NUCLEOSOME DYNAMICS AND ORGANIZATION IN 
*C. ELEGANS* GAMETES AND EMBRYOS

Tess Elizabeth Jeffers

A DISSERTATION 
PRESENTED TO THE FACULTY 
OF PRINCETON UNIVERSITY 
IN CANDIDACY FOR THE DEGREE 
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE 
BY THE PROGRAM IN 
QUANTITATIVE AND COMPUTATIONAL BIOLOGY

Adviser: Dr. Coleen Murphy

September 2016
ABSTRACT

Nucleosomes act as scaffolds on which DNA is structurally organized, allowing the cell to fit meters of DNA inside a micrometer-sized nucleus. In order to access the DNA sequence and perform regulatory events, however, nucleosomes must be evicted from their binding sites. A thorough description of nucleosome biology – of where nucleosomes are localized, the post-translational modifications they contain, the stability of their DNA interaction, and how these features are inherited from one generation to the next – is critical to understanding the complicated mechanism coordinating gene regulatory events. Next Generation Sequencing, and the myriad new technologies utilizing this platform, has revolutionized our understanding of gene regulation. My thesis describes the application of these technologies to C. elegans, a fabulous genetic model organism that previously lacked comprehensive genomic characterization.

Fundamental to the study of gene regulation is a thorough list of gene locations and transcription start sites. In Chapter 2, I describe work performed with the Ahringer lab to catalogue precise transcriptional start sites genome-wide in C. elegans embryos. In Chapter 3, I describe the results of the Model Organism Encyclopedia of DNA Elements project (modENCODE), a large-scale, multi-laboratory consortium effort to detail and compare the transcription factor, histone modification, and chromatin landscape among different model organisms.

Together, these works generated hundreds of publicly available datasets that will serve as the backdrop for future studies. In Chapter 4, I further investigated the dynamic nature of nucleosome stability in C. elegans embryos genome-wide. I found destabilized nucleosomes enriched at the promoters of lowly-expressed genes, and suggest that unstable nucleosomes act to poise genes for context-specific response. Finally, in Chapter 5, together with the Strome
lab I examined how nucleosomes are organized in *C. elegans* oocytes and sperm, and how chromatin modifications are inherited from gametes to the fertilized embryo.

Collectively, these experiments represent a significant improvement in our understanding of chromatin function and regulation in the model organism *C. elegans*, and how epigenetic signals are transmitted from one generation to the next. Future experiments will build on this platform to further investigate the mechanisms by which chromatin organization controls cell-type and tissue-specific events.
ACKNOWLEDGEMENTS

I would like to thank my advisers, Dr. Jason Lieb and Dr. Coleen Murphy. I am fortunate to have had two incredibly supportive labs in which to develop my skills as a computational biologist and geneticist. Jason taught me how to write and think like a scientist, and how to communicate my research in an engaging way. My PhD was long, independent, and difficult, and I am a stronger and better person today because of it. Coleen is the most dedicated and determined scientist I have ever met, and her success speaks for itself. She challenges her students to do rigorous, challenging, and exciting work that will transform human health and longevity. I am so lucky to have been a part of her lab.

I am particularly thankful for the incredible group of researchers that I encountered in the Lieb lab: I learned so much from all of you. From my first day, Erin Osborne Nishimura was a shining example of everything a biologist should be – optimistic and upbeat, hard working and dedicated, constantly curious about the wide world of biology. Her positivity is contagious. Colin Lickwar taught me everything I know about nucleosomes, and urged the lab to ponder the nature of information and scientific inquiry. His dedication to philosophy and ethics reminds us all to think deeply about the consequences of our actions. Sophia Tintori was my comrade and sounding block. She taught me the value of independence, and the many ways to pursue science, passion, and truth. Kohta Ikegami is a chromatin genius – my ChIPs would have never worked without his selfless and constant guidance. Sheera Adar taught me about the importance of family and work-life balance. Her warmth knows no bounds. Sebastian Pott devised the narrative of my paper, I can never thank him enough for lending me his scientific acumen. I should also thank Daniel McKay, for his sage wisdom and advice, and Zhuzhu Zhang and Jeremy Simon, for the fun and unwinding. Finally, I am indebted to Timothy Palpant for his tutoring in nucleosome computational analysis, and the fabulous set of tools he devised and shared with me for processing, analyzing, and visualizing high throughput sequencing data.
(https://github.com/timpalpant/java-genomics-toolkit). I would not be here today without his patience, compassion, and friendship.

Halfway through my graduate career, I was fortunate to find a second home in the Murphy lab. I am grateful for the advice of Dr. April Williams, a kindred spirit who shares my love of transcriptional regulation and doughnuts. I am so proud and amazed by Dr. Angelina Sylvain, who taught me what friendship looks like and how to have fun in Princeton. I am so excited to watch her transform the cultural landscape for women and minorities in STEM. Finally, I am thankful for my friendship with Race DiLoreto, who reminds me often never to say “I’m sorry”, to dare to have courage, and to be confident in myself. Alexi, Rachels, Nicole, Jasmine, Vanisha, Will, Siran, Cheng, Mohammad, Thomas, Rebecca, Savannah - I appreciate your kindness and welcoming support.

I would like to thank the members of the Teagle Seminar and the Princeton McGraw Center for Teaching and Learning for revitalizing my passion for teaching. The hours of conversation with Dr. Nic Voge and Dr. Geneva Stein have fueled my excitement to engage students and transform pedagogy.

None of this would have been possible without the constant love and support of my family. I would like to dedicate this work to my sister Abra, the trailblazer, who set a high bar for success; to my parents, Anne and Kevin, who taught me the value of education, perseverance, and curiosity; and finally, to my grandparents Janet and Kent, Roberta and Ed. I love you all.
## TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>x</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>xi</td>
</tr>
<tr>
<td>LIST OF ABBREVIATIONS</td>
<td>xiii</td>
</tr>
<tr>
<td>CHAPTER 1 - AN INTRODUCTION TO C. ELEGANS AND TRANSCRIPTIONAL REGULATION</td>
<td>1</td>
</tr>
<tr>
<td>NUCLEOSOME OCCUPANCY AND POSITIONING</td>
<td>1</td>
</tr>
<tr>
<td>DNA sequence influences on nucleosome positioning</td>
<td>2</td>
</tr>
<tr>
<td>ATP-dependent chromatin remodelers influence nucleosome positions</td>
<td>3</td>
</tr>
<tr>
<td>Nucleosome organization is conserved across eukaryotes</td>
<td>3</td>
</tr>
<tr>
<td>Nucleosome stability influences the dynamics of gene expression</td>
<td>4</td>
</tr>
<tr>
<td>An MNase digestion timecourse can identify unstable nucleosomes genome-wide</td>
<td>5</td>
</tr>
<tr>
<td>C. ELEGANS AS A MODEL ORGANISM FOR TRANSCRIPTIONAL REGULATION</td>
<td>7</td>
</tr>
<tr>
<td>Gene regulation in C. elegans</td>
<td>8</td>
</tr>
<tr>
<td>Transcription start sites in C. elegans</td>
<td>9</td>
</tr>
<tr>
<td>TRANSGENERATIONAL INHERITANCE OF CHROMATIN MODIFICATIONS</td>
<td>9</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>11</td>
</tr>
<tr>
<td>CHAPTER 2 - THE LANDSCAPE OF RNA POLYMERASE II TRANSCRIPTION INITIATION IN C. ELEGANS REVEALS PROMOTER AND ENHANCER ARCHITECTURES</td>
<td>15</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>16</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>17</td>
</tr>
<tr>
<td>RESULTS</td>
<td>19</td>
</tr>
<tr>
<td>Identification of C. elegans RNA polymerase II transcription initiation sites</td>
<td>19</td>
</tr>
<tr>
<td>Chromatin features of transcription initiation sites</td>
<td>22</td>
</tr>
<tr>
<td>Core promoter elements are correlated with different initiation types</td>
<td>24</td>
</tr>
<tr>
<td>Characterization of bidirectional transcription</td>
<td>27</td>
</tr>
<tr>
<td>Most transcription factor binding sites initiate transcription</td>
<td>28</td>
</tr>
<tr>
<td>Characterization of a novel widespread regulatory architecture</td>
<td>29</td>
</tr>
<tr>
<td>DISCUSSION</td>
<td>34</td>
</tr>
<tr>
<td>MATERIALS AND METHODS</td>
<td>37</td>
</tr>
<tr>
<td>Preparation of nuclear capped RNA-seq libraries</td>
<td>37</td>
</tr>
<tr>
<td>Mapping, clustering of 5' end tags, and shape score</td>
<td>38</td>
</tr>
<tr>
<td>TSS assignments</td>
<td>39</td>
</tr>
<tr>
<td>Motif plots</td>
<td>39</td>
</tr>
<tr>
<td>Defining enhancers</td>
<td>40</td>
</tr>
<tr>
<td>Data sets, processing, and visualization</td>
<td>41</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>43</td>
</tr>
<tr>
<td>CHAPTER 3 - COMPARATIVE ANALYSIS OF METAZOAN CHROMATIN ORGANIZATION</td>
<td>49</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>50</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>50</td>
</tr>
<tr>
<td>RESULTS</td>
<td>51</td>
</tr>
<tr>
<td>DISCUSSION</td>
<td>64</td>
</tr>
<tr>
<td>MATERIALS AND METHODS</td>
<td>65</td>
</tr>
<tr>
<td>Preprocessing of ChIP-seq data</td>
<td>65</td>
</tr>
<tr>
<td>Preprocessing of ChIP-chip data</td>
<td>66</td>
</tr>
</tbody>
</table>
NuclEsOnome FragIliGitY is associatEd wIth future transcriptional response to developmenTal cues and stress in C. elegans

ABSTRACT
INTRODUCTION
RESULTS
A digestion timecourse identifies nucleosomes with differential MNase sensitivity
Trans-factors increase nucleosome fragility
Nucleosome fragility increases throughout heat-shock genes upon induction
Nucleosome fragility or resistance is associated with stereotypic nucleosome locations within genes
Nucleosome fragility and resistance is not correlated with nucleosome occupancy
Nucleosome fragility near genes is anti-correlated with expression
Nucleosome fragility is correlated to cis-encoded DNA features
Nucleosome fragility is not associated with a specific epigenetic state
Fragile nucleosomes are associated with genes expressed in context-specific situations

DISCUSSION
MNase resistant nucleosomes tend to be in gene bodies and correlated with expression
MNase sensitive fragile nucleosomes are 5’ enriched and anti-correlated with expression

MATERIALS AND METHODS
Worm strains and growth in liquid culture
MNase Digestion Time Course
Heat shock
RNA isolation
Illumina Library Preparation
Illumina Sequencing and Post-Processing ................................................................. 124
Nucleosome analysis .............................................................................................. 124
Nucleosome Fragility and Resistance scores ......................................................... 125
Nucleosome occupancy scores ............................................................................. 126
Genes associated with fragile or resistant nucleosomes ....................................... 126
Gene ontology analysis ......................................................................................... 126
Stable and developmentally regulated genes ......................................................... 127
RNA-seq analysis ................................................................................................. 127
REFERENCES ......................................................................................................... 128

CHAPTER 5 - ............................................................................................................ 133
CHROMATIN PACKAGING IN C. ELEGANS SPERM AND OOCYTES ...................... 133
ABSTRACT ............................................................................................................... 134
INTRODUCTION ....................................................................................................... 135
RESULTS .................................................................................................................. 137
   Isolating pure populations of C. elegans sperm and oocytes ............................... 137
   Transcriptome of C. elegans gametes ................................................................. 138
   Nucleosome organization in C. elegans sperm and oocytes ............................... 141
   Characterization of histone isoform usage in mature sperm and oocytes ............ 146
DISCUSSION ............................................................................................................ 147
MATERIALS AND METHODS ................................................................................ 148
   N2 worm liquid culture and mixed stage embryo harvest ................................. 148
   MNase digestion ................................................................................................. 149
   Paired-end sequencing analysis ........................................................................ 149
   Isolating pure populations of C. elegans sperm and oocytes ............................... 150
   RNA isolation and transcriptome analysis ........................................................ 151
   Chromatin immunoprecipitation ...................................................................... 151
REFERENCES ......................................................................................................... 153

CHAPTER 6 - ............................................................................................................ 156
HOW HISTONE LEVELS CONTROL DEVELOPMENTAL EVENTS, CONCLUDING REMARKS, AND FUTURE OUTLOOK .......................................................... 156
ABSTRACT ............................................................................................................... 156
PERSPECTIVE ......................................................................................................... 158
   Chromatin changes during differentiation ......................................................... 158
   Pioneer TFs remodel and activate the quiescent genome ................................. 159
   Intrinsically unstable nucleosomes may play a role in ZGA ............................... 160
   Zelda bookmarks chromatin for genome activation ......................................... 160
   Canonical nucleosome organization ............................................................... 161
   Alternative mechanisms proposed to control the timing of ZGA ....................... 161
   Histone levels affect transcription in general .................................................... 162
   Competition between nucleosomes and TFs regulates gene activation ............. 163
   Histone levels are regulated with age ............................................................... 164
CONCLUSIONS AND FUTURE OUTLOOK ............................................................ 165
REFERENCES ......................................................................................................... 167
LIST OF TABLES

Table 2-1. Summary of Transcription Initiation Cluster (TIC) assignments..........................22
Table 2-2. GO-term analysis of genes associated with enhancers.......................................36
Table 3-1: Summary of key shared and organism-specific chromatin features in human, fly and
worm........................................................................................................................................54
LIST OF FIGURES

FIGURE 1-1. THE NUCLEOSOME IS THE FUNDAMENTAL REPEATING UNIT OF EUKARYOTIC CHROMATIN...... 2
FIGURE 1-2. SCHEMATIC OF NUCLEOSOME POSITIONS INFERRED USING MNase-SEQ. ....................... 4
FIGURE 1-3. THE STABILITY OF THE NUCLEOSOME-DNA INTERACTION CAN INFLUENCE TRANSCRIPTIONAL ACTIVATION OR REPRESSION .................................................................................. 5
FIGURE 1-4. DIFFERENT LENGTHS OF MNase DIGESTION CAN BE USED TO PREFERENTIALLY ISOLATE FRAGILE (GREEN) OR RESISTANT (ORANGE) NUCLEOSOMES. ............................................................. 6
FIGURE 2-1. TRANSCRIPTION INITIATION REGIONS (TICS) ARE OCCUPIED BY RNA POL II AND ENRICHED NEAR THE 5’ END OF CODING GENES. .................................................................................................. 20
FIGURE 2-2. EXAMPLES OF INITIATION AND ELONGATION PATTERNS AT TRANS-SPliced GENES .......... 21
FIGURE 2-3. CHROMATIN FEATURES OF PROTEIN CODING, ncRNA AND UNASSIGNED TSS ASSIGNMENT CLASSES...................................................................................................................... 23
FIGURE 2-4. COMPARISON OF capRNA TSSs TO WORMBASE TRANSCRIPT START SITES. .................... 24
FIGURE 2-5. TRANSCRIPTION INITIATION CLUSTERS (TICS) DISPLAY A RANGE OF SHAPES. ................. 25
FIGURE 2-6. PROTEIN-CODING PROMOTER SEQUENCE FEATURES. .................................................. 26
FIGURE 2-7. BIDIRECTIONAL TRANSCRIPTION IS WIDESPREAD IN C. ELEGANS.............................. 27
FIGURE 2-8. BIDIRECTIONAL TRANSCRIPTION AT DIVERGENT AND TANDEMLY ORIENTED PROTEIN CODING GENES. ...................................................................................................................... 28
FIGURE 2-9. TF-BINDING REGIONS INITIATE BIDIRECTIONAL TRANSCRIPTION ................................... 29
FIGURE 2-10. INTERGENIC TF-BINDING REGIONS OVERLAPPING TICS ARE ENRICHED FOR ENHANCER-LIKE CHROMATIN SIGNATURE. ....................................................................................... 30
FIGURE 2-11. VALIDATED EMBRYONIC ENHANCER REGIONS CONTAIN ENHANCERS IDENTIFIED BY OVERLAP BETWEEN TF-BINDING SITES AND TICS. ................................................................. 31
FIGURE 2-12. SEQUENCE AND CHROMATIN FEATURES OF ENHANCER TSSs. .................................... 32
FIGURE 2-13. DEFINITION OF A NOVEL REGULATORY ARCHITECTURE................................................. 33
FIGURE 3-1. OVERVIEW OF THE DATA SET. .......................................................................................... 52
FIGURE 3-2. CHROMATIN FEATURES AT TSSs AND GENE BODIES, AND CO-OCCURRENCE OF HISTONE MODIFICATIONS. .............................................................................................................. 53
FIGURE 3-3. SHARED AND ORGANISM-SPECIFIC CHROMATIN STATES .................................................. 56
FIGURE 3-4. GENOME-WIDE ORGANIZATION OF HETEROCHROMATIN.................................................. 58
FIGURE 3-5. HISTONE MODIFICATIONS IN HETEROCHROMATIN ............................................................ 60
FIGURE 3-6. ORGANIZATION OF SILENT DOMAINS .............................................................................. 63
FIGURE 4-1. AN MNase DIGESTION TIME COURSE ON C. ELEGANS EMBRYOS................................. 91
FIGURE 4-2. ACCOMPANYING FIGURE 4-1. .............................................................................................. 92
FIGURE 4-3. ACCOMPANYING FIGURE 4-1. .............................................................................................. 93
FIGURE 4-4. COMPETITION WITH TRANSCRIPTION FACTORS INFLUENCES NUCLEOSOME FRAGILITY ..... 96
FIGURE 4-5. ACCOMPANYING FIGURE 4-4. .............................................................................................. 98
FIGURE 4-6. HEAT SHOCK INCREASES NUCLEOSOME FRAGILITY AT THE PROMOTER AND GENE BODY OF UPREGULATED GENES .................................................................................. 99
FIGURE 4-7. ACCOMPANYING FIGURE 4-6. .............................................................................................. 101
FIGURE 4-8. FRAGILITY IS ENRICHED 5’ AND 3’ OF GENES AND IS ANTI-CORRELATED WITH GENE EXPRESSION. ............................................................................................................................... 103
FIGURE 4-9. ACCOMPANYING FIGURE 4-8. .............................................................................................. 106
FIGURE 4-10. ACCOMPANYING FIGURE 4-8. ............................................................................................ 107
FIGURE 4-11. FRAGILE NUCLEOSOMES CONTAIN AT-rich, NUCLEOSOME DISFAVORING, AND POORLY CONSERVED DNA. .......................................................................................................... 109
FIGURE 4-12. ACCOMPANYING FIGURE 4-11. .......................................................................................... 111
FIGURE 4-13. FRAGILE NUCLEOSOMES ARE ENRICHED AT GENES THAT WILL BE EXPRESSED IN THE FUTURE AND IN SPECIFIC CONTEXTS. .............................................................................. 114
Figure 4-14. We propose a model whereby nucleosome fragility is determined by two distinct mechanisms, one that operates in cis at all genes, and one that operates in trans at a subset of genes.

Figure 5-1. Isolation of pure populations of sperm and oocytes from fem-1;fer-15 females and him-5 males.

Figure 5-2. qRT-PCR was used to analyze identity and purity of isolated gametes.

Figure 5-3. Cross-correlation analysis of oocyte and sperm RNA-seq replicates.

Figure 5-4. Comparison of transcripts identified with different techniques.

Figure 5-5. Differential expression analysis of genes enriched in sperm and oocytes.

Figure 5-6. Gene Ontology Analysis of transcripts enriched in C. elegans oocytes and sperm.

Figure 5-7. C. elegans gamete nucleosome organization is distinct from mixed stage embryo chromatin.

Figure 5-8. Addressing functional significance of gamete chromatin organization in the early embryo.

Figure 5-9. Gamete-enriched histone variants.
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BP</td>
<td>Base Pair</td>
</tr>
<tr>
<td>ChIP</td>
<td>Chromatin Immunoprecipitation</td>
</tr>
<tr>
<td>ChIP-seq</td>
<td>Chromatin Immunoprecipitation followed by High-Throughput Sequencing</td>
</tr>
<tr>
<td>CTD</td>
<td>C-terminal Domain</td>
</tr>
<tr>
<td>MNase</td>
<td>Micrococcal Nuclease</td>
</tr>
<tr>
<td>KB</td>
<td>Kilobase Pairs</td>
</tr>
<tr>
<td>HOT</td>
<td>High Occupancy of Transcription Factors</td>
</tr>
<tr>
<td>HSF</td>
<td>Heat Shock Factor</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase Chain Reaction</td>
</tr>
<tr>
<td>ncRNA</td>
<td>Non-Coding RNA</td>
</tr>
<tr>
<td>NDR</td>
<td>Nucleosome Depleted Region</td>
</tr>
<tr>
<td>NELF</td>
<td>Negative Elongation Factor</td>
</tr>
<tr>
<td>NFR</td>
<td>Nucleosome Free Region</td>
</tr>
<tr>
<td>RNA Pol II</td>
<td>RNA Polymerase II</td>
</tr>
<tr>
<td>TBP</td>
<td>TATA Binding Protein</td>
</tr>
<tr>
<td>TFIID</td>
<td>General Transcription Factor II D</td>
</tr>
<tr>
<td>TF</td>
<td>Transcription Factor</td>
</tr>
<tr>
<td>TFBS</td>
<td>Transcription Factor Binding Site</td>
</tr>
<tr>
<td>TIC</td>
<td>Transcription Initiation Cluster</td>
</tr>
<tr>
<td>TN</td>
<td>Terminal Nucleosome</td>
</tr>
<tr>
<td>TSS</td>
<td>Transcription Start Site</td>
</tr>
<tr>
<td>UTR</td>
<td>Untranslated Region</td>
</tr>
</tbody>
</table>
Genetic information is stably encoded in deoxyribonucleic acid (DNA), which is inherited from parent cell to daughter cell, from one generation to the next. In order to grow or divide, cells transcribe segments of DNA called genes into messenger RNA (mRNA), and translate mRNA into proteins. Proteins are the building blocks used to grow, divide, or differentiate a skin cell from a muscle cell. The many steps along this process (DNA → RNA → Protein) are susceptible to regulation, and can be broadly described as the regulation of gene expression.

Of particular interest is the function of a set of proteins called histones. Together, two copies each of histone H2A, H2B, H3, and H4 wrap ~ 147 bp of DNA to form the nucleosome. Nucleosomes act as scaffolds on which DNA is structurally organized, allowing the cell to fit meters of DNA inside a micrometer-sized nucleus. In order to access the DNA sequence and perform regulatory events, however, nucleosomes must be evicted from their binding sites. A thorough description of nucleosome biology – of where nucleosomes are localized, the post-translational modifications they contain, the stability of their DNA interaction, and how these features are inherited from one generation to the next – is critical to understanding the complicated mechanism coordinating gene regulatory events.

NUCLEOSOME OCCUPANCY AND POSITIONING

147 bp of negatively charged DNA is wrapped ~ 1.7 superhelical turns around a positively charged histone octamer to form the nucleosome (Luger et al. 1997). The DNA entry and exit site from the nucleosome is often sealed by the linker histone, histone H1, to cover
approximately 160 bp of DNA (Happel and Doenecke 2009). Intervening between adjacent nucleosomes is linker DNA, which can range in size from 20 – 90 bp between different organisms and cell types (Collepardo-Guevara and Schlick 2014).

Figure 1-1. The nucleosome is the fundamental repeating unit of eukaryotic chromatin. Adapted from Darryl Leja, NHGRI. Approximately 147bp of DNA is wrapped 1.7 superhelical turns around an octamer of histone proteins. Nucleosomes are then coiled together to form the 30nm fiber, and eventually the chromosome.

DNA sequence influences on nucleosome positioning

Nucleosome positions determine the transcriptional output of the genome, by determining where and when DNA binding factors can reach their target sites. Nucleosome positions are determined through a combination of cis and trans factors. The intrinsic sequence preference encoded in cis for nucleosome formation can be tested using in vitro reconstitution of nucleosomes. In this experiment, genomic DNA is isolated, histone proteins are removed, and the DNA is digested using 4-base restriction enzymes, or randomly sheared using sonication.
Purified histone proteins are added to the sheared DNA, and salt concentrations are slowly increased until stable histone octamers form on sequence-favoring nucleosome positions. These experiments have shown that nucleosomes prefer to form on GC rich DNA, and fail to form on AT rich DNA or poly(dA:dT) stretches (Segal et al. 2006).

**ATP-dependent chromatin remodelers influence nucleosome positions**

Although the DNA sequence plays a major role in determining nucleosome positions, current evidence suggests that the majority of nucleosomes in vivo are instead positioned by chromatin remodelers (Clapier and Cairns 2009). Chromatin remodelers utilize ATP to slide, disassemble and reassemble, or otherwise influence nucleosome positions. In many cases, DNA sequence and *trans* factors work together to coordinate nucleosome positions. For example, although high AT content is encoded upstream of many genes, chromatin remodelers like CHD-1 in yeast play a critical role in maintaining the integrity of the upstream nucleosome depleted region (NDR) (Pointner et al. 2012; Smolle et al. 2012). At other locations, *trans* factors dominate: nucleosomes reconstituted in whole cell extract only form phased nucleosome arrays in the context of ATP, indicating that active chromatin remodeling is necessary for the formation of canonical nucleosome organization (Zhang et al. 2011).

**Nucleosome organization is conserved across eukaryotes**

All eukaryotic genomes encode histones and use nucleosomes to achieve higher order genome condensation in the nucleus. Similarly, nucleosome organization around coding regions is to some extent conserved across all species. The 5’ and 3’ end of genes are generally nucleosome-depleted, indicative of nucleosome eviction events for the purpose of transcriptional regulation (Ranjan et al. 2013). The first nucleosome downstream from the transcription start site (TSS) is called the +1 nucleosome. Nucleosomes occur in regularly spaced units, called nucleosome arrays, which are phased with respect to the +1 and terminal nucleosome (TN) (Jiang and Pugh 2009). The coding region of transcribed genes often shows
well positioned, and highly phased nucleosomes, while silenced genes and intergenic regions are typified by poorly positioned “fuzzy” nucleosomes. These data indicate that polymerase, chromatin remodelers, or other active processes are responsible for maintaining nucleosome positions (Venters and Pugh 2009).

Nucleosome positions are typically inferred using the enzyme micrococcal nuclease (MNase) (Axel 1975). MNase cleaves unprotected DNA, leaving only nucleosome-bound DNA intact. Nucleosome positions can be inferred genome-wide by isolating the mononucleosomal DNA fragments and preparing them for next generation sequencing using standard DNA library preparation methods and paired-end sequencing (Figure 1-2) (Kaplan et al. 2009). Reads are mapped back to the genome, and read pileups are quantified to assign a nucleosome occupancy score to each base pair in the genome (Gossett and Lieb 2012).

![Diagram of nucleosome positions inferred using MNase-seq.](image)

**Figure 1-2.** Schematic of nucleosome positions inferred using MNase-seq.

**Nucleosome stability influences the dynamics of gene expression**

Nucleosome-DNA binding dynamics can influence transcriptional activation and repression (Figure 1-3).
Breathing is an intrinsic feature of nucleosomes that potentiates accessibility to transcription factor binding sites. DNA at the nucleosome entry and exit site undergoes transient unwrapping: in the exposed state, regulatory proteins can bind and further unwrap DNA from the nucleosome core (Ettig et al. 2011). An alternative model proposes that cooperative binding of transcription factors to the DNA can outcompete nucleosomes to expose regulatory regions (Mirny 2010). Nucleosome stability has classically been studied in vitro using sensitivity to enzymatic digestion or salt concentration (Bloom and Anderson 1978; Burton et al. 1978; Li et al. 1993; Polach and Widom 1995; 1999; Jin and Felsenfeld 2007; Wu and Travers 2004). Genome-wide adaptations of these methods have been used to identify nucleosome position and stability in vivo.

An MNase digestion timecourse can identify unstable nucleosomes genome-wide

Studies in yeast, Drosophila, plants, and mammals have used varying concentrations of the enzyme micrococcal nuclease (MNase) to identify nucleosomes with differential sensitivity to MNase digestion (Xi et al. 2011; Weiner et al. 2010; Kubik et al. 2015; Henikoff et al. 2011; Chereji et al. 2015; Vera et al. 2014; Lombraña et al. 2013). Similar to the use of low and high
MNase concentrations, chromatin digestion with MNase for varying lengths of time should also probe nucleosomes of varying DNA stabilities (Figure 1-4). Nucleosomes sensitive to low concentrations of MNase have been labeled as “fragile”, and have been associated with transcription factor binding sites (Vera et al. 2014), active origins of replication (Lombrana et al. 2013), gene promoters (Xi et al. 2011), and genomic sequences with high AT content (Chereji et al. 2015).

