Distribution Agreement

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

Signature:

_____ _____

Donald Bryant Date

An Essential Role for MYST5 (CG1894) in Histone H4 Acetylation and RNA Polymerase II Pause Release

By

Donald Bryant
Master of Science

Biology

_____
Victor G. Corces, Ph.D.
Advisor

_____
Ronald Calabrese, Ph.D.
Committee Member

_____
John Lucchesi, Ph.D.
Committee Member

_____
Matthew Weinschenk, Ph.D.
Committee Member

Accepted:

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

_____
Date

An Essential Role for MYST5 (CG1894) in Histone H4 Acetylation and RNA Polymerase II Pause Release

By

Donald Bryant

Advisor: Victor G. Corces, Ph.D.

An abstract of
A thesis 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
Master of Science in Biology

2013

Abstract

The post-translational modification of histone tails is critical for transcriptional regulation in eukaryotes. One major type of covalent modification shown to be important for transcriptional regulation is the acetylation of lysine residues on the N-terminal tails of histones, and the enzymes that carry out this reaction are referred to as histone acetyltransferases (HATs). Here, I demonstrate that the *D. melanogaster* protein CG1894 (MYST5) is a member of the MYST (MOZ, Ybf2/Sas3, Sas2, TIP60) family of histone acetyltransferases and is essential for the acetylation of lysines 5 and 8 on the N-terminal tail of histone H4 (H4K5 and H4K8). Through genome-wide mapping studies, I show that MYST5 binding correlates with active transcription and that MYST5 binds the promoters of a subset of genes that are highly transcriptionally paused and enriched with a "stalled" RNA polymerase II. Moreover, I demonstrate that MYST5 colocalizes with *Drosophila* Myc, Mof (males absent on the first), and GAF (GAGA associated factor), which have all been implicated in the regulation of polymerase pausing. Finally, I show that loss of MYST5 results in a reduction in the levels of elongating polymerase. Collectively, these data indicate that MYST5 plays an active role in transcription by contributing to the regulation of RNA polymerase II pause release.

An Essential Role for MYST5 (CG1894) in Histone H4 Acetylation and RNA Polymerase II Pause Release

By

Donald Bryant

Advisor: Victor G. Corces, Ph.D.

A thesis 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
Master of Science in Biology

2013

Acknowledgements

I would like to thank Dr. Victor Corces, my advisor and mentor, for his outstanding guidance, mentorship, training, and support during my four years as a student in his lab. I truly would not be the scientist that I am without his help and advisement. I would also like to thank Dr. John Lucchesi, Dr. Ronald Calabrese, and Dr. Matthew Weinschenk for their helpful advice and for their willingness to take the time to be on my committee. I would also like to thank Yaeji Park in particular for her active role in this project, help with experiments, and for sharing data for this project. I want to thank Naomi Takenaka for my initial training in the Corces lab, her help with experiments, advice, and for sharing data used for this research. Similarly, I want to thank Dr. Jingping Yang for her extensive help with my bioinformatic analyses, willingness to answer all of my questions, and sharing data for this project. I want to thank Dr. Li Li for his statistical advice and help with my bioinformatic analyses. I also would like to thank Kevin Van Bortle for his comments on my work, his advice, and for contributing data used in this project's analyses. I want to extend my gratitude to all members of the Corces lab, both current and former, for making my lab experience nothing short of amazing. Finally, I want to thank Sevara Rakhimova and my family for standing by my side throughout my career and for their steady interest in my research.

This thesis is dedicated in loving memory to my mother

Pamela Bryant

(February 2nd, 1970 – December 2nd, 2010)

Table of Contents

Table of Figures

Introduction

Eukaryotic genomes are tightly packed into the nucleus within a compact and complex protein-DNA structure referred to as chromatin. Although chromatin allows an organism to efficiently package its genetic information, decades of research have revealed that the functionality of chromatin structure extends beyond simple storage and plays a crucial role in regulating the expression of an organism's genes. The fundamental unit of chromatin is the nucleosome, which is comprised of DNA wound around proteins named histones. In order to properly modulate gene expression at the transcriptional level in a dynamic environment and during complex biological processes such as those seen during development, eukaryotes need to precisely regulate interactions between the transcriptional machinery, histones, and DNA.

In eukaryotes, a basic model of transcription can be viewed as having a series of fundamental stages (Nechaev and Adelman 2011) (Figure 1). The first part of transcription involves processes that allow RNA polymerase II (RNA pol II) to bind the promoter of a gene and initiate transcription (Figure 1 Stage 1). In this stage, transcription factors and other proteins such as GAGA associated factor (GAF) bind near the transcription start site of a gene and recruit chromatin remodeling complexes, leading to the eviction of histones from the promoter region (Chen, Li et al. 1994, Tsukiyama, Becker et al. 1994, Cosma, Tanaka et al. 1999, Peterson and Workman 2000, Boehm, Saunders et al. 2003, Boeger, Griesenbeck et al. 2004). The transcription factors also facilitate the recruitment of RNA pol II to the promoter region (Cosma, Tanaka et al.

1999, Nechaev and Adelman 2011). The list of factors involved in the initiation of transcription and recruitment of polymerase continues to grow and includes the TATA binding protein (TBP), transcription factors IIh and IIe (TFIIH, & TFIIE), Mediator, and GAF (Imbalzano, Kwon et al. 1994, Ohkuma 1997, Wilkins and Lis 1997, Esnault, Ghavi-Helm et al. 2008). Once recruited, RNA pol II is phosphorylated at the serine 5 residue of its C-terminal tail by TFIIH and initiates transcription elongation (Akoulitchev, Makela et al. 1995, Nechaev and Adelman 2011).

In the next stage of transcription (Figure 1 Stage 2), RNA pol II associates with negative elongation factor (NELF) and DRB sensitivity inducing factor (DSIF) and halts downstream of the transcription start site after transcribing approximately 25 – 50 base pairs (bp) of DNA (Wu, Yamaguchi et al. 2003). In stage three (Figure 1 Stage 3) an intricate series of events take place, culminating in the recruitment of the positive transcription elongation factor b (P-TEFb). Once recruited, P-TEFb phosphorylates RNA pol II at its serine 2 residue, releasing it from its paused status (Lis, Mason et al. 2000, Ni, Saunders et al. 2008). After it is released from its stalled state, RNA pol II proceeds into productive transcription elongation and completes transcription of the gene (Figure 1 Stage 4).

The events leading up to polymerase pause release often include the covalent modifications of nucleosomal barriers near the pausing event (Kireeva, Hancock et al. 2005, Bondarenko, Steele et al. 2006, Carey, Li et al. 2006). For example, *Ivaldi et al* discovered that the phosphorylation of histone H3 at its serine 10 residue (H3S10ph) by the Jil-1 kinase is necessary for P-TEFb recruitment and pause release (Ivaldi, Karam et

al. 2007). An extension of these findings by *Karam et al* revealed that the acetylation of histone H3 at lysine 9 (H3K9) also occurred during polymerase pausing and was dependent on H3S10 phosphorylation (Karam, Kellner et al. 2010). Moreover, it was found that the histone aceyltransferase elongator protein 3 (Elp3) was recruited by 14-3-3 to acetylate H3K9 (Karam, Kellner et al. 2010).

Although there are several ways in which P-TEFb can be recruited during pausing to facilitate polymerase pause release, two common mechanisms by which this occurs are through recruitment by factors that bind chromatin and recruitment by additional transcription factors (Peterlin and Price 2006). An example by which P-TEFb is recruited by a chromatin-binding factor is illustrated in a recent study using human cell lines and induced gene expression. In human 293 cells, the acetylation of histone H4 at lysine 16 (H4K16) by males absent on the first (Mof) and H3K9 acetylation were found to occur downstream of H3S10 phosphorylation during serum induced gene activation (Zippo, Serafini et al. 2009). In turn, this study discovered that the bromodomain containing 4 (BRD4) protein bound to the acetylated H4K16 and H3K9 and recruited P-TEFb to facilitate polymerase pause release. In contrast to this mechanism of recruitment, other studies have found that P-TEFb could be recruited to regions of active transcription by transcription factors like c-Myc (Eberhardy and Farnham 2002). Taken together, these findings highlight the complexity of the pausing event as well as important roles for chromatin in regulating eukaryotic transcription.

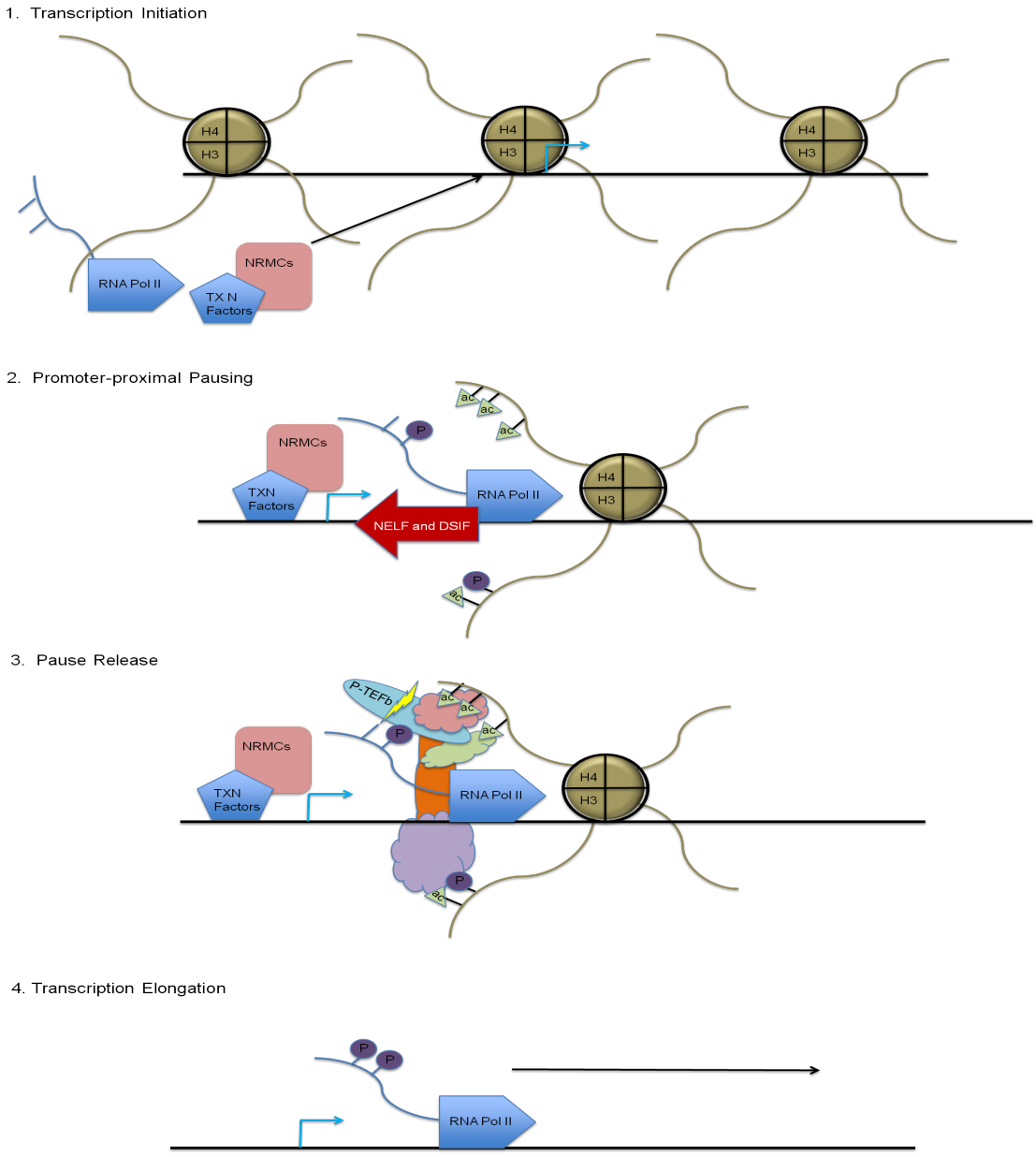


Figure 1 - A basic view of eukaryotic transcription
Multicellular eukaryotic transcription can be viewed as a process having four major stages, each involving several proteins and intricate regulatory mechanisms: (1) Transcription initiation, which involves general transcription factors (TXN Factors) such as TFIIH, nucleosome remodeling complexes (NRMCs), and histones (i.e. H3 and H4), (2) Polymerase pausing downstream of the transcription start site (blue arrow), which involves factors that establish the pause (e.g. NELF and DSIF), (3) Pausing release, which involves factors such as BRD4 that help recruit P-TEFb to the pausing site, and (4) Transcriptional elongation. Examples of histone modifications during this process are shown for histones H3 and H4. The above model is a simplified version of that described in (Nechaev and Adelman 2011).

The covalent modification of the N-terminal tails of histones is an important mechanism used to regulate eukaryotic transcription and to fine tune chromatin structure. The acetylation of lysine residues on the tails of histones is one of the most well characterized forms of post-translational histone modification, and numerous studies have indicated that this modification is strongly correlated with active gene transcription (Clayton, Hazzalin et al. 2006). Currently, it is believed that histone acetylation can influence transcription in two primary ways. First, the addition of an acetyl group to lysines can reduce the overall positive charge of histones, which can result in conformational changes of the histone tail as well as influence its interactions with the negatively charged DNA. This charge neutralization could lessen the attractions between the histones and DNA, allowing for a more open chromatin conformational state that favors transcription. Secondly, transcriptional regulators possessing bromodomains, such as BRD4, that bind the acetylated lysines of histone tails can be recruited to nucleosomes during different stages of transcription (Figure 1 Stage 3). Once bound to acetylated histones, these proteins can recruit factors like P-TEFb that further regulate transcription (Yang, Yik et al. 2005).

