Discovery of active *cis*-regulatory elements and transcription factor footprints in nematodes using functional genomics approaches

Thesis by

Margaret Ching Wai Ho

In Partial Fulfillment of the Requirements for

the Degree of

Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2015

(Defended May 27, 2015)

ACKNOWLEDGEMENTS

I am grateful to many people who have helped me along my way at Caltech and who believed in me. Working towards a PhD is like running a very long distance, and not many P.I.s are as good at advising graduate students through the process as my thesis advisor Paul Sternberg. I am extremely grateful for his mentorship and support through the entire process and value both his genuinely kind personality and his keen insight into the process of doing scientific research.

I would also like to thank my thesis committee members Ellen Rothenberg, Barbara Wold, and Alexei Aravin for providing important scientific criticism, good ideas, and deep knowledge of the field. Ellen for being wonderful to me starting from the minute I met her during graduate student interviews, and also demonstrating enthusiasm, tough scientific rigor and excellent classroom teaching. Barbara Wold for providing very insightful input as an expert in the fields of *cis*-regulation and functional regulatory genomics. Alexei Aravin for being the head of my committee and asking great questions that helped me think about my experimental methods and analysis.

Special thanks also goes to Ali Mortazavi who I remember took the time to get me started running analysis scripts on command line and has given me excellent scientific and life advice. Without his encouragement I don’t think I

would have had the confidence to pursue this research project and perform the analysis myself.

I would also like to thank many current and former members of the Sternberg lab especially Mihoko Kato, Andrea Choe, Hillel Schwartz, David Angeles, Pei Shih, Ping Hsueh, Adler Dillman, Erich Schwarz, Steven Kuntz, Oren Schaedel, Han Wang, Ryoji Shinya, James Lee, Ravi Nath, Jonathan Liu, Alli Akagi, John DeModena and Chris Cronin. I would also like to thank the people who kept the lab running smoothly, Mary Alvarez, Sarah Torres, Gladys Medina, Barbara Perry and Shahla Gharib. Everyone has been very friendly, supportive and gave me helpful input. I am grateful that all the postdocs remembered what it was like to be a graduate student and gave me advice on doing science, communicating research, and applying and interviewing for a postdoc. I would also like to thank Marissa Macchietto from the Mortazavi lab at UC Irvine for sharing with me her research and the joy of working with *Steinernema*. Adler Dillman and Steven Kuntz were both very talented graduate students whose previous work helped inform and guide my studies and who continued to be very helpful even after they graduated. Our lab is lucky to work closely with Wormbase. I would especially like to thank James Done, Chris Grove and Xiaodong Wang who helped me a lot. Their invaluable technical help made tough analysis easier to do.

I would like to thank my loving family: Mom and Dad, my sister Joyce and my brother Albert. Also, my whip-smart cousins Libby and Lise Ho who infect me with their enthusiasm for science and who I am sure will have excellent careers in STEM.

Several very close friends who have continually kept my spirits up are Maddalena Jackson, Paul Anzel, Lisa Van Etten, and James Maloney. Last but definitely not least I'd like to thank Jeffrey Smith, who I am especially lucky to have in my life.

ABSTRACT

High throughput DNA sequencing has emerged as a versatile and inexpensive readout of functional activity in biological organisms. In this study I describe the implementation of DNaseI hypersensitivity assays using deep sequencing (DNase-seq) to systematically identify *Caenorhabditis elegans cis*-regulatory modules (CRMs) in embryonic and L1 arrest larval life stages in an unbiased and *de novo* manner. We validated our data by comparison to many known enhancers of *lin-39/ceh-13* Hox complex and of *hlh-1*, *myo-2*, *myo-3*, *lin-26,* and other important developmental genes and are also able to predict novel *cis*-regulatory modules. We predict novel regulatory motifs from our DNase-seq data and predict potential regulatory functions using gene ontology and anatomy enrichment analysis. In addition, our data are high-resolution enough to identify binding sites of transcription factors in the genome. Our data provide support for many distal CRMs in *C. elegans* and for a significant portion of genes possessing multiple CRMs. DNase-seq data can also be used to refine prediction of tissue-specific genes such as those regulated by *C. elegans* pan-neuronal N1 and intestinal ELT-2 DNA motifs. Overall, we identify 24,128 putative CRMS containing over 55,000 footprints. In L1 arrest, we identify 15,841 putative CRMs in the L1 arrest larvae containing 32,000 TF footprints. From comparison of these datasets, we identify an additional 1,854 noncoding DHS that appear to be specific to the L1 arrest larvae condition. These genes include downstream targets of signaling pathways known to be regulated during L1 arrest such as insulin-like signaling via DAF-16/FOXO and Forkhead box transcription factor PHA-4/FOXA that impacts starvation survival in the L1 arrest condition. Having established the first proof-of-principle DNase-seq in nematodes using *C. elegans*, I am applying DNase-seq to a distantly related entomopathogenic nematode, *Steinernema carpocapsae*, with a recently sequenced genome and transcriptome. Finally, I am using a massively parallel reporter assay to test the functional activity of the CRMs we have discovered from DNase-seq using two reporter designs based on MPRA and STARR-seq and by performing DNA and RNA sequencing on transgenic *C. elegans*.

TABLE OF CONTENTS

NOMENCLATURE

Transcription Factor (TF). Proteins with a DNA binding domain that binds to specific sequences and can regulate target gene expression through activation or repression.

***Cis*-regulatory module (CRM)**. Genomic DNA sequence that contains binding sites for transcription factors and that regulates transcription of target genes on the same chromosome.

Enhancer. Orientation-independent CRM that can act at a distance to upregulate target gene expression.

DNaseI. Nuclease that cuts DNA preferentially in nucleosome-free regions and with relatively low sequence specificity

DNase-seq. Experimental technique that measures cleavage patterns in chromatin by DNaseI using high throughput sequencing to discover CRMs and TF binding sites.

DNase Hypersensitive Site (DHS). Genomic DNA sequence (roughly several hundred base pairs in length) that has been found to exhibit significantly increased DNaseI cleavage.

Noncoding DHS. DHS that have been annotated in non-coding regions of the genome and represent putative *cis*-regulatory modules (CRMs).

TF Footprint. In the context of DNase-seq, stretches of genomic DNA sequences between 6-40bp within noncoding DHS that show significantly lower read coverage and strand-shift in mapped reads and represent putative binding sites for TFs.

ChIP-seq. Experimental technique that detects binding sites for TFs using protein-DNA crosslinking, chromatin immunoprecipitation using antibodies against TFs of interest, and high throughput sequencing.

ATAC-seq. Experimental technique that uses Tn5 transposase integration of sequencing primers and high throughput sequencing to discover CRMs and TF binding sites.

Gene Ontology. Terms within a controlled vocabulary to describe characteristics of gene products in the domains of cellular localization and biological function.

Chapter 1

Evolving approaches to the discovery of *cis*-regulatory elements and transcription factor binding sites in *Caenorhabditis elegans* and other metazoans

Introduction

Approaches to discover and characterize *cis*-regulatory modules (CRMs) in diverse model organisms have evolved and improved greatly in the last decade, enabling high throughput analysis and characterization of functional activity of noncoding sequences in eukaryotic genomes. In this chapter I will review methods in this field of research from the perspective of trying to apply these methods to study *C. elegans* transcriptional regulation. The central question guiding this review and my thesis is: How can we systematically identify and characterize CRMs and their regulatory functions? I will examine this question through the lens of historical approaches in the field and more recent methods that use sequencing as a read out of chromatin accessibility, TF binding, and functional activity.

The nematode *Caenorhabditis elegans* as a model for studying transcriptional regulation and development

Nematodes represent a diverse phylum and are increasingly well-studied, not in small part due to the rapidly decreasing costs of sequencing entire nematode genomes (Dillman et al. 2012; Sommer and Streit et al. 2011; Kumar et al. 2012). The genetically best-studied nematode species is *Caenorhabditis elegans*, with one of the best annotated and complete metazoan genome sequences containing some 20,431 protein-coding genes (Hillier et al. 2005). C. *elegans* presents a fruitful system in which to study transcriptional gene regulation in the context of development and evolution. The embryonic and larval development of *C. elegans* is well-studied and large populations of individuals are easy to grow and synchronize in liquid culture, making it easy to isolate large amounts of chromatin from worms at distinct life stages (e.g Baugh et al. 2009; Figure 1.1). Studies of *cis*-regulation in *C. elegans* have given us insight into mechanisms of transcriptional regulation during development from the rapid activation of growth genes following recovery from developmentally arrested states mediated by RNA polymerase II pausing (Baugh et al. 2009) to the *cis*-regulatory architecture involved in specification of cell fates (reviewed by Maduro et al. 2010).

Studying *C. elegans* transcription has some unique considerations due to *trans*-splicing of mRNA transcripts. Around 70% of *C. elegans* transcripts are known to be *trans*-spliced, wherein the RNA transcript containing a 3' splice site is spliced to an SL1 or SL2 splice leader sequence (Krause and Hirsch, 1987; reviewed in Blumenthal et al. 2012). As a result, the transcription start sites (TSS) of *C. elegans* are not easily defined with conventional RNA-seq methods. Fortunately, recent studies have used 5'capped nuclear RNA sequencing (Chen et al. 2013) and similar GRO-cap sequencing (Kruesi et al. 2013) to generate TSS maps for C. *elegans*. Also of note is that >17% of *C. elegans* genes are present in operons (Allen et al. 2011). Genes in operons are transcribed together as a polycistronic primary transcript and processed by splicing machinery to generate multiple messenger RNA transcripts (Blumenthal 2004; reviewed in Blumenthal 2012).

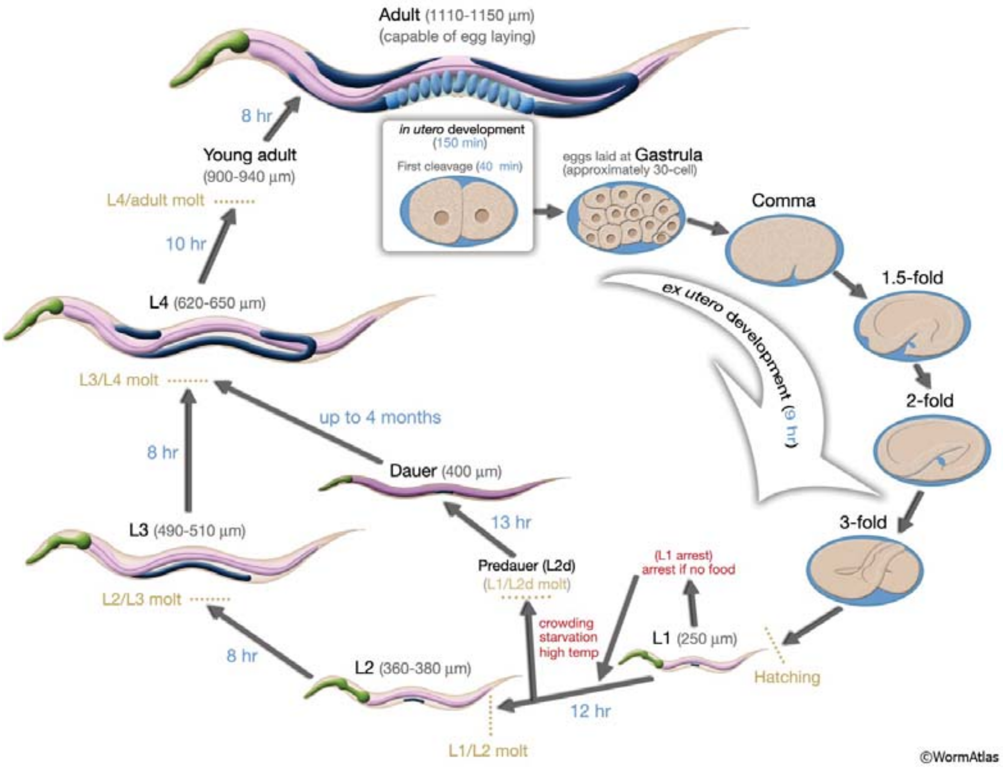


Figure 1.1 The life cycle of *Caenorhabditis elegans* (WormAtlas[1])

C. elegans is fast growing, with a lifecycle of ~2.5 days. An embryo undergoes about 11 hours of development to hatch. L1 larvae will arrest in the absence of available food. In the presence of food, L1 larvae will proceed to L2, but can be diverted to pre-dauer L2d in conditions of crowding, starvation and high temperature. L2 larvae will develop normally into L3, L4, and then into a reproductive adult.

[1] http://www.wormatlas.org/

C*is*-regulatory modules during development and the function of enhancers

The control of gene expression during development is critically dependent on the binding of transcription factor proteins to *cis*-regulatory modules (CRMs) in the genome to regulate transcription of target genes (Figure 1.2; reviewed in Noonan and McCallion et al., 2010; Borok et al. 2010). In the case of enhancers, which were first discovered in the SV40 simian virus as sequences that could drive the transcription of adjacent genes in an orientation-independent manner (Banerji et al. 1981; Benoist and Chambon, 1981), the binding of activator TFs to specific DNA motifs triggers recruitment of RNA polymerase II and drives transcription of the target gene according to specific spatiotemporal patterns. Repressor TF binding sites within the enhancer help to restrict spatiotemporal pattern of expression. Other CRMs such as silencers specifically block expression of target genes by binding repressor TFs or chromatin remodeling proteins such as Polycomb (Zhang and Bienz, 1992; Sengupta et al. 2004; reviewed in Ogbourne and Antalis, 1998). In *Drosophila* and mammals, insulators are a type of CRM that blocks transcription of a target gene in specific orientation (Kellum and Schedl, 1991; reviewed in Gaszner and Felsenfeld, 2007). Other examples of CRM types are the locus control regions (LCR), most notably in the β-globin locus and promoter tethering elements in the *Drosophila* bithorax complex (BX-C) (Akbari et al. 2008; Ho et al. 2011; Kwon et al. 2009).

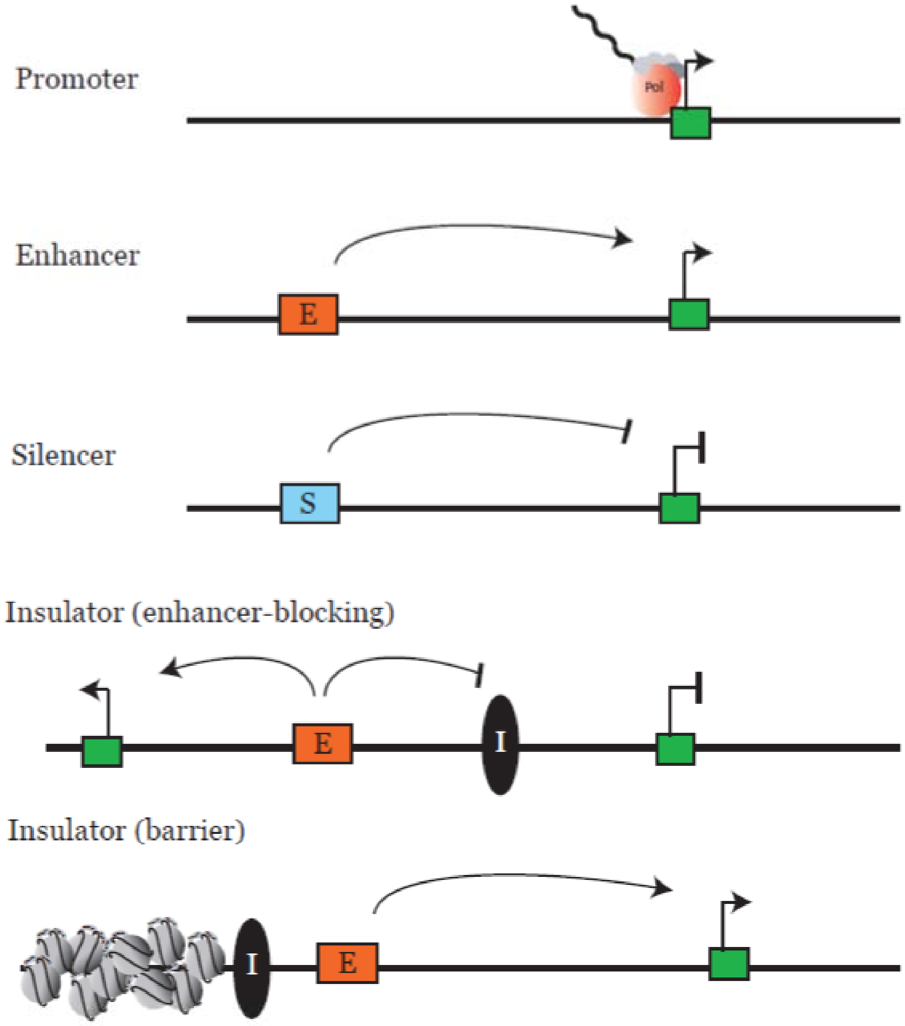


Figure 1.2 Major types of CRMs found in eukaryotes

Promoters (green) bind RNA Pol II and the basal transcription machinery to direct transcription. Enhancers (orange; which bind TFs) can act at a distance to upregulate target gene expression. Silencers (blue; which bind TFs) can also act at a distance to downregulate target gene expression. Insulators (black) can act either as enhancer blocking (EB) element or as a barrier to heterochromatin spread. Figure redrawn from Noonan and McCallion 2010.

In order to identify and study CRMs, it greatly helps to understand how they are situated within the context of the chromosome. In the nucleus, the DNA on chromosomes is wound around roughly 146bp of core histone octamer and packaged into nucleosomes (Kornberg 1977). Condensation of DNA into nucleosomes allows approximately 2 meters of DNA to be packaged into chromatin in the space of only a few microns in diameter. Regulation of the higher order structure of this chromatin is a complex process in three-dimensional space that renders parts of the DNA accessible or inaccessible to binding by DNA binding proteins such as transcription factors, RNA polymerase, and other chromatin-regulatory factors (Figure 1.3; reviewed in Cockerill 2011). Promoter and enhancer CRMS are often found in relatively nucleosome-free regions of the genome that are accessible to binding by TFs and other transcriptional machinery (reviewed in Shlyueva et al. 2014).

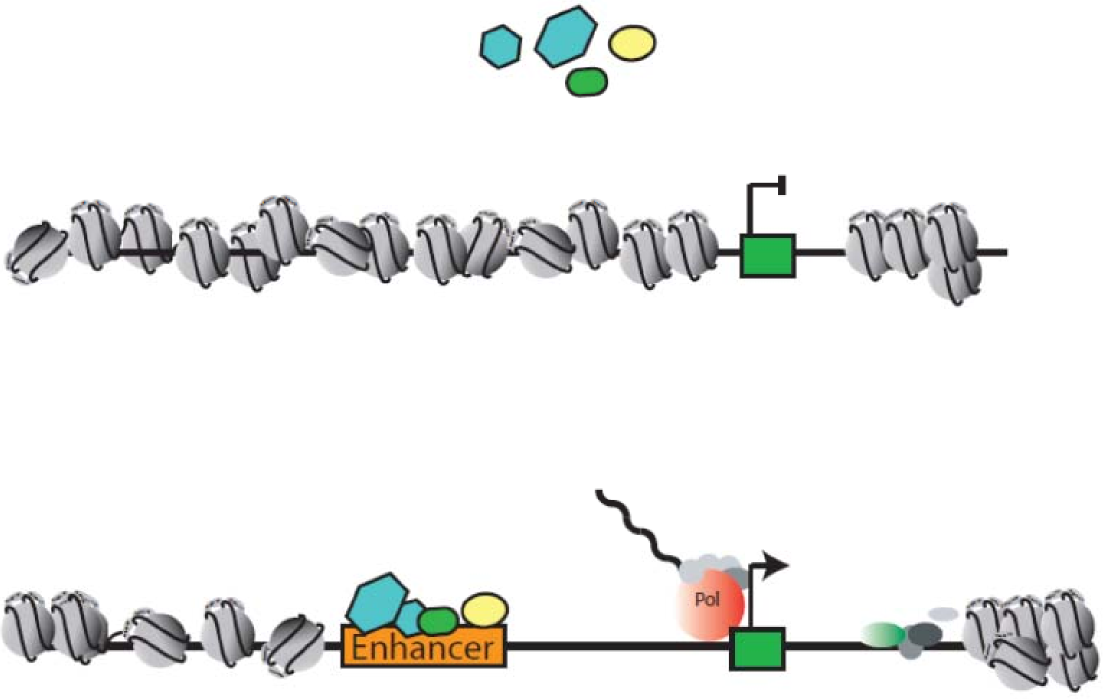


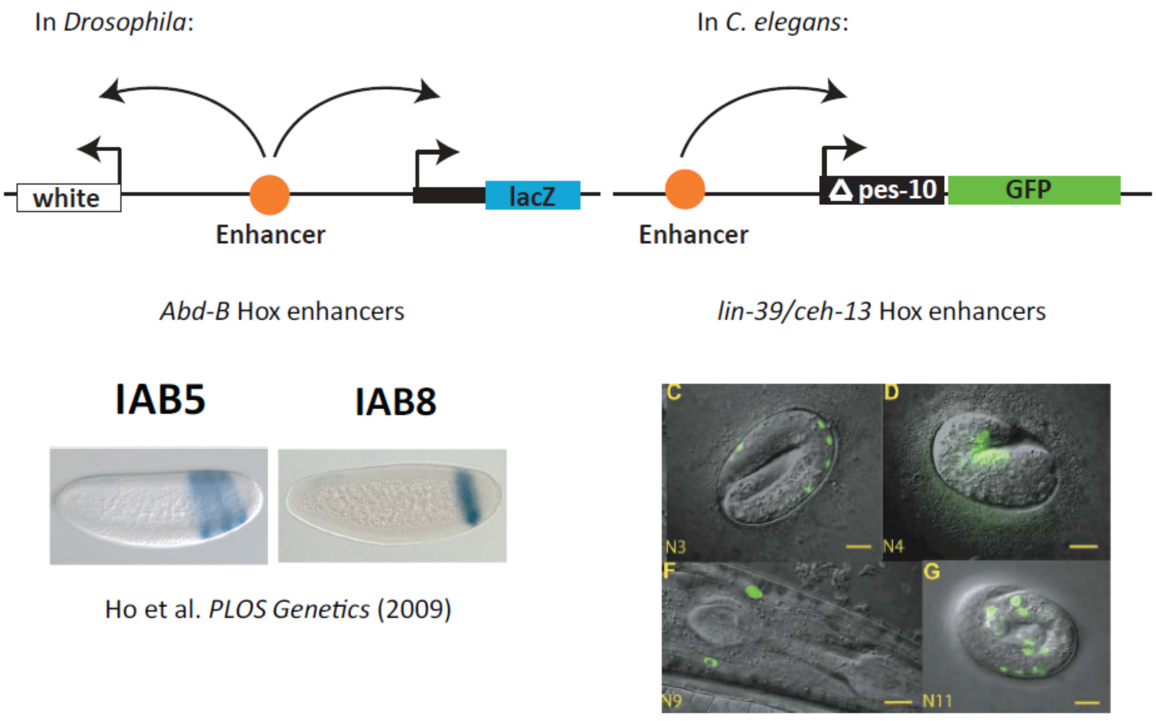
Figure 1.3 Chromatin as an accessibility barrier to binding by DNA-binding proteins to sequences such as enhancers and promoters.
CRMs such as enhancers (orange box) and promoters (green box) tend to be in relatively nucleosome-free regions where TFs (yellow and green ovals and blue hexagons) are able to access and bind to specific DNA binding motifs and recruit other DNA-binding proteins such as RNA polymerase II (red complex).

Many TFs use cooperativity in order to bind target binding sites in CRMs on nucleosomes (reviewed in Mirny 2010), but a subset of TFs, the pioneer TFs, are able to bind independently to nucleosomes and they do so earlier than most TFs (reviewed in Zaret and Carroll 2011). TF binding to target sites is not explained entirely by DNA sequence motifs (reviewed in Shlyueva et al. 2014) but also by local sequence features such as GC content (White et al. 2013) and perhaps also chromatin accessibility.

Additional factors help recruit TFs and transcriptional machinery to CRMs. These include CBP-1/P300 transcriptional activator (Visel et al. 2009) and the histone modification H3K4 methylation (Heintzmann et al 2007; Mikkelsen et al. 2007). Locations of these epigenetic marks have been used to locate enhancers (He et al. 2010). However, there is no consensus about exactly which marks are suitable and not all enhancers have marks (reviewed in Shlyueva et al. 2014).

Identifying CRMs using sequence conservation and limitations of these approaches

The gold standard method of testing enhancers has been to individually test sequences using a transgenic construct to determine whether these sequences are able to drive expression of a reporter gene such as lacZ or GFP (Figure 1.4).



Ho et al. *PLOS Genetics* (2009)

Kuntz et al. *Genome Research* (2008)

Figure 1.4 Testing enhancers for functional activity using transgenic reporter assays in *C. elegans* and *Drosophila*. Figures adapted from Ho et al. 2009; Kuntz et al. 2008.

Systematic interrogation of the genome by individually testing enhancers in transgenic reporter assays is laborious, requiring the cloning and injection of individual constructs for each test sequence. Having said that, systematic analysis has been performed for some large complex loci such as the *C. elegans lin-39/ceh-13* Hox locus (Kuntz et al. 2008) and the Hox genes in *Drosophila* BX-C (reviewed in Akbari et al. 2006 and Borok et al. 2010). Detection of CRMs in the *C. elegans* study by Kuntz et al. was greatly aided by the sequencing of many related *Caenorhabditis* species, allowing comparison of orthologous genomic sequences between species to identify regions exhibiting high sequence conservation, in an approach that is sometimes called phylogenetic footprinting. Kuntz and colleagues validated these conserved sequences as enhancers by testing them in transgenic reporter gene assays (Figure 1.4). The rationale behind this approach is that functional sequences such as regulatory sequences or protein-coding sequences are more likely to be conserved compared to genomic background because changes to these important sequences are likely to disrupt functional activity. This has been an approach that has helped to find many CRMs in *C. elegans* (e.g. Kirouac and Sternberg 2003; Wenick and Hobert 2004; Puckett-Robinson et al. 2013).