![Figure 1-4](image.png)

**Figure 1-4. Different lengths of MNase digestion can be used to preferentially isolate fragile (green) or resistant (orange) nucleosomes.**

Thus, both DNA-encoded sequence features and *trans* factors influence nucleosome fragility. However, the functional implication of nucleosome fragility remains unclear. For example, one study reported fragile nucleosomes at the promoters of repressed stress-response genes during normal growth (Xi et al. 2011), while another found fragile nucleosomes at the promoters of highly transcribed genes in yeast (Kubik et al. 2015). In Chapter 4, I
investigated fragile nucleosomes in *C. elegans* embryos. We were interested in understanding the relationship between fragility and gene activity in a developing multicellular organism. We found fragile nucleosomes associated with lowly expressed genes and genes expressed in a context-specific fashion. My data suggest that the fragility of nucleosome-DNA interactions acts to poise genes for induction in response to stress or developmental cues.

**C. ELEGANS AS A MODEL ORGANISM FOR TRANSCRIPTIONAL REGULATION**

Sydney Brenner famously proposed to his advisor that he “would like to tame a small metazoan organism to study development directly” (Letter to Max Perutz from Sydney Brenner, 1963). Because of his pioneering work, *C. elegans* has served as an excellent genetic and developmental model organism. *C. elegans* is primarily a hermaphrodite, with males occurring sporadically at a low percent in the population. Most beneficial for the study of development, *C. elegans* has an invariant lineage, where every animal develops from a fertilized egg into an adult with the same number of cells (Sulston et al. 1983).

*C. elegans* was the first model organism to have its genome sequenced, and subsequent transcriptomic analyses have thoroughly characterized gene expression during worm development, aging, response to stress, pathogens, and after loss of important genetic regulators (Hillier et al. 2005; Murphy 2012; Baugh 2003; Rodriguez et al. 2013). However, knowledge of the chromatin and epigenetic landscape of the worm has lagged behind other model organisms like Drosophila and yeast, where the study of gene regulation was spearheaded. In Chapter 3, a large consortium effort called the Model Organism Encyclopedia of DNA Elements (modENCODE) collectively generated over 650 datasets of histone, transcription factors, and RNA molecules at different life stages (Ho et al. 2014). These data will serve as the foundation for further investigation into the mechanism of transcriptional regulation in *C. elegans*. 

7
**Gene regulation in C. elegans**

Many of the principles learned from yeast and Drosophila apply to transcriptional regulation *C. elegans*. In yeast, transcription start sites are demarcated by the following factors: (1) a well-positioned +1 nucleosome, (2) an upstream nucleosome depleted region, (3) the DNA sequence TATAAA (TATA box), or (4) an initiator element YYANWYY (Lenhard et al. 2012). Together, these components direct TBP, TFIIID, and other general transcription factors to assemble at the promoter. In addition to the TATA box and IR elements, Kozak sequences, Sp1-like sites, T-blocks, and SL1-sites are frequently found in the promoters of *C. elegans* (Grishkevich et al. 2011).

RNA polymerase II assembles with the general transcription factors at the promoter to form the pre-initiation complex (PIC). Pol II release from the PIC is regulated by nucleoside triphosphates, strand separation and formation of the open promoter complex, and phosphorylation of the Pol II C-terminal domain (CTD). After initiation, Pol II then elongates into the gene body, generating mRNA. In higher eukaryotes, Pol II can pause after elongating a small number of nucleotides, generating a polymerase that is poised for action under context-specific induction. Pausing is thought to be enriched at genes involved in signal-response (Gilchrist et al. 2010).

Pol II post-recruitment regulation is also widely used in lower eukaryotes like yeast, Drosophila, and worms. During yeast stationary phase, Pol II is found docked on the promoters of early growth response genes in the absence of transcription (Radonjic et al. 2005). In Drosophila, Pol II binds the HSP70 locus, initiates and elongates ~ 20-40bp, before pausing and waiting for further heat shock signals (Buckley et al. 2014). Similarly, in *C. elegans*, a significant number of stress-response and growth genes exhibit Pol II pausing and “docking”, wherein Pol II accumulates at the promoter without first initiating (Maxwell et al. 2014). Although the majority of Pol II regulation is conserved in worms, *C. elegans* lack the negative elongation factor NELF.
(Baugh et al. 2009), and thus the factor responsible for inhibiting Pol II elongation remains unknown (Adelman and Lis 2012).

**Transcription start sites in C. elegans**

In addition to the absence of NELF, the other transcriptional oddity that occurs in C. elegans is called trans-splicing. Trans-splicing replaces the 5’ UTR with a 22 nucleotide leader sequence (SL1) for the large majority of Pol II transcribed genes (Allen et al. 2011; Krause and Hirsh 1987). Trans-splicing obscures the 5’ end of mature polyadenylated mRNAs, and thus the gene annotations historically used for C. elegans do not reflect the exact base pair position of transcription initiation. Lack of precise transcriptional start sites has hindered the comprehensive study of chromatin and gene regulatory mechanisms in C. elegans. In Chapter 2, together with the Ahringer lab, we sequenced RNA molecules containing 5’ caps (CapSeq) to define the locations of transcription start sites in C. elegans embryos. These data indicate that transcription initiation in C. elegans is much more similar to other eukaryotes than previously thought (Chen et al. 2013).

**TRANSGENERATIONAL INHERITANCE OF CHROMATIN MODIFICATIONS**

One question of particular relevance is how chromatin patterns are initiated in the early embryo and how they change during development. The early embryo represents a totipotent state from which all subsequent cell lineages are derived. As differentiation proceeds, totipotency is lost as cells commit to their various fates. Regulation of chromatin structure is a key mechanism by which cellular identity is determined and maintained. Additional restrictions on the chromatin landscape occur through nucleosome positioning, higher order chromatin structure, histone variants and posttranslational modifications. Much of the work in the developmental genomics field has centered on describing the gene expression and chromatin changes that arise as an organism develops from an embryo to an adult. However, our
understanding of how chromatin landscapes are reset for the next generation is just beginning to unfold.

In animals, germ cell progenitors are set-aside during the final stages of embryogenesis. After reaching sexual maturity, meiosis produces sperm and oocytes in preparation for fertilization. A genetically tractable system like *C. elegans* is needed to further understand the role of chromatin and nucleosome positions in these gametes, and how information is transferred and used in the next generation. In Chapter 5, together with the Strome Lab, we developed methods to isolate pure populations of *C. elegans* sperm and oocytes in large quantities. We used these gametes to profile nucleosome locations, RNA enrichment, and histone post-translational modifications. We found that *C. elegans* oocytes contain chromatin landscapes reminiscent of their transcriptional history in the germline, yet also poised for embryonic development. Future work will investigate the mechanism by which poising is established and its function for developmental robustness.
REFERENCES


Chapter 2 -

The landscape of RNA polymerase II transcription initiation in *C. elegans* reveals promoter and enhancer architectures

Author's Statement:

The work in Chapter 2 was previously published in *Genome Research* (Chen et al., 2013). I assisted in this work by participating in experimental design, embryo harvest, chromatin immunoprecipitation, library preparation, and sequencing analysis for HTZ-1. The majority of the work was performed by the Ahringer lab, which also wrote the manuscript.

Reference:

ABSTRACT

RNA polymerase transcription initiation sites are largely unknown in Caenorhabditis elegans. The initial 5’ end of most protein-coding transcripts is removed by trans-splicing, and noncoding initiation sites have not been investigated. We characterized the landscape of RNA Pol II transcription initiation, identifying 73,500 distinct clusters of initiation. Bidirectional transcription is frequent, with a peak of transcriptional pairing at 120 bp. We assign transcription initiation sites to 7691 protein-coding genes and find that they display features typical of eukaryotic promoters. Strikingly, the majority of initiation events occur in regions with enhancer-like chromatin signatures. Based on the overlap of transcription initiation clusters with mapped transcription factor binding sites, we define 2361 transcribed intergenic enhancers. Remarkably, productive transcription elongation across these enhancers is predominantly in the same orientation as that of the nearest downstream gene. Directed elongation from an upstream enhancer toward a downstream gene could potentially deliver RNA polymerase II to a proximal promoter, or alternatively might function directly as a distal promoter. Our results provide a new resource to investigate transcription regulation in metazoans.
INTRODUCTION

Transcription is a fundamental process that plays a central role in development and cellular responses. The regulation of transcription is complex, often involving interplay between promoters and regulatory elements such as enhancer regions. A prerequisite for understanding the mechanisms and principles of this process requires identification of these sites and their relationships.

*Caenorhabditis elegans* is a widely used and powerful model organism for functional genomic studies. The genome of 100 MB is well-annotated and 30× smaller than that of humans, making genomic studies highly accessible (The *C. elegans* Sequencing Consortium 1998). However, despite these advantages, study of transcription regulation has been hampered because the transcription initiation sites of most *C. elegans* genes are unknown. More than 70% of protein-coding genes are *trans*-spliced to a 22-nt leader RNA encoded from another region of the genome (Allen et al. 2011; Blumenthal 2012). The region from the initial 5′ end to the *trans*-splice site, termed the “outron,” is spliced off and degraded (Bektesh and Hirsh 1988; Blumenthal 1995). Consequently, the annotated start sites for most *C. elegans* genes mark *trans*-splice sites rather than transcription initiation sites because cDNAs and RNA-seq data derived from total RNA indicate the structure of the mature *trans*-spliced transcript. Using these positions or translational start sites as transcription start sites (TSSs), some promoter features have been described (Ooi et al. 2006, 2010; Gerstein et al. 2010; Liu et al. 2011b). However, the true sites and features of RNA polymerase II transcription initiation and the chromatin landscape at these sites remain to be discovered.

In eukaryotes, regulatory and chromatin features of protein-coding TSSs have been extensively documented (de Hoon and Hayashizaki 2008; Jiang and Pugh 2009; Venters and Pugh 2009). For example, transcription initiation sites often possess an initiator element (Inr),
and a subset of genes have a TATA-box 25 to 32 nt upstream of the TSS which has a role in focusing start-site selection (Weis and Reinberg 1992; Smale 1997; Juven-Gershon et al. 2008; Lenhard et al. 2012). An Sp1 motif has also been reported to play a role in transcript initiation (Segal et al. 1999; Wierstra 2008). Active promoters usually contain a nucleosome-depleted region (NDR) flanked by well-positioned nucleosomes (Bernstein et al. 2004; Lee et al. 2004; Nishida et al. 2006; Henikoff 2007; Jiang and Pugh 2009; Rhee and Pugh 2012). In addition, nucleosomes near active protein-coding TSSs are enriched for trimethylation of H3K4 and acetylation of H3K27 (Turner 2000; Barrera and Ren 2006). Furthermore, histone variants H3.3 and H2A.Z, associated with unstable nucleosomes, are also enriched in promoter regions (Zhang et al. 2005; Millar et al. 2006; Raisner and Madhani 2006; Whittle et al. 2008; Weber et al. 2010).

Recent high-throughput sequencing studies in other organisms have uncovered novel transcriptional phenomena at promoters and enhancers. For example, in yeast and mammalian cells, it has been observed that transcription at protein-coding promoters often initiates bidirectionally, whereas bidirectional transcription is infrequent in Drosophila (Core et al. 2008; Neil et al. 2009; Seila et al. 2009; Xu et al. 2009; Nechaev et al. 2010; Flynn et al. 2011; Kharchenko et al. 2011; Wei et al. 2011). How bidirectional transcription initiation drives productive transcription elongation primarily in the coding direction is not well understood (Seila et al. 2008, 2009; Flynn et al. 2011), but a recent report has suggested that gene loops can play a role (Tan-Wong et al. 2012). In addition, it has been shown that mammalian enhancer regions are transcribed, but transcription initiation within enhancer regions has not been well characterized (De Santa et al. 2010; Kim et al. 2010; Melgar et al. 2011; Ong and Corces 2011; Preker et al. 2011; Wang et al. 2011). The mechanism and functions of transcription at enhancers are unclear. Proposed models suggest regulatory functions or, alternatively, that the process of transcription could act to generate regions of open chromatin in order to enhance the
accessibility of downstream coding promoters (Travers 1999; Ong and Corces 2011; Natoli and Andrau 2012).

Here, we globally identify sites of RNA polymerase II transcription initiation and elongation in C. elegans. We assign transcription start sites to protein-coding and noncoding genes and find that bidirectional transcription is widespread. Furthermore, we discover 2361 mapped transcription factor binding sites that overlap transcription initiation clusters, defining a set of transcribed enhancers. We show that productive transcription elongation across these regions is usually oriented toward and in the same orientation as that of the nearest downstream gene. Our characterization of this regulatory architecture and provision of transcription initiation and elongation resources will facilitate studies of transcription regulation in metazoans.

RESULTS

Identification of C. elegans RNA polymerase II transcription initiation sites

To globally capture RNA polymerase II initiation sites, including nascent transcripts before they have been trans-spliced, we isolated short (20–100 nt) nuclear RNAs with a 5’ cap from mixed-stage embryos, and prepared and sequenced RNA-seq libraries (hereafter called short cap RNAs). We also made strand-specific RNA-seq libraries from capped nuclear RNAs over 200 nt long to assess nuclear expression of elongated transcripts (hereafter called long cap RNAs). In total, we obtained ~50 million uniquely mappable reads for the short capped RNA libraries and ~55 million uniquely mappable reads for the long capped RNA libraries. As expected for RNA polymerase II transcription initiation sites, there is strong enrichment for Ser5 phosphorylated RNA Pol II (the initiating form) near the 5’ ends of short cap RNA reads (Figure 2-1A).
Figure 2-1. Transcription initiation regions (TICs) are occupied by RNA Pol II and enriched near the 5’ end of coding genes. (A) Heat map displays of forward strand TICs (n = 36,662) ranked by enrichment of Pol II Ser5 ChIP-chip signal. All panels are plotted for the enrichment of signal in 2-kb windows centered at cap RNA TSSs. (B) Genome browser views of short cap RNA (orange) and long cap RNA (green) signals at a non-trans-spliced and a trans-spliced gene. (*) The outron region. (C) The fraction sites where at least two cap RNA starts were analyzed relative to WormBase transcript start sites of trans-spliced (blue) and non-trans-spliced (red) genes are plotted in an 800-bp window. (D) Mononucleosome signal (adult) anchored at either cap RNA TSSs (red) or WormBase transcript starts (gray).

We observed that short cap RNA reads often mapped near the 5’ ends of genes. For non-trans-spliced genes, reads initiated at or near the annotated starts of WormBase transcripts (Fig. 2-1B,C). In contrast, for trans-spliced genes, where the WormBase transcript starts usually mark the trans-splice site, short cap RNA reads initiated upstream, and we observed long cap RNA signals covering the outron region between the trans-splice site and the short cap RNA tags (Fig. 2-1B,C). The outron signals are not observed in RNA-seq libraries prepared from whole-cell (mainly cytoplasmic) poly(A)+ RNAs (Fig. 2-2). Initiation sites are broadly distributed upstream of trans-spliced transcripts (Fig. 2-1C), indicating that transcription does not initiate at a fixed position relative to trans-splice sites.
As in other organisms, we observed that the initiation sites of short cap RNAs are usually closely clustered (Fig. 2-2). We grouped the 5’ tags into clusters, which defined 73,500 RNA polymerase II transcription initiation clusters (TICs) (see Methods; Table 2-1, See also [Chen et al., 2013] Supplementary Table 2). To link TICs to annotated WormBase genes, TICs mapping from −200 bp to +100 bp of an annotated WormBase transcript start, or where long cap RNA signal was present between an annotated transcript start and a TIC, were assigned as potential transcription start sites for protein-coding genes ($n = 10,106$ TICs assigned to 7691 genes) (Table 2-1, See also [Chen et al., 2013] Supplementary Table 2). A further 14,810 TICs mapped within the body of protein-coding genes and might mark alternative 5′ ends or have regulatory functions, and 3637 TICs were assigned to ncRNA genes. The remaining 44,947 TICs were unassigned. Unassigned TICs map on the anti-sense strand in gene bodies or not within −200 bp to +100 bp of an annotated transcript start (from −200 bp to +100 bp).
**Table 2-1. Summary of Transcription Initiation Cluster (TIC) assignments**

<table>
<thead>
<tr>
<th></th>
<th>Number of TICs</th>
<th>Number of assigned genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coding TSS</td>
<td>10106</td>
<td>7691</td>
</tr>
<tr>
<td>Coding Body</td>
<td>14810</td>
<td>6192</td>
</tr>
<tr>
<td>ncRNA</td>
<td>3637</td>
<td>2603</td>
</tr>
<tr>
<td>unassigned</td>
<td>44947</td>
<td>-</td>
</tr>
</tbody>
</table>

**Chromatin features of transcription initiation sites**

To characterize the identified TICs, we analyzed the distributions of chromatin features that have well-characterized properties at promoters, transcribed regions, and enhancers using heat map plots. We selected the mode of each TIC as a representative TSS for analysis and ranked the TSSs based on the ratio of H3K4me3 to H3K4me1 ChIP signals to more easily visualize promoter and enhancer characteristics (Guenther et al. 2007; Heintzman and Ren 2007; Gerstein et al. 2010; Pekowska et al. 2011; Rada-Iglesias et al. 2011).

TSSs assigned to protein-coding genes display chromatin features typical of promoters in other organisms. Near the TSSs, we observed generally higher H3K4me3 than H3K4me1 and high levels of H3K27ac (Fig. 2-3), a signature of active protein-coding promoters (Ruthenburg et al. 2007; Lenhard et al. 2012). Productive elongation detected by long cap coverage is observed downstream from the assigned coding TSSs, and H3K36me3 signal, a mark of transcriptional elongation of protein-coding genes (Guenther et al. 2007; Sati et al. 2012), is usually also present (Fig. 2-3). Regions with the highest levels of H3K4me3 appear to be divergent promoters, as these additionally display H3K36me3 and reverse strand long cap RNA signals upstream of the TSSs.
Figure 2-3. Chromatin features of protein coding, ncRNA and unassigned TSS assignment classes. Enrichment signals of the indicated datasets was visualized by heat map for TIC modes of protein coding TSSs (n=5,206), ncRNA TSSs (n=1,726) and unassigned TSSs (n=22,241). Each dataset was ranked by H3K4me3/H3K4me1 ratio. The majority of protein coding TSSs (64%) has higher H3K4me3 than H3K4me1 whereas unassigned and ncRNA TSSs have the opposite pattern: 57% of unassigned and 64% of ncRNA TSSs have higher H3K4me1 than H3K4me3. The indicated signals are plotted in 2kb windows centered at TIC modes from only forward strand TSSs.

To further validate the protein-coding gene TSS assignments, we compared the patterns of chromatin signals anchored at the starts of WormBase transcripts to the same set of genes anchored at the newly assigned cap RNA TSSs (Fig. 2-1D; Fig. 2-4). We observed that promoter-associated chromatin signals are sharper at short cap TSSs (Fig. 2-4). Additionally, plotting nucleosome signal at short cap RNA TSSs reveals a wave-like pattern, indicating that nucleosomes are well-phased at C. elegans promoters (Fig. 2-1D; Fig. 2-4), as observed in other eukaryotic cells (Bernstein et al. 2004; Lee et al. 2004; Nishida et al. 2006; Henikoff 2007; Jiang and Pugh 2009; Rhee and Pugh 2012). Such positioning is not clear using WormBase TSSs.
Figure 2-4. Comparison of capRNA TSSs to Wormbase transcript start sites. Genes in the top 40% expression band and having an assigned capRNA TSS were used. The distribution of embryo HTZ-1, H3, and mononucleosome signal was plotted centered either at short cap RNA TSSs (blue) or Wormbase transcript start sites (green) for the same set of genes.

In contrast to the cap RNA TSSs assigned to protein-coding genes, unassigned TSSs and those assigned to noncoding RNA (ncRNA) genes are enriched for chromatin features typical of enhancers, with higher H3K4me1 relative to H3K4me3 and low levels of H3K36me3 (Fig. 2-3). ncRNA and unassigned TSSs are frequently embedded within long cap RNA signal, indicating they lie in regions with productive transcriptional elongation (Fig. 2-3). To summarize, TSSs assigned to protein-coding genes show conventional promoter chromatin signatures, whereas ncRNA and unassigned TSSs are enriched for enhancer-like chromatin signatures.

Core promoter elements are correlated with different initiation types

We observed that the TICs have a range of patterns, from a single strong peak to a broad distribution of initiation events (Fig. 2-5), as has been seen in humans and Drosophila (Carninci et al. 2006; Ni et al. 2010). We developed a shape score, which we defined as the fraction of reads associated with the highest point in the TIC, to categorize TICs.
containing at least 10 tags (see Methods). A sharp TIC where all tags map to the same position has a score of 1, whereas a broad TIC containing multiple starts of similar heights will have a low score. We defined sharp promoters as TICs with scores of 0.7–1.0 and broad ones as those between 0 and 0.2 (Fig. 2-5). The remaining TICs have intermediate characteristics.

Figure 2-5. Transcription initiation clusters (TICs) display a range of shapes. Plots display the distribution of short capRNA tags in 100 bp windows centered at the TIC mode (X-axis). Tags counts are shown on the Y-axis. Shape scores are calculated as the ratio of the number of tags at the mode TSS relative to the total number TSSs in the cluster (mode weight). Examples of shapes in each 0.1 score bin are shown, with the percentage and number of TICs in each shape bin shown at the right.
To look for fixed sequence features in the protein-coding promoters, we aligned TSSs of sharp and broad promoters using the WebLogo sequence analysis tool (Fig. 2-6A; Crooks et al. 2004). As in other organisms (Zhang and Dietrich 2005; Carninci et al. 2006), we found that transcription preferentially initiated at the adenine of a TCA sequence, with a stronger preference in sharp compared to broad promoters. In addition, in sharp promoters, there is a weak enrichment for T/A-rich sequence between −25 to −30, corresponding to the region where TATA-boxes reside (Juven-Gershon et al. 2008).

![Figure 2-6. Protein-coding promoter sequence features. (A) WebLogo analysis of sequence aligned at sharp and broad TSSs. (B) Inr, TATA, and Sp1 motif occurrence at sharp and broad protein-coding TSSs.](image)

To further investigate core promoter sequence features, we examined the occurrence of Inr and TATA motifs as well as that of Sp1, a widely active transcription factor, using published position weight matrices (Lenhard et al. 2012). We found that the Inr motif was present in both types of protein-coding promoters, with slightly higher frequency in sharp promoters (Fig. 2-6B). In contrast, the TATA motif was only associated with sharp promoters (Fig. 2-6B); a bias for
TATA occurrence in sharp promoters was also reported in human and *Drosophila* (Ni et al. 2010; Rach et al. 2011). Similar to the TATA motif, we found that the Sp1 motif was associated primarily with sharp promoters (Fig. 2-6B). Of note, sharp TSSs with a TATA box were substantially less likely to harbor an Inr motif at the TSS than those without a TATA box, suggesting some degree of redundancy between TATA and Inr (*P* = 5 × 10^{-15}; χ² test).

**Characterization of bidirectional transcription**

In the heat map analyses, we observed that reverse strand TSSs were often located a short distance upstream of forward strand TSSs, indicative of bidirectional transcription initiation (Fig. 2-3). To investigate the occurrence of bidirectional transcription, we plotted the density of forward and reverse strand short cap initiation sites across the three classes of TSSs. This analysis showed a strong peak of reverse strand signal ~120 bp upstream of the initiation sites of all TSS classes (protein-coding, noncoding, and unassigned) (Fig. 2-7); 39% of all TICs have an upstream reverse strand initiation within 200 bp. The bidirectional signal observed for protein-coding genes is not simply due to divergent protein-coding transcription (head to head configuration, where the upstream gene is transcribed in the opposite orientation) because it is also clearly evident at tandemly transcribed protein-coding genes (tail to head, where the upstream gene is in the same orientation) (Fig. 2-8). We conclude that transcription initiation in *C. elegans* is often bidirectional.

![Figure 2-7. Bidirectional transcription is widespread in C. elegans.](image)

*Figure 2-7. Bidirectional transcription is widespread in C. elegans.* The frequency of forward (red) and reverse (blue) transcription initiations were analyzed at (A) coding, (B) ncRNA, and (C) unassigned TSSs. The distribution of signal is plotted in 1-kb windows at mode TSSs of the TICs.
Figure 2-8. Bidirectional transcription at divergent and tandemly oriented protein coding genes. The density of forward (red) and reverse (blue) strand TSSs was plotted in a 1kb window centered at the mode TSSs of protein coding TICs.

**Most transcription factor binding sites initiate transcription**

Recent studies in mammalian cells identified transcripts and engaged RNA polymerase II within enhancer regions, indicating that these regions are transcriptionally active (De Santa et al. 2010; Kim et al. 2010; Melgar et al. 2011; Wang et al. 2011; Natoli and Andrau 2012). However, where transcription initiates within enhancer regions and the relationship between transcription factor binding sites and enhancer transcription initiation are unclear.

To investigate these questions, we used a set of *C. elegans* transcription factor (TF)-binding regions defined from ChIP mapping of 22 transcription factors (Gerstein et al. 2010). We focused on “factor-specific regions” bound by 1–4 TFs (n= 13,156, 79% of binding sites), as these are enriched for containing expected sequence motifs and being located near genes known to be under their regulation. As expected, we observed that most factor-specific TF regions have a classical enhancer signature of high H3K4me1 and low H3K4me3 (Fig. 2-9B), in contrast to protein-coding promoters, which have the opposite pattern (Fig. 2-9A).
Figure 2-9. TF-binding regions initiate bidirectional transcription. Comparison of the enrichment of Pol II (Ser5), H3K4me3, and H3K4me1 at TSSs of ubiquitously expressed protein-coding genes (n = 4282) (A), and factor-specific TF-binding regions (n = 13,516) (B). Heat map rows are ranked by H3K4me3/H3K4me1 ratio and all signals plotted in 2-kb windows centered at (A) TIC mode TSSs, or (B) the midpoint of TF-binding regions. (C) Forward (red) and reverse (blue) TIC coverage plotted in a 2-kb region centered at the midpoints of factor-specific TF-binding regions; 60% of sites have both forward and reverse strand signals.

Using our short and long cap RNA data, we next examined the relationship between transcription factor binding, transcription initiation, and productive transcription elongation. We first asked whether transcription initiates near TF-binding sites. Remarkably, we observed that 65% of factor-specific TF-binding regions overlap a TIC. Therefore, TF-binding sites frequently initiate transcription. To further explore the transcription initiation activity of TF-binding sites, we plotted the frequency and positions of transcription initiation events relative to TF-binding sites. We observed peaks of transcription initiation closely flanking the TF-binding sites, with 60% of transcribed binding sites having both forward and reverse strand short cap RNA signals (Fig. 2-9B,C). RNA Pol II is also clearly enriched at TF-binding regions (Fig. 2-9B). These results demonstrate that bidirectional transcription is initiated at TF-binding sites.

Characterization of a novel widespread regulatory architecture

We used the criterion of overlap between transcription factor binding regions and transcription initiation sites to define active enhancers (Melgar et al. 2011). We first selected intergenic TF-binding sites that were at least 500 bp from an annotated transcript start and that overlapped a TIC, then removed regions that appeared to be unannotated promoters based on having a signature of high H3K4me3/low H3K4me1 (Fig. 2-10). These intergenic TF-binding
regions that initiate transcription define 2361 transcribed enhancers. Supporting the view that these sites identify active enhancers (Fig. 2-11), they overlap previously published functionally validated embryonic enhancer regions (Landmann et al. 2004; Lei et al. 2009).

Figure 2-10. Intergenic TF-binding regions overlapping TICs are enriched for enhancer-like chromatin signature. Heat map analysis for the indicated enrichment signals displayed as 8 k-means clustered groups based on H3K4me3 and H3K4me1 (n = 3137). All signals were plotted in 2-kb windows centered at the midpoint of TF-binding regions. A promoter-like signature of high H3K4me3/low H3K4me1 is found in clusters 1–4 (n = 776). In contrast, clusters 5–8 show an enhancer-like chromatin signature, high H3K4me1 and low/under-enriched H3K4me3 (n = 2361).
Validated embryonic enhancer regions contain enhancers identified by overlap between TF-binding sites and TICs. (A) Enhancer regions identified in lin-26/lir-1 region (Landmann et al. 2004) and (B) hlh-1 (Lei et al. 2009) contain TF-binding sites that overlap TICs and that show an enhancer-like chromatin signature of high H3K4me1/low H3K4me3.

