Distribution Agreement

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

Yiwen Wang Ap ril 7, 2021

Identification of novel histone locus body components in *Drosophila melanogaster*

by

Yiwen Wang

Leila Rieder
Adviser

Biology

Leila Rieder

Adviser

Dr. Arri Eisen

Committee Member

Dr. Judith Fridovich-Keil

Committee Member

2021

Identification of novel histone locus body components in *Drosophila melanogaster*

By

Yiwen Wang

Leila Rieder

Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Biology

2021

Abstract

Identification of novel histone locus body components in *Drosophila melanogaster*
By Yiwen Wang

The conserved histone locus body (HLB) facilitates replication-dependent histone biogenesis during the early developmental stages in most eukaryotic organisms. However, many regulatory factors at the HLB have not been identified. In this paper, I use two research approaches, an unbiased proteomic screening method, and a bioinformatics-based protein candidate method, to identify potential protein factors that target the histone genes. For the unbiased proteomic screening, I adopted a dCas9 (inactive Cas9)-gRNA sequence-specific system that targets the histone locus in *D. melanogaster*. Future work with this system will discover new HLB factors by mass spectrometry. For the bioinformatics-based candidate approach, I mapped publicly available ChIP-seq datasets of potential HLB factors to a single copy of the histone gene array of *Drosophila melanogaster*. I identified SIN 187 and SIN 220, two major isoforms of the SIN3 complex, as candidates that target the histone locus. I expect to test my hypothesis through subsequent immunostaining experiments. Overall, through the two mentioned experimental approaches, I expect to contribute to the further understanding of HLB regulatory factors at the histone locus and their contribution to the coordinated histone biogenesis in *Drosophila melanogaster*.

Identification of novel histone locus body components in *Drosophila melanogaster*

By

Yiwen Wang

Dr. Leila Rieder

Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Biology

2021

Acknowledgements

First and foremost, I would like to thank Dr. Leila Rieder for her guidance, patience, insight, for answering my many questions, and for being a fantastic all-around mentor. If she had not given me the chance to talk with her and learn about her research project in the winter of 2018, I would not have the wonderful experience of working in her lab. I would also like to thank members of my Honors Committee, Dr. Arri Eisen and Dr. Judith Fridovich-Keil, for reviewing both my proposal and thesis and for giving me inspiring ideas on my research project. Besides, I would like to thank people in my lab for their support along the way. I want to thank Casey Schmidt for mentoring me for the past year and a half, assisting me with my research, and writing this thesis. I want to thank Gwyn Puckett, our lab manager, for providing fly food and being in charge of lab equipment, resources, and orders. I want to thank Skye Comstra for developing the bioinformatics pipeline in Galaxy that was used in my research project, and Lauren Hodkinson, for teaching other undergraduates and me in the lab how to use Galaxy during the CURE project in the fall semester of 2020. Thanks to them and other undergraduate students in the lab, I have enjoyed a research experience gaining invaluable assistance, support, and inspiration. Finally, I would like to thank my friends and family members who had provided me much mental support as I was completing the Honors Program.

Table of Contents

`

Table of Figures

Tables

Chapter 1: Introduction

Replication-dependent histone genes and the histone locus in *Drosophila melanogaster*

A eukaryotic nucleus is a very compact and crowded environment: billions of base pairs need to be fit into a nucleus with a diameter of 6 micrometers. Thus, in eukaryotes, the proper organization of the genome requires extensive three-dimensional configurations. The basic unit of genome organization is the nucleosome, which consists of DNA wrapped around the four core histone proteins, H2A, H2B, H3, and H4. During the early developmental stage, massive cellular differentiation and DNA replication give rise to a sudden increase in the need for the four core histone proteins and the linker histone, H1[1]. When cells undergo differentiation, the reorganization of genomic DNA results in the need for additional histones that is specific to each cell type. During the S phase of mammalian cells, approximately 4 × 10 ^ 8 molecules of each of the four core histone genes must be synthesized each cell cycle to package the newly-replicated DNA[2].

The replication-dependent histone genes have several unique features. First, animal genomes contain multiple copies of each of the five histone genes. Second, the five replication-dependent histone genes have remained relatively clustered throughout evolution. Third, the five histone genes encode the only non-poly adenylated mRNA transcripts in eukaryotes[3].

In *Drosophila melanogaster*, the single histone locus resides on chromosome 2L and consists of 107 tandem copies of a 5kb histone gene array[4]. Each array contains the five replication-dependent histone genes, H1, H2a, H2b, H3, and H4. In each 5kb array, H2a and H2b share a bidirectional promoter, as well as H3 and H4[5] (Fig. 1).

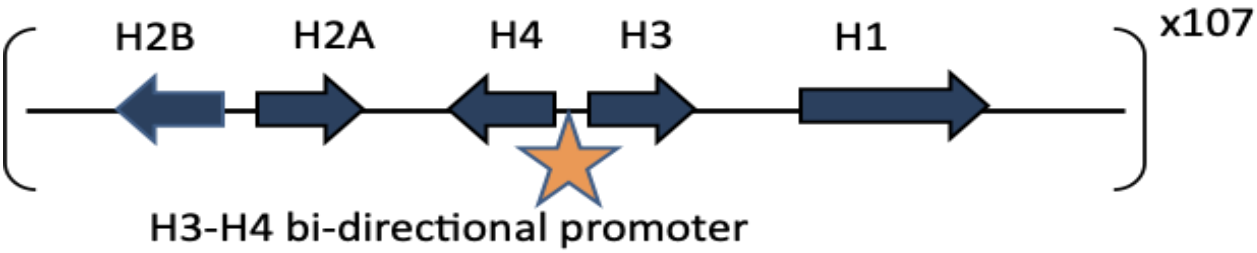


Figure 1. Configuration of the histone gene array in *Drosophila melanogaster*.
The histone locus is made of 107 tandem repeats of a ~5kb array consisting of five replication dependent histone genes[4]. Researchers have hypothesized that the H3/H4 bi-directional promoter (starred) is the trigger for the formation of the histone locus body, a group of known and unknown protein factors that regulate histone biogenesis.

The Histone Locus Body (HLB)

In the model organism *Drosophila melanogaster*, the sudden rise in the need for the five replication-dependent histone proteins in equal amounts is fulfilled by the highly conserved histone locus body (HLB) that assembles at the repetitive histone locus. To improve the efficiency of histone biogenesis, the HLB concentrates regulatory factors[6] and promotes interactions among them that would otherwise be stochastic. A set of transcription and pre-mRNA processing factors are found in the HLB, leading to highly coordinated histone biogenesis[2] (Table 1). Some components of the HLB localize at the histone locus throughout the cell cycle, and nascent histone mRNA transcripts undergo processing in the HLB before export to the cytoplasm.

Several factors at the HLB are necessary for histone biogenesis. In the absence of pre-mRNA processing factors U7 SnRNP, poly-adenylated histone transcripts can be detected by *in situ* hybridization assays[7]. The insufficient processing of pre-mRNA transcripts is also observed in

mutants in which another HLB processing factor, FLASH, is prevented from interacting with specific HLB components and localizing at the histone locus[6].

It is not known how the HLB assembles specifically at the histone locus, but previous research demonstrated that the bidirectional promoter between the H3 and H4 genes contains sequences required for HLB assembly at the histone[5,8]. HLB assembly at the histone locus is hypothesized to occur through a "seed and grow" mechanism. With the presence of the H3/H4 promoter, the formation of the HLB is an ordered process. The core components Mxc and FLASH are first recruited, followed by the recruitment of Mute and U7 snRNP[8]. However, this hypothesized stepwise mechanism does not rule out stochastic self-organization[9].

Table 1. An expanding list of known proteins at the HLB and their role in histone biogenesis. This list from Marzluff and Duronio (2017) summarizes proteins known in 2017 that concentrate at the HLB. Since 2017, several new protein candidates such as CLAMP[5] and Prp40[10], have been discovered in *Drosophila*, suggesting that this list is not comprehensive.

Component	Function during histone mRNA biosynthesis	Organism
NPAT / Mxc	Transcription initiation/HLB assembly	Human, Mouse, Drosophila
HiNF-P	Transcription initiation	Human
RNA Pol II	Transcription	Xenopus, Drosophila
TBP	Transcription initiation	Drosophila
TRF2	Transcription initiation	Drosophila
TFIIA	Transcription initiation	Drosophila
Myc	Transcription initiation	Drosophila
GAPDH	Transcription initiation	Drosophila
NELF	Transcription elongation	Human
Spt6	Transcription elongation	Drosophila
ARS2	Transcription elongation	Human
FLASH	mRNA 3′ end processing/HLB assembly	Human, Mouse, Drosophila
Symplekin	mRNA 3′ end processing	Xenopus, Drosophila
ZFP100	mRNA 3′ end processing	Human
U7 snRNP	mRNA 3′ end processing	Human, Drosophila, fish, frogs
Mute/YARP	Represses histone mRNA accumulation	Human, Drosophila
WGE	Represses histone mRNA accumulation	Drosophila
Abo	Represses histone mRNA accumulation	Drosophila
HERS	Represses histone mRNA accumulation	Drosophila
hCINAP	Unknown	Human
PARP	Unknown	Drosophila
Cpn10	Unknown	Human
WDR79	Unknown; likely a Cajal body component	Drosophila
Coilin	Unknown/not required	Xenopus, Drosophila
MPM-2	Detects Cyclin E/Cdk2 dependent phosphoepitope on Mxc	Drosophila

The Unknowns of the Histone Locus Body (HLB)

Despite previous efforts over the past 15 years to decipher the components, the assembly, and the function of the HLB, researchers have not learned the full story of it. One question that has remained unsolved is how do HLB components initially find and target the histone locus. While researchers have previously identified a set of protein factors that are involved in the transcription or the processing of histone transcripts, only one of these known HLB factors, CLAMP (Chromatin-linked adaptor for MSL proteins), is DNA-binding. At the *Drosophila* histone locus, CLAMP is bound specifically to a short but highly conserved GA repeat in the H3/H4 promoter[5]. However, since CLAMP also binds at other non-histone loci[11], additional regulatory factors and initiation factors must be present at the promoter region to confer specificity at the histone locus. In addition, since processing of immature histone mRNA is a complicated pathway, it is likely that there are still unknown processing factors that localize at the locus. Thus, the identification of novel regulatory factors at the HLB will enhance our understanding of the mechanisms of HLB formation as well as for histone biogenesis in *Drosophila melanogaster* (Fig. 2).