There are still many limitations to using sequence conservation since *cis*-regulatory sequences may not necessarily display increased sequence conservation compared to genomic background (Ho et al. 2009). Despite this lack of sequence conservation, orthologous enhancers from distantly related species have in many cases continued to function even after significant evolutionary sequence change (Hare et al. 2008; Ho et al. 2009). This appears to be due to conservation of TF binding site clusters and some flexibility in secondary binding sites, allowing sequences surrounding TF binding sites in the enhancer to change. Furthermore, at least in *Drosophila*, virtually all of the noncoding sequence can be considered conserved and so identifying regulatory elements based solely on sequence conservation is rather difficult (Peterson et al. 2009)

Algorithms to find clusters of TF binding sites have had some success in helping to predict the location of CRMs (Berman et al. 2002; Starr et al. 2011; Davidson et al. 2002; reviewed in Wasserman and Sandelin, 2004 and Su et al. 2010) but this is possible only if DNA binding motifs for TFs have been characterized beforehand and if activator and repressor TFs have been well defined for a particular locus or set of genes, as has been the case for well studied systems such as the *Drosophila* BX-C (Starr et al. 2011) and sea urchin endoderm gene regulatory network (Yuh et al. 1998).

Regardless, approaches relying solely on sequence conservation (Kuntz et al. 2008) or TF binding sites are still associated with significant false positives and negatives and better understanding of the constraints on sequence and function will likely help improve prediction of additional CRMs (Figure 1.5). Furthermore, there is a great need to increase the number of enhancer CRMs that are well-characterized. High throughput methods to identify and test CRMs would aid greatly in this endeavor.

Probing TF binding and chromatin accessibility with high-throughput sequencing: ChIP-seq, DNase-seq, FAIRE, ATAC-seq

Approaches utilizing high-throughput DNA sequencing technology to assay TF regulatory inputs and RNA output allow the investigation of *cis*-regulation genome wide (reviewed in Tsompana and Buck 2014). Studies of chromatin immunoprecipitation (ChIP) with antibodies against transcription factors of interest to isolate DNA bound by those TFs allows the measurement of TF binding sites in the genome (Ren et al. 2000; Johnson et al. 2007). Data from these ChIP-chip (wherein DNA is hybridized to microarrays) and ChIP-seq (wherein DNA is sequenced) studies can be mined to detect CRMs (Visel et al. 2009).

In *C. elegans*, ChIP-seq studies have helped identify binding sites for over 100 TFs of interest and the locations of chromatin regulatory marks such as H3K4 methylation, H3K27 acetylation, etc. (Araya et al. 2014; Zhong et al. 2010; Gerstein et al. 2010; Kuntz et al. 2012) as well as the transcriptional machinery of RNA polymerase II (Baugh et al. 2009). ChIP-seq is limited by the availability of high quality antibodies or GFP-tagged TFs of interest. Interestingly, not all TF sites bound in ChIP-seq are functional enhancers, raising the question of what, other than TF binding, determines the functional activity of sequences. This may be due to the need for cooperative binding of TFs, or local chromatin context such as histone marks and chromatin accessibility.

It has been known since the early 1980s that DNaseI, a nuclease with relatively low sequence specificity, will cut based on chromatin accessibility and thus preferentially in nucleosome-free CRMs (Gross and Garrard, 1988; reviewed in Cockerill et al. 2010). In fact, the CRMs of the β-globin locus were discovered using DNaseI hypersensitivity assays (Fraser et al. 1993; Tuan et al. 1985), and early studies showed that chromatin domains containing actively transcribed genes are at least twice as accessible to nuclease digestion as inactive genes (reviewed in Cockerill et al. 2011). Other older footprinting assays used chemicals such as potassium permanganate ($KMnO_4$) and dimethyl sulfate (DMS) to identify CRMs, based on selective oxidation of single-stranded thymine and differential methylation of guanine bound or unbound by TFs, respectively, but these are not scalable (Spicuglia et al. 2004; Drouin et al. 1997). Compared to earlier methods measuring footprinting in specific loci using northern blots, it possible to treat chromatin with DNaseI and size select and sequence the shortest fragments and measure chromatin accessibility over the entire genome in a method called DNase-seq (Figure 1.5; Hesselberth et al 2009; Thurman et al. 2012). Two methods of DNase-seq have been described (Boyle et al. 2011; Hesselberth et al. 2009), with the double-hit protocol from the Stamatoyannopoulos lab being primarily used by ENCODE (Consortium 2012).

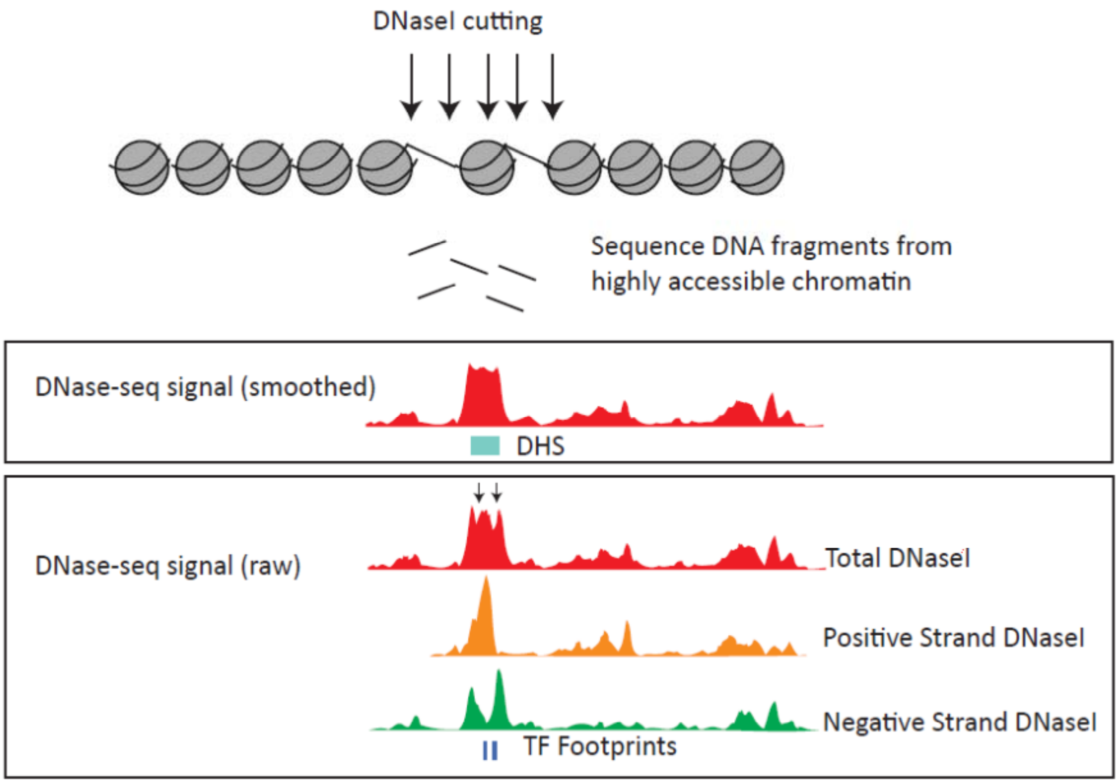


Figure 1.5 DNase-seq schematic.

TF binding within a CRM can also be detected in DNase-seq data. Within larger regions (hundreds of base pairs) showing DNaseI hypersensitivity (high read coverage), the presence of TFs will protect smaller regions (6-40bp) from being cut by DNaseI (low read coverage) and also cause a strand-shift in read coverage (Figure 1.6). An example of DNase-seq data from the *C. elegans* embryo is shown in Figure 1.7.

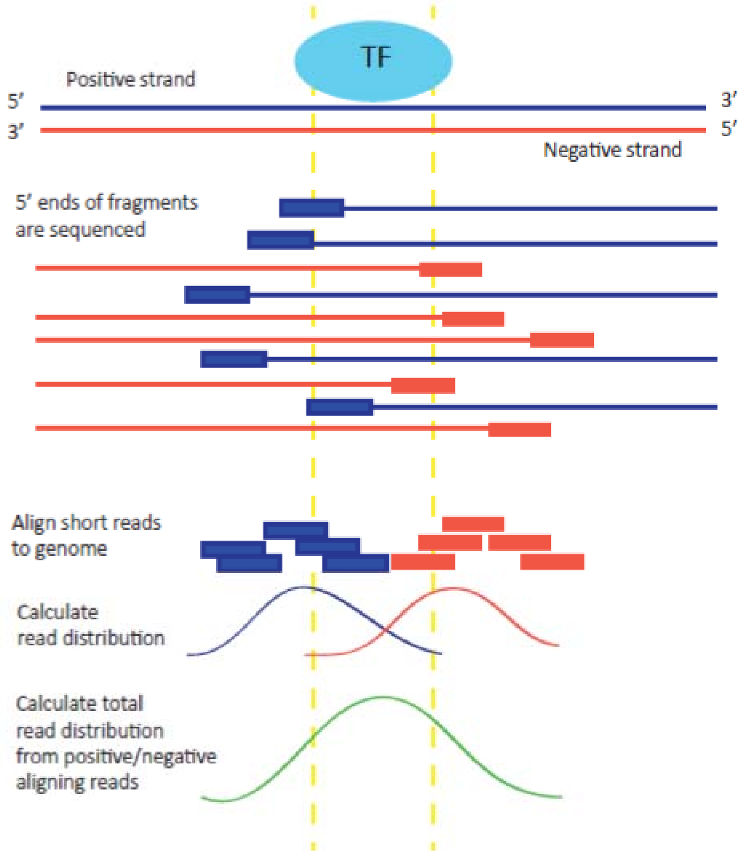

TF
Positive strand
5'
3'
3'
5'
Negative strand
5' ends of fragments are sequenced
Align short reads to genome
Calculate read distribution
Calculate total read distribution from positive/negative aligning reads

Figure 1.6 Strand shift in reads in ChIP-seq and DNase-seq due to TF binding. Sequencing by synthesis occurs in a 5' to 3' direction, yielding reads that align on opposite strands on either side of a bound TF (figure adapted from Park et al. 2009). Aligning reads from each strand results in peaks that flank the TF binding site.

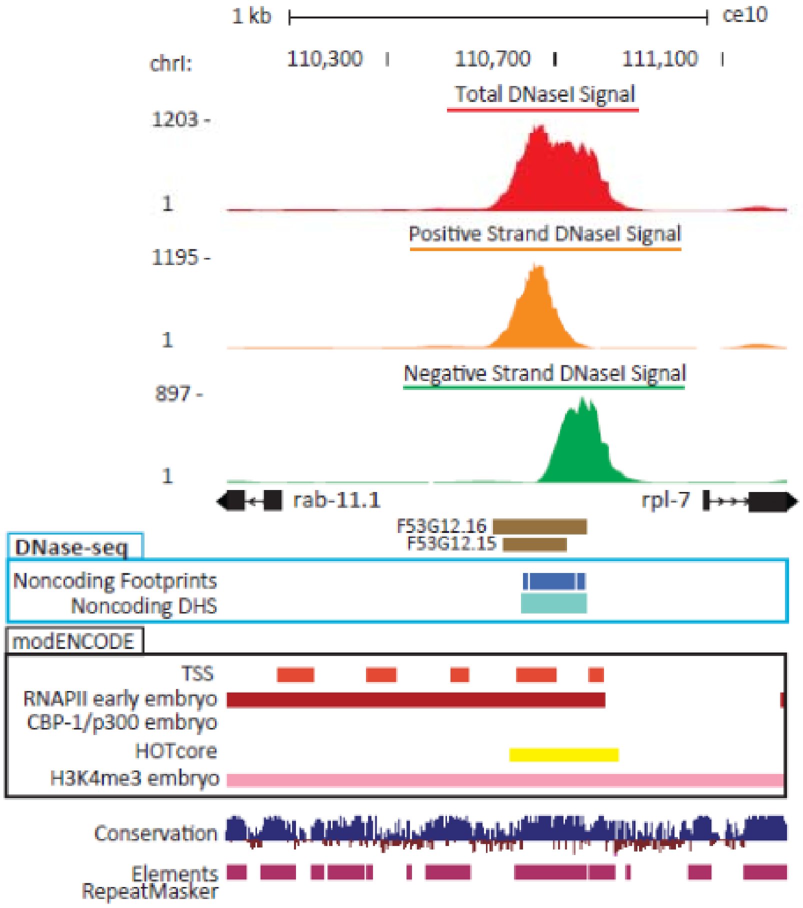


Figure 1.7 Example of DNase-seq data

Total DNaseI signal (red) can be separated in to positive (light orange) and negative strands (green). One noncoding DHS (light blue) and several TF footprints (dark blue) were found overlapping two noncoding transcripts (brown) between two embryo-expressed genes *rab-11.1* and *rpl-7* (black with arrows). Existing comparison data from modENCODE (Gerstein et al. 2010) shows ChIP peaks of RNA Pol II (dark red), H3K4me3 (pink), a highly occupied TF region (yellow, indicates more than 15 TFs binding) and TSS data from Chen et al. 2013 (dark orange). Conservation track across seven *Caenorhabditis* species is shown in dark blue and MULTIZ conserved elements in magenta.

The depth of sequencing required to probe chromatin accessibility depends on the desired features to be captured. TF footprinting with DNase-seq requires a higher depth of sequencing. Paired end or long reads are often preferred in genomes where there are many repeat elements or there is low complexity. However short read sequencing, which is less expensive, is often sufficient for chromatin accessibility studies (reviewed in Tsompana and Buck, 2014).

A similar method to DNase-seq, formaldehyde-assisted identification of regulatory element elements (FAIRE) can be used to make regulatory maps using formaldehyde crosslinking followed by phenol-chloroform extraction to isolate nucleosome-depleted regions in the aqueous layer for sequencing (Giresi et al. 2007; Giresi and Lieb 2009). FAIRE suffers from low signal to noise ratio and it does not provide the resolution needed to identify TF footprints within CRMs. Studies comparing DNase-seq and FAIRE show strong-cross-validation of putative CRMs identified (Song et al. 2011). FAIRE, being an orthogonal study, can still be useful to validate some DNase-seq results.

Another promising alternative to DNase-seq is transposase-accessible chromatin using sequencing, also known as ATAC-seq. ATAC-seq utilizes a Tn5 transposase to insert sequencing primers into the genome based on chromatin accessibility. In comparison to DNase-seq, several thousand cells are needed instead of 100,000 cells required for DNase-seq. ATAC-seq involves only two steps: Tn5 insertion followed by PCR (Buenrostro et al. 2013), and therefore reduces loss of sample material from gel extraction and adaptor ligation needed in DNase-seq. Maps of chromatin accessibility from ATAC-seq can be equal or close to the quality of DNase-seq and can also provide maps of nucleosome positioning near regions of accessibility.

Investigating TF footprinting in chromatin accessibility studies

Several recent TF footprint detection methods have been developed that use similar underlying statistical methods to detect lower read coverage and strand shift in reads indicative of TF binding sites. The Wellington algorithm (Piper et al. 2013) detects significantly lower read coverage in a region within a DHS compared to positive and negative shoulder regions of varying shoulder lengths and tests the null hypothesis that the number of reads in the footprint region is proportional to region length. P-values are then calculated for the footprints and TF footprints are chosen on the basis of a p-value threshold.

A more recent method, DNase2TF, has been shown to improve accuracy and sensitivity, and also provide a greater number of TF footprints when tested against orthogonally derived ChIP-seq TF binding sites using receiver-operating characteristic (ROC) curves (Sung et al. 2014). DNase2TF works by measuring cut count within a DHS and then adjusts the cut count by dinucleotide frequency bias (measured from the DNaseI sample) and mappability using measures of read mappability generated by PeakSeq (Rozowsky et al. 2009). Cut count depletion (indicating TF protection) is measured and modeled with a binomial distribution to assess the significance of local depletion with a z-score. This z-score compares cut count in the candidate region and in a surrounding window that is three times the size of the region. The more the candidate region is depleted of cutting, the lower its z-score and the greater its depth of TF protection. Footprints are merged if comparing the z-score between consecutive footprints shows that the z-score of the combined region is better than the individual regions. The location of reads mapping within each DHS are randomized, allowing an estimation of the false discovery rate (FDR) and a threshold z-score.

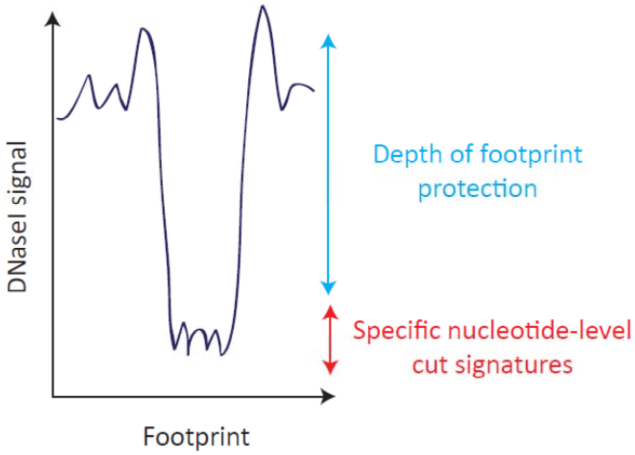

DNaseI signal
Depth of footprint protection
Specific nucleotide-level cut signatures
Footprint

Figure 1.8 DNaseI profiles are composed of two components: Protection from transiently bound TF and DNA-cut signature.

An important consideration in the analysis of DNase-seq has been raised by Sung and colleagues (2014). They found that, contrary to previous reporting of low sequence specificity for nuclease digestion by DNaseI, there is some DNaseI sequence specificity that impacts the observed profile of footprints for a given TF or sequence, and this is not dependent on TF-DNA contacts as was previously reported (Hesselberth et al. 2009). Instead, it appears critical that TF footprints are called on the basis of protection depth and not on their specific nucleotide-level cut signatures (Figure 1.8). These nucleotide-level cut signatures are in fact dependent on the use of DNaseI as the cutting nuclease and can be predicted by measuring and modeling the dinucleotide cut preferences of nucleases on naked DNA (Sung et al. 2014). This is likely also to prove an important caveat to similar analyses using ATAC-seq since it seems likely that no accessibility method is entirely immune to sequence bias.

Importance of the transgenic functional assay and need for higher throughput assays

Transgenic reporter assays continue to be the gold standard test for testing *cis*-regulatory activity, but new approaches using high throughput sequencing are enabling parallel testing of enhancers. Parallel assays have been previously described that use sequencing (Nam and Davidson 2012) and/or fluorescence-activated cell sorting (FACS) methods to test enhancers in bulk (Gisselbrecht et al. 2013; Dickel et al 2014).

In enhancer FACS-seq, libraries of putative enhancers are cloned upstream of fluorescent reporter genes and these constructs are injected to generate transgenic organisms. Dissociated cells from the transgenic organisms are selected for the fluorescent transgene with FACS and sequenced in order to determine active enhancer sequences (Gisselbrecht et al. 2013). Additional fluorescent reporters or specific cell-surface markers can be expressed to allow cell or tissue-specific sorting.

Another similar method called SIF-seq uses single-copy site-specific integration of putative enhancer libraries cloned upstream of a reporter gene with FAC-sorting and sequencing to test enhancer CRMs. Both these methods are effective but require the additional step of FAC sorting which may limit throughput. Hundreds of CRMs were tested in the case of eFS, whereas the use of fragmented BAC constructs in SIF-seq limited them to specific gene loci. These techniques are nevertheless still promising.

The use of custom oligo libraries traditionally used in the synthesis of microarrays to generate test sequences tagged with unique barcodes and distinct amplification primers have opened the doors to studies (MPRA and FIREWACh) testing many tens or hundreds of thousands of enhancers for functional activity (Melnikov et al. 2012; Murtha et al. 2014). Custom oligos are synthesized as a mixture and primers can be designed to amplify subsets of the oligo library. Using custom oligo library technology, parallel reporter assay constructs are designed such that a unique barcode (included in the oligo sequence) is expressed when the putative CRM (also on the custom oligo) is able to drive expression. Testing sequences for functional activity is accomplished by transfecting oligo library constructs into cell lines, and then simultaneously collecting RNA and genomic DNA to be sequenced using RNA-seq and DNA sequencing. Detection of unique barcodes in RNA-seq expression can therefore indicate that the sequence that is uniquely associated with the barcode is able to function as an enhancer. DNA sequencing enables the RNA-seq expression data to be normalized by the amount of transgene that is successfully transfected into cells. Thousands of enhancer sequences can thus be screened in a single experiment and, if found to direct expression in the sequencing data, they can be selected for further characterization in single transgene assays. Furthermore, using custom oligos allows for mutagenesis and manipulation of any part of the enhancer sequence to be tested (for example, mutations in TF binding sites) allowing analysis of enhancer function.

Another variation, STARR-seq, has the candidate sequence being tested for enhancer activity cloned downstream of the reporter gene, so that it is also transcribed (Arnold et al. 2013). This sequence is then detectable in the RNA-seq expression data. This can mitigate the need to have barcodes to distinguish each enhancer sequence (Figure 1.9). In this case, thousands of potential enhancers are isolated for cloning using genomic fragmentation. Application of a massively parallel reporter assay to *C. elegans* using transgenesis yielding extrachromosomal arrays is promising and would open up the system to large comparative studies of CRM function.

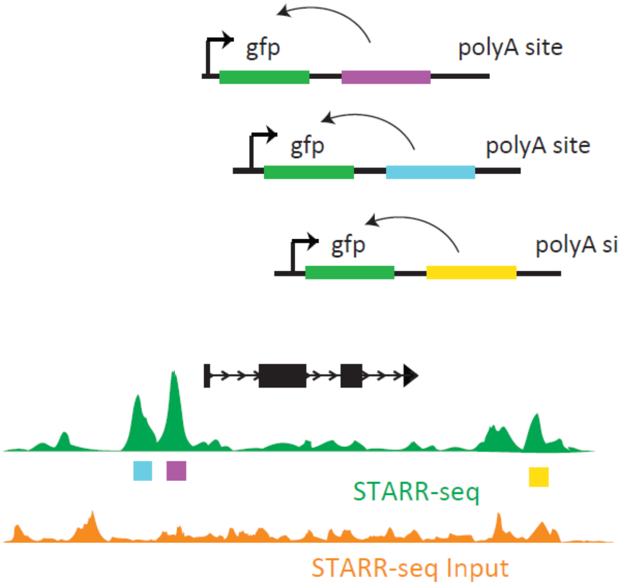


Figure 1.9 STARR-seq experimental design and resulting enhancer maps

Libraries of putative enhancers (purple, light blue, yellow) were isolated by genomic fragmentation and cloned downstream of GFP reporter. Once construct libraries are transfected in S2 cells, DNA and poly-A RNA are isolated and sequenced. The resulting STARR-seq of the RNA reads (dark blue) shows representation of the enhancer sequences, indicating sequences are able to drive reporter expression.

Comparative genomics of nematodes

Comparisons of enhancer CRMs in different species are useful to study their function and evolution, and thus the study of *C. elegans* transcriptional regulation will undoubtedly benefit from more comparisons with related species in the nematode phylum. To date, more than 80 nematode genomes have been published (according to WormBase ParaSite[2]), including some with transcriptome profiles by RNA-seq, which enables the annotation of protein-coding genes. The diversity of species being sequenced from all nematode clades represents a rich genomic toolkit with which to investigate nematode development, evolution, and behavior, especially as these nematodes have diverse ecology and lifestyles, ranging from free-living to parasitic, and reach evolutionary distances that span hundreds of millions of years (Dillman et al. 2012). Much of the comparative analysis of nematode genomes has focused on protein coding genes, such as protein families that appear to have expanded in the genomes of parasites and may play a role in host infection (Dillman et al. 2012; Dillman et al. 2013). However, future studies that delve into the noncoding regions of these genomes are likely to yield fascinating insights into the mechanism of regulation of important genes.

Evolution of the Hox gene complex and *cis*-regulatory elements in nematodes

The Hox genes are an ancient regulatory protein family and are involved in regulating critical developmental process across metazoans. Hox gene regulation has been studied by researchers over many decades (McGinnis et al. 1984; Lewis et al. 1978; reviewed in Pearson et al. 2005). Hox gene complexes have been studied across several nematodes, showing striking loss and sequence turnover compared to other metazoans (Aboobaker and Blaxter, 2003a,b). Among closely related

[2] http://parasite.wormbase.org

Caenorhabditis species, sequence conservation has been used successfully to identify CRMs such as those from *lin-39/ceh-13* (Kuntz et al. 2008). The latest data from *Steinernema* genomes show that many of the Hox genes present in *C. elegans* are also present in members of the *Steinernema* genus (Dillman, Macchietto et al. submitted; Figure 1.10). However, many additional unrelated protein-coding genes appear to be inserted between the Hox genes *lin-39* and *ceh-13*, increasing the intergenic distance to more than 40 kb. It remains to be seen whether the *cis*-regulatory regions found by Kuntz et al. (2008) in *C. elegans* are conserved in other nematodes such as those in *Steinernema* genus.

