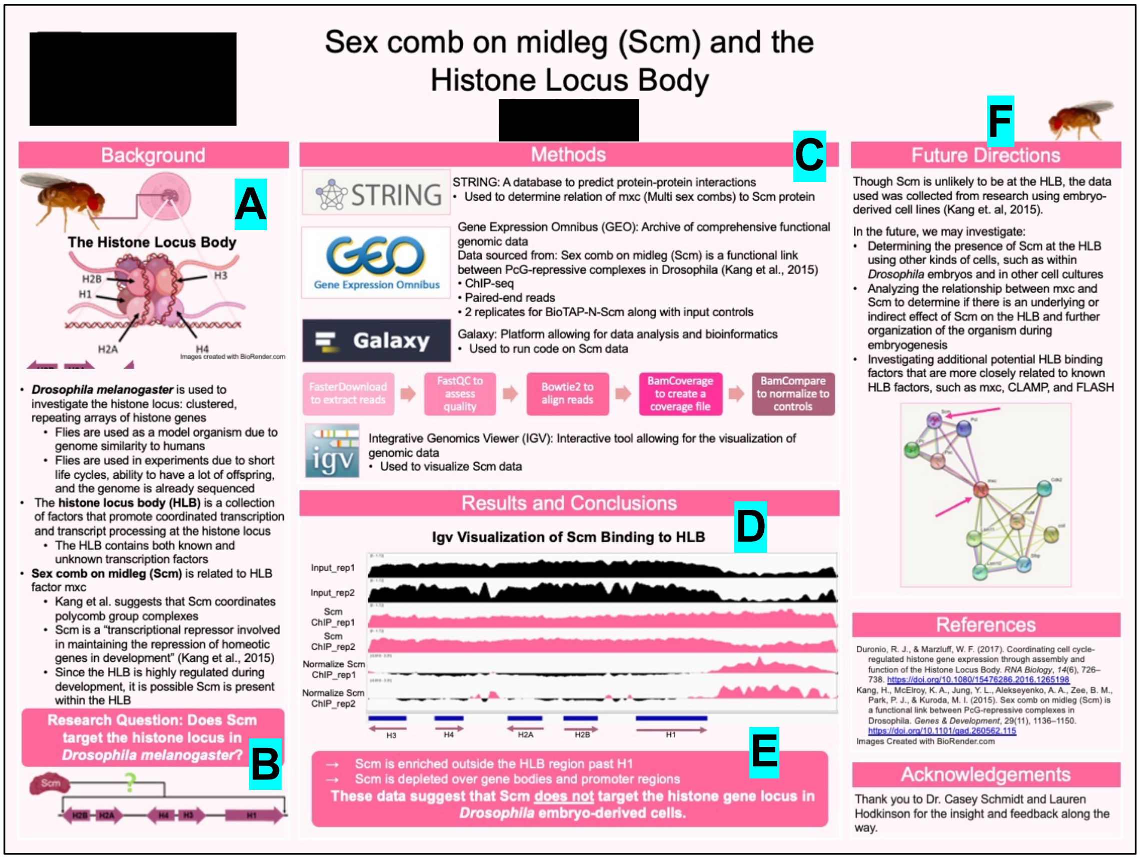
6.3.3.3 Work days

The next six discussion periods functioned as work sessions for students to carry out their bioinformatics analyses. Because the majority of NIH-funded high-throughput sequencing experiments are deposited into public databases such as the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/), many students identified ChIP-seq datasets for their candidate protein(s) by directly searching this database. Others located GEO accession numbers within primary literature. We also directed students to additional databases such as modENCODE (23), which contains ChIP-seq datasets for various transcription factors from numerous tissues and developmental timepoints in model organisms. In several cases, students formed strong hypotheses, but ChIP-seq data did not yet exist for their candidate protein. We instructed these students to choose another factor, and this did not bias their grade. Although some students went through this selection process several times, all found a unique factor to investigate. Once each student located usable data, they carried out the analysis pipeline in Galaxy (**Figure 6.2**) and subsequently generated alignment figures using IGV (Appendix 5). Several students had time to investigate multiple (often related) candidates based on the conclusions from their first hypothesis.

We dedicated two of the work sessions to poster design. We presented resources for crafting posters (for example, https://www.posternerd.com/tutorials) and shared our assessment rubric (Appendix 6). Students submitted a draft of the poster in week 13 (Figure 6.1), for which we provided written feedback and allowed students to revise before printing.

6.3.3.4 Poster session

We held the poster session on the last day of the discussion period and invited members of the Biology Department to attend. See **Figure 6.3** for an example student poster. We divided the students into two groups: while the first group presented their posters, each student in the second group filled out three peer review forms (Appendix 4), which served as a graded assessment. The students switched roles halfway through the session.



[redacted]

Sex comb on midleg (Scm) and the Histone Locus Body

[redacted]

Background

The Histone Locus Body
H2B
H1
H3
H2A
H4
A

Images created with BioRender.com

- ***Drosophila melanogaster*** is used to investigate the histone locus: clustered, repeating arrays of histone genes
  - Flies are used as a model organism due to genome similarity to humans
  - Flies are used in experiments due to short life cycles, ability to have a lot of offspring, and the genome is already sequenced
- The **histone locus body (HLB)** is a collection of factors that promote coordinated transcription and transcript processing at the histone locus
  - The HLB contains both known and unknown transcription factors
- **Sex comb on midleg (Scm)** is related to HLB factor mxc
  - Kang et al. suggests that Scm coordinates polycomb group complexes
  - Scm is a "transcriptional repressor involved in maintaining the repression of homeotic genes in development" (Kang et al., 2015)
  - Since the HLB is highly regulated during development, it is possible Scm is present within the HLB

Research Question: Does Scm target the histone locus in *Drosophila melanogaster*?

Methods

STRING: A database to predict protein-protein interactions
- Used to determine relation of mxc (Multi sex combs) to Scm protein

Gene Expression Omnibus (GEO): Archive of comprehensive functional genomic data

Data sourced from: Sex comb on midleg (Scm) is a functional link between PcG-repressive complexes in Drosophila (Kang et al., 2015)
- ChIP-seq
- Paired-end reads
- 2 replicates for BioTAP-N-Scm along with input controls

Galaxy: Platform allowing for data analysis and bioinformatics
- Used to run code on Scm data

FasterDownload to extract reads → FastQC to assess quality → Bowtie2 to align reads → BamCoverage to create a coverage file → BamCompare to normalize to controls

Integrative Genomics Viewer (IGV): Interactive tool allowing for the visualization of genomic data
- Used to visualize Scm data

Results and Conclusions

Igv Visualization of Scm Binding to HLB

→ Scm is enriched outside the HLB region past H1

→ Scm is depleted over gene bodies and promoter regions

These data suggest that Scm <u>does not</u> target the histone gene locus in *Drosophila* embryo-derived cells.

Future Directions

Though Scm is unlikely to be at the HLB, the data used was collected from research using embryo-derived cell lines (Kang et. al, 2015).

In the future, we may investigate:
- Determining the presence of Scm at the HLB using other kinds of cells, such as within *Drosophila* embryos and in other cell cultures
- Analyzing the relationship between mxc and Scm to determine if there is an underlying or indirect effect of Scm on the HLB and further organization of the organism during embryogenesis
- Investigating additional potential HLB binding factors that are more closely related to known HLB factors, such as mxc, CLAMP, and FLASH

References

Duronio, R. J., & Marzluff, W. F. (2017). Coordinating cell cycle-regulated histone gene expression through assembly and function of the Histone Locus Body. *RNA Biology*, *14*(6), 726–738. https://doi.org/10.1080/15476286.2016.1265198

Kang, H., McElroy, K. A., Jung, Y. L., Alekseyenko, A. A., Zee, B. M., Park, P. J., & Kuroda, M. I. (2015). Sex comb on midleg (Scm) is a functional link between PcG-repressive complexes in Drosophila. *Genes & Development*, 29(11), 1136–1150. https://doi.org/10.1101/gad.260562.115

Images Created with BioRender.com

Acknowledgements

Thank you to Dr. Casey Schmidt and Lauren Hodkinson for the insight and feedback along the way.

Figure 6.3: Example student poster. (A) The background section contains an image created in BioRender, information on the specific factor the student investigated, and information on the *Drosophila* histone gene locus. **(B)** The research question accurately summarizes the project. **(C)** The methods section lists the type of data analyzed (ChIP-seq, paired-end reads, 2 replicates), the programs/databases used, and the specific tools in Galaxy. **(D)** In the results section, the student re-labeled the tracks to descriptive titles and increased the default font size. The IGV tracks are also color-coded. **(E)** The conclusion section summarizes the data while considering limitations (i.e. the cell type used for the ChIP experiment). **(F)** At least 2 future experiments are proposed.

6.3.3.5 Notes and recommendations

Instructors may wish to have a set of “backup” datasets for students who have difficulty locating appropriate data. Students can map data from ChIP-seq variation techniques, such as ChIP-nexus (24), CUT&RUN (25), and CUT&Tag (26) using the same bioinformatic analysis as ChIP-seq. However, ChIP-chip (chromatin immunoprecipitation followed by microarray) datasets cannot be used with our pipeline (**Figure 6.2**), because microarrays utilize different analyses. Unfortunately, some datasets do not contain appropriate controls. For example, we routinely find ChIP-seq datasets that do not include an input or control immunoprecipitation (e.g. IgG) condition, which are important to normalize or compare to the experimental ChIP data. The lack of normalization can sometimes lead to misleading or false positive results, wherein small local peaks appear as positive signal (16). Although there is no way to rectify the lack of controls, it allows for important discussions with students on what conclusions one can draw from their datasets.

Suggestions for determining student learning

Student posters were the primary mode of assessment for our CURE (worth 25% of the discussion grade, plus 15% for the poster peer review assignment). The remaining 60% of the discussion grade was based on participation in the research, assessed through student-reported activity logs. It is sometimes difficult to assess inquiry-based research, and the bioinformatics component added additional hurdles for some students. For example, there may not exist appropriate datasets for a student’s selected candidate, Galaxy may perform slowly, or a dataset from a large study may contain many variables (e.g., environmental conditions, mutant genotypes, treatments, tissue types) such that students struggle to determine which samples are relevant (see Appendix 5). Therefore, we emphasized progress and effort over results and did not penalize students for things out of their control. At the end of each discussion session, the students filled out a Google form describing the activities they performed that day. These forms included space to upload a screenshot of Galaxy or IGV (Appendix 3). Through the Google forms, we assessed participation and monitored progress so that we could intervene if necessary. Specifically, we ensured that students had found a dataset for their candidate by week 9 (work session 2 – see **Figure 6.1**), and provided guidance if they had not.

An additional approach to determining student learning is to include formative assessments throughout the semester. For example, groups of students might complete a worksheet such as the Figure Facts template (27) that walks through a figure from a primary research article, which could be a graded formative assessment. Students could also gain presentation experience by sharing a research article that includes the dataset they plan to analyze. The instructor may choose to have students self-report their activities in graded lab notebooks. These assessments offer additional opportunities for instructor feedback but may be impractical in a larger class.

Sample data

We present example candidates (**Figure 6.4**). First, we identified a dataset for GAGA Factor (GAF) (28) and mapped ChIP-seq reads to the *Drosophila* histone array. We classify GAF as a "positive" candidate, due to the strong, broad peak between the *H3* and *H4* genes (**Figure 6.4A**). This result suggests that GAF targets this region of the histone array, and is a good candidate for wet-lab follow up experiments. Second, we mapped ChIP-seq data for the transcription factor Caudal (23), but did not observe meaningful signal (**Figure 6.4B**). Although the normalized panels appear to have signal, the peaks are not reflected in the ChIP panels, suggesting that they are an artifact of normalization and thus not true signal. Other students also observed this phenomenon (**Figure 6.3**). We classify Caudal as a "negative" candidate. The results from these and other CURE iterations are suitable for publication (14, 16).

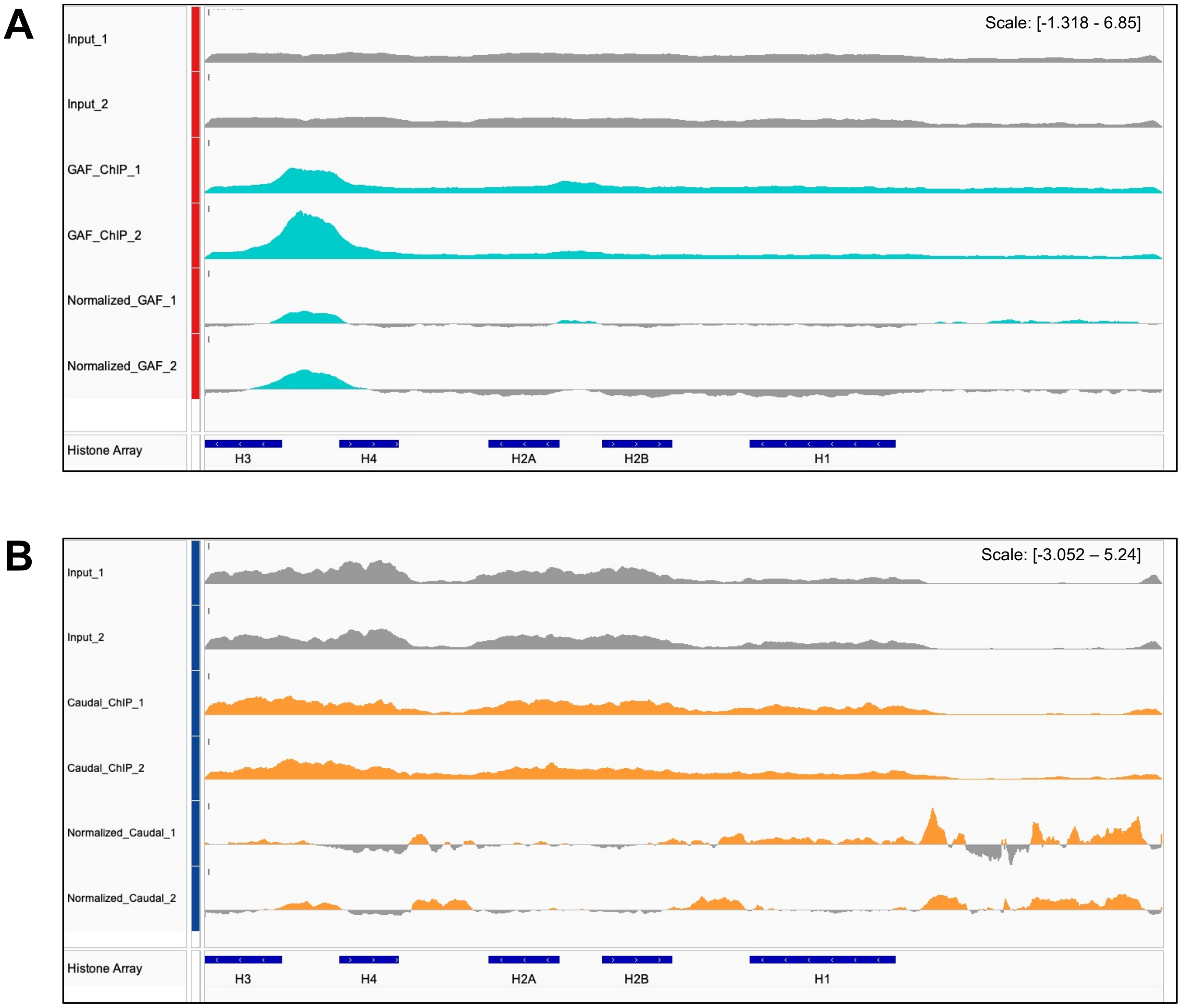


Figure 6.4: Sample data. (A) ChIP-seq alignment of GAGA Factor (GAF) in stage 3 *Drosophila* embryos (teal; input, gray). The figure shows two replicates from the same study. There is a clear peak between the *H3* and *H4* genes, suggesting that GAF localizes to this region. This finding was surprising, given that GAF does not target the histone gene array in cultured S2 cells or by immunofluorescence in early embryos (17). Data from (28), GEO accession

GSE152770. **(B)** ChIP-seq alignment of Caudal in 0-4hr embryos (orange; input, gray). The figure shows two replicates from the same study. Although there is a signal upstream of *H1* in the normalized panels, the peaks are not reflected in the ChIP panels, suggesting that they are not true signal. Thus, there is no clear enrichment of Caudal at the histone gene array. Data from (23), GEO accession GSE20000.