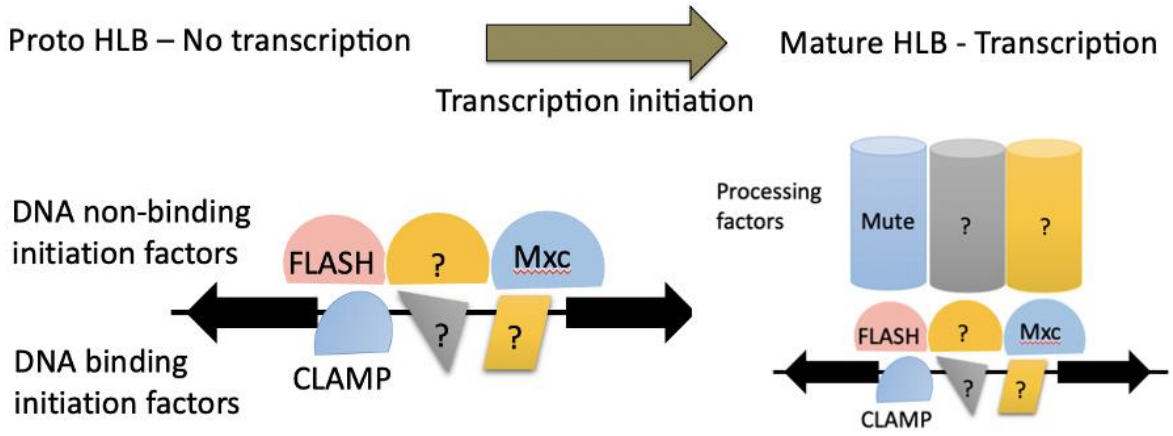


Figure 2. An assembly model of *Drosophila* HLB at H3/H4 promoter.
During embryonic nuclear cycle 10, DNA binding and non-binding initiation factors form a "proto-HLB" in the absence of transcription[8]. At cycle 11, transcription initiation at the histone locus recruits various processing factors to the proto-HLB, forming a mature HLB. Once established, HLB remains attached to the locus in the absence of histone gene expression. Some of the known HLB factors are labeled in the figure, including CLAMP, the DNA-binding factor, and non-DNA binding factors (FLASH, Mxc, and Mute) that are necessary for normal histone biogenesis.

Conventional protein identification methods have their limitations. Proteomics screening experiments failed to pull out many of the known HLB[9,12], suggesting unsaturated screening results. With such limitations in mind, it then becomes critical to develop novel screening techniques that can identify more HLB components at once.

As a member of Dr. Leila Rieder's lab at Emory University, my undergraduate research project aims to address the question of identifying novel HLB components through innovative approaches. In Chapter 2 of the thesis, I will outline an unbiased proteomic screening approach utilizing an inactive Cas9 - guide RNA (dCas9-gRNA) system. In Chapters 3, and 4, I will introduce a protein candidate approach which utilizes bioinformatics analysis and publicly available datasets.

Chapter 2: Identification of novel HLB candidates by dCas9-targeted locus specific protein isolation in *Drosophila melanogaster*

Introduction

The first approach I took was an unbiased proteomic screening approach. I developed a catalytically inactive Cas9 (dCas9) system that targets the H3/H4 promoters in the *D. melanogaster* histone locus, directed by guide RNAs[12] (Fig. 3A). To address any unspecific targeting of the dCas9 enzyme on the genome, I also constructed a control gRNA plasmid that contains gRNAs for the 5S rRNA loci. The 5S rDNA also exists as tandem repeats in *Drosophila*, sharing a similar genomic structure with the histone cluster (Fig. 3B). Since the mature HLB forms at nuclear cycle 11[8], I plan to perform the proteomic screening experiment in *Drosophila* embryos. Currently, I have integrated all three transgenes into the *Drosophila* genome in separate fly lines. In the future, I will cross the dCas9-expressing flies to those expressing the gRNAs. After verifying expression and correct positioning of dCas9 by western blot and polytene immunofluorescence staining, I will perform cross linked ChIP using antibodies that target the FLAG tag on the dCas9 protein in *Drosophila* embryos. I expect to isolate the H3/H4 promoter region as well as protein factors in close proximity. The final step will be to proceed to mass spectrometry to identify and compare proteins localized at the H3/H4 promoter and 5S rRNA loci (Fig. 4). A newly identified HLB candidate would be a protein that is enriched specifically at the H3/H4 promoter.

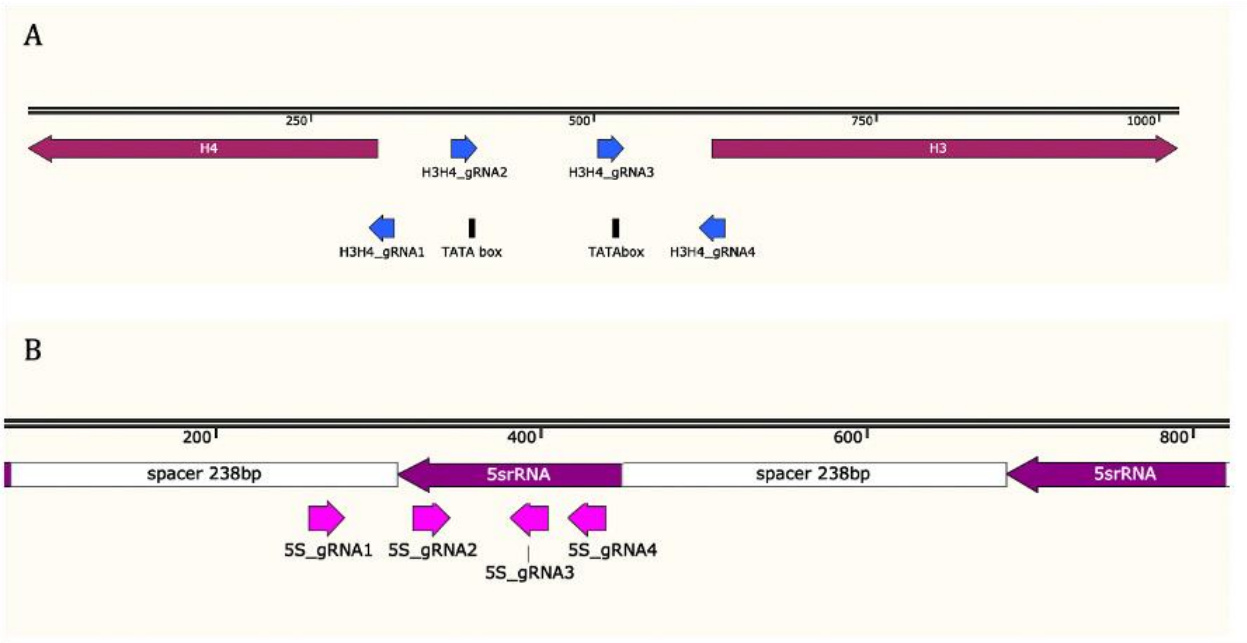


Figure 3. Location of the gRNAs that target dCas9 to the H3/H4 promoter (A) and 5S rRNA (B) loci.

The binding sites of the gRNAs were carefully selected to overlap with DNAse I hypersensitivity sites[13] so that they will not interfere with the normal binding of regulatory factors at these sequences. (A) The 4 gRNAs have a good coverage of the H3/H4 promoter region. (B) gRNA targeting 5S -rRNA loci in Drosophila is the control in this study. In Drosophila, the 5S rRNA genes exist as a gene clustser made of tandem repeats, a structure similar to the histone gene cluster. Figures generated in Snapgene.

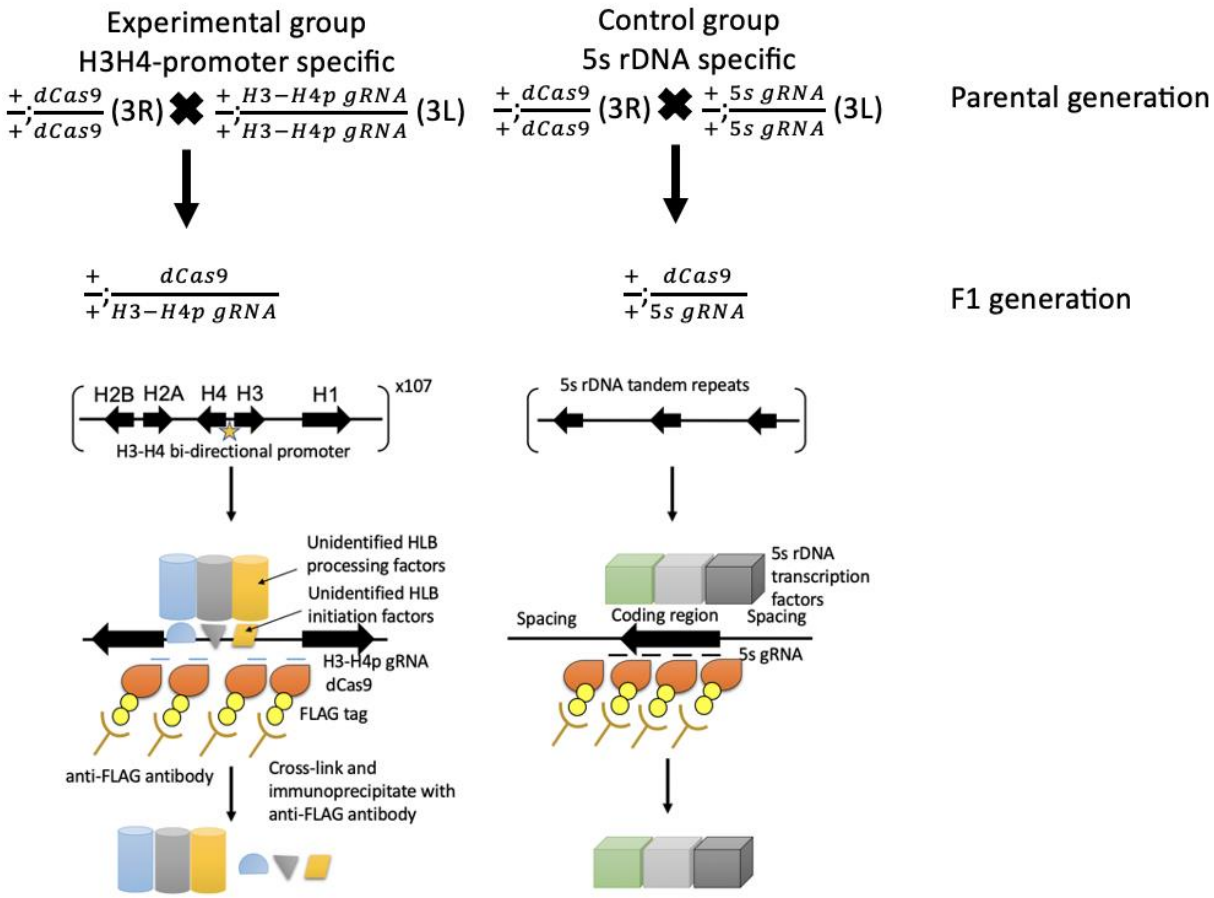


Figure 4. gRNAs guide dCas9 to specific repetitive loci in the *Drosophila* genome. H3/H4p gRNAs will guide dCas9 to the histone promoter region of interest while 5s gRNAs will guide the enzyme to the 5S rDNA. After immunoprecipitation using the anti-FLAG antibody, I expect to identify different proteins from the two experiments. I expect to identify unknown components of the HLB from animals that express both dCas9 and H3H4p - gRNA.