The enzymes that carry out the acetylation of the histone tails are collectively referred to as histone acetyltransferases and constitute a remarkably diverse class of enzymes that play significant roles in transcriptional regulation (Roth, Denu et al. 2001). For instance, dosage compensation in *D. melanogaster* is heavily dependent on the acetylation of histone H4 at lysine 16 (H4K16) by Mof, a member of the MYST family of histone acetyltransferases (Hilfiker, HilfikerKleiner et al. 1997). In this system, the

activity of Mof allows male fruit flies to transcribe their single X chromosome at a higher rate such that gene expression is similar to that in female fruit flies, which actively transcribe both of their X chromosomes.

The importance of histone acetylation in regulating key transcriptional events was highlighted even further in a recent study showing that blocking the ability of BRD4 to bind acetylated histones disrupted levels of RNA pol II phosphorylated at serine 2 at enhancers. Specifically, this study revealed that the small molecule inhibitor JQ1 disrupted BRD4's ability to bind acetylated H4K5 and H4K8 in human CD4+ T-cells (Zhang, Prakash et al. 2012). In turn, this led to a reduction in RNA pol II phosphorylated at its serine 2 residue at specific enhancers as well as a reduction in the expression of T-cell lineage specific genes (Zhang, Prakash et al. 2012). These findings indicate an intricate role for histone acetyltransferases in regulating transcription.

Given the importance of histone acetyltransferases in regulating transcription at multiple levels I investigated whether or not CG1894 (MYST5), a putative histone acetyltransferase, also plays a role in transcriptional regulation. By performing analyses of MYST5's binding on a genome-wide scale, I discovered that nearly half of MYST5 sites localized to the promoters of highly paused and highly expressed genes. I also found that MYST5 associates with GAF, Myc, and Mof, which have been shown to function in regulating polymerase pausing. I show that MYST5 functions downstream of the Jil-1 kinase and is essential for H4K5 and H4K8 acetylation. Finally, I demonstrate that MYST5 is critical for proper polymerase pause release. Taken together, these data

strongly support a role for MYST5 in regulating polymerase pause release during transcription on a genome-wide scale.

Materials and Methods

Fly Stocks

All flies were maintained at 25°C, and the Oregon R (OR) strain of flies was used for wild type controls in all experiments. Flies containing the *Jil-1* null allele (*Jil-1*^{z2}) were a kind gift from Dr. Kristen Johansen, and flies harboring a p-element transformation in the *CG1894* (*CG1894*^{f06204}) gene were obtained from the Bloomington Stock Center at Indiana University.

Antibodies

The CG1894 antibody used in all experiments was raised in guinea pigs against amino acids 36 – 176 of the CG1894 protein. The α-rat SuHw antibody used in immunofluorescence experiments was previously generated by members of the Corces lab. Two α-rabbit antibodies that recognize both isoforms of Fs(1)h as well as the long isoform of Fs(1)h were a kind gift from Dr. Igor Dawid at the National Institutes of Health. The α-rabbit Histone H3 (abcam GR25302-1), α-rabbit H4K8 acetyl (abcam 760872), α-rabbit H4K16 acetyl (Millipore DAM1675759), α-rabbit H3K14 acetyl (Upstate 07-353), α-rabbit H3K9 acetyl (Millipore DAM1767487), α-rabbit H4K12 acetyl (06-761), α-rabbit H4K5 acetyl (06-759), anti-Pol II ser 2 phospho Mouse IgM (Covance, H5), and anti-Pol II ser 5 phospho Mouse IgM (Covance, H14) were purchased and used as described elsewhere in these methods.

Western Analyses

Wandering 3rd instar larvae were washed three times in 1X PBS and homogenized in Laemmli Sample Buffer supplemented with 1mM PMSF, 0.1 M DTT, and phosphatase inhibitors (10 μL per 5 mg protein). Western blotting was performed as in (Karam, Kellner et al. 2010). Antibodies used in these experiments included guinea pig anti-CG1894 (1:500), α-rabbit H3 (1:10,000), α-rabbit H4K12 acetyl (1:1000), α-rabbit H3S10 phospho (1:1000), α-rabbit H4K8 acetyl (1:1000), α-rabbit H3K14 acetyl (1:1500), α-rabbit H4K5 acetyl (1:1000), α-rabbit H4K16 acetyl (1:500), Mouse IgM anti-Pol II ser 2 phospho (1:500), and Mouse IgM anti-Pol II ser 5 phospho (1:500). Images for H4K5 acetyl western blots had high background and needed adjustment; the contrast and brightness were adjusted for whole images with Microsoft Powerpoint.

Polytene Chromosome and Immunofluorescence Analyses

The salivary glands from wandering 3rd instar larvae were isolated in sterile saline and fixed for 3 minutes in 1X PBS/1% Triton X-100/ 3.2 % paraformaldehyde. Glands were then fixed for 3 additional minutes in 50% acetic acid/3.2% paraformaldehye. Glands were then squashed using standard procedures. Immunofluorescence analysis of salivary glands was performed as described in (Ivaldi, Karam et al. 2007). In short, salivary glands were incubated with primary antibodies [α-rabbit Fs(1)h long and short isoform (1:200), α-rabbit Fs(1)h long isoform (1:100), α-rat SuHw (1:100), mouse IgM anti-Pol II ser 5 phospho (1:50), or guinea pig anti-CG1894 (1:100)] in 1X PBS/0.1% Triton X-100/1% BSA at 4 °C overnight. Polytene chromosomes were washed in 1X PBS/0.1% Triton X-100 and incubated with the appropriate secondary antibodies for 1-2

hours at room temperature. Slides were then incubated with DAPI for 5 minutes and treated with Vectashield H-1000 (X0113) prior to visualization.

Chromatin-immunoprecipitation and DNA Library Preparation

Approximately 3.0 X 10^7 Kc167 cells were cross-linked in 1 % formaldehyde for 10 minutes at room temperature. The cross-linking reaction was quenched with 0.125 M glycine for 5 minutes. Cells were washed twice in cold 1X PBS and suspended in 5 mL of cell lysis buffer (5 mM PIPES, 85 mM potassium chloride, and 0.5 % Nonidet P40) supplemented with Roche cocktail protease inhibitors for 15 minutes on ice. Cells were spun down at 4000 rpm, and the nuclei were resuspended in 1 mL of cold nuclei lysis buffer (50 mM Tris HCl, 10 mM Na_2EDTA, and 1% SDS) with protease inhibitors for 20 minutes on ice. Samples were then diluted with cold IP dilution buffer (0.01% SDS, 1.1% Triton X-100, 1.2 mM Na_2EDTA, 16.7 mM Tris HCl, and 167 mM NaCl) with protease inhibitors and sonicated 14 times for 10-second intervals to generate DNA fragments ranging in length from 200 – 800 bp. Samples were then pre-cleared with GE Healthcare Protein A Sepharose Beads (#10090512) overnight at 4 °C. After pre-clear, 30 µL of guinea pig anti-CG1894 were added to the samples, and the samples were incubated overnight at 4 °C. Samples were next incubated with 50 µL of Protein A Sepharose Beads for 2 hours at 4 °C. The beads containing the immunoprecipitated chromatin were then washed in low salt buffer (0.1% SDS, 1 % Triton X-100, 2 mM Na_2EDTA, 20 mM Tris HCl, and 150 mM NaCl), high salt buffer (0.1 % SDS, 1 % Trition-X-100, 2 mM Na_2EDTA, 20 mM TrisHCl, and 500 mM NaCl), LiCl buffer (10 mM TrisHCl, 1 mM Na_2EDTA, 0.25 M LiCl, 1% NP40, and 1% sodium deoxycholate),

and eluted twice in 0.1M $NaHCO_3$/1% SDS. Crosslinking was reversed overnight at 65 °C, and samples were incubated with Proteinase K for 2 hours at 50 °C. DNA was isolated using standard phenol:chloroform extraction followed by ethanol precipitation.

The ChIP DNA was prepared for Illumina Adaptor Ligation using the End-It DNA End Repair Kit (ER0720) and the addition of "A" nucleotides to the ends of the fragments using Klenow 3'-5' exo- (M0212S). Illumina adaptors were titrated according to the concentration of sample DNA and ligated with T4 ligase for 2 hours at 16 °C. Samples were amplified via PCR with Phusion polymerase for 18 cycles, and size selected for 200 – 400 bp fragments using standard gel extraction procedures. Library samples were sequenced with an Illumina HiSeq2000 sequencer at the Hudson Alpha Institute for Biotechnology.

ChIP-seq Data Processing

Raw sequences were mapped to the *Drosophila* genome (dm3, Flybase 5.27) using BOWTIE 0.12.5 with default settings (Langmead, Trapnell et al. 2009). After normalizing the tag count for ChIP samples and input samples, MACS peak analysis software was used to find statistically significant peaks with a p-value cutoff of 1.0×10^{-10} (Zhang, Liu et al. 2008). Mof ChIP-chip data in Kc167 cells was downloaded from the Model Organism Encyclopedia of DNA Elements (modEncode) (www.modencode.org). ChIP-chip is similar to ChIP-seq except sequences are obtained via microarray. Gene expression data for Kc167 cells were also obtained from modEncode. ChIP-seq data for *Drosophila* Myc, both isoforms of Fs(1)h, the long isoform of Fs(1)h, RNA polymerase II, GAF, Jil-1, and 14-3-3 were generated previously by the Corces lab.

ChIP-seq Data Analysis

MYST5 peaks were intersected with the genomic regions described in Figure 2, (i.e. promoters, exons, introns, etc) which were obtained from the UCSC genome browser with the online Galaxy server using a minimum overlap of 1 bp (Giardine, Riemer et al. 2005, Goecks, Nekrutenko et al. 2010). The Galaxy server was also used to determine the degree of overlap between MYST5, GAF, Myc, and Mof sites and to identify promoters bound by these factors. All representative views of raw sequence data were generated using the Integrative Genomics Viewer (Robinson, Thorvaldsdottir et al. 2011).

For expression analyses, all *D. melanogaster* genes were sorted into 5 groups of roughly equal size with expression cutoffs as described in (Jingping Yang 2013). MYST5-bound genes were then assigned to one of five groups. The proportion of genes bound by MYST5 in each group was then plotted against the expected proportion of genes for each group.

The pausing index of each gene was calculated as described in (Jingping Yang 2013) using RNA pol II ChIP-chip data from Modencode. Briefly, the log2-transformed average amount of RNA Polymerase in the 200 bp region surrounding the transcription start site (TSS; TSS +/- 100 bp) was calculated. From this sum, the log2-transformed mean enrichment of RNA polymerase in the gene body and the 200 bp region following the end of the gene was subtracted to generate the pausing index. To calculate MYST5's genome-wide association with pausing, the log2-transformed sum of MYST5 enrichment

at each promoter was plotted against the pausing index of that promoter. Pearson's correlation test was used to test the correlation between MYST5 binding and pausing.

To assess differences in the pausing indices of different subsets of genes, a cumulative frequency distribution of pausing indices was generated for each group of genes. A Kolmogorov-Smirnov test was then used to assess statistical differences between each pausing distribution. To further assess the pausing characteristics of genes bound by MYST5, all genes were sorted into five groups on the basis of pausing index. In a manner similar to that described for expression analysis, the genes bound by MYST5 were sorted into one of these five groups. The proportion of genes belonging to each category was calculated and compared to the proportion of all genes belonging to each category. The pausing index cutoffs for each group are as follows: Group 5: 0.82 – 3.76; Group 4: 0.33 – 0.81; Group 3: 0.084 – 0.32; Group 2: -0.12 – 0.083; and Group 1: < -0.12.

To assess the binding enrichment of different proteins at MYST5 sites as well as the enrichment of RNA Pol II, MYST5, GAF, Mof, and Myc at the promoters of genes bound by MYST5, the wig data for each of these proteins was binned into 100 bp bins. The amount of reads in each 100 bp bin extending from the summit of each site of interest to 1000 bp upstream or downstream of the site was extracted for each site and placed in a matrix using in-house R scripts. Heatmaps were generated from the extracted matrices using Java Treeview software (Saldanha 2004).

Motif analysis of DNA sequences bound by MYST5 was performed with MEME Chip (Multiple Em for Motif Elicitate, (Machanick and Bailey 2011)) using default

parameters. In brief, the central 100 bp DNA sequence surrounding the summit of each MYST5 peak was analyzed to find the described consensus sequence.

Gene ontology (GO) analysis of the genes bound by MYST5, Myc, Mof, and GAF was performed using the Database for Annotation, Visualization and Integrated Discovery (DAVID, (Dennis, Sherman et al. 2003)). Each set of genes bound by GAF, Myc, or Mof was divided into two groups: Group 1: Genes bound only by a given factor (i.e. GAF) and Group 2: Genes bound by both a given factor and MYST5 (i.e. GAF and MYST5). Next, these groups were submitted to DAVID, which was used to determine the statistical enrichment of different GO terms in each group of genes. The three most significant terms enriched in each group were used to analyze differences in gene ontologies between each group. A Chi-square test with a two by two contingency table was used to determine if statistical differences in the number of genes in each group assigned to each GO term are present. As an example, consider the generic term "Development". A certain number of genes bound only by GAF can be assigned to this group. The rest of the genes are described as not belonging to this group. The same calculation can be made for genes bound by both GAF and MYST5. Based on the number of genes in each group that either belong or do not belong to this category, an expected frequency for inclusion can be calculated and used to determine if one group (i.e. genes with GAF and MYST5) contains a significantly higher or lower number of genes can be assigned the GO term "Development".

Results

CG1894 is a MYST histone acetyltransferase that binds the promoters of actively expressed genes.