We find that features of enhancer transcription initiation sites differ from those of protein-coding promoter TSSs. First, although enhancer TSSs are enriched for the Inr initiation element, TATA and Sp1 elements are essentially absent (Fig. 2-12). Second, chromatin features also differ at enhancer and promoter TSSs. In addition to differences in relative levels of H3K4me3 and H3K4me1 described above, HTZ-1 is strongly enriched at protein-coding TSSs, but nearly absent at enhancer TSSs (Fig. 2-12C). Further, whereas histone H3 and mononucleosomes are depleted at protein-coding TSSs, these show weak enrichment at enhancer TSSs (Fig. 2-12C). In summary, promoter sequence elements and the chromatin signature of enhancer TSSs differ from those of protein-coding genes.
We next examined whether transcription initiation at enhancers was productively elongated or, alternatively, whether initiation events generally produced short transcripts. To answer this question, we analyzed the abundance of nuclear long cap RNA signals (generated from nuclear capped RNAs > 200 nt) on each strand of intergenic enhancers. We find that enhancers generally produce elongated transcripts and that there is a striking bias for
transcription to be predominantly on one strand (Fig. 2-13A,B). Therefore, although most enhancers initiate bidirectional transcription, productive elongation is primarily in one direction.

**Figure 2-13. Definition of a novel regulatory architecture.** (A) Genome browser image of signals indicating enhancers, TICs (red: forward strand; blue: reverse strand), H3K4me3, H3K4me1, whole-cell (mainly cytoplasmic) RNA, and nuclear long cap RNA from the plus and minus strands, with gene annotation below. (B) RNA Pol II elongation across enhancers is in the same orientation as that of the nearest downstream gene. Plots show forward strand long cap RNA-seq read count (log_{10} scale) relative to reverse strand signal over enhancers upstream of forward (red) or reverse (blue) strand tandem genes. The red dots showing enhancers upstream of the tol-1 gene (forward strand) shown in A were circled. (C) Proposed model illustrating the architecture and potential regulatory roles of transcribed enhancers.

Since enhancers are often located upstream of their target genes, we investigated the relationship between enhancer and downstream gene transcription. We assigned intergenic enhancers to the nearest downstream protein-coding gene, and asked whether the direction of transcription across the enhancer was related to that of the downstream gene. To be confident in the assignment of the downstream gene, we limited this analysis to intergenic enhancer regions upstream of tandemly transcribed genes, excluding divergently transcribed genes. Remarkably, for 90% of enhancer regions, the majority of elongated transcription is in the same
orientation as that of the downstream gene (Fig. 2-13B). We also found that the level of transcription across enhancers is positively correlated with that of the downstream gene ($\rho = 0.45$, Spearman's $P < 10^{-7}$). Together, these results support the view that transcribed enhancers are active regulatory regions for downstream gene expression (Fig. 2-13C).

DISCUSSION

Here, we characterized the global landscape of RNA polymerase II transcription initiation in *C. elegans*. As in other eukaryotic cells (Lenhard et al. 2012), we find that transcription initiation in *C. elegans* is often bidirectional. In addition to identifying transcription initiation sites at protein-coding genes, we also documented abundant intergenic and intronic transcription initiation sites and defined thousands of active enhancers based on the transcriptional activity of mapped transcription factor-binding sites. The global identification of transcription start sites will facilitate future studies of gene expression regulation in *C. elegans* and allow comparative analyses with other eukaryotes.

Mammalian enhancer regions have also been shown to be transcribed, but this phenomenon is not well-characterized, and its function is not yet known (De Santa et al. 2010; Kim et al. 2010; Koch et al. 2011; Melgar et al. 2011; Wang et al. 2011; Natoli and Andrau 2012). In addition, both unidirectional and bidirectional transcription has been observed. The relationships between the different reports of mammalian enhancer transcription are unclear because the studies used different cell lines and methods and assayed different types of sites: DNAse I hypersensitive sites, CBP-binding sites, TF-binding sites, or regions with enhancer-like chromatin signatures. In *C. elegans*, we observed that although transcriptional initiation at enhancers is usually bidirectional, productive elongation across them is predominantly in one direction, usually oriented toward the downstream gene. These enhancer transcripts are either unstable or not exported from the nucleus, as they are primarily apparent in libraries made from nuclear RNA and usually undetectable in libraries generated from whole-cell poly(A)+ (mainly
cytoplasmic) RNA. We note that since the binding sites for only a small percentage of *C. elegans* transcription factors are available, the set of enhancers identified here is unlikely to be complete.

Oriented transcription across enhancers may seem to be at odds with the classical view that enhancers act in an orientation-independent manner. However, oriented enhancer function has been documented (Swamynathan and Piatigorsky 2002; Hozumi et al. 2013). Furthermore, although transcription across *C. elegans* enhancer regions is oriented, we do not know whether the enhancer function of these binding sites is oriented as well. For example, orientation could be determined through interactions with other regulatory elements associated with the downstream gene (e.g., the promoter).

Transcription from enhancer regions toward proximal promoters might create an open chromatin environment (Hirota et al. 2008) or could potentially deliver RNA polymerase II to the promoter (Vernimmen et al. 2011). Another possibility is that upstream enhancers could function as alternative promoters for protein-coding genes. Such a function would be in line with a recent report in mouse showing that active upstream intergenic enhancers can function as alternative tissue-specific promoters (Kowalczyk et al. 2012). *C. elegans* genes associated with upstream transcribed enhancers are enriched for having developmental, signaling, or neuronal functions, types of genes that usually show tissue or temporally restricted expression, and underenriched for genes with general cellular functions (Table 2-2).

**Table 2-2.** GO-term analysis of genes associated with enhancers. Enhancers were assigned to the nearest downstream genes. Genes were analysed for enrichment or underenrichment of GO terms using GOstat (significance p<0.01) (Beissbarth and Speed 2004) (http://gostat.wehi.edu.au/). These GO lists were reduced based on functional redundancies by REVIGO (Supek et al. 2011) (http://revigo.irb.hr/).
Direct tests will be necessary to address whether enhancers have promoter activity because trans-splicing would remove upstream transcribed regions from the mature transcript. However, if upstream enhancers do function as promoters, they appear to have different characteristics than proximal promoters (Fig. 2-12). In the future, it will be of interest to functionally analyze enhancer transcription and explore the similarities and differences in regulatory architecture between different organisms. Our finding of a novel widespread
regulatory architecture in *C. elegans* and generation of new resources will facilitate future studies of gene expression regulation in metazoans.

**MATERIALS AND METHODS**

**Preparation of nuclear capped RNA-seq libraries**

Worms were grown in liquid culture and two million adults bleached to obtain mixed staged embryos as previously described (Stiernagle 2006). Nuclei were isolated from embryos using the method of Ooi et al. (2010). Nuclei were further purified by centrifugation through a 1.8 M sucrose cushion in 10 mM Tris pH 7.5, 10 mM MgCl2. RNA was extracted from the purified nuclei using Tripure (Roche).

To prepare short cap RNA-seq libraries, we followed the published method by Nechaev et al. (2010) with minor modification. In brief, total nuclear RNA (20 to 25 μg) was run on a polyacrylamide gel and a size range of 20 to 100 nt extracted. To enrich for capped RNA, the purified short nuclear RNA was incubated with 5’ to 3’ RNA nuclease, Terminator (Epicentre), to deplete noncapped RNA. The enriched short capped RNA was cloned and PCRed using Illumina short RNA-seq adapters according to the manual instruction. To construct long cap RNA-seq libraries, total nuclear RNA was incubated with DNase I (Ambion) to remove genomic DNA contamination, followed by Qiagen Mini-Elute columns (cut-off size > 200 nt RNAs). These cleaned-up “long” nuclear RNAs were treated with Terminator nuclease to enrich for capped RNA. Strand-specific long cap RNA libraries were made using the dUTP replacement method (Levin et al. 2010) and Illumina paired end adapters.

Short cap RNA and long cap RNA-seq libraries were sequenced using on Illumina GIIA instruments (SE36 and PE36, respectively). Two biological replicates were prepared for each library type.
Mapping, clustering of 5′ end tags, and shape score

Short cap reads were mapped to the WS220 reference sequence using BWA 0.5.9 (Li and Durbin 2009). The 36-bp reads were first mapped in their entirety, and those that did not align were then trimmed of any 3′ adaptor match and then retried. Mappings with quality < 10 were discarded at this point. Mapped reads were split by strand, with forward and reverse strand mappings analyzed independently, and mappings with the same strand and 5′ end position (cap-stacks) were combined. To define cap clusters, all cap-stacks containing two or more tags were clustered using a single-linkage approach, merging two or more stacks if they lie within 50 bp of one another. Singleton cap-stacks (not considered in the initial clustering step) were added to a cluster if they lay within the clustered region or 25 bp on either side. The final boundary coordinates of each cluster were then defined to include any singleton stacks. Finally, clusters overlapping rRNA, tRNA, miRNA, snRNA, snoRNA, or snlRNA genes (from Ensembl release 61/WS220) were excluded from the set. Using pooled cap RNA-seq data sets (two biological and one technical replica; 55,021,362 mapped reads), we obtained 73,500 clusters defining transcription initiation regions (TICs). For analyses where a single TSS position was required, we considered the distribution of cap 5′ ends within the TIC, and selected the position with the most tags (the mode). In the case of a tie (two or more positions with the same number of tags), we selected the mode closest to the median of the TIC. TIC shape score was calculated as the fraction of total tags in the cluster present at the mode position (the “mode weight”). Only TICs with at least 10 tags were used for the shape analysis. To monitor elongation, long cap RNA reads were mapped as read pairs using BWA 0.5.9 and filled in. For browser analyses of RNA-seq data, 36-bp long cap RNA sequence reads were mapped using BWA 0.5.9 in single-end mode.
TSS assignments

TICs were assigned to WormBase TSSs (via the Ensembl release 61/WS220 gene set) using a stepwise approach. In all cases, assignments were strand-specific (i.e., forward-strand TICs could only be assigned to forward-strand genes). First, any TIC overlapping a window from −199:+100 relative to a WormBase TSS was assigned (type “wormbase_tss”) to the gene associated with that TSS. Second, the remaining TICs were assigned (type “transcript_body”) to a gene if they lie anywhere within its body. Third, long-cap RNA mappings were assembled into continuous “rafts” of overlapping tags from the same strand. Rafts that overlap the TSS of exactly one gene (on the correct strand) were associated with that gene. The remaining (intergenetic) TICs were assigned (type “raft_to_wormbase_tss”) if they overlap a window from −199:+100 relative to the 5′ end of a transcript raft on the appropriate strand. All remaining TICs were left unassigned. [Chen et al., 2013] Supplementary Table S2 gives the list of TICs and associated information.

Motif plots

Motif plots were performed on a subset of TICs made nonredundant by randomly selecting one representative from any set of TICs lying within 250 bp (taking both strands into account). The following position weight matrices (PWMs) were used: Inr (Chalkley and Verrijzer 1999), TATA (V$TATA_01 from TRANSFAC [Wingender 2008]), and Sp1 (V$SP1_Q6 from TRANSFAC [Wingender 2008]). We plotted the fraction of sequences with a match for the motif at base-pair resolution (for TATA and Inr) and at 20-bp resolution for the much-more-sparingly distributed Sp1. For TATA and Sp1, we consider the first position in the PWM to be the position of the motif match. For Inr, we follow convention in taking the strongly constrained A (third position in the PWM) as the location of the motif match.
**Defining enhancers**

Intergenic regions bound by one to four transcription factors were selected from Niu et al. (2011), discarding those lying within 500 bp of the TSS of a protein-coding gene ($n = 4914$). Regions overlapping at least one TIC (excluding TICs assigned to the TSSs of protein-coding genes) were identified ($n = 3137$). The Cistrome k-means clustering tool (Liu et al. 2011a) was used to generate eight groups of H3K4me3 and H3K4me1 histone modification signals to identify and discard likely unannotated promoters (clusters 1–4 in Fig. 2-10). The remaining 2361 regions were defined as enhancers (See [Chen et al., 2013] Supplementary Table S3; and clusters 5–8 in Fig. 2-10).

Each enhancer was associated with its two neighboring protein-coding genes. If both these genes are in the same orientation (tandem genes), the enhancer was assigned to the downstream gene. Enhancers lying between transcribed genes were associated with the nearest downstream gene. Enhancers lying between convergently transcribed genes were left unassigned. Assignments are given in [Chen et al., 2013] Supplementary Table S3.

To assess the directionality of transcription across enhancers assigned to tandem genes, we counted the number of long-cap fragments mapping in each direction over each enhancer, then calculated the moderated log ratio of forward vs. reverse reads, i.e., $\log_{10}(f+1/r+1)$, where $f$ is the number of forward-strand reads and $r$ is the number of reverse-strand reads. Moderated log-ratios were used because there is a substantial number of enhancers with zero reads on one or the other strand. We then plotted forward/reverse ratios for enhancers between tandem forward-strand genes and enhancers between tandem reverse-strand genes.
**Data sets, processing, and visualization**

MNase-digested mononucleosome data for embryos (GSM514735) (Ooi et al. 2010) and adults (GSM777719) (Steiner et al. 2012) were downloaded from Gene Expression Omnibus. Transcription factor ChIP binding region data sets are from Niu et al. (2011). The following embryo ChIP-seq data were obtained from modENCODE (http://www.modencode.org/), with the following accession numbers: H3K4me3 (modENCODE_5166), H3K4me1 (modENCODE_5158), H3K36me3 (modENCODE_5165), H3K27ac (modENCODE_5159), H3K27me3 (modENCODE_5163), and HTZ-1 (modENCODE_6218). ChIP-chip RNA Pol II 4H8 data is from modENCODE_4148. Data are available from http://submit.modencode.org/submit/public/download/accession_number (e.g., for H3K4me3 data: submit.modencode.org/submit/public/download/5166).

Raw ChIP-seq data sets obtained from modENCODE were aligned using BWA with default settings (Li and Durbin 2009), normalized using BEADS (Cheung et al. 2011), then converted to log₂ ratios of BEADS scores (enrichment relative to input), standardized so the mean of the autosomal signal was 0 and the standard deviation 1, and then z-scored. For RNA Pol II 4H8 ChIP-chip data, log₂ ratios of IP/Input were obtained and standardized so the signal had mean 0 and standard deviation 1.

The heat map analyses were performed using the heat map analysis function in Cistrome (Liu et al. 2011a). The IGV Genome Browser was used for visualization (Robinson et al. 2011).

Nuclear protein-coding gene expression levels were calculated using the method of Hiller et al. (2009), calculating depth of coverage per million reads (dcpm) on the appropriate strand from the long cap RNA sequence data.
The GO term analysis was performed on tandem genes associated with enhancers at http://gostat.wehi.edu.au/ (Beissbarth and Speed 2004) and GO terms summarized using REVIGO (http://revigo.irb.hr/) (Supek et al. 2011).

To generate graphical representations of multiple sequence alignments at TSSs, extracted sequence data were uploaded to WebLogo (Crooks et al. 2004), a web-based application for sequence logos (http://weblogo.threeplusone.com/). The plots were generated using the default settings with adjustment of 36% GC composition for the C. elegans genome.
REFERENCES


Chapter 3 -

Comparative analysis of metazoan chromatin organization

Author Statement:
The work in Chapter 3 was previously published in Nature (Ho et al., 2014). I assisted in this work by participating in experimental design, embryo harvest, chromatin immunoprecipitation, library preparation, and sequencing analysis for HTZ-1. This work was performed as part of the consortium effort of modENCODE.

Reference:
ABSTRACT

Genome function is dynamically regulated in part by chromatin, which consists of the histones, non-histone proteins and RNA molecules that package DNA. Studies in *Caenorhabditis elegans* and *Drosophila melanogaster* have contributed substantially to our understanding of molecular mechanisms of genome function in humans, and have revealed conservation of chromatin components and mechanisms\(^1\)\(^,\)\(^2\)\(^,\)\(^3\). Nevertheless, the three organisms have markedly different genome sizes, chromosome architecture and gene organization. On human and fly chromosomes, for example, pericentric heterochromatin flanks single centromeres, whereas worm chromosomes have dispersed heterochromatin-like regions enriched in the distal chromosomal ‘arms’, and centromeres distributed along their lengths\(^4\)\(^,\)\(^5\). To systematically investigate chromatin organization and associated gene regulation across species, we generated and analysed a large collection of genome-wide chromatin data sets from cell lines and developmental stages in worm, fly and human. Here we present over 800 new data sets from our ENCODE and modENCODE consortia, bringing the total to over 1,400. Comparison of combinatorial patterns of histone modifications, nuclear lamina-associated domains, organization of large-scale topological domains, chromatin environment at promoters and enhancers, nucleosome positioning, and DNA replication patterns reveals many conserved features of chromatin organization among the three organisms. We also find notable differences in the composition and locations of repressive chromatin. These data sets and analyses provide a rich resource for comparative and species-specific investigations of chromatin composition, organization and function.

INTRODUCTION

We used chromatin immunoprecipitation followed by DNA sequencing (ChIP-seq) or microarray hybridization (ChIP-chip) to generate profiles of core histones, histone variants, histone modifications and chromatin-associated proteins (Fig. 3-1, Fig. 3-2 and see also [Ho et
al., 2014] Supplementary Tables 1 and 2). Additional data include DNase I hypersensitivity sites in fly and human cells, and nucleosome occupancy maps in all three organisms. Compared to our initial publications\(^1\),\(^2\),\(^3\), this represents a tripling of available fly and worm data sets and a substantial increase in human data sets (Fig. 3-1b, c). Uniform quality standards for experimental protocols, antibody validation and data processing were used throughout the projects\(^6\). Detailed analyses of related transcriptome and transcription factor data are presented in accompanying papers\(^7\),\(^8\).

**RESULTS**

We performed systematic cross-species comparisons of chromatin composition and organization, focusing on targets profiled in at least two organisms (Fig. 3-1). Sample types used most extensively in our analyses are human cell lines H1-hESC, GM12878 and K562; fly late embryos, third instar larvae and cell lines S2, Kc, BG3; and worm early embryos and stage 3 larvae. Our conclusions are summarized in Table 3-1.
Figure 3-1. **Overview of the data set.** a, Histone modifications, chromosomal proteins and other profiles mapped in at least two species (see Supplementary Fig. 1 for the full data set and Supplementary Table 1 for detailed descriptions). Different protein names for orthologues are separated by ‘/’ (see Supplementary Table 2). b, The number of all data sets generated by this and previous consortia publications\(^\dagger\) (new, 815; old, 638). Each data set corresponds to a replicate-merged normalized profile of a histone, histone variant, histone modification, non-histone chromosomal protein, nucleosome or salt-fractionated nucleosome. c, The number of unique histone marks or non-histone chromosomal proteins profiled.
Figure 3-2. Chromatin features at TSSs and gene bodies, and co-occurrence of histone modifications. a. Comparative analysis of promoter architecture at transcription start sites (TSSs). From the top, H3K4me3 (human GM12878, fly L3 and worm L3), DNase I hypersensitivity sites (DHSs), GC content, and nascent transcript (GRO-seq in human IMR90 and fly S2 cells). Human promoters, and to a lesser extent worm promoters (as defined using recently published capRNA-seq data), exhibit a bimodal enrichment for H3K4me3 and other active marks around TSSs. In contrast, fly promoters clearly exhibit a unimodal distribution of active marks, downstream of TSSs. As genes that have a neighbouring gene within 1 kb of a TSS or TES (transcription end site) were removed from this analysis, any bimodal histone modification pattern cannot be attributed to nearby genes. This difference is also not explained by chromatin accessibility determined by DHS, or by fluctuations in GC content around the TSSs, although the GC profiles are highly variable across species. b. Average gene body profiles of histone modifications on protein-coding genes in human GM12878, fly L3 and worm L3. c. Genome-wide correlations between histone modifications show intra- and inter-species similarities and differences. Top left, pairwise correlations between marks in each genome, averaged across all three species. Bottom right, pairwise correlations, averaged over cell types and developmental stages, within each species (pie chart), with inter-species variance (grey-scale background) and intra-species variance (grey-scale small rectangles) of correlation coefficients for human, fly and worm. Modifications enriched within or near actively transcribed genes are consistently correlated with each other in all three organisms. In contrast, we found a major difference in the co-occurrence pattern of two key repressive chromatin marks (black cell in bottom left): H3K27me3 (related to Polycomb (Pc)-mediated silencing) and H3K9me3 (related to heterochromatin). These two marks are strongly correlated at both developmental stages analysed in worm, whereas their correlation is low in human (r = -0.24 to -0.06) and fly (r = -0.03 to -0.1).
Not surprisingly, the three species show many common chromatin features. Most of the genome in each species is marked by at least one histone modification (See [Ho et al., 2014] Supplementary Fig. 2), and modification patterns are similar around promoters, gene bodies, enhancers and other chromosomal elements (See [Ho et al., 2014] Supplementary Figs 3 –12). Nucleosome occupancy patterns around protein-coding genes and enhancers are also largely

Table 3-1: Summary of key shared and organism-specific chromatin features in human, fly and worm

<table>
<thead>
<tr>
<th>Chromatin features</th>
<th>Human</th>
<th>Fly</th>
<th>Worm</th>
<th>Figures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Promoters</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H3K4me3 enrichment pattern around TSS</td>
<td>Bimodal peak</td>
<td>Unimodal peak*</td>
<td>Weak bimodal peak</td>
<td>ED1a,b,S12</td>
</tr>
<tr>
<td>Well positioned +1 nucleosome at expressed genes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>S13</td>
</tr>
<tr>
<td>Gene bodies</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lower H3K9me3 in specifically expressed genes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>S21-S23</td>
</tr>
<tr>
<td>Enhancers</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High H3K27ac sites are closer to expressed genes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>S5-S6</td>
</tr>
<tr>
<td>Higher nucleosome turnover at high H3K27ac sites</td>
<td>Yes</td>
<td>Yes</td>
<td>ND</td>
<td>S7</td>
</tr>
<tr>
<td>Nucleosome positioning</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10-bp periodicity profile</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>S19a</td>
</tr>
<tr>
<td>Positioning signal in genome</td>
<td>Weak</td>
<td>Weak</td>
<td>Less weak</td>
<td>S19b</td>
</tr>
<tr>
<td>LADs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histone modification in short LADs</td>
<td>H3K27me3</td>
<td>H3K27me3</td>
<td>H3K27me3</td>
<td>S17</td>
</tr>
<tr>
<td>Histone modification in long LADs</td>
<td>H3K9me3 internal, H3K27me3 borders</td>
<td>ND</td>
<td>H3K9me3 + H3K27me3</td>
<td>S15</td>
</tr>
<tr>
<td>Associated with late replication in S-phase</td>
<td>Yes</td>
<td>Yes</td>
<td>ND</td>
<td>S16</td>
</tr>
<tr>
<td>Genome-wide correlation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correlation between H3K27me3 and H3K9me3</td>
<td>Low</td>
<td>Low</td>
<td>High (in arms)</td>
<td>ED1c,ED3a</td>
</tr>
<tr>
<td>Chromatin state maps</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Similar marks and genomic features at each state</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>2,529-32</td>
</tr>
<tr>
<td>Silent domains: constitutive heterochromatin</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Composition</td>
<td>H3K9me3</td>
<td>H3K9me3</td>
<td>H3K9me3 + H3K27me3</td>
<td>2,ED3b</td>
</tr>
<tr>
<td>Predominant location</td>
<td>Pericentric + chrY</td>
<td>Pericentric + chrY</td>
<td>Arms</td>
<td>3a,ED3b</td>
</tr>
<tr>
<td>Depletion of H3K9me3 at TSS of expressed genes</td>
<td>Yes</td>
<td>Yes</td>
<td>Weak</td>
<td>ED2c</td>
</tr>
<tr>
<td>Silent domains: Polycomb-associated</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Composition</td>
<td>H3K27me3</td>
<td>H3K27me3</td>
<td>H3K27me3</td>
<td>2</td>
</tr>
<tr>
<td>Predominant location</td>
<td>Arms</td>
<td>Arms + chr4</td>
<td>Arms + centers</td>
<td>3a,ED3b</td>
</tr>
<tr>
<td>Topological domains</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Active promoters enriched at boundaries</td>
<td>Yes</td>
<td>Yes</td>
<td>ND</td>
<td>S38</td>
</tr>
<tr>
<td>Similar chromatin states are enriched in each domain</td>
<td>Yes</td>
<td>Yes</td>
<td>ND</td>
<td>S39</td>
</tr>
</tbody>
</table>
similar across species, although we observed subtle differences in H3K4me3 enrichment patterns around transcription start sites (TSSs) (Fig. 3-2a and See [Ho et al., 2014] Supplementary Figs 12–14). The configuration and composition of large-scale features such as lamina-associated domains (LADs) are similar (See [Ho et al., 2014] Supplementary Figs 15–17). LADs in human and fly are associated with late replication and H3K27me3 enrichment, suggesting a repressive chromatin environment (See [Ho et al., 2014] Supplementary Fig. 18). Finally, DNA structural features associated with nucleosome positioning are strongly conserved (See [Ho et al., 2014] Supplementary Figs 19 and 20).

Although patterns of histone modifications across active and silent genes are largely similar in all three species, there are some notable differences (Fig. 3-2b). For example, H3K23ac is enriched at promoters of expressed genes in worm, but is enriched across gene bodies of both expressed and silent genes in fly. H4K20me1 is enriched on both expressed and silent genes in human but only on expressed genes in fly and worm (Fig. 3-2b). Enrichment of H3K36me3 in genes expressed with stage or tissue specificity is lower than in genes expressed broadly, possibly because profiling was carried out on mixed tissues (See [Ho et al., 2014] Supplementary Figs 21–23; see [Ho et al., 2014] Supplementary Methods). Although the co-occurrence of pairs of histone modifications is largely similar across the three species, there are clearly some species-specific patterns (Fig. 3-2c and See [Ho et al., 2014] Supplementary Figs 24 and 25).

Previous studies showed that in human9,10 and fly1,11 prevalent combinations of marks or ‘chromatin states’ correlate with functional features such as promoters, enhancers, transcribed regions, Polycomb-associated domains, and heterochromatin. ‘Chromatin state maps’ provide a concise and systematic annotation of the genome. To compare chromatin states across the three organisms, we developed and applied a novel hierarchical non-
parametric machine-learning method called hiHMM (See [Ho et al., 2014] Supplementary Methods) to generate chromatin state maps from eight histone marks mapped in common, and compared the results with published methods (Fig. 3-3 and See [Ho et al., 2014] Supplementary Figs 26–28). We find that combinatorial patterns of histone modifications are largely conserved. Based on correlations with functional elements (See [Ho et al., 2014] Supplementary Figs 29–32), we categorized the 16 states into six groups: promoter (state 1), enhancer (states 2 and 3), gene body (states 4–9), Polycomb-repressed (states 10 and 11), heterochromatin (states 12 and 13), and weak or low signal (states 14–16).

Figure 3-3. Shared and organism-specific chromatin states. Sixteen chromatin states derived by joint segmentation using hiHMM (See [Ho et al., 2014] Supplementary Methods) based on enrichment patterns of eight histone marks. The genomic coverage of each state in each cell type or developmental stage is also shown (See [Ho et al., 2014] Supplementary Figs 26–32 for detailed analysis of the states). States are named for putative functional characteristics.