Methods

Lead contacts and materials availability

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Leila Rieder (leila.rieder@emory.edu). All newly generated plasmids (dCas9 and gRNA) are available upon request.

Original Cas9 plasmid

We used a plasmid, Hsp70 Cas9 3x FLAG, (Plasmid #46294) provided by Addgene for the construction of the dCas9 gene. This Cas9 transgene is codon-optimized for expression in *Drosophila* and its expression is driven by an endogenous, inducible *hsp70* promoter. While the induction of the *hsp70* promoter normally depends on the binding of heat-shock factors[14], it possesses a basal leakiness at permissive temperatures that allows expression of the Cas9 transgene in *Drosophila* embryos before zygotic genome activation[15]. We did not know the exact length of the plasmid, so we did two restriction enzyme diagnoses using known cutting sites for BglI and SacI and reached an estimation of ~9kb.

PCR mutagenesis

In order to catalytically inactive the Cas9 protein into dCas9, we introduced two point mutations into the coding sequence of the Cas9 gene using New England Biolab Q5 site-directed PCR mutagenesis kit. The first mutation introduced a D10A substitution and the second mutation introduced a H841A substitution[12]. We introduced one mutation at a time. For the D10A mutation, we used a forward primer dCas9_D10A_ F, ATCGGCCTGGCCATCGGCACC,

along with a reverse primer dCas9_D10A_R, GCTGTACTTCTTGTCGGCTGC. The codon change made by these primers is a point mutation from GAC (D) to GCC (A). For the H841 mutation, we used a forward primer dCas9_H841A_F, CGATGTGGACGCCATCGTGCCTCAGAG, and a reverse primer, dCas9_H841A_R, TAGTCGGACAGCCGGTTG. The codon change made by these primers is a three-point mutation from CAG (H) to GCC (A). For both site-directed mutageneses, we designed primers with NEBase changer. We obtained primers from IDT.

Verification of PCR mutagenesis

After each site-directed mutagenesis, we sequenced the whole plasmid and verified the newly introduced mutations (GeneWiz). The sequencing primer for D10A was a reverse primer D10A_seq_R, CAGATCCGGTTCTTCCGTCTG. The sequencing primer for H841A was a forward primer H841A_seq_F, CTGAAAGAACACCCCGTGGAAAAC.

Subclone the dCas9 segment into a pCFD3.1 histone plasmid (Fig 5)

Since the original dCas9 plasmid did not contain any phenotypic marker for *Drosophila*, we subcloned the Hsp70-dCas9-3xFLAG into a new histone vector obtained from Dr. Robert Duronio's lab. The backbone contains a *mini-white* marker that encodes red eye color. We first amplified the complete dCas9 segment using the aforementioned primers by PCR and introduced two restriction enzyme cutting sites, one for BsiWI and another for NotI, to either end of the amplified segment using NEB Q5 polymerase. The primer that introduced the BsiWI cutting site was BsiWI_dCas9_F, GAGGTCGTACGGGACAATATCCCCCTAGAATCCCAAAACAAAC, and the primer that introduced the NotI cutting site was dCas9_NotI_R1,

AATTGCGCGGCCGCGGATCC. The length of the final PCR product was 5756bp. We then cut out the Hsp70-dCas9-3xFLAG segment from the PCR product using NotI and BsiWI and digested the destination histone vector (where two restriction cut sites were originally included in the plasmid), both using BsiWI and NotI. After purifying the dCas9 segment and the linearized histone backbone, we ligated the two parts together using NEB quick ligation kit.

Original plasmids for H3/H4/5s gRNA (Fig 5)

We used a pCFD5:U6:3-t::gRNA plasmid (#73914 Addgene) for expression of one or multiple tRNA-flanked Cas9 gRNAs under the control of the strong, ubiquitous RNA pol III promoter, U6:3. This plasmid also contained a *vermilion* eye color marker that could be used for recognition of successful transformants after injection. We incorporated multiple gRNA oligonucleotides into the pCFD5 plasmid by Gibson Assembly using the protocol provided with the plasmid and confirmed successful cloning products using gel electrophoresis. The expected size of the gRNAs was 1134bp. Samples with the correct size were sent for Sanger sequencing at GeneWiz using a forward primer U63seqfwd, ACGTTTTATAACTTATGCCCCTAAG, and a reverse primer pCFDseqrev, GCACAATTGTCTAGAATGCATAC.

Transform all three plasmids into a competent E. coli strain and verify successful transformants

I transformed all three newly made plasmids into NEB 5-alpha Competent *E. coli.* This *E. coli* strain was ideal for subcloning efficiency. Each newly made plasmids contained an Ampicillin resistant gene. I plated newly transformed *E. coli* culture on Carbenicillin plates at 37 Celcius overnight. I viewed surviving colonies as candidates for successful transformants. I picked 10 of

the survived dCas9 colonies, 7 of the survived 5s gRNA colonies, and 6 of the survived H3H4 colonies to make overnight cultures and a streak plate. For the dCas9 colonies, I performed miniprep and nanodrop on overnight cultures, confirmed the correct ligation using restriction diagnosis with the SacI enzyme, and sent verified samples for Sanger sequencing at Genewiz. For colonies containing the gRNA plasmids, I performed miniprep and nanodrop on overnight cultures, verified the existence of the transgene through PCR genotyping, and sent verified samples for Sanger sequencing at Genewiz.

Glycerol stocks

We made a 1mL glycerol stock for all three plasmids, which contained 500uL of the overnight culture of the plasmid that had been transformed into the Carbenicillin resistant bacterial cultures and 500uL of 50% glycerol. We stored glycerol stocks in a -80 degrees C freezer.

Midiprep

We midiprepped all three sequence-verified plasmids using Qiagen Midiprep Kit and achieved a concentration above 500ng/uL.

Drosophila melanogaster injection

We sent 10uL of midiprepped samples for injection into *Drosophila* embryos at Genetivision. The company also screened for the *vermilion* marker in gRNA transformants. All plasmids were introduced into transgenic flies through the attB-attP system at site VK27, an open region on chromosome 3R (89E11). The plasmids contained attB sites (Fig 5), which is integrated by PhiC31 integrase into attP sites in the defined genomic locations.

Sequencing

All sequencing mentioned in this study was carried out by GeneWiz. For samples under 10,000bp, we used 10uL of purified samples at 80ng/uL and 5uL of respective primers at 5ng/uL.

Table 2. Key resources table.

REAGENT or RESOURCE	**SOURCE**	**IDENTIFIER**
Bacterial and Virus Strains		
	NEB	
Chemicals, Peptides, and Recombinant Proteins		
FRESH Proteinase K		
Critical Commercial Assays		
Miniprep	Qiagen	Cat No./ID: 27115
Midiprep	Qiagen	Cat No./ID: 12143
PCR clean up kit	Qiagen	Cat No./ID: 28115
Gibson Assembly	NEB	Cat No.: E5510S
Quick ligation kit	NEB	Cat No.: M2200
Q5 site-directed mutagenesis	NEB	Cat No.: E0554S

Experimental Models: Organisms/Strains		
D. melanogaster: For H3H4 and 5s-rDNA injection: VK31 attP	Bloomington *Drosophila* Stock Center	# 9748
For dCas9 injection: VK27 attP	Bloomington Drosophila Stock Center	# 9744
E. coli: For dCas9 plasmid transformation: NEB 5- alpha Competent *E. coli*	NEB	Cat No.: C2988J
Oligonucleotides		
H3H4 gRNA1: ATAGAACAGTGAAAAATGAC	IDT	N/A
H3H4 gRNA2: TTCTGTGTGCCCCTATTTAT	IDT	N/A
H3H4 gRNA3: TCCACGATTGCTATATAAGT	IDT	N/A
H3H4 gRNA4: GTACGAGCCATCTCCGATTT	IDT	N/A
5s gRNA1: TGCTAAAATAAAAACTAAGA	IDT	N/A
5s gRNA2: TGGACGAGGCCAACAACACG	IDT	N/A

5s gRNA3: CACCGAAATTAAGCAGCGTC	IDT	N/A
5s gRNA4: CCATACCACGCTGAATACAT	IDT	N/A
Primer: dCas9_D10A_ F: ATC GGC CTG GCC ATC GGC ACC	IDT	N/A
Primer: dCas9_D10A_R: GCT GTA CTT CTT GTC GGC TGC	IDT	N/A
Primer: dCas9_H841A_F, CGA TGT GGA CGC CAT CGT GCC TCA GAG	IDT	N/A
Primer: dCas9_H841A_R: TAG TCGA CAGCCGGTTG	IDT	N/A
Recombinant DNA		
Plasmid: Hsp70-dCas9-3xFLAG plasmid	This report	N/A
Plasmid: pRD384_pATTB	Dr. Robert Duronio	
Plasmid: H3H4 gRNA-pCFD5: U6:3-t::gRNA	This report	N/A
Plasmid: 5s gRNA-pCFD5: U6:3-t::gRNA	This report	N/A
Other		

Results

Verified constructed plasmids used to generate recombinant Drosophila melanogaster

We successfully constructed and verified the two gRNA-pCFD5 plasmids and the dCas9-pRD384-pATTB histone plasmid. Both pCFD5 plasmids contain a *vermilion* marker (Fig 5A) which encodes for vermilion eye color in flies, so successful transformants can be distinguished from the normal red-eye wild type flies by screening for difference in eye colors. The dCas9 plasmid contains a *mini-white* gene (Fig 5B), which encodes for normal red eye color. In this case, successful transformants with red eyes differ from the non-transgenic flies, which have white eyes.