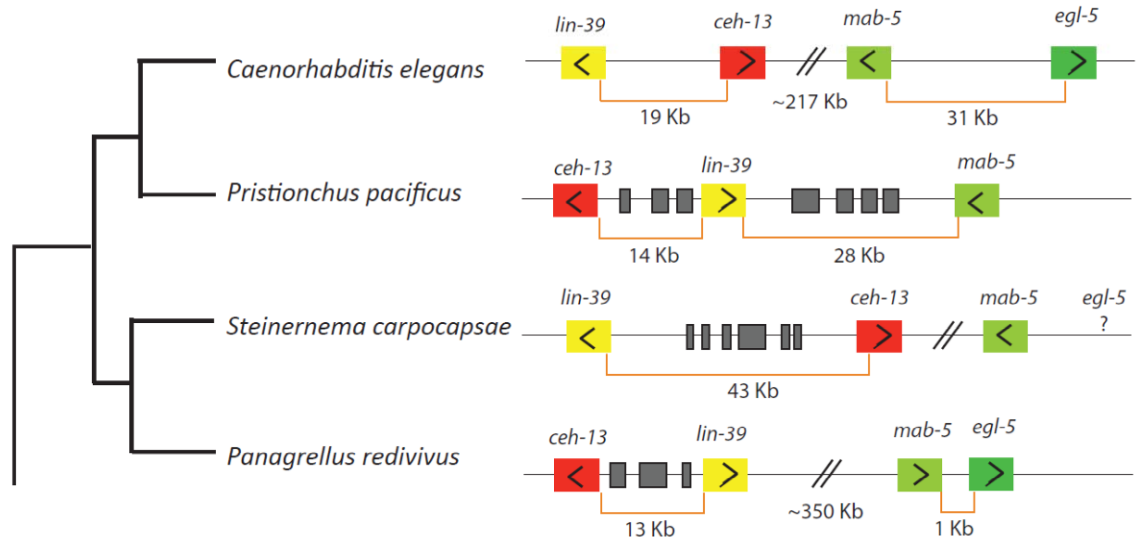


Figure 1.10 Nematode Hox gene clusters (Adapted from Dillman, Macchietto et al. submitted)

Improving our knowledge of *cis*-regulation in *C. elegans* and other nematodes will help to address questions about function and flexibility in the evolution of

CRMs. The more examples that we have of characterized CRMs in *C. elegans* and other nematode species, the better we are able to understand the underlying mechanisms of CRM function, evolutionary change, and species diversity.

References

Aboobaker AA, Blaxter ML. 2003a. Hox Gene Loss during Dynamic Evolution of the Nematode Cluster. Curr Biol. 13(1):37-40.

Aboobaker A, Blaxter M. 2003b. Hox gene evolution in nematodes: novelty conserved. Curr Opin Genet Dev. 13(6):593-8.

Akbari OS, Bae E, Johnsen H, Villaluz A, Wong D, Drewell RA. 2008. A novel promoter-tethering element regulates enhancer-driven gene expression at the bithorax complex in the Drosophila embryo. Development. 135(1):123-31.

Akbari OS, Bousum A, Bae E, Drewell RA. 2006. Unraveling cis-regulatory mechanisms at the abdominal-A and Abdominal-B genes in the Drosophila bithorax complex. Dev Biol. 293(2):294-304.

Araya CL, Kawli T, Kundaje A, Jiang L, Wu B, Vafeados D, Terrell R, Weissdepp P, Gevirtzman L, Mace D, et al. 2014. Regulatory analysis of the *C. elegans* genome with spatiotemporal resolution. Nature 512:453-6.

Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science. 339(6123):1074-7.

Banerji J, Rusconi S, Schaffner W. 1981. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. Cell 27, 299–308.

Benoist C, Chambon P. 1981. In vivo sequence requirements of the SV40 early promotor region. Nature. 290(5804):304-10.

Baugh LR, Demodena J, Sternberg PW. 2009. RNA Pol II accumulates at promoters of growth genes during developmental arrest. Science 324:92-4.

Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, et al. 2002. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc Natl Acad Sci U S A 99: 757–762.

Blumenthal, T. 2004. Operons in eukaryotes. Briefings in Functional Genomics and Proteomics 3:199–211.

Blumenthal T. 2012. Trans-splicing and operons in *C. elegans*. WormBook. 1-11. doi: 10.1895/wormbook.1.5.2.

Borok MJ, Tran DA, Ho MC, Drewell RA. 2010. Dissecting the regulatory switches of development: lessons from enhancer evolution in Drosophila. Development. 137(1):5-13.

Boyle AP, Song L, Lee BK, London D, Keefe D, Birney E, Iyer VR, Crawford GE, Furey TS. 2011. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. Genome Res. 21(3):456-64.

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. 10(12):1213-8.

Chen RA, Down TA, Stempor P, Chen QB, Egelhofer TA, Hillier LW, Jeffers TE, Ahringer J. 2013. The landscape of RNA polymerase II transcription initiation in *C. elegans* reveals promoter and enhancer architectures. Genome Res 23:1339-47.

Cockerill PN.2011. Structure and function of active chromatin and DNase I hypersensitive sites. FEBS J. 278(13):2182-210

Davidson, EH. Rast JP, Oliveri P, Ransick A, Calestani C, Yuh CH, Minokawa T, Amore G, Hinman V, Arenas-Mena C, Otim O, et al. A genomic regulatory network for development. Science 295, 1669–1678 (2002).

Dickel DE, Zhu Y, Nord AS, Wylie JN, Akiyama JA, Afzal V, Plajzer-Frick I, Kirkpatrick A, Göttgens B, Bruneau BG et al. 2014. Function-based identification of mammalian enhancers using site-specific integration. Nat Methods. 11(5):566-71.

Dillman AR, Mortazavi A, Sternberg PW. 2012. Incorporating genomics into the toolkit of nematology. J Nematol. 44(2):191-205.

Dillman AR, Minor PJ, Sternberg PW. 2013. Origin and evolution of dishevelled. G3 (Bethesda). 3(2):251-62.

Dillman AR, Macchietto M, Porter CF, Rogers A, William BA, Antoshechkin I, Lee M, Goodwin Z, Lu X, Lewis EE, Goodrich-Blair H, et al. Comparative genomics of Steinernema reveals deeply conserved regulatory networks in nematodes. Submitted.

Drouin R, Angers M, Dallaire N, Rose TM, Khandjian EW, Rousseau F. 1997. Structural and functional characterization of the human FMR1 promoter reveals similarities with the hnRNP-A2 promoter region. Hum Mol Genet. 6(12):2051-60.

ENCODE Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414):57–74

Fraser P, Pruzina S, Antoniou M, Grosveld F. 1993. Each hypersensitive site of the human beta-globin locus control region confers a different developmental pattern of expression on the globin genes. Genes Dev 7:106-13.

Gaszner M, Felsenfeld G. 2006. Insulators: exploiting transcriptional and epigenetic mechanisms. Nat Rev Genet. 7(9):703-13.

Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, et al. 2010. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. Science 330:1775-87.

Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. 2007. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res. 17(6):877-85

Giresi PG, Lieb JD. 2009. Isolation of active regulatory elements from eukaryotic chromatin using FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements). Methods. 48(3):233-9

Gisselbrecht SS, Barrera LA, Porsch M, Aboukhalil A, Estep PW 3rd, Vedenko A, Palagi A, Kim Y, Zhu X, Busser BW, et al. 2013. Highly parallel assays of tissue-specific enhancers in whole Drosophila embryos. Nat Methods. 10(8):774-80.

Gross DS, Garrard WT. 1988. Nuclease hypersensitive sites in chromatin. Annu Rev Biochem 57:159-97.

Hare EE, Peterson BK, Iyer VN, Meier R, Eisen MB. 2008. Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. PLoS Genet. 4(6):e1000106.

He HH, Meyer CA, Shin H, Bailey ST, Wei G, Wang Q, Zhang Y, Xu K, Ni M, Lupien M, et al. 2010. Nucleosome dynamics define transcriptional enhancers. Nat Genet. 42(4):343-7.

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39:311-8.

Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, Thurman RE, Neph S, Kuehn MS, Noble WS et al. 2009. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nat Methods 6:283-9.

Hillier LW, Coulson A, Murray JI, Bao Z, Sulston JE, Waterston RH. 2005. Genomics in *C. elegans*: so many genes, such a little worm. Genome Res. 15(12):1651-60.

Ho MC, Johnsen H, Goetz SE, Schiller BJ, Bae E, Tran DA, Shur AS, Allen JM, Rau C, Bender W, Fisher WW, Celniker SE, Drewell RA. Functional evolution of cis-regulatory modules at a homeotic gene in Drosophila. PLoS Genet. 2009 Nov;5(11):e1000709

Ho MC, Schiller BJ, Akbari OS, Bae E, Drewell RA. 2011. Disruption of the abdominal-B promoter tethering element results in a loss of long-range enhancer-directed Hox gene expression in Drosophila. PLoS One. 6(1):e16283.

Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. Science. 316(5830):1497-502.

Kellum, R., Schedl, P. 1991. A position-effect assay for boundaries of higher order chromosomal domains. Cell 64, 941–950

Kirouac M, Sternberg PW. 2003. cis-Regulatory control of three cell fate-specific genes in vulval organogenesis of *Caenorhabditis elegans* and C. briggsae. Dev Biol. 257(1):85-103.

Kornberg RD. 1977. Structure of chromatin. Annu Rev Biochem. 1977;46:931-54.

Krause M, Hirsh D. 1987. A trans-spliced leader sequence on actin mRNA in *C. elegans*. Cell.. 49(6):753-61.

Kruesi WS, Core LJ, Waters CT, Lis JT, Meyer BJ. 2013. Condensin controls recruitment of RNA polymerase II to achieve nematode X-chromosome dosage compensation. Elife. 2:e00808.

Kumar S, Koutsovoulos G, Kaur G, Blaxter M. 2012. Toward 959 nematode genomes. Worm. 1(1):42-50.

Kuntz SG, Schwarz EM, DeModena JA, De Buysscher T, Trout D, Shizuya H, Sternberg PW, Wold BJ. 2008. Multigenome DNA sequence conservation identifies Hox cis-regulatory elements. Genome Res 18:1955-68.

Kuntz SG, Williams BA, Sternberg PW, Wold BJ. 2012. Transcription factor redundancy and tissue-specific regulation: evidence from functional and physical network connectivity. Genome Res 22:1907-19.

Kwon D, Mucci D, Langlais KK, Americo JL, DeVido SK, Cheng Y, Kassis JA. 2009. Enhancer-promoter communication at the Drosophila engrailed locus. Development. 136(18):3067-75.

Lewis EB. 1978. A gene complex controlling segmentation in Drosophila. Nature. 276(5688):565-70.

Ludwig, M. Z., Bergman, C., Patel, N. H. & Kreitman, M. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 403, 564–567.

Ludwig, M. Z. Palsson A, Alekseeva E, Bergman CM, Nathan J, Kreitman M. 2005 Functional evolution of a cis-regulatory module. PLoS Biol. 3, e93.

Maduro MF· 2010. Cell fate specification in the *C. elegans* embryo. Dev Dyn. 239(5):1315-29.

McGinnis W, Levine MS, Hafen E, Kuroiwa A, Gehring WJ. 1984. A conserved DNA sequence in homoeotic genes of the Drosophila Antennapedia and bithorax complexes. Nature. 308(5958):428-33.

Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG Jr, Kinney JB, et al. 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat Biotechnol. 30(3):271-7.

Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, et al. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature. 448(7153):553-60.

Mirny LA. 2010. Nucleosome-mediated cooperativity between transcription factors. Proc Natl Acad Sci U S A. 107(52): 22534–22539.

Murtha M, Tokcaer-Keskin Z, Tang Z, Strino F, Chen X, Wang Y, Xi X, Basilico C, Brown S, Bonneau R,et al. 2014. FIREWACh: high-throughput functional detection of transcriptional regulatory modules in mammalian cells. Nat Methods. 11(5):559-65.

Nam J, Davidson EH. 2012. Barcoded DNA-tag reporters for multiplex cis-regulatory analysis. PLoS One. 7(4):e35934.

Noonan JP, McCallion AS. 2010. Genomics of long-range regulatory elements. Annu Rev Genomics Hum Genet 11:1-23.

Niu W, Lu ZJ, Zhong M, Sarov M, Murray JI, Brdlik CM, Janette J, Chen C, Alves P, Preston E, et al. 2011. Diverse transcription factor binding features revealed by genome-wide ChIP-seq in *C. elegans*. Genome Res 21:245-54.

Ogbourne S, Antalis TM. 1998. Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. Biochem J. 1998 Apr 1;331 (Pt 1):1-14.

Park PJ. 2009. ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet. 10(10):669-80.

Peterson BK, Hare EE, Iyer VN, Storage S, Conner L, Papaj DR, Kurashima R, Jang E, Eisen MB.2009. Big genomes facilitate the comparative identification of regulatory elements. PLoS One. 4(3):e4688

Piper J, Elze MC, Cauchy P, Cockerill PN, Bonifer C, Ott S. 2013. Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. Nucleic Acids Res. 41(21):e201.

Puckett-Robinson C, Schwarz EM, Sternberg PW. Identification of DVA interneuron regulatory sequences in *Caenorhabditis elegans*. PLoS One. 2013;8(1):e54971.

Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, et al. 2000. Genome-wide location and function of DNA binding proteins. Science 290:2306-9.

Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB. 2009. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat Biotechnol. 27(1):66-75.

Ruvinsky I, Ruvkun G. 2003. Functional tests of enhancer conservation between distantly related species. Development. 130(21):5133-42.

Sengupta AK, Kuhrs A, Müller J. 2004. General transcriptional silencing by a Polycomb response element in Drosophila. Development. 2004 May;131(9):1959-65.

Shlyueva D, Stampfel G, Stark A. 2014. Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet. 15(4):272-86.

Spicuglia S, Kumar S, Chasson L, Payet-Bornet D, Ferrier P. 2004. Potassium permanganate as a probe to map DNA-protein interactions in vivo. J Biochem Biophys Methods. 59(2):189-94.

Sommer RJ, Streit A. Comparative genetics and genomics of nematodes: genome structure, development, and lifestyle. Annu Rev Genet. 2011;45:1-20. doi: 10.1146/annurev-genet-110410-132417.

Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee BK, Sheffield NC, Gräf S, Huss M, Keefe D, et al. 2011. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res 21:1757-67.

Starr MO, Ho MC, Gunther EJ, Tu YK, Shur AS, Goetz SE, Borok MJ, Kang V, Drewell RA. 2011. Molecular dissection of cis-regulatory modules at the Drosophila bithorax complex reveals critical transcription factor signature motifs. Dev Biol. 359(2):290-302.

Su J, Teichmann SA, Down TA. 2010. Assessing computational methods of cis-regulatory module prediction. PLoS Comput Biol. 6(12):e1001020.

Sung MH, Guertin MJ, Baek S, Hager GL. 2014. DNase footprint signatures are dictated by factor dynamics and DNA sequence. Mol Cell. 56(2):275-85.

Thomas S, Li XY, Sabo PJ, Sandstrom R, Thurman RE, Canfield TK, Giste E, Fisher W, Hammonds A, Celniker SE, et al. 2011. Dynamic reprogramming of chromatin accessibility during Drosophila embryo development. Genome Biol 12:R43.

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B et al. 2012. The accessible chromatin landscape of the human genome. Nature. 489(7414):75-82.

Tsompana M, Buck MJ. 2014. Chromatin accessibility: a window into the genome. Epigenetics Chromatin. 7(1):33.

Tuan D, Solomon W, Li Q, London IM. 1985. The "beta-like-globin" gene domain in human erythroid cells. Proc Natl Acad Sci U S A. 82(19):6384-8.

Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, et al. 2009. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature 457:854-8.

Wasserman WW, Sandelin A. 2004. Applied bioinformatics for the identification of regulatory elements. Nature Reviews Genetics 5, 276-287.

Wenick AS, Hobert O. 2004. Genomic cis-regulatory architecture and trans-acting regulators of a single interneuron-specific gene battery in *C. elegans*. Dev Cell. 6(6):757-70.

White MA, Myers CA, Corbo JC, Cohen BA. 2013. Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. Proc Natl Acad Sci U S A. 110(29):11952-7.

Yuh CH, Bolouri H, Davidson EH. 1998. Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. Science. 279(5358):1896-902.

Zaret KS, Carroll JS. 2011. Pioneer transcription factors: establishing competence for gene expression. Genes Dev. 25(21):2227-41. doi: 10.1101/gad.176826.111.

Zhang CC, Bienz M. 1992. Segmental determination in Drosophila conferred by hunchback (hb), a repressor of the homeotic gene Ultrabithorax (Ubx). Proc Natl Acad Sci U S A. 89(16):7511-5.

Zhong M, Niu W, Lu ZJ, Sarov M, Murray JI, Janette J, Raha D, Sheaffer KL, Lam HY, Preston E, et al. 2010. Genome-wide identification of binding sites defines distinct functions for *Caenorhabditis elegans* PHA-4/FOXA in development and environmental response. PLoS Genetics 6:e1000848.

Chapter 2

Genome-wide discovery of active regulatory elements and transcription factor footprints in *Caenorhabditis elegans* embryos and L1 arrest larvae using DNase-seq

Margaret C. W. Ho and Paul W. Sternberg*

Division of Biology and Bioengineering, Howard Hughes Medical Institute,
California Institute of Technology, Pasadena, CA

*Corresponding author: pws@caltech.edu

Abstract

Deep sequencing of size-selected DNaseI-treated chromatin (DNase-seq) allows high resolution measurement of chromatin accessibility to DNaseI cleavage, permitting identification of *de novo* active CRMs and individual TF binding sites. We adapted DNase-seq to nuclei isolated from *C. elegans* embryos and L1 arrest larvae to generate high-resolution maps of TF binding. Over half of embryonic DNaseI hypersensitive sites (DHS) were annotated in noncoding sequences, with 23% in intergenic, 11% promoter regions and 21% in introns, with similar statistics in L1 arrest data. Noncoding DHS exhibit high evolutionary sequence conservation and are enriched in marks of enhancer activity and transcription. We mined the data to identify putative active CRMs, TF footprints, and 57 novel *cis*-regulatory motifs. We validated noncoding DHS against a previously investigated set of enhancers from *lin-39/ceh-13*, *myo-2, myo-3, hlh-1, elt-2* and *lin-26/lir-1* gene loci and were able to recapitulate 22 of 29 known enhancers and predict novel CRMs. Our DNase-seq data was able to improve predictions of tissue-specific expression compared to motifs alone. Overall, we provide experimental annotation of 26,644 putative CRMs in the embryo containing 55,890 TF footprints, and 15,841 putative CRMs in the L1 arrest larvae containing 32,685 TF footprints. Comparative analysis shows 1,854 condition-specific DHS in L1 arrest, representing putative CRMs of genes targeted by DAF-16 and PHA-4 and which respond to starvation.

Keywords: cis-regulatory modules, gene regulation, enhancers, nematode development, transcription, DNase, hypersensitivity

Introduction

Prior research in metazoans has described several important types of *cis*-regulatory modules (CRMs) such as enhancers, repressors and insulators that can be located far from target genes (reviewed in Noonan and McCallion 2010). Enhancers upregulate expression of target gene(s) in a specific spatiotemporal pattern during development. Repressors restrict expression of target gene(s). Insulators act in a direction-dependent manner to block inappropriate target gene expression and/or block spreading of heterochromatin. These CRMs are thought to function by action of sequence-specific transcription factor (TF) binding which helps recruit RNA polymerase II to the target gene in the case of enhancers or prevent its association in the case of repressors. Enhancers may serve to recruit RNA polymerase II to target genes by physical association with promoters of target genes (reviewed in Bulger and Groudine 2010; Krivega and Dean 2012).

The nematode *Caenorhabditis elegans* has a well-annotated genome, well-studied development and many genetic tools available (Harris et al. 2014; Boulin and Hobert 2012). *C. elegans* provides an excellent case to study transcriptional regulation within a multicellular organism, especially as it is easy to collect synchronized populations of worms in distinct developmental stages (e.g. Baugh et al. 2009).

Rapid establishment of cell fate is transcriptionally regulated during *C. elegans* embryogenesis, as most cell lineages are determined by the 51-cell stage, shortly after eggs have been laid (Edgar 1992). Studies of early embryonic transcription regulation have described a mid-blastula transition that occurs shortly before this period, around the 26-cell stage, when transcription transitions from maternal to zygotic (Baugh et al. 2003) and where embryonic control is underway by the 40-cell stage after initiation of gastrulation. At the end of embryogenesis, the hatched larva has 558 cells (Sulston et al. 1983). When *C. elegans* L1 larvae hatch in the absence of food, they remain in a developmentally arrested state that is resistant to environmental stress (reviewed in Baugh, 2013). Developmental arrest of L1 depends on the insulin-like signaling (IlS) pathway of *C. elegans* (Baugh and Sternberg 2006). Mutants strongly defective in the sole insulin receptor of *C. elegans, daf-2*, are L1 arrest constitutive (Gems et al. 1998), while mutants of the downstream transcriptional effector of the insulin-like signaling pathway, *daf-16*, result in defects in L1 arrest and reduce survival of worms when subjected to starvation (Munoz and Riddle 2003; Baugh and Sternberg 2006). In addition, starvation survival of L1 arrest worms is dependent on the Tor signaling pathway of *C. elegans*, resulting in changes in gene expression mediated by the transcription factor Forkhead/PHA-4 (Zhong et al. 2010). The *C. elegans* embryo and L1 arrest larvae thus provide interesting conditions in which to examine the control of transcription during development.

A number of enhancer CRMs have been characterized in *C. elegans*, of which many are located close (less than 2 kb away) to the promoter of the target gene (e.g. Okkema and Krause 2005). This preponderance of closely-located enhancers is likely due to experiments focusing mostly on promoter-proximal regions of genes. A few studies have identified more distantly located CRMs (reviewed in Gaudet and McGhee 2010). These include AIY-dependent enhancers located in the intron of the neighboring gene or 6kb upstream of the target gene (Wenick and Hobert 2004), the CHE-1 binding site 5kb upstream of *cog-1* (O'Meara et al. 2009) identified through a forward mutagenesis screen, and the TRA-1 repressor element located 6 kb downstream of *egl-1* (Conradt et al. 1999). Studies of the *lin-39/ceh-13* Hox locus have also identified many distant enhancers, such as N7, located 7kb away from its target gene *lin-39*, and N2, N3, N4 enhancers located 18-20kb away from their target *ceh-13 (*Kuntz et al. 2008). Systematic identification of *C. elegans* CRMs as a whole has proved difficult, since most studies have focused on identifying noncoding regions that are conserved on the sequence level and individually testing for functional activity in reporter assays.

ChIP-seq can be used to measure binding of a specific TF of interest to the genome (Robertson et al. 2007; Visel et al. 2009; Ren et al. 2000). ChIP-seq in *C. elegans* (e.g. Baugh et al. 2009; Kuntz et al. 2012; Araya et al. 2014; Gerstein et al. 2010; Zhong et al. 2010; Niu et al. 2011) has generated data can be mined to identify CRMs regulated by TFs of interest; nevertheless a general view of simultaneous TF binding in the genome that allows the discovery of CRMs and regulatory motifs *de novo*, without prior knowledge of TFs and need of specific antibodies or GFP-tagging, is desirable in *C. elegans*.

Hypersensitivity to cleavage by DNaseI has been long known as a property of active *cis*-regulatory regions (Gross and Garrard 1988). CRMs of the β-globin locus, including the locus control region and insulators, were discovered through DNaseI hypersensitivity assays (Fraser et al. 1993; Tuan et al. 1985). Studies in yeast, mammals, *Drosophila,* and *Arabidopsis* have utilized deep sequencing of DNaseI-treated chromatin to map protein-DNA interactions *de novo* (Hesselberth et al. 2009; Boyle et al. 2011; Thomas et al. 2011; Song et al. 2011; Sullivan et al. 2014). In addition to identifying DNaseI-hypersensitive (DHS) regions that may act as putative CRMs, deep sequencing allows sufficient resolution to identify shorter sequences within DHS protected from DNaseI cleavage. These protected regions or footprints represent putative TF binding sites. These DHS and footprint regions can be computationally analyzed to discover novel regulatory motifs. While a previous study looked at DNaseI hypersensitivity in *C. elegans* young adults by hybridizing to DNA tiling arrays and was able to identify 7095 large DNaseI hypersensitive regions that ranged from 46 bp to 754 bp long, the data did not give sufficient resolution to identify TF footprints and it was not clear whether the authors had indeed located known CRMs (Shi et al. 2009).

In this study we describe the mapping of *cis*-regulatory protein-DNA binding within the *C. elegans* genome in embryos and L1 arrest larvae using deep sequencing of DNA extracted from DNaseI-treated chromatin. Our studies identify 41,825 and 23,674 reproducible DHS peaks in embryos and L1 arrest larvae, respectively, using samples that on average comprise 30 million Illumina HiSeq 50-76bp single reads, giving 15X coverage of the 100 million base pair *C. elegans* genome.

Results

A DNase-seq method for *C. elegans*

To identify DNaseI hypersensitivity sites in *C. elegans*, we performed DNaseI treatment on three and four high-quality biological replicate samples of embryos and L1 arrest larvae, respectively. We then isolated DNA fragments less than 500bp that represent chromatin regions most accessible to DNaseI cleavage (Figure 1.1A; see methods for details). QPCR was used to identify DNaseI treatment conditions that resulted in the highest enrichment of regulatory regions in the DNase-seq sample, using primers designed against conserved known enhancers from the *lin-39/ceh-13* Hox cluster (Kuntz et al. 2008) and negative control regions lacking any known regulatory activity (see Methods). DNase-seq samples were sequenced to 15X coverage of the *C. elegans* genome and the read data were used to identify regions with increased hypersensitivity across 150 bp consecutive nucleotides using HOTSPOT DNaseI peak-calling software (John et al. 2011) (Figure 1.1.C). Raw peak calls were filtered using the irreproducibility discovery rate (IDR) framework developed for ENCODE, which uses a non-parametric copula mixture model to filter peaks into reproducible or irreproducible categories (Li et al. 2011). Peaks were selected on combination of rank or score and consistency across replicates to yield 41,825 embryonic and 23,674 L1 arrest DNaseI hypersensitive sites (DHS).