Safety issues

Because this activity does not involve a traditional laboratory setup, we do not foresee any safety issues.

6.4 Discussion

6.4.1 Field testing

We began this bioinformatics project as a strategy to engage our junior laboratory members in remote work during the early COVID-19 pandemic. During the fall of 2020, undergraduates at our institution were not permitted to work in research buildings. Instead, our undergraduate laboratory researchers collectively learned basic bioinformatics skills. Four students each chose a protein to study, identified datasets, mapped data to the histone gene array, and presented their findings to the larger laboratory group. After this first pilot, we recruited nine naive undergraduates from our institution to remotely study the chromatin landscape of the *Drosophila* histone gene array in the spring of 2021. For this iteration, students chose a histone post-translational modification and mapped ChIP-seq data from the modENCODE project (23). The students presented their findings to a wider audience via a virtual poster session. Three students from this group joined our wet laboratory when we returned to in-person instruction and carried out independent projects.

Our laboratory also sponsored a remote high school student that continued the bioinformatics project during the summer of 2021. This student investigated several early *Drosophila* embryo patterning factors (**Figure 6.4**), providing our wet laboratory with candidates for follow-up studies. Most recently, we implemented the project as an in-person CURE in a 200-level genetics course with 25 students.

The class size will likely contribute to the effectiveness of this CURE. Our weekly discussion period was split into two 50-minute sections, with 14 students in one and 11 in the other. This small size allowed us to grant individual attention to each student. Because several of our students ran into difficulties finding appropriate ChIP-seq datasets for their chosen candidate factor, we found that this one-on-one time was necessary to ensure the success of all students, and we recommend a ratio of one instructor to no more than 15 students. If individual conversations are not feasible, the instructor could employ additional experienced TAs to consult with the students, or students could operate in small groups.

6.4.2 Evidence of student learning

We primarily evaluated the CURE learning objectives through the student posters, which served as a summative assessment (**Figure 6.5**). Learning objectives 1-7 were reflected in the poster rubric (Appendix 6). The posters were worth 50 points in total. We gave all students ungraded feedback on their poster before the final submission by providing written comments. Student grades for the poster ranged between 92-100%. Most deductions were related to data presentation, as we instructed students to change the default labels and font size in the IGV plots (Figure 6.5, Appendix 6).

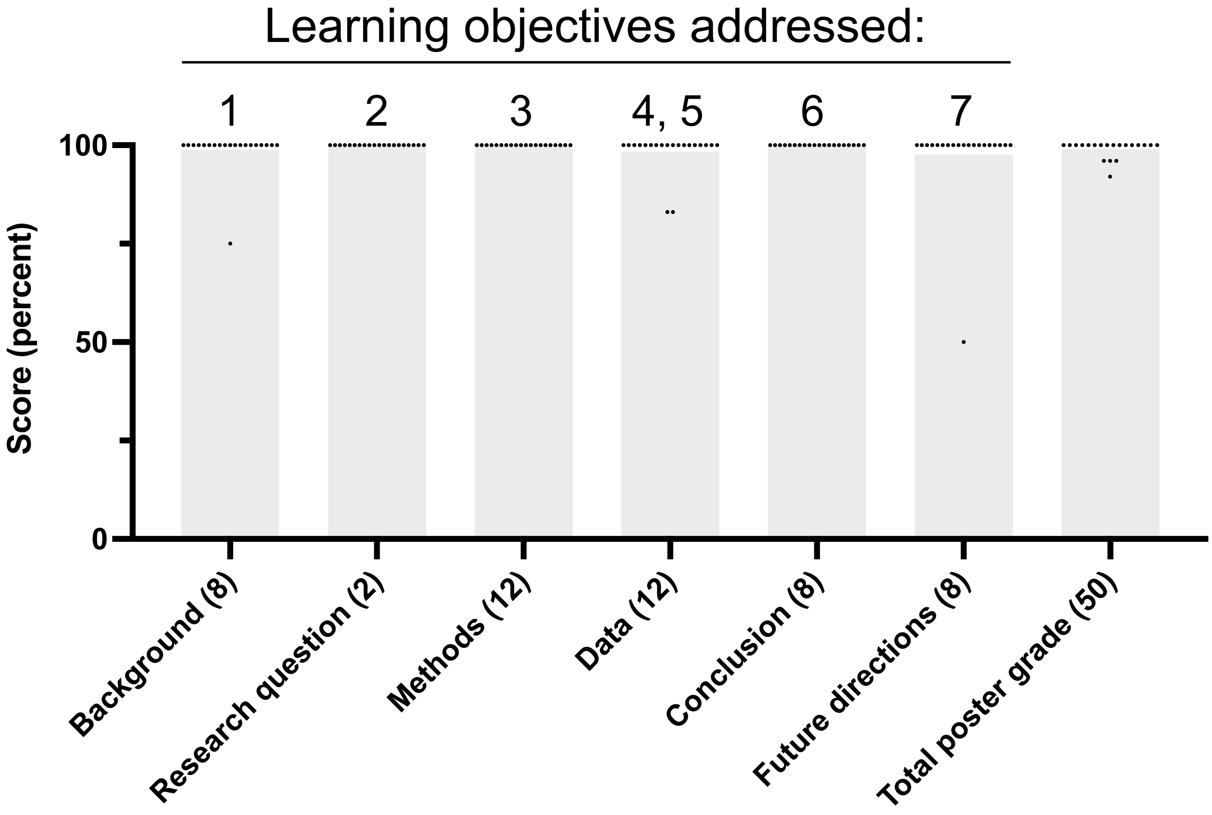
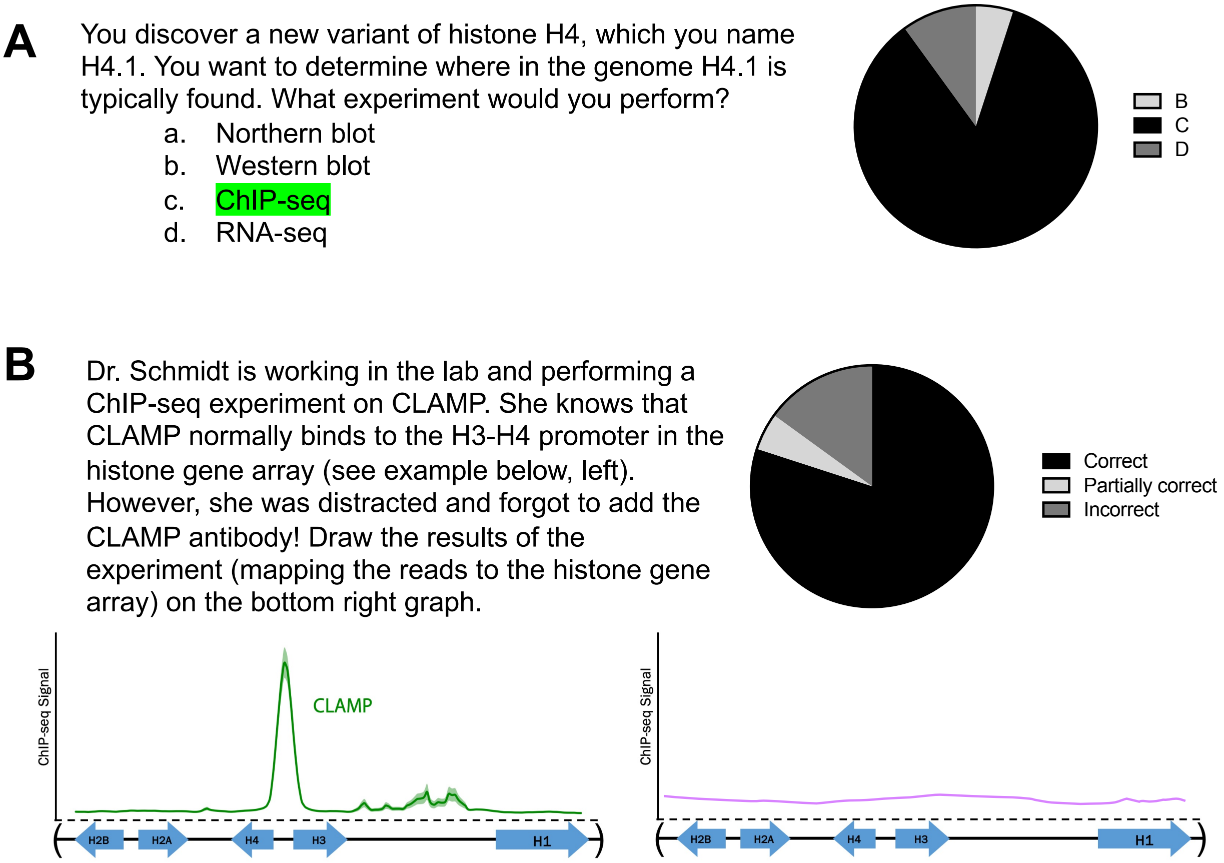


Figure 6.5: The primary form of summative assessment for this CURE was the students' posters. The bar graph represents the score (as a percent) for individual poster sections and the entire poster. Each dot (black) represents an individual student's score. Each bar (gray) represents the average of the dots. The point value of each poster section is listed in parentheses. Learning objectives addressed by each poster section are listed above the bars. Data obtained from consenting students.

We also documented student learning in CURE-related exam questions, which at least 80% of students answered correctly (**Figure 6.6**). For example, we asked what experiment a student would perform to determine the genomic localization of a hypothetical new histone variant protein. This question, which we classify in the “Apply” level of Bloom’s taxonomy (29), required students to recall that histones are DNA binding proteins and to differentiate between types of experiments (**Figure 6.6A**). In addition, we asked students to draw the results of a ChIP-seq experiment if the researcher forgot to add the primary antibody (**Figure 6.6B**). We classify this question in the “Analyze” level of Bloom’s taxonomy, because it addresses the role of different reagents in an experiment. Collectively, these results demonstrate that our students displayed higher-order reasoning on CURE-related topics in their exams.



A You discover a new variant of histone H4, which you name H4.1. You want to determine where in the genome H4.1 is typically found. What experiment would you perform?

a. Northern blot
b. Western blot
c. ChIP-seq
d. RNA-seq

B Dr. Schmidt is working in the lab and performing a ChIP-seq experiment on CLAMP. She knows that CLAMP normally binds to the H3-H4 promoter in the histone gene array (see example below, left). However, she was distracted and forgot to add the CLAMP antibody! Draw the results of the experiment (mapping the reads to the histone gene array) on the bottom right graph.

Figure 6.6: CURE-related exam questions. (A) A multiple-choice exam question answered correctly by 85% of students; the correct answer is highlighted in green. (B) An open-ended exam question answered correctly by 80% of students; a correct answer is drawn on the right panel by the instructor in purple. CLAMP ChIP-seq data from (17). Data obtained from consenting students.

6.4.3 Possible modifications

Because bioinformatics research does not require a wet laboratory setup, this CURE can be implemented remotely and/or asynchronously. We held our CURE pilots synchronously over Zoom during the early COVID-19 pandemic and used the platform SpatialChat (https://spatial.chat) to hold a virtual poster session. In addition, instructors can adapt this CURE to study any genomic locus of interest (for example, an enhancer region that might attract regulatory factors) in any species with an annotated genome. The workflow is particularly suitable for repetitive regions (such as the histone or ribosomal gene arrays) because these regions are often excluded from genome-wide analyses in prior publications. Galaxy contains many built-in genomes, but instructors can also provide a custom genome. We used a custom genome that contains a single copy of the histone gene array (30) because the sequences of the ~100 array copies are nearly identical in the *Drosophila melanogaster* genome (31). This approach amplifies the ChIP-seq signal (**Figure 6.4**) (17). Furthermore, this CURE can be used to map other types of high-throughput data. -For example, students could examine chromatin landscape data, such as ATAC-seq or FAIRE-seq, and compare to histone modification ChIP-seq datasets that correlate with different chromatin states at a particular locus (32).

An exciting follow-up to the bioinformatics CURE is to confirm positive candidates with wet lab experiments. *Drosophila melanogaster* is a particularly useful model organism for these follow-up studies due to the wealth of available mutant and RNAi lines in public stock centers, as well as established protocols for staining tissues. There are also numerous custom antibodies that researchers can request from individual laboratories or purchase from stock centers such as the Developmental Studies Hybridoma Bank (https://dshb.biology.uiowa.edu/). These wet-lab experiments can provide a platform for future studies: for example, testing histone gene expression in the absence of a validated protein that targets the histone gene locus (17).

6.5 Summary

The data generated from this CURE will ultimately add to the growing body of knowledge regarding transcription factor targeting of genomic loci. In addition, the CURE provides students with an authentic research experience, especially in situations where in-person wet laboratory research is not feasible. Students also gain transferable skills that are important for STEM education, including: (A) reading and interpreting primary literature; (B) forming hypotheses based on prior research; (C) navigating complex databases; (D) drawing conclusions from data; and (E) proposing future studies. Furthermore, students interested in continuing bioinformatics research will require less training because they have learned basic bioinformatics techniques. The skills gained during this CURE are crucial to both research science and critical thinking.

6.6 Acknowledgements

We are grateful to the Emory University students who participated in our first (Dabin Cho, Gregory Kimmerer, Mary Wang, and Mellisa Xie) and second (Eric Albanese, Yono Bulis, Edgar Hsieh, Shaariq Khan, Andre Mijacika, Sean Parker, Rohan Ramdeholl, Annalise Weber, and Kelly Yoon) remote pilots, as well as the students who participated in the in-person CURE. This study was determined by the Emory University Institutional Review Board to be exempt from further review (STUDY00005976), and grade data in figures 4 and 5 was obtained only from consenting students. We also thank Dr. Karen Resendes, Dr. Kelsey Gray, Dr. Jennifer Gresham, Dr. Ethan Rundell, and Dr. Michaelyn Hartmann for critically reviewing this manuscript. This work was supported by K12GM00068 to CAS and HSC; F32GM140778 to CAS; T32GM00008490 and F31HD105452 to LJH; and R00HD092625 and R35GM142724 to LER.

6.7 References

1. Eagan MK, Hurtado S, Chang MJ, Garcia GA, Herrera FA, Garibay JC. 2013. Making a Difference in Science Education: The Impact of Undergraduate Research Programs. Am Educ Res J 50:683–713.

2. Szteinberg GA, Weaver GC. 2013. Participants' reflections two and three years after an introductory chemistry course-embedded research experience. Chem Educ Res Pr 14:23–35.

3. Bangera G, Brownell SE. 2014. Course-Based Undergraduate Research Experiences Can Make Scientific Research More Inclusive. CBE—Life Sci Educ 13:602–606.

4. Auchincloss LC, Laursen SL, Branchaw JL, Eagan K, Graham M, Hanauer DI, Lawrie G, McLinn CM, Pelaez N, Rowland S, Towns M, Trautmann NM, Varma-Nelson P, Weston TJ, Dolan EL. 2014. Assessment of Course-Based Undergraduate Research Experiences: A Meeting Report. CBE—Life Sci Educ 13:29–40.