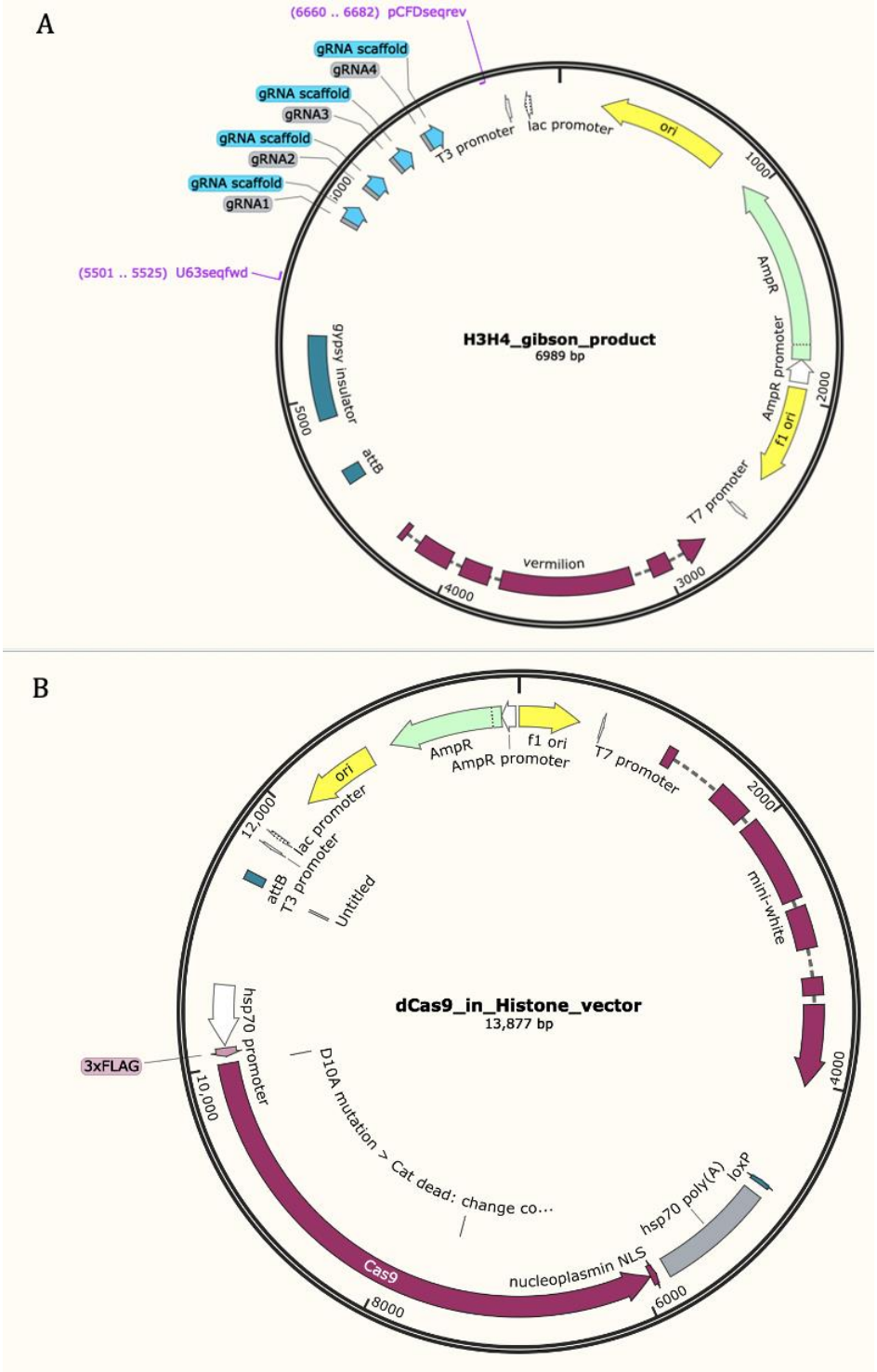


Figure 5. Configuration of the H3/H4gRNA-pCFD5 plasmid and the dCas9-pRD384 histone plasmid.

Each plasmid vector contains an ampicillin resistance marker, an attB site and a phenotypic marker for eye color in *Drosophila*. (A) The H3/H4 gRNA plasmid has a final length of 6989 bp

with a *vermilion* eye color marker. (B) The dCas9 plasmid has a final length of 13,877bp with a *mini-white* gene encoding red eye color.

Validation of gRNA transformants by genotyping

Out of the 8 randomly picked transformants for validation from the H3H4 gRNA strain, 2 of them obtained the full gRNA sequence. Out of the 8 randomly picked transformants from the 5S rDNA strain, all of them obtained the full 5S gRNA (Fig. 6). These results from gel electrophoresis were further sequence-validated by GeneWiz.

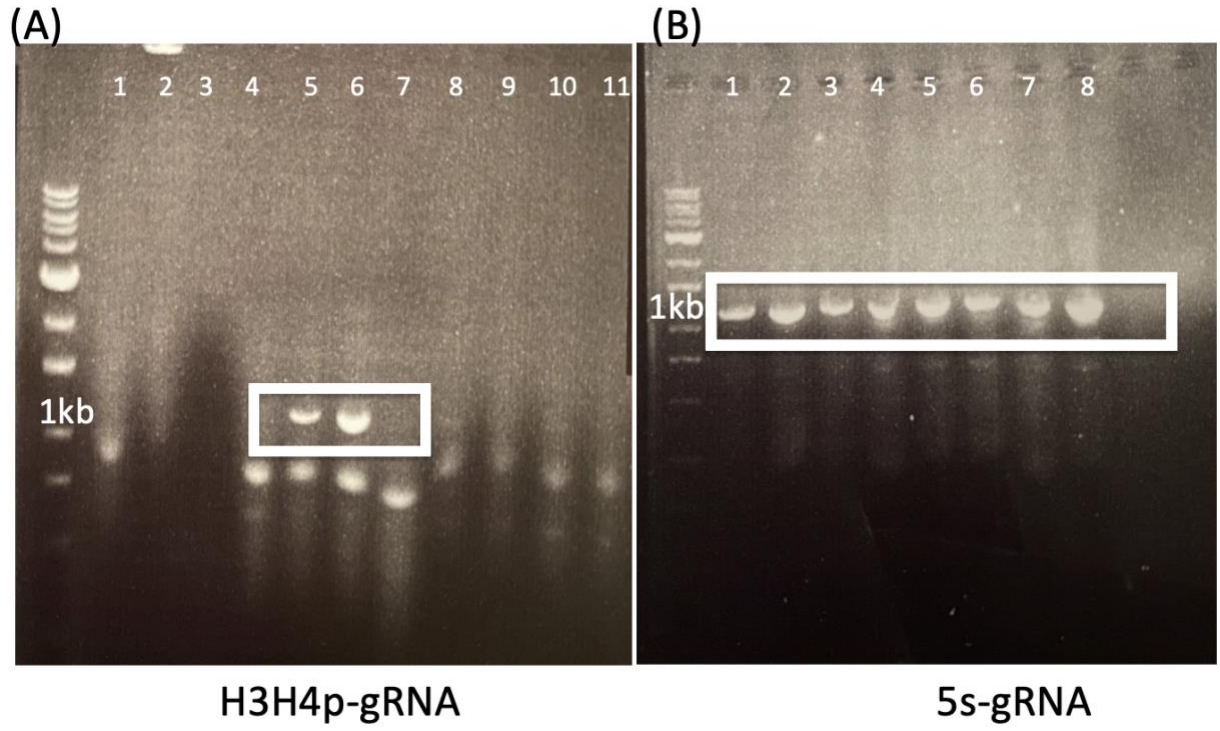


Figure 6. Verification of transgenic animals.
I extracted genomic DNA from animals transgenic for the H3/H4 gRNAs (A) and 5S gRNAs (B) and performed PCR using primers specific to the transgenes. Both gRNA PCR products are expected to have a length of 1134 bp. The 1Kb ladder is labeled for reference. Boxed samples are the successful transformants. (A) Gel electrophoresis of PCR products from genomic DNA extracted from H3/H4 gRNA transformants sequenced using the sequencing primers mentioned in the Method section. Two (boxed) out of 8 candidates contained the validated H3/H4

promoter targeting gRNA (Samples 5 and 6). (B) Gel electrophoresis of PCR products from genomic DNA extracted from 5S gRNA transformants. Eight (boxed) out of 8 candidates contained the complete 5S rDNA targeting gRNA.

Validation of dCas9 transformants by genotyping

Three different transformant lines were produced from injection. A male and a virgin female from each line were genotyped by Casey Schmidt, a post-doc who helped with my research project along the way. Out of the 6 transformants for validation from the dCas9 strain, all of them contained the dCas9 sequence. Neither of the 2 wild-type flies contained the dCas9 sequence (Fig. 7).

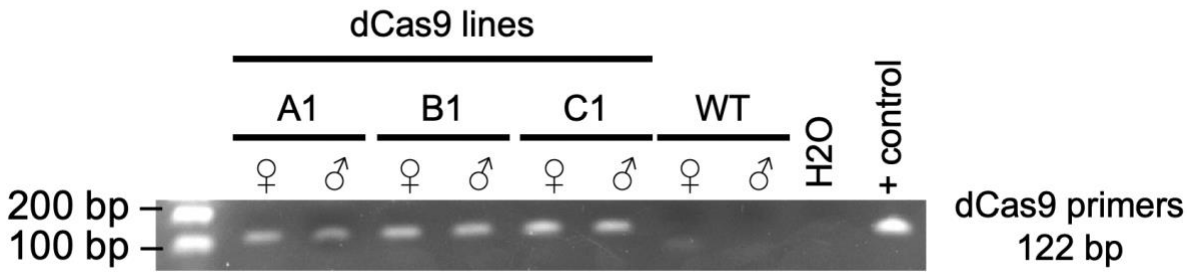


Figure 7. Verification of the dCas9 transgene in genomic DNA extracted from potential transformants.
One male and one virgin female from the three lines of transformants were picked for PCR genotyping for the dCas9 transgene. PCR products are expected to have a length of 122bp. The 100bp and 200bp ladders are labeled for reference. All six parents contained the validated dCas9 sequence. One male and one female Wild Type (WT) fly were genotyped and served as the negative control. A PCR reaction with no template was also used as the negative control and the injected dCas9 plasmid was used as the positive control.

Discussions

I successfully constructed and injected plasmids carrying H3/H4 and 5S gRNAs as well as the one carrying the dCas9 transgene into *Drosophila melanogaster*. I verified transformants with the two gRNAs through PCR genotyping and Casey genotyped transformants with the dCas9 transgene. However, I had to leave the development of this dCas9-gRNA protein identification approach after April 2020 due to the COVID restrictions at Emory University. I plan to continue this project in the summer of 2021, when I should be allowed to work in the lab.

Since mature HLBs form during nuclear cycle 11[8], follow up experiments will be to validate the expression of the dCas9 protein in embryos. In order to do that, we can prepare protein lysate from dCas9 transgenic embryos and proceed to western blot, or do RNA extraction, reverse transcribe RNA extracts into cDNA and run RT-qPCR, which quantifies mRNA level. Since the dCas9 protein is FLAG tagged, we can use anti-FLAG during western blot.

Once we confirm expression of dCas9 in transgenic fly embryos, we can cross the gRNA flies with the dCas9 flies to validate successful guidance of the dCas9 by both gRNAs to the target loci. In this step, we plan to co-stain for HLB-specific proteins such as Mxc and the dCas9 protein and confirm colocalization of signals on polytene chromosomes extracted from the salivary glands of *Drosophila* third instar larvae. Other known HLB factors that can be used for co-staining include FLASH, CLAMP and Mute. Colocalization of signals will not only enhance the reliability of the dCas9-gRNA system, but also confirm that the dCas9 enzyme doesn't interfere with *Drosophila* histone biosynthesis. If dCas9 does interfere with histone biogenesis, the

solution will be to design another set of gRNAs that will guide the dCas9 enzyme to the intergenic region instead of the H3/H4 promoter.