Heterochromatin is a classically defined and distinct chromosomal domain with important roles in genome organization, genome stability, chromosome inheritance and gene regulation. It is typically enriched for H3K9me3 (ref. 12), which we used as a proxy for identifying heterochromatic domains (Fig. 3-4a and See [Ho et al., 2014] Supplementary Figs 33 and 34). As expected, the majority of the H3K9me3-enriched domains in human and fly are concentrated in the pericentromeric regions (as well as other specific domains, such as the Y chromosome
and fly 4th chromosome), whereas in worm they are distributed throughout the distal chromosomal ‘arms’\(^{11,13,14}\) (Fig. 3-4a). In all three organisms, we find that more of the genome is associated with H3K9me3 in differentiated cells and tissues compared to embryonic cells and tissues (Fig. 3-5a). We also observe large cell-type-specific blocks of H3K9me3 in human and fly\(^{11,14,15}\) (See [Ho et al., 2014] Supplementary Fig. 35). These results suggest a molecular basis for the classical concept of ‘facultative heterochromatin’ formation to silence blocks of genes as cells specialize.
Figure 3-4. Genome-wide organization of heterochromatin. a, Enrichment profiles of H3K9me1, H3K9me2, H3K9me3, and H3K27me3, and identification of heterochromatin domains based on H3K9me3 (illustrated for human H1-hESC, fly L3 and worm L3). For fly chr 2, 2L, 2LHet, 2RHet and 2R are concatenated (dashed lines). C, centromere; Het, heterochromatin. b, Genome-wide correlation among H3K9me1, H3K9me2, H3K9me3 and H3K27me3 and H3K36me3 in human K562 cells, fly L3 and worm L3; no H3K9me2 profile is available for human. c, Comparison of Hi-C-based and chromatin-based topological domains in fly LE. Heat maps of similarity matrices for histone modification and Hi-C interaction frequencies are juxtaposed (See [Ho et al., 2014] Supplementary Fig. 40).

Two distinct types of transcriptionally repressed chromatin have been described. As discussed above, classical ‘heterochromatin’ is generally concentrated in specific chromosomal
regions and enriched for H3K9me3 and also H3K9me2 (ref. 12). In contrast, Polycomb-associated silenced domains, involved in cell-type-specific silencing of developmentally regulated genes11-14, are scattered across the genome and enriched for H3K27me3. We found that the organization and composition of these two types of transcriptionally silent domains differ across species. First, human, fly and worm display significant differences in H3K9 methylation patterns. H3K9me2 shows a stronger correlation with H3K9me3 in fly than in worm (r = 0.89 versus r = 0.40, respectively), whereas H3K9me2 is well correlated with H3K9me1 in worm but not in fly (r = 0.44 versus r = −0.32, respectively) (Fig. 3-4b). These findings suggest potential differences in heterochromatin in the three organisms (see below). Second, the chromatin state maps reveal two distinct types of Polycomb-associated repressed regions: strong H3K27me3 accompanied by marks for active genes or enhancers (Fig. 3-3, state 10; perhaps due to mixed tissues in whole embryos or larvae for fly and worm), and strong H3K27me3 without active marks (state 11) (See also [Ho et al., 2014] Supplementary Fig. 31). Third, we observe a worm-specific association of H3K9me3 and H3K27me3. These two marks are enriched together in states 12 and 13 in worm but not in human and fly. This unexpectedly strong association between H3K9me3 and H3K27me3 in worm (observed with several validated antibodies; Fig. 3-5b) suggests a species-specific difference in the organization of silent chromatin.
Figure 3-5. Histone modifications in heterochromatin. a. Genomic coverage of H3K9me3 in multiple cell types and developmental stages. Embryonic cell lines or stages are marked with an asterisk and a black bar. b. Evidence that overlapping H3K9me3 and H3K27me3 ChIP signals in worm are not due to antibody cross-reactivity. ChIP-chip experiments were performed from early embryo (EE) extracts with three different H3K9me3 antibodies (from Abcam, Upstate and H. Kimura) and three different H3K27me3 antibodies (from Active Motif, Upstate and H. Kimura). The H3K9me3 antibodies show similar enrichment profiles (top panel) and high genome-wide correlation coefficients (bottom left). The same is true for...
H3K27me3 antibodies. There is significant overlap between the H3K9me3 and H3K27me3 ChIP signal, especially on chromosome arms, resulting in relatively high genome-wide correlation coefficients (Extended Data Fig. 1c). The Abcam and Upstate H3K9me3 antibodies showed low level cross-reactivity with H3K27me3 on dot blots24, and the Abcam H3K9me3 ChIP signal overlapped with H3K27me3 on chromosome centres. The Kimura monoclonal antibodies against H3K9me3 and H3K27me3 showed the least overlap and smallest genome-wide correlation. In enzyme-linked immunosorbent assay (ELISA) using histone H3 peptides containing different modifications, each Kimura H3K9me3 or H3K27me3 antibody recognized the modified tail against which it was raised and did not cross-react with the other modified tail25, 26, providing support for their specificity. Specificity of the Kimura antibodies was further analysed by immunostaining germlines from wild type, met-2 set-25 mutants (which lack H3K9 histone methyltransferase (HMT) activity16), and mes-2 mutants (which lack H3K27 HMT activity27) in the bottom right panel. Staining with anti-HK9me3 was robust in wild type and in mes-2, but undetectable in met-2 set-25. Staining with anti-HK27me3 was robust in wild type and in met-2 set-25, but undetectable in mes-2. Finally, we note that the laboratories that analysed H3K9me3 and H3K27me3 in other systems used Abcam H3K9me3 (for human and fly) and Upstate H3K27me3 (for human), and in these cases observed non-overlapping distributions. Another paper also reported non-overlapping distributions of H3K9me3 and H3K27me3 in human fibroblast cells using the Kimura antibodies26. The overlapping distributions that we observe in worms using any of those antibodies suggest that H3K9me3 and H3K27me3 occupy overlapping regions in worms. Those overlapping regions may exist in individual cells or in different cell sub-populations in embryo and L3 preparations. c. Average gene body profiles of H3K9me3 and H3K27me3 on expressed and silent genes in euchromatin and heterochromatin in human K562 cells, fly L3 and worm L3.

We also compared the patterns of histone modifications on expressed and silent genes in euchromatin and heterochromatin (Fig. 3-5c and See [Ho et al., 2014] Supplementary Fig. 36). We previously reported prominent depletion of H3K9me3 at TSSs and high levels of H3K9me3 in the gene bodies of expressed genes located in fly heterochromatin14, and now find a similar pattern in human (Fig. 3-5c and See [Ho et al., 2014] Supplementary Fig. 36). In these two species, H3K9me3 is highly enriched in the body of both expressed and silent genes in heterochromatic regions. In contrast, expressed genes in worm heterochromatin have lower H3K9me3 enrichment across gene bodies compared to silent genes (Fig. 3-5c and See [Ho et al., 2014] Supplementary Figs 36, and 37). There are also conspicuous differences in the patterns of H3K27me3 in the three organisms. In human and fly, H3K27me3 is highly associated with silent genes in euchromatic regions, but not with silent genes in heterochromatic regions. In contrast, consistent with the worm-specific association between H3K27me3 and H3K9me3, we observe high levels of H3K27me3 on silent genes in worm
heterochromatin, whereas silent euchromatic genes show modest enrichment of H3K27me3 (Fig. 3-5c and See [Ho et al., 2014] Supplementary Fig. 36).

Our results suggest three distinct types of repressed chromatin (Fig. 3-6). The first contains H3K27me3 with little or no H3K9me3 (human and fly states 10 and 11, and worm state 11), corresponding to developmentally regulated Polycomb-silenced domains in human and fly, and probably in worm as well. The second is enriched for H3K9me3 and lacks H3K27me3 (human and fly states 12 and 13), corresponding to constitutive, predominantly pericentric heterochromatin in human and fly, which is essentially absent from the worm genome. The third contains both H3K9me3 and H3K27me3 and occurs predominantly in worm (worm states 10, 12 and 13). Co-occurrence of these marks is consistent with the observation that H3K9me3 and H3K27me3 are both required for silencing of heterochromatic transgenes in worms\textsuperscript{16}. H3K9me3 and H3K27me3 may reside on the same or adjacent nucleosomes in individual cells\textsuperscript{17,18}; alternatively the two marks may occur in different cell types in the embryos and larvae analysed here. Further studies are needed to resolve this and determine the functional consequences of the overlapping distributions of H3K9me3 and H3K27me3 observed in worm.
Figure 3-6. Organization of silent domains. a, The correlation of H3K27me3 and H3K9me3 enrichment for human K562 (left panel), fly L3 (second panel), and worm EE chromosome arms (third panel) and centres (right panel) with a 10-kb bin (top) and a 1-kb bin (bottom). The density was calculated as a frequency of bins that fall in the area in the scatter plot (darker grey at a higher frequency). \( r \) indicates Pearson correlation coefficients between binned H3K27me3 fold enrichment (log\(_2\)) and H3K9me3 fold enrichment (log\(_2\)). Worm chromosome arms have a distinctly high correlation between H3K27me3 and H3K9me3. The lower correlation in worm chromosome centres is due to the overall absence of H3K9me3 in these regions. b, Schematic diagrams of the distributions of silent domains along the chromosomes in human (H1-hESC), fly (S2) and worm (EE). In human and fly, the majority of the H3K9me3-enriched
domains are located in the pericentric regions (as well as telomeres), while the H3K27me3-enriched domains are distributed along the chromosome arms. H3K27me3-enriched domains are negatively correlated with H3K36me3-enriched domains, although in human, there is some overlap of H3K27me3 and H3K36me3 in bivalent domains. CENP-A resides at the centromere. In contrast, in worm the majority of H3K9me3-enriched domains are located in the arms, whereas H3K27me3-enriched domains are distributed throughout the arms and centres of the chromosomes and are anti-correlated with H3K36me3-enriched domains. In arms and centres, domains that are permissive for CENP-A incorporation generally reside within H3K27me3-enriched domains.

Genome-wide chromatin conformation capture (Hi-C) assays have revealed prominent topological domains in human \(^\text{19}\) and fly \(^\text{20, 21}\). Although their boundaries are enriched for insulator elements and active genes \(^\text{19, 20}\) (See [Ho et al., 2014] Supplementary Fig. 38), the interiors generally contain a relatively uniform chromatin state: active, Polycomb-repressed, heterochromatin, or low signal \(^\text{22}\) (See [Ho et al., 2014] Supplementary Fig. 39). We found that chromatin state similarity between neighbouring regions correlates with chromatin interaction domains determined by Hi-C (Fig. 3-4c, See [Ho et al., 2014] Supplementary Fig. 40 and Supplementary Methods). This suggests that topological domains can be largely predicted by chromatin marks when Hi-C data are not available (See [Ho et al., 2014] Supplementary Figs 41 and 42).

**DISCUSSION**

*C. elegans* and *D. melanogaster* have been used extensively for understanding human gene function, development and disease. Our analyses of chromatin architecture and the large public resource we have generated provide a blueprint for interpreting experimental results in these model systems, extending their relevance to human biology. They also provide a foundation for researchers to investigate how diverse genome functions are regulated in the context of chromatin structure.
MATERIALS AND METHODS

Preprocessing of ChIP-seq data

Raw sequences were aligned to their respective genomes (hg19 for human; dm3 for fly; and ce10 for worm) using bowtie28 or BWA29 following standard preprocessing and quality assessment procedures of ENCODE and modENCODE6. Validation results of the antibodies used in all ChIP experiments are available at the Antibody Validation Database (http://compbio.med.harvard.edu/antibodies/). Most of the ChIP-seq datasets were generated by 36 bp (in human and worm), 42 bp (in worm), or 50 bp (in fly and worm) single-end sequencing using the Illumina HiSeq platform, with an average of ~20 million reads per sample replicate (at least two replicates for each sample). Quality of the ChIP-seq data was examined as follows. For all three organisms, cross-correlation analysis was performed, as described in the published modENCODE and ENCODE guidelines6. This analysis examines ChIP efficiency and signal-to-noise ratio, as well as verifying the size distribution of ChIP fragments. The results of this cross-correlation analysis for the more than 3000 modENCODE and ENCODE ChIP-seq data sets are described elsewhere30. In addition, to ensure consistency between replicates in the fly data, we further required at least 80% overlap of the top 40% of peaks in the two replicates (overlap is determined by number of bp for broad peaks, or by number of peaks for sharp peaks; peaks as determined by SPP31 etc). Library complexity was checked for human. For worm, genome-wide correlation of fold enrichment values was computed for replicates and a minimum threshold of 0.4 was required. In all organisms, those replicate sets that do not meet these criteria were examined by manual inspection of browser profiles to ascertain the reasons for low quality and, whenever possible, experiments were repeated until sufficient quality and consistency were obtained. To enable the cross-species comparisons described in this paper, we have reprocessed all data using MACS32. Due to the slight differences in the peak-calling and input normalization steps, there may be slight discrepancies between the fly profiles analyzed here.
and profiles available at the project data portal (http://data.modencode.org; http://encodeproject.org) or modMine (http://modmine.org, which redirects to http://intermine.modencode.org). For every pair of aligned ChIP and matching input-DNA data, we used MACS33 version 2 to generate fold enrichment signal tracks for every position in a genome:

macs2 callpeak -t ChIP.bam -c Input.bam -B --nomodel --shiftsize 73 --SPMR -g hs -n ChIP
macs2 bdgcmp -t ChIP_treat_pileup.bdg -c ChIP_control_lambda.bdg -o ChIP_FE.bedgraph -m FE

Depending on analysis, we applied either log transformation or z-score transformation.

**Preprocessing of ChIP-chip data**

For the fly data, genomic DNA Tiling Arrays v2.0 (Affymetrix) were used to hybridize ChIP and input DNA. We obtained the log-intensity ratio values (M-values) for all perfect match (PM) probes: \( M = \log_2(\text{ChIP intensity}) - \log_2(\text{input intensity}) \), and performed a whole genome baseline shift so that the mean of \( M \) in each microarray is equal to 0. The smoothed log intensity ratios were calculated using LOWESS with a smoothing span corresponding to 500 bp, combining normalized data from two replicate experiments. For the worm data, a custom Nimblegen two-channel whole genome microarray platform was used to hybridize both ChIP and input DNA. MA2C34 was used to preprocess the data to obtain a normalized and median centered log2 ratio for each probe. All data are publicly accessible through the modENCODE data portal or modMine.

**Preprocessing of GRO-seq data**

Raw sequences of the fly S2 and human IMR90 datasets were downloaded from NCBI Gene Expression Omnibus (GEO) using accession numbers GSE2588735 and GSE1351836 respectively. The sequences were then aligned to the respective genome assembly (dm3 for fly...
and hg19 for human) using bowtie28. After checking for consistency based on correlation analysis and browser inspection, we merged the reads of the biological replicates before proceeding with downstream analyses. Treating the reads mapping to the positive and negative strands separately, we calculated minimally-smoothed signals (by a Gaussian kernel smoother with bandwidth of 10 bp in fly and 50 bp in human) along the genome in 10 bp (fly) or 50 bp (human) non-overlapping bins.

**Preprocessing of DNase-seq data**

Aligned DNase-seq data were downloaded from the modENCODE data portal and the ENCODE UCSC download page (http://encodeproject.org/ENCODE/). Additional Drosophila embryo DNase-seq data were downloaded from37. After confirming consistency, reads from biological replicates were merged. We calculated minimally-smoothed signals (by a Gaussian kernel smoother with bandwidth of 10 bp in fly and 50 bp in human) along the genome in 10 bp (fly) or 50 bp (human) non-overlapping bins.

**Preprocessing of MNase-seq data**

The MNase-seq data were analyzed as described previously38. In brief, tags were mapped to the corresponding reference genome assemblies. The positions at which the number of mapped tags had a Z-score > 7 were considered anomalous due to potential amplification bias. The tags mapped to such positions were discarded. To compute profiles of nucleosomal frequency around TSS, the centers of the fragments were used in the case of paired-end data. In the case of single-end data, tag positions were shifted by the half of the estimated fragment size (estimated using cross-correlation analysis39 toward the fragment 3'-ends and tags mapping to positive and negative DNA strands were combined). Loess smoothing in the 11- bp window, which does not affect positions of the major minima and maxima on the plots, was applied to reduce the high-frequency noise in the profiles.
**GC-content and PhastCons conservation score**

We downloaded the 5bp GC% data from the UCSC genome browser annotation download page (http://hgdownload.cse.ucsc.edu/downloads.html) for human (hg19), fly (dm3), and worm (ce10). Centering at every 5 bp bin, we calculated the running median of the GC% of the surrounding 100 bp (i.e., 105 bp in total). PhastCons conservation score was obtained from the UCSC genome browser annotation download page. Specifically, we used the following score for each species:

<table>
<thead>
<tr>
<th>Target Species</th>
<th>PhastCons scores generated by multiple alignments with:</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. elegans (ce10)</td>
<td>6 Caenorhabditis nematode genomes</td>
</tr>
<tr>
<td></td>
<td><a href="http://hgdownload.cse.ucsc.edu/goldenPath/ce10/phastCons7way/">URL</a></td>
</tr>
<tr>
<td>D. melanogaster (dm3)</td>
<td>15 Drosophila and related fly genomes</td>
</tr>
<tr>
<td></td>
<td><a href="http://hgdownload.cse.ucsc.edu/goldenPath/dm3/phastCons15way/">URL</a></td>
</tr>
<tr>
<td>H. Sapiens (hg19)</td>
<td>45 vertebrate genomes</td>
</tr>
<tr>
<td></td>
<td><a href="http://hgdownload.cse.ucsc.edu/goldenPath/hg19/phastCons46way/vertebrate/">URL</a></td>
</tr>
</tbody>
</table>

Both GC and phastCons scores were then binned into 10 bp (fly and worm) or 50 bp (human) non-overlapping bins.

**Genomic sequence mappability tracks**

We generated empirical genomic sequence mappability tracks using input-DNA sequencing data. After merging input reads up to 100M, reads were extended to 149 bp which corresponds to the shift of 74 bp in signal tracks. The union set of empirically mapped regions was obtained. They are available at the ENCODE-X Browser ([http://encodex.med.harvard.edu/data_sets/chromatin/](http://encodex.med.harvard.edu/data_sets/chromatin/)).
Coordinates of unassembled genomic sequences

We downloaded the “Gap” table from UCSC genome browser download page (http://hgdownload.cse.ucsc.edu/downloads.html). The human genome contains 234 Mb of unassembled regions whereas fly contains 6.3 Mb of unassembled genome. There are no known unassembled (i.e., gap) regions in worm.

Gene annotation

We used human GENCODE version 10 (http://www.gencodegenes.org/releases/10.html) for human gene annotation40. For worm and fly, we used custom RNA-seq-based gene and transcript annotations generated by the modENCODE consortium7.

Worm TSS definition based on capRNA-seq (capTSS)

We obtained worm TSS definition based on capRNA-seq from Chen et al.23. Briefly, short 5'-capped RNA from total nuclear RNA of mixed stage embryos were sequenced (i.e., capRNAseq) by Illumina GAIIA (SE36) with two biological replicates. Reads from capRNA-seq were mapped to WS220 reference genome using BWA29. Transcription initiation regions (TICs) were identified by clustering of capRNA-seq reads. In this analysis we used TICs that overlap with wormbase TSSs within -199:100bp. We refer these capRNA-seq defined TSSs as capTSS in this study.

Gene expression data

Gene expression level estimates of various cell-lines, embryos or tissues were obtained from the modENCODE and ENCODE projects7. The expression of each gene is quantified in terms of RPKM (reads per kilobase per million reads). The distribution of gene expression in each cell line was assessed and a cut-off of RPKM=1 was determined to be generally a good threshold to separate active vs. inactive genes. This definition of active and inactive genes was used in the construction of meta-gene profiles.
Genomic coverage of histone modifications

To identify the significantly enriched regions, we used SPP R package (ver.1.10)31. The 5’end coordinate of every sequence read was shifted by half of the estimated average size of the fragments, as estimated by the cross-correlation profile analysis. The significance of enrichment was computed using a Poisson model with a 1 kb window. A position was considered significantly enriched if the number of IP read counts was significantly higher (Zscore > 3 for fly and worm, 2.5 for human) than the number of input read counts, after adjusting for the library sizes of IP and input, using SPP function get.broad.enrichment.cluster.

Genome coverage in each genome is then calculated as the total number of base pair covered by the enriched regions or one or more histone marks. It should be noted that genomic coverage reported in Supplementary Fig. 2 refers to percentage of histone mark coverage with respective to mappable region. A large portion (~20%) of human genome is not mappable based on our empirical criteria. These unmappable regions largely consist of unassembled regions, due to difficulties such as mapping of repeats. Furthermore, some unmappable regions may be a result of the relatively smaller sequencing depth compared to fly and worm samples. Therefore it is expected that empirically determined mappability is smaller in human compared to fly and worm.

Identification and analysis of enhancers

We used a supervised machine learning approach to identify putative enhancers among DNaseI hypersensitive sites (DHSs) and p300 or CBP-1 binding sites, hereafter referred collectively as “regulatory sites”. The basic idea is to train a supervised classifier to identify H3K4me1/3 enrichment patterns that distinguish TSS distal regulatory sites (i.e., candidate enhancers) from proximal regulatory sites (i.e., candidate promoters). TSS-distal sites that carry these patterns are classified as putative enhancers.

Human DHS and p300 binding site coordinates were downloaded from the ENCODE
UCSC download page (http://genome.ucsc.edu/ENCODE/downloads.html). When available, only peaks identified in both replicates were retained. DHSs and p300 peaks that were wider than 1 kb were removed. DHS positions in fly cell lines were defined as the 'high-magnitude' positions in DNase I hypersensitivity identified by Kharchenko et al11. We applied the same method to identify similar positions in DNase-seq data in fly embryonic stage 14 (ES14)37, which roughly corresponds to LE stage. Worm MXEMB CBP-1 peaks were determined by SPP with default parameters. CBP-1 peaks that were identified within broad enrichment regions wider than 1 kb were removed. For fly and human cell lines, DHS and p300 data from matching cell types were used. For fly late embryos (14-16 h), the DHS data from embryonic stage 14 (10:20–11:20 h) were used. For worm EE and L3, CBP-1 data from mixed-embryos were used.

To define the TSS-proximal and TSS-distal sites, inclusive TSS lists were obtained by merging ensemble v66 TSSs with GENCODE version 10 for human, and modENCODE transcript annotations for fly and worm, including all alternate sites. Different machine learning algorithms were trained to classify genomic positions as a TSS-distal regulatory site, TSS-proximal regulatory site or neither, based on a pool of TSS-distal (>1 kb) and TSS–proximal (<250 bp) regulatory sites and a random set of positions from other places in the genome. The random set included twice as many positions as the TSS-distal site set for each cell type. Five features from each of the two marks, H3K4me1 and H3K4me3, were used for the classification: maximum fold-enrichment within +/-500 bp, and four average fold enrichment values in 250 bp bins within +/-500 bp. The pool of positions was split into two equal test and training sets. The performance of different classifier algorithms was compared using the area under Receiver Operator Characteristics (ROC) curves. For human and fly samples, the best performance was obtained using the Model-based boosting (mboost) algorithm41, whereas for the worm data sets, the Support Vector Machine (SVM) algorithm showed superior performance. TSS-distal sites that in turn get classified as “TSS-distal” make up our enhancer set. In worm, the learned
model was used to classify sites within 500-1000 bp from the closest TSS, and those classified as TSS-distal were included in the final enhancer set to increase the number of identified sites. Our sets of putative enhancers (hereafter referred to as ‘enhancers’) include roughly 2000 sites in fly cell lines and fly embryos, 400 sites in worm embryos, and 50,000 sites in human cell lines.

It should be noted that while enhancers identified at DHSs (in human and fly) or CBP-1 binding sites (in worm) may represent different classes of enhancers, for the purpose of studying the major characteristics of enhancers, both definitions are a reasonable proxy for identifying enhancer-like regions. We repeated all human enhancer analysis with p300 sites (worm CBP-1 is an ortholog of p300 in human). Half of the p300-based enhancers overlap with DHS-based enhancers (See [Ho et al., 2014] Supplementary Table 3). In addition, all the observed patterns were consistent with the enhancers identified using DHSs (See [Ho et al., 2014] Supplementary Fig. 3), including the association of enhancer H3K27ac levels with gene expression (See [Ho et al., 2014] Supplementary Figs. 4-6), patterns of nucleosome turnover (See [Ho et al., 2014] Supplementary Figs. 7-9) and histone modifications and chromosomal proteins (See [Ho et al., 2014] Supplementary Fig. 10). The results based on DHS-based enhancers were validated by analyzing p300-based enhancers (See [Ho et al., 2014] Supplementary Fig. 11).

For [Ho et al., 2014] Supplementary Fig. 3-6, the enrichment level of a histone mark around a site (DHS or CBP-1 enhancer) is calculated based on the maximum ChIP fold enrichment within +/- 500 bp region of the site. These values are also used to stratify enhancers based on the H3K27ac enrichment level. For [Ho et al., 2014] Supplementary Fig. 10, we extracted histone modification signal +/- 2 kb around each enhancer site in 50 bp bins. ChIP fold enrichment is then averaged across all the enhancer sites in that category (high or low H3K27ac). These average signals across the entire sample (e.g., human GM12878) are then
subjected to Z-score transformation (mean = 0, standard deviation = 1). All z-scores above 4 or below -4 are set to 4 and -4 respectively.

In terms of analysis of average expression of genes that are proximal to a set of enhancers (See [Ho et al., 2014] Supplementary Fig. 5), we identify genes that are located within 5, 10, 25, 40, 50, 75, 125, 150, 175 and 200 kb away from the center of an enhancer in both directions, and take an average of the expression levels of all of the genes within this region.

**Analysis of HiC-defined topological domains**

We used the genomic coordinates of the topological domains defined in the original publication on fly late embryos20, and human embryonic stem cell lines19. The human coordinates were originally in hg18. We used UCSC's liftOver tool (http://genome.ucsc.edu/cgi-bin/hgLiftOver) to convert the coordinates to hg19.

**Analysis of chromatin states near topological domain boundaries**

For each chromatin state, the number of domain boundaries where the given state is at a given distance to the boundary is counted. The random expected value of counts is calculated as the number of all domain boundaries times the normalized genomic coverage of the chromatin state. The ratio of observed to expected counts is presented as a function of the distance to domain boundaries.

**Analysis of chromatin states within topological domains**

In [Ho et al., 2014] supplementary Fig. 39, the interior of topological domains is defined by removing 4 kb and 40 kb from the edges of each topological domain for fly and human Hi-C defined domains respectively. To access the chromatin state composition of each topological domain, the coverage of the domain interior by each chromatin state is calculated in bps and normalized to the domain size, yielding a measure between 0 and 1. Then the matrix of values
corresponding to chromatin states in one dimension and topological domains in the second dimension is used to cluster the chromatin states hierarchically. Pearson correlation coefficients (1-r) between domain coverage values of different chromatin states are taken as the distance metric for the clustering. The clustering tree is cut as to obtain a small number of meaningful groups of highly juxtaposed chromatin states. The coverage of each chromatin state group is calculated by summing the coverage of states in the group. Each topological domain is assigned to the chromatin state group with maximum coverage in the domain interior.

**Definition of lamina associated domains (LADs)**

Genomic coordinates of LADs were directly obtained from their original publications, for worm, fly and human. We converted the genomic coordinates of LADs to ce10 (for worm), dm3 (for fly) and hg19 (for human) using UCSC's liftOver tool with default parameters (http://genome.ucsc.edu/cgi-bin/hgLiftOver). For [Ho et al., 2014] Supplementary Fig. 16b, the raw fly DamID ChIP values were used after converting the probe coordinates to dm3.

**LAD chromatin context analysis**

In [Ho et al., 2014] Supplementary Fig. 15, scaled LAD plot, long and short LADs were defined by top 20% and bottom 20% of LAD sizes, respectively. For a fair comparison between human and worm LADs in the figure, a subset of human LADs (chromosomes 1 to 4, N = 391) was used, while for worm LADs from all chromosomes (N= 360) were used. 10 kb (human) or 2.5 kb (worm) upstream and downstream of LAD start sites and LAD ending sites are not scaled. Inside of LADs is scaled to 60 kb (human) or 15 kb (worm). Overlapping regions with adjacent LADs are removed.