CG1894 has been commonly referred to as a member of the MYST family of histone acetyltransferases (Sapountzi and Cote 2011). In support of this, I used the National Center for Biotechnology Information's Basic Local Alignment Search Tool (NCBI BLAST) to analyze CG1894's amino acid sequence and found that it possesses a MOZ/Sas acetyltransferase domain, which is characteristic of the MYST family of histone acetyltransferases (Supplementary Figure 1A). Within this catalytic domain is a motif that allows the enzyme to bind acetyl-CoA as well as a C2HC-type zinc finger motif (Sapountzi and Cote 2011). Further analysis of CG1894's amino acid sequence with Clustal Omega (Sievers, Wilm et al. 2011) revealed that CG1894, when compared to *D. melanogaster* histone acetyltransferases from different enzyme families, clusters with Tat interactive protein 60 kDa (Tip60), Chameau, and Mof, which are well characterized members of the MYST family of histone acetyltransferases (Supplementary Figure 1B). Taken together, these data indicate that CG1894 may possess catalytic activity and that it is a MYST histone acetyltransferase. Given these findings, I henceforth refer to CG1894 as MYST5 because the names MYST1, MYST2, MYST3, and MYST4 have been assigned to other proteins.

Because of the extensive roles that MYST HATs play in transcription, I sought to examine if MYST5 is also involved in regulating gene expression. To gain a global view of MYST5's function, I generated a genome-wide binding profile by performing chromatin immunoprecipitation with high throughput sequencing (ChIP-seq) on Kc167

cells with an antibody against MYST5. A representative view of the raw sequencing data is shown in Figure 2A. These data indicate that MYST5 has discrete binding patterns and is found near genomic regions that possess genes. After aligning the raw sequence data to the *D. melanogaster* genome using BOWTIE and performing peak analysis with MACS, 1583 binding sites were identified for MYST5. From the binding data, I discovered that nearly half (~ 48%) of MYST5's binding regions overlap with gene promoters (defined as 200 bp upstream of the transcription start site) (Figure 2B), indicating that MYST5 may play a role in regulating the expression of these genes. MYST5's binding sites also overlapped considerably with exons (11%), introns (19%), and distal intergenic regions (22%) that are more than 200 bp away from the start and end sites of genes. These data suggest that MYST5 may play additional roles outside of direct transcriptional regulation.

Overall, MYST5 binds the promoters of 816 unique genes. To gain a greater insight into MYST5's function at these promoters, I analyzed the expression values of the genes that correspond to promoters bound by MYST5. Using criteria established by *Yang et al*, genes bound by MYST5 were placed into five different expression groups with group 1 genes possessing the lowest expression and group 5 genes having the highest expression values (Jingping Yang 2013). Approximately 45% of genes with MYST5 in the promoter belong to the top 20% of most highly expressed genes while roughly 70% of genes with MYST5 belong to the top 40% of genes (Groups 4 and 5) in terms of expression (Figure 2C). This result is indicative that MYST5 associates with actively transcribed genomic regions and that it may promote active transcription.

Many proteins have been shown to play a role in promoting gene expression in *D. melanogaster*. Among these factors are *Drosophila* Myc, Mof, the Jil-1 kinase, 14-3-3 proteins, and RNA polymerase II, which were described above. In further support of a role for MYST5 in promoting transcription, I examined the binding distributions of the above-mentioned factors at MYST5 sites and discovered that they colocalize extensively at MYST5 sites (Figure 2D). Additionally, I found that MYST5 colocalized with both isoforms of *Drosophila* female sterile 1 homeotic (Fs(1)h), which is the ortholog of BRD4 described above (Figure 2D). Moreover, there was a general binding correspondence between these factors and MYST5 on a genome-wide scale (Figure 2D). The colocalization and correspondence between MYST5 and these factors suggest that MYST5 may associate with these proteins to play an active role in transcription.

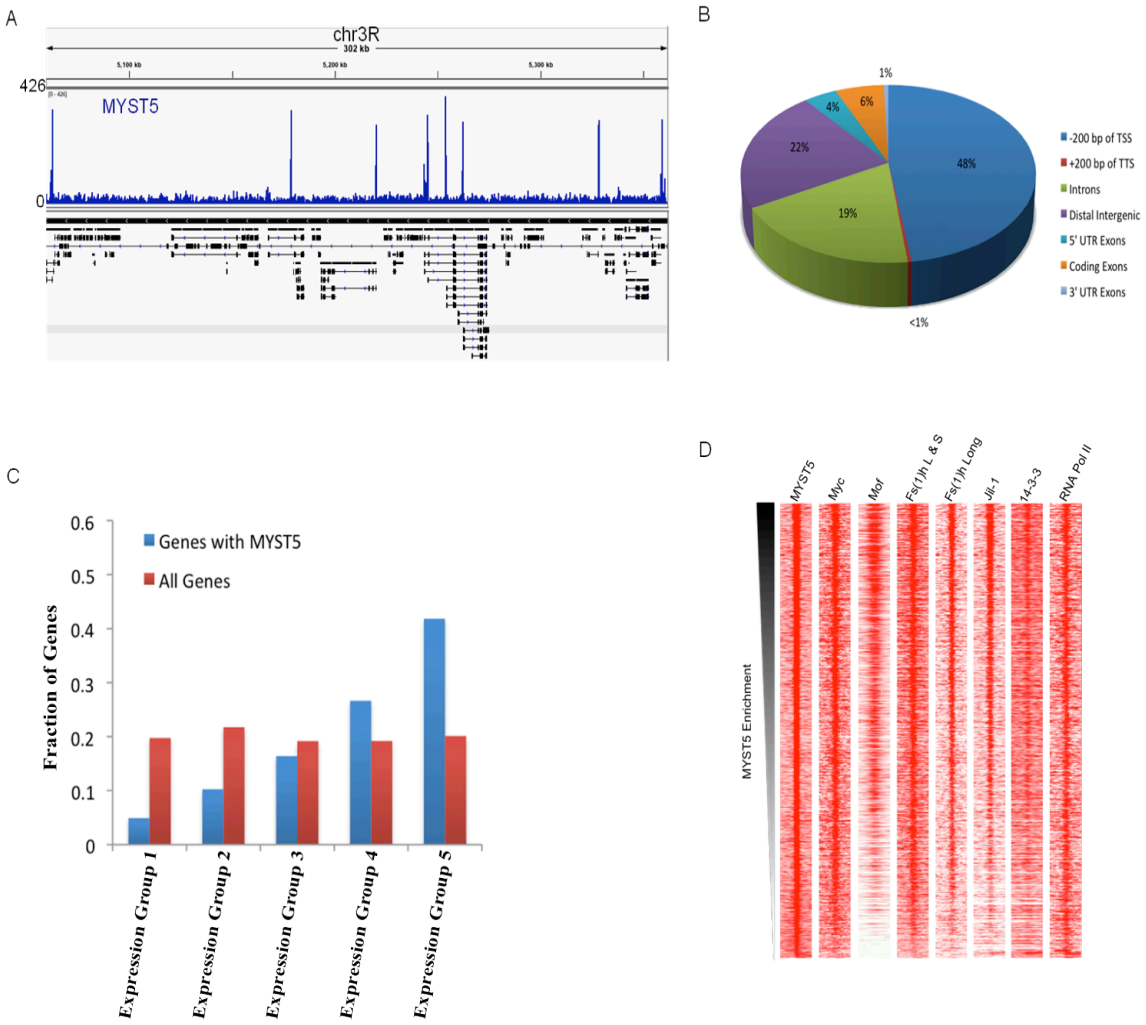


Figure 2- The genomic binding patterns of MYST5 correlate with active transcription

(A) Representative view of raw (wig) MYST5 binding data on a 300 kb segment of chromosome 3L. The Y-axis represents the MYST5 read count enrichment, and the X-axis represents linear genomic regions. The lines and bars below the X-axis represent genes. The data indicate that there is a strong enrichment of MYST5 near genes. (B) Genomic distribution of MYST5 binding sites showing that almost half of MYST5 binding sites overlap with promoters. Peaks were intersected using the online Galaxy server. (C) Graph showing the proportion of genes bound by MYST5 belonging to each expression group. All genes were divided into five subgroups of roughly equal size based on expression. A higher than expected proportion of genes bound by MYST5 belongs to the top expression groups, implicating a role for MYST5 in active gene expression. (D) Heatmap showing association of MYST5 with other factors known to promote transcription. Heatmaps represent a 2 kb window centered on the summits of MYST5 binding regions and were ordered according to increasing MYST5 enrichment. Darker red regions correspond to regions of higher read enrichment.

MYST5 binds a highly paused subset of genes

The discovery that RNA polymerase often stalls downstream of the transcription start site shortly after initiating transcription at most genes suggests that this is a key regulatory event for a given gene's expression. Although initially suspected to result in a lower level of transcription, polymerase pausing is now widely believed to serve as a mechanism to fine-tune expression (Adelman and Lis 2012). Upon closer inspection of the genes with MYST5 enriched in their promoters, I noticed that they exhibited a substantial enrichment for RNA polymerase II near their transcription start sites relative to the average gene (Figure 3A). This observation led to the speculation that MYST5 could possibly regulate transcription by playing a more specific role in polymerase pausing. To discern a possible role for MYST5 in polymerase pausing on a global level, MYST5 binding in promoter regions was plotted against the pausing indices of all genes. The pausing index of a gene measures the difference between the total amounts of polymerase near its promoter versus that in the gene body. A high pausing index is indicative of a great degree of polymerase retention in the promoter region whereas a lower pausing index indicates that there is more polymerase in the gene body region.

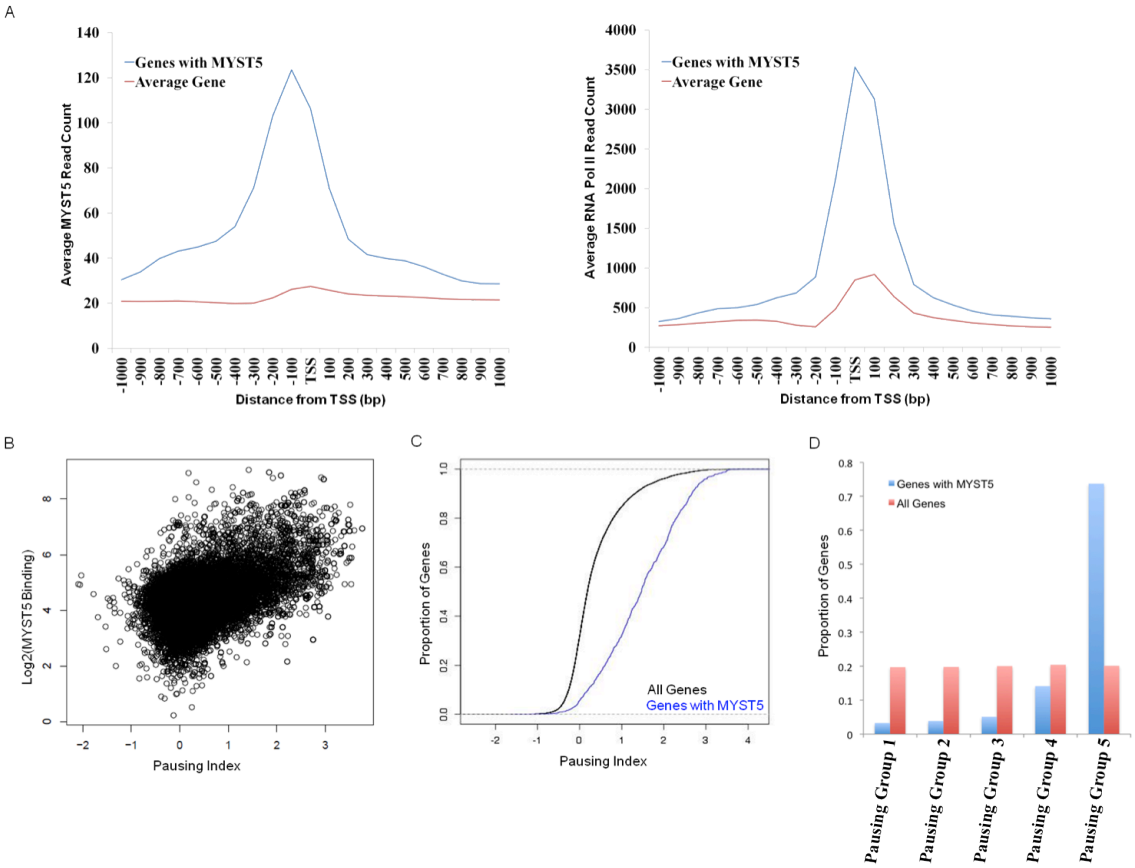


Figure 3 – MYST5 associates with genes that are highly paused
(A) Density distributions of the average MYST5 read count (left panel), and the average RNA pol II read count (right panel) near the promoter regions of genes bound by MYST5. Distributions are centered around the transcription start site. Relative to the average gene, MYST5 is enriched at approximately -100 bp from the transcription start site, confirming that it is at the promoters of these genes. RNA pol II is highly enriched at the transcription start site indicating that MYST5-bound genes may be paused. (B) Scatterplot of the logarithmically transformed MYST5 read count in promoters plotted against the pausing indices of all genes. Results indicate a positive correlation between MYST5 binding and pausing (Pearson's correlation coefficient = 0.497, p-value < 2.2 X 10^{-16}. (C) Cumulative frequency distribution of the pausing indices of all genes (black line) and genes bound by MYST5 (blue line). The pausing indices of MYST5-bound genes are significantly higher than that of all genes (KS-test, p-value < 2.2 X 10^{-16}). (D) Graph showing the proportion of genes bound by MYST5 belonging to different groups of genes based on pausing index. The data indicate that almost all (~80%) of MYST5 bound genes belong to group 5 which comprises the top 20% most highly paused genes. This is four times higher than expected.