5. Evans CJ, Olson JM, Mondal BC, Kandimalla P, Abbasi A, Abdusamad MM, Acosta O, Ainsworth JA, Akram HM, Albert RB, Alegria-Leal E, Alexander KY, Ayala AC, Balashova NS, Barber RM, Bassi H, Bennion SP, Beyder M, Bhatt KV, Bhoot C, Bradshaw AW, Brannigan TG, Cao B, Cashell YY II, Chai T, Chan AW, Chan C, Chang I, Chang J, Chang MT, Chang PW, Chang S, Chari N, Chassiakos AJ, Chen IE, Chen VK, Chen Z, Cheng MR, Chiang M, Chiu V, Choi S, Chung JH, Contreras L, Corona E, Cruz CJ, Cruz RL, Dang JM, Dasari SP, De La Fuente JRO, Del Rio OMA, Dennis ER, Dertsakyan PS, Dey I, Distler RS, Dong Z, Dorman LC, Douglass MA, Ehresman AB, Fu IH, Fua A, Full SM, Ghaffari-Rafi A, Ghani AA, Giap B, Gill S, Gill ZS, Gills NJ, Godavarthi S,

Golnazarian T, Goyal R, Gray R, Grunfeld AM, Gu KM, Gutierrez NC, Ha AN, Hamid I, Hanson A, Hao C, He C, He M, Hedtke JP, Hernandez YK, Hlaing H, Hobby FA, Hoi K, Hope AC, Hosseinian SM, Hsu A, Hsueh J, Hu E, Hu SS, Huang S, Huang W, Huynh M, Javier C, Jeon NE, Ji S, Johal J, John A, Johnson L, Kadakia S, Kakade N, Kamel S, Kaur R, Khatra JS, Kho JA, Kim C, Kim EJ-K, Kim HJ, Kim HW, Kim JH, Kim SA, Kim WK, Kit B, La C, Lai J, Lam V, Le NK, Lee CJ, Lee D, Lee DY, Lee J, Lee J, Lee J, Lee J-Y, Lee S, Lee TC, Lee V, Li AJ, Li J, Libro AM, Lien IC, Lim M, Lin JM, Liu CY, Liu SC, Louie I, Lu SW, Luo WY, Luu T, Madrigal JT, Mai Y, Miya DI, Mohammadi M, Mohanta S, Mokwena T, Montoya T, Mould DL, Murata MR, Muthaiya J, Naicker S, Neebe MR, Ngo A, Ngo DQ, Ngo JA, Nguyen AT, Nguyen HCX, Nguyen RH, Nguyen TTT, Nguyen VT, Nishida K, Oh S-K, Omi KM, Onglatco MC, Almazan GO, Paguntalan J, Panchal M, Pang S, Parikh HB, Patel PD, Patel TH, Petersen JE, Pham S, Phan-Everson TM, Pokhriyal M, Popovich DW, Quaal AT, Querubin K, Resendiz A, Riabkova N, Rong F, Salarkia S, Sama N, Sang E, Sanville DA, Schoen ER, Shen Z, Siangchin K, Sibal G, Sin G, Sjarif J, Smith CJ, Soeboer AN, Sosa C, Spitters D, Stender B, Su CC, Summapund J, Sun BJ, Sutanto C, Tan JS, Tan NL, Tangmatitam P, Trac CK, Tran C, Tran D, Tran D, Tran V, Truong PA, Tsai BL, Tsai P-H, Tsui CK, Uriu JK, Venkatesh S, Vo M, Vo N-T, Vo P, Voros TC, Wan Y, Wang E, Wang J, Wang MK, Wang Y, Wei S, Wilson MN, Wong D, Wu E, Xing H, Xu JP, Yaftaly S, Yan K, Yang E, Yang R, Yao T, Yeo P, Yip V, Yogi P, Young GC, Yung MM, Zai A, Zhang C, Zhang XX, Zhao Z, Zhou R, Zhou Z, Abutouk M, Aguirre B, Ao C, Baranoff A, Beniwal A, Cai Z, Chan R, Chien KC, Chaudhary U, Chin P, Chowdhury P, Dalie J, Du EY, Estrada A, Feng E, Ghaly M, Graf R, Hernandez E, Herrera K, Ho VW, Honeychurch K, Hou Y, Huang JM, Ishii M, James N, Jang G-E, Jin D, Juarez

J, Kesaf AE, Khalsa SK, Kim H, Kovsky J, Kuang CL, Kumar S, Lam G, Lee C, Lee G, Li L, Lin J, Liu J, Ly J, Ma A, Markovic H, Medina C, Mungcal J, Naranbaatar B, Patel K, Petersen L, Phan A, Phung M, Priasti N, Ruano N, Salim T, Schnell K, Shah P, Shen J, Stutzman N, Sukhina A, Tian R, Vega-Loza A, Wang J, Wang J, Watanabe R, Wei B, Xie L, Ye J, Zhao J, Zimmerman J, Bracken C, Capili J, Char A, Chen M, Huang P, Ji S, Kim E, Kim K, Ko J, Laput SLG, Law S, Lee SK, Lee O, Lim D, Lin E, Marik K, Mytych J, O'Laughlin A, Pak J, Park C, Ryu R, Shinde A, Sosa M, Waite N, Williams M, Wong R, Woo J, Woo J, Yepuri V, Yim D, Huynh D, Wijiewarnasurya D, Shapiro C, Levis-Fitzgerald M, Jaworski L, Lopatto D, Clark IE, Johnson T, Banerjee U. 2021. A functional genomics screen identifying blood cell development genes in Drosophila by undergraduates participating in a course-based research experience. G3 GenesGenomesGenetics 11:jkaa028.

6. Olson JM, Evans CJ, Ngo KT, Kim HJ, Nguyen JD, Gurley KGH, Ta T, Patel V, Han L, Truong-N KT, Liang L, Chu MK, Lam H, Ahn HG, Banerjee AK, Choi IY, Kelley RG, Moridzadeh N, Khan AM, Khan O, Lee S, Johnson EB, Tigranyan A, Wang J, Gandhi AD, Padhiar MM, Calvopina JH, Sumra K, Ou K, Wu JC, Dickan JN, Ahmadi SM, Allen DN, Mai VT, Ansari S, Yeh G, Yoon E, Gon K, Yu JY, He J, Zaretsky JM, Lee NE, Kuoy E, Patananan AN, Sitz D, Tran P, Do M-T, Akhave SJ, Alvarez SD, Asem B, Asem N, Azarian NA, Babaesfahani A, Bahrami A, Bhamra M, Bhargava R, Bhatia R, Bhatia S, Bumacod N, Caine JJ, Caldwell TA, Calica NA, Calonico EM, Chan C, Chan HH-L, Chang A, Chang C, Chang D, Chang JS, Charania N, Chen JY, Chen K, Chen L, Chen Y, Cheung DJ, Cheung JJ, Chew JJ, Chew NB, Chien C-AT, Chin AM, Chin CJ, Cho Y, Chou MT, Chow K-HK, Chu C, Chu DM, Chu V, Chuang K, Chugh AS, Cubberly MR, Daniel MG,

Datta S, Dhaliwal R, Dinh J, Dixit D, Dowling E, Feng M, From CM, Furukawa D, Gaddipati H, Gevorgyan L, Ghaznavi Z, Ghosh T, Gill J, Groves DJ, Gurara KK, Haghighi AR, Havard AL, Heyrani N, Hioe T, Hong K, Houman JJ, Howland M, Hsia EL, Hsueh J, Hu S, Huang AJ, Huynh JC, Huynh J, Iwuchukwu C, Jang MJ, Jiang AA, Kahlon S, Kao P-Y, Kaur M, Keehn MG, Kim EJ, Kim H, Kim MJ, Kim SJ, Kitich A, Kornberg RA, Kouzelos NG, Kuon J, Lau B, Lau RK, Law R, Le HD, Le R, Lee C, Lee C, Lee GE, Lee K, Lee MJ, Lee RV, Lee SHK, Lee SK, Lee S-LD, Lee YJ, Leong MJ, Li DM, Li H, Liang X, Lin E, Lin MM, Lin P, Lin T, Lu S, Luong SS, Ma JS, Ma L, Maghen JN, Mallam S, Mann S, Melehani JH, Miller RC, Mittal N, Moazez CM, Moon S, Moridzadeh R, Ngo K, Nguyen HH, Nguyen K, Nguyen TH, Nieh AW, Niu I, Oh S-K, Ong JR, Oyama RK, Park J, Park YA, Passmore KA, Patel A, Patel AA, Patel D, Patel T, Peterson KE, Pham AH, Pham SV, Phuphanich ME, Poria ND, Pourzia A, Ragland V, Ranat RD, Rice CM, Roh D, Rojhani S, Sadri L, Saguros A, Saifee Z, Sandhu M, Scruggs B, Scully LM, Shih V, Shin BA, Sholklapper T, Singh H, Singh S, Snyder SL, Sobotka KF, Song SH, Sukumar S, Sullivan HC, Sy M, Tan H, Taylor SK, Thaker SK, Thakore T, Tong GE, Tran JN, Tran J, Tran TD, Tran V, Trang CL, Trinh HG, Trinh P, Tseng H-CH, Uotani TT, Uraizee AV, Vu KKT, Vu KKT, Wadhwani K, Walia PK, Wang RS, Wang S, Wang SJ, Wiredja DD, Wong AL, Wu D, Xue X, Yanez G, Yang Y-H, Ye Z, Yee VW, Yeh C, Zhao Y, Zheng X, Ziegenbalg A, Alkali J, Azizkhanian I, Bhakta A, Berry L, Castillo R, Darwish S, Dickinson H, Dutta R, Ghosh RK, Guerin R, Hofman J, Iwamoto G, Kang S, Kim A, Kim B, Kim H, Kim K, Kim S, Ko J, Koenig M, LaRiviere A, Lee C, Lee J, Lung B, Mittelman M, Murata M, Park Y, Rothberg D, Sprung-Keyser B, Thaker K, Yip V, Picard P, Diep F, Villarasa N, Hartenstein V, Shapiro C, Levis-Fitzgerald M, Jaworski L, Loppato D, Clark

IE, Banerjee U. 2019. Expression-Based Cell Lineage Analysis in Drosophila Through a Course-Based Research Experience for Early Undergraduates. G3 GenesGenomesGenetics 9:3791–3800.

7. Delventhal R, Steinhauer J. 2020. A course-based undergraduate research experience examining neurodegeneration in Drosophila melanogaster teaches students to think, communicate, and perform like scientists. PLOS ONE 15:e0230912.

8. Genné-Bacon EA, Wilks J, Bascom-Slack C. 2020. Uncovering Factors Influencing Instructors' Decision Process when Considering Implementation of a Course-Based Research Experience. CBE—Life Sci Educ 19:ar13.

9. Fernandes PA, Passos Ó, Ramos MJ. 2022. Necessity is the Mother of Invention: A Remote Molecular Bioinformatics Practical Course in the COVID-19 Era. J Chem Educ 99:2147–2153.

10. Anderson N, Wilch M. 2021. Online Instruction – Bioinformatics Lesson for a COVID-19 Vaccine. Am Biol Teach 83:464–471.

11. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A. 2005. Galaxy: A platform for interactive large-scale genome analysis. Genome Res 15:1451–1455.

12. Kirkpatrick C, Schuchardt A, Baltz D, Cotner S. 2019. Computer-Based and Bench-Based Undergraduate Research Experiences Produce Similar Attitudinal Outcomes. CBE—Life Sci Educ 18:ar10.

13. Turner AN, Challa AK, Cooper KM. 2021. Student Perceptions of Authoring a Publication Stemming from a Course-Based Undergraduate Research Experience (CURE). CBE—Life Sci Educ 20:ar46.

14. Hodkinson LJ, Smith C, Comstra HS, Albanese EH, Ajani BA, Arsalan K, Daisson AP, Forrest KB, Fox EH, Guerette MR, Khan S, Koenig MP, Lam S, Lewandowski AS, Mahoney LJ, Manai N, Miglay J, Miller BA, Milloway O, Ngo VD, Oey NF, Punjani TA, SiMa H, Zeng H, Schmidt CA, Rieder LE. 2023. A bioinformatics screen reveals Hox and chromatin remodeling factors at the Drosophila histone locus. bioRxiv https://doi.org/10.1101/2023.01.06.523008.

15. Duronio RJ, Marzluff WF. 2017. Coordinating cell cycle-regulated histone gene expression through assembly and function of the Histone Locus Body. RNA Biol 14:726–738.

16. Xie M, Comstra S, Schmidt C, Hodkinson L, Rieder LE. 2022. Max is likely not at the Drosophila histone locus. bioRxiv https://doi.org/10.1101/2022.09.11.507040.

17. Rieder LE, Koreski KP, Boltz KA, Kuzu G, Urban JA, Bowman SK, Zeidman A, Jordan WT, Tolstorukov MY, Marzluff WF, Duronio RJ, Larschan EN. 2017. Histone locus regulation by the Drosophila dosage compensation adaptor protein CLAMP. Genes Dev 31:1494–1508.

18. Deutch CE. 2018. Mendel or Molecules First: What is the Best Approach for Teaching General Genetics? Am Biol Teach 80:264–269.

19. Jenkins VK, Larkin A, Thurmond J. 2022. Using FlyBase: A Database of Drosophila Genes and Genetics, p. 1–34. *In* Dahmann, C (ed.), Drosophila: Methods and Protocols. Springer US, New York, NY.

20. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT, Legeay M, Fang T, Bork P, Jensen LJ, von Mering C. 2021. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res 49:D605–D612.

21. Horard B, Loppin B. 2015. Histone storage and deposition in the early Drosophila embryo. Chromosoma 124:163–175.

22. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. Nat Biotechnol 29:24–26.

23. The modENCODE Consortium, Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, Lin MF, Washietl S, Arshinoff BI, Ay F, Meyer PE, Robine N, Washington NL, Stefano LD, Berezikov E, Brown CD, Candeias R, Carlson JW, Carr A, Jungreis I, Marbach D, Sealfon R, Tolstorukov MY, Will S, Alekseyenko AA, Artieri C, Booth BW, Brooks AN, Dai Q, Davis CA, Duff MO, Feng X, Gorchakov AA, Gu T, Henikoff JG, Kapranov P, Li R, MacAlpine HK, Malone J, Minoda A, Nordman J, Okamura K, Perry M, Powell SK, Riddle NC, Sakai A, Samsonova A, Sandler JE, Schwartz YB, Sher N, Spokony R, Sturgill D, Baren M van, Wan KH, Yang L, Yu C, Feingold E, Good P, Guyer M, Lowdon R, Ahmad K, Andrews J, Berger B, Brenner SE, Brent MR, Cherbas L, Elgin SCR, Gingeras TR, Grossman R, Hoskins RA, Kaufman TC, Kent W, Kuroda MI, Orr-Weaver T, Perrimon N, Pirrotta V, Posakony JW, Ren B,

Russell S, Cherbas P, Graveley BR, Lewis S, Micklem G, Oliver B, Park PJ, Celniker SE, Henikoff S, Karpen GH, Lai EC, MacAlpine DM, Stein LD, White KP, Kellis M. 2010. Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE. Science 330:1787–1797.

24. He Q, Johnston J, Zeitlinger J. 2015. ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. Nat Biotechnol 33:395–401.

25. Skene PJ, Henikoff S. 2017. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. eLife 6:e21856.

26. Kaya-Okur HS, Wu SJ, Codomo CA, Pledger ES, Bryson TD, Henikoff JG, Ahmad K, Henikoff S. 2019. CUT&Tag for efficient epigenomic profiling of small samples and single cells. 1. Nat Commun 10:1930.