To correlate H3K9me3, H3K27me3 and EZH2/EZ with LADs, the average profiles were obtained at the boundaries of LADs with a window size of 120 kb for human, 40 kb for fly and 10 kb for worm. The results at the right side of domain boundaries were flipped for [Ho et al., 2014] Supplementary Fig. 16a.
LAD Replication Timing analysis

The repli-seq BAM alignment files for the IMR90 and BJ human cell lines were downloaded from the UCSC ENCODE website. Early and late RPKM signal was determined for nonoverlapping 50 kb bins across the human genome, discarding bins with low mappability (i.e., bins containing less than 50% uniquely mappable positions). To better match the fly repli-seq data, the RPKM signal from the two early fractions (G1b and S1) and two late fractions (S4 and G2) were each averaged together. The fly Kc cell line replication-seq data was obtained from GEO. Reads were pooled together from two biological replicates (S1: GSM1015342 and GSM1015346; S4: GSM1015345 and GSM1015349), and aligned to the Drosophila melanogaster dm3 genome using Bowtie28. Early and late RPKM values were then calculated for each non-overlapping 10 kb bin, discarding low mappability bins as described above. To make RPKM values comparable between both species, the fly RPKM values were normalized to the human genome size. All replication timing bins within a LAD domain were included in the analysis. An equivalent number of random bins were then selected, preserving the observed LAD domain chromosomal distribution.

<table>
<thead>
<tr>
<th>Cell Type</th>
<th>Phase</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMR90</td>
<td>G1b</td>
<td><a href="http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwRepliSeq/wgEncodeUwRepliSeqImr90G1bAlnRep1.bam">Link</a></td>
</tr>
<tr>
<td>IMR90</td>
<td>S1</td>
<td><a href="http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwRepliSeq/wgEncodeUwRepliSeqImr90S1AlnRep1.bam">Link</a></td>
</tr>
<tr>
<td>IMR90</td>
<td>S4</td>
<td><a href="http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwRepliSeq/wgEncodeUwRepliSeqImr90S4AlnRep1.bam">Link</a></td>
</tr>
<tr>
<td>IMR90</td>
<td>G2</td>
<td><a href="http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwRepliSeq/wgEncodeUwRepliSeqImr90G2AlnRep1.bam">Link</a></td>
</tr>
<tr>
<td>BJ</td>
<td>G1b</td>
<td><a href="http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwRepliSeq/wgEncodeUwRepliSeqBjG1bAlnRep2.bam">Link</a></td>
</tr>
<tr>
<td>BJ</td>
<td>S1</td>
<td><a href="http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwRepliSeq/wgEncodeUwRepliSeqBjS1AlnRep2.bam">Link</a></td>
</tr>
<tr>
<td>BJ</td>
<td>S4</td>
<td><a href="http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwRepliSeq/wgEncodeUwRepliSeqBjS4AlnRep2.bam">Link</a></td>
</tr>
<tr>
<td>BJ</td>
<td>G2</td>
<td><a href="http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwRepliSeq/wgEncodeUwRepliSeqBjG2AlnRep2.bam">Link</a></td>
</tr>
</tbody>
</table>
Analysis of DNA structure and nucleosome positioning

The ORChID2 algorithm was used to predict DNA shape and generate consensus profiles for paired-end MNase-seq fragments of size 146-148 bp as previously described \(^4^5\). Only 146-148 bp sequences were used in this analysis to minimize possible effect of over- and underdigestion in the MNase treatment. The ORChID2 algorithm provides a more general approach than often-used investigation of mono- or dinucleotide occurrences along nucleosomal DNA since it can capture even degenerate sequence signatures if they have pronounced structural features.

For individual sequence analyses, we used the consensus profile generated above and trimmed three bases from each end to eliminate edge effects of the prediction algorithm, and then scanned this consensus against each sequence of length 146-148 bp. We retained the maximum correlation value between the consensus and individual sequence, and compared this to shuffled versions of each sequence (See [Ho et al., 2014] Supplementary Figs. 19-20). To estimate the sequence effect on nucleosome positioning we calculated the area between the solid lines and normalized by the area between the dashed lines (See [Ho et al., 2014] Supplementary Fig. 20a; upper panel) and reported this result in [Ho et al., 2014] Supplementary Fig. 19b.

Construction of meta-gene profiles

We defined transcription start site (TSS) and transcription end site (TES) as the 5' most and 3' most position of a gene, respectively, based on the modENCODE/ENCODE transcription group's gene annotation \(^7\). To exclude short genes from this analysis, we only included genes with a minimum length of 1 kb (worm and fly) or 10 kb (human). To further alleviate confounding signals from nearby genes, we also excluded genes which have any neighboring genes within 1 kb upstream of its TSS or 1 kb downstream of its TES. The ChIP enrichment in the 1 kb region upstream of TSS or downstream of TES, as well as 500 bp downstream of TSS or upstream of
TES, were not scaled. The ChIP-enrichment within the remaining gene body was scaled to 2 kb. The average ChIP fold enrichment signals were then plotted as a heat map or a line plot.

**Analysis of broadly and specifically expressed genes**

For each species, we obtained RNA-seq based gene expression estimates (in RPKM) of multiple cell lines or developmental stages from the modENCODE/ENCODE transcription groups. Gene expression variability score of each gene was defined to be the ratio of standard deviation and mean of expression across multiple samples. For each species, we divide the genes into four quartiles based on this gene expression variability score. Genes within the lowest quartile of variability score with RPKM value greater than 1 is defined as "broadly expressed". Similarly, RPKM>1 genes within the highest quartile of variability score is defined as "specifically expressed". We further restricted our analysis to protein-coding genes that are between 1 and 10 kb (in worm and fly) or between 1 and 40 kb (in human) in length. For ChIP-chip analysis of BG3, S2 and Kc cells, ChIP signal enrichment for each gene was calculated by averaging the smoothed log intensity ratios from probes that fall in the gene body. For all other cell types, ChIP-seq read coordinates were adjusted by shifting 73 bp along the read and the total number of ChIP and input fragments that fall in the gene body were counted. Genes with low sequencing depth (as determined by having less than 4 input tags in the gene body) were discarded from the analysis. ChIP signal enrichment is obtained by dividing (library normalized) ChIP read counts to Input read count. The same procedure was applied to calculate enrichment near TSS of genes, by averaging signals from probes within 500 bp of TSSs for BG3 cells and using read counts within 500 bp of TSSs for ChIPseq data.

**Genome-wide correlation between histone modifications**

In Fig. 2-1c and [Ho et al., 2014] Supplementary Fig. 24, eight histone modifications commonly profiled in human (H1-hESC,GM12878 and K562), fly (LE, L3 and AH), and worm (EE and L3), were used for pairwise genome-wide correlation at 5 kb bin resolution.
Unmappable regions and regions that have fold enrichment values less than 1 for all 8 marks (low signal regions) were excluded from the analysis. To obtain a representative correlation value for each species, an average Pearson correlation coefficient for each pair of marks was computed over the different cell types and developmental stages of each species. The overall correlation (upper triangle of Fig. 2-2c) was computed by averaging the three single-species correlation coefficients. Intra-species variance was computed as the average within-species variance of correlation coefficients. Inter-species variance was computed as the variance of the within-species average correlation coefficients. For the large correlation heatmaps in Supplementary Fig. 25, 10 kb (worm and fly) or 30 kb (human) bins were used with no filtering of low-signal regions.

**Chromatin segmentation using hiHMM**

We performed joint chromatin state segmentation of multiple species using a hierarchically linked infinite hidden Markov model (hiHMM). In a traditional HMM that relies on a fixed number of hidden states, it is not straightforward to determine the optimal number of hidden states. In contrast, a non-parametric Bayesian approach of an infinite HMM (iHMM) can handle an unbounded number of hidden states in a systematic way so that the number of states can be learned from the training data rather than be pre-specified by the user\(^{46}\). For joint analysis of multi-species data, the hiHMM model employs multiple, hierarchically linked, iHMMs over the same set of hidden states across multiple species - one iHMM per species. More specifically, within a hiHMM, each iHMM has its own species-specific parameters for both transition matrix \(\pi(\cdot)\) and emission probabilities for \(c=\{\text{human, fly, worm}\}\). Emission process was modeled as a multivariate Gaussian with a diagonal covariance matrix such that where represents \(m\) dimensional vector for observed data from \(m\) chromatin marks of species \(c\) at genomic location \(t\), and represents the corresponding hidden state at \(t\). The parameters correspond to the mean signal values from state \(k\) in species \(c\), and is the species-specific
covariance matrix. To take into account the different self-transition probabilities in different species, we also incorporate an explicit parameter that controls the self-transition probability. In the resulting transition model, we have . Each row of the transition matrix across all the species follows the same prior distribution of the so-called Dirichlet process that allows the state space to be shared across species. Using this scheme, data from multiple species are weakly coupled only by a prior. Therefore hiHMM can capture the shared characteristics of multiple species data while still allowing unique features for each species. This hierarchically linked HMM has been first applied to the problem of local genetic ancestry from haplotype data46 in which the same modeling scheme for the transition process but a different emission process has been adopted to deal with the SNP haplotype data.

This hierarchical approach is substantially different from the plain HMM that treats multispecies data as different samples from a homogeneous population. For example, different species data have different gene length and genome composition, so one transition event along a chromosome of one species does not equally correspond to one transition in another species. So if a model has just one set of transition probabilities for all species, it cannot reflect such difference in self-transition or between-state transition probabilities. Our model hiHMM can naturally handle this by assuming species-specific transition matrices. Note that since the state space is shared across all the populations, it is easy to interpret the recovered chromatin states.

Since hiHMM is a non-parametric Bayesian approach, we need Markov chain Monte-Carlo (MCMC) sampling steps to train a model. Instead of Gibbs sampling, we adopted a dynamic programming scheme called Beam sampling47, which significantly improves the mixing and convergence rate. Although it still requires longer computation time than parametric methods like a finite-state HMM, this training can be done once offline and then we can approximate the decoding step of the remaining sequences by Viterbi algorithm using the trained HMM parameters.
ChIP-seq data were further normalized before being analyzed by hiHMM. ChIP-seq normalized signals were averaged in 200 bp bins in all three species. MACS2 processed ChIP-seq fold change values were log2 transformed with a pseudocount of 0.5, i.e., \( y = \log_2(x+0.5) \), followed by mean-centering and scaling to have standard deviation of 1. The transformed fold enrichment data better resemble a Gaussian distribution based on QQ-plot analysis.

To train the hiHMM, the following representative chromosomes were used:

- Worm (L3): chrII, chrIII, chrX
- Fly (LE and L3): chr2L, chr2LHet, chrX, chrXHet
- Human (H1-hESC and GM12878): chr1, chrX

It should be noted that H4K20me1 profile in worm EE is only available as ChIP-chip data. This is why worm EE was not used in the training phase. In the inference phase, we used the quantile-normalized signal values of the H4K20me1 EE ChIP-chip data.

One emission and one transition probability matrix was learned from each species. We also obtained the maximum a priori (MAP) estimate of the number of states, \( K \). We then used Viterbi decoding algorithm to generate a chromatin state segmentation of the whole genome of worm (EE and L3), fly (LE and L3) and human (H1-hESC and GM12878). To avoid any bias introduced by unmappable regions, we removed the empirically determined unmappable regions before performing Viterbi decoding. These unmappable regions are assigned a separate “unmappable state” after the decoding.

The chromatin state definition can be accessed via the ENCODE-X Browser (http://encodex.med.harvard.edu/data_sets/chromatin/).

**Chromatin segmentation using Segway**

We compared the hiHMM segmentation with a segmentation produced by Segway48, an existing segmentation method. Segway uses a dynamic Bayesian network model, which includes explicit representations of missing data and segment lengths. Segway models the
emission of signal observations at a position using multivariate Gaussians. Each label \( k \) has a corresponding Gaussian characterized by a mean vector and a diagonal covariance matrix. At locations where particular tracks have missing data, Segway excludes those tracks from its emission model. For each label, Segway also includes a parameter that models the probability of a change in label. If there is a change in label, a separate matrix of transition parameters models the probability of switching to every other label. Given these emission and transition parameters, Segway can calculate the likelihood of observed signal data. To facilitate modeling data from multiple experiments with a single set of parameters, we performed a separate quantile normalization on each signal track prior to Segway analysis. We took the initial unnormalized values from MACS2’s log-likelihood ratio estimates. We compared the value at each position to the values of the whole track, determining the fraction of the whole track with a smaller value. We then transformed this fraction, using it as the argument to the inverse cumulative distribution function of an exponential distribution with mean parameter. We divided the genome into 100 bp non-overlapping bins, and took the mean of the transformed values within each bin. We then used these normalized and averaged values as observations for Segway in place of the initial MACS2 estimates. We trained Segway using the Expectation-Maximization algorithm and data from all three species: a randomly-sampled 10% of the human genome (with data from H1-hESC and GM12878) and the entire fly (LE and L3) and worm (EE and L3) genomes. Using these data sets jointly, we trained 10 models from 10 random initializations. In every initialization, we set each mean parameter for label \( i \) and track \( k \) by sampling from a uniform distribution defined in \( \sigma \), where \( \sigma \) is the empirical standard deviation of track \( k \). We placed a Dirichlet prior on the self-transition model to make the expected segment length 100 kb. We always initialized transition probability parameters with an equal probability of switching from one label to any other label. While these parameters changed during training, we increased the likelihood of a flatter transition matrix by including a Dirichlet prior of 10
pseudocounts for each ordered pair of labels. To increase the relative importance of the length components of the model, we exponentiated transition probabilities to the power of 3. After training converged, we selected the model with the highest likelihood. We then used the Viterbi algorithm to assign state labels to the genome in each cell type of each organism.

**Chromatin segmentation using ChromHMM**

We also compared hiHMM with another existing segmentation method called ChromHMM49. ChromHMM uses a hidden Markov model with multivariate binary emissions to capture and summarize the combinatorial interactions between different chromatin marks. ChromHMM was jointly trained in virtual concatenation mode using 8 binary histone modification ChIP-seq tracks (H3K4me3, H3K27ac, H3K4me1, H3K79me2, H4K20me1, H3K36me3, H3K27me3 and H3K9me3) from two developmental stages in worm (EE, L3), two developmental stages in fly (LE, L3) and two human cell-lines (GM12878 and H1-hESC). The individual histone modification ChIP-seq tracks were binarized in 200 bp non-overlapping, genome-wide, tiled windows by comparing the ChIP read counts (after shifting reads on both strands in the 5’ to 3’ direction by 100 bp) to read counts from a corresponding input-DNA control dataset based on a Poisson background model. A p-value threshold of 1e-3 was used to assign a presence/absence call to each window (0 indicating no significant enrichment and 1 indicating significant enrichment). Bins containing < 25% mappable bases were considered unreliable and marked as ‘missing data’ before training. In order to avoid a human-specific bias in training due to the significantly larger size of the human genome relative to the worm and fly genomes, the tracks for both the worm and fly stages were repeated 10 times each, effectively upweighting the worm and fly genomes in order to approximately match the amount of training data from the human samples. ChromHMM was trained in virtual concatenation mode using expectation maximization to produce a 19 state model which was found to be an optimal trade-off between model complexity and interpretability. The 19 state model was used to compute a
posterior probability distribution over the state of each 200 bp window using a forward-backward algorithm. Each bin was assigned the state with the maximum posterior probability.

The states were labeled by analyzing the state-specific enrichment of various genomic features (such as locations of genes, transcription start sites, transcription end sites, repeat regions etc.) and functional datasets (such as transcription factor ChIP-seq peaks and gene expression). For any set of genomic coordinates representing a genomic feature and a given state, the fold enrichment of overlap was calculated as the ratio of the joint probability of a region belonging to the state and the feature to the product of independent marginal probability of observing the state in the genome and that of observing the feature. Similar to the observations of hiHMM states, there are 6 main groups of states: promoter, enhancer, transcription, polycomb repressed, heterochromatin, and low signal.

**Heterochromatin region identification**

To identify broad H3K9me3+ heterochromatin domains, we first identified broad H3K9me3 enrichment region using SPP31, based on methods get.broad.enrichment.cluster with a 10 kb window for fly and worm and 100 kb for human. Then regions that are less than 10 kb of length were removed. The remaining regions were identified as the heterochromatin regions. The boundaries between pericentric heterochromatin and euchromatin on each fly chromosome are consistent with those from lower resolution studies using H3K9me214 (See [Ho et al., 2014] Supplementary Fig. 34).

**Genome-wide correlation analysis for heterochromatin-related marks**

For heterochromatin related marks in Fig. 2-4b, the pairwise genome-wide correlations were calculated with 5 kb bins using five marks in common in the similar way as described above. Unmappable regions or regions that have fold enrichment values < 0.75 for all five marks were excluded from the analysis.
Chromatin-based topological domains based on Principal Component Analysis

We respectively partitioned the fly and worm genomes into 10 kb and 5 kb bins, and assign average ChIP fold enrichment of multiple histone modifications to each bin (See below for the list of histone modifications used). Aiming to reduce the redundancy induced by the strong correlation among multiple histone modifications, we projected histone modification data onto the principal components (PC) space. The first few PCs, which cumulatively accounted for at least 90% variance, were selected to generate a "reduced" chromatin modification profile of that bin. Typically 4-5 PCs were selected in the fly and worm analysis. Using this reduced chromatin modification profile, we could then calculate the Euclidean distance between every pair of bin in the genome. In order to identify the boundaries and domains, we calculated a boundary score for each bin (See [Ho et al. 2014], Supplemental Methods). If a bin has larger distances between neighbors, in principle, it would have a higher boundary score and be recognized as a histone modification domain boundary. The boundary score cutoffs are set to be 7 for fly and worm. If the boundary scores of multiple continuous bins are higher than the cutoff, we picked the highest one as the boundary bin. The histone marks used are H3K27ac, H3K27me3, H3K36me1, H3K36me3, H3K4me1, H3K4me3, H3K79me1, H3K79me2, H3K9me2 and H3K9me3 for fly LE and L3, and H3K27ac, H3K27me3, H3K36me1, H3K36me3, H3K4me1, H3K4me3, H3K79me1, H3K79me2, H3K79me3, H3K9me2 and H3K9me3 for worm EE and L3.
REFERENCES


Chapter 4 -

Nucleosome fragility is associated with future transcriptional response to developmental cues and stress in *C. elegans*

**Author's Statement:**

The work in Chapter 4 has been previously presented publicly at the *C. elegans* International Worm Meeting (2015) and at the NIH Syposium on Chromosome Biology (2015), and is under review for publication at *Genome Research*. I would like to thank Coleen Murphy and members of the Lieb laboratory for feedback and comments on the manuscript. Experiments were designed by TEJ and JDL. Experiments and bioinformatics analyses were performed by TEJ with guidance and feedback from JDL. Manuscript was prepared by TEJ, and edited by JDL.

**Reference:**

Jeffers, TE and Lieb, JD. Nucleosome fragility is associated with future transcriptional response to developmental cues and stress in *C. elegans*. Submitted.
ABSTRACT

Nucleosomes have structural and regulatory functions in all eukaryotic DNA-templated processes. The position of nucleosomes on DNA and the stability of the underlying histone-DNA interactions affect the access of regulatory proteins to DNA. Both stability and position are regulated through DNA sequence, histone post-translational modifications, histone variants, chromatin remodelers, and transcription factors. Here, we explored the functional implications of nucleosome properties on gene expression and development in C. elegans embryos. We performed a time-course of micrococcal nuclease (MNase) digestion, and measured the relative sensitivity or resistance of nucleosomes throughout the genome. Fragile nucleosomes were defined by nucleosomal DNA fragments that were recovered preferentially in early MNase-digestion time points. Nucleosome fragility was strongly and positively correlated with the AT content of the underlying DNA sequence. There was no correlation between promoter nucleosome fragility and the levels of histone modifications or histone variants. Genes with fragile nucleosomes in their promoters tended to be lowly expressed and expressed in a context-specific way, operating in neuronal response, the immune system, and stress response. In addition to DNA-encoded nucleosome fragility, we also found fragile nucleosomes at locations where we expected to find destabilized nucleosomes, for example at transcription factor binding sites where nucleosomes compete with DNA-binding factors. Our data suggest that in C. elegans promoters, nucleosome fragility is in large part DNA-encoded, and that it poises genes for future context-specific activation in response to environmental stress and developmental cues.
INTRODUCTION

The fundamental unit of eukaryotic chromatin is the nucleosome, which consists of 147 bp of DNA wrapped around an octamer of histone proteins (Luger et al. 1997). Nucleosomes have important structural and regulatory functions in organizing the genome and restricting access of regulatory factors to the DNA sequence (Luger et al. 1997; Henikoff 2008). As such, the interactions between nucleosomes and DNA strongly influence the regulation of gene expression by determining DNA accessibility for transcription factors and RNA polymerase. In addition to regulated nucleosome assembly and disassembly through the action of histone chaperones and chromatin remodelers, nucleosome stability is influenced by histone modifications, histone variants, DNA features encoded in cis, and competition with DNA-binding factors in trans. A complete picture of the mechanisms governing nucleosome stability is fundamental to understanding how gene expression is dynamically regulated.

Nucleosome stability has been studied in vitro using sensitivity to enzymatic digestion or salt concentration (Bloom and Anderson 1978; Burton et al. 1978; Li et al. 1993; Polach and Widom 1995; 1999; Jin and Felsenfeld 2007; Wu and Travers 2004). Genome-wide adaptations of these methods have been used to identify nucleosome position and stability in vivo. Studies in yeast, Drosophila, plants, and mammals have used varying concentrations of the enzyme micrococcal nuclease (MNase) to identify nucleosomes with differential sensitivity to MNase digestion in vivo (Xi et al. 2011; Weiner et al. 2010; Kubik et al. 2015; Henikoff et al. 2011; Chereji et al. 2015; Vera et al. 2014; Lombraña et al. 2013). Nucleosomes sensitive to low concentrations of MNase have been labeled as “fragile”, and have been associated with transcription factor binding sites (Vera et al. 2014), active origins of replication (Lombraña et al. 2013), gene promoters (Xi et al. 2011), and genomic sequences with high AT content (Chereji et al. 2015). Thus, both DNA-encoded sequence features and trans factors influence nucleosome fragility. However, the functional implication of nucleosome fragility remains unclear. For
example, one study reported fragile nucleosomes at the promoters of repressed stress-response genes during normal growth (Xi et al. 2011), while another found fragile nucleosomes at the promoters of highly transcribed genes in yeast (Kubik et al. 2015).

We performed a timecourse of MNase digestion in C. elegans mixed-stage embryos to study the relationship between fragility and gene activity in a developing multicellular organism. In our study, fragile nucleosomes were associated with lowly expressed genes and genes expressed in a context-specific fashion. Although we found that competition with trans factors promoted nucleosome fragility, our data suggest that nucleosome fragility in the C. elegans embryo is driven by cis features encoded in the DNA sequence. Our data suggest that the fragility of nucleosome-DNA interactions acts to poise genes for induction in response to stress or developmental cues.

RESULTS

A digestion timecourse identifies nucleosomes with differential MNase sensitivity

We postulated that functionally distinct nucleosomes in the C. elegans could be distinguished by the length of time it took them to be liberated from bulk chromatin by MNase digestion. Previous studies using this approach defined nucleosomes released early in the timecourse as “fragile” and those released later in the timecourse as “resistant” (Xi et al. 2011). To identify nucleosomes of differential sensitivity genome-wide, we isolated mixed-stage embryos from C. elegans, treated them with formaldehyde to cross-link the chromatin, isolated nuclei, and digested the chromatin with MNase (Fig. 1A). After 2, 4, 8, 15, and 30 minutes of digestion we removed a chromatin aliquot and performed paired-end Illumina sequencing on the mononucleosomal fragments liberated at each time point (Fig. 4-1B). We performed two replicate experiments on native chromatin and two replicates on formaldehyde-fixed chromatin samples. Results from the native and fixed chromatin were very similar (Fig. 4-2). We therefore
focused our downstream analysis on fixed chromatin for maximum compatibility with previously-generated datasets. Although the genome-wide occupancy profiles of mononucleosomal fragments were globally similar across the timepoints (Fig. 4-1B, Fig. 4-2), there were a number of substantial differences in the nucleosome maps among the timepoints (Fig. 4-1C, Fig. 4-3F).

**Figure 4-1. An MNase digestion time course on *C. elegans* embryos.** (A) Mixed-stage embryos were collected from gravid hermaphrodites by bleach treatment. Dissociated nuclei from mixed-stage embryos were incubated with MNase for 2, 4, 8, 15, or 30 minutes. (B) Paired-end reads from each timepoint were mapped to the *C. elegans* genome, normalized, and Gaussian smoothed for display. High signals represent regions of the genome protected from MNase digestion. Region plotted: chr IV position 12,074,951 to 12,084,347. (C) Calculation of fragility and resistance scores. Fragility: for each nucleosome, the average occupancy of the intermediate timepoints is subtracted from the 2 minute timepoint. Resistance: for each nucleosome, the average occupancy of the intermediate timepoints is subtracted from the 30 minute timepoint. Intermediate timepoints are 4, 8, and 15 minutes. Region plotted: chr IV position 12,076,980 to 12,078,364. (D) Distribution of fragility and resistance scores at all nucleosomes. The top 10% of each class (shaded in green and orange, respectively) were considered “Fragile” or “Resistant.”
Figure 4.2. Accompanying Figure 4-1. (A) Representative image of an N2 embryo MNase digestion timecourse after gel electrophoresis. For each timepoint, mononucleosome-sized fragments were excised from the gel (white box) and used for paired-end Illumina DNA sequencing. Size markers (“M”) are indicated. (B,C) Mononucleosome fragments are shorter with increasing Mnase digestion time, in 10 bp increments. Comparison of nucleosome sizes between native (B) and formaldehyde-fixed (C) timecourse. (D-G) Heatmaps showing distribution of (D) fragile and (E) resistant nucleosomes in native mixed stage embryos, as well as (F) fragility and (G) resistance in fixed mixed stage embryos. Genes are sorted by gene length, maximum gene length shown is 3500 bp. Genes are aligned at the first nucleosome downstream from the transcript start site (+1 nucleosome), named the 5′ boundary nucleosome. Yellow line cartoons the center of the +1 and terminal nucleosomes. (H) Line plot comparing fragility and resistance in native (dotted lines) and fixed (solid lines) embryos around the 5′ (left) and 3′ (right) boundary nucleosomes.
Figure 4-3. Accompanying Figure 4-1. (A,B) Cross – timecourse and cross – replicate Spearman correlation in 150 bp windows between two replicates of mixed-stage embryo MNaseTC (Jee et al. 2011). RT = replicate 1, 2RT = replicate 2. (A) Unclustered timepoints. (B) Hierarchical clustering of Spearman correlation between timepoints. (C-F) Heatmaps from (C) 2 minute, (D) average of 4, 8, and 15 minute nucleosomes, and (E) 30 minute nucleosomes, sorted by length. Genes are aligned at “0”, the center of the 1st nucleosome downstream from the transcript start site. Nucleosome dyad occupancy is shown in blue – yellow. (F) Line plot average of dyad density around the 5’ and 3’ boundary nucleosomes for the early (2 minute), intermediate (average of 4, 8, and 15 minute), and late (30 minute) timepoints. (G) Scatterplot comparing fragility and resistance scores at 448,600 nucleosomes. Fragility Score = 41.2 is the threshold above which a nucleosome is considered a “fragile nucleosome”. Resistance Score = 37.9 is the threshold above which a nucleosome is considered a “resistant nucleosome”. Plot is colored by density of points. Purple = least dense, Red = most dense.

To systematically study nucleosomes of differential sensitivity to MNase, we assigned
each nucleosome both a fragility and a resistance score as follows (Fig. 4-3G). For each timepoint, we first called nucleosome positions and then assigned each nucleosome an occupancy score (see Methods for details). The fragility score for a nucleosome is defined by subtracting the average occupancy score of the intermediate timepoints (4, 8, and 15 min) from the occupancy score of the 2 min timepoint. Conversely, a resistance score is computed by subtracting the average occupancy score of the intermediate timepoints from that of the 30 min timepoint (Fig. 4-1C). Thus, fragility and resistance scores were generally reciprocal to each other at a given nucleosome, but not necessarily so. We defined the top 10% of nucleosomes with the highest fragility or resistance scores as “fragile” or “resistant” nucleosomes, respectively (Fig. 4-1D).