After the above validation experiments, our ultimate goal is to perform cross linking immunoprecipitation and mass spectrometry on embryos from the control and the experimental groups. Novel HLB candidates will be those factors that are significantly more enriched at the H3/H4 promoter. Proteins that result from the unspecific hitting will be eliminated by comparing proteins that are present at both the H3/H4 promoter and the 5s rDNA locus. Known HLB factors (e.g., CLAMP, FLASH, Mxc) will be used as the positive control for the mass spectrometry experiment and we expect to observe significant enrichment of many of these factors at the H3/H4 promoter. Overall, this novel protein identification method will allow us to obtain a comprehensive view on all proteins at the histone locus in *Drosophila melanogaster* during the early embryonic development stage. Identifying unknown proteins at this locus will enhance our understanding of the initiation of the HLB and provide new insight on coordinated gene expression in *Drosophila* and other animals.

Chapter 3: Identification of SIN3 as an HLB candidate by a bioinformatics-based protein candidate approach

Introduction

Another approach I used is a bioinformatics-based protein candidate approach. I utilized the online bioinformatics platform Galaxy[16] to map publicly available ChIP-seq datasets to a single copy of the *Drosophila* histone array[17]. Although there are over 100 copies of the histone array at the histone locus, they are all nearly identical in sequence[4]. Thus, signal over a single array represents the cumulative signal from all 100 copies of the array. Then, I visualized the mapping results in IGV, the Integrative Genomics Viewer[18] (Fig. 8). For a protein to be a candidate HLB factor, I expect to see substantial and steady enrichment of alignment signals at the single copy histone locus in IGV.

To identify protein candidates for bioinformatic analysis, I searched primary literature as well as the protein-protein interaction database STRING[19]. I first identified the two isoforms of the global regulatory factor, SIN3 complex, as candidates. The multisubunit SIN3 complex is a global transcription factor in *Drosophila melanogaster.* The single Sin3A gene encodes different isoforms of the complex, of which SIN187 and SIN220 as the two major isoforms[20]. Previous studies suggested structural similarities and physical interactions between the SIN3 complex and proteins of the myc family[21,22]. Myc is a known HLB factor responsible for the initiation of the histone transcription[23]. These observations led me to hypothesize that SIN3 might localize to the histone locus.

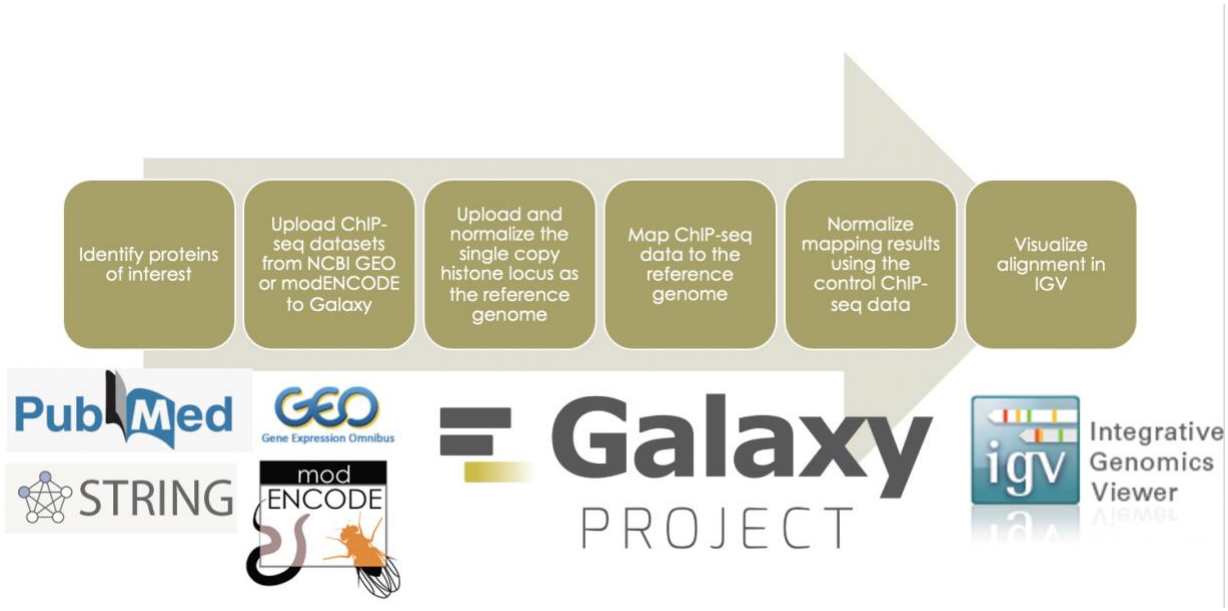


Figure 8. A graphical workflow of the Galaxy-based bioinformatics protein candidate method.

Methods

Obtained ChIP-seq datasets for SIN187 and SIN220 from the NCBI GEO database.

I obtained ChIP-seq data for SIN187 and SIN220 from the study by Saha *et al*. 2016. Researchers performed ChIP-seq experiments in non-synchronized *Drosophila* S2 cells cultured in *Drosophila* Schneider's media (1X) + L-glutamine supplemented with 10 % heat-inactivated fetal bovine serum (Invitrogen) . Chromatin incubated with pre-immune IgG was used as the non-specific ChIP control[20]. I found ChIP-seq data from the experiments on the NCBI Gene Expression Omnibus with the accession number GSE72171.

Uploaded Chip-seq datasets to the Galaxy web platform, and analyzed datasets using the public server at usegalaxy.org

I processed, analyzed, and mapped all ChIP-seq data to the *Drosophila melanogaster* histone array using Galaxy, an open source, web-based platform for data intensive biomedical research. After uploading ChIP-seq datasets to the Galaxy web platform, I analyzed ChIP-seq data following a workflow that was designed by Dr. Skye Comstra. Tools used in the workflow are introduced below.

Extracted Chip-seq datasets to the Galaxy web platform through the Faster Download and Extract Reads in FASTQ format from NCBI SRA function in Galaxy

I first added all ChIP-seq data from GEO to Galaxy through the SRA Run Selector. After adding data to the Galaxy web platform, I extracted ChIP-seq data in fastq format from the SRA list based on the SRA Accession Number for each ChIP-seq sample. The version of the Faster Download function I used in the study was Galaxy Version 2.10.8 + galaxy 0.

Quality checked extracted FastQ data through the FASTQC function in Galaxy

The FASTQC function was used as a quality control for ChIP-seq data that was generated by others. Two outputs were produced, one being a webpage of quality reports and the other being the raw quality check data. A summary of sequence quality was provided. Users can also be informed whether the sequence has been trimmed.

Uploaded a single copy of the 5kb histone array to Galaxy as the reference genome and normalized through the NormalizeFAST function

I used a single copy of the 5kb histone array as the custom reference genome. The custom genome was adapted from the single copy histone locus from Mckay *et al*. 2015 paper[17]. I uploaded the custom genome to Galaxy and normalized it using NormalizeFAST so that the histone locus was compatible for the downstream alignment step. In Galaxy, I set the line length used for output to 80.

Mapped the ChIP-seq data of the two SIN3 isoforms to the custom reference genome through Bowtie2 in Galaxy

I mapped the single-end ChIP-seq data to the single copy histone array through Bowtie2. The output files were in BAM format. The mapping of both single-end and paired-end data against a reference genome is supported. To maximize mapping accuracy, the mapping was conducted based on the very sensitive, end-to end, preset. I used the Galaxy version 2.3.4.3 + galaxy 0 of Bowtie 2 alignment in the study.

Normalized ChIP-seq alignment data by subtracting IgG control from the alignment of the experimental groups through BAMCompare in Galaxy

Two BAM files are needed for the normalization of the alignment data of the experimental group through BAMCompare. I set the bin size to 1. The output was a bigwig file that could be further visualized in the Integrative Genomics Viewer (IGV).

Visualized the alignment of ChIP-seq data to the single copy histone locus in the Integrative Genomics Viewer

The Integrative Genomics Viewer (IGV) is a free, high-performance, easy-to-use, interactive tool for the visual exploration of genomic data. I downloaded a local IGV software for this study. I downloaded BAMCompare results from Galaxy and directly imported them into IGV. To visualize mapped ChIP-seq data to the histone array, I first loaded the custom genome into IGV through the Load Genome From File function under Genome. I also needed a physical representation of the five histone genes (as the features of the histone array) and loaded it into IGV through the Load From File function under File. The feature file was a bed. file. After loading the mapping results of the protein of interest, the histone genome and the features, I could see whether SIN3 isoforms were enriched at the histone locus. I was able to adjust the height of the track of each mapping result, as well as the color and the scale of each track.

Results

ChIP-seq datasets showed enrichment of two major SIN3 isoforms components at the histone locus in Drosophila melanogaster

I mapped ChIP-seq datasets of the two major isoforms of the SIN3 complex, SIN187 and SIN220, to the single copy of the histone array. ChIP-seq profiles of the two SIN3 isoforms were normalized against the ChIP-seq profile of pre-immuned IgG chromatins as control. Alignment at the single copy histone locus results was visualized in IGV. Mapping results suggested enrichment of both isoforms at both the H3/H4 and the H2a/H2b promoters. Such observation supported the hypothesis that SIN3 localizes at the histone locus (Fig. 9).

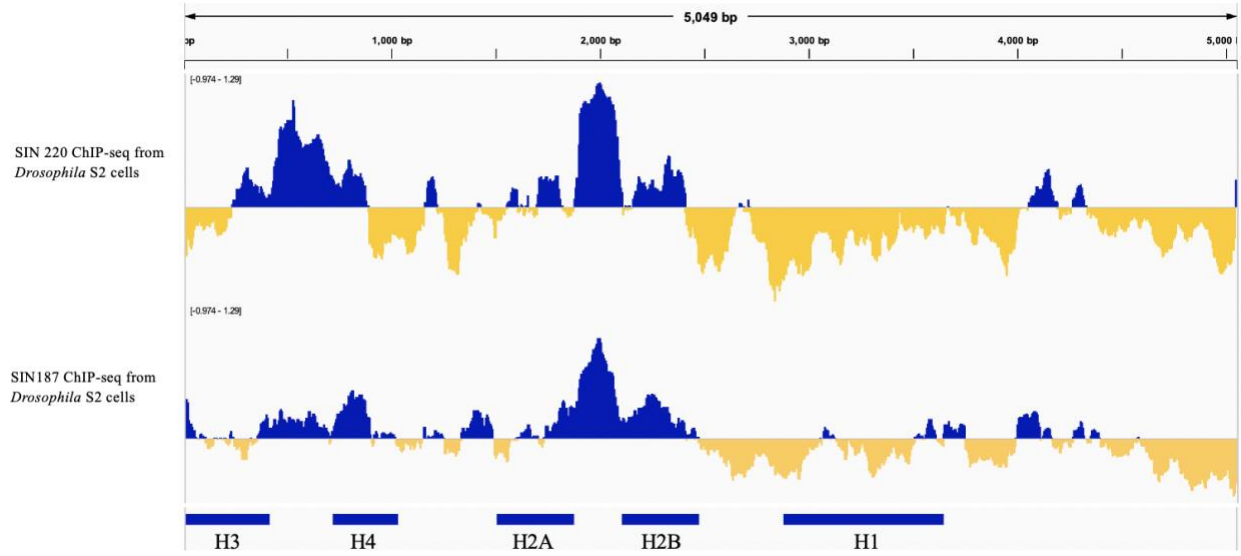


Figure 9. ChIP-seq profiles of two major isoforms of the SIN3 complex, SIN187 and SIN220, at a single histone array in *Drosophila* S2 cells.
Binding of SIN187 and SIN220 are depicted as tracks on the integrated genomic viewer (IGV). The five replication dependent histone genes are shown in solid blue bars at the bottom of the figure. Both SIN187 and SIN220 are enriched at the H3/H4 promoter and the H2a/H2b promoter.