27. Round JE, Campbell AM. 2013. Figure Facts: Encouraging Undergraduates to Take a Data-Centered Approach to Reading Primary Literature. CBE—Life Sci Educ 12:39–46.

28. Gaskill MM, Gibson TJ, Larson ED, Harrison MM. 2021. GAF is essential for zygotic genome activation and chromatin accessibility in the early Drosophila embryo. eLife 10:e66668.

29. 1956. Taxonomy of Educational Objectives Handbook I: The Cognitive Domain. Longman, New York.

30. McKay DJ, Klusza S, Penke TJR, Meers MP, Curry KP, McDaniel SL, Malek PY, Cooper SW, Tatomer DC, Lieb JD, Strahl BD, Duronio RJ, Matera AG. 2015. Interrogating the Function of Metazoan Histones using Engineered Gene Clusters. Dev Cell 32:373–386.

31. Bongartz P, Schloissnig S. 2019. Deep repeat resolution—the assembly of the Drosophila Histone Complex. Nucleic Acids Res 47:e18–e18.

32. Filion GJ, Bemmel JG van, Braunschweig U, Talhout W, Kind J, Ward LD, Brugman W, Castro IJ de, Kerkhoven RM, Bussemaker HJ, van Steensel B. 2010. Systematic Protein Location Mapping Reveals Five Principal Chromatin Types in Drosophila Cells. Cell 143:212–224.

Chapter 7

Discussion

7.1 Overview: Coordinated gene regulation is a heavy burden

Coordinated gene regulation is a heavy burden for the nucleus. Coordination serves many purposes, including allowing groups of related or even unrelated genes to be expressed together in order to meet strict spatial or temporal regulatory needs (Michalak 2008; Nair *et al.* 2022)(Michalak 2008; Nair *et al.* 2022). Because coordinated gene expression can involve co-transcription at specific times or in specific places, the combinations of transcription factors that are responsible for regulating these genes also require a high level of orchestration to target genes at the correct times. A current knowledge gap is what cues factors integrate to determine where, when, and how to regulate their gene targets. Investigating coordinated gene expression by transcription factors presents an experimental challenge because there can be tens or hundreds of genes and factors involved in these large-scale processes. As a model, I focused on the *Drosophila melanogaster* histone locus for interrogating both coordinated gene regulation and transcription factor function.

7.2 The *Drosophila melanogaster* histone locus is inherently absurd

The *D. melanogaster* histone locus exists as a comically perfect paradox for understanding coordinated gene regulation. On one hand, it is a hallmark example of coordinated regulation in both gene arrangement and temporal coupling: the histone locus comprises highly regular tandemly repeated gene arrays at a single locus (diagramed in **Figure 7.1**) and the histone genes have strict expression requirements coupled to the cell cycle (Duronio and Marzluff 2017; Bongartz and Schloissnig 2018). On the other hand, because of these qualities, the *D. melanogaster* histone locus is unlike almost any other repetitive or coordinated gene family. Although there are other examples of genes that are repetitive and that are clustered together in arrays for regulatory purposes such as the rDNA loci or the Hox genes, neither of these mentioned examples are quite as extreme in their arrangement as the *D. melanogaster* histone locus. In flies, there two rDNA loci which are organized into repetitive units but, these units are somewhat less regularly organized than the histone array and are also present on both the X and Y chromosomes which presents some different organization based on sex (Tartof and Dawid 1976). The Hox genes are clustered together like the histone genes however the arrays function to allow for specific expression at the right time and in the right areas of the developing embryo rather than to be expressed together (Pearson *et al.* 2005). Furthermore, the histone genes themselves are clustered across many different organisms however none of them share the characteristic of having a single histone locus like *D. melanogaster*. As the famous Dr. Leila Rieder often says: "if someone were to ask me how I would construct the perfect genomic locus, I would want it to be organized like the *D. melanogaster* histone locus." Needless to say, the *D. melanogaster* histone locus is an extreme example of clustered, coordinated genes.

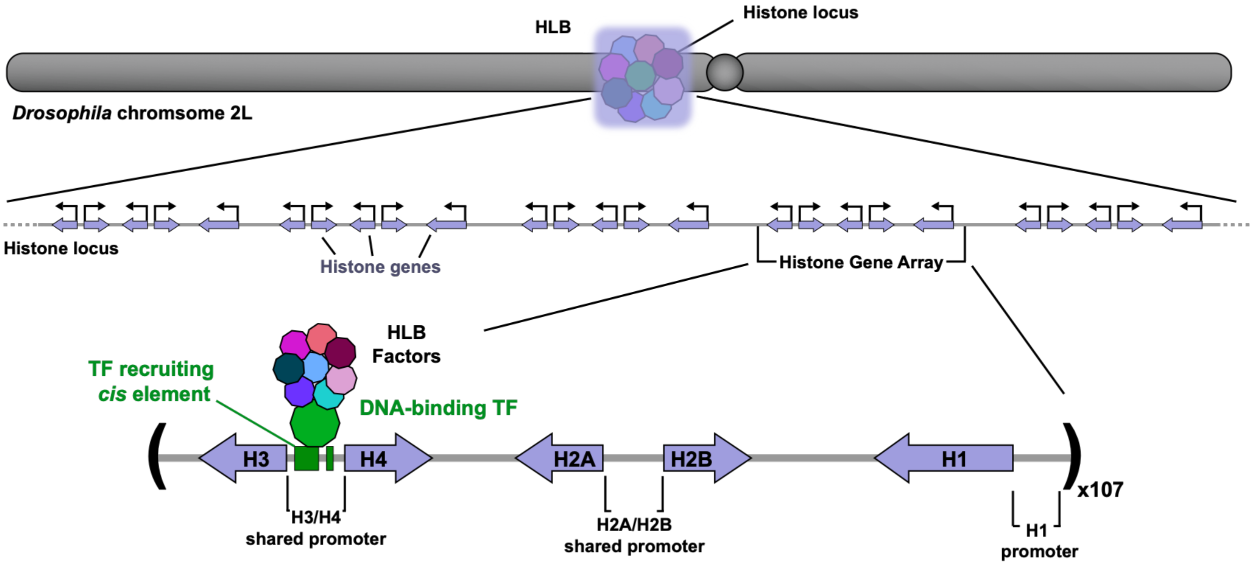


Figure 7.1 A diagram of the *D. melanogaster* histone locus and HLB. The histone locus body (HLB) forms as a concentration of phase separated factors around the histone locus. The histone locus is located on chromosome 2L near the centromere and contain ~107 highly organized tandemly repeated gene arrays. Each array is ~5 kb and contains all 5 canonical histone genes (bottom) along with their respective promoters and regulatory elements.

We have the luxury of working in the well-developed *D. melanogaster* model system in which we can leverage powerful transgenic models and genetic tools to understand the nuanced regulation of histone gene expression (McKay *et al.* 2015). In Chapter 2, I discovered that the only variable sequence within the 107 histone gene arrays is the CLAMP recruiting GA-repeat however that this variability does not suggest that the arrays are differentially regulated due to the impact on CLAMP binding. In **Chapter 3**, I leveraged the single histone gene array transgene in which I induced subtle DNA manipulations to explore how transcription factor function at the histone locus can be impacted by *cis* element sequence and identity. I discovered that exchanging the CLAMP recruiting GA-repeat in the histone gene array for another CLAMP recruiting *cis* element changes its function meaning that CLAMP *cis* elements with similar sequence are not simply interchangeable. Furthermore, we discovered the GA-repeat must reside in the bidirectional *H3/H4* promoter suggesting that CLAMP is using a combination of cues from *cis* element sequence and flanking sequence, possibly cofactors that bind those sequences, to determine function at the histone locus.

Although our results give insight into the cues required by transcription factors or their tolerance for sequence changes, our specific discoveries may be unique to the coordination of histone gene expression. The long, perfect GA-repeat that is imperative to CLAMP factor recruitment in *D. melanogaster* (Rieder *et al.* 2017)), is distinctly different from that of *Drosophila virilis*, which I explore in **Chapter 4**. This observation sparks many additional questions and suggests that transcription factors that target the *D. virilis* histone loci use different cues to achieve their histone locus functions, or alternatively, the regulatory mechanism for *D. virilis* histone genes is completely unique from that of *D. melanogaster.*

While there are likely factors that serve different regulatory functions in histone biogenesis across organisms, there remain some regulatory mechanisms that are conserved to facilitate histone gene expression. Mxc, or the orthologous NPAT, is completely unique to the HLB meaning it has only ever been shown to associate with the histone genes and conserved across *Drosophila* species and humans (Gao *et al.* 2003; Duronio and Marzluff 2017). Even more broadly speaking, although there is no human ortholog to CLAMP, there is likely a factor or set of factors that assists in the targeting the histone gene arrays and recruitment of these conserved, critical factors like Mxc/NPAT to the histone locus. Our observations about the general cues that impact transcription factor recruitment and function at the histone locus can be applied to these systems despite how specific CLAMP's function may be *D. melanogaster* histone biogenesis.

7.3 The impact of *cis* element length on transcription factor recruitment

The clustering of *Drosophila* histone genes at one (*D. melanogaster*) or a few (*D. virilis*) loci raises additional questions about how the genes might be differentially regulated, especially as they experience similar transcription factor nuclear microenvironments. In the genomes of other animals like humans (Marzluff *et al.* 2002) and sea urchins (Marzluff *et al.* 2006), histone genes are clustered, sometimes only loosely, in several sets and these sets are differentially regulated to ensure the required concentrations of the histones are produced in different ratios or at different developmental time points.

In **Chapter 2**, we therefore sought to understand how the histone gene arrays in *D. melanogaster* might be differentially regulated. The 107 histone gene arrays present in the *D. melanogaster* genome are virtually identical in sequence other than the GA-repeat *cis* element within the *H3/H4* promoter, which we discovered varies in length. We also showed that this GA-repeat is targeted by CLAMP (**Chapter 3**, (Rieder *et al.* 2017)). CLAMP contains six DNA-interacting zinc-fingers and we do not yet understand how CLAMP physically interacts with its GA-repeat targets. When tested by *in vitro* binding assays, a CLAMP DNA interaction domain with four zinc fingers bound similar 15 bp long GA-rich DNA targets when compared to six zinc-fingers but did have weaker affinity for these targets (Kuzu *et al.* 2016). It is currently unknown if or when CLAMP uses all six zinc fingers to bind DNA targets or if the length of the GA-repeat can impact this interaction. Furthermore, the length of the GA-repeat does not seem to influence CLAMP binding at different histone gene arrays, although the GA-repeat itself is important for HLB factor recruitment (**Chapter 2, 3**). We also investigated other GA-repeat binding factors, GAF and Psq, and discovered that they both target the embryonic histone locus by ChIP-seq (**Chapter 2**). It is possible that CLAMP competes for binding at the histone locus with these two other GA-repeat binding factors. However, previous work indicated that GAF is not present at the histone locus in embryos nor in polytenes by immunofluorescence unless CLAMP has been depleted by RNAi (Rieder *et al.* 2017), implying that GAF may be a "backup" transcription factor that can bind the free GA repeats at the histone locus when it is not being out competed by CLAMP. This is true of the cites on the X chromosome as well; when CLAMP is depleted, GAF will localize to GA-rich motifs it normally does not occupy on the X chromosome

(Kaye *et al.* 2018). This may be due to the fact that GA-repeats that are not occupied by another binding factor are targeted by GAF opportunistically rather than for a true functional “backup” mechanism. Still, future studies can explore whether a low level of GAF undetectable by normally immunofluorescent assays targets a subset of histone arrays. We could construct a barcoded 12-array transgene where we could differentiate the expression of the histone genes from each array while also genetically manipulating the length of the GA-repeat in the *H3/H4* promoter to assess whether GAF targets and is functionally active at histone gene arrays based on GA-repeat length or CLAMP targeting.

More than one CLAMP protein can bind a GA-repeat if it contains enough “GA” nucleotides to stoichiometrically accommodate more than just a single CLAMP protein (Kuzu *et al.* 2016; Kaye *et al.* 2018). Bioinformatic and immunofluorescence assays are unable to resolve this detail therefore we do not yet know if more than one CLAMP molecule is interacting with the GA-repeat in the histone gene array or how this may impact histone gene expression within each array. The repetitiveness of the histone locus also creates a barrier to resolving this question. The variable length of the GA-repeat likely influences the number of transcription factors that can simultaneously bind this region, which also may impact the regulation of the histone genes and lead to differential expression. Future experiment could include completion electrophoretic mobility shift assays with histone promoter probes and recombinant proteins that associate with these regions. We expect CLAMP to be able to bind all GA-repeat lengths within the *H3/H4*p in these assays and then could utilize antibodies to identify of additional factors such as GAF or other candidates can bind in combination with CLAMP to histone locus-like DNA probes.

7.4 A combination of cues can influence transcription factor function

We now know that the length of the GA-repeat does not seem to impact CLAMP targeting of histone gene arrays; however, CLAMP has several functions across the genome. CLAMP is a context dependent transcription factor that not only functions as a critical factor at the histone locus but also functions in dosage compensation. CLAMP targets GA-rich *cis* elements across the genome; on the X chromosome it binds GA-rich regions called MREs and recruits dosage compensation factors. However, at the histone locus on chromosome 2L CLAMP targets a long, strict GA-repeat in the *H3/H4* promoter and recruits HLB specific factors. CLAMP, therefore, can bind similar *cis* elements on the X chromosome versus the autosomes and retain the ability to function uniquely in dosage compensation and histone biogenesis respectively. How transcription factors integrate cues at different loci and retain unique functions is currently a large knowledge gap.

To interrogate the cues that impact transcription factor function, we leveraged a transgenic histone gene array system with which we could manipulate the CLAMP binding *cis* elements. We found that X chromosome GA-rich MREs do not functionally substitute for the GA-repeat in the histone gene array and therefore changes in *cis* element sequence and local structure are enough to impact TF function modeled in **Figure 7.2** (Hodkinson *et al.* 2023b). Additionally, we found that the GA-repeat must reside within the *H3/H4*p for proper CLAMP function and HLB factor recruitment. These data suggest that there may be additional *cis* elements and, therefore, additional TFs, that aid in providing contextual cues for CLAMP to function. Zelda is a strong candidate for this early acting transcription factor that might function in combination or separate from CLAMP to regulate histone gene expression based on an already established relationship with CLAMP and Zelda (Duan *et al.* 2021). We found that there are

TAGteam motifs within the histone gene array and that Zelda targets the histone locus early in development by immunofluorescence assays. Despite these data, based on our current observations, Zelda is dispensable at the histone locus and depleting Zelda does not impact histone mRNA levels or HLB factor recruitment (O'Haren *et al.* 2023) implying Zelda may not be this secondary mechanism of regulation that works in tandem with CLAMP. This leaves a large open question about what other transcription factors could be good candidates for understanding the additional factors that aid in CLAMP's targeting and regulation of histone gene expression. In **Chapter 5**, I discuss how we have started screening for additional DNA-binding factors that may fill this roll and have several HLB-member candidates to explore in the future.