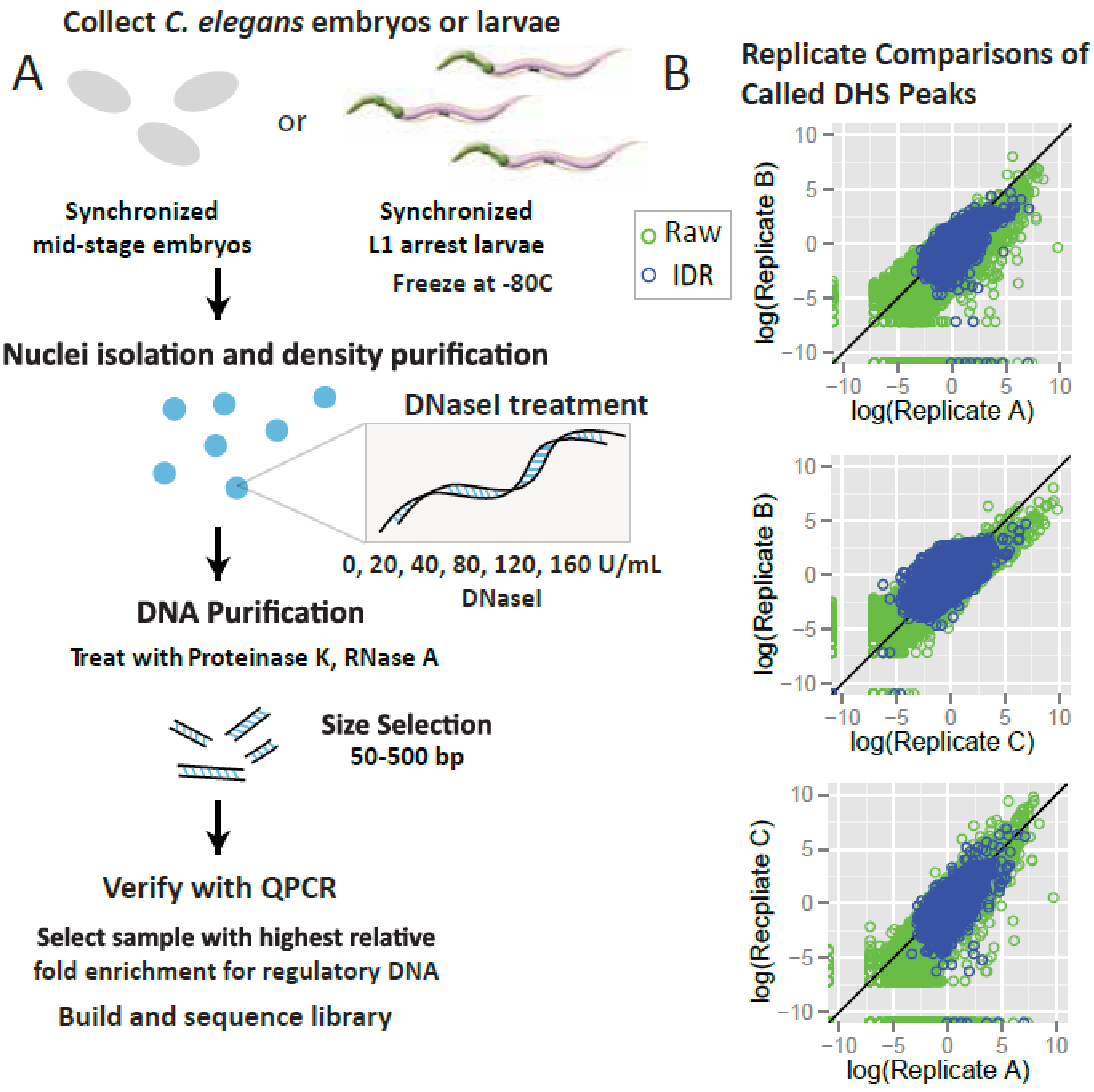
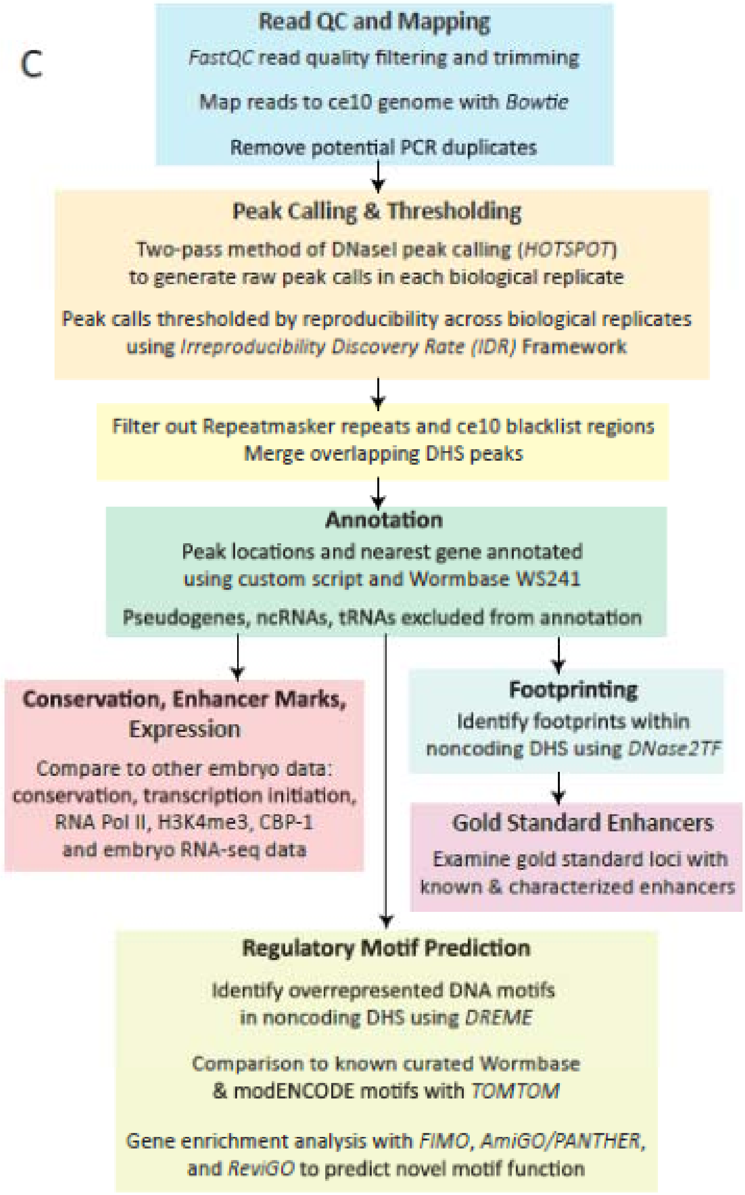


Figure 1.1 Experimental method and reproducibility

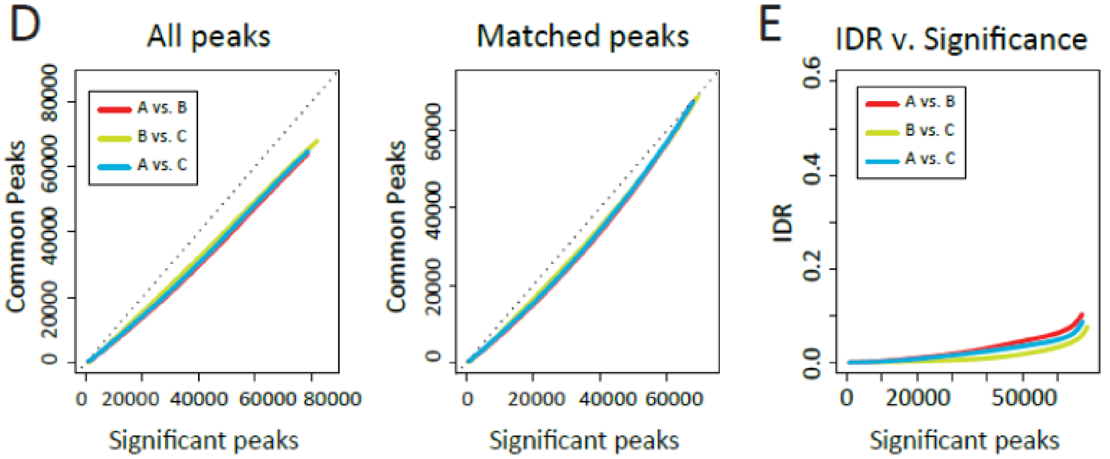
A. Experimental Method. Wild-type N2 worms were grown synchronously for at least two generations. Embryos at around the 40-cell stage or L1 arrest larvae were collected and frozen at -80C. Freeze-thaw cycles in a nuclei purification buffer and a Dounce homogenizer were used to isolate nuclei. Nuclei (blue) were purified by spinning on Optiprep density gradient medium and visualized with DAPI (see Methods). Nuclei were divided into aliquots and DNaseI treatment was performed at 0, 20, 40, 80, 120, 160 U/mL DNaseI concentration. Resulting DNA was isolated by treatment with Proteinase K, RNaseA, column purification, and size selection by gel extraction. DNA was quantified using Qubit fluorescence. Enrichment in regulatory regions was verified using

QPCR designed against *lin-39/ceh-13* Hox enhancers. The sample with highest relative fold enrichment for regulatory regions was selected for library construction and sequencing.

B. Reproducibility of read coverage over DHS in embryo biological replicates. Pair-wise comparisons of embryo biological replicate DNase signal across all identified Raw (green) and IDR-filtered (blue) DHS show good reproducibility. Signal is measured in log2 of reads per base pair. Black diagonal line represents the ideal case of perfect reproducibility.


C
Read QC and Mapping
FastQC read quality filtering and trimming
Map reads to ce10 genome with Bowtie
Remove potential PCR duplicates
Peak Calling & Thresholding
Two-pass method of DNaseI peak calling (HOTSPOT) to generate raw peak calls in each biological replicate
Peak calls thresholded by reproducibility across biological replicates using Irreproducibility Discovery Rate (IDR) Framework
Filter out Repeatmasker repeats and ce10 blacklist regions
Merge overlapping DHS peaks
Annotation
Peak locations and nearest gene annotated using custom script and Wormbase WS241
Pseudogenes, ncRNAs, tRNAs excluded from annotation
Conservation, Enhancer Marks, Expression
Compare to other embryo data: conservation, transcription initiation, RNA Pol II, H3K4me3, CBP-1 and embryo RNA-seq data
Footprinting
Identify footprints within noncoding DHS using DNase2TF
Gold Standard Enhancers
Examine gold standard loci with known & characterized enhancers
Regulatory Motif Prediction
Identify overrepresented DNA motifs in noncoding DHS using DREME
Comparison to known curated Wormbase & modENCODE motifs with TOMTOM
Gene enrichment analysis with FIMO, AmiGO/PANTHER, and ReviGO to predict novel motif function

C. Computational analysis workflow. Italics indicate the software packages used (see Methods).



D. Biological replicates show reproducibility of matched peaks. Comparison between number of common peaks and significant peaks in pairs of biological replicates when all raw peaks are assessed together (All Peaks) or peaks matching in replicates (Matched Peaks). Pair-wise comparisons of biological replicates: A vs. B (red), B vs. C (green) and A vs. C (blue) are shown.

E. Observed relationship between irreproducible discovery rate (IDR) threshold and number of significant peaks called in biological replicates. 69,155 reproducible embryo DHS peaks remained after IDR filtering using threshold 0.1. Filtering for ce10 blacklist regions and repeat regions resulted in 41,825 embryo DHS peaks (see Appendix Figure 2.5 B-C for L1 arrest data).

Regions with high enrichment of reads in one DNase-seq replicate are generally observed to have high enrichment in other biological replicates from the same condition (Figure 1.1B shows embryo data, see Appendix Figure 2.5 B-C for L1 arrest data). Comparing raw peaks from HOTSPOT to DHS peaks filtered by IDR, we observed that filtering by IDR successfully removes peaks with low read coverage and some very high scoring peaks that did not pass replicate consistency requirements. We observe a robust correlation between numbers of significant peaks and common peaks at most levels of peak calling for the overlapping peaks (the subset of peaks that overlap in replicates) compared to total peaks (Figure 1.1C). The irreproducible discovery rate (IDR) compared to the statistical significance of the peaks shows a shallow slope, indicating that we are able to call a large number of significant peaks at a low IDR (Figure 1.1D-E) (Li et al. 2011; Landt et al. 2012).

After comparison with the WS241 *C. elegans* genome annotation, we found that 26,644 and 15,841 of these embryonic and L1 arrest DHS, respectively, overlap with noncoding regions of the genome and represent putative active CRMs in these conditions. To identify regions within noncoding DHS that could be footprints of TF binding sites, we searched for signatures of TF footprints (protection from DNaseI cleavage and positive-to-negative strand shift in reads) using DNase2TF, which has been shown to perform significantly better and recover more accurate peaks compared to other algorithms such as Wellington and DNaseR (Sung et al. 2014). We were thus able to discover 55,890 and 32,685 putative DNaseI TF footprints within these noncoding DHS in the *C. elegans* embryo and L1 arrest, respectively. Comparing the embryo and L1 arrest datasets, we observe 1,854 condition-specific DHS in L1 arrest harboring 2,964 TF footprints.

DHS peaks are most abundant in noncoding regions and DNaseI hypersensitivity correlates with expression

Annotation of peaks with WormBase WS241 gene models revealed that DHS peaks are most abundant (55%) in noncoding regions (Figure 2.2A). Less than half

(45%) occur within exons, which is expected as DHS were found throughout exons of actively transcribed genes (Mercer et al. 2013). Above half were observed in noncoding regions, with 23% in intergenic regions, 11% in promoters (defined as less than 300 bp of exon start), and 21% in introns. Noncoding DHS residing in introns, intergenic and promoter regions, by being accessible to DNaseI, may thus represent candidate CRMs. Similar statistics were observed in L1 arrest larvae with 67% of DHS in noncoding regions of the genome; with 28% in intergenic regions, 13% in promoters, and 27% in introns (Appendix Figure 2.5A).

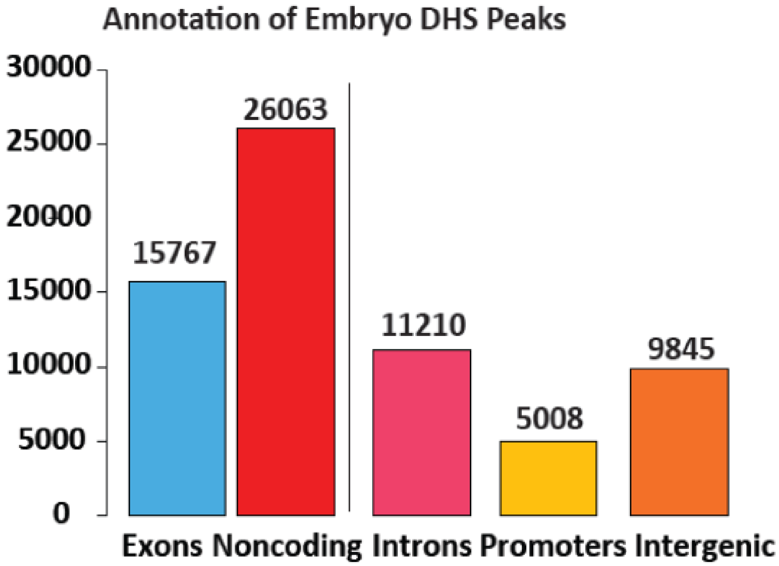


Figure 2.2.A Genomic location of embryo DHS shows abundance of noncoding DHS. Embryo DHS were annotated according to position relative to WormBase WS241 protein-coding genes: exons (blue) and noncoding (red). Noncoding DHS are further subdivided into introns (pink), promoter (defined as less than 300bp 5' of ATG; yellow) and intergenic (orange) regions.

Most genes exhibit a uniform distribution of reads over the gene body and surrounding sequence with an average of 20 mapped reads per bp (Figure 2.2B). However, about 9% of genes exhibit much higher read coverage and show a pattern of three peaks of read enrichment reaching as high as 120 mapped reads per bp. These peaks correspond to the 5' upstream region, gene body, and 3' downstream region. We observe that this subset of genes with higher and tri-modal patterns of read enrichment are 66% more highly expressed in embryo than genes with lower and uniform pattern (two-sided Kolmogorov-Smirnov (KS) test, p = 1.1x10^{-8}) (Figure 2.2B).

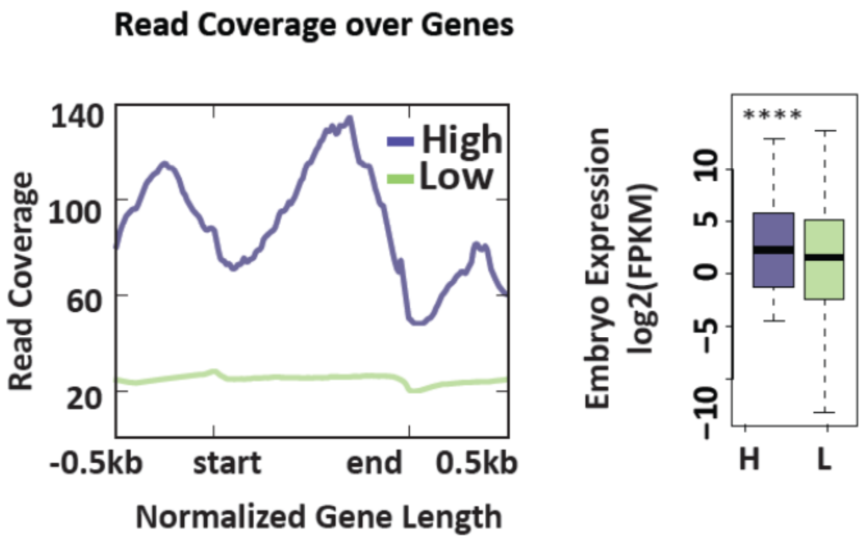


Figure 2.2.B Protein-coding genes with higher DNase accessibility are more highly expressed. Read coverage (total DNaseI signal across biological replicates) was measured for length-normalized WormBase WS241 protein-coding genes and 1kb of surrounding sequence. *k*-means clustering of genes by read coverage was used to find genes with higher (High) and lower read coverage (Low). Embryo expression (from Zhong et al. 2010) which is measured in log2 of fragments per kilobase of exon per million fragments mapped (FPKM) was compared between higher read coverage (H) vs. lower (L) read coverage genes.

Noncoding DHS are twice as conserved as expected by random chance and DNaseI hypersensitivity is strongly correlated with sequence conservation

Comparing median DNaseI hypersensitivity and sequence conservation in a 2kb region surrounding noncoding DHS, we find that levels of DNaseI hypersensitivity strongly correlate with sequence conservation on a per nucleotide basis (Figure 2.2C). Both DNaseI hypersensitivity and sequence conservation peak at the midpoint of noncoding DHS and are centered in a 400bp region surrounding the site. If we compare with levels of sequence conservation of known enhancer CRMs such as those in the *lin-39/ceh-13* Hox complex we find that median phyloP sequence conservation was 0.543 for true positive enhancers in Kuntz et al. (2008), suggesting a typical size for CRMs of *C. elegans* of about 200bp. True negatives in the same study showed phyloP sequencing conservation of about 0.43 (see Methods for details). A typical size noncoding DHS of 150bp thus captures the bulk of both the DNaseI hypersensitivity and sequence conservation. DHS peaks in noncoding regions are on average twice as conserved on a per nucleotide basis than expected by chance (two-sided KS test, $p < 3 \times 10^{-16}$).

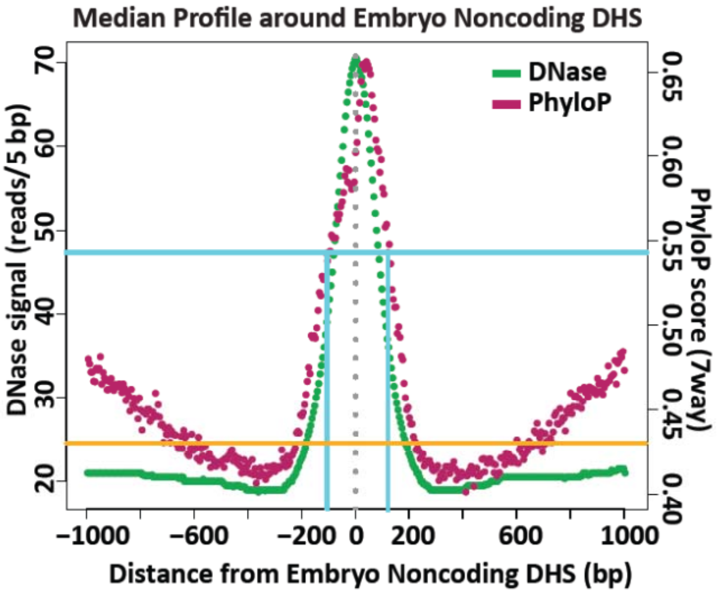


Figure 2.2C. DNase accessibility and phyloP sequence conservation both reach a peak in embryo noncoding DHS. Median DNase signal (green; measured in 5bp windows) and phyloP score (pink; 7 way) are measured across 2kb of sequence surrounding embryo noncoding DHS, and peak in the embryo noncoding DHS at 70.5 reads in a 5 bp window and at 0.66 for phyloP sequence conservation. The level of phyloP conservation for known true positive *lin-39/ceh-13* enhancers is 0.54 (blue) and for negative control non-enhancer regions is 0.43 (orange line; see Methods).

Noncoding DHS are highly enriched in marks of enhancer activity and transcription

Embryo DHS peaks are significantly enriched in embryonic sites of transcription initiation (TSS) (4.2 fold, two-sided KS test, $p < 3 \times 10^{-16}$) (Chen et al. 2013) and overlap many annotated noncoding RNAs. The average DNase profile of these TSS shows enrichment of read coverage in the surrounding 400bp sequence, demonstrating high accessibility to DNaseI cleavage, with even higher accessibility within the noncoding DHS themselves (Figure 2.2D). Comparison with data from a different study using GRO-cap sequencing to identify *C. elegans* TSS also showed that embryo and L1 arrest larvae are 7.9 and 7.7-fold enriched, respectively, in stage-matched sites of transcription identified by this study (Kruesi et al. 2013) (two-sided KS tests, $p < 3 \times 10^{-16}$).

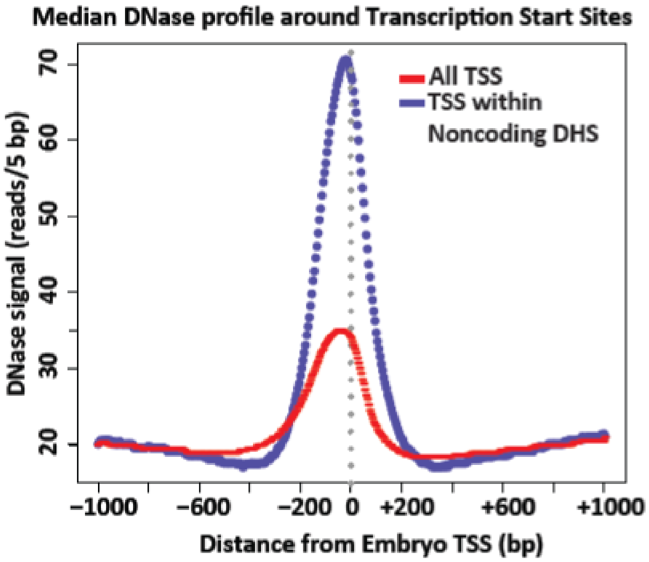


Figure 2.2D Median DNase signal peaks in *C. elegans* transcription start sites and shows 5' bias. Median DNase signal (measured in 5bp windows) is measured in 2 kb of sequence surrounding embryo transcription start sites (TSS) (locations from Chen et al. 2013), with the orientation placed according to the direction of transcription (5' to 3' is shown from left to right) and the center of the TSS indicated at 0 bp (gray dotted line). DNase signal peaks at All TSS (red) and at TSS within noncoding DHS (purple) and shows strongest DNase accessibility just 5' to the TSS.

Comparing embryonic noncoding DHS to stage-matched H3K4me3 ChIP-seq and *C. elegans* p300 homolog CBP-1 ChIP-chip peaks from modENCODE, they are significantly enriched in marks associated with potential enhancer regulatory activity in eukaryotic genomes (2.8 fold, $p < 3 \times 10^{-16}$) (Heintzman et al. 2007). Also, 199 (65%) of 304 high occupancy target (HOT) genomic core regions bound by fifteen or more TFs tested by modENCODE overlap with embryo DHS (5.1 fold, $p < 3 \times 10^{-16}$; Gerstein et al. 2010). 57% of RNA polymerase II binding regions identified in early embryos by modENCODE overlap our observed embryo DHS (1.38 fold, $p < 3 \times 10^{-16}$) (Gerstein et al. 2010).