Consistent with a function for MYST5 in polymerase pausing, the degree of MYST5 promoter occupancy correlates significantly with the pausing status of genes on a global level (Figure 3B). Furthermore, when compared to all *Drosophila* genes, the subset of genes bound by MYST5 had significantly higher pausing indices (Figure 3C, p-value $< 2.2 \times 10^{-16}$). A closer examination of individual genes bound by MYST5 revealed that MYST5 is enriched in the promoter region of heat shock genes (Supplementary Figure 2A), which numerous studies have demonstrated to be highly paused in *Drosophila (Rougvie and Lis 1988)*. MYST5 also binds the promoters of the *patched* (*ptc*) and *frizzled* (*fz*) genes, which are enriched for polymerase near the transcription start site and have pausing indices that are more than five-fold higher than the genome-wide average (Supplementary Figure 2B).

In a manner similar to that for my expression analyses, I subdivided genes into five different groups based on pausing index and assigned genes bound by MYST5 into each group. Group 5 consisted of the top 20% most highly paused genes whereas group 1 contained the lowest 20% of genes in terms of pausing index. Strikingly, the results of this analysis indicated that MYST5 almost exclusively binds (~80% of genes) the promoters of genes that belong to group 5 (Figure 3D). These data along with those described above strongly indicate that MYST5 not only associates with highly paused genes, but that it also may play an active role in regulating polymerase pausing at this highly paused subset of genes.

To gain a more global view of the processes that MYST5-bound genes function in, I performed Gene Ontology analyses on these genes using the Database for

Annotation, Visualization, and Integrated Discovery (DAVID). From these analyses, I discovered that genes occupied by MYST5 are highly enriched in developmental terms with the three most significant terms being "Imaginal Disc Development", "Post-embryonic Development", and "Tissue Morphogenesis" (Table 1). Because many developmental genes are paused (Zeitlinger, Stark et al. 2007), these findings are consistent with the finding that MYST5 binds highly paused genes and implicate a role for MYST5 in regulating gene expression during developmental processes.

MYST5 colocalizes with factors that function in pausing

To date, many factors have been shown to play a role in polymerase pausing. For instance, NELF, DSIF and GAF have been shown to play functional roles in establishing polymerase pausing (Wu, Yamaguchi et al. 2003, Lee, Li et al. 2008). In contrast, Myc and Mof may play roles in the release of paused polymerase (Zippo, Serafini et al. 2009, Rahl, Lin et al. 2010).

An interesting finding from the analysis of MYST5's amino acid sequence is that it is not predicted to have a canonical DNA-binding domain or chromodomain, which allow many chromatin-associated factors to bind to chromatin (Supplementary Figure 1A). Thus, it is likely that MYST5 associates with other proteins in order to exert its regulatory functions, if any, on polymerase pausing. In an effort to identify some of these factors, I analyzed the 100 bp DNA sequences surrounding the summits of MYST5 binding sites with MEME ChIP in order to discover motifs bound by other factors involved in transcription.

With this analysis, I found that MYST5 sites were centrally and significantly enriched for a motif commonly bound by GAF (also known as trithorax-like) (Figure 4A), suggesting that it may colocalize with this protein. GAF has been shown to play a role in a variety of processes including enhancer blocking, insulator bypass, and promoter proximal pausing (Ohtsuki and Levine 1998, Melnikova, Juge et al. 2004, Lee, Li et al. 2008). In further support of an association between MYST5 and GAF, I found that 80% of MYST5 sites overlap with sites bound by GAF (Figure 4B). I also examined whether or not MYST5 associates with other factors known to play a role in promoter pausing and found that nearly half of MYST5 sites intersect sites bound by Myc and a little over two-thirds of MYST5 sites overlap regions bound by Mof (Figure 4B).

Given the substantial overlap between MYST5 and these pausing factors, I next investigated whether these factors also localize to promoter regions bound by MYST5. I found a high degree of correspondence between the binding of MYST5, GAF, Myc, and Mof at the highly paused promoters occupied by MYST5 (Figure 4C and Supplementary Figure 3A). As expected, I also observed an enrichment of RNA polymerase II at these highly paused genes that tended to correspond with the pausing index of these genes (Figure 4C). These data suggest that MYST5 could possibly function with GAF, Myc, and Mof to regulate polymerase pausing.

GO Term	% of Genes	P-value	Fold Enrichment	False Discovery Rate
Imaginal Disc Development	12.6	1.5×10^{-25}	3.4	2.6×10^{-22}
Post-embryonic Development	12.5	3.0×10^{-23}	3.2	5.1×10^{-20}
Tissue Morphogenesis	9.1	6.9×10^{-23}	4.2	1.2×10^{-19}
Epithelium Development	8.7	8.1×10^{-22}	4.2	1.4×10^{-18}
Morphogenesis of an Epithelium	8.4	2.6×10^{-21}	4.2	4.5×10^{-18}
Instar Larval or Pupal Development	11.6	1.3×10^{-20}	3.1	2.2×10^{-17}
Post-embryonic Organ Development	9.6	1.7×10^{-20}	3.6	2.9×10^{-17}
Post-embryonic Morphogenesis	10.4	3.9×10^{-20}	3.3	6.7×10^{-17}
Instar Larval or Pupal Morphogenesis	10.3	6.2×10^{-20}	3.3	1.1×10^{-16}
Metamorphosis	10.4	1.7×10^{-19}	3.2	3.0×10^{-16}

Table 1- Gene ontology analysis of genes bound by MYST5
Table showing the top 10 gene ontology categories for genes bound by MYST5. Unique flybase I.D.s for each gene bound by MYST5 were submitted to DAVID and analyzed using default conditions. The table shows that MYST5 associates with genes that may play critical roles in development.

Although a substantial portion of MYST5 sites (~ 50% - 80%) overlap with GAF sites, Myc sites, or Mof sites, MYST5 binding regions only overlap with roughly 20% of each of these proteins' binding sites (Figure 4B). Thus, it is possible that MYST5 may identify a unique subset of GAF, Myc, and Mof sites throughout the genome. To explore this further, I identified genes bound by GAF, Myc, and Mof and intersected these genes with MYST5 binding regions to discover genes bound by both MYST5 and either GAF, Myc, or Mof. Overall, I discovered that MYST5 bound 725 of 2786 GAF-bound genes, 376 of 2311 Myc-bound genes, and 664 of 5588 genes bound by Mof.

I next asked how the pausing distributions of genes bound by GAF and MYST5, Myc and MYST5, and Mof and MYST5 compared to all genes bound by GAF, Myc, or Mof, respectively. In all three cases, I found that the pausing distributions of genes that were bound by GAF and MYST5, Myc and MYST5, and Mof and MYST5 were significantly greater than those of all genes bound by either GAF, Myc, and Mof (p-value $< 2.2 \times 10^{-16}$) (Figure 4D). The pausing indices of genes bound by both GAF and MYST5 as well as those bound by Mof and MYST5 were not significantly different from those of all genes bound by MYST5 (Figure 4D), indicating that MYST5 may bind a highly paused subset of genes that are also bound by GAF and Mof. Interestingly, the pausing indices of genes bound by both Myc and MYST5 were significantly greater than for all genes bound by MYST5 (p-value < .05) (Figure 4D). This finding suggests that a combination of MYST5 and Myc binding identifies a subset of genes that are more highly paused than all genes bound by MYST5 or all genes bound by Myc.

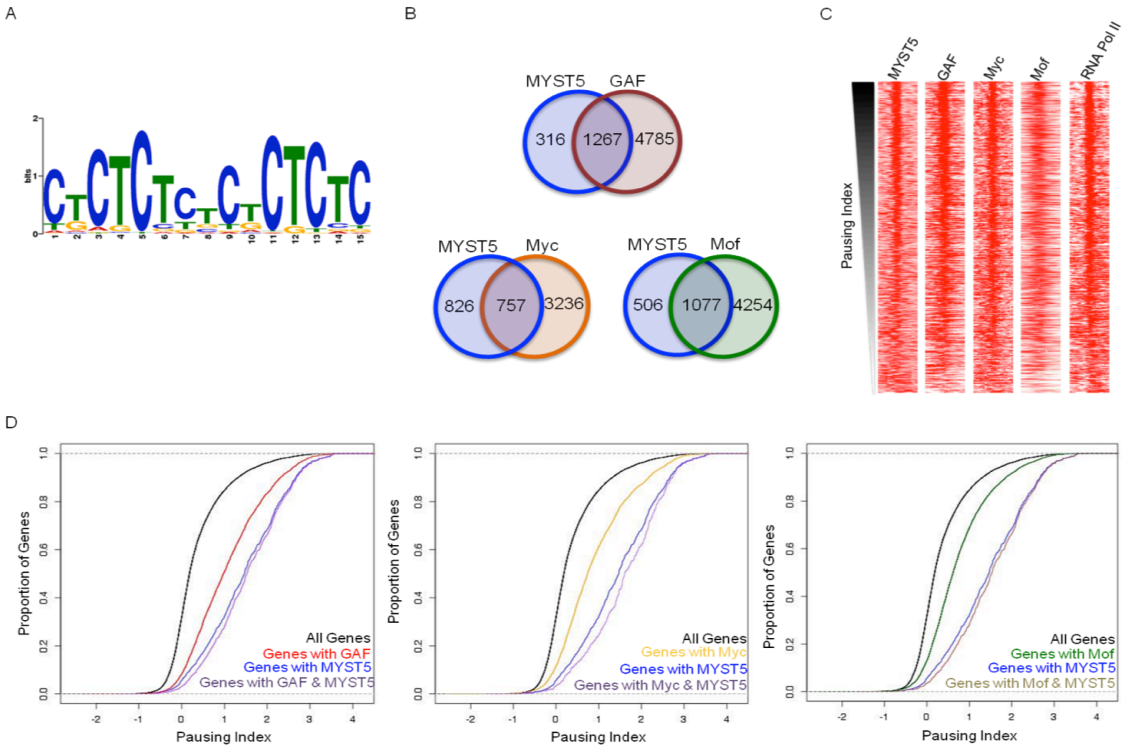


Figure 4- MYST5 may interact with other pausing factors at a subset of genes
(A) The consensus sequence identified for MYST5 in the 100 bp region surrounding MYST5 summits is the previously identified GAF binding motif (E-value < 1.0 X 10^{-895} ; p-value < 1.0 X 10^{-5}). (B) Venn diagrams showing genome-wide binding site overlaps between MYST5 and GAF, MYST5 and Myc, and MYST5 and Mof. The number of overlapping sites was determined using Galaxy and a 1 bp minimum overlap. (C) Heatmap showing read enrichment for MYST5, GAF, Myc, Mof, and RNA pol II at genes bound by MYST5. Heatmap is arranged according to increasing pausing index of the set of genes and represent a 2 kb window centered around the transcription start site of each gene. Data indicate an enrichment and correspondence between these factors near the promoters of genes bound by MYST5. (D) Cumulative frequency distributions of the pausing indices of different categories of genes. The left panel consists of a cumulative pausing distribution for all genes, all genes bound by GAF, all genes bound by MYST5, and genes bound by both GAF and MYST5. The middle and right panel are similar except they analyze the genes bound by Myc, and Mof respectively. Genes bound by both GAF and MYST5, Myc and MYST5, and Mof and MYST5 have significantly higher pausing indices (p-value < 2.2 X 10^{-16}) relative to all genes bound by GAF, Myc, and Mof, which themselves have higher pausing indices (p-value < 2.2 X 10^{-16}) relative to all genes. Genes bound by GAF and MYST5 as well as genes with Mof and MYST5 do not have significantly higher pausing indices compared to MYST5 genes. Genes bound by Myc and MYST5 are significantly (p-val < .05) more paused relative to all genes bound by MYST5. The data indicate that MYST5 may identify a highly paused subset of GAF and Mof genes and binds a subset of Myc genes that are significantly more paused than all MYST5-bound genes and all Myc-bound genes.

Given the finding that genes with a combination of either GAF, Myc, or Mof and MYST5 are more paused than the set of genes having only GAF, Myc, or Mof, I next explored the functions of these genes to see if combinatorial binding by these factors also identified a different functional subset of genes. For this analysis, I split all genes bound by GAF, Myc, or Mof into two groups each: genes only bound by GAF and genes bound by both GAF and MYST5, genes only bound by Myc and genes bound by both Myc and MYST5, and genes bound by only Mof and genes bound by both Mof and MYST5. I then used DAVID to perform GO analysis on the genes belonging to each of the above-mentioned categories, examining the three most significant terms identified for each of these groups.

I found that genes bound by both Myc and MYST5 as well as those bound by both Mof and MYST5 were enriched for developmental terms (Supplementary Figure 3B, top half of middle and right panels), whereas genes bound by only Myc or only Mof were enriched mostly for housekeeping terms (Supplementary Figure 3B, top half of middle and right panels). This is consistent with a recent finding by the Becker lab that Mof almost exclusively binds to genes involved in housekeeping functions (Feller, Prestel et al. 2012). I also found that genes bound only by GAF as well as those with both GAF and MYST5 were also enriched for terms associated with development (Supplementary Figure 3B, left panel). Using a Chi-square test with a two by two contingency table, I discovered that terms associated with development were significantly more enriched for genes bound by Myc and MYST5, and Mof and MYST5 relative to genes bound only by Myc and Mof, respectively (Supplementary Figure 3B, middle and

right panel). This analysis also revealed that genes bound both by GAF and MYST5 were more enriched in some, but not all, developmental categories relative to those only bound by GAF (Supplementary Figure 3B, left panel). Collectively, these data indicate that MYST5 may identify subsets of developmental genes bound by GAF, Myc, and Mof.

MYST5 is essential for H4K5 and H4K8 acetylation, associates with active transcription, and functions downstream of the Jil-1 kinase during transcription.