Discussion

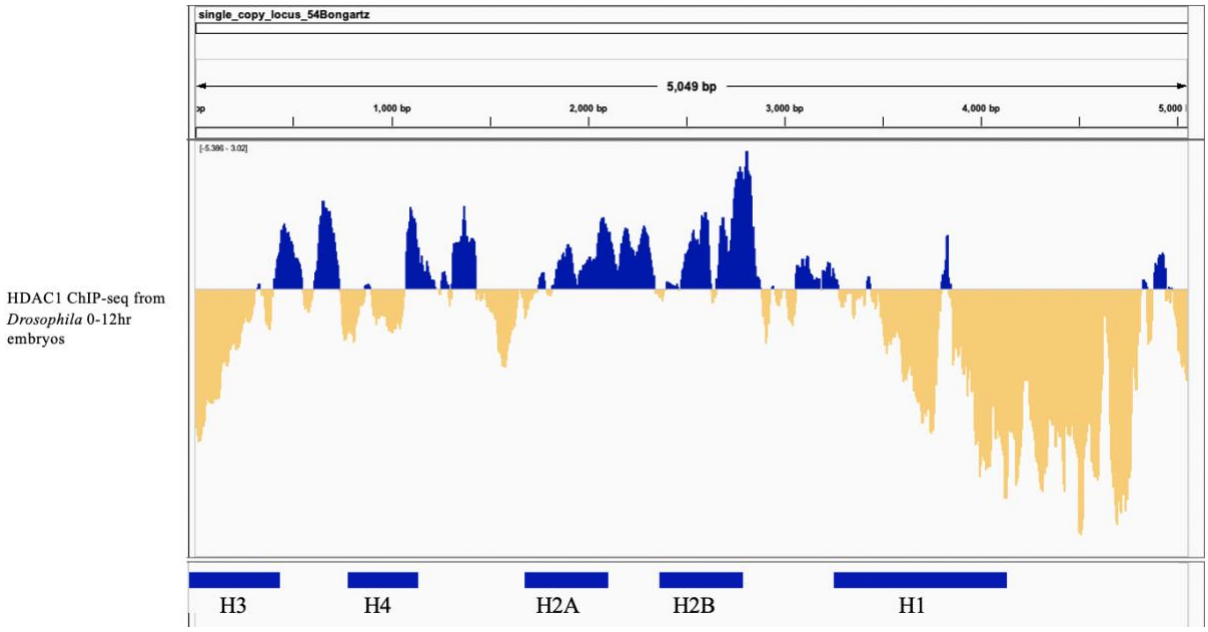
The multisubunit SIN3 complex was first identified as a global transcription factor in *Drosophila melanogaster*[24]. Previous studies have identified SIN3 as the master transcriptional adapter protein that interacts with the deacetylase RPD3 and other accessory proteins[25]. The dataset from Saha *et al*. (2016)[20] used for this analysis is the first dataset that included ChIP-seq data for individual isoforms of the SIN3 complex, namely, SIN187 and SIN220. SIN3 220 is the predominantly expressed isoform in proliferative cells such as larval imaginal disc cells, whereas SIN3 187 is the prevalent isoform during the latter stages of embryonic development and in adults[26].

Mapping results visualized in IgV showed enrichment of both isoforms at the H3/H4 promoter and the H2a/H2b promoter (Fig. 9), suggesting a binding pattern of both isoforms at the promoters of the histone locus.

To further test the hypothesis that SIN3 localizes at the histone array, wet-lab experiments need to be done. Co-staining *Drosophila* embryos at nuclear cycle 11 using the available antibodies for the two isoforms and antibodies for a known HLB factor like Mxc and FLASH can validate the colocalization of the SIN3 complex with the HLB.

The SIN3 complex interacts with proteins involved in histone post-translational modifications. The study by Kadamb *et al.* (2013) suggested the complex role as the scaffold for the Rpd3/HDAC1 histone deacetylase complex[27]. However, my mapping results in Galaxy failed to

show an enrichment of the HDAC1 enzyme anywhere at the histone locus (Sup Fig. 1).



Supplementary Figure 1. ChIP-seq profiles of HDAC at a single histone array in *Drosophila* embryos.
Transient interactions of HDAC1 with histone proteins in 0-12hr embryos are depicted as tracks on the integrated genomic viewer (IGV). Datasets were aligned to the single copy of the histone array at the bottom. The five replication dependent histone genes are shown in solid blue bars at the bottom. ChIP-seq datasets were obtained from the Drosophila modENCODE project on GEO with the Accession Number of GSE20000. Alignment of HDAC1 to the single copy histone locus was normalized using an input dataset as the control. No obvious pattern for the interaction of HDAC1 at the histone protein was shown. Thus, the mapping results do not support the hypothesis that the HDAC1 localizes at the histone locus in *Drosophila melanogaster*.

Chapter 4: Mapping of ChIP-seq data in Galaxy fails to show localization of the gypsy insulator complex or the M1BP insulator to the histone array

Introduction

Besides SIN3, another group of protein candidates at the histone locus are chromatin insulators, proteins that establish higher order-independent DNA domains to influence transcriptional regulation. They can block communication between an enhancer and a promoter and also act as a barrier between heterochromatin and euchromatin. In *Drosophila melanogaster*, CLAMP, a DNA-binding factor that targets the H3/H4 promoter in the histone array, colocalizes with cp190, Su(hw), and Mod(mdg4), three protein components of the *gypsy* insulator complex[28]. Thus, I was prompted to hypothesize that members of the gypsy insulator complex localize to the histone array. Another insulator that I looked at is M1BP (Motif-1 binding protein), which colocalizes with CP190 (Bag *et al*. 2020, preprint).

Methods

Mapped ChIP-seq datasets were to a single copy of the histone array

I mapped ChIP-seq datasets to the single copy of the histone locus of interest in the online bioinformatics platform Galaxy and visualized in IGV as introduced in the Methods section of Chapter 3.

Obtained ChIP-seq datasets for CLAMP, cp190, Su(hw), and mod(mdg4) from the NCBI GEO database

I obtained ChIP-seq data for CLAMP, cp190, Su(hw), and mod(mdg4) from the study by Bag *et al*. 2017[28]. ChIP-seq experiments were performed in Drosophila Kc167 cell lines. Data from the experiments can be found on the NCBI Gene Expression Omnibus with the accession number GSE103601.

Obtained ChIP-seq datasets for M1BP from the NCBI GEO database.

I obtained ChIP-seq data for MIBP from the study by Bag *et al*. 2020. ChIP-seq experiments were performed in Drosophila Kc167 cell lines. Data from the experiments can be found on the NCBI Gene Expression Omnibus with the accession number GSE142531.

Results

ChIP-seq datasets failed to show localization of any gypsy complex components to the histone array in Drosophila melanogaster

I mapped ChIP-seq datasets of CLAMP, cp190, Mod(mdg4), and Su(hw) to the single copy of the histone array and the results were visualized in IGV. ChIP-seq data of each protein was normalized by an ChIP pre-immune input as the control. Since CLAMP is a known HLB component, I considered the mapping results of CLAMP as the positive control. Visualization in IGV showed enrichment of CLAMP specific to the H3/H4 promoter, agreeing with previous reports[5,29]. However, none of the three components of the gypsy insulator complex showed enrichment signals that are stronger than background or consistent at both H3/H4 and

H2a/H2b promoters. While cp190 showed some positive enrichment at the H2a/H2b promoter, the amplitude of the peak was not big enough to be considered as a valid localization compared with the enrichment signals of CLAMP from the same study. Mapping results of Mod(mdg4) showed a slight depletion of the protein at the H3/H4 promoter (Fig. 10).

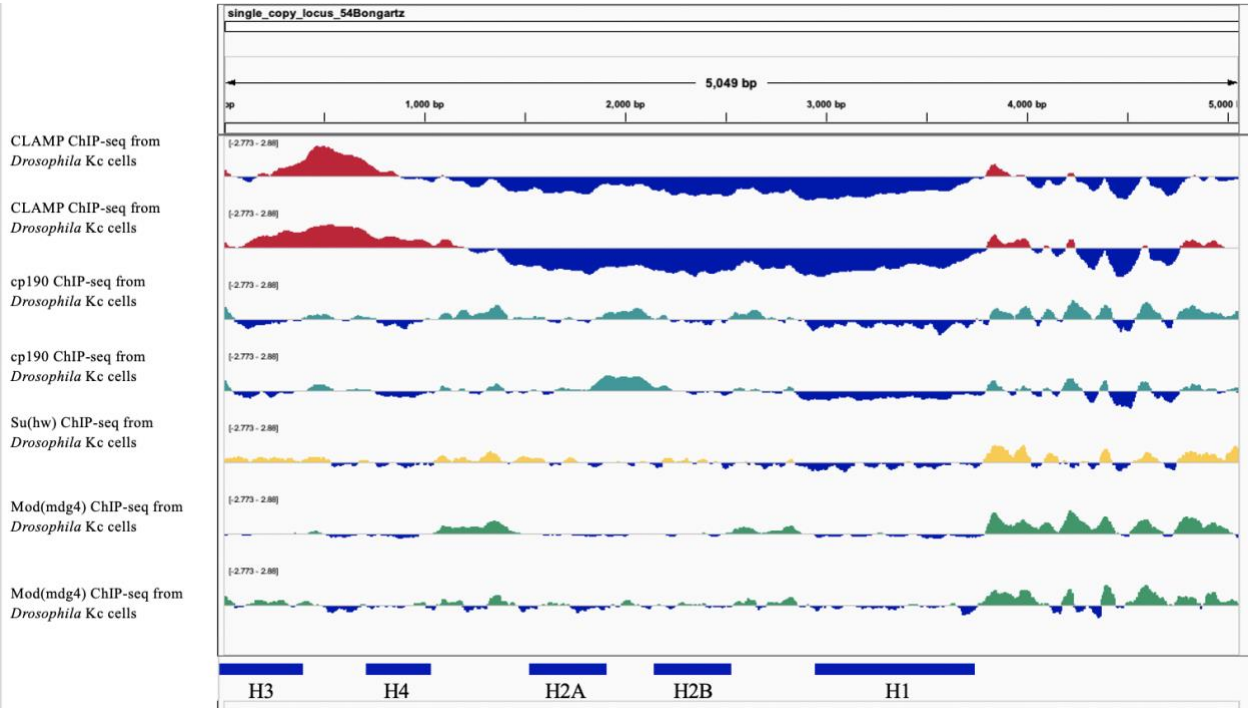


Figure 10. ChIP-seq profiles of CLAMP, cp190, Su(hw), and Mod(mdg4) over a single histone array in *Drosophila* Kc cells.
Binding of CLAMP and the three insulators are depicted as tracks on the integrated genomic viewer (IGV). Data were aligned to the single copy of the histone array at the bottom. The five replication dependent histone genes are shown in solid blue bars at the bottom. CLAMP (top 2 tracks) is enriched at the H3/H4 promoter, as previously reported, but little enrichment was observed for the three insulators (each indicated by a different color). Alignment results failed to support the localization of the gypsy insulator complex at the histone locus in *Drosophila* embryos.