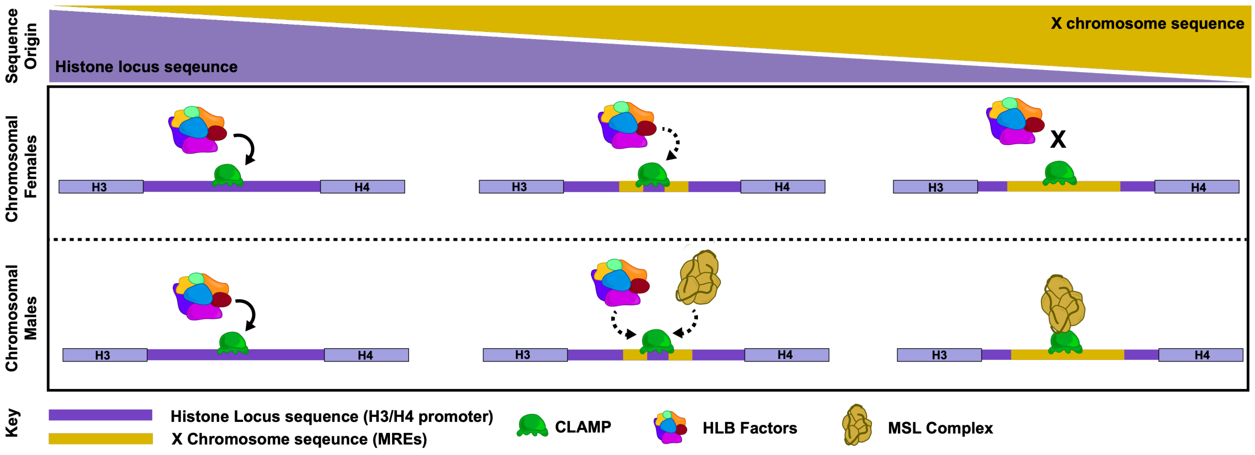


Figure 7.2 Model depicting how CLAMP incorporates information from *cis* element sequence and flanking DNA sequence to determine function at the histone array. CLAMP modulates its function as more X chromosome sequence (yellow, increases left to right) is introduced to the H3/H4 promoter (Purple, deceases left to right). In chromosomal females (top) CLAMP does not recruit histone locus body factors to the histone array as more X chromosome sequence is

introduced. In chromosomal males, CLAMP recruits X chromosome factors to the histone array (autosomal) as more X chromosome sequence is introduced.

Considering our current model where CLAMP function is influenced by the H3/H4 promoter DNA sequence, a recent study presents a conundrum. A 12-array transgene in which the H3/H4 promoters are replaced with the H2A/H2B promoter, which does not contain any CLAMP binding GA-repeats, is able to support histone gene expression and viability in an endogenous histone locus deletion background (Koreski *et al.* 2020). CLAMP is present at this transgene by immunofluorescence assays, but does not interact with any particular DNA sequence by ChIP-seq (Koreski *et al.* 2020). These data imply that, in dire situations where the only source for histones is from the genes within the 12-array transgene, there is an additional mechanism or cofactor that assists CLAMP in targeting the histone genes. Although these data are still a mystery given that we found CLAMP gleans information from *cis* element sequence and the flanking sequence of the *H3/H4* promoter itself (**Chapter 3**), these data support the idea that there are important cofactors that can heavily influence CLAMP function, even helping chaperone it to a locus where it has no *cis* binding targets. These data also prompt a question of additional CLAMP functions at the histone locus suggesting that CLAMP may be participating in protein-protein interactions that are additional to its function of binding the *H3/H4* promoter, opening chromatin as a pioneer factor, and seeding the HLB for other factors to then be recruited (Duan *et al.* 2021). Future studies should focus on trying to understand CLAMP and its potential for protein-protein interactions by looking at *in vitro* protein binding or co-immunoprecipitation assays.

Additionally, based on our observations the GA-repeat and other possible *cis* element information in the *H3/H4* promoter is important, but there is likely a “backup” mechanism of HLB formation, related to factors that can target the *H2A/H2B* promoter, to ensure histone gene expression is maintained. To date, CLAMP is the only known DNA binding factor with direct evidence of physically binding DNA at the histone locus in *Drosophila* and have yet to confirm factors that bind to the *H2A/H2B* promoter. There may be one or several other DNA-binding factors that target the histone gene arrays that impact CLAMP function or provide necessary contextual cues for CLAMP to distinguish its function at the histone locus from its other functions.

7.5 Regulatory mechanisms of histone gene expression vary between species

The histone locus in *Drosophila melanogaster* is a unique system for studying histone gene regulation because all the replication dependent histone genes are clustered at a single locus. Even within the genus *Drosophila*, there is substantial variability in the number of histone genes and their genomic organization, further complicating our understanding of histone gene regulation. In **Chapter 4**, we sought to explore the sequence differences and regulatory mechanisms of other *Drosophila* species to expand our comprehensive understanding of histone gene expression.

D. virilis, a species ~40 MYa diverged from *D. melanogaster,* has two histone loci each with somewhat unique characteristics and almost better represent histone gene organization since it is closer to the way the human histone genes are organized. As discussed in Chapter 4, the major locus located on chromosome 2 is composed of 27 quartet histone gene arrays that include *H3, H4, H2A,* and *H2B* and 5 quintet arrays that also include *H1* however this locus is not regularly spaced and is interrupted by spacing and other gene fragments based on our annotation. The minor locus of chromosome 4, however, only includes 5 regularly spaced quintet arrays. We sought to understand how these different loci may be differentially regulated and focused on attempting to identify factors that may influence differences in expression between the two loci. Similar to *D. melanogaster,* there exists a GA-repeat in the *D. virilis H3/H4* promoter that is targeted by CLAMP, however the *virilis* GA-repeat is considerably shorter and resembles the MRE GA-rich motifs CLAMP targets on the X chromosome for male dosage compensation. Interestingly, CLAMP recruits MSL2, a member of the MSLc, is recruited to the major histone locus in *D. virilis* but not to the minor locus or any loci in other *Drosophila* species we examined (Xie *et al.* 2022).

Our observations imply not only that there may be distinct regulatory differences between the two histone loci in *D. virilis* but that there are also significant differences between the regulation of histone genes among *Drosophila* species. Furthermore, this introduces the possibility that *D. virilis* MSL2 serves a function outside of male dosage compensation, evolving additional or altogether different functions from *D. melanogaster* MSLc members. Future work should focus on exploring whether CLAMP is recruiting MSL2 to the *D. virilis* major locus which can be assessed by disrupting the known MSL2 interaction domain of CLAMP to see the impact on MSL2 at the histone locus. Additionally, we could assess MSL2 functionality at the histone locus related to histone gene regulation by looking at expression of the histone genes by qPCR when MSL2 is not at the locus.

I leveraged the *D. melanogaster* system because of the extensive genetic tool repertoire we have at our disposal to answer these nuanced questions about histone gene regulation. Based on the differences we already know in other *Drosophila* species such as multiple histone loci and different factor recruitment, to gain a comprehensive understanding of histone gene regulation we need to focus on looking in other *Drosophila* species. We already have some evidence that implies histone gene regulation and HLB formation use different mechanisms when there are multiple histone loci. A recent model suggests that recruitment of HLB factors and competition for those facts between loci is a mechanism for coordinated, differential regulation of the histone genes from different clusters to maintain the correct concentrations of histone transcripts, and subsequent proteins (Chaubal *et al.* 2023). The fact that many other *Drosophila* species do not have completely published genomes coupled with the fact that the genetic tools that are utilized in *D. melanogaster* may not translate in other systems, we are faced with a large challenge in studying other *Drosophila* species. Thinking more broadly, future studies should involve characterizing sequence differences of other *Drosophila* species, such as *D. simulins* and *D. yakuba.* We recently conducted a preliminary annotation of the histone gene arrays in *D. simulins* and *D. yakuba* and found that not only the number of histone loci differed, but regularity of the arrays at those loci are distinct from *D. melanogaster* (Sisi Falcone and Annalise Weber, unpublished data). Additionally, there are differences in the *H3/H4* GA-repeat which we know is critical for CLAMP binding and seeding the HLB in *D. melanogaster.* In future experiments, we could leverage the barcoded 12-array transgene system to differentiate histone gene expression from each gene array and probe how these different *cis* elements or array structures influence differential regulation of the histone genes. We could also utilize this barcoded 12-array transgene system to create animals with histone arrays from both *D. simulins* and *D. yakuba* probing how differences in array structure can influence HLB factor recruitment, or even competition of recruitment, and histone gene expression.

7.6 The HLB is a wild unknown entity

Our observations suggest that CLAMP is not the only DNA binding factor that might target the histone locus and, furthermore, that there may be additional cofactors that impact histone gene regulation. We therefore performed a bioinformatics screen to identify novel HLB candidates. The Hox factors Ubx, Abd-A, and Abd-B emerged as strong candidates targeting the *H3/H4* promoter (Hodkinson *et al.* 2023a). Hox factors are expressed early during embryogenesis and are responsible for body patterning based on where in the embryo they are spatially expressed (Pearson *et al.* 2005). Since histone gene expression is both non-cell specific and critical in the dividing syncytial embryo, we might expect other transcription factors acting early in the embryo, like Hox factors, to also regulate histone gene expression.

Our bioinformatics screen likely masks nuanced information about the factor under investigation. For example, a Hox factor might only target a subset of the 107 histone gene arrays, or that the number and identity of arrays targeted might change through embryogenesis. There are several possibilities of how proteins like the Hox factors target the locus that we are unable to resolve by aligning to the custom single histone gene array genome, modeled in **Figure 7.3.** Besides the Hox factors, we identified several other DNA-binding candidates for future wet lab studies. These experiments could include performing immunofluorescence staining for these Hox factors in the early embryo where they may be associating with the histone locus and in different cycling tissues where Hox factors are active.

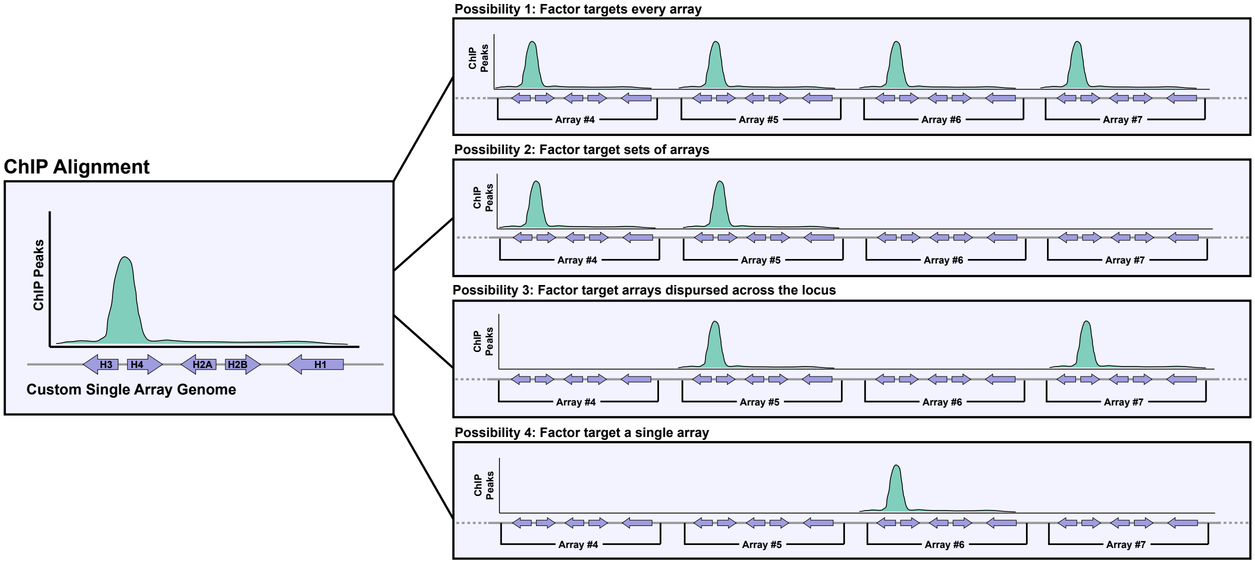

ChIP Alignment
ChIP Peaks
H3
H4
H2A
H2B
H1
Custom Single Array Genome
Possibility 1: Factor targets every array
ChIP Peaks
Array #4
Array #5
Array #6
Array #7
Possibility 2: Factor target sets of arrays
ChIP Peaks
Array #4
Array #5
Array #6
Array #7
Possibility 3: Factor target arrays dispursed across the locus
ChIP Peaks
Array #4
Array #5
Array #6
Array #7
Possibility 4: Factor target a single array
ChIP Peaks
Array #4
Array #5
Array #6
Array #7

Figure 7.3 Challenges with aligning ChIP data to the single histone gene array. We cannot resolve detailed information about transcription factor targeting at the histone locus using the custom single histone array genome in ChIP alignment. ChIP-sequencing alignment for a given transcription factor to the single histone array might show a peak at the H3/H4 promoter (left panel) however, there are several possibilities for how that factor targets the arrays *in vivo.* Out of the ~107 histone gene arrays that comprise the histone locus, a given transcription factor may target all arrays (possibility #1), only a set of neighboring arrays (possibility #2), a series of arrays distributed across the locus (possibility #3), or as few as one histone gene array (possibility #4).

Identifying other DNA-binding factors that target the histone gene arrays is imperative to understanding the regulation of the histone locus. It may be that CLAMP is the primary DNA-binding factor that orchestrates the initial regulation of the histone locus early in the embryo; maternally deposited CLAMP is important for proper zygotic histone gene regulation. However, it is likely that other factors target the locus at different times during embryogenesis or related to the cell cycle. Furthermore, based on our observations from **Chapter 2**, GA-repeat length does not impact CLAMP targeting, and therefore CLAMP may not contribute to the differential regulation of the histone gene arrays. It is possible that some of our candidate DNA-binding factors identified in this screen are influencing that differential regulation, providing another layer of coordination for the expression of the histone genes in *D. melanogaster.*

7.7 Histone gene regulation is way more complicated than you think

Histone gene expression and regulation in *D. melanogaster* is far more complicated than just considering how transcription factors are functioning to express a gene family. The three-dimensional and spatial qualities of the histone locus and the body of factors that regulate the histone genes are complex and, currently, less defined. The histone locus body itself is a phase separated nuclear body; it is a membraneless concentration of factors that is phase separated from the rest of the nuclear material and because there is no barrier, components can be readily exchanged (Mitrea and Kriwacki 2016; Tatomer *et al.* 2016; Duronio and Marzluff 2017). Because there are over 100 histone gene arrays that may have different TFs interacting with different regions of the histone locus at any given time, the histone locus body may act like a sink for transcription factors, theoretically pulling all the necessary regulatory factors to the histone locus. This characteristic could potentially explain why GAF is recruited to the histone locus in the absence of CLAMP (Rieder *et al.* 2017) since the histone locus would contain unbound GA-repeats and the phase separated HLB could pull GAF into the concentration of factors already present at the histone genes.

Mxc is an HLB scaffolding protein that targets the histone locus in the early embryo prior to zygotic genome activation. Mxc has a self-interaction domain that facilitates oligomerization, and the C terminus directly interacts with the C terminus of FLASH, another key component of the HLB (Kemp *et al.* 2021). Together, they form a core-shell configuration. Besides these data, we know about distinct protein-protein interactions between the factors that process histone mRNAs which have unique needs since they lack both introns and polyA tails, and instead have a 3' end hairpin structure that needs to be bound and cleaved (Marzluff *et al.* 2008; Tatomer *et al.* 2016). FLASH is also critical for histone mRNA processing factor and FLASH can be mutated in a way that prevents it from being recruited to the phase separated HLB but does not change its protein levels (Tatomer *et al.* 2016). By simply disrupting the ability for FLASH to localize, histone biogenesis was completely disrupted. These data provide evidence that nuclear bodies concentrate factors to make regulatory events more efficient and in turn, make an unusual environment to make unique processes efficient.