Motivated by my findings from the ChIP-seq data, I investigated the mechanism by which MYST5 regulates transcription. I first sought to determine if MYST5 is necessary for maintaining normal levels of certain histone modifications. To this end, the levels of H3K9ac, H3S10ph, H3K14ac, H4K5ac, H4K8ac, H4K12ac, and H4K16ac in 3rd instar *D. melanogaster* larvae that were null for the MYST5 protein (Supplementary Figure 4A) were compared to those of wild type larvae. I found that the levels of H4K5ac and H4K8ac were reduced in mutants relative to wild type larvae (Figure 5A). The levels of all other histone modifications remained relatively unchanged (Supplementary Figure 4B). These findings implicate a role for MYST5 in the acetylation of H4K5 and H4K8.

In accordance with the ChIP-seq data, MYST5 localizes to regions of active transcription marked by RNA pol II phosphorylated at serine 5 during both non-heat shock and heat shock conditions on *Drosophila* polytene chromosomes (Figure 5B). These findings substantiate a role for MYST5 in active transcription. Furthermore, localization to the promoters of heat shock genes provides an independent confirmation

of the ChIP-seq data, which indicate that MYST5 is enriched at the promoter regions of heat shock genes (Supplementary Figure 2A).

Because loss of MYST5 did not affect levels of H3S10ph and H3K9ac, which have been shown to be downstream of transcription initiation by RNA polymerase II phosphorylated at its serine 5 residue and of the Jil-1 kinase, it is possible that MYST5 may function downstream of these events. To test this hypothesis, the giant polytene chromosomes of *D. melanogaster* were used to assess the ability of MYST5 to bind under different conditions. Polytene chromosomes are advantageous for this purpose as they allow for direct visualization of MYST5's ability to bind chromatin in the absence of Jil-1.

If MYST5 functions downstream of Jil-1 during active transcription, then its ability to localize to chromatin should be diminished in the absence of Jil-1. Analyses of polytenes from 3rd instar larvae null for the Jil-1 kinase revealed that MYST5 binding to polytene chromosomes is lost in these mutants (Figure 5C). These data indicate that MYST5's role in regulating transcription is downstream of the activity of Jil-1.

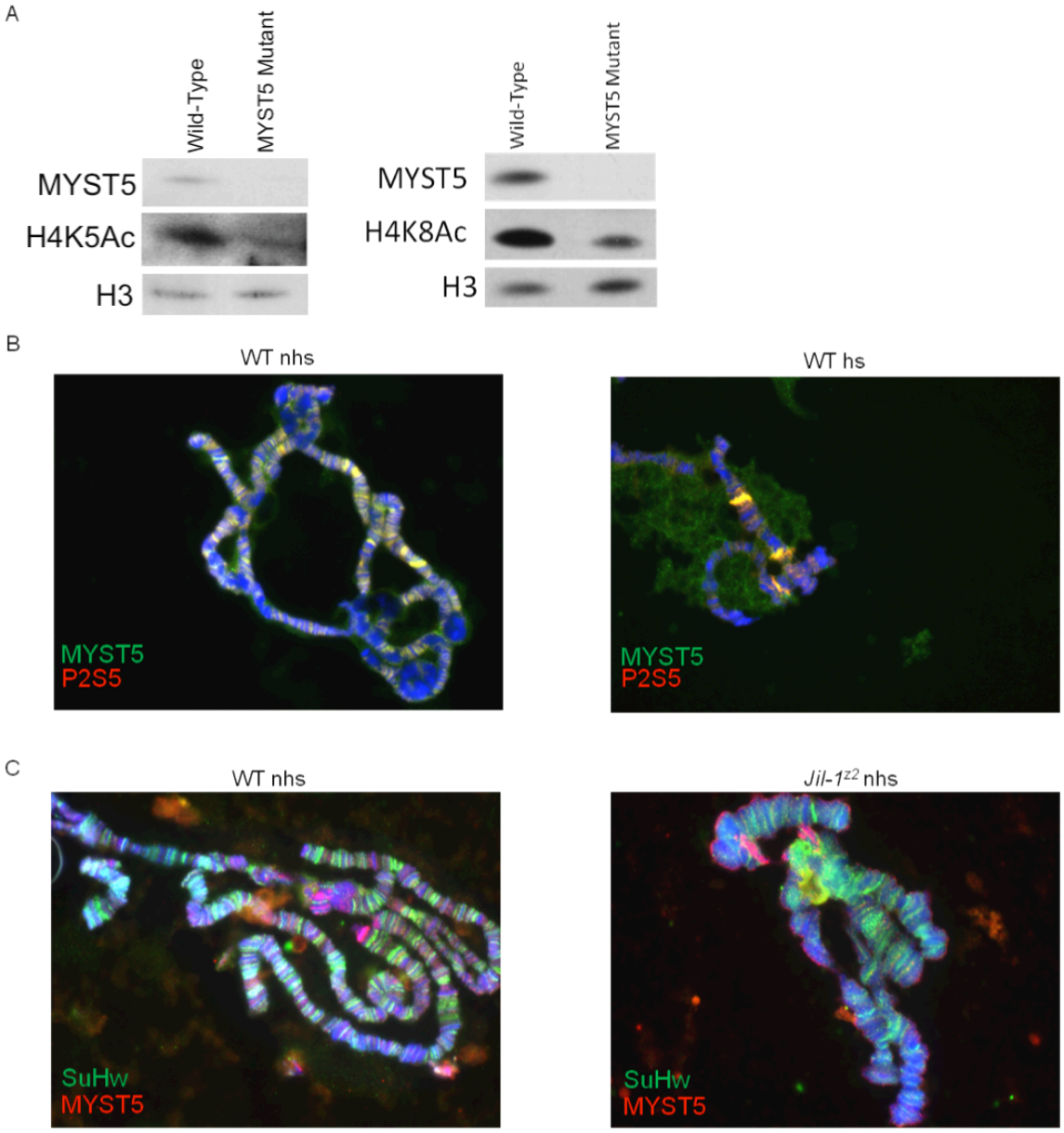


Figure 5 – MYST5 is essential for H4K5 and H4K8 acetylation and functions downstream of Jil-1 during active transcription.
(A) Western analyses indicate that overall levels of acetylated H4K5 and H4K8 are reduced in larvae that are null for MYST5. These data indicate a role for MYST5 in H4K5 and H4K8 acetylation. (B) Polytene chromosomes showing that MYST5 localizes to active regions as indicated by its overlap (yellow bands) with RNA pol II phosphorylated at serine 5 (initiating polymerase). (WT = wild type, nhs = non-heat shock, hs = heat shock) (C) Polytene chromosomes showing that MYST5 binding to chromatin is lost in *Jil-1*^{z2} null mutants. Suppressor of Hairy Wing (SuHw), previously shown to be unaffected by loss of Jil-1, was used as an internal control for fluorescence.

MYST5 is essential for polymerase pause release but not for binding of Fs(1)h

I next aimed to characterize MYST5's specific role in promoter-proximal polymerase pausing. Extensive research has shown that the phosphorylation of RNA polymerase II on serine 2 of its C-terminal domain by the P-TEFb kinase is an indication of pause release and signals the start of productive transcription elongation (Figure 1 Stage 3) (Ni, Saunders et al. 2008). To exactly determine how MYST5's activity regulates polymerase pausing, I examined the levels of RNA polymerase II phosphorylated at its serine 2 residue (elongating polymerase) in 3rd instar larvae.

If MYST5 functions in establishing the polymerase pausing event, then the levels of elongating polymerase would be expected to increase relative to the levels seen in wild type larvae as more polymerases will be able to escape the pause event. Conversely, if MYST5 is essential for pausing release, then loss of this protein would result in lower levels of elongating polymerase as fewer polymerases will be phosphorylated on their serine 2 residue due to their inability to escape pausing. Alternatively, MYST5 may not function in polymerase pausing, and the levels of RNA pol II serine 2 phosphorylation would remain constant in MYST5 mutants. In support of the second hypothesis, the levels of elongating polymerase were decreased in larvae that lacked MYST5 as evidenced by a reduction in the form of polymerase phosphorylated at the serine 2 residue (Figure 6A). The ability of polymerase to initiate transcription remained relatively unaffected as revealed by examination of levels of RNA pol II serine 5 phosphorylation, H3S10 phosphorylation, and H3K9 acetylation, although some mutant larvae exhibited fluctuating RNA pol II serine 5 phosphorylation levels (Supplementary

Figure 4B). Collectively, these data indicate that MYST5 is essential for the release of paused polymerase and not transcription initiation.

I next attempted to establish a connection between the activity of MYST5 and RNA pol II serine 2 phosphorylation. The modifications on the N-terminal tails of histones can serve as binding platforms for a multitude of chromatin regulatory factors, including many that function in transcription. *Zhang et al* demonstrated that BRD4 binds to acetylated H4K5 and H4K8 to regulate RNA pol II serine 2 phosphorylation at enhancers in human CD4 + T cells (Zhang, Prakash et al. 2012). The *Drosophila* ortholog of BRD4 is Fs(1)h, and it has two major isoforms. Although it has been shown that BRD4 binds to acetylated H3K9 and H4K16, the above study demonstrates that BRD4 can differentially bind to different histone acetylation patterns to exert differential regulatory functions. Thus, it is possible that Fs(1)h could bind acetylated H4K5 and H4K8 during polymerase pausing. Analyses of the ChIP-seq data indicate that there is a strong association between both Fs(1) h isoforms at MYST5 binding sites (Figure 2D), suggesting that these proteins could possibly serve as the connection between MYST5's acetyltransferase activity and RNA pol II serine 2 phosphorylation.

To determine if Fs(1)h is the connection between H4K5 and H4K8 acetylation by MYST5 and polymerase pause release, I examined the ability of both isoforms of Fs(1)h to bind to polytene chromosomes in MYST5 mutants. Using an antibody that recognizes both the long and short isoforms of Fs(1)h, I did not observe an appreciable difference in the binding of Fs(1)h to polytene chromosomes in MYST5 mutants relative to that seen in wild type larvae (Figure 6B). Although this suggested that Fs(1)h could still bind to

chromatin in the absence of MYST5, it is possible that only one of the two isoforms binds acetylated H4K5 and H4K8 and that a reduction in the binding of one isoform may not be noticeable due to the antibody recognizing the other isoform. Therefore, I also examined the binding of the long version of Fs(1)h with an antibody that specifically recognizes the long isoform. As before, I observed no substantial reduction in the ability of the long variant of Fs(1)h to bind chromatin in the absence of MYST5 activity (Figure 6C). Collectively, these findings suggest that MYST5's regulation of RNA pol II serine 2 phosphorylation does not involve Fs(1)h, although there is still a possibility that the signal generated by the long Fs(1)h isoform in experiments with an antibody examining the binding of both isoforms could be masking the signal for the short isoform.

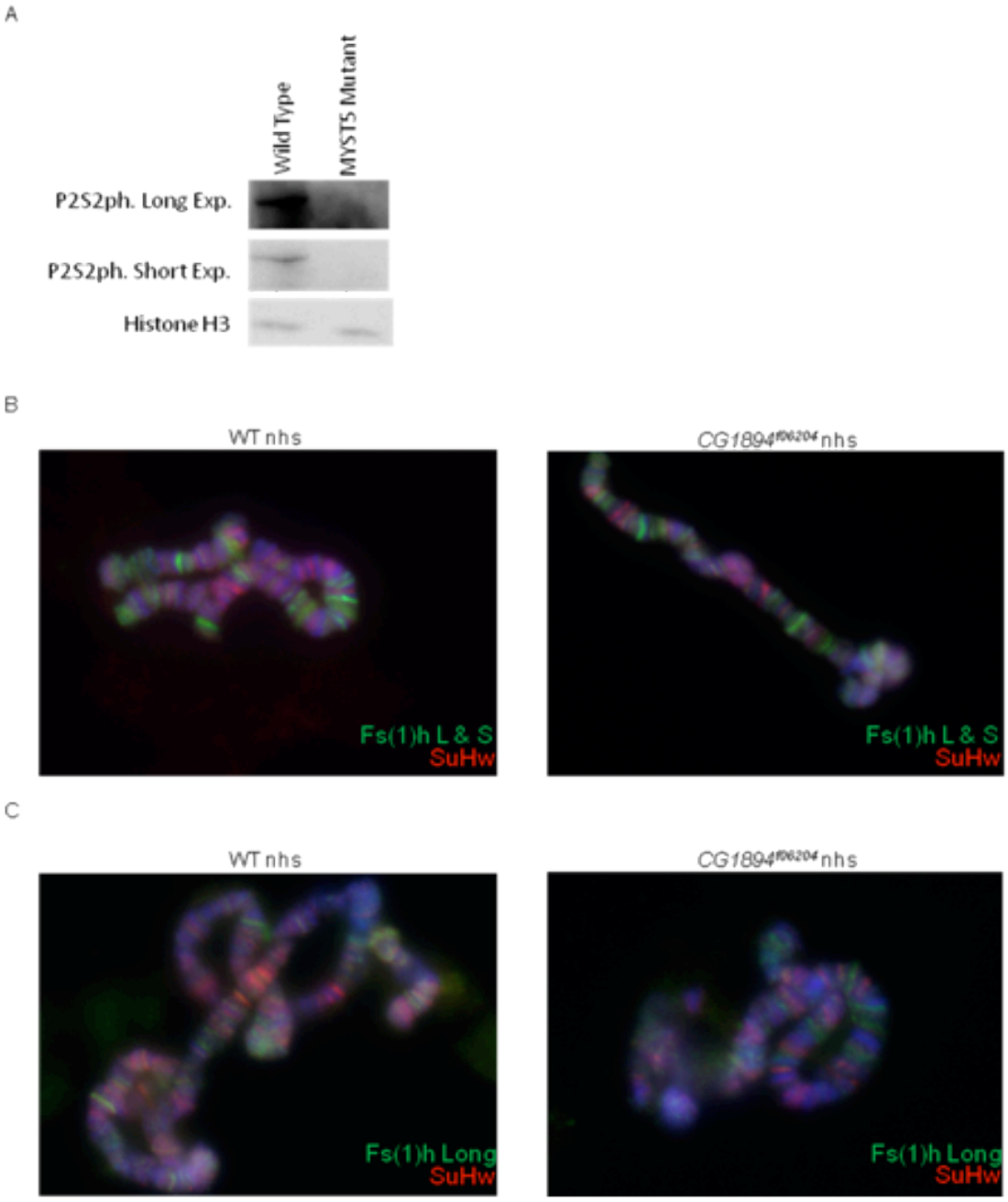


Figure 6 – MYST5 is necessary for proper pause release but not Fs(1)h binding
(A) Western blot showing that larvae that are null for MYST5 have reduced levels of elongating polymerase as evidence by a reduction in global levels of RNA pol II serine 2 phosphorylation (exp = exposure). (B) Polytene chromosomes from wild type and MYST5 mutant larvae showing that the binding of both isoforms of Fs(1)h (L = long, S = short) appear unaffected in larvae null for MYST5 (WT = wild type, nhs = non-heat shock). (C) Polytene chromosomes from wild type and larvae null for MYST5 show that the binding of the long Fs(1)h is unaffected by loss of MYST5.