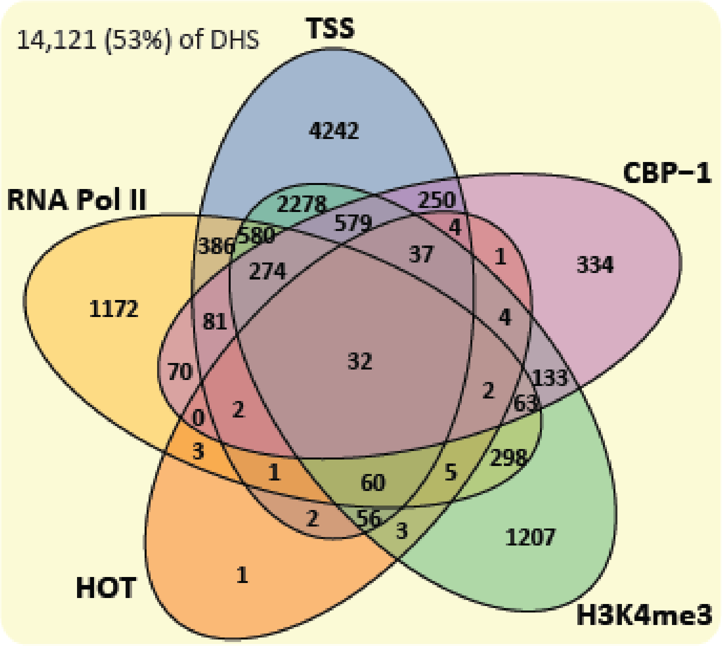


Figure. 2.3A Half of embryo noncoding DHS coincide with transcription start sites (TSS), histone marks, CBP-1, and HOT regions. 47% of noncoding DHS with marks of enhancer activity such as RNA Polymerase II (RNA Pol II; yellow), transcription start site (TSS; blue), CBP-1 (pink), H3K4me3 (green) observed in embryos and modENCODE high occupancy transcription factor regions (HOT; orange). TSS data are from Chen et al. (2013) and remaining data are from modENCODE (Gerstein et al. 2010).

Nearly half (46%, 12,160) of noncoding DHS overlap with one or more marks of transcription (initiation sites, CBP-1 transcriptional co-activator, RNA polymerase II, H3K4me3 histone marks) or high TF occupancy (modENCODE HOT regions) from stage-matched samples (Figure 2.3A). 14,484 (54%) noncoding DHS do not overlap with any such marks. Of noncoding DHS that do overlap with these marks, most (57%, 6,956) overlap with just one mark, while 3,424 overlap with two marks, 1,373 overlap with three marks, 375 overlap with four marks, and 32 overlap with five marks. Genes associated with noncoding DHS possessing one or more marks are on average 8.9-fold more highly expressed in embryos compared to genes with noncoding DHS lacking any marks *(*p $< 2.2x10^{-16}$ two-sided KS test) (Figure 2.3B). Moreover, genes associated with embryo noncoding DHS overlapping with greater numbers of marks correlates with increased embryonic expression, up to three marks (5.1-fold higher expression compared to one mark, $p < 3 \times 10^{-14}$).

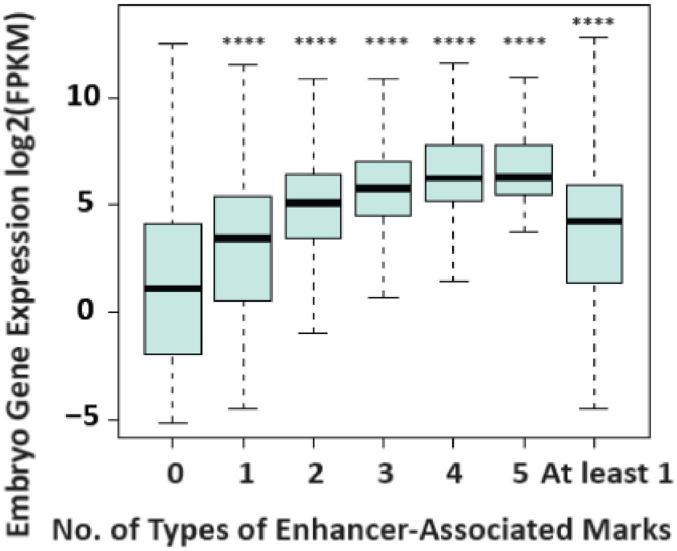


Figure 2.3.B Genes associated with any noncoding DHS harboring enhancer-associated marks are 9-fold more highly expressed than those with DHS lacking any marks. Genes near embryo noncoding DHS with any number of marks (at least one, two, three, four, or five type(s) of enhancer-associated mark) exhibit, on average, 8.9-fold higher levels of embryo expression (measured in log2 of FPKM, data from Zhong et al. 2010) compared to those with embryo noncoding DHS lacking marks ($p < 3 \times 10^{-16}$, two sided KS test). With each additional mark, median observed expression increases, up to three marks (5.1-fold higher expression compared to one mark, $p < 3 \times 10^{-14}$). No significant difference is observed between genes near noncoding DHS with three, four or five marks.

Presence of at least one noncoding DHS peaks is correlated with gene expression

10,890 (53%) protein-coding genes were assigned at least one DHS nearby according to our annotation that assigned the nearest gene to each DHS (Appendix Figure 2.3B). 9,822 (47%) protein-coding genes did not possess nearby noncoding DHS. The presence of at least one embryo noncoding DHS near a gene is associated with 4.5-fold higher embryo expression compared to genes lacking DHS (*p* < 3 x 10^{-16}, two-sided KS test; Figure 2.3C). There is 54% increase in embryo expression between one and two noncoding DHS near a gene and 44% increase from two to three (two-sided KS tests, p < 3x10^{-6} and p < 0.007, respectively). Additional increases in noncoding DHS beyond three DHS per gene do not increase expression. Genes with DHS that do not have any marks are still 2.3-fold more expressed compared to genes lacking any DHS (*p* < 3 x 10^{-16}, two-sided KS test) (Figure 2.3D).

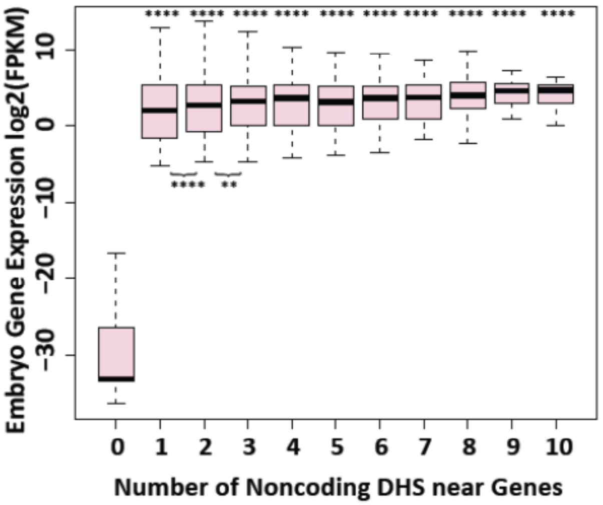


Figure 2.3.C The presence of at least one embryo noncoding DHS near a gene is correlated with 4.5-fold higher embryo expression. Comparison of embryo expression between genes with zero and one to ten noncoding DHS peaks shows that the presence of at least one embryo noncoding DHS is associated with 4.5 fold higher embryo expression compared to none (p<3x10^{-16}). Embryo expression, measured as log2 of the fragments per kilobase of exon per million reads mapped (FPKM; data from Zhong et al. 2010) increases 54% from one to two embryo noncoding DHS (p < 3x10^{-6}) and 44% from two to three (p < 0.007). However further increases in DHS are not correlated with expression.

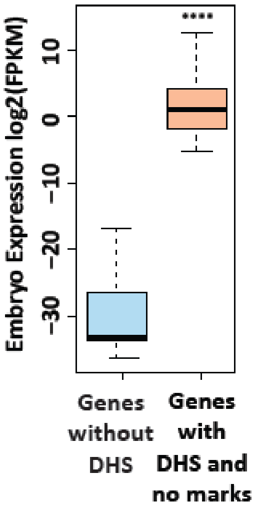


Figure 2.3.D. Genes associated with embryo noncoding DHS and lacking marks are still twice as highly expressed as genes without DHS. Genes with embryo noncoding DHS lacking enhancer-associated marks (orange) show 2.3-fold higher embryo expression compared to genes lacking DHS (blue; $p < 3 \times 10^{-16}$).

Within noncoding embryo DHS peaks, we identified 55,890 potential TF binding sites (TFBS) using DNase2TF (Sung et al., 2014). Regions between 6-40 bp within noncoding DHS that showed less coverage than neighboring nucleotides and exhibited a strand shift in mapped reads characteristic of TF binding were identified as potential TF footprints using an FDR cutoff of 0.05. Most (21857, 82%) of noncoding DHS possess detectable footprints, whereas 18% (4787) do not (Appendix Figure 2.3A). This pattern largely holds true even when we subdivide noncoding DHS according to overlap by marks (TSS, H3K4me3, RNAPII, CBP-1 and HOT) (Appendix Figure 2.3A). We did not detect any difference in expression between genes associated with DHS that do have detectable footprints and those that do not. These data fit the model that DHS peaks represent potential CRMs, with many of the hallmarks of CRM activity including sequence conservation, active transcription, H3K4me3, and TF occupancy.

Embryo noncoding DHS peaks and footprints coincide with many known CRMs

To investigate whether the locations of previously investigated enhancers can be identified by our DNase-seq method, we examined several *C. elegans* genetic loci harboring known enhancers, particularly those of the *lin-39/ceh-13* Hox locus and genes active in embryos. The genetic locus containing Hox anterior-posterior patterning genes *ceh-13* and *lin-39* and lincRNA *linc-57* is known to harbor numerous enhancers that are as far away as 20 kb from target genes. A previous study identified enhancers using MUSSA (multi-species sequence analysis using ungapped transitive alignments) to find conserved sequences across several *Caenorhabditis* species and characterized their expression patterns in transgenic reporter assays (Kuntz et al. 2008). Within these large enhancer regions, ranging from 591 bp to 1120 bp, they also identified smaller 15-33 bp MUSSA conserved sub-regions.

We observed several noncoding DHS in our embryo data that overlap these previously identified *lin-39/ceh-13* enhancers (Figure 2.4 A, B; Appendix Figure 2.1A). Specifically, observed noncoding DHS peaks pinpointed core MUSSA conserved regions of seven (N1, N2, N3, N4, N8, N10, and N11 enhancers) of the nine enhancers previously identified. We also observed noncoding DHS within two “false negative” regions (I4 and I8) able to drive expression in the Kuntz et al. study (2008). Regarding potential false positives, we found one noncoding DHS in N5 that does not appear to drive reporter expression (Figure 2.4B). We also observe TF footprints within several noncoding DHS. Footprints were detected within noncoding DHS peaks corresponding to six (N1, N2, N3, N8, N10 and N11) of the nine enhancers (Appendix Figure 2.1A). We also find footprints in noncoding DHS found within the I4 and I8 enhancers reported as a “false negative” by Kuntz et al. (2008; Figure 2.4B) Surprisingly, while some of these enhancers do not apparently drive reporter expression until later in development, our data raise the possibility that the chromatin surrounding these regions is already accessible to DNaseI in the early embryo. These examples include N1, which drives expression in L4 through adulthood, but which we observe to be hypersensitive in embryos (Kuntz et al. 2008) (Figure 2.4A).

Figure 2.4 Noncoding DHS coincide with known CRMs
Total DNaseI signal (red) from both strands of embryo read data shown, as well as individual DNaseI signal from positive (orange) and negative (green) strands. Noncoding DHS (light blue boxes) and all DHS (medium blue boxes) and TF footprints (dark blue boxes) detected. Additional tracks are *C. elegans* RefSeq genes (black boxes with arrows), noncoding transcripts (brown boxes), and phyloP conservation (very dark blue). Other tracks (if shown) include TSS (dark orange boxes; Chen et al. 2013), RNAP II ChIP-seq (red boxes), H3K4me3 (pink) and CBP-1 (lavender boxes), ChIP-chip and HOT regions (yellow boxes) from modENCODE embryo data. MULTIZ conserved elements (magenta boxes) and Repeatmasker elements (black boxes) are also shown.

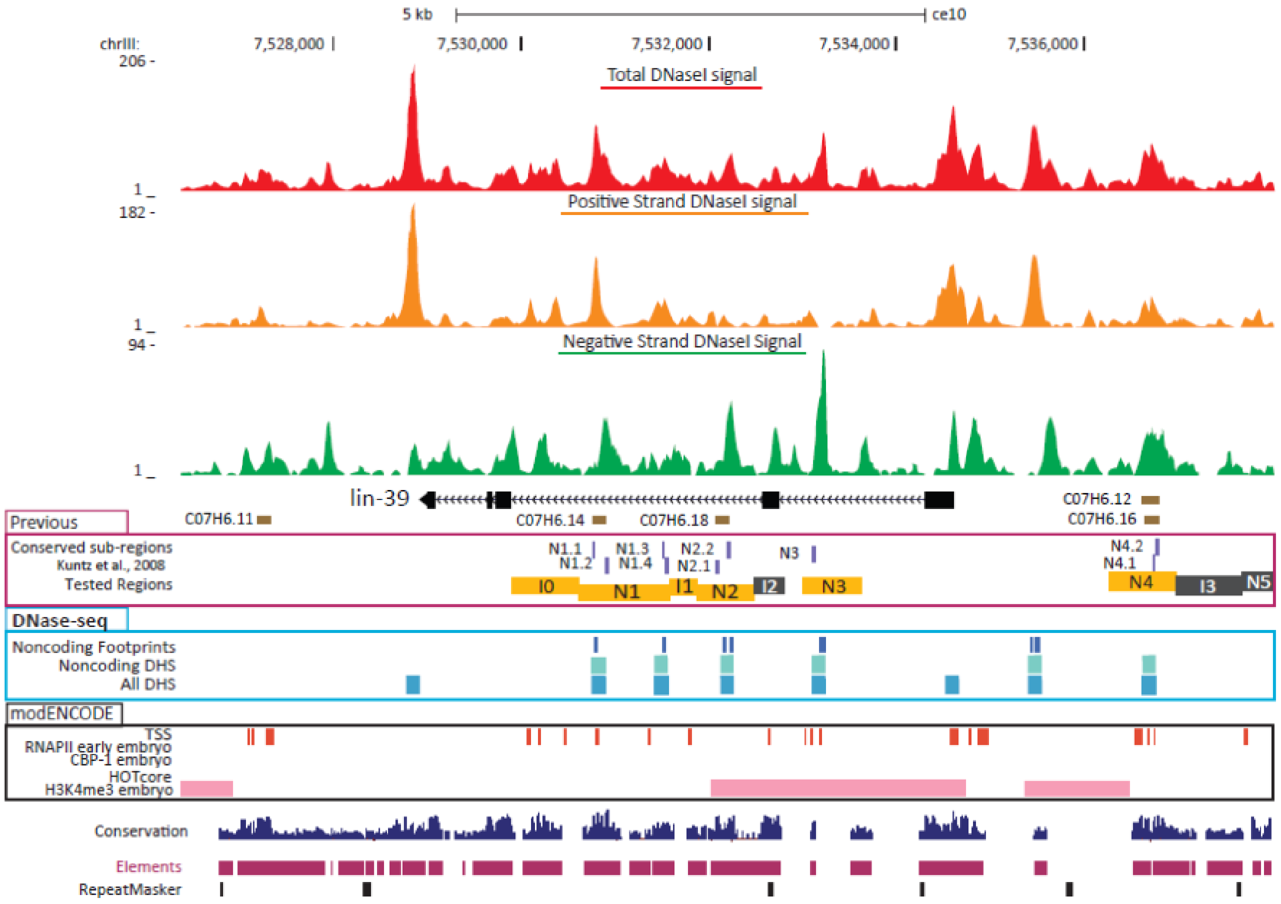


Figure 2.4A. N1, N2, N3, and N4 enhancers of *lin-39* are all recapitulated by embryo noncoding DHS. Conserved MUSSA regions (dark purple boxes), regions from Kuntz et al. 2008 that drove reporter expression (yellow boxes), regions that did not drive reporter expression (dark gray boxes) are indicated. Two noncoding DHS peaks are detected in N1 intronic enhancer of *lin-39* overlapping with N1_1 and N1_2 MUSSA conserved sub-regions and N1_3 and N1_4 MUSSA regions, respectively. One noncoding DHS peak is detected in N2 and overlaps with N2_1 and N2_2 MUSSA regions, and another overlaps with the conserved MUSSA region of the N3 enhancer in the first intron of *lin-39*. Another noncoding DHS peak overlaps with N4_1 and N4_2 conserved MUSSA regions of N4 enhancer of *lin-39*. TF footprints are detected overlapping MUSSA regions N1.1, N1.3, N2.1 and in enhancers N1-N3. A noncoding DHS is detected in the *lin-39* promoter, along with TF footprints.

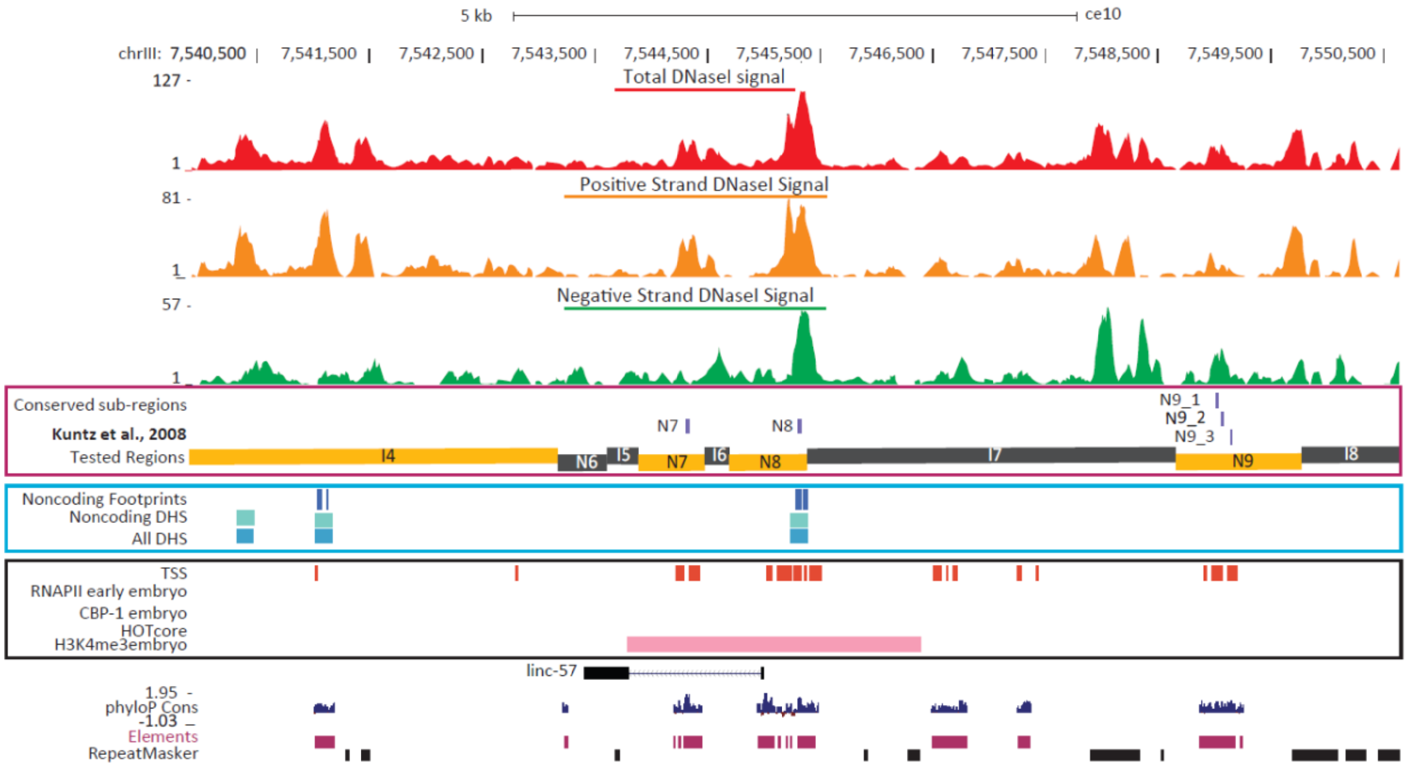


Figure 2.4.B I4 ("False negative" in the Kuntz study) detected, along with N8 enhancer/promoter of *linc-57* long intergenic noncoding RNA. Two noncoding DHS were detected in I4 region that drives transgenic reporter expression (reported as "false negative" in Kuntz et al. 2008). Of these, the noncoding DHS that overlaps a MULTIZ conserved element exhibits a TF footprint and overlaps an observed TSS. Another noncoding DHS overlaps with N8 in 5' promoter region of *linc-57* and its conserved MUSSA region, and also harbors TF footprints. Enhancers N7 and N9 were not detected.

We then examined well-studied gene loci representing major tissue regulators or structural genes expressed during embryonic development. The epithelial differentiation factor *lin-26* begins to be expressed in early embryos in all epithelial cells of the ectoderm and is responsible for somatic gonad differentiation (Landmann et al. 2004). *elt-2* is an intestinal terminal differentiation TF (McGhee et al. 2009) whose expression first appears in mid 2E-cell stage (Fukushige et al. 1998). *myo-3* is a myosin heavy chain gene that begins expression during the pre-comma stage and is eventually expressed in all muscle cells outside of the pharynx (Fox et al. 2007; Okkema et al. 1993). *myo-2* is a myosin heavy chain gene whose expression begins later in the 2-fold stage embryo and is expressed in all pharyngeal muscle cells (Okkema and Fire 1994; Gaudet and Mango 2002). These embryonic expression patterns led us to expect that some of their CRMs would exhibit DNaseI hypersensitivity in embryos.

A previous study identified sequences required for proper expression of *lin-26* upstream of the gene in an 11kb region spanning the first intron of *lir-1* (Landmann et al. 2004). We are able to detect at least one noncoding DHS and multiple footprints in each of the five previously described enhancer regions corresponding to the A+B (Late), C+D (Late), E (Intermediate), F+G (Intermediate) and H (Early) enhancers. Of these enhancers, A+B (Late) and C+D (Late) are bound and regulated by PHA-4. The noncoding DHS and footprints we detect in these two enhancers correspond to the locations of PHA-4 ChIP-seq peaks previously observed in embryos (Zhong et al. 2010) (Figure 2.4C).

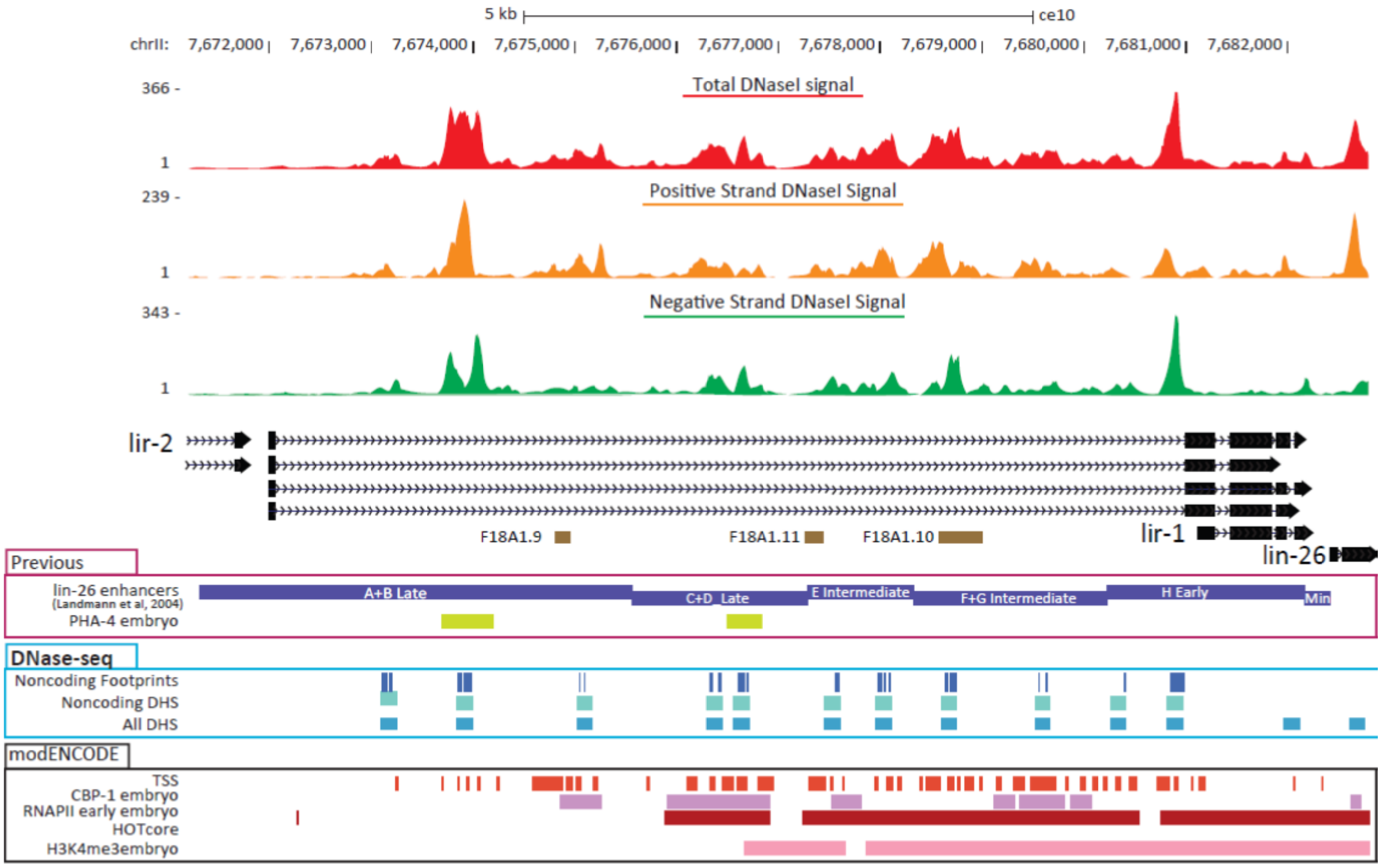


Figure 2.4.C All five known enhancers of *lin-26* in 11kb first intron of *lir-1* are recovered, each harboring at least one embryo noncoding DHS and footprint. Multiple noncoding DHS detected in 11kb region upstream *lin-26*, in the first intron of *lir-1*, which harbors many enhancers active in embryos (purple): A+B Late, C+D Late, E Intermediate, F+G Intermediate, and H Early (Landmann et al. 2004). Noncoding DHS and footprints were detected in each known enhancer, and in the case of A+B and C+D Late, the noncoding DHS and footprints overlap with PHA-4 binding sites (light green; Zhong et al. 2010).