Trans-factors increase nucleosome fragility

We sought to address whether nucleosome fragility was a consequence of competition with DNA binding proteins and other trans factors. Trans-acting factors disrupt nucleosomes by competing with histones for binding to the DNA sequence (Simpson 1990; Adams and Workman 1995). We first examined regions of the genome where we expected to find nucleosomes destabilized by competition with other DNA-binding factors, for example at transcription factor binding sites (TFBS). We collected a set of 35,062 TFBS bound at any stage of C. elegans development, as identified by transcription factor (TF) ChIP-seq from the modENCODE consortium (Araya et al. 2014). The DNA immediately surrounding TFBS in the C. elegans genome on average show strong affinity to histones in vitro (Locke et al. 2013). A nucleosome occupancy model based solely on DNA sequence also predicted C. elegans TFBS to be nucleosome bound (Kaplan et al. 2009) (Fig. 4-4A). In vivo, however, these sites show a local decrease in nucleosome occupancy, consistent with the footprint of TF binding. Moreover, TFBS had high fragility scores on average (Fig. 4-4B). These data are in agreement with
previous reports from yeast to humans that transcription factors compete with nucleosomes for access to DNA (Wang et al. 2012; Ozonov and van Nimwegen 2013; Barozzi et al. 2014). To further investigate the relationship between TF binding and fragility we broke TFBS into groups depending on the number of TFs bound at a site. Although the majority of TFBS identified in C. elegans are bound by a single factor, some sites are bound by many TFs (Chen et al. 2014; Araya et al. 2014; Boyle et al. 2014). Fragility scores increased with the number of TFs bound at a single TFBS (Fig. 4-4C).
Figure 4.4. Competition with transcription factors influences nucleosome fragility. (A) Average reconstituted nucleosome occupancy (Locke et al. 2013) and computational nucleosome occupancy model scores (Kaplan et al. 2009) at 35,062 regions bound at any stage by any number of transcription factors. (B) Average fragility, resistance, and intermediate nucleosome occupancy scores are plotted around the same set of intervals from (A). (C) Boxplot of average fragility or resistance scores at groups of sites bound by different numbers of transcription factors. N\text{factors} = number of transcription factors bound. N\text{sites} = number of regions in each category. (D) Cartoon characterization of how embryo-specific and L4-specific HOT regions were identified. (E) Model to distinguish whether trans (top) or cis (bottom) effects result in nucleosome fragility at a given nucleosome in the embryo. Hypothetical fragility scores are represented. (F) Fragility, resistance, and nucleosome occupancy scores measured in the embryo at 119 embryo-specific HOT regions. (G) Fragility, resistance, and intermediate nucleosome occupancy scores at 88 L4-specific HOT regions.
We found that TFBS had high fragility scores despite their intrinsic preference for nucleosome formation in vitro (Fig. 4-4A). One possible explanation is that transcription factors destabilize nucleosomes at their binding sites, causing the fragility at TFBS. Alternatively, TFBS may contain DNA sequence features that promote nucleosome fragility through a mechanism that is independent of sequence influence on nucleosome occupancy. To distinguish among these possibilities, we identified a set of TFBS specifically bound at different developmental stages (Fig. 4-4D). We hypothesized that if active competition with TFs increases nucleosome fragility then TFBS bound only in the embryo should be fragile in embryos, whereas TFBS bound only in the L4 larval stage should not be fragile in embryos (Fig. 4-4E, top). Alternatively, if DNA sequence influences nucleosome fragility then the embryo-specific and L4-specific TFBS should be equally fragile in embryos (Fig. 4-4E, bottom). Due to their high fragility scores and dynamic nature, we focused our analysis on HOT regions, TFBS where significant enrichment (false discovery rate <5%) in multiple transcription factor binding sites are observed (Araya et al. 2014). We found that embryo-specific HOT regions had high nucleosome fragility and low nucleosome occupancy in our dataset, which was obtained in embryos (Fig. 4-4F, Fig. 4-5). By contrast, L4-specific HOT regions showed lower fragility and higher nucleosome occupancy in the embryonic samples (Fig. 4-4G). These results support the hypothesis that active competition with transcription factors in vivo contributes to nucleosome fragility despite their intrinsically nucleosome favoring properties in vitro.
Figure 4-5. Accompanying Figure 4-4. Ubiquitously HOT regions are fragile in vivo. Average fragility, resistance, and intermediate nucleosome occupancy plotted around the center of 847 ubiquitously HOT regions identified in Araya et al (Araya et al. 2014).

Nucleosome fragility increases throughout heat-shock genes upon induction

The preceding analysis found a correlation between TF binding and nucleosome fragility. We next sought to test the relationship between fragile nucleosomes and trans factors more explicitly. At extremely highly transcribed genes, such as heat shock responsive genes after induction, it has been proposed that RNA polymerase II (Pol II) molecules occupy the entire gene body (Schwabish and Struhl 2004; Merz et al. 2008; Cole et al. 2014). We hypothesized that fragility would increase at gene bodies after inducing high levels of transcription, as a result of nucleosome competition with transcribing Pol II. To test whether we could induce nucleosome fragility, we designed a heat shock experiment in conjunction with an MNase-seq timecourse (Fig. 4-6A).
Figure 4-6. Heat shock increases nucleosome fragility at the promoter and gene body of upregulated genes. (A) Experimental overview. Mixed-stage embryos were either incubated at room temperature (RT) or heat shocked at 34°C (HS) for 20 minutes. Subsequently, embryos were fixed and used for an MNase-seq timecourse or stored in TRIzol and used for RNA-seq. (B) mRNA-seq identifies differentially expressed genes after HS at 34°C. Significantly differentially expressed genes (p_adj < 0.1) shown in red. (C) RNA, fragility, resistance, and nucleosome occupancy scores with and without HS are plotted in the region surrounding F33E5.4 and F33E5.5, two divergently transcribed hsp-70 orthologues. Region plotted: chr II position 11,749,925 to 11,770,394. (D) Nucleosome fragility and nucleosome occupancy at 14 significantly differentially expressed genes with and without HS. (E) Nucleosome fragility, resistance, and occupancy scores at all 20,785 coding genes with and without HS.
Heat shock in *C. elegans* activates HSF-1 and HSF-2, two homologues of the mammalian HSF1 transcription factor, which bind heat shock elements (HSE) in the promoters of heat shock-responsive genes to upregulate their expression (Åkerfelt et al. 2010). Using RNA-seq, we identified 14 genes that are rapidly upregulated after a brief (20 minute) heat shock at 34 °C (Fig. 4-6B, Fig. 4-7A). We then analyzed how fragility scores changed at those genes after heat shock (Supplemental Fig. 4-6B, 4-6C). Though nucleosome occupancy decreased at heat shock-responsive genes, we found nucleosome fragility dramatically increased both 5′ and 3′ of heat-shock genes, as well as in the gene body itself (Fig. 4-6C, 4-6D). Notably, promoter and +1 nucleosome fragility increased on average genome-wide, although gene-body fragility was specific to the set of heat shock-induced genes (Fig. 4-6E).
Figure 4-7. Accompanying Figure 4-6. (A) RNA-seq cross-correlation plot for two replicates of room temperature and two replicates of heat shock embryos. (B) Spearman correlation between two replicates of heat shock and two replicates of room temperature control mixed stage embryo MNaseTC experiments, unclustered (B) or hierarchically clustered (C) (Jee et al. 2011).
High transcription rates have been suggested to remove the entire histone octamer, superseding FACT-mediated H2A-H2B recycling (Kireeva et al. 2002; Kulaeva et al. 2010). But nucleosomes can also be removed from gene bodies independent of transcription. Previous studies of the Hsp70 locus in Drosophila have shown that heat shock induces rapid and transcription-independent loss of gene-body nucleosomes (Petesch and Lis 2008). Finally, it is possible that the 14°C temperature increase itself perturbs nucleosomes. Chereji et al. observed a related effect: nucleosomes in Drosophila S2 cells cultured at 18 °C are more stable than when cultured at 27 °C (Chereji et al. 2015). Future experiments may clarify the exact mechanism by which gene body nucleosomes become fragile after heat shock.

Nucleosome fragility or resistance is associated with stereotypic nucleosome locations within genes

We found high fragility scores at genomic locations where we expected to find destabilized nucleosomes, like transcription factor binding sites and the gene bodies of newly induced genes. We next investigated the genome-wide distribution of fragile nucleosomes (nucleosomes with the highest 10% of fragility scores; Fig. 4-1D) in detail (Fig. 4-8). Fragile nucleosomes were enriched 5’ and 3’ of genes, specifically at the promoter -2, -1, and +1 nucleosomes, and at the terminal nucleosome (TN) and TN+1 nucleosomes (Fig. 4-8B, Fig. 4-2H). Resistant nucleosomes (nucleosomes with the highest 10% of resistance scores (Fig. 4-1D) were enriched in gene bodies (Fig. 4-8C, Fig. 4-2H).
Figure 4-8. Fragility is enriched 5′ and 3′ of genes and is anti-correlated with gene expression. (A) Log$_2$ DESeq2-normalized number of reads measured by mRNA-seq at 20,785 genes, ordered by their relative expression. (B) Heatmap of fragility scores (green) at genes ordered as in (A). Genes were aligned at the center of the first nucleosome downstream from the transcript start site, known as the +1 or 5′ boundary nucleosome (yellow line). (C) Same as in (B), except resistance scores are plotted in orange. (D) Same as (B), except for HTZ-1 input-normalized ChIP-seq signals (Ho et al. 2014). (E) Same as (B), except the average GC content (as a percentage of 100%) in 5 bp windows is plotted. (F) Fragility, (G) resistance, and (H) nucleosome occupancy scores around the 5′ and 3′ boundary nucleosomes averaged over expression quintiles (highest expressed 20% in dark orange, green, or black, lowest expressed 20% in lightest orange, green, or black). Quintile 1: 0 to 4.5 normalized counts. Quintile 2: 4.5 to 65. Quintile 3: 65 to 619. Quintile 4: 619 to 2209. Quintile 5: > 2209.

Nucleosome fragility and resistance is not correlated with nucleosome occupancy

To investigate whether nucleosome fragility or resistance scores were a consequence of nucleosome positioning or occupancy, we asked whether fragility or resistance scores were correlated with nucleosome occupancy or the standard deviation of the nucleosome center (“fuzziness”). Nucleosome fragility scores were not correlated with nucleosome occupancy in the intermediate MNase timepoint (R = -0.02) or with nucleosome occupancy as measured by an independent histone H3 ChIP (R = -0.14) (Supplemental Results and Fig. 4-9). In addition,
neither nucleosome fragility nor nucleosome resistance scores were correlated with the fuzziness of the nucleosome at the intermediate timepoint (fragility vs. fuzziness: $R = 0.03$, resistance vs. fuzziness: $R = -0.06$), suggesting that susceptibility to MNase digestion is an independent feature of nucleosomes, and is not a direct consequence of nucleosome occupancy or positioning (Fig. 4-9).
Figure 4-9. Accompanying Figure 4-8. (A) Scatterplots comparing intermediate nucleosome occupancy (average of 4, 8, and 15 minute nucleosomes) and fragility scores, 2 minute nucleosome occupancy and fragility scores, intermediate nucleosome occupancy and resistance scores, 30 minute nucleosome occupancy and resistance scores, 2 minute nucleosome fuzziness and fragility scores, intermediate nucleosome fuzziness and fragility scores, intermediate nucleosome fuzziness and resistance scores, and 30 minute nucleosome fuzziness and resistance scores. (B) Nucleosome occupancy scores from histone H3 ChIP-seq around the 5′ and 3′ boundary nucleosomes averaged over expression quintiles (highest expressed 20% in black, lowest expressed 20% in lightest gray) (Ho et al. 2014a). Quintile 1: 0 to 4.5 normalized counts. Quintile 2: 4.5 to 65. Quintile 3: 65 to 619. Quintile 4: 619 to 2209. Quintile 5: > 2209. (C) Same as in (B), except nucleosome occupancy scores from the intermediate (average of 4, 8, and 15 minute timepoints) as measured in this study by MNase-seq are plotted. (D) Log2 DESeq2-normalized number of reads measured by mRNA-seq at 20,785 genes, ordered by their relative expression. (E) Heatmap of H3 ChIP-seq scores at genes ordered as in (D). Genes were aligned at the center of the first nucleosome downstream from the transcript start site, known as the +1 or 5′ Boundary Nucleosome. ChIP-seq data from Ho et al., 2014. (F) Same as in (E), except intermediate nucleosome occupancy scores measured by MNase-seq are plotted. (G) Scatterplots comparing nucleosome occupancy at 448,600 nucleosomes as measured by histone H3 ChIP-seq or by MNase-seq performed in this study. (H) Scatterplots comparing nucleosome occupancy as measured by histone H3 ChIP-seq and nucleosome fragility or resistance scores measured in this study.

Nucleosome fragility near genes is anti-correlated with expression

Gene expression levels on average were anti-correlated with nucleosome fragility at both promoters and gene bodies (Fig. 4-8A, 4-8B, 4-8F, Fig. 4-10B. R = -0.17) and positively correlated with nucleosome resistance (Fig. 4-8A, 4-8B, 4-8G, Fig. 4-10B. R = 0.11). In contrast to our earlier observation at the heat shock genes, we found no correlation between expression and nucleosome fragility at the gene bodies of the most highly transcribed genes (compare Fig. 4-6D to Fig. 4-10C). It is possible that only newly-induced genes display gene-body fragility, or extremely high levels of transcription are required to induce fragility in gene bodies.
Figure 4-10. Accompanying Figure 4-8. (A) Transcription factor binding sites in the embryo are biased towards expressed genes. Heatmaps sorted by gene expression (highest expressed genes at top). modENCODE ChIP-seq of transcription factors from embryo stages are shown for comparison. EM: embryo. LE: late embryo. EE: early embryo. mxE: mixed-stage embryo. ChIP data from Araya et al. 2014 (Araya et al. 2014), expression data from mixed-stage embryos, this study. (B) Scatterplots displaying the correlation between nucleosome fragility (top) or nucleosome resistance (bottom) and expression (left), GC content (middle), and HTZ-1 occupancy (right). Expression is measured as the number of normalized RNA-seq counts that mapped to a nucleosome. HTZ-1 occupancy is from (Ho et al. 2014). (C) The gene body of highly expressed genes is not fragile. Fragility scores averaged at the highest 20% (green), highest 5% (pink), or highest 1% (yellow) of expressed genes in the embryo. Genes are aligned at the +1 nucleosome. (D) H2A.Z containing nucleosomes are not enriched at fragile or resistant nucleosomes. Of the total 5,042 H2A.Z nucleosomes, 1,486 were also identified to be fragile, while 2614 were identified as resistant (Ho et al. 2014a).
Although nucleosome fragility scores were generally high at the 5’ and 3’ ends of all genes, including at the majority of TFBS (Fig. 4-4), fragile nucleosomes occurred preferentially at the promoters of lowly-expressed genes (Fig. 4-8B, 4-8F, Fig. 4-10A). Examination of the existing ChIP data did not yield clues to the mechanism underlying the preferential nucleosome fragility at low-expressing gene promoters. No single transcription factor profiled in the embryo significantly overlapped the distribution of fragile nucleosomes (Fig. 4-10A). Further, although previous reports have suggested that the histone variant H2A.Z may act to promote nucleosome instability (Jin and Felsenfeld 2007; Jin et al. 2009; Xi et al. 2011), we did not observe a significant overlap between previously-identified H2A.Z-containing nucleosomes (Ho et al.) and fragile nucleosomes (Fig. 4-8D, Supplemental Fig. 4-10B, 4-10D). Our data, placed in the context of the existing literature, suggests that two separate mechanisms account for nucleosome fragility, depending on the genomic context. In places where nucleosomes are directly in competition with transcription factors (Fig. 4-4), or in the bodies of exceptionally highly expressed or newly induced genes (Fig. 4-6), fragility arises through competition with transcription factors or other DNA-binding proteins. By contrast, fragility at the 5’ and 3’ end of genes at locations with few TF binding events appears to be determined by another mechanism, which we explored next.

**Nucleosome fragility is correlated to cis-encoded DNA features**

We hypothesized that *cis* features may be responsible for the fragility of nucleosomes at the promoters of lowly-expressed genes. We examined the DNA sequences occupied by fragile and resistant nucleosomes, and compared these to sequences occupied by all nucleosomes in the genome (Fig. 4-11). Compared to the set of all nucleosomes, DNA sequences occupied by fragile nucleosomes had lower GC sequence content on average, a feature favoring nucleosome formation (Fig. 4-11A). We then asked whether these sequences were likely to
form nucleosomes based on a previously reported *in vitro* reconstitution assay (Locke et al. 2013). DNA sequences occupied by fragile nucleosomes in the embryo were generally less occupied *in vitro* (Fig. 4-11C). Finally, we observed that sequences occupied by fragile nucleosomes were less conserved across nematodes than DNA sequences occupied by the set of all nucleosomes (Fig. 4-11E).

**Figure 4-11.** Fragile nucleosomes contain AT-rich, nucleosome disfavoring, and poorly conserved DNA. (A) Histogram of average GC content at fragile (green) or all nucleosomes (grey). (B) Same as (A) for resistant (orange) or all nucleosomes (grey). (C) Histogram of *in vitro* nucleosome occupancy scores at fragile or all nucleosomes. (D) Same as (C) for resistant nucleosomes or all nucleosomes. (E) Histogram of PhastCons 7-way conservation score at fragile or all nucleosomes. (F) Same as (E) for resistant or all nucleosomes.

Poly(dA:dT) tracts disrupt nucleosome formation and tend to increase transcription of downstream genes (Raveh-Sadka et al. 2012), while TATA box motifs in yeast are associated with bendable promoters sensitive to chromatin remodelers (Tirosh et al. 2007; Albert et al.
2007). In *C. elegans*, the number of T-block motifs (3 to 5 consecutive thymine nucleotides, often spaced at 10 bp periodicity) have been positively correlated with expression: genes with more than 5 T-blocks have fivefold higher expression than genes with fewer than 4 T-blocks (Grishkevich et al. 2011), presumably through a reduction in promoter nucleosome occupancy. T-blocks were not enriched at fragile or resistant nucleosomes, whereas TATA box motifs were enriched at fragile nucleosomes (Fig. 4-12A). Taken together, our data indicate that fragile nucleosomes in gene promoters are correlated with high AT content and TATA box motifs. It seems that most promoters are fragile at least in part due to high AT content (see residual fragility in Fig. 4-4G), but that this *cis* effect of DNA sequence becomes apparent only at sites where the observation is not confounded by TF binding.
Figure 4-12. Accompanying Figure 4-11. (A) Fragile nucleosomes are enriched for TATA box motifs; both fragile and resistant nucleosomes are underrepresented with T-block motifs. 146 bp sequences underneath fragile, resistant, and random nucleosomes were searched for perfect matches to TATAAA, TTTTTT (T6), or TTTTTTTTT (T9) motifs using FIMO (Grant et al. 2011). Numbers of perfect matches were normalized to the number of nucleosomes used: fragile nucleosomes, 48,110; resistant...
nucleosomes, 48,089; randomly chosen nucleosomes, 48,000. All comparisons were significant by chi-square test with a p-value < 2.2E-16. (B) Spearman correlation between histone modifications, histone variants, and other ChIP-seq experiments in 150 bp sliding windows. 80 mM and 350 mM extracted nucleosomes from (Ooi et al. 2010). LEM-2 from (Ikegami et al. 2010). All other data from (Ho et al. 2014a). (C) Fragile nucleosomes are associated with 80 mM salt-extracted nucleosomes. Resistant nucleosomes are associated with H3.3, H3K36me1/2/3, H3K79me1. Histone post-translational modifications score at all nucleosomes (gray), fragile nucleosomes (green), or resistant nucleosomes (orange). Histone modifications from (Ho et al. 2014a). H3.3 and 80 mM extracted chromatin from (Ooi et al. 2010). (D) Enrichment of fragile and resistant nucleosomes at chromatin states. Chromatin states from (Ho et al. 2014b). Blank box: no enrichment of histone post-translational modification. +, ++, ++++, +++++: relative enrichment of histone post-translational modification in the chromatin state. (E) Longer linkers are weakly correlated with (left) higher nucleosome fragility and (right) decreased nucleosome resistance. Flanking linker length: total length of linker surrounding a nucleosome (upstream + downstream). Fragility quartiles: all nucleosomes are assigned a fragility score. Nucleosomes are broken into equally sized quartiles. Q1 = 120,205 nucleosomes with lowest fragility scores, Q4 = 120,205 nucleosomes with highest fragility scores. Same process applied for resistance quartiles. Statistical significance evaluated with chi square test. * p-value < 0.05, ** p-value < 1e-08, *** p-value < 1e-12. All other comparisons not significant. (F) GC content is the strongest predictor of nucleosome fragility and resistance scores. Relative importance of each feature for predicting nucleosome fragility and resistance scores was calculated using the ‘lmig’ metric from the R package “relimpo”: the R² contribution averaged over orderings among regressors (Gromping 2006). (Top) The model composed of %gc + genic + expr + fStd + iStd + linker + HeiT + MGW + OCR2 + ProT + Roll + WW + SS explains 23.54% of the variance in nucleosome fragility scores. Of the explained variance, GC content is the most explanatory. (Bottom) The model composed of %gc + genic + expr + fStd + iStd + linker + HeiT + MGW + OCR2 + ProT + Roll + WW + SS explains 8.13% of the variance in nucleosome resistance scores. Of the explained variance, GC content is the most explanatory. %gc = average GC content, gene = genic or non genic sequence, expr = expression level, fStd = nucleosome fuzziness at 2m timepoint, iStd = nucleosome fuzziness at average of intermediate timepoints (4, 8, 15m), linker = length of flanking linker, fOcc = nucleosome occupancy at 2m timepoint, iOcc = nucleosome occupancy at average of intermediate timepoints (4, 8, 15m), HeiT = helical twist, MGW = major groove width, ORC2 = hydroxyl radical cleavage ORChID2, ProT = propeller twist, Roll = DNA roll, WW = frequency of (AA, AT, TA, or TT) dinucleotides, SS = frequency of (GG, GC, CG, CC) dinucleotides. DNA shape properties are from (Chiu et al. 2015). (G) Correlation between observed and predicted fragility (top) and resistance (bottom) scores using the model fragility score = %gc + genic + expr + fStd + iStd + linker + HeiT + MGW + OCR2 + ProT + Roll + WW + SS. Model was trained on 90% of the data and tested on the remaining 10%. (H,I) Normalized signal from 80 mM salt-extracted nucleosomes and (J) H3.3 ChIP’d nucleosomes from (Ooi et al. 2010). Genes are sorted by expression level as measured in the current study. (J,K) Correlation between 80 mM salt extract nucleosome occupancy and fragility scores measured in this study. (L) Correlation between H3.3 occupancy and resistance scores measured in this study.

**Nucleosome fragility is not associated with a specific epigenetic state**

Through comparison with previously generated datasets, we asked whether any histone post-translational modifications, histone variants, or chromatin states were positively associated with nucleosome fragility or resistance (Ooi et al. 2010; Ho et al. 2014b). No specific chromatin modification or combination of modifications were associated with nucleosome fragility. Only “low signal” chromatin states and chromatin extracted with 80 mM salt (another method
proposed to identify unstable nucleosomes) were associated with fragile nucleosomes (Fig. 4-12) (Ooi et al. 2010; Ho et al. 2014b). Although longer linkers were weakly correlated with increased fragility levels (Fig. 4-12E), the GC content of the nucleosome was the strongest predictor of overall nucleosome fragility score (Fig. 4-12F).

**Fragile nucleosomes are associated with genes expressed in context-specific situations**

To infer potential functional implications of nucleosome fragility in the developing embryo, we next asked which genes were significantly associated with fragile nucleosomes. We identified two sets of genes that contain fragile and resistant nucleosomes, and then looked for enriched gene ontology (GO) terms in each of the gene sets. Genes with fragile nucleosomes were enriched for GO terms related to neuronal response, immune response, and stress response genes (“sensory perception of chemical stimulus”, “defense response”, “pharynx development”, “immune system process”) (Fig. 4-13A). In contrast, genes with resistant nucleosomes were enriched for embryogenesis and cell cycle related terms (“mitotic cell cycle”, “RNA processing”, “regulation of developmental process”, “organic substance transport”) (Fig. 4-13B). Because of their anti-correlation with expression and transcription factor binding, this second class of fragile nucleosomes in the embryo is unlikely to be fragile due to the action of *trans* factors. Rather, these fragile nucleosomes were associated with lowly-transcribed genes that are expressed in a context-specific fashion during stress response or development.
To confirm the association between fragile nucleosomes and future context-specific expression with an independent method, we used the modENCODE transcriptome data from seven different life stages to define a set of “developmentally regulated” genes and a set of “stably expressed” genes (Fig. 4-13C) (Spencer et al. 2011; Gerstein et al. 2014; Pérez-Lluch et al. 2015). We hypothesized that if promoter nucleosome fragility is related to context-specific expression as our GO analysis suggested, then we should find higher fragility signals near developmentally regulated genes. When we plotted the average fragility scores around these genes, we indeed saw higher nucleosome occupancy and fragility signals at developmentally regulated genes as compared to the set of stably expressed genes (Fig. 4-13D). While both sets
of genes have fragile promoters, our data indicate that fragility is enriched at genes that tend to be expressed specifically during development, stress, or environmental stimulus response. Together, we suggest that these sequences may reflect a specialized promoter architecture that is primarily determined by high AT content, which acts to allow future disruption of nucleosome stability and thereby the rapid induction of gene expression in a context-specific fashion (Fig. 4-14).

**Figure 4-14.** We propose a model whereby nucleosome fragility is determined by two distinct mechanisms, one that operates in cis at all genes, and one that operates in trans at a subset of genes. **Left:** Competition in trans with transcription factors and polymerase machinery destabilizes nucleosomes at the promoters of stably expressed genes. **Right:** Condition-specific and developmentally regulated genes contain promoters with high levels of nucleosome fragility, determined primarily in cis by high AT content. Green line: high AT content is sequence-encoded at all promoters, but is highest at condition-specific genes. Orange cylinders: resistant nucleosomes found in the gene body of highly and stably expressed genes. Green cylinders: fragile nucleosomes compete (single arrow) with transcription factors and RNA Pol II at stably expressed genes. Fragile nucleosomes at condition-specific genes “treadmill” on the DNA (three arrows) due to destabilizing DNA elements like TATA-box motifs and high AT content.

**DISCUSSION**

We performed an MNase digestion timecourse, a simple modification to the traditional MNase digestion assay, in *C. elegans* embryos. Our experiment measured which individual nucleosomes were most quickly released from their polynucleosome context after exposure to MNase. Sensitivity to MNase digestion, and thereby fragility or resistance as defined in this study, could be determined by a number of factors. These include (1) a DNA sequence that is
preferentially cut by MNase; (2) longer linker regions; (3) low DNA-histone affinity; or (4) competition with transcription factors. Two lines of evidence suggest that nucleosome fragility reflects nucleosome instability. First, we found that nucleosomes can be made fragile by competition with transcription factors and Pol II. Second, we observed that nucleosomes can be made unstable in cis by being wrapped around nucleosome-disfavoring DNA sequences with high AT content and TATA-box motifs. All of these factors have been shown to cause nucleosome instability in previous studies (Widom 2002; Ozonov and van Nimwegen 2013).

We performed our experiments in nuclei derived from whole embryos, which reflect a mixture of cell types. This creates challenges for data interpretation. For example, from this heterogeneous mixture of cells, we observed that promoter nucleosome fragility decreases with increasing gene expression. In our data, ubiquitously expressed genes are more likely to fall into the 'highly expressed' category, while a gene expressed in a cell-type specific fashion is more likely be part of the 'lowly-expressed' category, due to its lower expression on average in the mixture of many cell types. Therefore, for genes expressed in every cell (or most cells), our data are simpler to interpret relative to nucleosome behavior.

Further, we may be less able to detect fragile nucleosomes that occur in a small fraction of the cell population. This may include nucleosomes at the promoters of genes expressed in a cell-type specific manner, which are likely to differ in their fragility between expressing and non-expressing cells. For example, promoter nucleosomes of myosin genes may be fragile in muscle cells and non-fragile in neuronal or intestinal cell types. As we observed in the heat shock experiment, response to a stimulus can affect nucleosome fragility. Thus, fragile nucleosomes at these cell-type specific genes are potentially under-represented in our data. Future experiments using embryo dissociation and FACS enrichment for a certain cell or tissue types may be able to test this hypothesis.
MNase resistant nucleosomes tend to be in gene bodies and correlated with expression

We found a class of nucleosomes that required relatively long durations of MNase digestion to be removed from chromatin. Traditional expectations might be that unstable nucleosomes would be found in the body of transcribed genes, and stable nucleosomes in silent, heterochromatic genomic regions. However, recent reports illustrate that nucleosomes in the gene body of transcribed genes are consistently well-positioned and highly occupied due to a number of factors, including the activity of Pol II and the histone chaperone FACT (FAcilitates Chromatin Transcription) (Jiang and Pugh 2009; Bai and Morozov 2010). Indeed, we found stable nucleosomes enriched in the gene body of actively transcribed housekeeping genes. Histone modifications such as H3K36me3, which we found positively correlated with stable nucleosomes, are also thought to contribute to nucleosome stability and maintenance of transcription fidelity (Lieb and Clarke 2005; Lickwar et al. 2009). Together, our measurements agree with an emerging picture of highly regulated nucleosome stability throughout the genome, which is likely critical for regulation of DNA templated events like transcription, splicing, and DNA replication (Bintu et al. 2011; Kwak et al. 2013; Tilgner et al. 2009; Eaton et al. 2010).