ChIP-seq datasets showed depletion of M1BP at the histone array in Drosophila melanogaster

ChIP-seq datasets of M1BP in Kc 167 cells were mapped to the single copy of the histone array and the results were visualized in IGV. ChIP-seq alignment data of M1B was normalized by an

ChIP pre-immune input as the control. Mapping results showed depletion of M1BP at both H3/H4 and H2a/H2b promoters at the histone locus (Fig. 11).

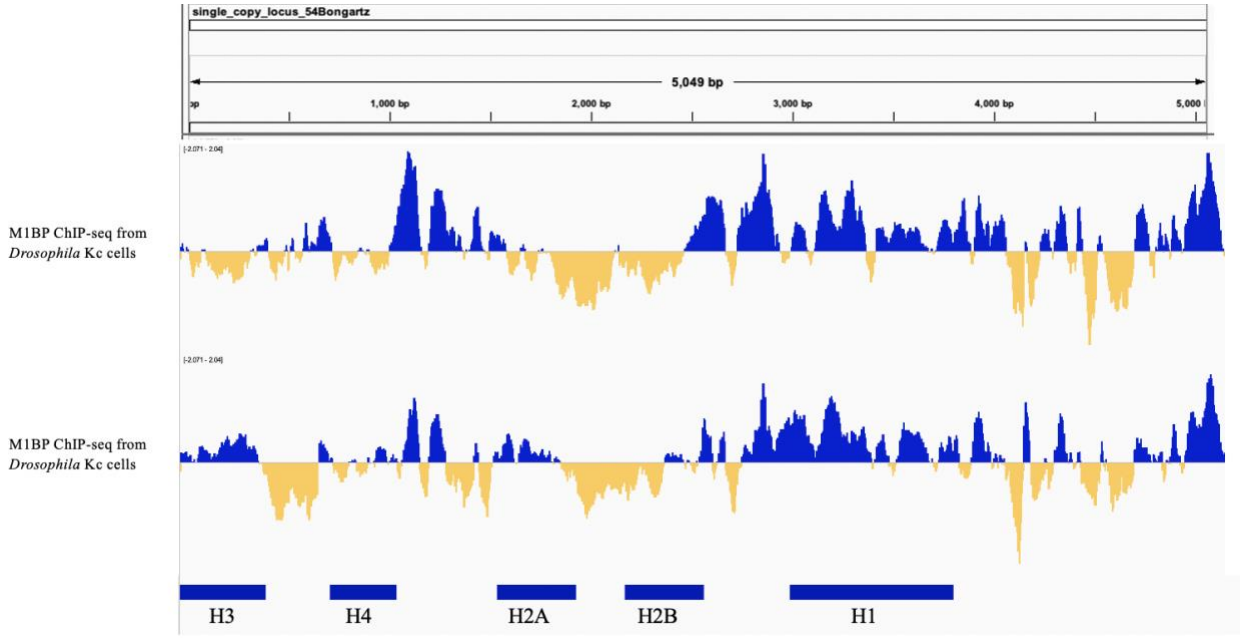


Figure 11. ChIP-seq profiles of M1BP over a single histone array in *Drosophila* Kc cells. Binding of all four insulators are depicted as tracks on the integrated genomic viewer (IGV). Data were aligned to the single copy of the histone array at the bottom. The five replication dependent histone genes are shown in solid blue bars at the bottom. M1BP was depleted at both H3/H4 and H2a/H2b promoters.

Discussion

Insulators are proteins that can modulate eukaryotic gene expression by affecting higher-order chromatin structure and regulating interactions between *cis*-regulatory elements and promoters. While the study by Bag *et al* (2017) suggested physical interactions of all three components of the gypsy insulator complex with CLAMP through Co-IP[28], mapping results (Figs. 10-11) failed to show enrichment of signals of any of the three components over the histone array. This absence of enrichment at the histone locus can be explained due to the ubiquitous nature of CLAMP as a zinc-finger regulatory factor that binds across the *Drosophila* genome.

CLAMP is known to be enriched at the evolutionary conserved GA repeats on the *Drosophila* genome[30]. One of such locations is on the male X-chromosome where CLAMP recruits the dosage compensation complex MSL[11]. Since M1BP has only been shown to interact with cp190 (Bag *et al*. 2020, preprint), the depletion of M1BP at the histone locus becomes less surprising. However, mapping results obtained from this study do not reject the possibility of localization of the four insulators to the histone array. Since ChIP-seq data used for analysis came from experiments performed in Kc167 cells cultured *in vitro*, the rapid division activity of the cell line might not accurately represent biological activities *in vivo*. Thus, it is possible that some of the four insulators localize at the HLB in embryos but not in Kc cell lines. Another explanation can be that the ChIP-seq experiments were done in *in vitro* Kc cells that are asynchronous. Thus, enrichment of insulators at the histone locus in a subset of Kc cells could be washed out by the depletion of the same insulators in another set of the Kc cells.

Chapter 5: Discussions and Future Directions

Coordinated gene expression is one of the major ways to regulate precise spatial and temporal expression of specific genes. In *Drosophila melanogaster*, the coordinated biogenesis of the five replication dependent histone genes is facilitated by the histone locus body (HLB). To improve the efficiency of gene expression, the HLB localize regulatory factors and promote interactions among them that would otherwise be stochastic[6]. However, not every HLB factor has been identified, and new protein identification approaches are needed to achieve a more comprehensive understanding of factors of the HLB.

In this thesis, I used two novel approaches to identify protein candidates at the histone locus. The dCas9-gRNA system is less biased than the conventional protein identification method (e.g., immunoprecipitation) since the ubiquitous Cas9 enzyme, rather than a known HLB component, is used as the target during the isolation of proteins associated with the H3/H4 promoter. The bioinformatics protein candidate approach using Galaxy and IGV is innovative in that it offers an opportunity to take advantage of existing high-throughput datasets and map sequencing results to the single copy *Drosophila* histone locus.

Limitations

For the proteomic screening approach, one potential challenge is driving the expression of the Cas9 enzyme in *Drosophila* embryos. The expression of the Cas9 transgene is currently driven by a leaky heat-shock promoter (Fig 5B). However, Western blot on *Drosophila* embryos laid by

heat-shocked mothers failed to show expression of the FLAG-tagged dCas9 enzyme as supposed to. Another potential challenge might be preventing the exogenous dCas9 enzyme from interfering with the normal histone biogenesis mechanism at the histone locus. While gRNAs were designed to avoid overlapping to the TATA box in the H3/H4 promoter, whether localization of the dCas9 enzyme at proximal regions on the genome would interrupt the function of critical regulatory factors at the promoter remains unknown.

Another potential limitation with the approach is the risk of running into ascertainment bias. If localization of the inactive Cas9 enzyme does interfere with histone biogenesis, some embryos may die or be arrested during development. Since only live embryos containing both the dCas9 and the gRNA transgene will be chosen for mass spectrometry (Fig. 4), Nevertheless, this limitation will not prevent us from achieving the goal of identifying HLB novel candidates through this approach.

My bioinformatics workflow has its own limitations. First, I do not include a way to quantitatively verify peaks shown in IGV. Enrichment of signals over the histone array may be artifacts of the ChIP-seq workflow that I did not control. One solution to this problem is to map the protein of interest to another sequence it is known to bind at and compare the amplitude of the peak between that verified sequence and the signal from the histone array. Besides, the data that can be mapped to the histone array is limited by existing datasets since there is not a high-throughput dataset available for every protein with an input or ChIP-seq data with IgG as the control. Without subtracting the baseline enrichment from ChIP-seq mapping data of the

protein of interest, one will not be able to tell if enrichment at the histone locus is due to the localization of the specific protein or the abundance of background reads. In addition, since ChIP-seq data represents an average of the protein bound to the histone array over a period of time, mapping results might fail to catch transient binding (such as the transient interaction between the HDAC1 protein and histone octamers). Finally, since the histone locus used for reference only contains one copy of the histone array, I will not be able to observe variations in ChIP-seq profiles of a protein between multiple histone arrays for now.

Future directions

Much more can be done with both of the two approaches introduced in this paper. For the proteomic screening approach, to resolve the problem of driving dCas9 expression in *Drosophila* embryos, our lab has recently obtained a dCas9 transgene that is driven by nanos, a maternal germline driver in *Drosophila*, from Dr. Nicole Crown at Case Western University. We expect to see a large amount of dCas9 enzymes deposited in embryos with the new transgene and plan to validate its function through protein extraction and Western blot. If we see expression of dCas9 in *Drosophila* embryos, we will cross this strain of flies to each of the two gRNA strains. Further verification experiments can be done as described in the discussion section of Chapter 2.

Besides cross-linking proteins at the histone locus directly with the dCas9 enzyme, the dCas9 enzyme can be further integrated with a novel proximal labeling approach, which utilizes the biomolecule biotin as the “tag” to put on proteins near the histone locus[31]. With the assistance

of the exogenous biotin ligase, the biotin molecule forms covalent bonds with neighboring proteins, resulting in stronger affinity of the tag on proteins and the ability to capture transient interactions between proteins at the histone locus. The biotin ligase can be incorporated into the embryos as a transgene through a fusion protein with the dCas9. Biotinylated proteins can be precipitated using streptavidin beads and can be sent for proteomic analysis.

For the bioinformatics-based protein candidate approach, one can explore how to map other types of high-throughput datasets to the histone array. Another group of protein candidates is histone post-translational modification marks that are known to be related to the activation or the suppression of gene expression. Since massive histone transcription takes place when the HLB remains at the locus, a combination of active and repressive histone marks might also contribute to the temporal regulation of histone biogenesis. This bioinformatics approach also raises new questions. Through mapping datasets of the same protein coming from samples of different tissues at different development timepoints, one can visualize if the targeting of an HLB component is changing during development or in different tissues. Since the need for histone proteins is critical at the S/G1 phase of the cell cycle, analysis of ChIP-seq profiles of a known HLB factor at different embryonic development time points may reveal further details on how HLB formation is coupled to development. In addition, cells in different tissues divide at different paces; thus, the formation and maintenance of HLB and the localization of some of its factors may also be partly tissue-specific. While it is easier to observe enrichment signals of an HLB component in synchronous embryonic cells, it is also possible to see enrichment of the

same protein in developing wing discs, another proliferative tissue where cells are asynchronous.