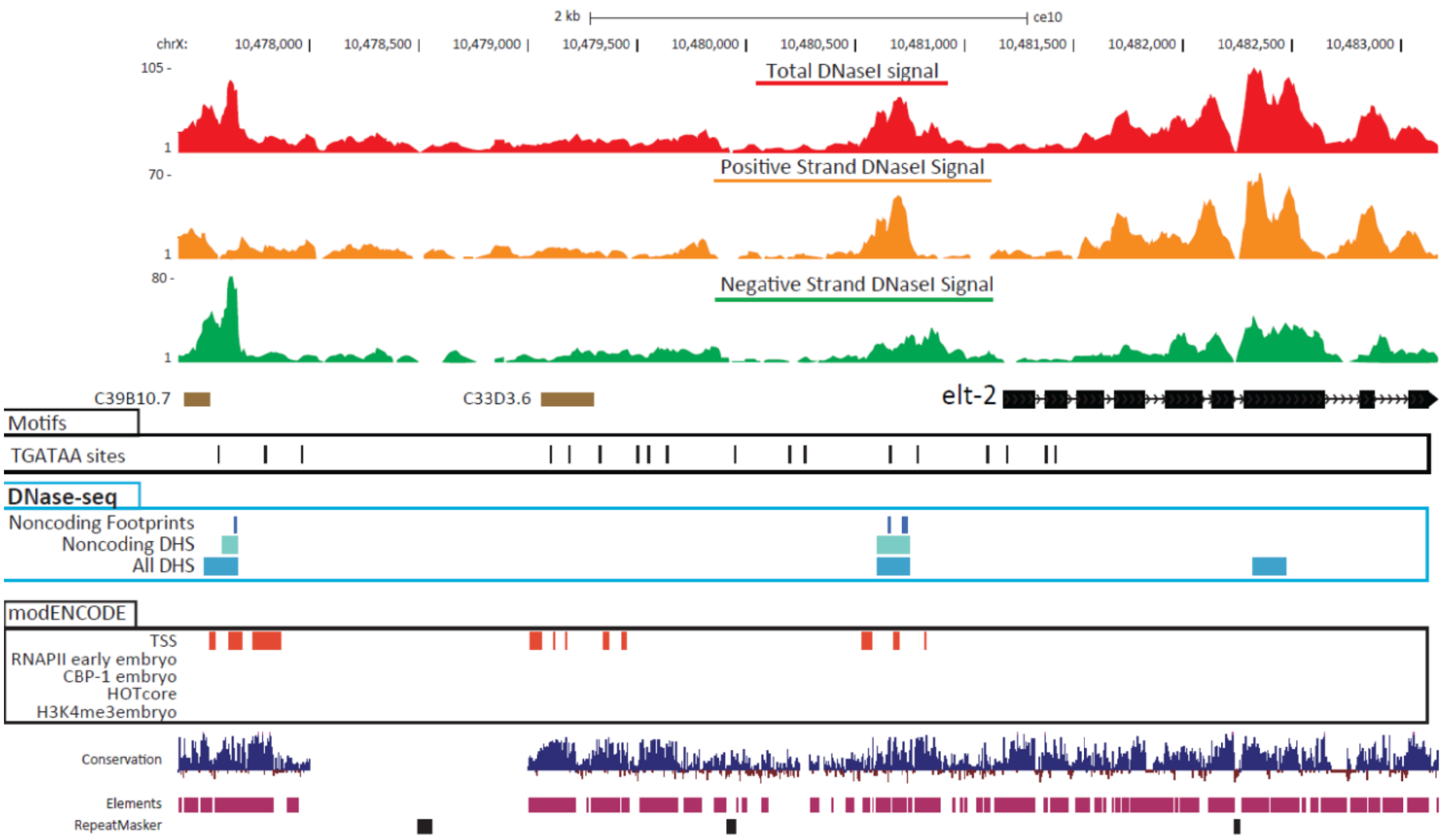


Figure 2.4.D Two *elt-2* enhancer regions detected including regions with *elt-2* ChIP-seq signal. Noncoding DHS detected upstream of *elt-2* in two known *elt-2* CRMs (Fukushige et al. 1999) including promoter where ELT-2 itself binds (E. Osborne Nishimura and J. McGhee, personal communication). Two footprints are detected in the *elt-2* promoter, with one overlapping ELT-2 binding sites (TGATAA sites; black). A TF footprint is also found in the distal noncoding DHS. Known CRM overlapping the C33D3.6 noncoding transcript was not detected.

The promoter and 5' upstream region of *elt-2* shows several DHS that coincide with *elt-2* ChIP-seq peak data (E. Osborne Nishimura and J. McGhee, personal communication). Studies have shown that *elt-2* is auto-regulated by binding to its own promoter in embryos (Fukushige et al. 1999). In addition, two TF footprints are detected within the distal enhancer and promoter of *elt-2* (Figure 2.4D).

Regulation of *myo-2* expression by its A, B, and C sub-elements has been extensively dissected (Okkema and Fire 1994). We observe one noncoding DHS and associated footprint that overlap with the minimal *myo-2* promoter bound by PHA-4 in embryos, corresponding to a pan-pharyngeal element (Kalb et al. 1998). Another noncoding DHS detected in our study overlaps with the B and C sub-elements that drive pharyngeal expression in reporter assays (Figure 2.4E). In particular, we detect a putative TF footprint in the sub-element C which binds PHA-4 (Kalb et al. 1998; Okkema and Fire 1994) through genetic evidence and PHA-4 ChIP-seq data (Zhong et al. 2010). Noncoding DHS peaks are observed in both the first intron and upstream region of *myo-3*, coinciding with three enhancers MC186, MC197, and MC165 previously reported to drive reporter expression (Okkema et al. 1993). Noncoding DHS that coincide with these enhancers possess several TF footprints (Appendix Figure 2.1B).

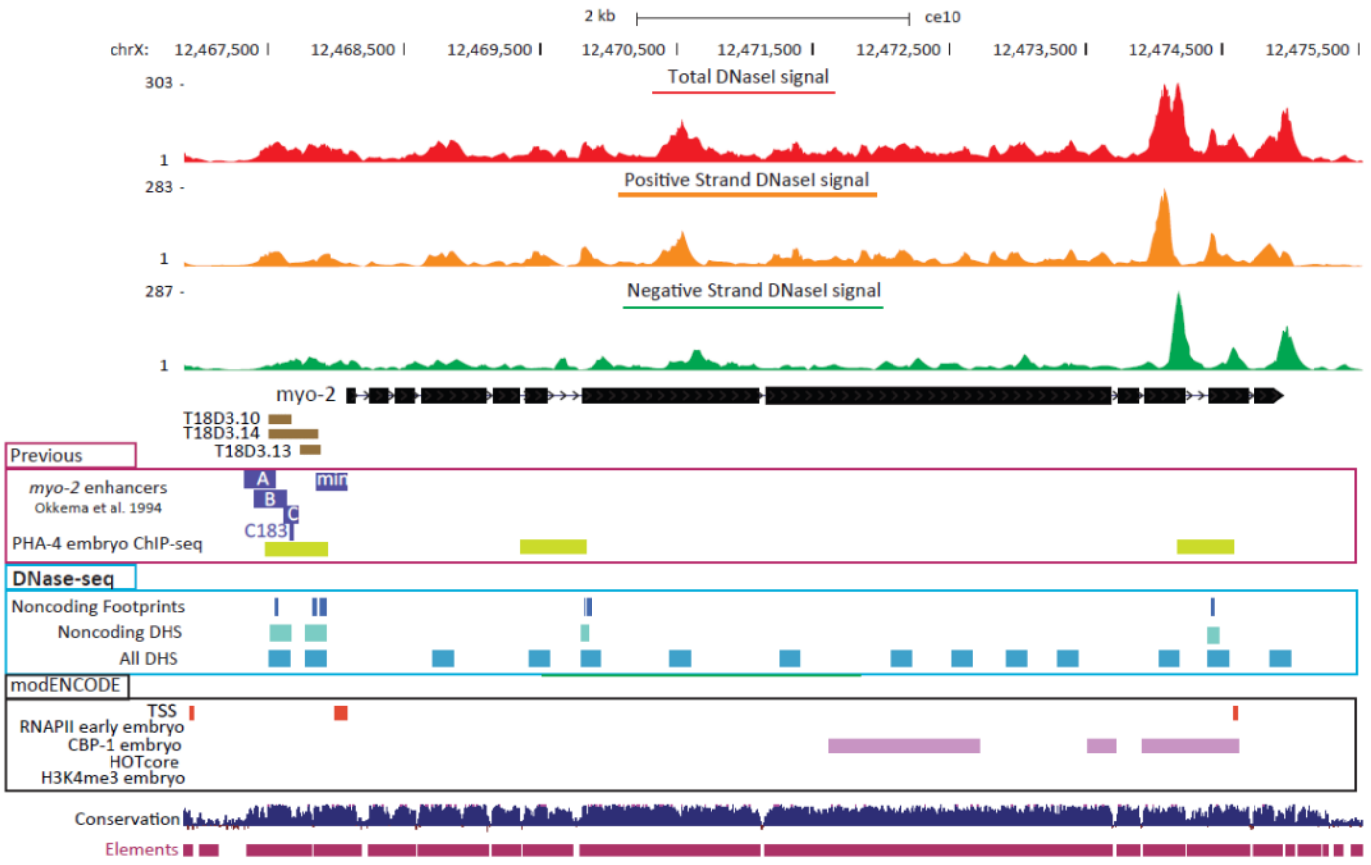


Figure 2.4.E. ***myo-2*****, regulated by PHA-4, possesses multiple upstream and intron embryo noncoding DHS.** Noncoding DHS and footprints were detected in 1.5 kb upstream region of *myo-2* including a pan-pharyngeal element within the *myo-2* promoter. In addition, a noncoding DHS is detected in the 6th and 10th introns of *myo-2*, both of which harbor TF footprints. The four noncoding DHS and associated footprints coincide with PHA-4 ChIP binding sites (light green; Zhong et al. 2010).

Embryo noncoding DHS partially recapitulate enhancers defined in another *C. elegans* locus encoding *hlh-1*, a major bHLH TF of body wall muscle (BWM) that begins expression in embryos (Krause et al. 1994; Lei et al. 2009). The noncoding DHS and TF footprints that we observe at this locus overlap with the enh1 region and enh2 regulatory regions reported to drive expression in BWM precursors D+C and MS+D+C, respectively (Appendix Figure 2.1C). However, the specific P1 and E1 regions that bind PAL-1 and HLH-1, respectively, within enh1 and the enh3 regions are closely located to but do not overlap with our identified noncoding DHS. This discrepancy may be partly due to weak and broad DNaseI signal at the locations, which were not called by our peak calling method as part of the DHS. Our data also do not detect the enh4 enhancer.

To investigate whether the noncoding DHS we observe in the *C. elegans* embryo may represent not only enhancers but also potential repressors or sites of negative regulation, we examined the intergenic region between *col-43* dauer collagen and *sth-1*, which is expressed in spermatheca. Two homeodomain proteins MAB-18 (also known as VAB-3) and CEH-14 are required to insulate *col-43* from activation by the adjacent promoter of *sth-1* (Bando et al. 2005) and are anteriorly expressed in the early embryo (Chisholm and Horvitz 1995; Kagoshima et al. 2013). Homeodomain binding sites HB1 and HB2 for MAB-18 and CEH-14 or MAB-18 alone, respectively, reside in the intergenic region. We observed one embryo noncoding DHS with a TF footprint overlapping the HB1 site that is part of the spermathecal enhancer (Bando et al. 2005). Another noncoding DHS harboring a TF footprint overlaps the HB2 site and an embryo TSS (Chen et al. 2013) (Appendix Figure 2.1D).

DNase-seq data predict additional novel enhancers and distant-acting CRMs

Even within the well-studied gene loci we investigated, we detected several novel regulatory elements. Some of these predictions include noncoding DHS in the first intron of and downstream of *ceh-13*, which were not tested in the Kuntz et al. (2008) study, but are conserved and transcribed and which may represent additional *ceh-13* regulatory elements (Appendix Figure 2.1A). Another example is that of footprints and noncoding DHS observed in the 6th and 10th introns of *myo-2* that overlap with other PHA-4 ChIP binding sites. Since PHA-4 is a transcriptional regulator of pharynx expression, these noncoding DHS may represent additional PHA-4 regulated enhancers of *myo-2* (Figure 2.4E). In addition, we observe a noncoding DHS in the 1st intron of *hlh-1* corresponding to a region that is bound by PHA-4 in embryos (Zhong et al. 2010). We expect that *hlh-1* is repressed by PHA-4 in the pharynx through this putative CRM.

Our data also provide additional evidence for distant-acting regulatory elements in *C. elegans*. Nearly half (6,312) of the noncoding DHS detected in the *C. elegans* embryo are situated less than 500bp to the nearest gene (Figure 2.5A). However, 4,724 (43%) are between 500bp and 2kb from the nearest gene and 3,895 (26%) are over 2kb away, up to 11kb away.

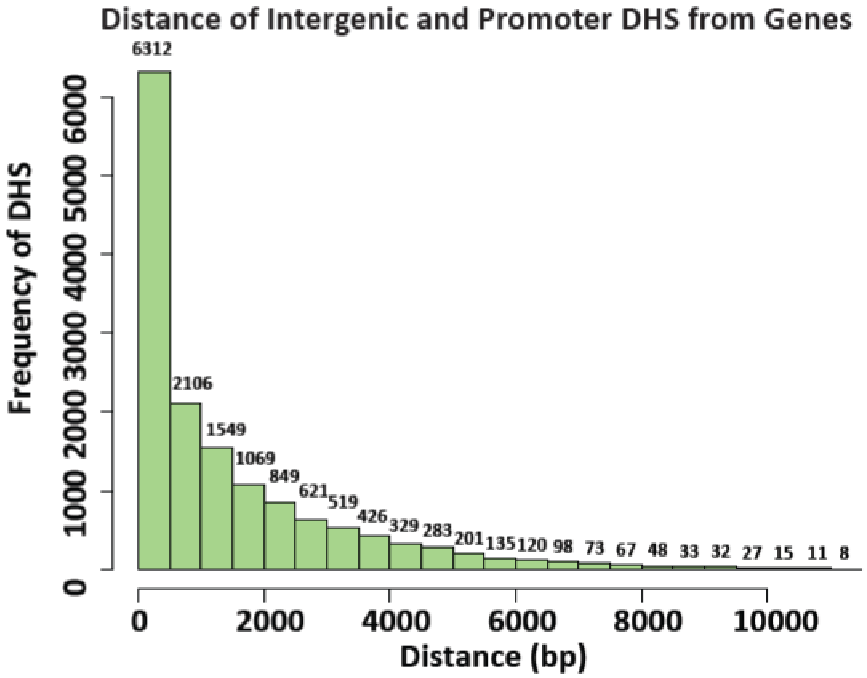


Figure 2.5A Distance of intergenic and promoter DHS to nearest protein-coding gene shows additional evidence for relatively distant CRMs. A little over half (56%; 8,418) of intergenic and promoter DHS are found within 1kb of the nearest protein-coding gene, and most (74%; 11,036) are within 2kb. However, a minority (26%, 3,895) of intergenic and promoter DHS are greater than 2kb away and 10% (1,480) are more than 4kb away.

Discriminative motif discovery within noncoding DHS peaks recovers many known promoter and TF regulatory motifs

We performed discriminative motif discovery to identify overrepresented motifs within noncoding DHS peaks and putative TF footprints using DREME (Bailey et al. 2011). We surmised that overrepresented motifs within these noncoding peaks and footprints might represent sites of TF binding and regulatory activity. Many known *C. elegans* regulatory motifs matched overrepresented motifs found in noncoding DHS, including SL1, Kozak and T-blocks regulatory motifs that were previously described in *C. elegans* core promoters (Grishkevich et al. 2011). We also detect DNA binding motifs of pharyngeal TF PHA-4 expressed in embryos (Figure 2.5B) (Gaudet et al. 2004; Zhong et al. 2010). In addition, we find DNA binding motifs for embryonic regulators EFL-1 (Page et al. 2001), GEI-11 (Tsuboi et al. 2002), EGL-27 (Solari et al. 1999), which regulate embryonic asymmetry, ventral enclosure, and embryonic patterning respectively, and the motif for neuronal nuclear receptor UNC-55 (Zhou and Walthall, 1998).

Among motifs situated in the promoter (<300 bp upstream of ATG start), intergenic, and intron DHS, we detected additional *C. elegans* motifs. In promoter DHS we recover *C. elegans* TATA-box and SP1 canonical promoter motifs (Grishkevich et al. 2011), as well as binding motifs for ELT-2 intestinal TF (McGhee et al. 2009), and CEH-28, a NK-2 homeodomain TF expressed in the M4 neuron and other extra-pharyngeal neurons in embryos (Ray et al. 2008). In intergenic DHS, we find the intestinal TF SLR-2 (Kirienko and Fay 2010), the N1 pan-neuronal regulatory motif (Ruvinsky et al. 2007), and motifs of EGL-5, a TF expressed in the posterior half of the embryo (Ferreira et al. 1999; Baum et al. 1999), and NHR-6, a nuclear hormone receptor with several roles in development including embryo morphology (Gissendanner et al. 2008).

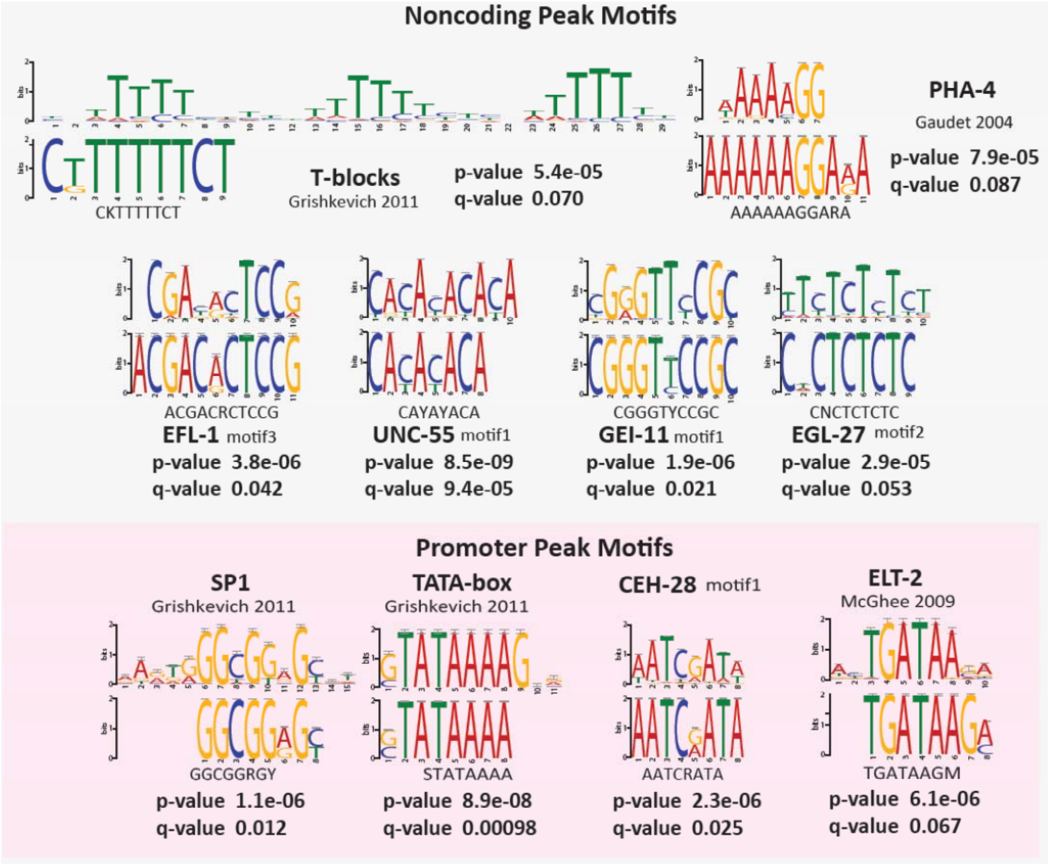
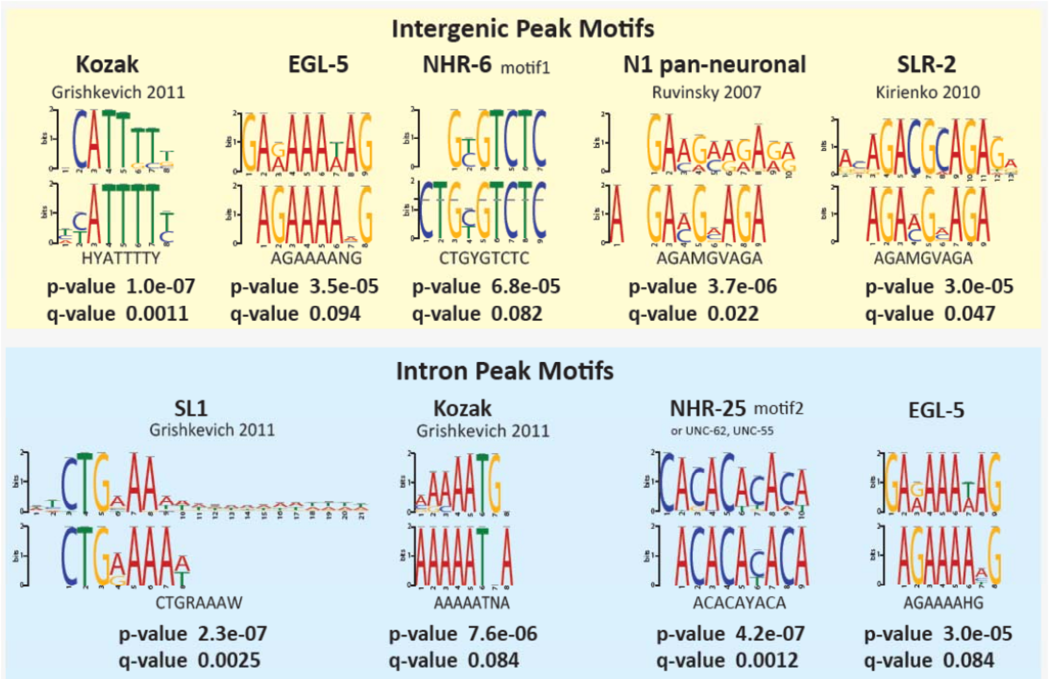


Figure 2.5B. Known motifs recovered from noncoding DHS regions

Unless otherwise specified, comparison motifs are from modENCODE (Gerstein et al. 2010; Araya et al. 2014).

Noncoding peak motifs. Motifs match many *C. elegans* regulatory motifs, including promoter T-blocks (Grishkevich et al. 2011), and TF binding motifs of PHA-4 (Gaudet et al. 2004), EFL-1, UNC-55, GEI-11, and EGL-27**.**

Promoter peak motifs. Motifs match SP1 and TATA-box core promoter motifs (Grishkevich et al. 2011), and TF binding motifs of ELT-2 (McGhee et al. 2009), and CEH-28.



Intergenic peak motifs. Motifs match the Kozak motif (Grishkevich et al. 2011), and TF binding motifs of SLR-2 (Kirienko and Fay 2010), N1 pan-neuronal motif (Ruvinsky et al. 2007), EGL-5, and NHR-6.

Intron peak motifs. Motifs match several regulatory motifs, including SL1 core promoter motif and Kozak sequences (Grishkevich et al. 2011), and EGL-5 TF binding motif.

From noncoding DHS associated with gut-specific genes we recovered DNA motifs resembling binding motifs of known intestinal differentiation factors ELT-2 and SLR-2 (McGhee et al. 2009; Kirienko et al. 2008; Kirienko and Fay 2010) (Appendix Figure 2.3C). In the noncoding DHS associated with neuronal-specific genes we found the TF binding motif for EGL-5 which is involved in development of the posterior nervous system (Ferreira et al. 1999; Baum et al. 1999) (Appendix Figure 2.3D).

Nucleotide-level DNaseI cleavage accessibility across *C. elegans cis*-regulatory motifs

We measured the pattern of DNaseI cleavage accessibility across predicted *cis*-regulatory DNA motifs on a nucleotide level. We focused our attention on known motifs recovered in our study (Figure 2.5B). When we mapped average DNaseI cleavage in a window surrounding motif sites identified within 2kb upstream regions of protein-coding genes, almost all the motifs showed patterns characteristic of TF footprints, with a lower read coverage centering around the DNA motif indicating protection from DNaseI cleavage and a symmetric shift between reads aligning to positive and negative strands of the genome (Figure 2.6A; Appendix Figure 2.4).