MNase sensitive fragile nucleosomes are 5′ enriched and anti-correlated with expression

Differential MNase digestion and salt fractionation have been previously used to probe nucleosome-DNA stability. Results from yeast (Xi et al. 2011; Weiner et al. 2010; Kubik et al. 2015), plants (Vera et al. 2014), mouse (Lombraña et al. 2013; Deng et al. 2015; Iwafuchi-Doi et al. 2016; Mieczkowski et al. 2016), worm (Ooi et al. 2010), and fly (Chereji et al. 2015; Henikoff et al. 2009) have identified highly labile nucleosomes in 5′ and 3′ “nucleosome free” regions. In yeast, Xi et al. observed that fragile nucleosomes were associated with H2A.Z containing promoter nucleosomes, believed to be involved in stress response (Zhang et al. 2005; Li et al. 2005). In vertebrates, individual nucleosomes containing both H3.3 and H2A.Z histone variants
are unstable (Jin and Felsenfeld 2007; Jin et al. 2009). We did not observe a correlation between H2A.Z incorporation and nucleosome fragility in *C. elegans* as measured by our assay. Rather, H2A.Z is distribution is strongly biased towards active genes (Whittle et al. 2008; Liu et al. 2011). This distinction could be due to a divergence in H2A.Z properties between yeast and *C. elegans* (Zlatanova and Thakar 2008). It is also possible that the H2A.Z-containing nucleosomes used for this study (measured in Ho et al.) was comprised of the particularly stable, homotypic type of H2A.Z nucleosomes (Ishibashi et al. 2009).

Previous reports disagree about the relationship between nucleosome fragility and expression. Studies in yeast, *C. elegans*, *Drosophila*, and Maize have used salt profiling and different MNase concentrations to identify a positive correlation between promoter fragility and expression (Kubik et al. 2015; Ooi et al. 2010; Henikoff et al. 2009; Vera et al. 2014). However, Xi et al used the same approach and observed the opposite effect. One possible explanation for the discrepancy is that some techniques may recover fragments from nucleosome-depleted loci, like the nucleosome-free regions in the promoters of active genes. A recent study performed a careful comparison of hypersensitive (fragile) nucleosomes and histone occupancy and concluded that fragile nucleosomes can come from one of two classes: from (1) nucleosome-depleted open-chromatin regions that are easily cleaved by light MNase, or (2) from truly nucleosome occupied regions with destabilized nucleosomes (Iwafuchi-Doi et al. 2016). It is possible that the fragile nucleosomes we observe at highly transcribed genes are of the first class: nucleosome depleted regions that are rapidly digested during low MNase conditions. In contrast, the highly fragile nucleosomes observed at the promoters of lowly transcribed genes – where we see the highest fragility signals – are more likely to be highly occupied nucleosomes that also have high fragility scores. This interpretation may help to clarify the apparently contradictory reports of nucleosome fragility enriched at either highly transcribed genes or at
lowly transcribed genes. Further, slight variations in MNase digestion conditions may bias the capture of these two different classes of fragile nucleosomes.

In addition to sensitivity to MNase, other methods have been proposed to identify unstable or fragile nucleosomes. Washing chromatin with different ionic-strength buffers can be used to isolate nucleosomes of fundamentally different stabilities: unstable nucleosomes are liberated at low salt concentrations, whereas more stable nucleosomes require higher salts to be disrupted. Ooi et al. used differential salt profiling followed by tiling microarray to investigate where 80 mM salt-extracted nucleosomes are located on chromatin isolated from C. elegans embryos (Ooi et al. 2010). Both fragile nucleosomes and 80 mM salt extracted nucleosomes are enriched 5’ and 3’ of genes (Supplemental Fig. S7H-K). However, 80 mM salt nucleosomes are more highly recovered from highly expressed genes. In contrast, fragile nucleosomes observed in this study show the opposite enrichment, with higher fragile nucleosome recovery at lowly expressed genes.

Xi and colleagues identified nucleosome fragility at nearly one-third of all promoters of protein-coding genes, and enriched at the promoters of genes involved in stress response. When we assessed the types of functional annotations that were enriched at promoters with fragile nucleosomes in C. elegans embryos, we identified GO terms related to context-specific expression: sensory perception of chemical stimulus, defense response, immune system process. Based on our findings in conjunction of those with Xi et al., we propose that nucleosome fragility may serve to poise genes for rapid activation in response to developmental or external stimuli. This is consistent with previous work investigating the transcriptional activation of mammalian primary response genes, where unstable nucleosomes are used to achieve rapid induction independent of chromatin remodeling complexes (Ramirez-Carrozzi et al. 2009).
We found high fragility scores at the -2, -1, and +1 nucleosomes of developmentally regulated genes in comparison to stably expressed housekeeping genes. Previous work found that the promoters of developmentally regulated genes lack the histone post-translational modifications associated with active genes, like H3K4me3 (Pérez-Lluch et al. 2015). Likewise, Iwafuchi-Doi noticed few histone modifications associated with sensitive nucleosomes at tissue-specific and stage-specific enhancers. Similarly, we were unable to find an association between fragile nucleosomes and any histone post-translational modifications examined by the modENCODE group. Given the increased fragility of these nucleosomes, it is possible that (1) these nucleosomes at developmentally regulated genes were lost from standard chromatin preparation protocols and are thus underrepresented in the histone ChIP, or (2) developmentally regulated genes use promoter nucleosome fragility as a mechanism for gene regulation. Iwafuchi-Doi proposes that the activity of pioneer transcription factors like FoxA1/A2 are responsible for destabilizing nucleosomes or displacing the stabilizing linker histone H1. We observed no correlation between transcription factor binding and fragile nucleosomes at cell-type specific genes. This could be for a number of reasons, including (1) the TF ChIP was also performed on a heterogenous mixture of cells, making it difficult to recover signal from cell-type specific binding events, or (2) the cell-type specific / pioneering TF was not a member of the modENCODE ChIP collection.

Our results are reminiscent of previous reports from yeast, which propose that promoter structures can generally be classified as containing depleted proximal nucleosomes (DPN) or occupied proximal nucleosomes (OPN) (Tirosh and Barkai 2008). In yeast, DPN genes have low transcriptional plasticity (defined as the capacity to modulate transcription levels upon changing conditions), well positioned nucleosomes, and are enriched for TF binding sites and H2A.Z. In contrast, OPN genes have high transcriptional plasticity, higher evolutionary divergence, higher nucleosome turnover, and are sensitive to chromatin regulation (Lickwar et
al., 2014. The yeast DPN genes may correspond to the set of stably-expressed genes we defined in *C. elegans*, which have depleted proximal nucleosomes. The yeast OPN genes may correspond to the set of developmentally regulated genes we defined in *C. elegans*, which have high promoter fragility and highly occupied proximal nucleosomes. To our knowledge, OPN and DPN-type promoters have not been previously described or defined in *C. elegans*. Our results are consistent with a model in which nucleosome fragility is encoded at the promoters of OPN-type genes, potentiating the high transcriptional plasticity observed at these sites. The presence of these promoter structures in yeast, human, and now *C. elegans* suggests a well-conserved strategy that uses nucleosome architecture to regulate the dynamics of gene expression.
MATERIALS AND METHODS

Worm strains and growth in liquid culture

Wild type N2 worms were obtained from the Caenorhabditis Genome Center and maintained at 20°C in liquid culture as previously described (Whittle et al. 2008; Ercan et al. 2011). Mixed-stage embryos were isolated from gravid adults by bleach hypochlorite treatment and fixed with 2% formaldehyde for 30 minutes at room temperature.

MNase Digestion Time Course

MNase digestion was performed as previously described (Ercan et al. 2011), with slight alterations. Micrococcal nuclease (Worthington LS004798) was resuspended in water at 50 U/uL and frozen in individual aliquots at -80°C. To control for variability in enzyme activity, individual aliquots were removed from the freezer and thawed on ice, and never reused. Mixed-stage embryos were incubated with chitinase (Sigma Cat #C6137), washed, and dounced in dounce buffer (0.35 M sucrose, 15 mM HEPES-KOH pH 7.5, 0.5 mM EGTA, 5 mM MgCl₂, 10 mM KCl, 0.1 mM EDTA, 1 mM DTT, 0.5% TritonX-100, 0.25% NP-40) to extract nuclei. Nuclei were pelleted and washed with MNase digestion buffer (110 mM NaCl, 40 mM KCl, 2 mM MgCl₂, 1 mM CaCl₂, 50 mM HEPES-KOH pH 7.5). MNase was added, and at each timepoint (0, 2, 4, 8, 15, or 30 minutes after enzyme addition) a fraction of the reaction was removed, quenched with EDTA, and stored on ice. Samples were treated with Proteinase K for 2 hours at 55°C, then incubated overnight at 65°C to reverse crosslinks. DNA was isolated from RNA and proteins using phenol:chloroform extraction and RNase A treatment for 1 hour at 37°C. Mononucleosome-sized fragments (100 to 200 bp) were extracted from a 2% agarose gel and purified using a Qiagen gel extraction kit.
**Heat shock**

Mixed-stage embryos were isolated as described and split into two pools. One pool was incubated at 34°C for 20 minutes with intermittent brief mixing, while the other pool nutated at room temperature. After 20 minutes, an aliquot from each pool was saved for RNA-seq, while the remaining embryos were fixed for 30 minutes in 2% formaldehyde at room temperature.

**RNA isolation**

Embryos were dropped into TRIzol (Life Technologies) and flash frozen in liquid nitrogen after incubation for 20 minutes at room temperature or 34°C heat shock. Embryos were homogenized by thawing at 37°C and refreezing in liquid nitrogen 3x. Total RNA was isolated using a TRIzol/chloroform extraction followed by RNeasy Mini (Qiagen) preparation with On Column DNaseI Digestion (Qiagen).

**Illumina Library Preparation**

Individual libraries were prepared with unique barcodes for each timepoint from the timecourse. MNaseTC libraries were prepared from 100 ng of gel-extracted DNA using the Illumina TruSeq DNA library preparation kit v2 (FC-121-2001) according to manufacturer instructions. Ampure beads were used for the final purification in lieu of gel purification. RNA-seq libraries were prepared from 2 ug of total RNA using the Illumina TruSeq RNA library preparation kit v2 (RS-122-2001) according to manufacturer instructions. Individual samples were barcoded with unique 6bp index sequences contained within the sequencing adapters. Individual libraries were then pooled at equimolar ratios for paired-end multiplex sequencing.
Illumina Sequencing and Post-Processing

Paired-end sequencing was performed by the Princeton University Sequencing Core Facility according to Illumina protocols. Paired end reads were mapped to the UCSC Oct. 2010 (WS220/ce10) genome release using Bowtie (v1.1.2) with stringent multimapping parameters:

bowtie -q -X 2000 --fr -p 1 -S -n 2 -e 70 -l 28 --pairtries 100 --maxbts 125 -k 1 -m 1 --un /Unmapped_Reads.fastq --phred33-quals /ce10 -1 /read1.fastq -2 /read2.fastq

Nucleosome analysis

Reads with insert sizes between 100 and 250 bp were kept for downstream analysis. Replicates were first processed individually, then pooled after confirming a high degree of correlation between replicates. Nucleosome analysis was performed as described previously (Kaplan et al. 2010; Gossett and Lieb 2012).

Coverage: Nucleosome coverage was calculated by extending the filtered mapped reads to their fragment length and measuring the sum of reads covering each bp. To normalize for variation between samples, nucleosome coverage was scaled by 1/(mean coverage), yielding a mean nucleosome coverage of 1.0.

Dyad Occupancy: Dyads are approximated as the center of a paired-end fragment. The number of dyads at each base pair was scaled by 1/(mean dyad density), then Gaussian smoothed with a standard deviation of 20 bp.

Nucleosome calls: Nucleosome positions were identified from dyad density maps using a previously reported greedy algorithm (Albert et al. 2007; Gossett and Lieb 2012). Using the local maxima of the dyad density as the nucleosome center \( p \), the size of the nucleosome (the nucleosome-protected region) was determined by measuring the average length of all reads.
that covered the nucleosome center. The standard deviation of the nucleosome center (the nucleosome “fuzziness”) was calculated for each called nucleosome as the standard deviation of dyads around the mean. Nucleosome occupancy was defined as the number of dyads that fell within 50 bp of the nucleosome center.

**Boundary nucleosomes:** Using these called nucleosome positions 5′ and 3′ boundary nucleosomes were identified for the 20,578 RefSeq annotated genes. 5′ +1 nucleosomes were identified as the first nucleosome call with a dyad coordinate downstream of the 1st coding exon. Similarly, the 3′ boundary nucleosome was identified as the first nucleosome call with a dyad coordinate upstream of the TTS. Because *C. elegans* utilizes trans-splicing, the 5′ end of mature polyadenylated mRNAs, and the RefSeq annotations used in this analysis, do not reflect the exact base pair position of transcription initiation. Although recent studies have used novel methods to identify the true transcription initiation sites (Chen et al. 2013; Kruesi et al. 2013; Saito et al. 2013), the TSS annotations are only known for a subset of expressed genes in a small number of stages (Chen et al. only tested embryos, only 31.7% of genes had a TSS in at least 1 of 3 stages tested in Kruesi et al., Saito et al., only tested embryo and adult). For completeness, we chose to instead investigate the full set of known genes using their first coding exon as an alignment point.

**Nucleosome Fragility and Resistance scores**

The pooled “intermediate” nucleosome profile was generated by pooling the reads from the 4, 8, and 15 minute time points from each replicate. The pooled reads were used to generate average nucleosome positions, fuzziness scores, and occupancies. To identify regions of the genome that were liberated earlier or later than average, we subtracted the occupancy of the pooled sample from either the 2 minute (2m – pool = Fragility score) or the 30 minute samples
(30m – pool = Resistance score). To highlight regions significantly enriched with this signal, we considered the 10% of nucleosomes with the highest fragility or resistance scores as Fragile or Resistant Nucleosomes.

**Nucleosome occupancy scores**

Reads from the 4, 8, and 15 minute digestion timepoints were pooled and used to calculate nucleosome coverage and nucleosome dyad occupancy scores (see above). Nucleosome dyad occupancy is referred to as “Nucleosome Occupancy”.

**Genes associated with fragile or resistant nucleosomes**

We found fragile nucleosomes enriched at the -1 nucleosome and resistant nucleosomes were enriched at the +1 nucleosome (Fig. 4-4). Given the strong association with the gene start, we identified two sets of genes that contain fragile and resistant nucleosomes, respectively, +/- 500 bp from their transcript start site.

**Gene ontology analysis**

Gene lists were uploaded to the FatiGO web server (babelomics.bioinfo.cipf.es) and compared against the background set of all *C. elegans* genes (Al-Shahrour et al. 2004). P-values were calculated using the Fisher’s exact test, and corrected for multiple testing using the FDR procedure of Benjamini and Hochberg (Benjamini and Hochberg 1995). Corrected p-values and GO terms were then input into REVIGO to reduce and visualize significantly enriched GO clusters (Supek et al. 2011).
Stable and developmentally regulated genes

Pre-normalized transcriptome sequencing data was downloaded from: https://www.encodeproject.org/comparative/transcriptome/ (Gerstein et al. 2014; Spencer et al. 2011). For each gene, we calculated the coefficient of variation (CV): $c_v = \frac{\sigma}{\mu}$. We took the 1000 genes with the highest CVs as the set of developmentally regulated genes, and the set of 1000 genes with the lowest CVs as the set of stably expressed genes.

RNA-seq analysis

Unstranded mRNA libraries were prepared from total RNA for RNA-seq using the Illumina TruSeq RNA Library Preparation Kit v2 (RS-122-2001). RNA-seq reads were mapped to the C. elegans WS220 Gene Annotation Model using Tophat2 (v0.7) (Trapnell et al. 2012). The resulting alignment files were quantified using HT-Seq (v0.4.1) and the RefSeq gene annotations for WS220 (Anders et al. 2015). Total read counts per gene were normalized for differential expression using DESeq2 (v1.0.19) in R (v3.0.1) (Love et al. 2014).
REFERENCES


Locke G, Haberman D, Johnson SM, Morozov AV. 2013. Global remodeling of nucleosome


Chapter 5 -

Chromatin packaging in C. elegans sperm and oocytes

Author's statement:

This chapter reflects ongoing work done in collaboration between myself, Tomoko Tabuchi, PhD, Andres Rechsteiner, PhD, and Susan Strome, PhD. I was responsible for experimental design, oocyte ChIPs, and data analysis. Portions of this work have been presented publicly at the C. elegans International Worm Meeting (2013) and the NIH Symposium on Chromosome Biology (2013).
ABSTRACT

Nucleosome occupancy and higher-order chromosomal organization are highly regulated patterns that indicate the underlying functional identity of the cell. During gametogenesis, massive changes in the chromatin landscape prepare the genome for fertilization and early developmental events. Using isolated populations of him-5 sperm and fem-1;fer-15 oocytes, we assembled nucleosome, histone modification, and transcriptome maps to determine how gamete chromatin is reconfigured to establish a totipotent fate. Comparisons between male and female gametes, and between germ cell and somatic chromatin will identify 1) where nucleosomes are positioned to prepare for early transcriptional events and 2) if the maternal and paternal genomes transfer distinct nucleosome positions to the fertilized zygote. Differences identified between sperm and oocyte nucleosome organization may either be a passive result of transcription during gametogenesis, or may function to “prime” the developing embryo for early zygotic transcription. This work will establish where active (H3K36me3, H3K4ac), and repressive (H3K27me3, H3K9me3) histone posttranslational modifications are deposited in sperm and oocytes, as well as RNA polymerase II and the histone variant H2A.Z. Comparing these modifications to those in the adult gonad and the early embryo will determine if the modifications are programmatic or passive.
INTRODUCTION

Both the male and female gametes carry a chromatin contribution to the zygote. Studies have highlighted the critical importance of proper chromatin organization to maintain cell identity and, in the case of gametes, reproductive fertility. Chromatin remodeling events are highly regulated during gametogenesis. During yeast sporulation, H3K4 demethylation by JHD2 globally represses noncoding transcription. This highly regulated silencing is critical to proper postmeiotic development: JHD2 mutants produce stress-sensitive spores and show aberrant differentiation\(^1\). *C. elegans* are similarly dependent on proper H3K4 methylation levels for successful gametogenesis. Loss of SPR-5, an H3K4me2 demethylase, exhibits a progressive sterility phenotype, specifically misregulating spermatogenesis-expressed genes\(^2\).

In higher eukaryotes, additional remodeling events further silence sperm chromatin. Somatic histones are hyperacetylated and replaced with the transition nuclear proteins TNP1 and TNP2, then subsequently replaced by the protamines PRM1 and PRM2\(^3\). If hyperacetylation of the elongating spermatid is prevented by loss of the histone acetyltransferase CHD5, proper histone-to-protamine transition fails, leaving the mouse infertile\(^4\). In mouse and humans, although the majority of nucleosomes are exchanged for protamines, a number of nucleosomes are retained on mature sperm chromatin. Retained nucleosomes are enriched at developmental loci, a potential mechanism by which paternal epigenetic information is transferred to the embryo\(^5\)\(^-\)\(^7\).

In mammals, two oocyte-specific histone variants have been identified, H1foo and macroH2A. *H1FOO* mRNA is heavily polyadenylated, and the protein is high in lysines, suitable for methylation or acetylation. Loss of H1Foo in growing oocytes severely decreases the capacity for the oocyte to resume meiosis \(^8\)\(^,\)\(^9\). MacroH2A has been found associated with chromatin in the growing oocyte, as well as with the maternal chromatin post fertilization, and may play a role in chromatin condensation\(^10\).
While intriguing, teasing apart the molecular mechanisms behind chromatin remodeling and proper gamete function remains difficult. Isolation of sufficient quantities of mammalian oocytes for genomic studies is not yet possible, nor is genetic perturbation by transgenesis feasible in a rapid fashion. We instead decided to perform experiments in a readily tractable genetic, developmental, and genomic model organism: C. elegans.

In C. elegans, spermatogenesis can occur in either the hermaphrodite or male body (XX or XO). The mature C. elegans spermatozoa lacks a flagellum, and is able to crawl using a pseudopod activated by major sperm proteins. Besides gross morphology, little is known about how chromatin condensation proceeds in C. elegans sperm chromatin\textsuperscript{11}. Although male sperm has traditionally been considered to lack nucleosomes and histone posttranslational modification\textsuperscript{12}, recent evidence indicates that lack of antibody staining by immunofluorescence was a technical error rather than biological fact (S. Strome, unpub.). Whereas sperm undergo a multitude of chromatin remodeling events in preparation for condensation, oocytes primarily remain somatic in their epigenetic decorations. C. elegans oogenesis begins shortly after reaching the young adult stage. Germ cells first mitotically proliferate at the distal tip of the gonad, then progress through the early stages of meiosis prophase I in assembly-line fashion. Oogenesis maintains H3K8/16ac, H3K9/K14ac, H3K4me, and H3K27me modifications from the distal gonad to the maternal pronucleus in the 1-cell embryo\textsuperscript{12}. Homologues of the oocyte-specific mammalian macroH2A or H1Foo have not yet been identified in C. elegans oogenesis\textsuperscript{11}.

We are interested in understanding the epigenetic landscape of C. elegans sperm and oocytes, and how it changes after fertilization to generate a new animal. First, we optimized a gamete isolation procedure in C. elegans mutant worms. We have performed the first genome-wide map of nucleosome organization in C. elegans sperm and oocytes, and have found
significant differences in global structural organization between sperm and oocytes, and between gametes and somatic cells. Subsequent experiments will complete the characterization of the gamete chromatin landscape and identify the mechanistic relationship between gamete histone variant usage and proper gamete chromatin packaging.

RESULTS

Isolating pure populations of *C. elegans* sperm and oocytes

It remains unclear how nucleosomes are positioned in gamete chromatin in order to properly package the genome for fertilization, as well as to maintain proper gene regulation. To investigate, we used two classes of mutants to obtain pure populations of *C. elegans* oocyte and sperm. *Him* mutants increase the percentage of males in the population, while *Fem* mutants show a feminization of the hermaphrodite germline. *Him-5* (e1490V, strain DR466) is a non-disjunction mutant. *him*-5 populations have ~40% true males (rather than 0.1% in wild type N2). *Fem-1* is required for male trait development: mutants do not make sperm, and unfertilized, immature oocytes accumulate in the gonad.

To isolate intact, unovulated oocytes, feminized *fem-1; fer-15* worms can be chopped repeatedly with a razor blade in egg salts buffer. Alternatively, oocytes can be ovulated using serotonin and levamisole treatment, but this procedure produces fewer oocytes with a higher percentage of endomitotic nuclei. For these reasons, we instead chose to use the chopping protocol. Due to the permeable nature of the oocyte membrane, composition of the buffer is critical to maintain the protein and DNA content of the oocyte after isolation. Oocytes were then separated from worm carcasses using a series of filtration steps to remove debris and contaminating cell types (Figure 5-1A). Sperm can be isolated from males using physical pressure or chopping to extrude the only unattached cells from the male body (Figure 5-1B).
Transcriptome of C. elegans gametes

Previous work has used subtractive microarray analysis to identify transcripts enriched in C. elegans oogenic and spermatogenic germline. We chose to perform transcriptome profiling on isolated pools of oocytes and sperm to specifically identify mRNA molecules found inside the gametes themselves. First, we confirmed using qRT-PCR the identity and purity of the isolated gametes (Figure 5-2). Next, we performed three replicates of oocyte RNA-seq and two replicates of sperm RNA-seq. Overall, the replicates were well correlated (Figure 5-3). When we compared our transcriptome data to previous microarray and mass spectrometry data, we found a large set of shared transcripts, as well as a number of differences (Figure 5-4). These differences may reflect biases from the method, or true biological differences between the samples used. Regardless, we have confidence that our transcriptome is valid and useful for further analysis.
Figure 5-2. qRT-PCR was used to analyze identity and purity of isolated gametes. act-1, actin-1; ama-1, large subunit of RNA polII; ges-1, gut-specific esterase; myo-2, pharynx specific muscle myosin; spe-9, sperm-specific transmembrane protein; oma-1, oocyte maturation factor; hil-2/3 & 4/5, histone H1-like protein; htas-1, sperm-specific histone H2A variant.

Figure 5-3. Cross-correlation analysis of oocyte and sperm RNA-seq replicates.
Figure 5-4. Comparison of transcripts identified with different techniques. Transcriptome data (this study) to previous microarray and mass-spectrometry data generated using subtractive analysis (microarray – Renike et al. Mol Cell 2000, Development 2004) or gamete isolation (mass-spec – Chu et al. Nature 2006).

We used differential expression analysis to identify transcripts specifically enriched in sperm or oocytes. DESeq2 identified a set of ~1200 transcripts enriched in in oocytes, and ~700 enriched in sperm. We confirmed that previously identified sperm-specific genes like sss-2, htas-1, msp-10, spe-9, were enriched in our sperm samples. Similarly, we confirmed that known oocyte-enriched genes were found in our oocyte samples: pbrm-1, mex-3, oma-1/2, cav-1 (Figure 5-5).

Figure 5-5. Differential expression analysis of genes enriched in sperm and oocytes. MA-plot of average sperm vs average oocyte transcriptome abundance. In total, 759 transcripts were found significantly enriched versus 1235 significantly enriched transcripts in oocyte.
We next asked what classes of genes tended to be enriched in oocyte and sperm. To do this, we used FatiGO (Al-Shahrour et al., NAR 2007) to search for over-represented gene ontology terms in the list of genes identified in either sperm or oocytes. We next used ReviGO (Supek et al., PLOS ONE 2011) to reduce GO term redundancy and visualize relationships between them (Figure 5-6). We found oocytes enriched in GO terms like cell fate specification, post-embryonic development, chromatin organization, and mitotic cell cycle progression. In contrast, sperm samples were enriched in GO terms like actin filament-based process, spermatogenesis, and phosphorylation. We suspect that these GO terms reflect the large difference in mRNA contribution made by the oocyte and sperm to the developing embryo: oocytes are highly packaged with genes necessary for early embryo development, while sperm mostly contain mRNA products for their own activation and motility.

**Figure 5-6. Gene Ontology Analysis of transcripts enriched in C. elegans oocytes and sperm.**
Genes were submitted to FatiGO analysis to identify over-represented biological processes. GO term redundancy was reduced using REVIGO.

**Nucleosome organization in C. elegans sperm and oocytes**

The mature sperm and oocyte cells represent a unique opportunity to investigate nucleosome positioning *in vivo* in a naturally transcriptionally quiescent system. Using the method described in Figure 5-1, we performed three replicate experiments to identify how nucleosomes are positioned in sperm and oocytes. Preliminary analysis found significant
differences between sperm and oocytes, and between gametes and somatic cells (Figure 5-7A). Where somatic cells traditionally show more well-positioned nucleosome arrays at highly transcribed genes, neither sperm nor oocytes show this transcription-dependent packaging (compare Figure 5-7B top to bottom with 5-7C&D). In addition, sperm nucleosomes show very little evidence of nucleosome array formation, as evidenced by a lack of the canonical striped pattern when genes are aligned at their transcriptional start sites (compare Figure 5-7B&C to 5-7D). When averaged together, this pattern is maintained (Figure 5-7E). One possible explanation is the relatively small, yet widely distributed size of the mononucleosomal fragments recovered in this experiment (140 bp on average, Figure 5-7F). Given that the nucleosome size still ‘peaks’ at the canonical 150bp, and given the immunohistochemistry histone post-translational modification staining observed in mature sperm, we presume that sperm chromatin do truly contain nucleosomes. We suspect that the smaller fragment sizes are due to the tightly packaged nature of *C. elegans* sperm chromatin, which makes mononucleosomes difficult to recover (and thus smaller fragments are preferentially represented). Alternatively, it has been shown that *C. elegans* contain protamine homologues, which may wrap significantly smaller DNA lengths. Future work may clarify the exact nature of chromatin packaging in *C. elegans* sperm.
C. *elegans* gamete nucleosome organization is distinct from mixed stage embryo chromatin. (A) Spearman correlation calculated in 100 bp windows between nucleosome positioning experiments (Ercan et al. 2011; Steiner et al. 2012; Kaplan et al. 2009) (B) Nucleosome dyad positions around the transcriptional start site reflect transcription-dependent packaging and phasing emanating from the +1 nucleosome. Each row represents a different transcription start site ordered by decreasing gene expression (Chen et al. 2013). Each row is a position relative to the +1 nucleosome. (C) Oocyte nucleosomes remain highly phased with respect to the +1 nucleosome. (D) Sperm chromatin shows neither significant nucleosome phasing nor transcription-dependent packaging. (E) A global average over all ~8000 transcripts for each cell type is shown. Nucleosome “peaks” reoccur every ~147 bp with 20 bp
spacing, except in the sperm sample. (F) Loss of nucleosome phasing in the sperm sample could be due to a relatively small area protected in the mononucleosomal fragment.