Discounting mRNA processing, there is little known about other protein-protein HLB interactions that would give more insight into HLB composition as well as overall histone gene transcription. Furthermore, because of the organelle-like nature of the phase separated HLB, it may be that cofactors do not always need to directly bind to one another but rather can simply be in close proximity to influence histone gene regulation.

Beyond just the less defined protein-protein interactions in the HLB, the three-dimensional interactions of the DNA within the histone locus (*D. melanogaster)* and between the histone loci (*D. virilis* and other organisms) remains an open-ended area of study in histone gene regulation. The major histone cluster in humans spans over several megabases and has multiple subclusters where the histone genes are concentrated but not arrayed like in *Drosophila*.

According to Hi-C, these subclusters physically interact and create hubs for histone gene regulatory factors to congregate and control expression of many histone genes at the same time (Carty *et al.* 2017; Ghule *et al.* 2023). We currently do not know if there are inter array or intra array interactions occurring between within the *D. melanogaster* histone locus or how this might impact histone gene regulation. There is some preliminary evidence of bother intra and inter array interactions of the *D. melanogaster* histone locus. Driving CLAMP to the *H3/H4* promoter promotes accessibility and expression across the entire histone gene array, not just at the H3 and H4 genes (Rieder *et al.* 2017). A recent paper also shows preliminary findings where, in a 12-array transgene where all the *H3/H4* promoters have been replaced with the *H2A/H2B* promoter except for one, that "active" array containing the *H3/H4* promoter can induce the expression of the adjacent histone gene arrays that would otherwise be "inactive" or have no expression (Koreski *et al.* 2020). These data suggest that expression of genes within individual arrays can influence the expression of the other histone gene within that array as well as genes within flanking arrays.

I would argue the three-dimensional organization and interactions of the *D. melanogaster* histone genes is the largest area for future studies in histone gene regulation. Future experiments could include using newer technologies such as DamID which is a system where a DNA adenine methyltransferase can be fused to a given transcription factor and subsequently methylate the DNA around where that factor targets (Aughey *et al.* 2019). This technique could give insight into what arrays or histone genes are actively targeted, and perhaps expressed, within the histone locus and if that targeting impacts the activity of neighboring genes. We would also resolve some of the inter-array interactions based on the methylation patterns. The epigenetic landscape of the histone gene arrays, which could also give insight to both the activity and interactions within the histone locus, is a large gap in our knowledge about histone locus organizations. Here, the repetitiveness of the histone gene arrays is, again, a challenge for sequencing and alignment meaning traditional methods of identifying histone marks such as ChIP-seq or Cut&RUN are unusable. Future experiments using the newly developed long read technology DiMeLo-seq, where a methyltransferase can methylate DNA around target area where a transcription factor targets or histone mark resides based on antibody binding (Altemose *et al.* 2022; Maslan *et al.* 2023), will give insight into overall epigenetic organization and activity of histone gene arrays. Overall, these future experiments are necessary in order to begin closing the large gaps in our understanding of the complicated regulation of the *D. melanogaster* histone genes outlined in **Figure 7.4**.

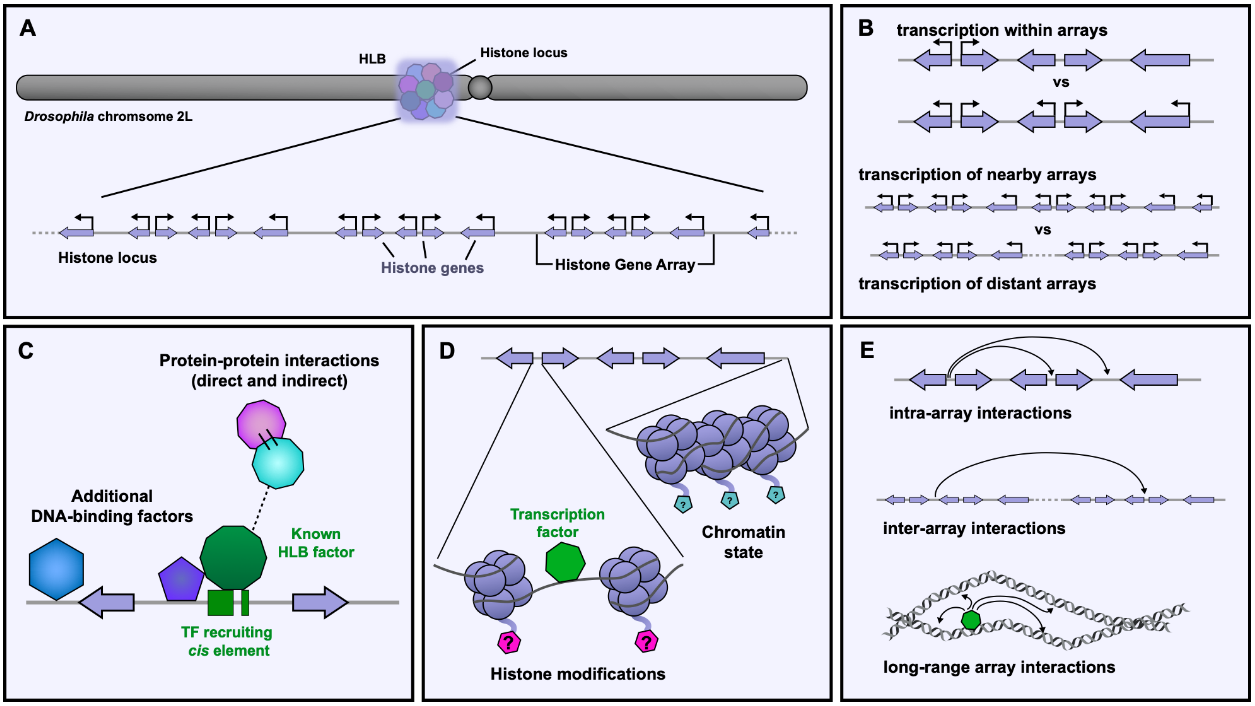


Figure 7.4 Summary of HLB unknowns. (A) A schematic of the different HLB features including the factors that regulate histone gene expression and the organization of the histone gene arrays. **(B)** Despite the organization of the arrays at the single locus, it is currently unknown if histone gene arrays are differentially expressed or how histone gene expression looks across all ~100 arrays. The histone gene array may be an arbitrary unit and expression of the histone genes may be based on individual gene sets rather than entire arrays. **(C)** While some histone locus factors have been confirmed to regulate histone gene expression, there are additional factors that have not been identified and the interactions between these factors are still poorly defined. (**D)** The chromatin landscape, including histone modifications, are virtually undefined across the histone locus due to limitations in sequencing and aligning repetitive DNA. **(E)** Finally, the three-dimensional architecture of the histone locus including potential intra- and/or inter-histone gene array contacts remain undefined.

7.8 The future of studying coordinated histone gene regulation: closing remarks

I began my dissertation by describing how coordinated gene regulation is a heavy burden for the nucleus. Coordinated gene expression involves tight orchestration of time, space, transcription factor function and regulatory mechanisms that need to be in place to achieve true coordination. The histone locus in *D. melanogaster* emerges as a seemingly perfect model to understand coordinated gene regulation; a single locus of ordered, tandemly repeated arrays containing essential genes that need to be expressed at extremely specific concentrations and strict times. The repetitiveness of the histone locus provides significant barriers for future experiments however with the development of new labeling methods and long read sequencing technologies, coupled with the large repertoire of genetic tools we have available in *Drosophila* such as our histone array transgenic systems, we can truly explore the nuances of histone gene regulation. The work in this dissertation provides more insight into the cues and information transcription factors at the histone locus incorporate to regulate and express the histone genes, there are still several open-ended questions that will fuel countless future experiments and lead to a more comprehensive understanding of the coordinated regulation of the histone genes. **Once you've explored the complexities of histone gene regulation, you never really forget it.**

7.9 References

Altemose N., A. Maslan, O. K. Smith, K. Sundararajan, R. R. Brown, *et al.*, 2022 DiMeLo-seq: a long-read, single-molecule method for mapping protein-DNA interactions genome wide. Nat Methods 19: 711–723. https://doi.org/10.1038/s41592-022-01475-6

Aughey G. N., S. W. Cheetham, and T. D. Southall, 2019 DamID as a versatile tool for understanding gene regulation. Development 146: dev173666. https://doi.org/10.1242/dev.173666

Bongartz P., and S. Schloissnig, 2018 Deep repeat resolution—the assembly of the Drosophila Histone Complex. Nucleic Acids Research 47: e18–e18. https://doi.org/10.1093/nar/gky1194

Carty M., L. Zamparo, M. Sahin, A. González, R. Pelossof, *et al.*, 2017 An integrated model for detecting significant chromatin interactions from high-resolution Hi-C data. Nat Commun 8: 15454. https://doi.org/10.1038/ncomms15454

Chaubal A., J. M. Waldern, C. Taylor, A. Laederach, W. F. Marzluff, *et al.*, 2023 Coordinated expression of replication-dependent histone genes from multiple loci promotes histone homeostasis in Drosophila. MBoC 34: ar118. https://doi.org/10.1091/mbc.E22-11-0532

Duan J., L. Rieder, M. M. Colonnetta, A. Huang, M. Mckenney, *et al.*, 2021 CLAMP and Zelda function together to promote Drosophila zygotic genome activation. eLife.

Duronio R. J., and W. F. Marzluff, 2017 Coordinating cell cycle-regulated histone gene expression through assembly and function of the Histone Locus Body. RNA Biol 14: 726–738. https://doi.org/10.1080/15476286.2016.1265198

Gao G., A. P. Bracken, K. Burkard, D. Pasini, M. Classon, *et al.*, 2003 NPAT expression is regulated by E2F and is essential for cell cycle progression. Mol Cell Biol 23: 2821–2833. https://doi.org/10.1128/MCB.23.8.2821-2833.2003

Ghule P. N., J. R. Boyd, F. Kabala, A. J. Fritz, N. A. Bouffard, *et al.*, 2023 Spatiotemporal higher-order chromatin landscape of human histone gene clusters at histone locus bodies during the cell cycle in breast cancer progression. Gene 872: 147441. https://doi.org/10.1016/j.gene.2023.147441

Hodkinson L. J., C. Smith, H. S. Comstra, B. A. Ajani, E. H. Albanese, *et al.*, 2023a A bioinformatics screen reveals hox and chromatin remodeling factors at the Drosophila histone locus. BMC Genom Data 24: 54. https://doi.org/10.1186/s12863-023-01147-0

Hodkinson L. J., J. Gross, C. A. Schmidt, P. P. Diaz-Saldana, T. Aoki, *et al.*, 2023b Sequence reliance of a Drosophila context-dependent transcription factor. 2023.12.07.570650.

Kaye E. G., M. Booker, J. V. Kurland, A. E. Conicella, N. L. Fawzi, *et al.*, 2018 Differential Occupancy of Two GA-Binding Proteins Promotes Targeting of the Drosophila Dosage Compensation Complex to the Male X Chromosome. Cell Rep 22: 3227–3239. https://doi.org/10.1016/j.celrep.2018.02.098

Kemp J. P., X.-C. Yang, Z. Dominski, W. F. Marzluff, and R. J. Duronio, 2021 Superresolution light microscopy of the Drosophila histone locus body reveals a core–shell organization associated with expression of replication–dependent histone genes. MBoC 32: 942–955. https://doi.org/10.1091/mbc.E20-10-0645

Koreski K. P., L. E. Rieder, L. M. McLain, W. F. Marzluff, and R. J. Duronio, 2020 Drosophila Histone Locus Body assembly and function involves multiple interactions. bioRxiv 2020.03.16.994483. https://doi.org/10.1101/2020.03.16.994483

Kuzu G., E. G. Kaye, J. Chery, T. Siggers, L. Yang, *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 Genet 12: e1006120. https://doi.org/10.1371/journal.pgen.1006120

Marzluff W. F., P. Gongidi, K. R. Woods, J. Jin, and L. J. Maltais, 2002 The human and mouse replication-dependent histone genes. Genomics 80: 487–98.

Marzluff W. F., S. Sakallah, and H. Kelkar, 2006 The sea urchin histone gene complement. Developmental Biology 300: 308–320. https://doi.org/10.1016/j.ydbio.2006.08.067

Marzluff W. F., E. J. Wagner, and R. J. Duronio, 2008 Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail. Nat Rev Genet 9: 843–854. https://doi.org/10.1038/nrg2438

Maslan A., N. Altemose, R. Mishra, J. Marcus, L. D. Brennan, *et al.*, 2023 Mapping protein-DNA interactions with DiMeLo-seq. 2022.07.03.498618.

McKay D. J., S. Klusza, T. J. Penke, M. P. Meers, K. P. Curry, *et al.*, 2015 Interrogating the function of metazoan histones using engineered gene clusters. Dev Cell 32: 373–86. https://doi.org/10.1016/j.devcel.2014.12.025

Michalak P., 2008 Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. Genomics 91: 243–248. https://doi.org/10.1016/j.ygeno.2007.11.002

Mitrea D. M., and R. W. Kriwacki, 2016 Phase separation in biology; functional organization of a higher order. Cell Communication and Signaling 14: 1. https://doi.org/10.1186/s12964-015-0125-7

Nair R. R., E. Pataki, and J. E. Gerst, 2022 Transperons: RNA operons as effectors of coordinated gene expression in eukaryotes. Trends in Genetics 38: 1217–1227. https://doi.org/10.1016/j.tig.2022.07.005

O'Haren T., T. Aoki, and L. E. Rieder, 2023 Zelda is dispensable for Drosophila melanogaster histone gene regulation. 2023.12.19.572383.

Pearson J. C., D. Lemons, and W. McGinnis, 2005 Modulating Hox gene functions during animal body patterning. Nat Rev Genet 6: 893–904. https://doi.org/10.1038/nrg1726

Rieder L. E., K. P. Koreski, K. A. Boltz, G. Kuzu, J. A. Urban, *et al.*, 2017 Histone locus regulation by the Drosophila dosage compensation adaptor protein CLAMP. Genes Dev 31: 1494–1508. https://doi.org/10.1101/gad.300855.117

Tartof K. D., and I. G. Dawid, 1976 Similarities and differences in the structure of X and Y chromosome rRNA genes of Drosophila. Nature 263: 27–30. https://doi.org/10.1038/263027a0

Tatomer D. C., E. Terzo, K. P. Curry, H. Salzler, I. Sabath, *et al.*, 2016 Concentrating pre-mRNA processing factors in the histone locus body facilitates efficient histone mRNA biogenesis. Journal of Cell Biology 213: 557–570. https://doi.org/10.1083/jcb.201504043

Xie M., L. J. Hodkinson, H. S. Comstra, P. P. Diaz-Saldana, H. E. Gilbonio, *et al.*, 2022 MSL2 targets histone genes in Drosophila virilis. 2022.12.14.520423.