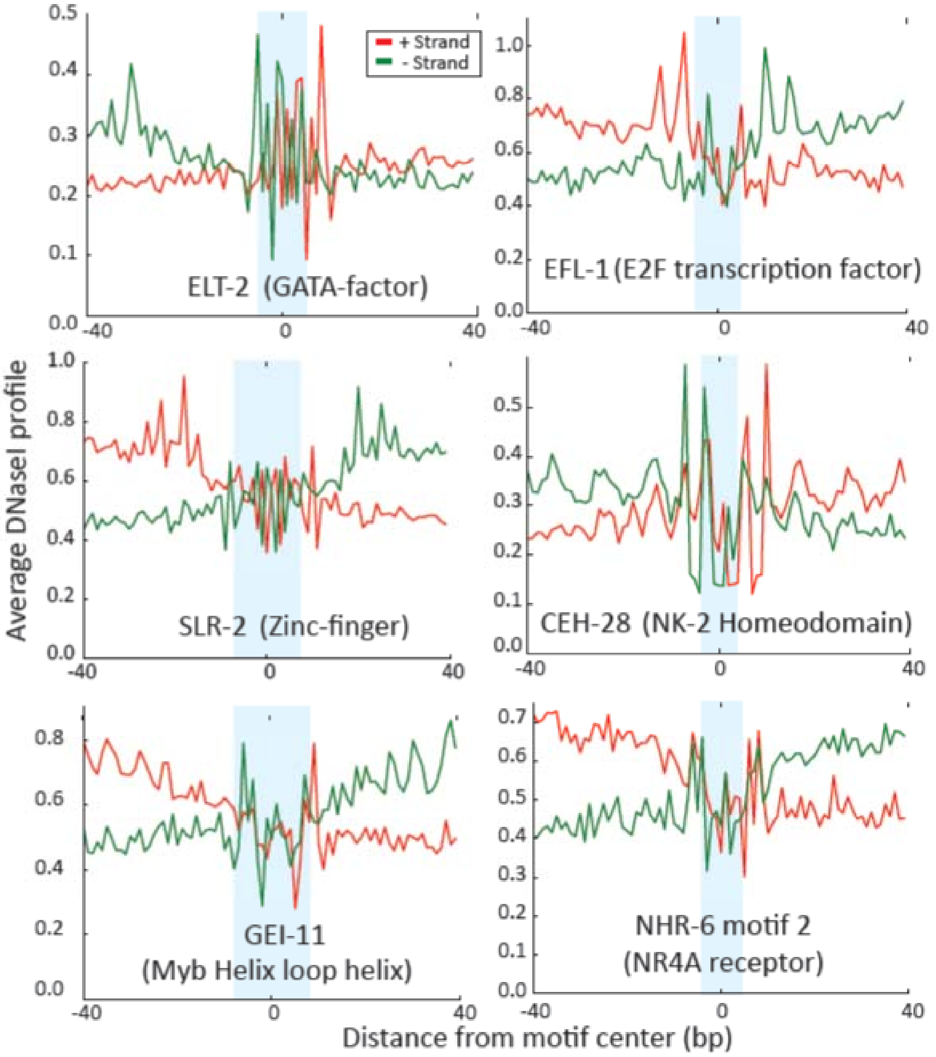


Figure 2.6A. Average DNase profile over *C. elegans* motif sites. *C. elegans* motif sites show characteristic patterns of DNaseI cleavage accessibility and demonstrate strand-shift in reads that is indicative of TF footprints. Average DNase profile is calculated over thousands of predicted motif sites within 2 kb upstream region of genes using start sites of reads across 80bp region surrounding motifs. Positive (red) and negative strand (green). Light blue shading shows base pair position of motif: ELT-2 (10bp motif), EFL-1 (10bp), SLR-2 (13bp), CEH-28 (8bp), GEI-11 (16bp motif), and NHR-6 motif 1 (7bp).

Prediction of novel *cis*-regulatory motifs

We also found many other novel motifs overrepresented in noncoding DHS for which there were no known functions. Some of these matched conserved DNA motifs found by two prior studies in *C. elegans* and other nematodes using alignment-based approaches (Ihuegbu et al. 2012) and gene orthologs (Elemento and Tavazoie 2005). We performed Gene Ontology and anatomy enrichment analysis on genes associated with these noncoding motifs in order to predict function (Table 2.1; Appendix Table 2.3). A variety of GO annotations of biological function were enriched, including response to stimulus (e.g. AAAATTCMAAA enriched in head neurons; MAACAACAACAA enriched in ventral cord neurons) and hormone signaling (e.g. ACTACAAACTAC enriched in excretory cell). Regulation of localization was enriched in several motif associated genes (CGCGCAAATGA; GCRGCCGACA enriched in intestine and muscle including vulval and body wall). Selected motifs are outlined in Table 2.1, with additional motifs in Appendix Table 2.3.

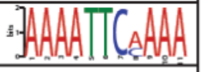
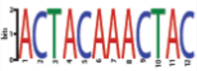
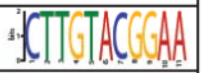
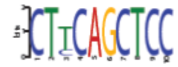
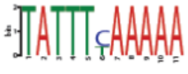
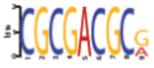
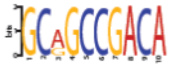
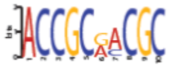
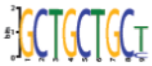
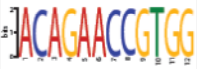
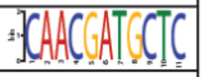
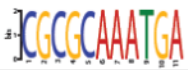
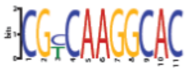
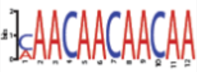
	IUPAC Motif	p-val	E-val	Prior Match?	Threshold	#	Example Genes	Gene Ontology Enrichment [Anatomy Enrichment]	Motif Logo
Intergenic	AAAATTCMAAA	1.10E-08	0.022	Stormo F01D5.10.8	0.05	77	*unc-13 srw-90 ceh-36* *nduf-7 nhr-52 unc-75* *clec-2 nhr-196 cnb-1*	response to chemical stimulus cellular metabolism [head neurons]	
	ACTACAAACTAC	3.70E-09	0.0073	Several incl. Stormo C39D10.7.2	0.025	24	*hmr-1 pph-5 pmr-1* *nhr-62 pad-1 pph-5* *oac-1 fbxb-69 clec-172*	response to chemical stimulus steroid hormone mediated signaling response to Ca2+ ion organic substance metabolism dauer entry [seam cell, spermatheca, excretory cell]	
	CTTGTACGGAA	1.50E-08	0.029	None	0.05	18	*ccg-1 nhr-196* *gst-35 nhr-52* *ncx-9 srw-90*	ion membrane transport actin myosin filament sliding	
	CTYCAGCTCC	2.50E-09	0.0049	None	0.05	27	*asna-1 nhr-258 pde-4* *avr-14 nhr-97 ptr-19* *hlh-14 oac-1 unc-9*	establishment of localization ion transport vesicle mediated transport transmembrane transport synaptic transmission cell-cell signaling	
	TATTTYAAAAA	4.00E-04	8.70E-02	None	0.05	12	*dmd-3 nhr-196* *gem-1 nhr-52* *mir-1823 pup-3*	signal transduction cell response to stimulus signaling cell communication activation of RasGTPase activity	
Promoter	CGCGACGCR	7.50E-13	9.10E-07	None	0.05	50	*act-4 clic-1 hut-1* *cit-1.1 fust-1 lin-53* *clec-87 hrp-1 lsy-2*	reproduction organ development reproductive structure development aging [head, tail, spermatheca]	
	GCRGCCGACA	1.10E-10	0.00013	None	0.05	33	*dpy-18 nlp-4 sri-33* *oac-12 psa-3 sri-78* *irld-65 ptc-1 stau-1*	transport establishment of localization maintenance of location lipid localization lipid storage vesicle mediated transport membrane organization [intestine, vulval muscle, BWM, anal dep. muscle]	
Intron	ACCGCRMCGC	1.40E-28	2.70E-22	None	0.025	71	*che-6 egas-2 nhr-95* *clec-126 egas-3 pde-1* *clec-84 fln-2 slo-1* *cogc-1 gcy-19 srr-2* *dyf-2 lgc-34 sru-47* *egas-1 nhr-79 vamp-7*	regulation of signal transduction	
	GCTGCTGCY	2.00E-19	4.00E-13	Elemento motif 151	0.05	93	*abcf-2 mca-3 sax-3* *abts-3 mpz-1 sax-7* *ceh-44 nhr-27 sos-1* *egl-4 pac-1 tol-1* *inso-1 prbm-1 unc-2* *lin-11 pde-4 unc-3* *lin-17 ptp-3 unc-76* *lin-40 rgs-2 unc-89* *lit-1 sax-2 vab-1*	cellular response to stimulus signal transduction regulation of response to stimulus signaling cell communication protein metabolism catabolism regulation of metabolism	
Noncoding	ACAGAACCGTGG	4.60E-10	0.0015	Stormo F45F2.11.3	0.025	32	*chaf-1 fkb-8 mes-4* *dct-17 haf-2 ruvb-1* *erfa-3 let-607 ubc-6*	cellular component organization apoptotic process cellular component organization or biogenesis aging anatomical structure development reproduction [excretory cell, coelomocyte, germline]	
	CAACGATGCTC	4.60E-10	0.0015	Stormo F55A3.1.4	0.05	29	*ark-1 map-2 set-16* *ash-2 nhr-60 set-33* *gpa-16 nra-1 vab-1*	reproduction	
	CGCGCAAATGA	7.40E-09	0.024	Elemento motif 95	0.05	26	*apb-1 ptr-24* *pha-1 nlp-40* *srh-48 odr-10*	localization embryo development embryo devt ending in birth or egg hatching protein glycosylation	
	CGYCAAGGCAC	1.10E-16	3.60E-10	Stormo T03F7.5.4, Elemento motif 367	0.025	32	*nhr-70 nlp-42* *nhr-79 oac-49* *nhr-95 pqn-18*	protein dephosphorylation regulation of vesicle-mediated transport	
	MAACAACAACAA	3.70E-11	0.00012	Stormo F58H7.3.1	0.025	42	*cdh-4 egl-8 unc-31* *daf-4 hlh-30 unc-53* *egl-10 lin-42 vab-1*	reproduction response to external stimulus regulation of cellular process positive regulation of locomotion regulation of locomotion positive regulation of biological process [ventral cord neurons]	

Table 2.1. Selected novel predicted regulatory motifs

Selected novel predicted regulatory motifs and gene ontology analysis of motif-associated genes. Left border shows category of noncoding DHS where motif is overrepresented. p-values (p-val) and erased E-value (E-val) of each of the identified motifs are shown, along with whether motif matches previously identified Stormo or Elemento motifs (Ihuegbu et al. 2012; Elemento et al. 2005), and FIMO threshold (Threshold) used to select motif-associated genes. Number of motif-associated genes (#Genes) used in GO enrichment analysis. Gene names of some motif-associated genes (Example Genes) are shown. Both IUPAC motif and motif logos are shown. Blue background indicates related GO terms. Top enriched GO terms are shown (see methods). Enriched anatomy terms, if present, are shown in square brackets.

DNase-seq data refines prediction of tissue-specific genes by regulatory DNA motifs

We investigated whether DNase-seq data would be able to improve our ability to predict the tissue-specific expression of genes regulated by known DNA motifs. The N1 pan-neuronal regulatory motif predicts genes expressed widely in neuronal cells (Ruvinsky et al. 2007). Similarly, the ELT-2 motif is found near intestinally expressed genes (McGhee et al. 2007). Other important TFs with known roles in the intestine include *C. elegans* homolog of Homothorax/Meis UNC-62 (Van Nostrand et al. 2013; McGhee et al. 2007; Van Auken et al. 2002) and SLR-2 (Kirienko et al. 2008; Kirienko and Fay 2010). We compared the percentage of genes correctly predicted to be expressed in neuronal or intestinal tissues using the presence of predicted DNA motifs alone versus the presence of DNA motifs within noncoding DHS (see Methods). In both cases we were able to improve prediction accuracy using noncoding DHS together with motifs, from 41% to 55% of genes in FACS-sorted neuronal tiling array (McGhee et al. 2009) and

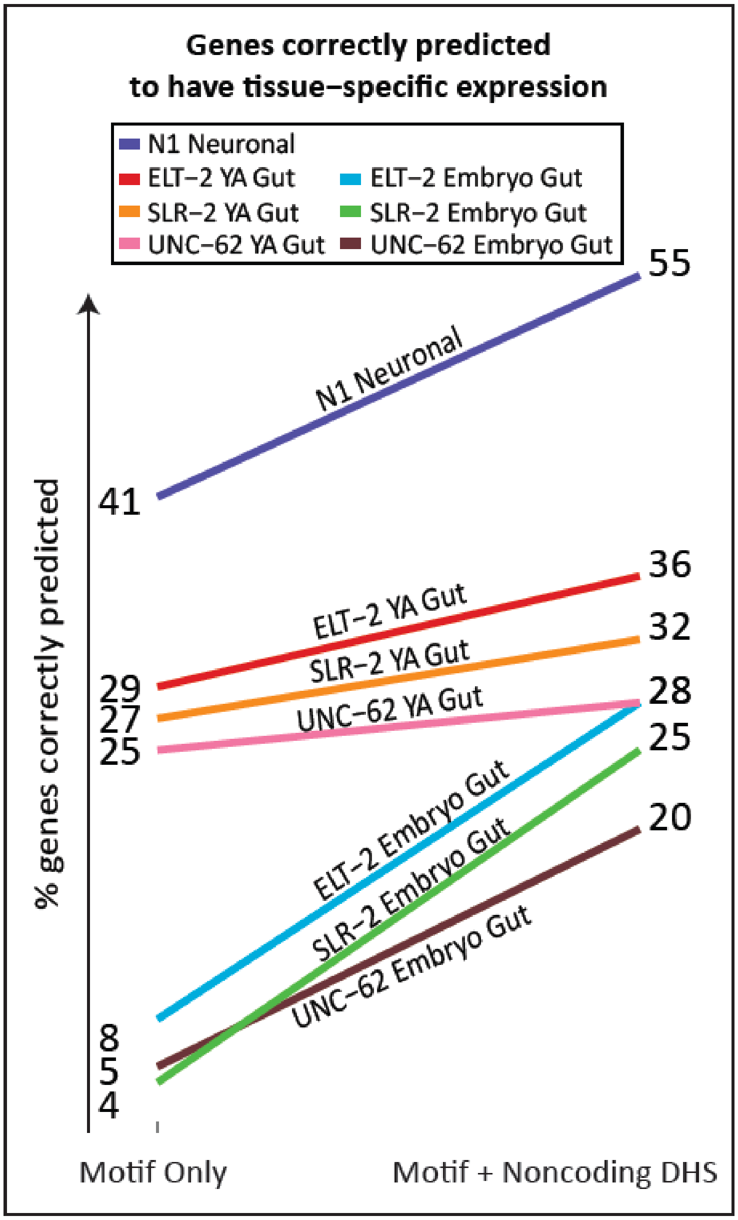


Figure 2.6B. Using regulatory motifs found within noncoding DHS to refine prediction of tissue specific gene expression. The percentage of genes correctly predicted to be expressed in the tissue expression dataset from the presence of DNA regulatory motif (Motif Only) was compared to presence of DNA regulatory motif within noncoding DHS (Motif + Noncoding DHS). Taking into account the presence of N1 motifs within noncoding DHS improves prediction accuracy of neuronal expression from 41% to 55% (data from McGhee et al. 2009; purple). Taking into account gut regulatory TF ELT-2 (blue), SLR-2 (green) and UNC-62 (brown) motifs located specifically within noncoding DHS improves prediction accuracy of embryonic intestinal expression (FACS data from Spencer et al. 2011) from 8%, 4% and, 5%, respectively to 28%, 25%, and 20%, respectively. Taking into account ELT-2 (red), SLR-2 (orange) and UNC-62 (pink) motifs within noncoding DHS slightly improves prediction accuracy of expression in adult dissected intestines (data from McGhee et al. 2007) from 29%, 27%, and 25%, respectively to 36%, 32%, and 28%, respectively.

from 8%, 4%, 5%, respectively using ELT-2, SLR-2, and UNC-62 motif only to 28%, 25%, 23%, respectively of genes in FACS embryonic intestine using ELT-2, SLR-2, and UNC-62 motifs within noncoding DHS (Figure 2.6B) (data from Spencer et al. 2011). We also show smaller improvement from 29% (ELT-2), 27% (SLR-2), and 26% (UNC-62) to 36%, 32%, 31%, respectively in adult dissected gut (data from McGhee et al. 2007). The result of these analyses using DNase-seq data is a smaller but more accurately predicted set of genes expressed in neurons or intestine.

Most L1 arrest regulatory elements discovered by DNase-seq are also found in the embryo, whereas 12% appear to be L1 arrest condition-specific and reflect higher gene expression

Comparing DNase-seq data between L1 arrest and embryo conditions, we find that most (88%) of the 16,084 noncoding DHS found during the L1 arrest stage were also found in the embryo. However 1,854 (12%) appear to be specific to the L1 arrest condition, when compared to *C. elegans* embryo DNase-seq data. We are also able to identify 9,359 putative transcription factor footprints in L1, with 2,946 TF footprints residing in L1 condition-specific elements. Genes with L1 condition-specific regulatory elements have 12.5% higher expression in the 6hr L1 starved larvae compared to the embryo (Appendix Figure 2.5E, expression data from Baugh et al. 2009, two-sided KS test, $p < 1.6 \times 10^{-8}$).

L1 arrest condition-specific noncoding DHS are found in many genes upregulated in L1 arrest larvae

Of the 1,854 L1 arrest condition-specific regulatory elements, 44% (817) are associated with at least one category of genes we expect to be involved in the regulation of L1 arrest: those targeted by DAF-16 and/or PHA-4, genes responsive to starvation in the L1, and genes highly upregulated in L1 starved vs. embryo (Appendix Figure 2.5D; see Methods for defining these genes). 14% (256) of these genes with L1 condition-specific DHS are top DAF-16 targets (Tepper et al. 2013), 22% are PHA-4 targets (Zhong et al. 2010), 18% are genes most responsive to starvation in L1 larvae (Baugh et al. 2009) and 17% are genes highly upregulated in L1 arrest larvae compared to embryos (Baugh et al. 2009). Furthermore, all DHS and noncoding DHS from L1 arrest larvae are 1.7-fold and 2.4-fold enriched, respectively, in PHA-4 ChIP binding sites from stage-matched samples of starved L1 larvae (two-sided KS test, $p < 3 \times 10^{-16}$), suggesting that we are able to recapitulate CRMs for targets of PHA-4, a TF regulator of L1 starvation survival.

We are able to detect L1 arrest condition-specific DHS in targets of DAF-16 and PHA-4 regulated genes and other genes differentially regulated in L1 arrest by investigating individual gene loci. For example, *icl-1* (also known as *gei-7*) is a key enzyme of the glyoxylate cycle, is involved in the breakdown of fats into carbohydrates, and is a known target of DAF-16 (Murphy et al. 2003; Tepper et al., 2013). Expression of *icl-1* is highly upregulated in *daf-2* mutants (Murphy et al. 2003) and in response to starvation (7.9 fold; Baugh et al. 2009; Van Gilst et al.

2005) and in L1 arrest compared to embryos (1.9 fold; Baugh et al., 2009). It also appears to be regulated by PHA-4 during embryo and L1 arrest according to ChIP data (Zhong et al., 2010). We detect one L1 condition-specific noncoding DHS harboring TF footprints which overlap both a DAF-16 binding motif ($p < 1x10^{-4}$) and PHA-4 motif ($p < 5x10^{-5}$) in the first intron of *icl-1* (Figure 2.7A). Three other L1 arrest noncoding DHS were found near *icl-1* coinciding with PHA-4 ChIP binding peaks detected in L1 starved larvae (Zhong et al., 2010).

Figure 2.7. L1 arrest condition-specific noncoding DHS detected in genes upregulated during L1 arrest. Total DNaseI signal (red) from both strands of L1 arrest DNase-seq read data shown, as well as individual DNaseI signal from positive (orange) and negative (green) strands. Total DNase signal (light blue) from both strands of embryo DNase-seq read data is also shown. L1 arrest noncoding DHS (red) and associated TF footprints (pink), as well as embryo noncoding DHS (light blue boxes) and associated TF footprints (dark blue boxes) were detected. Additional tracks are *C. elegans* RefSeq genes (black boxes with arrows), noncoding transcripts (brown boxes), 12hr Starved L1 mRNA-seq tracks (black) from Maxwell et al. (2012), and phyloP conservation (dark blue) are also shown. Other tracks include PHA-4 ChIP-seq binding peaks from embryo (light green) and starved L1 larvae (signal shown in purple; peaks shown as purple boxes; Zhong et al. 2010). PHA-4, DAF-16, DAF-19 binding motifs (if relevant) are shown in purple, orange or magenta boxes, respectively. TSS previously found by L1 Starved GRO-cap sequencing (if relevant; data from Kruesi et al. 2013) is shown as dark green boxes.

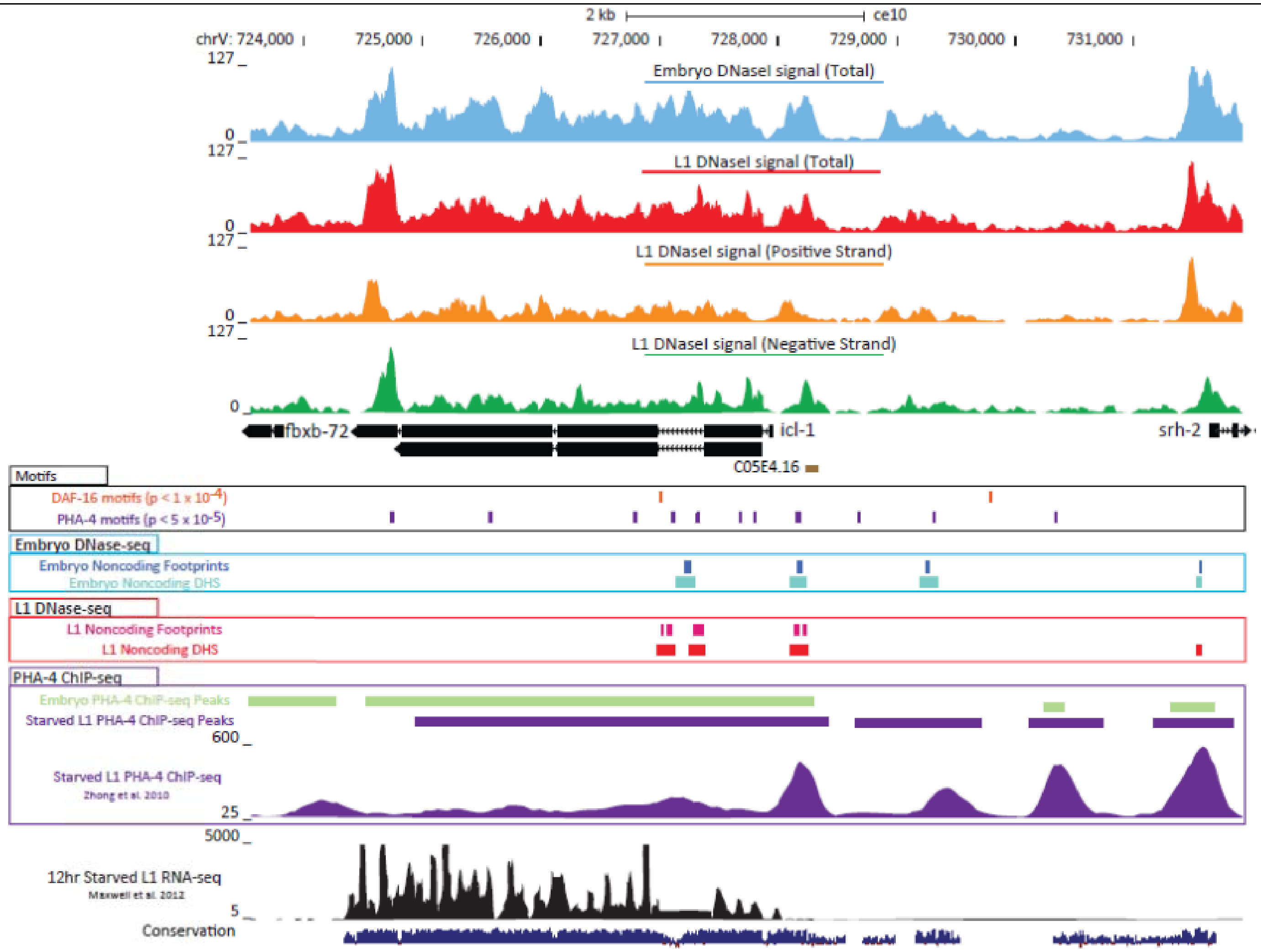


Figure 2.7A. Noncoding DHS of *icl-1/gei-7*, which is regulated by both DAF-16 and PHA-4. The *icl-1* gene harbors one L1 condition-specific noncoding DHS in the first intron with a DAF-16 binding motif ($p < 1 \times 10^{-4}$ threshold) and a PHA-4 motif ($p < 5 \times 10^{-5}$). TFs footprints are detected within this noncoding DHS and overlap the DAF-16 binding motif. Three other noncoding DHS are detected in both L1 and embryo which coincide with PHA-4 ChIP-seq binding detected in L1 starved larvae (Zhong et al., 2010) and two of them harbor TF motifs. Two additional upstream regions bound by PHA-4 in L1 starved larvae were not detected.

Another example is *pha-4*, a TF which plays a role in L1 starvation survival and autoregulates its own promoter (Zhong et al., 2010). We detected multiple L1 noncoding DHS upstream of *pha-4* coinciding with PHA-4 ChIP binding regions during L1 arrest (Figure 2.7C). One of these DHS coincides with the TSS of *pha-4c*, the shortest isoform, which was observed in a previous study using GRO-cap in both embryo and starved L1 larvae (Maxwell et al. 2014). Another TSS far upstream of the longest isoform *pha-4a* was previously observed in embryos but only weakly in the L1 starved larvae (Maxwell et al., 2014) and coincides with a noncoding DHS in our embryo DNase-seq data but not in L1 arrest. We also detect a L1 condition-specific noncoding DHS directly upstream of *pha-4a* containing multiple TF footprints. While this DHS harbors some PHA-4 binding motifs, it is unclear whether this noncoding DHS reflects PHA-4 binding (which appears to weakly bind this region) or binding of another TF.