Preliminary analysis identified promoter clusters that self-segregate using k-means clustering (Figure 5-8A, sperm clusters not shown). These promoter clusters are gamete-specific (Figure 5-8B), and have different tendencies towards being associated with the histone H2A variant HTZ-1 post fertilization (Figure 5-8C). To assign a functional significance to these promoter categories, we next asked if the genes from each cluster had different transcriptional behavior in the early embryo (Figure 5-8D), or were enriched in a particular gene ontology term (Figure 5-8E). By interrogating the use of additional histone modifications and histone variants in the gametes, we hope to address how the chromatin landscape at fertilization is uniquely poised to initiate precise developmental events. For example, we found that those genes most highly transcribed in the early embryo have the highest nucleosome density in the oocyte (Figure 5-8F). This trend is much less pronounced in sperm chromatin (Figure 5-8G), indicating that oocyte chromatin may be the template upon which early transcriptional events are based.
Figure 5-8 Addressing functional significance of gamete chromatin organization in the early embryo. (A) K-means clustering (k=6) on the promoter nucleosomes was used (sperm clusters not shown). (B) Embryo nucleosomes were sorted based on the promoter clusters identified in (A). The difference in promoter organization in B means that the promoter clusters are not sequence-dependent. (C) Promoter organization in the gamete is predictive of H2A.Z incorporation in the embryo. HTZ-1 ChIP-seq at genes sorted in the same order as (A). (D) Early embryo expression for genes in each promoter cluster from (A). (E) Top enriched GO category from each promoter cluster identified in (A). (F) Oocyte and (G) sperm nucleosome occupancy is linearly correlated with embryo expression. Genes expressed in the early embryo were binned into the top (1st quartile), mid-top, mid-low, and lowest expressed genes (4th quartile). Average nucleosome occupancy around the +1 nucleosome in the oocyte or sperm is plotted.
Characterization of histone isoform usage in mature sperm and oocytes

Eukaryotic chromatin contains a fifth “linker” histone. Known as either histone H1 or H5, the linker histone does not bind the nucleosome core particle itself, but rather binds the linker DNA emanating from the nucleosome entry and exit site (Figure 5-9A). The linker histone has been implicated in higher-order chromatin condensation, gene silencing, and regulation of histone H1 can play a major role in cellular events. For example, H1 is heavily phosphorylated by CDK1 during mitosis to increase H1 affinity for the mitotic chromosome. Increased affinity aids in mitotic chromosome compaction and is required for M-phase completion\(^29\). Humans have at least 11 different linker histone isoforms, while nine H1 orthologues or H1-like proteins have been identified in *C. elegans*. Evidence exists for H1 as the molecular ancestor to protamines\(^30\). Gamete enriched histone variants have been identified in both mammalian oocytes (H1foo) and sperm (HILS1), and are critical for meiotic maturation and chromatin remodeling\(^9,31\). The role and regulation of H1 isoforms besides the major variant in *C. elegans* (HIS-24) is unknown\(^32,33\). The second histone variant of interest is HTAS-1, a sperm-specific H2A variant specifically incorporated in the sperm chromatin at the condensation zone. HTAS-1 varies from the canonical H2A at the C-terminus, which is also the histone H1 docking domain\(^34\). Reconstituted HTAS-1 nucleosomes are incredibly stable (Chu lab, unpublished). Where, and what role HTAS-1 plays during spermatogenesis is unknown.

A recent proteomics experiment identified chromatin-associated proteins in *C. elegans* sperm and oocyte\(^54\). Two factors of interest include the histone-like proteins HIL-2/3 and HIL-4/5, and a sperm specific H2A, named HTAS-1 (Figure 5-9C). Preliminary qRT-PCR studies in isolated oocyte and sperm indicated differential abundance of the histone isoforms HTAS-1, hil-2/3, and hil-4/5 (Figure 5-2). RNA-seq studies confirmed differential enrichment (Figure 5-9D). While gamete specificity has not yet been shown, reports indicate limited expression in other cell types\(^55\). Previous knockdown experiments using RNAi have identified phenotypes that hint
at HIL-2/3/4/5 function in gametogenesis, including abnormal gonad morphology and meiotic defects. Future work investigating how these histones are utilized will clarify how gamete chromatin affects the process of fertilization and early developmental events.

**Figure 5-9. Gamete-enriched histone variants.** (A) Histone H1 seals the nucleosome entry and exit site in a non-sequence specific fashion. (B) All histone H1-like proteins have high identity to the ubiquitous C. elegans H1 isoform, his-24. (C) Previous proteomics experiment identified proteins associated with sperm and oocyte chromatin. The top-enriched histone variants of interest are plotted (Chu et al. 2006). (D) RNA-seq enrichment of mRNAs in isolated sperm and oocytes.

**DISCUSSION**

In this chapter I investigated the transcriptome and chromatin landscape of *C. elegans* sperm and oocytes by isolating pure populations of gametes. The ability to isolate large numbers of pure male and female gametes is a challenge in many organisms, like Drosophila and mouse. The ability to do so in *C. elegans*, and their ready use for chromatin and next-generation sequencing assays will serve useful in future studies investigating the nature of the chromatin inheritance from one generation to the next.

We performed transcriptome analysis using RNA-seq and nucleosome mapping via MNase-seq. We found ~700 transcripts enriched in sperm, ~1200 enriched in oocytes. Sperm were enriched with transcripts involved in phosphorylation pathways, while oocytes were enriched for genes related to GO terms like cell-fate specification and gene regulation. These
data confirm and support previous microarray and mass spectrometry-based analyses of the gamete transcriptomes in *C. elegans*.

In many higher eukaryotes, histone proteins are removed during spermatogenesis, and the chromatin is instead packaged around highly basic proteins called protamines. As such, comparative nucleosome positioning data in male and female gametes of the same organism has never before been performed. We were interested to understand (1) whether the maternal and paternal genomes contain similar or different nucleosome positions, and (2) if there are differences, how they are resolved after fertilization. We found that although oocytes and sperm contain well positioned +1 nucleosomes, sperm nuclei showed poor nucleosome phasing. This suggests that paternal and maternal genomes contain different nucleosome positions, potentially due to their different size constraints: sperm are approximately $1/100^{th}$ the size of an oocyte.

Given the differences in nucleosome positioning and occupancy between the gametes, we were curious to know how gamete chromatin compared to chromatin observed in the early embryo. We found that gamete nucleosome occupancy was correlated to embryonic gene expression, suggesting nucleosomes are retained at genes important for early zygotic development. This result has been observed previously in humans and mouse, suggesting a conserved mechanism to activate early developmental processes.

**MATERIALS AND METHODS**

**N2 worm liquid culture and mixed stage embryo harvest**

Wild type (N2, Bristol strain) worms were obtained from the *Caenorhabditis* Genetic Center. To obtain a large number of highly synchronized worms, adult gravid worms are first bleached in a basic hypochlorite solution. Embryos are left to hatch overnight in M9 buffer without food, resulting in an L1 arrest. Larval stage L1 worms are collected through a 20 um
nytex filter and put into liquid culture for massive, synchronized growth. Worms are grown in liquid culture supplemented with concentrated HB101 E. coli food at 20degC for ~65 hours with constant shaking at approximately 200 rpm. At this stage, the worms have completed their life cycle and are again gravid. Gravid adults are collected on 35 um nytex filter and washed into a 50 mL conical tube. Mixed-stage embryos are collected by bleaching, and either crosslinked in 2% formaldehyde for 30 minutes or immediately flash frozen in liquid nitrogen.

**MNase digestion**

Frozen, crosslinked or native embryos were thawed on ice and used as input for MNase digestion. The tough eggshell was removed with chitinase, nuclei were separated by douncing, and MNase enzyme was added to the nuclei at 25degC. 250 uL aliquots were removed, quenched with 0.5 M EDTA, and the DNA was prepared as described in Chapter 4.

**Paired-end sequencing analysis**

Paired-end sequencing relies on ligation of the Illumina adapter molecule to each end of the mononucleosomal (~130 to 200 bp in length) fragment. 50bp reads sequence each end of the fragment, which allows direct calculation of 1) the center of the nucleosome (nucleosome dyad) and 2) the length of each individual fragment (nucleosome size). After sequencing, paired-end reads are mapped to the C. elegans genome (WS220, UCSC build ce10) using bowtie in paired-end alignment mode, while filtering for PCR duplicates and reads in non-uniquely mappable regions. Mapped reads are scaled by 1/mean number of reads covering each basepair, resulting in scaled coverage. Alternatively, the unique position of each nucleosome dyad is agglomerated and gaussian smoothed to approximate the nucleosome dyad density at each position.
Isolating pure populations of C. elegans sperm and oocytes

We used two classes of mutants to obtain pure populations of C. elegans oocytes and sperm. *Him* mutants increase the percentage of males in the population, while *Fem* mutants show a feminization of the hermaphrodite germline. *Him*-5 (e1490V, strain DR466) is a non-disjunction mutant. *him*-5 populations have ~40% true males (rather than 0.1% in wild type N2). *Fem-1* is required for male trait development: mutants do not make sperm, and unfertilized, immature oocytes accumulate in the gonad.

Previous methods have described utilizing pressure to release sperm from the male worm body. This method relies on the fact that >50% of cells, and the only unattached cell in the male body, are gametes. Briefly, *him*-5 worms are grown in liquid culture. Males are separated from hermaphrodites at the young adult stage by filtration through a 35 um filter. Males are vigorous swimmers and have smaller bodies, and are able to swim through the filter while hermaphrodites cannot. Males are collected at the bottom of a conical tube by gravity. Approximately 200 uL of males are pipetted onto a watch glass. Two 25 5/8-gauge precision-glide needles are used to slice through the center of worms en masse. Carcasses and media are removed by filtering the sperm through a 10 um nytex filter, pelleted, and flash frozen in liquid nitrogen. In a similar fashion, worms with a *fem-1* temperature sensitive allele are grown in liquid culture at the restrictive temperature until the young adult stage. Adults are collected on a 35 um filter and transferred to a 50 mL conical tube in K buffer with serotonin and levamisole to induce oocyte laying. Every 30 minutes worms are collected in a 15mL syringe and forced through a 35 uM filter. Oocytes laid in the media can pass through, while adults cannot. Oocytes are collected by centrifugation and flash frozen in liquid nitrogen. This process is repeated ~3x or until no additional oocytes are liberated from the female body. Later iterations of this technique instead relied on physical disruption of the female body by repeatedly chopping with a razor blade.
RNA isolation and transcriptome analysis

After isolation, whole embryos or gametes were dropped directly into Trizol buffer and flash-frozen in liquid nitrogen. Cells were lysed through a series of freeze-thaw and vortex steps. To confirm RNA extraction and cell identity, first strand cDNA was made and used as input to qRT-PCR. Data were normalized to act-1, an arbitrary choice, as it was unclear if this transcript was expected to be stored in either cell type. Two negative controls, ges-1 and myo-2, a gut-specific esterase and a pharynx specific muscle myosin, respectively, suggested that the cell preparations were relatively free of contaminating cell types. Transcripts for spe-9, a sperm-specific transmembrane protein, were found highly enriched in the sperm sample, while oma-1, an oocyte maturation factor was significantly oocyte enriched.

Subsequently, polyadenylated mature mRNAs were extracted by hybridization to poly-dT beads. Polyadenylated mRNAs were fragmented, reverse transcribed into a cDNA library, and used as input for traditional Illumina HiSeq high throughput sequencing\textsuperscript{52}. RNA-seq data processing was performed using DESeq\textsuperscript{53}. Three biological replicates of oocyte RNA and two replicates of sperm RNA have been performed, with highly concordant results (cross-replicate correlation coefficients $> 0.92$).

Chromatin immunoprecipitation

Chromatin is prepared in a similar fashion from isolated gametes as described above. Gametes or embryos are fixed in formaldehyde, then the cell is lysed and chromatin is sheared to an appropriately sized fragment. Antibodies recognizing the histone post-translational modification or DNA-bound protein of interest are incubated with the sheared chromatin, and the DNA-antibody complex is isolated with Dynabead magnetic beads. ChIP-seq has been performed by our group for many chromatin-associated factors and histone post-translational
modifications at many different stages of the *C. elegans* lifecycle\textsuperscript{22}. The chromatin landscape is dynamic and complex, and still very little is known about what modifications are seeded on the genome in preparation for fertilization. I will utilize ChIP-seq to identify where HTZ-1, H3K27me3, H3K36me3, H3K4me3, and H3K4ac histone post-translational modifications are deposited in sperm and oocytes, as well as paused RNA polymerase II.
REFERENCES


20. Xi, Y., Yao, J., Chen, R., Li, W. & He, X. Nucleosome fragility reveals novel functional states
27. Zhang, Y. *et al.* Canonical nucleosome organization at promoters forms during genome activation. *Genome Research* (2013). doi:10.1101/gr.157750.113
Chapter 6 -

How histone levels control developmental events,

concluding remarks, and future outlook

Author's Statement:
Portions of this chapter were written in collaboration with Dr. Amanda Amodeo, with the intent to publish a review discussing the role of histones in controlling developmental events.
ABSTRACT

Broadly speaking, the work in this thesis has centered on firstly, creating informative datasets to further the usefulness of *C. elegans* as a model organism in the study of chromatin, epigenetics, and transcriptional regulation. Secondly, this work has investigated the dynamic nature of the histone-DNA interaction, and has posited a functional role for nucleosome instability in activating genes in signal-response pathways. This has led to a broader understanding and perspective on the role of histone levels in timing developmental events, like cell cycle, zygotic genome activation, and aging.
PERSPECTIVE

Chromatin changes during differentiation

At a gross morphological level, researchers have observed a change in chromatin configuration during the transition from maternal to zygotic control, with a general change in chromatin from “open” to “closed” (Yuzyuk et al. 2009; Ner and Travers 1994; Meshorer and Misteli 2006). Macrostructural observations made using microscopy describe pluripotent cells as containing relatively 'open' chromatin when compared to images of differentiated cell lineages. This open chromatin state is thought to contribute to the maintenance of pluripotency, and that differentiation progressively restricts accessibility of regulatory regions, in effect "locking in" the epigenetic landscape (Meshorer and Misteli 2006).

Correlated with this morphological change in chromatin compaction is a parallel shift in abundance of linker histones and other proteins that contribute to the overall degree of chromatin condensation. Ner and Travers identified that histone H1 is not present until transcriptional activation, while HMG-D protein levels are high in the rapidly cycling, transcriptionally quiescent embryo. They propose that the regulation of linker histone levels may play an important role in the coordinated shift from decompacted to condensed chromatin. Moreover, in alignment with previous observations made in Xenopus (Dimitrov et al. 1993), they suggest that ratio of H1/HMG-D proteins is a critical determinant of transcriptional competence.

The macromolecular definition of “open” chromatin is distinct from the nucleosome-level description of open chromatin as measured by DNase hypersensitivity, FAIRE, or ATAC-seq (Gaspar-Maia et al. 2011). In addition to histological observations, more recent studies have begun to investigate the fine-grained changes in nucleosome organization that occur during zygotic genome activation. In agreement with the macromolecular data, digestion of chromatin with DNase I and micrococcal nuclease (MNase) also indicates that differentiated chromatin is less accessible than that of pluripotent cell types (Schaniel et al. 2009). Studies of mammalian
ESC differentiation have found that nucleosome repeat length increases during differentiation by 5-7 bp (Teif et al. 2012). Of note, specific changes in nucleosome occupancy and positioning are limited, but are often co-localized with the binding sites of developmental regulators like Klf4, Oct4/Sox2 and c-Myc. ESC enhancers increase in nucleosome occupancy in the somatic cells, indicating significant remodeling of the regulatory landscape after differentiation (West et al. 2014). Together, these data may suggest that while individual nucleosomes undergo small local shifts, higher order chromatin compaction is dramatically remodeled during the transition from pluripotency to committed cell fate.

Similar to mammalian cells, the early nuclear divisions of zebrafish, C. elegans, and Drosophila embryogenesis also show limited heterochromatin formation, and nuclei staining indicate a more 'open', decompacted chromatin state. Thus, these model organisms serve as excellent systems to investigate how genome activation (or the transcription of lineage specific genes) can occur in the context of progressively closing chromatin. Here, we will discuss four mechanisms by which early zygotic genes are activated: the action of pioneer TFs, bookmarking with Zelda, higher order chromatin structures, and fragile nucleosomes.

**Pioneer TFs remodel and activate the quiescent genome**

Maternally provided pioneer transcription factors are thought to bind to chromatin in a sequence-specific fashion by out-competing a nucleosome or by recruiting additional chromatin remodelers to induce nucleosome loss. Nanog, Sox2, and Oct4 are some of the best-characterized pioneer TFs (Zaret and Carroll 2011), and have been shown to bind DNA motifs while they are still nucleosome-bound (Soufi et al. 2015). Although originally characterized for their role during stem cell reprogramming (Takahashi and Yamanaka 2006), it's presumed that the OSN TFs behave similarly during genome activation, and thus contribute to remodeling and activating the quiescent genome.
Intrinsically unstable nucleosomes may play a role in ZGA

Intrinsically unstable nucleosomes have been proposed to work in combination with pioneer transcription factors to regulate the expression of lineage-specific genes. By using different concentrations of the enzyme micrococcal nuclease (MNase), researchers are able to characterize "fragile" or unstable nucleosomes that are typically lost from canonical nucleosome experiments (Xi et al. 2011). Iwafuchi Doi and colleagues observed that unstable nucleosomes in mouse liver tend to occur at tissue-specific promoters, and proposed that the pioneer transcription factor FoxA2 may evict the linker histone H1, destabilizing the nucleosome and preparing tissue-specific genes for lineage-specific induction (Iwafuchi-Doi et al. 2016). A similar mechanism may exist to control early zygotic events. Unstable nucleosomes are found at TATA-box containing promoters (Xi et al. 2011) Jeffers & Lieb, submitted), which are also over-enriched in the promoters of early zygotic genes in Drosophila (Lee et al. 2014). Unstable nucleosomes that treadmill on and off the DNA may be the most sensitive to titration, and may thereby represent a transcription factor-independent mechanism to activate transcription.

Zelda bookmarks chromatin for genome activation

In addition to the pioneering activity of certain transcription factors, other DNA-binding proteins have been proposed to serve a role in genome activation. In Drosophila, the genome activator Vielfältig (Vfl), also known as Zelda (Zld), has been proposed to 'bookmark' promoters and enhancers of early developmental genes. Indeed, without the presence of Zld, binding of the early patterning transcription factor Dorsal (Dl) decreases (Sun et al. 2015). Together with GAGA factor, Zelda has the capacity to open chromatin and remodel the cis regulatory landscape of the early Drosophila embryo (Sun et al. 2015; Schulz et al. 2015). The mechanism by which Zelda performs these roles remains unclear, whether it itself is a chromatin remodeler
or modifying enzyme, or whether it recruits additional factors that performs these activities. Chromatin 'bookmarks' similar to Zelda have not yet been identified in other organisms.

**Canonical nucleosome organization**

In addition to nucleosome depletion at enhancers and promoters, a 'canonical' organization of regularly phased nucleosomes has been observed to form at promoters during genome activation. Intriguingly, the formation of a canonical nucleosome organization is independent of RNA pol II binding and transcriptional activity in general (Zhang et al. 2014; Weicksel et al. 2013). Together, these data suggest that chromatin remodelers, chromatin modifying enzymes, or other trans-acting factors are responsible for local and subtle changes that occur concurrent with transcriptional activation of the zygotic genome.

**Alternative mechanisms proposed to control the timing of ZGA**

A “zygotic clock” has been proposed as one possible mechanism by which the timing of zygotic genome activation is determined (Howe et al. 1995), which could involve the progressive gain of activity of a maternally provided factor, either through the accumulation of an active mark or through the progressive loss of a repressive mark. The earliest stages of ZGA in the mouse occurs ~24 hours after fertilization, regardless of the stage of the cell cycle. Factors proposed as zygotic time sensors include TBP and other general transcription factors, because RNA polymerase I, II, and III-dependent promoters all show the same time-dependent sensitivity. Maternally provided mRNAs are progressively poly-adenylated in the early embryo (Aanes et al. 2011; Harvey et al. 2013; Collart et al. 2014; Richter and Lasko 2011), and thus translational efficiency steadily increases after fertilization (Qin et al. 2007; Lee et al. 2014). Thus, protein accumulation from maternally provided mRNAs encoding factors required for transcriptional competence could act as a switch to regulate ZGA.
The length of the cell cycle has also been proposed to control zygotic genome activation, based primarily on the observation that genome activation is concomitant with cell cycle lengthening. In addition, the first zygotically transcribed genes tend to be short and intronless, such that RNA polymerase II (RNA pol II) would have sufficient time for initiation, elongation, and processing during the rapid cell cycles of the early embryo. However, alteration of the cell cycle is not sufficient to initiate early genome activation. Studies in zebrafish have shown that early activity of Chk1 in pre-MBT embryos can lead to longer cell cycles, but this cell cycle elongation does not lead to precocious genome activation. Rather, genome activation slows in parallel with the N:C ratio (Zhang et al. 2015).

**Histone levels affect transcription in general**

Histone protein levels are tightly regulated to prevent DNA damage and to ensure proper transcription and DNA replication. Excess histones in *S. cerevisiae* are degraded in a Rad53 dependent manner, and an undersupply of histones is rapidly sensed to upregulate histone mRNA transcription and translation. Cells are sensitive to total histone dosage, as well as to the stoichiometric balance between the four core histones H2A, H2B, H3, and H4.

Defects in nucleosome assembly behind the replication fork due to histone insufficient histone dosage leads to DNA double-strand breaks and S phase arrest (Ye et al. 2003). During transcription, proper histone levels are important to direct RNA polymerase machinery to the transcription start site, for preventing cryptic transcription (Lickwar et al. 2009; Li et al. 2007), and for exon recognition (Tilgner et al. 2009). *In vivo* depletion of histones in yeast leads to cell cycle arrest (Kim et al. 1988), de-repression of conditionally expressed genes, and large scale changes in the transcriptome (Wyrick et al. 1999). *In vitro*, increasing the histone:DNA ratio can cause aberrant chromatin segregation (Meeks-Wagner and Hartwell 1986), chromatin aggregation, and prevent transcription (Steger and Workman 1999). It remains unclear how
these deleterious phenotypes are avoided in the early embryo where histones are in high excess over DNA.

In addition to proper overall histone levels, the cell is also sensitive to loss of stoichiometry between H2A, H2B, H3, and H4. Overexpression of histone pairs (H2A and H2B or H3 and H4, rather than all four core histones together) can cause the same transcription-related phenotypes as loss-of-function mutations in those genes (Clark-Adams et al. 1988; Clark-Adams and Winston 1986; Malone et al. 1991; Swanson et al. 1991).

**Competition between nucleosomes and TFs regulates gene activation**

Nucleosomes play activating and repressive roles in transcription. Nucleosomes must be evicted from enhancers and upstream regulatory regions to create sufficient space for the transcriptional pre-initiation complex to form. Histone levels have long been understood to play a critical role in determining expression timing during the MBT. Most strikingly, Almouzni and Wolffe (Almouzni and Wolffe 1995) found that the action of the pre-ZGA "transcriptional inhibitor" depends on the association of core histones with the DNA. What’s more, addition of exogenous TBP overcame transcriptional repression, suggesting that increasing competition between TBP and a DNA binding inhibitor molecule, like core histones, can serve to overcome transcriptional repression.

Because some sets of genes contain promoters that are more unstable, these are also genes that putatively will be most sensitive to titration of histone levels in the embryo. Indeed, transcription from class III genes is most sensitive to titration of histone levels by adding exogenous nonspecific DNA. Class III genes have been proposed to have inherently destabilized nucleosomes. In addition, certain class II genes enriched with TATA elements and other nucleosome destabilizing motifs are also sensitive to histone titration (Prioleau et al. 1994;
Workman et al. 1988; Stringer et al. 1990). It is possible that DNA-encoded promoter elements may act as a TF-independent mechanism for early genome activation.

**Histone levels are regulated with age**

Other important developmental events and cellular states are influenced by histone levels. For example, old yeast cells show a significant loss of histone protein levels and histone incorporation into chromatin, primarily through decreased histone protein synthesis (Feser et al. 2010). There was a significant overlap in nucleosome occupancy and gene expression changes between old yeast and yeast where histone H3 was artificially shutoff, suggesting that loss of histones may drive the yeast aging phenotypes (Hu et al. 2014a; Gossett and Lieb 2012). This observation is also true in human fibroblasts. ASF1 and SLBP levels are altered in late-passage human fibroblasts, indicating that altered histone incorporation on the DNA is related to aging (O'Sullivan et al. 2010; 2014). Similar loss of histones is observed in vivo and in vitro in senescent human cells (Ivanov et al. 2013). DNA damage has been proposed as a driving force for histone loss.

Is histone loss a causal factor of functional decline observed with age, or a consequence of age-related chromatin re-organization? In yeast, nucleosome occupancy decreased by 50% across the whole genome during replicative aging. Nucleosomes become less well positioned, and move towards sequence-predicted regions. Loss of histones induces transcription of all genes. Genes normally repressed by promoter nucleosomes were the most strongly enriched. They found elevated levels of DNA double strand breaks, mitochondrial DNA transfer to the nuclear genome, and large scale chromosomal alterations. There is some evidence that histone loss is causal: Increasing histone levels is sufficient to extend lifespan in yeast (Hu et al. 2014b).
CONCLUSIONS AND FUTURE OUTLOOK

*C. elegans* has lagged behind other organisms as a model for transcriptional regulation for a number of reasons. First, trans-splicing has obscured the true position of transcriptional start sites, leaving it previously unclear whether *C. elegans* uses the same mode of transcriptional regulation as other higher eukaryotes. In Chapter 2, we annotated the transcriptional start sites of *C. elegans*, and showed that nucleosome positioning, chromatin modifications, and other transcriptional mechanisms are truly conserved in worms. These annotations will significantly benefit the future study of transcriptional regulation in *C. elegans*.

Comparative data sets are the boon of any research program. Discoveries happen faster and progress is made more quickly when high quality data is publicly available for the entire community to use. Large numbers of datasets detailing the yeast transcriptome and epigenome make it easy to compare new experiments against those already performed. In contrast, before the year 2010 and the advent of the modENCODE consortium, few *C. elegans* datasets existed (5979 total *C. elegans* datasets on Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo), only 747 of which are not modENCODE affiliated). In Chapter 3, we generated over 650 datasets describing the transcription factor, transcriptome, and chromatin landscape from over 7 different stages from the *C. elegans* lifecycle. These datasets have already been used and cited by thousands of researchers worldwide to further investigate the nature of chromatin organization in *C. elegans*.

One final challenge has historically limited the utility of *C. elegans* as a model: difficulty in isolating cells or tissues for investigating tissue-specific behavior. Now, with dissociation and tissue-specific sorting techniques pioneered by the Murphy lab, even this challenge has been overcome. In fact, *C. elegans* offers an opportunity to collect and profile cells that are rare in other organisms, like the oocytes we used in Chapter 5. *C. elegans* has taught us much about
genetics and development. In the coming years, the powerful genomics of the worm will come full force to advance our mechanistic understanding of human health and disease.
REFERENCES