Discussion

Prior to this study, little was known about MYST5's involvement in histone acetylation and transcription regulation. Although previous studies speculated that MYST5 could be a histone acetyltransferase and belonged to the MYST family of histone acetyltransferases on the basis of amino acid sequence analysis, no research has yet established an *in-vivo* connection between MYST5 and histone acetylation. Even though it cannot exclude an indirect relationship between MYST5 and histone acetylation, the present study strongly suggests that MYST5 is essential for H4K5 and H4K8 acetylation (Figure 5A). On the basis of this and other studies, there is a strong possibility that the acetylation of H4K5 and H4K8 may be critical to regulating transcription, specifically in the context of promoter-proximal pausing.

My analyses of MYST5's genomic localization revealed a previously unknown relationship between this factor and active transcription. I found that MYST5 is enriched in the promoter region of 816 genes, and that a majority of these genes (~ 70%) can be grouped within the top two quintiles of genes in terms of their expression (Figure 2C). I also discovered that numerous factors involved in promoting transcription are strongly enriched along with RNA polymerase II at numerous sites bound by MYST5 (Figure 2D). These data, along with the finding that MYST5 localizes to regions of active transcription on polytene chromosomes, highlight an important function for MYST5 in regulating transcription.

The strong enrichment of RNA polymerase II at the promoters of genes bound by MYST5 led to the finding that these genes are highly paused relative to all genes. In fact,

the genes bound by MYST5 belong almost exclusively (~80%) to the top quintile of paused genes (Figure 3D). MYST5 also colocalizes with GAF, Myc, and Mof, which function in regulating polymerase pausing, at the promoters of a subset of genes bound by these factors. Finally, western analyses reveal that MYST5 is critical for proper polymerase pause release. Therefore, it is clear that MYST5 plays an important role in transcription through polymerase pause release.

The current study led to the proposed the model of MYST5's function outlined in Figure 7. This model begins with RNA polymerase II that has already initiated transcription and been paused according to the mechanisms, which include the binding of GAF, described in Figure 1. In the first stage of this model the Jil-1 kinase phosphorylates H3S10 on a nearby nucleosome, leading to the subsequent recruitment of 14-3-3 and Elp3, which acetylates H3K9 (Figure 7 part 1) (Karam, Kellner et al. 2010). Also in response to H3S10 phosphorylation, Mof acetylates H4K16. Because MYST5 functions downstream of Jil-1, I propose that its recruitment to chromatin and regulation, either direct or indirect, of H4K5 and H4K8 acetylation occur after H3S10 phosphorylation (Figure 7 part 2). This hypothesis is supported by data indicating that MYST5 is unable to bind chromatin in Jil-1 null larvae (Figure 5C). Although the exact mechanism remains to be elucidated, it is possible that the acetylation of H4K5 and H4K8 may contribute to the recruitment of P-TEFb to the pause site (Figure 7 part 3). The recruitment of P-TEFb then results in the phosphorylation of RNA pol II at its serine 2 residue, triggering its release into productive transcription elongation (Figure 7 part 4).

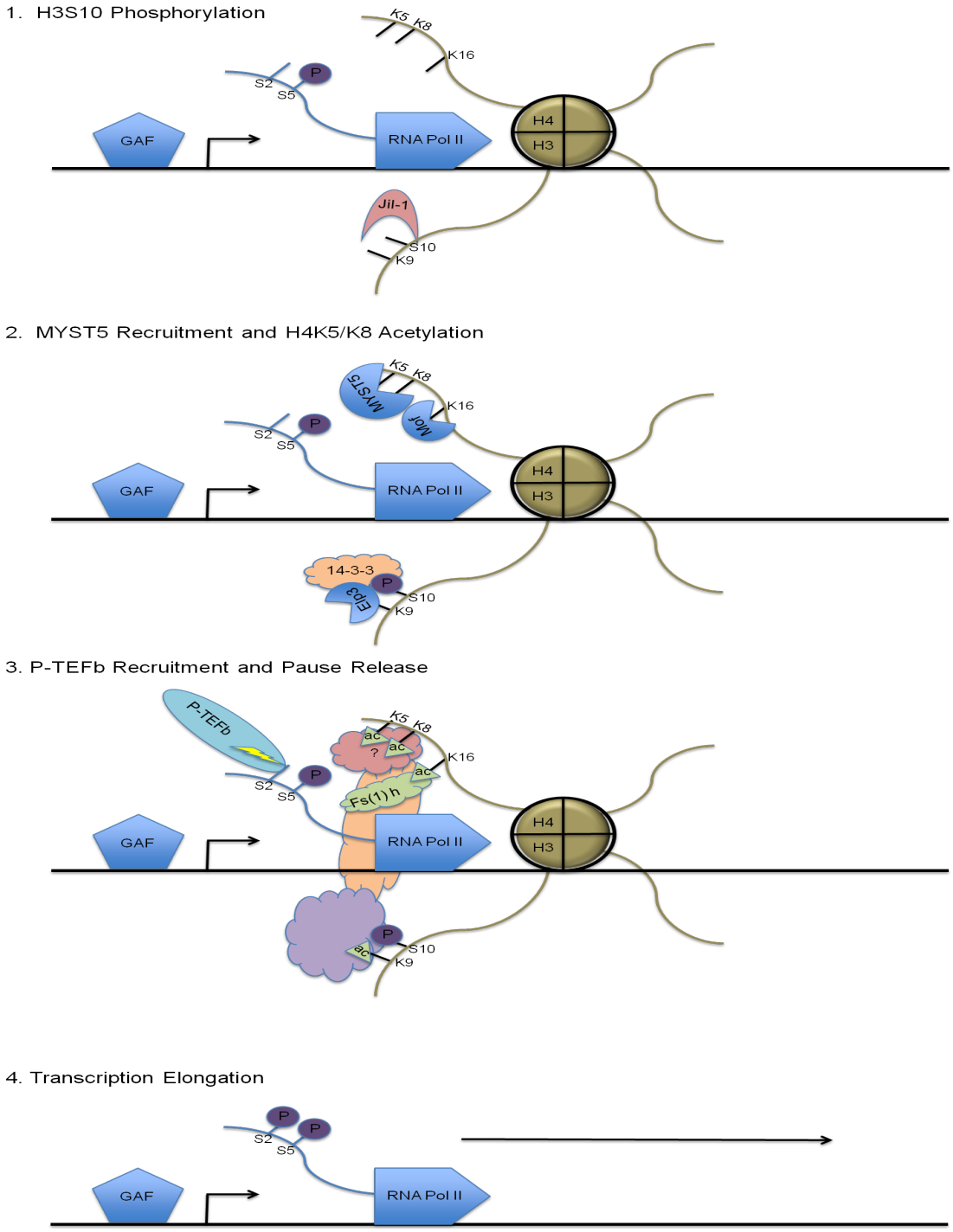


Figure 7 – Model of MYST5's role during transcription
Based on the data, MYST5 appears to function downstream of the Jil-1 kinase during transcription. Either directly or indirectly, MYST5 is essential for H4K5 and H4K8 acetylation and this mark may be essential for normal pause release. The exact mechanism by which MYST5 influences pause release is unknown, but it is clear that it plays an important role in the regulation of this event.

One interesting question that arises from this study is the mechanism by which MYST5 activity could lead to phosphorylation of RNA pol II at its serine 2 residue. One possibility by which this occurs is based on studies indicating that P-TEFb can be recruited to release polymerase pausing by chromatin binding factors such as BRD4. In this model, BRD4 or some other factor could recognize and bind to the MYST5-dependent acetylated H4K5 and H4K8 residues of nucleosome near the pause site. P-TEFb could then be recruited to the pausing event by these factors where it can phosphorylate the paused polymerase.

To test this possibility, I examined the ability of the *Drosophila* ortholog of BRD4 to bind polytene chromosomes in MYST5 null mutants and found that loss of MYST5 had no noticeable effect on the binding of either Fs(1)h isoform. Although this finding cannot completely rule out this role for Fs(1)h as it did not attempt to distinguish between Fs(1)h sites localized to active transcription from sites not associated with active transcription, it indicates that Fs(1)h may not link MYST5 activity with polymerase pause release. It is also possible that the signal generated by the long isoform of Fs(1)h could be masking any losses in binding by the short isoform. This finding does not exclude the possibility that another factor exists that can bind acetylated H4K5 and H4K8 and facilitate pause release either directly or through the recruitment of P-TEFb. To completely rule out these alternatives, experiments that examine the recruitment of P-TEFb as well as the exact localization of Fs(1)h in the absence of MYST5 are needed.

Another possibility through which MYST5 could act is through the acetylation of transcription factors, which then promote promoter pause release. Although beyond the

scope of this study, the acetylation of transcription factors by histone acetyltransferases have been documented. Specifically, the MYST histone acetyltransferase Tip60 has been shown to acetylate the c-Myc transcription factor, which can recruit P-TEFb under certain conditions (Patel, Du et al. 2004). Moreover, acetylation of p53 by Tip60 has been shown to be important for p53's ability to regulate the expression of genes involved in apoptosis (Tang, Luo et al. 2006). Currently, there is no evidence to support a role for MYST5 in acetylation of transcription factors, although this possibility could be the topic of future research.

This study also raises the question of why MYST5 binds the promoters of genes that are among the most highly paused in the *Drosophila* genome. A recent study by Gilchrist and colleagues demonstrated that the top 25 % most highly paused genes have promoter sequences that favor nucleosome occupancy (Gilchrist, Dos Santos et al. 2010). Thus, it is possible that MYST5 could participate in nucleosomal eviction at the promoters of these highly paused genes by acetylating H4K5 and H4K8. However, the findings of this study argue against such a role due to the ability of RNA pol II to initiate transcription in the absence of MYST5 (Supplementary Figure 4B,C). If MYST5 was necessary for nucleosome remodeling in the early stages of transcription, then its loss would be expected to decrease the efficiency of nucleosome remodeling and decrease transcription initiation due to the inability of RNA pol II to bind to the transcription start site.

The recruitment of MYST5 to genes could also serve as a way to differentially regulate polymerase pausing at these genes. Extensive research has shown that histone

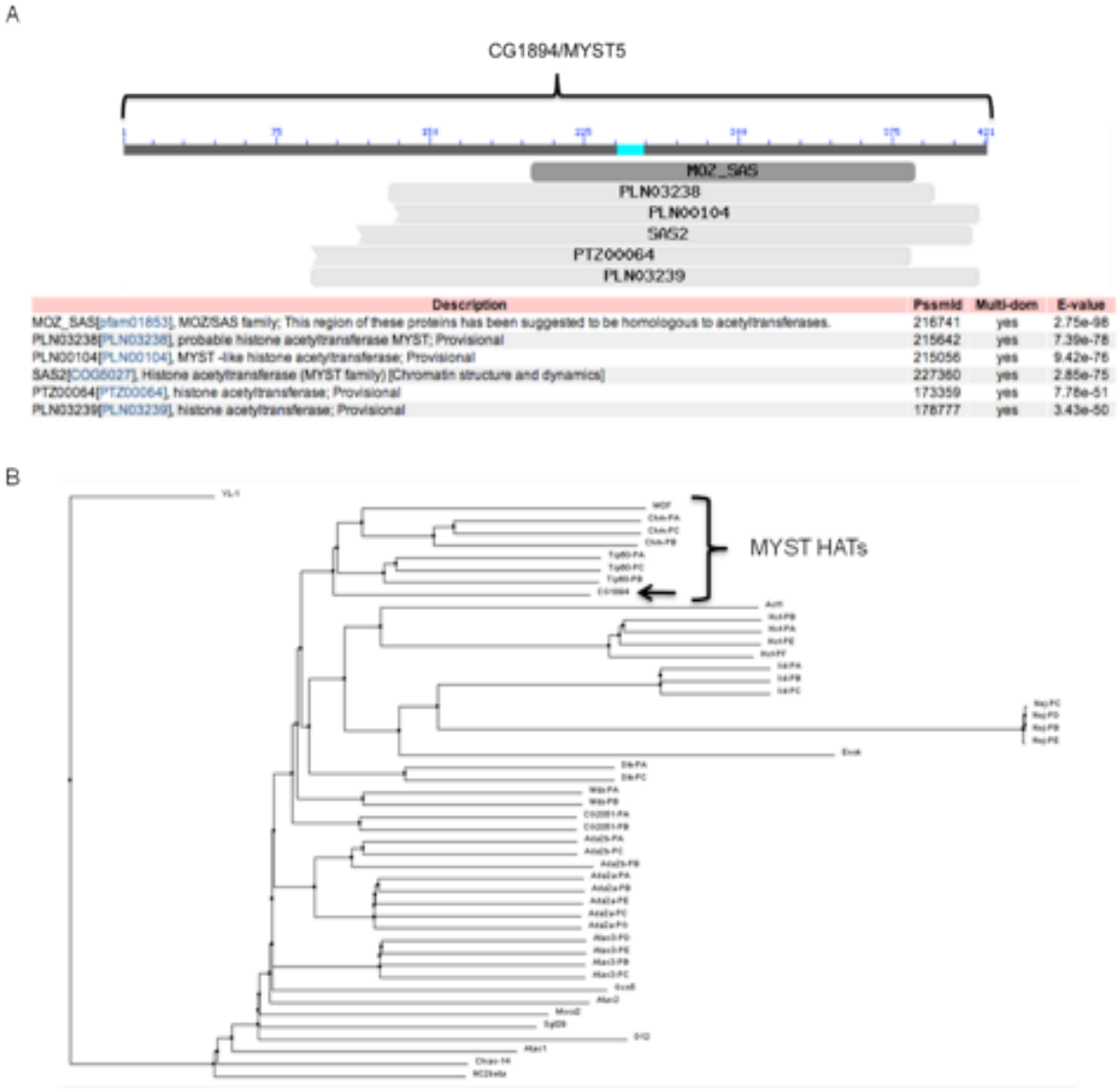
acetyltransferases rarely function in isolation. In contrast, these enzymes are commonly incorporated into large multi-subunit complexes that can play substantially different roles in gene expression depending on the specific proteins that are combined in the complex (Roth, Denu et al. 2001). Given the lack of a signature chromatin-binding domain in MYST5, it is not unreasonable to speculate that it interacts with other proteins during transcription or even that it could be incorporated into a regulatory complex. If this were to occur, then the addition of MYST5 to a multi-subunit complex at these genes could result in the formation of a regulatory complex that specifically regulates polymerase pause release at this set of genes during different biological processes. This is also plausible in light of observations that MYST5 appears to bind genes that are almost exclusively among the most highly paused genes in the genome, identifies a highly paused subset of genes bound by GAF, Myc, and Mof, and binds genes that are specifically involved in developmental processes (Table 1 and Supplementary Figure 3B). Nevertheless, how MYST5 specifically regulates pausing at these genes remains to be discovered and is the topic of future research.