Appendix A

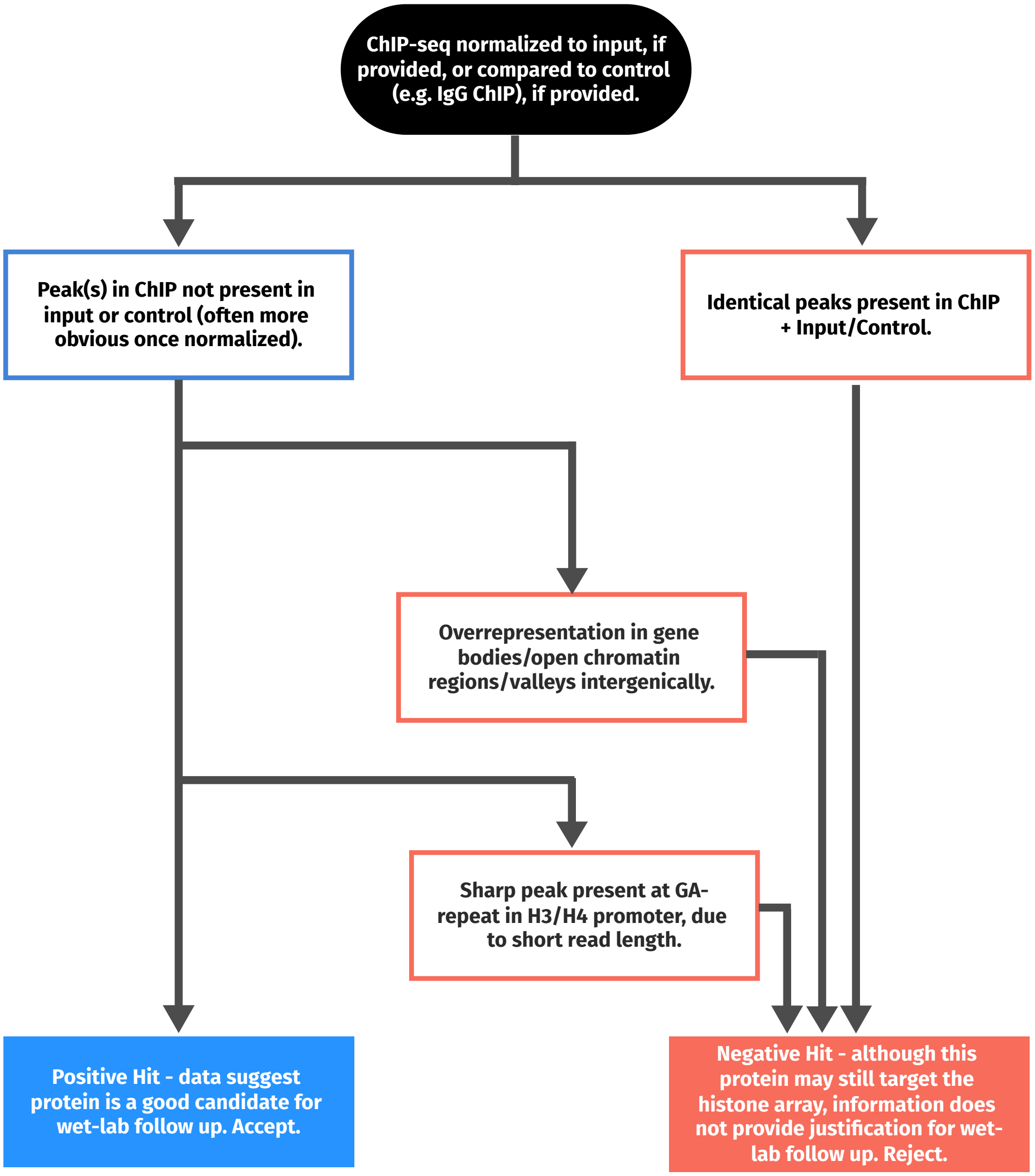
Supplementary Data

A bioinformatics screen reveals Hox and chromatin remodeling factors at the *Drosophila* histone locus

Reproduced with permission by:

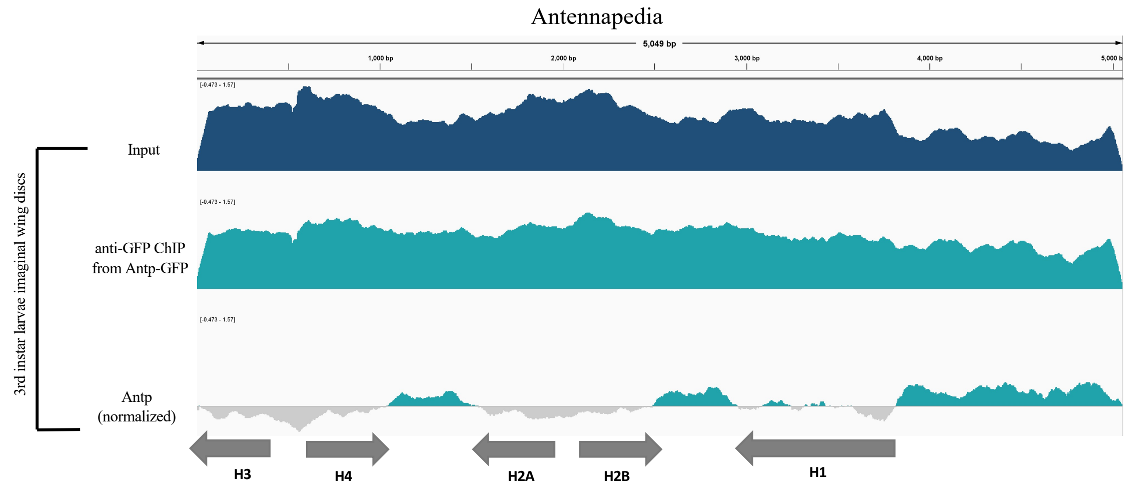
Lauren J. Hodkinson[1]*, Connor Smith[2]*, H. Skye Comstra[2], Bukola A. Ajani[2], Eric H. Albanese[2], Kawsar Arsalan[2], Alvaro Perez Daisson[2], Katherine B. Forrest[2], Elijah H. Fox[2], Matthew R. Guerette[2], Samia Khan[2], Madeleine P. Koenig[2], Shivani Lam[2], Ava S. Lewandowski[2], Lauren J. Mahoney[2], Nasserallah Manai[2], JonCarlo Miglay[2], Blake A. Miller[2], Olivia Milloway[2], Nhi Ngo[2], Vu D. Ngo[2], Nicole F. Oey[2], Tanya A. Punjani[2], HaoMin SiMa[2], Hollis Zeng[2], **Casey A. Schmidt[2]*, Leila E. Rieder[2]*** 2023 A bioinformatics screen reveals hox and chromatin remodeling factors at the *Drosophila* histone locus. BMC Genomic Data 24: 54. https://doi.org/10.1186/s12863-023-01147-0

*Authors contributed equally


ChIP-seq normalized to input, if provided, or compared to control (e.g. IgG ChIP), if provided.
Peak(s) in ChIP not present in input or control (often more obvious once normalized).
Identical peaks present in ChIP + Input/Control.
Overrepresentation in gene bodies/open chromatin regions/valleys intergenically.
Sharp peak present at GA-repeat in H3/H4 promoter, due to short read length.
Positive Hit - data suggest protein is a good candidate for wet-lab follow up. Accept.
Negative Hit - although this protein may still target the histone array, information does not provide justification for wet-lab follow up. Reject.

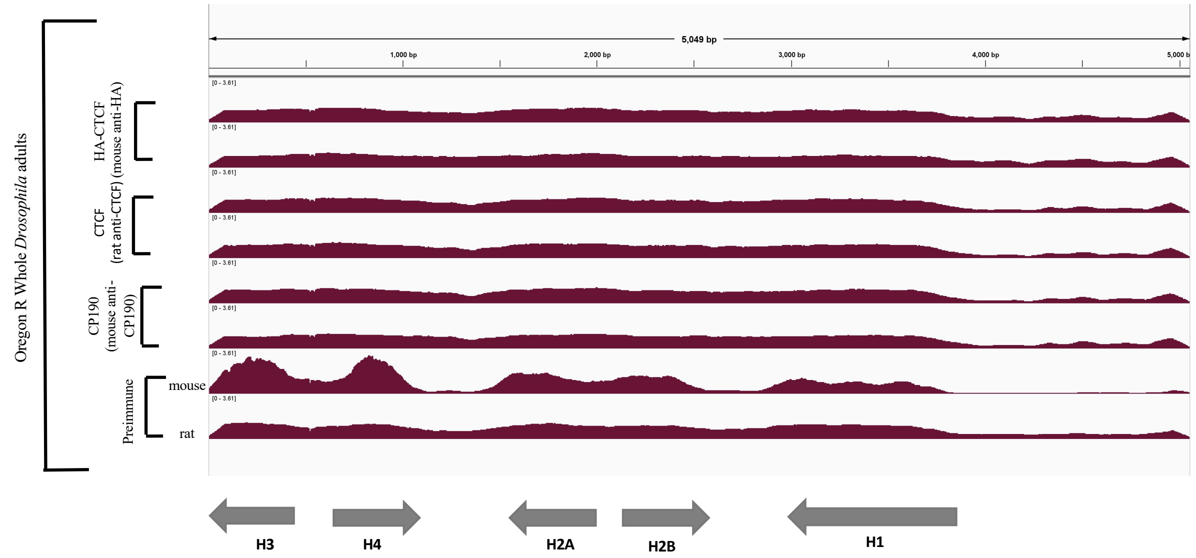
Supplemental Figure 5.1: Qualitative assessment for scoring candidates as positive or negative.

Supplemental Figure 5.2: Factors considered negative hits



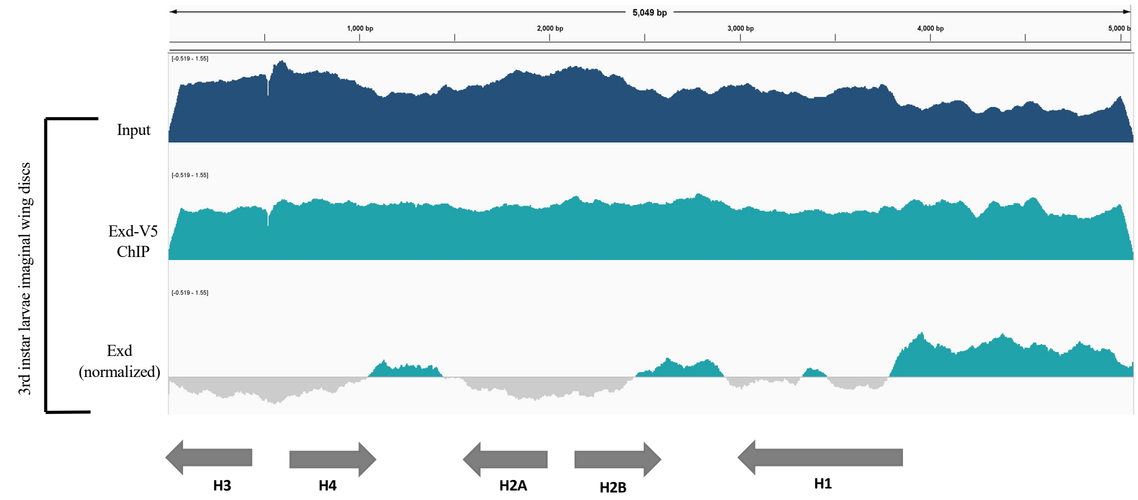
We mapped Antennapedia (Antp) ChIP-seq (cyan) and input (navy) data (Kribelbauer, *et al.* 2020) from 3rd instar larvae imaginal wing discs to the histone gene array. Experiment used an anti-GFP antibody to immunoprecipitate Antp-GFP. Antp does not show convincing localization to the histone gene array when compared to input.

CTCF

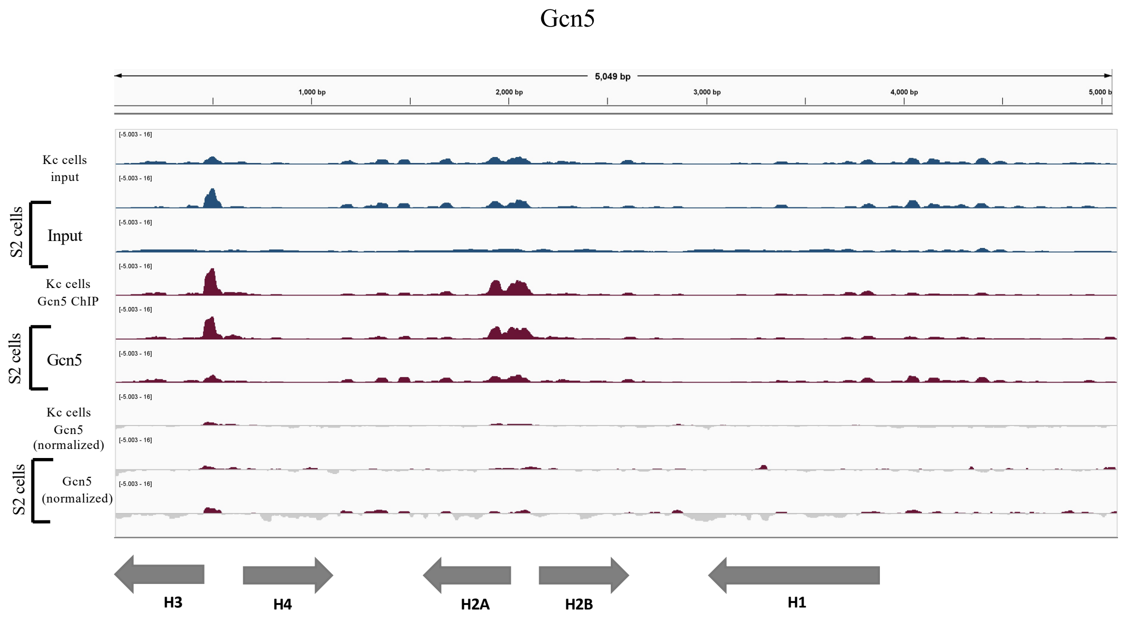


We mapped CTCF-HA ChIP, CP190 ChIP, and preimmune ChIP-seq data (Kyrchanova, *et al.* 2021) from OregonR whole *Drosophila* adults to the histone gene array. CTCF and CP190 do not show convincing localization to the histone gene array when compared to preimmune.

Extradenticle

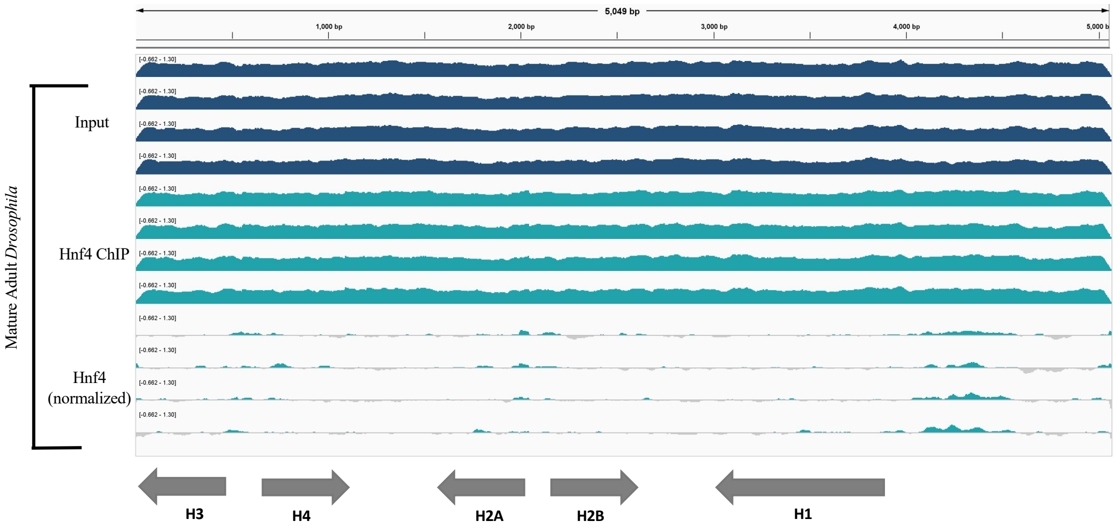


We mapped Extradenticle-V5 (Exd) ChIP-seq (cyan) and input (navy) data (Kribelbauer, *et al.* 2020) from 3rd instar larvae imaginal wing discs to the histone gene array. Experiment used an anti-V5 antibody to immunoprecipitate Exd-GFP. Exd does not show convincing localization to the histone gene array when compared to input.