Lastly, it's critical to recognize that both approaches only answer the question of whether a factor localizes at the histone locus. To be recognized as a HLB component, further experiments are needed to verify that a protein is functionally related to histone biogenesis. A verification experiment that can be done is to perform RNAi in wild type embryos, reduce the expression of an identified HLB candidate, and test if it affects histone biogenesis or HLB formation at the locus. Another approach is to generate a heterozygous mutant strain with a loss of one copy of the gene encoding for the protein of interest and then quantify the histone proteins expressed compared to that from a wildtype strain.

Regulation of histone biogenesis in *Drosophila melanogaster* is a complicated process whose mechanism has not been fully understood. I wish the two methods presented in this paper can, in the long run, contribute to the identification of novel HLB candidates and facilitate the understanding of the regulatory mechanisms of histone biogenesis.

References:

1. Marzluff WF, Wagner EJ, Duronio RJ. Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail. *Nat Rev Genet*. 2008 Nov; 9(11): 843-54. doi: 10.1038/nrg2438.

2. Duronio RJ, Marzluff WF. Coordinating cell cycle-regulated histone gene expression through assembly and function of the Histone Locus Body. *RNA Biol.* 2017 Jun 3;14(6):726-738. doi: 10.1080/15476286.2016.1265198.

3. Marzluff WF, Koreski KP. Birth and Death of Histone mRNAs. *Trends Genet*. 2017 Oct;33(10):745-759. doi: 10.1016/j.tig.2017.07.014. Epub 2017 Aug 31. PMID: 28867047; PMCID: PMC5645032.

4. Bongartz P, Schloissnig S. Deep repeat resolution-the assembly of the Drosophila Histone Complex. *Nucleic Acids Res*. 2019 Feb 20;47(3):e18. doi: 10.1093/nar/gky1194.

5. Rieder LE, Koreski KP, Boltz KA, Kuzu G, Urban JA, Bowman SK, Zeidman A, Jordan WT 3rd, Tolstorukov MY, Marzluff WF, Duronio RJ, Larschan EN. Histone locus regulation by the Drosophila dosage compensation adaptor protein CLAMP. *Genes Dev*. 2017 Jul 15;31(14):1494-1508. doi: 10.1101/gad.300855.117.

6. Tatomer DC, Terzo E, Curry KP, Salzler H, Sabath I, Zapotoczny G, McKay DJ, Dominski Z, Marzluff WF, Duronio RJ. Concentrating pre-mRNA processing factors in the histone locus body facilitates efficient histone mRNA biogenesis. *J Cell Biol*. 2016 Jun 6;213(5):557-70. doi: 10.1083/jcb.201504043.

7. Anamika K, Gyenis À, Poidevin L, Poch O, Tora L. RNA polymerase II pausing downstream of core histone genes is different from genes producing polyadenylated transcripts. *PLoS One*. 2012;7(6):e38769. doi: 10.1371/journal.pone.0038769.

8. Salzler HR, Tatomer DC, Malek PY, McDaniel SL, Orlando AN, Marzluff WF, Duronio RJ. A sequence in the Drosophila H3-H4 Promoter triggers histone locus body assembly and biosynthesis of replication-coupled histone mRNAs. *Dev Cell.* 2013 Mar 25;24(6):623-34. doi: 10.1016/j.devcel.2013.02.014.

9. White AE, Leslie ME, Calvi BR, Marzluff WF, Duronio RJ. Developmental and cell cycle regulation of the Drosophila histone locus body. *Mol Biol Cell*. 2007 Jul;18(7):2491-502. doi: 10.1091/mbc.e06-11-1033. Epub 2007 Apr 18. PMID: 17442888; PMCID: PMC1924828.

10. Prieto-Sánchez S, Moreno-Castro C, Hernández-Munain C, Suñé C. Drosophila Prp40 localizes to the histone locus body and regulates gene transcription and development. *J Cell Sci*. 2020 Apr 7;133(7):jcs239509. doi: 10.1242/jcs.239509. PMID: 32094262.

11. Larschan E, Soruco MM, Lee OK, Peng S, Bishop E, Chery J, Goebel K, Feng J, Park PJ, Kuroda MI. Identification of chromatin-associated regulators of MSL complex targeting in Drosophila dosage compensation. *PLoS Genet.* 2012;8(7):e1002830. doi: 10.1371/journal.pgen.1002830. Epub 2012 Jul 26. PMID: 22844249; PMCID: PMC3405997.

12. Tsui C, Inouye C, Levy M, Lu A, Florens L, Washburn MP, Tjian R. dCas9-targeted locus-specific protein isolation method identifies histone gene regulators. *Proc Natl Acad Sci U S A.* 2018 Mar 20;115(12):E2734-E2741. doi: 10.1073/pnas.1718844115. Epub 2018 Mar 5. PMID: 29507191; PMCID: PMC5866577.

13. Worcel A, Gargiulo G, Jessee B, Udvardy A, Louis C, Schedl P. Chromatin fine structure of the histone gene complex of Drosophila melanogaster. *Nucleic Acids Res*. 1983 Jan 25;11(2):421-39. doi: 10.1093/nar/11.2.421. PMID: 6402757; PMCID: PMC325723.

14. Westwood JT, Clos J, Wu C. Stress-induced oligomerization and chromosomal relocalization of heat-shock factor. *Nature*. 1991 Oct 31;353(6347):822-7. doi: 10.1038/353822a0. PMID: 1944557.

15. Hans S, Kaslin J, Freudenreich D, Brand M. Temporally-controlled site-specific recombination in zebrafish. *PLoS One*. 2009;4(2):e4640. doi: 10.1371/journal.pone.0004640. Epub 2009 Feb 27. PMID: 19247481; PMCID: PMC2645673.

16. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, Chilton J, Clements D, Coraor N, Grüning BA, Guerler A, Hillman-Jackson J, Hiltemann S, Jalili V, Rasche H, Soranzo N, Goecks J, Taylor J, Nekrutenko A, Blankenberg D. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res*. 2018 Jul 2;46(W1):W537-W544. doi: 10.1093/nar/gky379. PMID: 29790989; PMCID: PMC6030816.

17. McKay DJ, Klusza S, Penke TJ, Meers MP, Curry KP, McDaniel SL, Malek PY, Cooper SW, Tatomer DC, Lieb JD, Strahl BD, Duronio RJ, Matera AG. Interrogating the function of metazoan histones using engineered gene clusters. *Dev Cell.* 2015 Feb 9;32(3):373-86. doi: 10.1016/j.devcel.2014.12.025. PMID: 25669886; PMCID: PMC4385256.

18. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. *Nat Biotechnol.* 2011 Jan;29(1):24-6. doi: 10.1038/nbt.1754. PMID: 21221095; PMCID: PMC3346182.

19. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, Mering CV. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 2019 Jan 8;47(D1):D607-D613. doi: 10.1093/nar/gky1131. PMID: 30476243; PMCID: PMC6323986.

20. Saha N, Liu M, Gajan A, Pile LA. Genome-wide studies reveal novel and distinct biological pathways regulated by SIN3 isoforms. *BMC Genomics.* 2016 Feb 13;17:111. doi: 10.1186/s12864-016-2428-5. PMID: 26872827; PMCID: PMC4752761.

21. Ayer DE, Lawrence QA, Eisenman RN. Mad-Max transcriptional repression is mediated by ternary complex formation with mammalian homologs of yeast repressor Sin3. *Cell*. 1995 Mar 10;80(5):767-76. doi: 10.1016/0092-8674(95)90355-0. PMID: 7889570.

22. Schreiber-Agus N, Chin L, Chen K, Torres R, Rao G, Guida P, Skoultchi AI, DePinho RA. An amino-terminal domain of Mxi1 mediates anti-Myc oncogenic activity and interacts with a homolog of the yeast transcriptional repressor SIN3. *Cell*. 1995 Mar 10;80(5):777-86. doi: 10.1016/0092-8674(95)90356-9. PMID: 7889571.

23. Daneshvar K, Khan A, Goodliffe JM. Myc localizes to histone locus bodies during replication in Drosophila. *PLoS One*. 2011;6(8):e23928. doi: 10.1371/journal.pone.0023928. Epub 2011 Aug 23. PMID: 21886841; PMCID: PMC3160328.

24. Silverstein RA, Ekwall K. Sin3: a flexible regulator of global gene expression and genome stability. Curr Genet. 2005 Jan;47(1):1-17. doi: 10.1007/s00294-004-0541-5. Epub 2004 Nov 23. PMID: 15565322.

25. Grzenda A, Lomberk G, Zhang JS, Urrutia R. Sin3: master scaffold and transcriptional corepressor. *Biochem Biophys Acta*. 2009 Jun-Aug;1789(6-8):443-50. doi: 10.1016/j.bbagrm.2009.05.007. Epub 2009 Jun 6. PMID: 19505602; PMCID: PMC3686104.

26. Sharma V, Swaminathan A, Bao R, Pile LA. Drosophila SIN3 is required at multiple stages of development. *Dev Dyn*. 2008 Oct;237(10):3040-50. doi: 10.1002/dvdy.21706. PMID: 18816856.

27. Kadamb R, Mittal S, Bansal N, Batra H, Saluja D. Sin3: insight into its transcription regulatory functions. *Eur J Cell Biol*. 2013 Aug-Sep;92(8-9):237-46. doi: 10.1016/j.ejcb.2013.09.001. Epub 2013 Oct 9. PMID: 24189169.

28. Bag I, Dale RK, Palmer C, Lei EP. The zinc-finger protein CLAMP promotes gypsy chromatin insulator function in Drosophila. *J Cell Sci.* 2019 Mar 8;132(5):jcs226092. doi: 10.1242/jcs.226092. Erratum in: *J Cell Sci.* 2019 Mar 26;132(6): PMID: 30718365; PMCID: PMC6432716.

29. Koreski KP, Rieder LE, McLain LM, Chaubal A, Marzluff WF, Duronio RJ. *Drosophila* histone locus body assembly and function involves multiple interactions. *Mol Biol Cell*. 2020 Jul 1;31(14):1525-1537. doi: 10.1091/mbc.E20-03-0176. Epub 2020 May 13. PMID: 32401666; PMCID: PMC7359574.

30. 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

31. Branon TC, Bosch JA, Sanchez AD, Udeshi ND, Svinkina T, Carr SA, Feldman JL, Perrimon N, Ting AY. Efficient proximity labeling in living cells and organisms with TurboID. *Nat Biotechnol.* 2018 Oct;36(9):880-887. doi: 10.1038/nbt.4201. Epub 2018 Aug 20. Erratum in: Nat Biotechnol. 2020 Jan;38(1):108. PMID: 30125270; PMCID: PMC6126969.