Lastly, this study found that only half of MYST5 sites were present at promoters, suggesting that MYST5 may function in other regulatory processes outside of polymerase pausing (Figure 2B). It is possible that MYST5 could bind enhancer elements, but analyses of non-promoter MYST5 sites did not reveal any chromatin signatures that characterize enhancers at these sites (data not shown). Given MYST5's high degree of colocalization with GAF, which also has insulator-like functions, it may be possible that

MYST5 could function in regulating higher chromatin architectural processes. However, further analyses of these sites are needed in order to draw definitive conclusions.

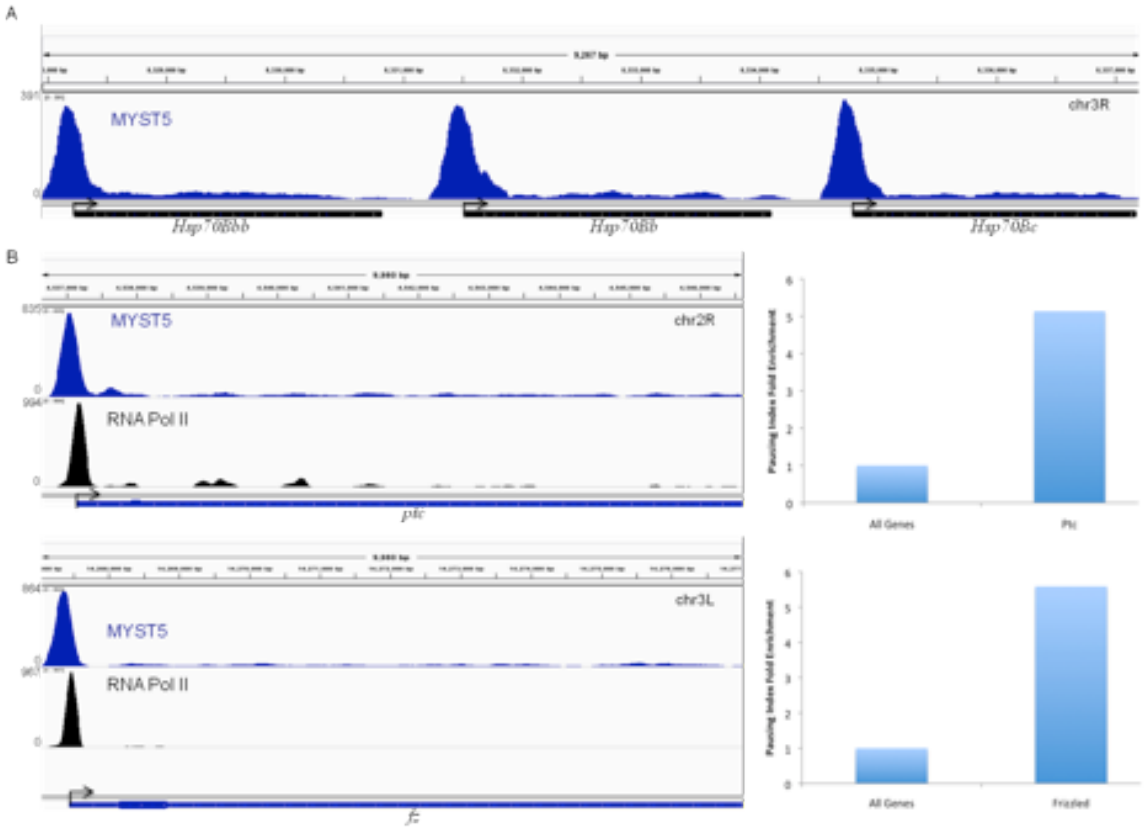
To date, five major families of histone acetyltransferases have been identified. Of the five major families of histone acetyltransferases, the MYST histone acetyltransferases constitute one of the largest and most diverse groups of histone acetyltransferases and include males absent on the first (Mof or MYST1), histone acetyltransferase bound to ORC (Hbo1 or MYST2), Monocytic leukemia zinc finger protein (Moz or MYST3), Querkopf (Qkf or MYST4), and Tat interactive protein 60 kDa (Tip60) (Sapountzi and Cote 2011). This study found that MYST5 (CG1894) possesses the catalytic MOZ/SAS domain that characterizes the MYST family of histone acetyltransferases and is necessary for the acetylation of H4K5 and H4K8 and proper transcriptional regulation. In light of these observations, the title "MYST5" is an appropriate designation for this protein and future studies of this enzyme have the potential to provide significant insight into how histone acetylation and chromatin structure allow for precise regulation of gene expression.

Supplementary Figures



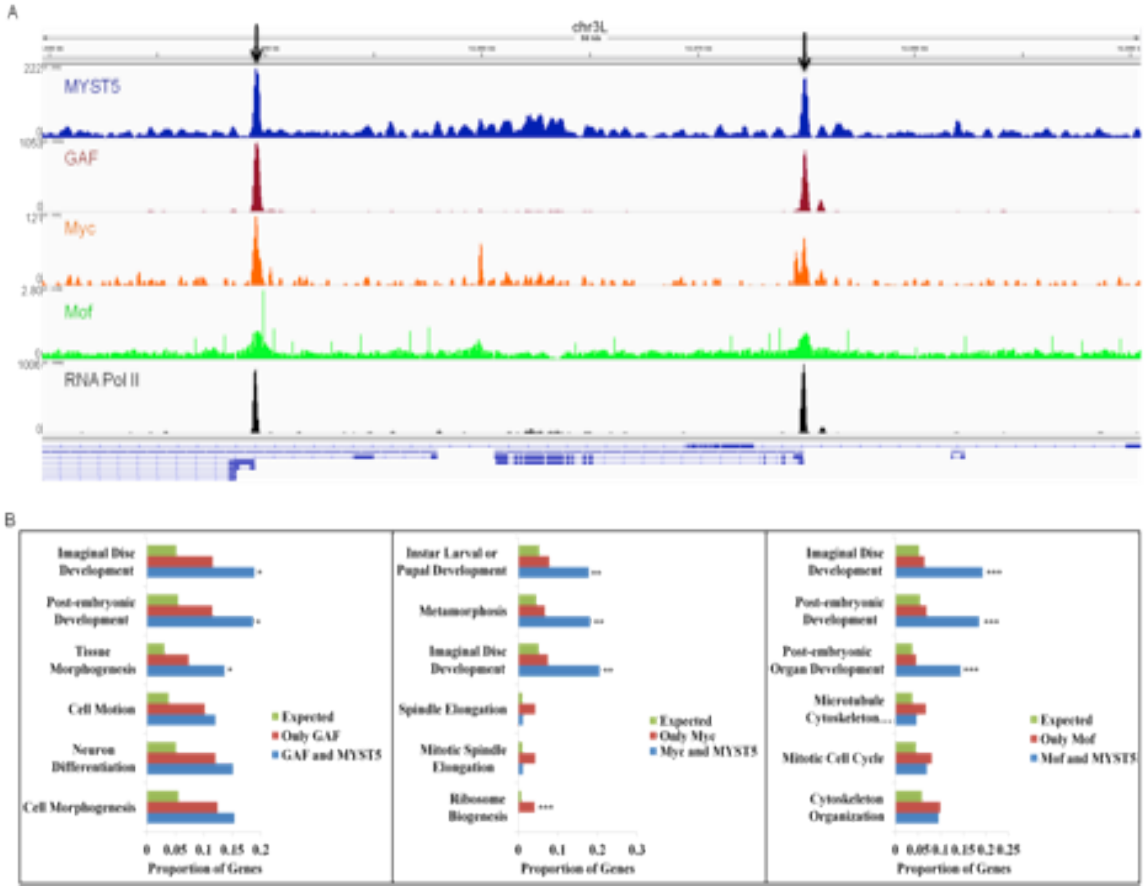
Description	Pssmid	Multi-dom	E-value
MOZ_SAS[pfam01853], MOZ/SAS family; This region of these proteins has been suggested to be homologous to acetyltransferases.	216741	yes	2.75e-98
PLN03238[PLN03238], probable histone acetyltransferase MYST; Provisional	215642	yes	7.39e-78
PLN00104[PLN00104], MYST -like histone acetyltransferase; Provisional	215056	yes	9.42e-76
SAS2[COG5027], Histone acetyltransferase (MYST family) [Chromatin structure and dynamics]	227360	yes	2.85e-75
PTZ00064[PTZ00064], histone acetyltransferase; Provisional	173359	yes	7.78e-51
PLN03239[PLN03239], histone acetyltransferase; Provisional	178777	yes	3.43e-50

Supplementary Figure 1- CG1894 is a MYST histone acetyltransferase
(A) An analysis of MYST5's amino acids adapted from the NCBI website. These data indicate that MYST5 contains a MOZ/SAS domain, which is characteristic of MYST histone acetyltransferases. (B) A sequence similarity tree showing that CG1894 (MYST5) clusters with other *Drosophila* MYST histone acetyltransferases (Mof, Tip60, and Chameau) on the basis of sequence. The arrow indicates the position of CG1894. Results were generated using Clustal Omega.

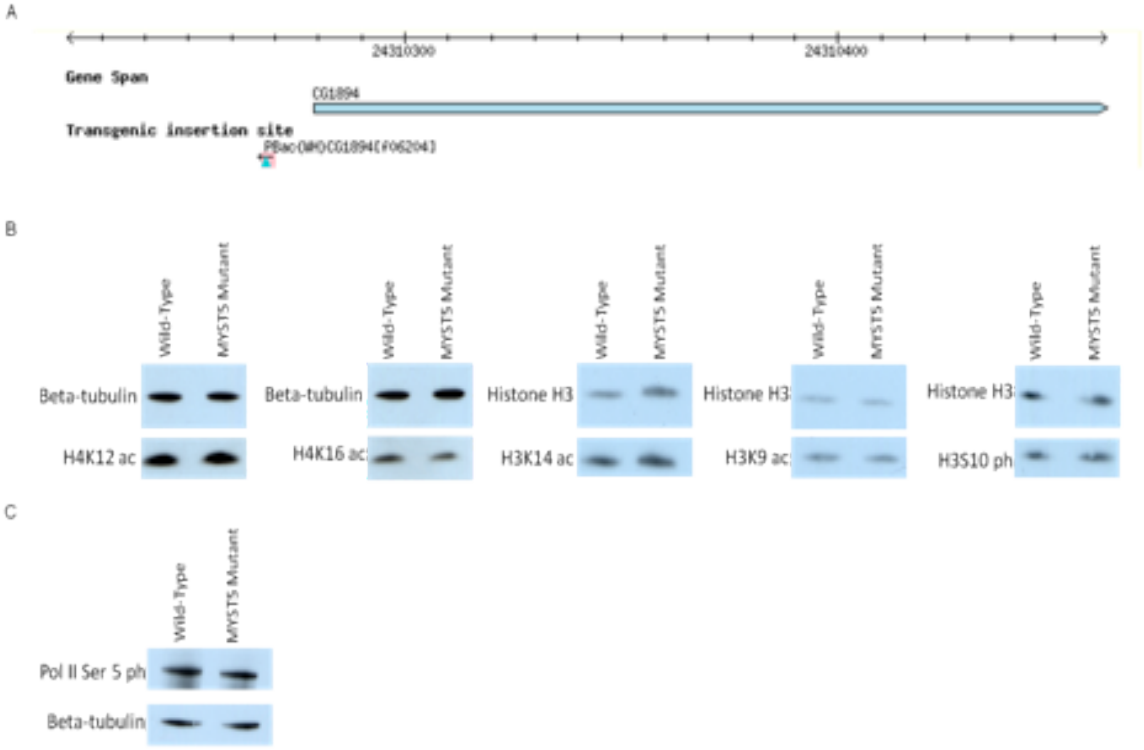


Supplementary Figure 2- MYST5 binds genes known to be paused and genes with high pausing indices

(A) A roughly 9 kb wide region on chromosome 3R showing that MYST5 localizes, from left to right, at the Heat Shock Protein 70 Bbb (HSP70 Bbb), HSP70 Bb, and HSP70 Bc loci. Extensive studies have shown that the heat shock genes are paused (Adelman and Lis 2012). (B) Two 10 kb regions of chromosome 2R and chromosome 3L showing that MYST5 binds the promoters of the *patched* (*ptc*) gene (top panel) and the *frizzled* (*fz*) gene (bottom panel). The promoters display an enrichment of RNA pol II near their transcription start site as well as pausing indices that are more than five-fold higher than the genome-wide average pausing index.