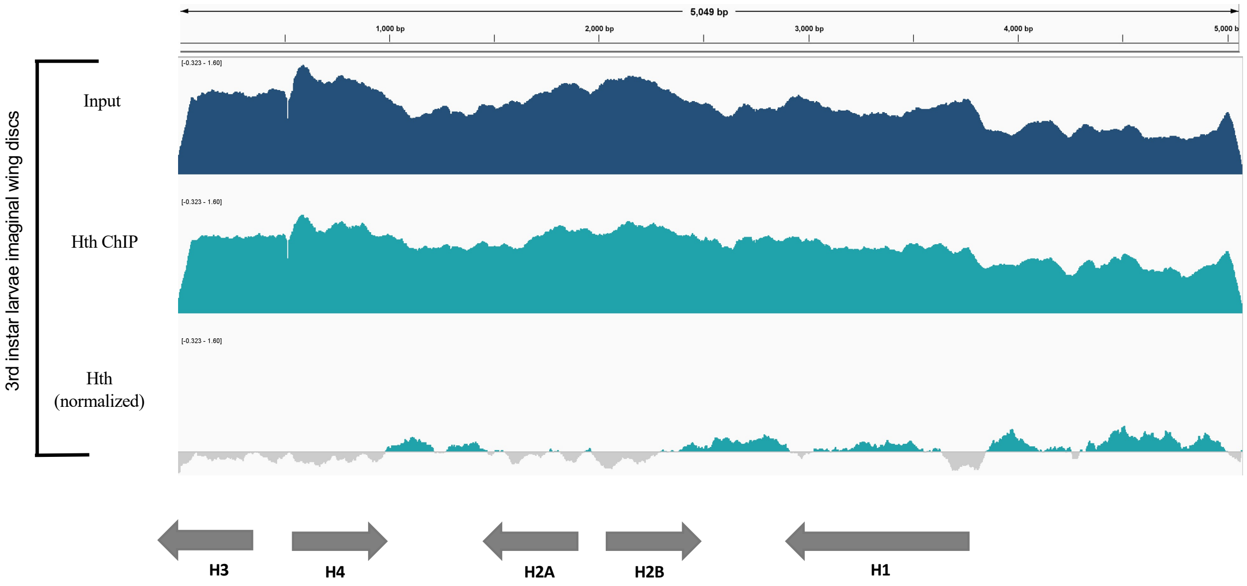
We mapped Gcn5 acetyltransferase (Gcn5) ChIP-seq (maroon) and input (navy) data (Ali, *et al.* 2017) from both Kc cells (one replicate) and S2 cells (two replicates) to the histone gene array. Gcn5 does not show convincing localization to the histone gene array when compared to the corresponding input.

Hepatocyte nuclear factor 4



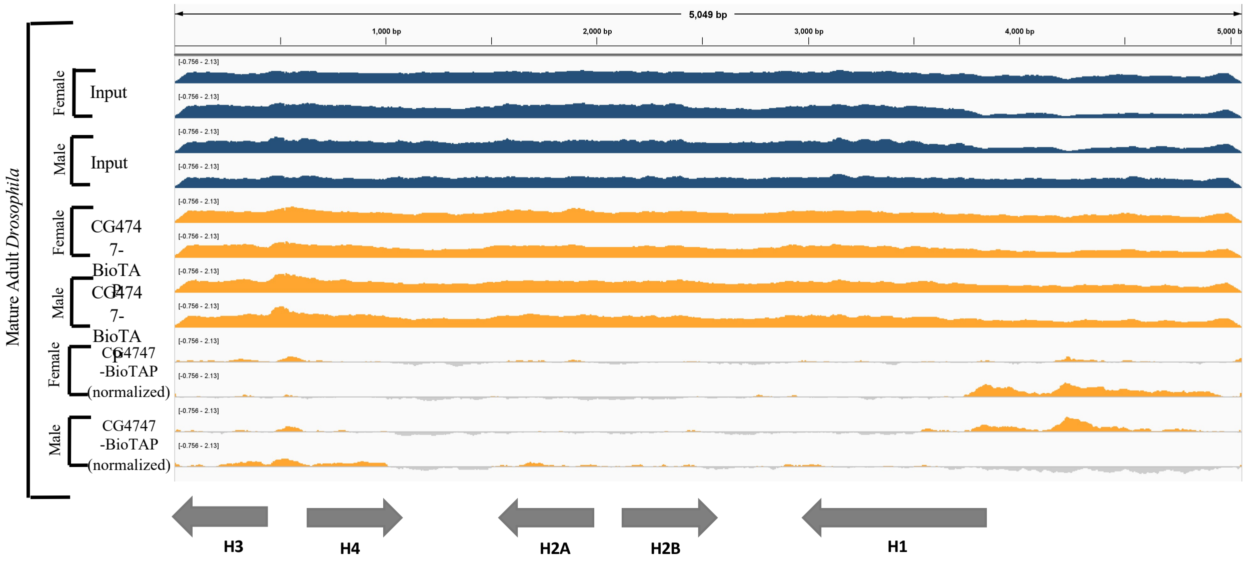
We mapped Hepatocyte nuclear factor 4 (Hnf4) ChIP-seq (cyan) and input (navy) data (Thummel, *et al.* 2015) from whole mature adult *Drosophila* to the histone gene array. Hnf4 does not show convincing localization to the histone gene array when compared to input.

Homothorax

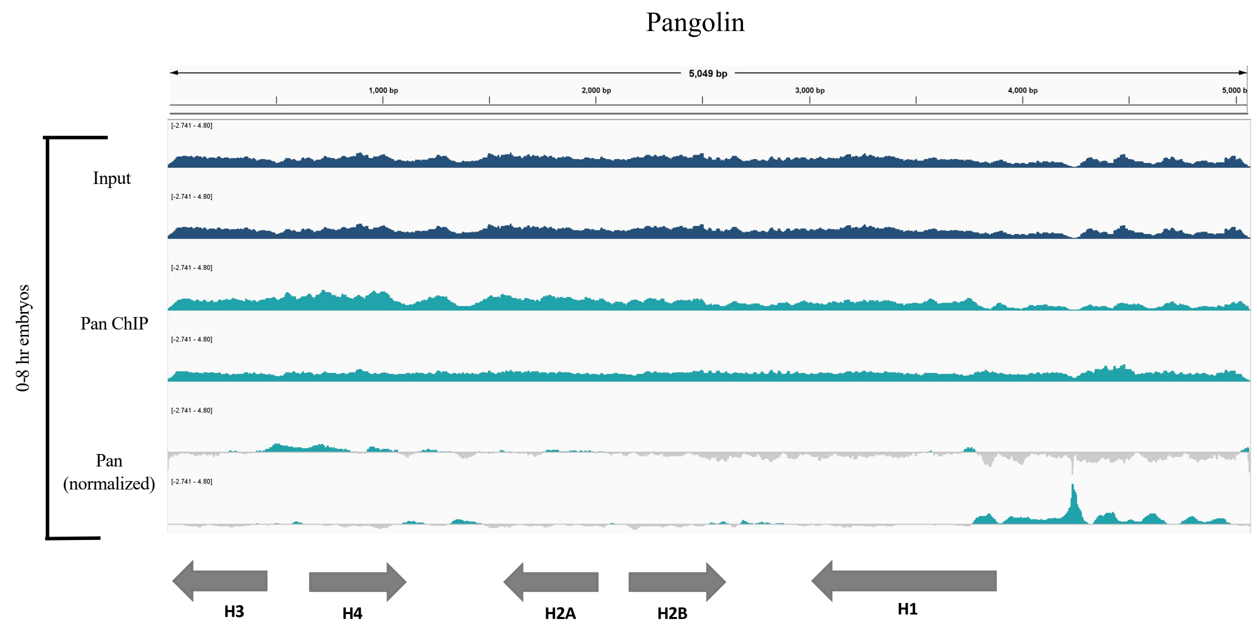


We mapped Homothorax (Hth) ChIP-seq (cyan) and input (navy) data (Kribelbauer, *et al.* 2020) from 3rd instar larvae imaginal wing discs to the histone gene array. Hth does not show convincing localization to the histone gene array when compared to input.

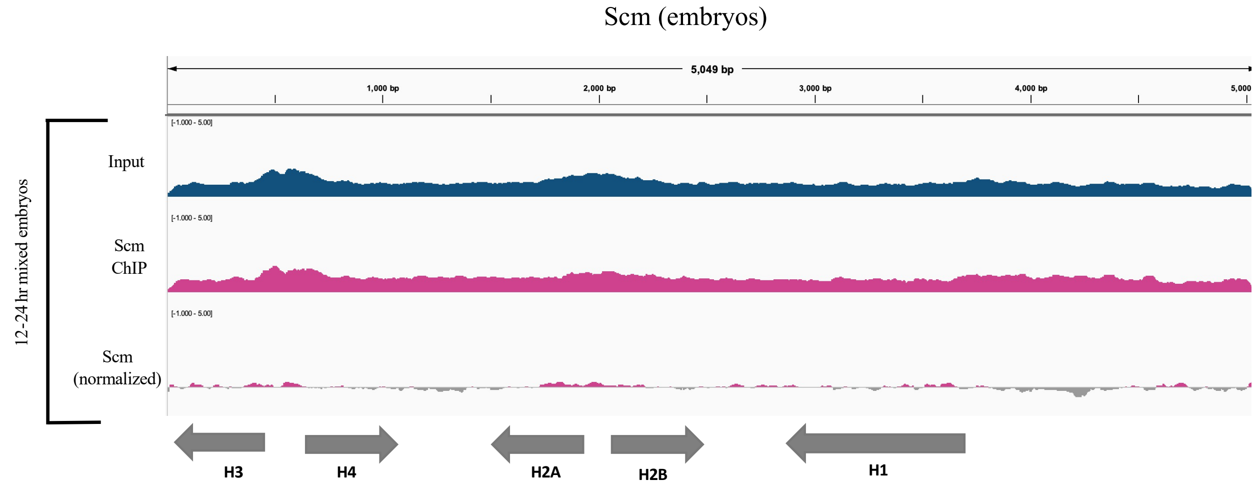
Nucleosome-destabilizing factor/CG4747



We mapped Nucleosome-destabilizing factor-BioTAP (Ndf/CG4747) ChIP-seq (yellow) and input (navy) data (GSE42025) from whole mature adult *Drosophila* to the histone gene array. Experiment used an anti-Bio-TAP antibody in animals expressing CG47474-BioTAP. Ndf/CG47474 does not show convincing localization to the histone gene array when compared to input.



We mapped Pangolin (Pan) ChIP-seq (cyan) and input (navy) data (ModENCODE) from 0-8 hr embryos from the *y,cn,bw,sp* genotype to the histone gene array. Pan does not show convincing localization to the histone gene array when compared to input.



We mapped Suppressor of zeste 12 (Su(z)12) ChIP-seq (maroon) and input (navy) data (Herz *et al*., 2012) from 3rd instar larvae to the histone gene array. Su(z)12 does not show convincing localization to the histone gene array when compared to input.

Candidate	Category	Rationale	Tissue/Timing	Localization	Region
Abd-A Abdominal A	Early development transcription factor	Continuous expression post 0-2hour developmental stage, with highest RNA expression levels between 2-4hrs	Kc cells	Yes	*H3/H4* promoter
Abd-B abdominal B	Early development transcription factor	Continuous expression post 0-2hour developmental stage, with highest RNA expression levels between 2-4hrs	Kc cells	Yes	*H3/H4* promoter
Antp Antennapedia	Early development transcription factor	Expressed in the early embryo	Imaginal wing disk		
CTCF CCCTC-binding factor	Chromatin structure/remodeler	Key chromatin architecture protein, serves as an insulator and allows for distant DNA interacts	Mixed adults		
CP190 Centrosomal protein 190kD	Chromatin structure/remodeler	Insulates centromeric heterochromatin	Kc/S2 cells		
Exd extradenticle	Early development transcription factor	Expressed in the early embryo	5-6 days wing		
Fs(1)h female sterile (1) homeotic	Chromatin structure/remodeler	short isoform binds at promoters and enhancers and long isoform binds at chromatin insulators	Kc cells	Yes (long isoform only)	*H3/H4* promoter *H2A/H2B* promoter
Gcn5	Chromatin structure/remodeler	Acetyltransferase, critical for oogenesis and morphogenesis and associates with insulators	Kc cells		
Hnf4 Hepatocyte nuclear factor 4	Early development transcription factor	Important for major events in embryogenesis, works in cell metabolism pathways	Mixed adults		
Hr78 Hormone-receptor-like in 78	Early development transcription factor	Continuous expression throughout embryogenesis and during metamorphosis	8-16h embryos	Yes	*H3/H4* promoter
Hth homothorax	Early development transcription factor	Expressed in the early embryo	Imaginal wing disc		
JIL-1	Dosage compensation/ X-chromosome associated factor	Kinase phosphorylation of H3S10 Enriched on X chromosome for dosage compensation Serine 10 mark prevents heterochromatin spreading	3rd instar larvae	Yes	*H2A/H2B* promoter
M1BP Motif 1 Binding Protein	Chromatin structure/remodeler	Interacts with TRF2 and related to insulator activity by associating with CP190	Kc/S2 cells		
MSL1 Male specific Lethal 1	Dosage compensation factor	Associates with the known HLB factor CLAMP	S2 cells		
Ndf (CG4747) Nucleosome-destabilizing factor	Dosage compensation associated factor	H3K36me3-binding protein that is important for MSLc localization	larval		
Nej Nejire	Hit from previous screen for HLB factors	Acetyltransferase and early developmental transcription factor	2-4 hr embryos S2 cells	Yes (embryos only)	*H3/H4* promoter *H2A/H2B* promoter
Opa Odd Paired	Early development transcription factor	Continuous expression post 0-2hour developmental stage, with highest RNA expression levels between 2-4hr when ZGA/ histone gene expression is high (73,79)	3h embryo 4h embryo		
Pan Pangolin	Early development transcription factor	Expressed in the early embryo/early in development	0-8h embryo		
Pnt Pointed	Hit from previous screen for HLB factors	Early development transcription factor	Stage 11 embryo		
Psc Posterior sex combs	Chromatin structure/remodeler	Polycomb member	S2 cells		

Scm Sex comb on midleg	HLB-associated factor	Genetically interacts with known HLB factor Mxc	Embryo 12-24h S2 cells		
su(z)12 suppressor of zeste 12	Chromatin structure/remodeler	Polycomb repressive complex member Highest RNA expression levels during 0-2h and 4-8h development stages	3rd instar larvae		
TAF1 TBP-associated factor 1	General transcription factor	TATA-box-binding protein known to associate with TBP	S2 cells	Yes	*H3/H4* promoter
TFIIB Transcription Factor II B	General transcription factor	TATA-box binding protein complex member, known to associate with TBP	OregonR Embyos	Yes	*H3/H4* promoter *H2A/H2B* promoter
TFIIF Transcription Factor II F	General transcription factor	TATA-box binding protein complex member, known to associate with TBP	OregonR Embyos	Yes	*H3/H4* promoter *H2A/H2B* promoter *H1* promoter
TRF2 TATA box binding protein-related factor 2	General transcription factor	TATA-less promoter binding activity at *H1* promoter	S2 cells	Yes	*H1* promoter
Ubx Ultrabithorax	Early development transcription factor	Continuous expression post 0-2hour developmental stage, with highest RNA expression levels between 2-4hrs	Kc cells, imaginal wing disc embryos	Yes (all)	*H3/H4* promoter

Supplementary Table 5.1: All candidates categories, function, and tissue details.