Supplementary Figure 3- MYST5 may identify subsets of genes bound by GAF, Myc, and Mof that are involved in developmental processes.
(A) A 50 kb region of chromosome 3L showing that MYST5, GAF, Myc, Mof, and RNA pol II colocalize near the promoters of genes (indicated by black arrows). (B) Gene ontology analysis of genes bound by combinations of GAF, Myc, and Mof and MYST5. These analyses show that sites with both GAF and MYST5, Myc and MYST5, and Mof and MYST5 are significantly more enriched for terms involved in development (Chi-square test, * indicates p-value < 2.0 X 10^{-4} , ** denotes p-value < 2.0 X 10^{-10}, and *** denotes p-value <2.0 X 10^{-15}).



Supplementary Figure 4- Many histone and polymerase modifications remain unchanged in larvae null for MYST5

(A) Schematic of the *CG1894* mutant allele in all larvae used. The mutants were generated via a p-element insertion upstream of the transcription start site of the *CG1894* gene (blue and pink square). Schematic was adapted from Flybase.org. (B) Western analyses showing that the levels of many histone modifications remain relatively constant in MYST5 mutant larvae. Beta-tubulin and Histone H3 were used as loading controls for experiments. (C) Pol II serine 5 phosphorylation remains relatively unchanged in most MYST5 mutant larvae although some MYST5 null larvae exhibit fluctuating levels of this modification relative to wild type larvae. Beta-tubulin was used as a loading control.

References

Adelman, K. and J. T. Lis (2012). "Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans." Nature Reviews Genetics **13**(10): 720-731.

Akoulitchev, S., T. P. Makela, R. A. Weinberg and D. Reinberg (1995). "Requirement for Tfiih Kinase-Activity in Transcription by Rna-Polymerase-Ii." Nature **377**(6549): 557-560.

Boeger, H., J. Griesenbeck, J. S. Strattan and R. D. Kornberg (2004). "Removal of promoter nucleosomes by disassembly rather than sliding in vivo." Molecular Cell **14**(5): 667-673.

Boehm, A. K., A. Saunders, J. Werner and J. T. Lis (2003). "Transcription factor and polymerase recruitment, modification, and movement on dhsp70 in vivo in the minutes following heat shock." Molecular and Cellular Biology **23**(21): 7628-7637.

Bondarenko, V. A., L. M. Steele, A. Ujvari, D. A. Gaykalova, O. I. Kulaeva, Y. S. Polikanov, D. S. Luse and V. M. Studitsky (2006). "Nucleosomes can form a polar barrier to transcript elongation by RNA polymerase II." Molecular Cell **24**(3): 469-479.

Carey, M., B. Li and J. L. Workman (2006). "RSC exploits histone acetylation to abrogate the nucleosomal block to RNA polymerase II elongation." Molecular Cell **24**(3): 481-487.

Chen, H., B. Y. Li and J. L. Workman (1994). "A Histone-Binding Protein, Nucleoplasmin, Stimulates Transcription Factor-Binding to Nucleosomes and Factor-Induced Nucleosome Disassembly." Embo Journal **13**(2): 380-390.

Clayton, A. L., C. A. Hazzalin and L. C. Mahadevan (2006). "Enhanced histone acetylation and transcription: A dynamic perspective." Molecular Cell **23**(3): 289-296.

Cosma, M. P., T. U. Tanaka and K. Nasmyth (1999). "Ordered recruitment of transcription and chromatin remodeling factors to a cell cycle- and developmentally regulated promoter." Cell **97**(3): 299-311.

Dennis, G., B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane and R. A. Lempicki (2003). "DAVID: Database for annotation, visualization, and integrated discovery." Genome Biology **4**(9).

Eberhardy, S. R. and P. J. Farnham (2002). "Myc recruits P-TEFb to mediate the final step in the transcriptional activation of the cad promoter." Journal of Biological Chemistry **277**(42): 40156-40162.

Esnault, C., Y. Ghavi-Helm, S. Brun, J. Soutourina, N. Van Berkum, C. Boschiero, F. Holstege and M. Werner (2008). "Mediator-dependent recruitment of TFIIH modules in preinitiation complex." Molecular Cell **31**(3): 337-346.

Feller, C., M. Prestel, H. Hartmann, T. Straub, J. Soding and P. B. Becker (2012). "The MOF-containing NSL complex associates globally with housekeeping genes, but activates only a defined subset." Nucleic Acids Research **40**(4): 1509-1522.

Giardine, B., C. Riemer, R. C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, W. Miller, W. J. Kent and A. Nekrutenko (2005). "Galaxy: A platform for interactive large-scale genome analysis." Genome Research **15**(10): 1451-1455.

Gilchrist, D. A., G. Dos Santos, D. C. Fargo, B. Xie, Y. A. Gao, L. P. Li and K. Adelman (2010). "Pausing of RNA Polymerase II Disrupts DNA-Specified Nucleosome Organization to Enable Precise Gene Regulation." Cell **143**(4): 540-551.

Goecks, J., A. Nekrutenko, J. Taylor and G. Team (2010). "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences." Genome Biology **11**(8).

Hilfiker, A., D. HilfikerKleiner, A. Pannuti and J. C. Lucchesi (1997). "mof, a putative acetyl transferase gene related to the Tip60 and MOZ human genes and to the SAS genes of yeast, is required for dosage compensation in Drosophila." Embo Journal **16**(8): 2054-2060.

Imbalzano, A. N., H. Kwon, M. R. Green and R. E. Kingston (1994). "Facilitated Binding of Tata-Binding Protein to Nucleosomal DNA." Nature **370**(6489): 481-485.

Ivaldi, M. S., C. S. Karam and V. G. Corces (2007). "Phosphorylation of histone H3 at Ser10 facilitates RNA polymerase II release from promoter-proximal pausing in Drosophila." Genes & Development **21**(21): 2818-2831.

Jingping Yang, E. S., Paul G. Donlin-Asp, & Victor G. Corces (2013). "A subset of *Drosophila* Myc sites remain associated with mitotic chromosomes colocalized with insulator proteins." Nature Communications **4**(1464).

Karam, C. S., W. A. Kellner, N. Takenaka, A. W. Clemmons and V. G. Corces (2010). "14-3-3 Mediates Histone Cross-Talk during Transcription Elongation in Drosophila." Plos Genetics **6**(6).

Kireeva, M. L., B. Hancock, G. H. Cremona, W. Walter, V. M. Studitsky and M. Kashlev (2005). "Nature of the nucleosomal barrier to RNA polymerase II." Molecular Cell **18**(1): 97-108.

Langmead, B., C. Trapnell, M. Pop and S. L. Salzberg (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." Genome Biology **10**(3).

Lee, C. Y., X. Y. Li, A. Hechmer, M. Eisen, M. D. Biggin, B. J. Venters, C. Z. Jiang, J. Li, B. F. Pugh and D. S. Gilmour (2008). "NELF and GAGA factor are linked to promoter-proximal pausing at many genes in Drosophila." Molecular and Cellular Biology **28**(10): 3290-3300.

Lis, J. T., P. Mason, J. Peng, D. H. Price and J. Werner (2000). "P-TEFb kinase recruitment and function at heat shock loci." Genes & Development **14**(7): 792-803.
Machanick, P. and T. L. Bailey (2011). "MEME-ChIP: motif analysis of large DNA datasets." Bioinformatics **27**(12): 1696-1697.

Melnikova, L., F. Juge, N. Gruzdeva, A. Mazur, G. Cavalli and P. Georgiev (2004). "Interaction between the GAGA factor and Mod(mdg4) proteins promotes insulator bypass in Drosophila."

Proceedings of the National Academy of Sciences of the United States of America **101**(41): 14806-14811.

Nechaev, S. and K. Adelman (2011). "Pol II waiting in the starting gates: Regulating the transition from transcription initiation into productive elongation." Biochimica Et Biophysica Acta-Gene Regulatory Mechanisms **1809**(1): 34-45.

Ni, Z. Y., A. Saunders, N. J. Fuda, J. Yao, J. R. Suarez, W. W. Webb and J. T. Lis (2008). "P-TEFb is critical for the maturation of RNA polymerase II into productive elongation in vivo." Molecular and Cellular Biology **28**(3): 1161-1170.

Ohkuma, Y. (1997). "Multiple functions of general transcription factors TFIIE and TFIIH in transcription: Possible points of regulation by trans-acting factors." Journal of Biochemistry **122**(3): 481-489.

Ohtsuki, S. and M. Levine (1998). "GAGA mediates the enhancer blocking activity of the eve promoter in the Drosophila embryo." Genes & Development **12**(21): 3325-3330.

Patel, J. H., Y. P. Du, P. G. Ard, C. Phillips, B. Carella, C. J. Chen, C. Rakowski, C. Chatterjee, P. M. Lieberman, W. S. Lane, G. A. Blobel and S. B. McMahon (2004). "The c-MYC oncoprotein is a substrate of the acetyltransferases hGCN5/PCAF and TIP60." Molecular and Cellular Biology **24**(24): 10826-10834.

Peterlin, B. M. and D. H. Price (2006). "Controlling the elongation phase of transcription with P-TEFb." Molecular Cell **23**(3): 297-305.

Peterson, C. L. and J. L. Workman (2000). "Promoter targeting and chromatin remodeling by the SWI/SNF complex." Current Opinion in Genetics & Development **10**(2): 187-192.

Rahl, P. B., C. Y. Lin, A. C. Seila, R. A. Flynn, S. McCuine, C. B. Burge, P. A. Sharp and R. A. Young (2010). "c-Myc Regulates Transcriptional Pause Release." Cell **141**(3): 432-445.

Robinson, J. T., H. Thorvaldsdottir, W. Winckler, M. Guttman, E. S. Lander, G. Getz and J. P. Mesirov (2011). "Integrative genomics viewer." Nature Biotechnology **29**(1): 24-26.

Roth, S. Y., J. M. Denu and C. D. Allis (2001). "Histone acetyltransferases." Annual Review of Biochemistry **70**: 81-120.

Rougvie, A. E. and J. T. Lis (1988). "The Rna Polymerase-Ii Molecule at the 5' End of the Uninduced Hsp70 Gene of Drosophila-Melanogaster Is Transcriptionally Engaged." Cell **54**(6): 795-804.

Saldanha, A. J. (2004). "Java Treeview-extensible visualization of microarray data." Bioinformatics **20**(17): 3246-3248.

Sapountzi, V. and J. Cote (2011). "MYST-family histone acetyltransferases: beyond chromatin." Cellular and Molecular Life Sciences **68**(7): 1147-1156.

Sievers, F., A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Z. Li, R. Lopez, H. McWilliam, M. Remmert, J. Soding, J. D. Thompson and D. G. Higgins (2011). "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega." Molecular Systems Biology **7**.

Tang, Y., J. Y. Luo, W. Z. Zhang and W. Gu (2006). "Tip60-dependent acetylation of p53 modulates the decision between cell-cycle arrest and apoptosis." Molecular Cell **24**(6): 827-839.

Tsukiyama, T., P. B. Becker and C. Wu (1994). "Atp-Dependent Nucleosome Disruption at a Heat-Shock Promoter Mediated by Binding of Gaga Transcription Factor." Nature **367**(6463): 525-532.

Wilkins, R. C. and J. T. Lis (1997). "Dynamics of potentiation and activation: GAGA factor and its role in heat shock gene regulation." Nucleic Acids Research **25**(20): 3963-3968.

Wu, C. H., Y. Yamaguchi, L. R. Benjamin, M. Horvat-Gordon, J. Washinsky, E. Enerly, J. Larsson, A. Lambertsson, H. Handa and D. Gilmour (2003). "NELF and DSIF cause promoter proximal pausing on the hsp70 promoter in Drosophila." Genes & Development **17**(11): 1402-1414.

Yang, Z. Y., J. H. N. Yik, R. C. Chen, N. H. He, M. K. Jang, K. Ozato and Q. Zhou (2005). "Recruitment of P-TEFb for stimulation of transcriptional elongation by the bromodomain protein brd4." Molecular Cell **19**(4): 535-545.

Zeitlinger, J., A. Stark, M. Kellis, J. W. Hong, S. Nechaev, K. Adelman, M. Levine and R. A. Young (2007). "RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo." Nature Genetics **39**(12): 1512-1516.

Zhang, W. S., C. Prakash, C. Sum, Y. Gong, Y. H. Li, J. J. T. Kwok, N. Thiessen, S. Pettersson, S. J. M. Jones, S. Knapp, H. Yang and K. C. Chin (2012). "Bromodomain-containing Protein 4 (BRD4) Regulates RNA Polymerase II Serine 2 Phosphorylation in Human CD4+T Cells." Journal of Biological Chemistry **287**(51): 43137-43155.

Zhang, Y., T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nussbaum, R. M. Myers, M. Brown, W. Li and X. S. Liu (2008). "Model-based Analysis of ChIP-Seq (MACS)." Genome Biology **9**(9).

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