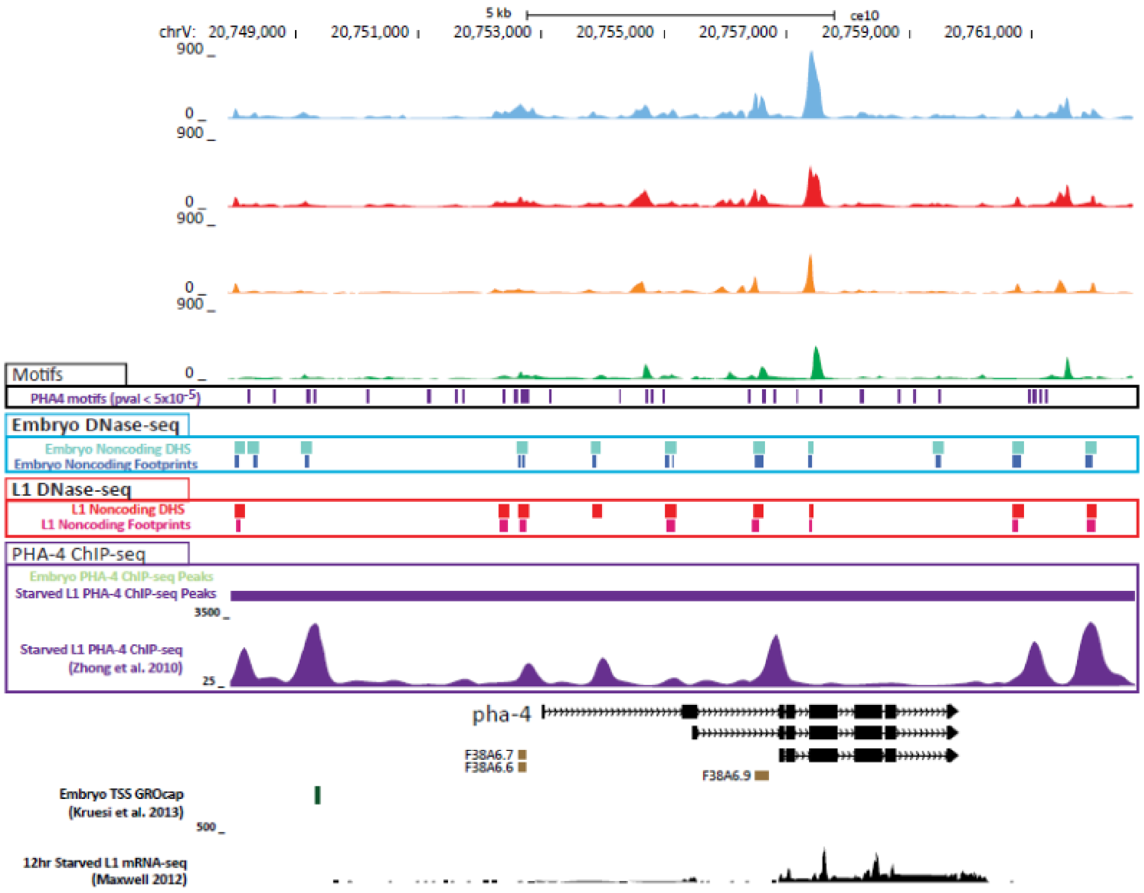


Figure 2.7B. Known and novel CRMs of *pha-4*. Upstream of the longest transcript, pha-4a, we observe four embryo and three L1 arrest noncoding DHS and an embryo condition-specific noncoding DHS overlapping a TSS previously detected in embryos but not L1 arrest by GRO-cap sequencing (Kruesi et al. 2013). Directly upstream of pha-4a is also an L1 arrest condition specific noncoding DHS harboring TF footprints and overlap PHA-4 TF binding sites. It is possible that is bound weakly by PHA-4 (Zhong et al. 2010).

An example of a gene whose role in L1 arrest is less well understood, but in which we found evidence supporting differential regulation, is the nuclear hormone receptor *nhr-4*. It is expressed in ciliated sensory amphid neurons and is directly regulated by RFX/DAF-19 TF (Burghoon et al 2012). Expression of *nhr-4* is upregulated 2.3 fold in response to starvation L1 and in L1 arrest compared to embryos (1.5 fold; Baugh et al. 2009). We detect four L1 noncoding DHS upstream of *nhr-4*, two of which are specific to the starved L1 larvae condition (Figure 2.7C). Of these, one overlaps an annotated TSS previously detected by GRO-cap sequencing in starved L1 (Maxwell et al. 2013). The other has footprints which coincide with both a DAF-19 motif and a PHA-4 motif, and which appears to be weakly bound by PHA-4 in starved L1 (Figure 2.7C; Zhong et al. 2010). The other two noncoding DHS detected in both the embryo and L1 arrest overlap PHA-4 ChIP peaks from both conditions.

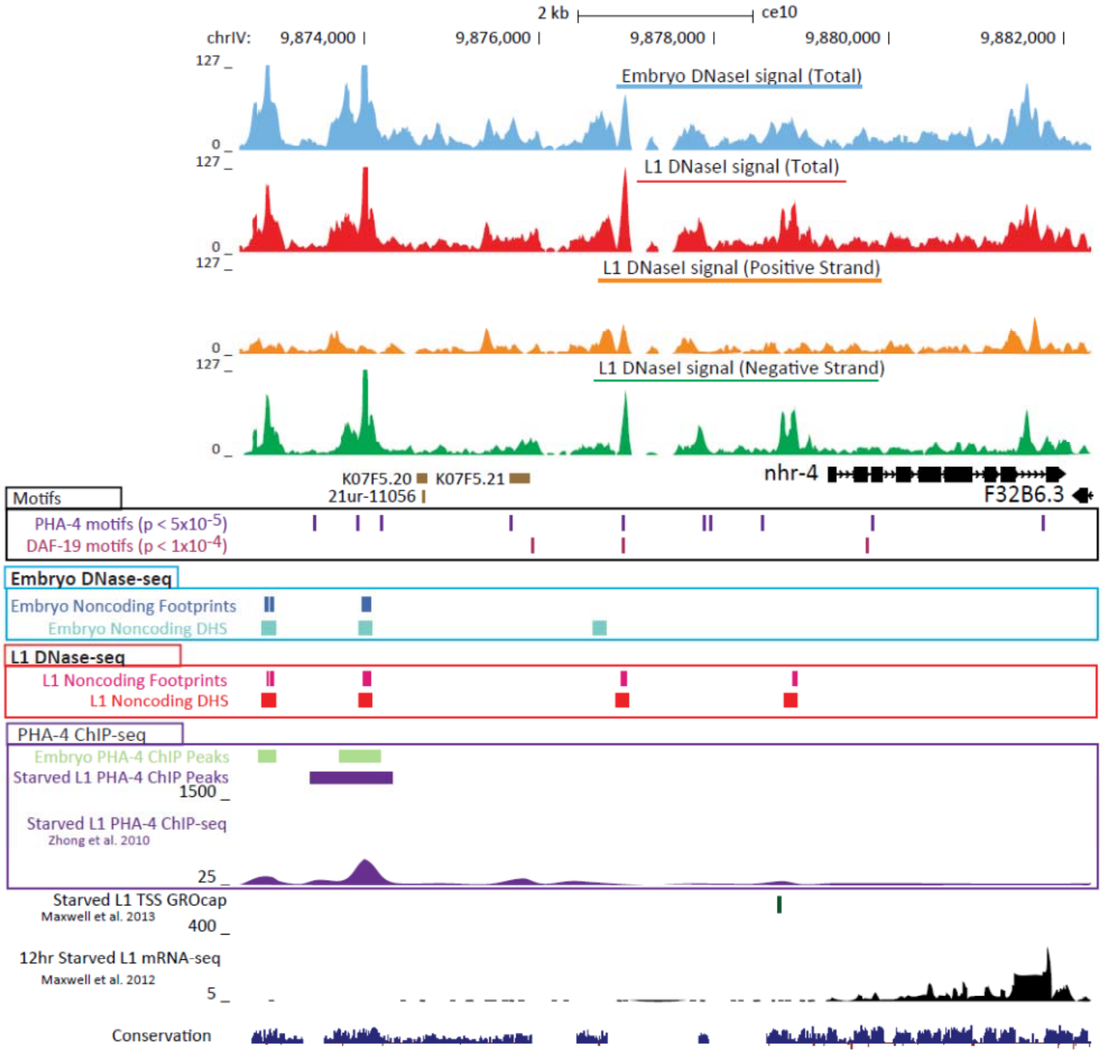


Figure 2.7C. Noncoding DHS of nuclear hormone receptor *nhr-4*. Four noncoding DHS are detected upstream of *nhr-4* in L1 arrest, all harboring TF footprints. The most proximal two are present only in L1 arrest and not in the embryo. The one located in the presumed promoter of *nhr-4* overlaps an *nhr-4* TSS that is detected in L1 arrest but not in the embryo by GRO-cap sequencing (Kruesi et al. 2013). The other overlaps both DAF-19 (p < 1x10-4) and PHA-4 (p < 5x 10-5) TF binding motifs, which does appear to be weakly bound by PHA-4 according to ChIP-seq (Zhong et al. 2010).

Discussion

We have identified 26,644 embryo noncoding DHS harboring 55,890 TF footprints and 15,841 L1 arrest noncoding CRMs harboring 32,685 TF footprints, through a genome-wide systematic study of *cis*-regulatory regions and TF binding in *C. elegans*. We are able to profile *cis*-regulatory sites without specifying particular prior TFs of interest and using chromatin accessibility as our guide. We have shown that we can recapitulate many known and functionally characterized enhancer regions and, in many cases, have refined the boundaries of the enhancer regions that were previously tested in transgenic reporter assays or detected through the relatively broad widths of ChIP-seq peaks. The DNaseI peaks identified here are typically only 150 bp and will be useful to define boundaries of many *cis*-regulatory modules (CRMs). We identified many known enhancers and TF footprints of *C. elegans* genes including *lin-39/ceh-13*, *hlh-1*, *myo-2*, *myo-3*, *elt-2*, and *lir-1/lin-26*. Our data were able to recapitulate 22 of 29 known enhancers within these loci. In addition to correctly identifying known enhancers and TF footprints, our data also predict potential novel CRMs and many smaller TF footprints. For instance, the data predict regions downstream of *ceh-13* and other regions that coincide with PHA-4 binding in the locus of *hlh-1,* where we surmise that PHA-4 may act to repress *hlh-1* where it is expressed in the pharynx, similar to its role in repressing *lin-26* in the pharynx. We also recovered known negative regulatory homeodomain sites in the *col-43/sth-1* locus, suggesting that we are also able to find repressor CRMs.

It is common practice in *C. elegans* to regard immediate sequence 5' of TSS as reflecting endogenous expression (Dupuy et al. 2007). However, there are numerous documented cases (reviewed in Gaudet and McGhee 2010) in which gene regulation is complex, being regulated from intronic, 3', or distant 5' sequences. Another study showed that while most (62%) *C. elegans* transcript and translation fusion reporter expressions replicated, expression was often observed in additional cells or in restricted patterns, suggesting other CRMs were involved (Murray et al. 2012). While we observed that most 74% (11,036) promoter and intergenic DHS are within 2kb of the nearest protein-coding gene, a significant proportion (26%; 3,895) are greater than 2kb, and 10% (1,480) are more than 4kb away (Figure 2.5A). Although it is difficult to definitively assign target genes to CRMs, even the nearest gene to a noncoding DHS can be far away. Furthermore, 53% (10,890) of protein-coding genes have at least one noncoding DHS in the embryo, and of these 17% (1,901) have complex regulation, with more than four noncoding DHS (Appendix Figure 2.3B). We thus provide additional evidence that *C. elegans* transcriptional regulation can be complex and controlled by relatively distant CRMs.

Our data are highly resolved enough to identify protection from DNaseI cleavage in noncoding DHS and across *C. elegans cis*-regulatory motifs within them that appear to be sites of TF binding (Appendix Figure 2.4). 82% and 84% of embryo and L1 arrest noncoding DHS, respectively, were found to harbor TF footprints. We find numbers of noncoding DHS on the same order of magnitude as

Drosophila DNase-seq (roughly 20,000 noncoding DHS per stage) with similar depths of sequencing (Thomas et al. 2011). Our finding that L1 arrest noncoding DHS are 88% shared with embryo noncoding DHS are also similar to findings from *Drosophila* showing that most noncoding DHS are also similar to previous findings that show 78% concordance of DHS between Stage 5 and Stage 11 *Drosophila* embryos (Thomas et al. 2011).

It is difficult to estimate the cellular resolution of DNase-seq data that we have generated from entire embryos or L1 arrest larvae. We were able to recover overrepresented motifs in DHS representing binding sites of TF regulators of the three most abundant tissues in *C. elegans*: muscle, neuronal, and intestine (Appendix3 D-E) as well as motifs that occur in a smaller number of tissues. Naturally, these data are likely composed of an average of DNase hypersensitivity profiles of different tissues. We were able to find novel regulatory motifs that at least, according to anatomy enrichment profiles, appeared to be enriched in relatively specific areas (e.g. AAAATTCMAAA enriched in head neurons; MAACAACAACAA enriched in ventral cord neurons; ACTACAAACTAC enriched in excretory cell; Table 2.1) but it is very difficult to specifically attribute changes in gene regulation to a given spatial region within the embryo or L1 larvae without additional information. Thus we have evaluated our noncoding DHS in gene loci in the context of global changes in transcriptional regulation that are occurring between L1 arrest and embryo and in gene loci whose expression and regulation has been studied in the embryonic or L1 arrest context. In order to probe gene activity within a small number of specific cell types we suspect it will become more feasible in the future to isolate tissues and use a similar technique such as ATAC-seq which can work with smaller amounts of starting material compared to DNase-seq.

These DNase-seq maps of DHS and TF footprints will be useful for exploring and dissecting genome-wide regulation of genes active in the embryo and to discover novel regulatory factors and their potential sites of action. For example, we were able to use DNaseI data to refine and improve the prediction of tissue-specific genes by focusing on N1 (neuronal) and ELT-2, UNC-62, and SLR-2 (intestinal) DNA motifs present within noncoding DHS in embryos. Putative CRMs and TF binding site data from this study will be available through WormBase.

Comparative analysis of L1 arrest condition-specific noncoding DHS indicate many potential sites of *cis*-regulatory action in genes whose expression differs between the L1 arrest larvae and the embryo, as well as genes implicated in starvation response of L1 larvae and in specific target genes of DAF-16 and PHA-4 transcriptional regulators downstream of signaling pathways involved in L1 arrest.

Using our noncoding DHS, we identified 57 novel regulatory DNA motifs involved in developmental processes ranging from aging and reproduction to signal transduction, cell-cell-signaling, and behavior. Future experiments will be needed to assay the functional activity of these noncoding DHS and the role of TF footprints in controlling activity. DNase-seq may be applied to other nematode species whose genomes and transcriptomes are known, but whose regulation has not yet been explored and for which transgenic assays will be extremely difficult.

Methods

C. elegans culture and nuclei isolation

C. elegans wild-type N2 worms were synchronized and grown in liquid culture (10 worms/uL and 20 mg/mL *E. coli* HB101 in S-complete) over at least two generations. Embryos around the 40-cell stage were obtained by bleaching adult worms and then frozen at -80°C. To obtain L1 arrest larvae, bleached embryos were resuspended in S. complete and allowed to hatch in the absence of food. Starved L1 arrest larvae were collected at 10 hours and frozen at -80°C. To isolate nuclei, samples were thawed and ground to fine powder with mortar and pestle over dry ice. Samples were reconstituted in nuclei purification buffer (0.1% Triton-X, spermine, spermidine, and protease inhibitor) and dounced for 30 strokes (nuclei isolation protocol from INTACT method; Steiner and Henikoff et al. 2015). Nuclei were collected by spinning 10 minutes at 0.1 *g* to separate from debris and visualized using DAPI. Nuclei were further purified by spinning 10 minutes at 1000 *g* over a cushion of Optiprep (60% iodixanol) at 4°C.

DNaseI treatment, DNA purification, and size-selection

Embryo and L1 arrest larvae nuclei were treated with 0, 20, 40, 80, 120, 160 U/mL DNaseI in 1X DNaseI digestion buffer (containing $CaCl_2$, spermine, spermidine, protease inhibitor) each for 3 minutes at 37°C. DNaseI treatment follows the conditions from the Stamatoyannopoulos lab protocol (Thurman et al. 2012). DNaseI treatment was quenched with STOP buffer containing 20mg/mL

Proteinase K and incubated 55°C overnight. After treating with 45ug/mL boiled RNase A for 30 minutes, DNA was purified and concentrated using column purification. The DNA sample was run on 1% agarose, stained with Sybr Gold, and the gel piece containing DNA fragments less than 500bp was purified. DNA yield was measured using a Qubit fluorometer. See Appendix 1 for adapted DNaseI protocol.

QPCR quality control and measuring enrichment in regulatory region

QPCR primers were designed against the conserved MUSSA regions of "true positive" N1, N2, N3, N4, N5, N7, N8, N9, N11 *lin-39/ceh-13* enhancers and N5 and N6 negative control non-enhancer regions studied by Kuntz et al. (2008). (Appendix Table 2.1). QPCRs were performed with calibration of duplicate genomic DNA standards and absolute derivative measurement of C_p. Relative fold enrichment was compared within samples by normalizing measured concentration of each region vs. mean of negative controls (Appendix Figure 2.2). The sample from the DNaseI concentration harboring the highest measure of regulatory enrichment from each biological replicate was prepared into a library and multiplex sequenced on Illumina HiSeq to yield 50bp single end reads.

Read alignment and quality control

Reads were analyzed using FastQC and filtered using quality threshold Q20 (Appendix Table 2.1). 50bp single-end reads from embryo replicates B, C, D and L1 arrest X, Y, Z replicates were trimmed to 45bp and mapped to WS220 (ce10) version of *C. elegans* genome using Bowtie 1.0.0 (Langmead et al. 2009) using settings that did not allow alignments with more than two mismatches, disallowing reads with more than two read alignments, and only permitting alignments in the best alignment "stratum". 76bp single-end reads from embryo replicate A and L1 arrest replicates W and V did not need trimming and were mapped using identical settings. Potential PCR duplicates were removed using software SAMtools (Li et al. 2009). 50bp single end reads are of sufficient length for mapping reads to the *C. elegans* genome.

Identification of DNaseI hypersensitivity peaks and TF footprints and annotation

Raw DNaseI hypersensitive peaks were identified by detecting read enrichment in 150bp consecutive nucleotides using HOTSPOT peak caller specifically designed for DNase-seq (version 3; John et al. 2011). We filtered raw peak calls obtained from HOTSPOT using the irreproducibility discovery rate (IDR) framework developed for ENCODE, which uses a non-parametric copula mixture model to filter peaks into reproducible or irreproducible categories (Li et al. 2011; Landt et al. 2012). Peaks are selected on the combination of their rank or score as well as their consistency across replicates. Peaks overlapping Repeatmasker repeats were omitted. In addition, blacklist regions from ENCODE that represent known ce10 genomic regions exhibiting signal artifacts in ChIP-seq experiments were filtered (ENCODE Project Consortium 2012). Overlapping peaks were also merged. 41,825 and 23,670 DHS peaks were thus found across embryo and L1 arrest biological replicates, respectively. DHS peak locations were annotated in exons (if 75% of region was located in exon), introns, promoter (<300bp from ATG), and intergenic regions (>300bp from ATG) using custom scripts and WormBase WS241 gene models. Pseudogenes, tRNAs, and ncRNAs were excluded from annotation.

Footprints were identified using DNase2TF software package (FDR threshold 0.05) (Sung et al. 2014) and BAM alignment files for each biological replicate in order to identify decreased read coverage within noncoding DHS in regions between 6-40bp with a strand shift in reads. Replicate data within each stage were merged and used to identify additional TF footprints.

Annotation, statistics and data analysis were performed with custom scripts using Python, Ruby, R, Bash scripting, Bedtools (Quinlan and Hall 2010), Bedops (Neph et al. 2012), and pyBedTools (Dale et al. 2011). Visualization of read coverage over normalized gene lengths and *k*-means clustering was performed using DeepTools (Ramirez et al. 2014).

Evaluating enrichment of enhancer marks, sequence conservation in noncoding DHS and gene expression

Sequence conservation is measured by phyloP score across seven related *Caenorhabditis* species. 10,000 randomizations of noncoding DHS from embryo and L1 arrest larvae were performed and compared with observed median phyloP score. Fold enrichment of conservation was calculated against the 97.5th percentile of median phyloP of randomizations. 10,000 randomizations of noncoding DHS from embryo and L1 arrest larvae and overlap with TSS (Chen et al. 2013), modENCODE HOT, CBP-1 embryo H3K4me3 and RNAP II regions (Gerstein et al. 2010) was performed on each randomization and compared with observed median overlap. Fold enrichment of different types of marks (TSS, CBP-1, HOT, RNA Pol II and H3K4me3) in noncoding DHS is calculated against the 97.5th percentile of median overlap from randomizations. Null hypothesis testing was performed with one-sample, two-sided Kolmogorov-Smirnov (KS) tests.

Embryo expression data (Zhong et al. 2010) measured in log2 of fragments per kilobase of exon per million fragments mapped (FPKM) was used to compare expression of higher vs. lower read coverage genes and between genes associated with different categories of noncoding DHS. Genes with varying numbers of noncoding DHS and with or without promoter-enhancer-associated marks were compared by measuring fold changes in expression in the embryo. In order to conservatively estimate magnitude of fold changes of expression, we adjust genes whose expression is below 0.01 FPKM to a more reasonably low level of 0.01 FPKM.

Refining prediction of genes expressed in neuronal and intestinal datasets using *cis*-regulatory motifs located within embryo noncoding DHS

FIMO (Grant et al. 2011) was used to identify sites of known *cis*-regulatory motifs N1 (Ruvinsky et al. 2007), ELT-2 (McGhee et al. 2009), SLR-2 (Kirienko and Fay 2010) and UNC-62 (Van Nostrand al. 2013) within the 2kb 5' and intron regions of *C. elegans* protein-coding genes using threshold $p < 1x10^{-4}$. Of these motif sites, those that were located within noncoding DHS were noted. Genes associated with motif sites were compared against genes enriched in neuronal and intestinal expression datasets (neuronal tiling array data from McGhee et al. 2009; dissected adult intestinal SAGE data from McGhee et al. 2007; FACS embryo intestine tiling array data from Spencer et al. 2011). Percentage of genes correctly predicted by the presence of at least one motif (Motif Only) was compared to that of the presence of at least one motif located within noncoding DHS (Motif and Noncoding DHS). Regardless of number of motifs or noncoding DHS, each gene was counted only once if at least one was present.

Average DNase read profile mapping across *C. elegans* cis-regulatory motifs

FIMO (Grant et al. 2011) was used to identify sites of known *cis*-regulatory motifs within the 2kb 5' regions of protein-coding genes in the *C. elegans* genome using threshold $p < 1x10^{-4}$. For each site, DNase cleavage was measured from start of read alignment (taking into account strand orientation of each read alignment) across a window of 80bp surrounding and including the motif, using scripts included in the pyDNase package (Piper et al. 2013). In this manner, average DNase cleavage was calculated across thousands of sites for a given motif.

Motif discovery

Motifs were identified within DHS peaks and footprints using DREME (Bailey et al. 2011) using E-value threshold 0.05. Entire sequences of DHS peaks and footprints greater than 10bp were used to identify motifs. For footprints less than 10bp, we included 5bp of neighboring genomic sequence. Motifs were compared to curated WormBase *C. elegans* motifs and promoter motifs from Grishkevich et al. (2011) using TOMTOM (Gupta et al. 2007) at thresholds of $q<0.1$ and 0.05. Motif occurrences within noncoding DHS peaks were identified with FIMO using thresholds of $q<0.05$ and 0.025 (Grant et al. 2011).

Gene Ontology and anatomy enrichment analysis

Gene Ontology (GO) analysis was performed on the nearest gene using AmiGO (Gene Ontology Consortium 2000) using p-value threshold 0.05. Only the 50 most enriched terms were considered. Enriched terms were parsed with ReviGO (Supek et al. 2011) to visualize term relatedness and predict biological and molecular function (Appendix Table 2.3).

Anatomy term enrichment was measured using a permutation test for motif-associated genes. Anatomy annotation was obtained from WormBase and only terms with at least 100 genes associated with them were considered. We measured N number of motif-associated genes and counted anatomy terms associated with each gene. For each motif, we performed 10^5 permutations, randomly selecting N genes from the dataset, and measured the number of associated anatomy terms. We then calculated anatomy enrichment probability for each motif, corresponding to the probability that the anatomy term appeared as or more frequently at random compared to observed value. Since lower probability indicates higher enrichment, we used a 0.05 probability threshold to select enriched anatomy terms for each motif.

Differential condition comparison of gene expression between embryo and L1 arrest

Differences in gene expression were analyzed by comparing normalized FPKM data from microarray datasets from Baugh et al. (2009) to compare expression between 6hr L1 starved larvae and embryos, and between 6hr starved L1 larvae and 6hr fed L1 larvae. From expression comparisons we generated lists of genes: top quartile of genes upregulated in 6hr starved L1 larvae versus embryo, and top deciles of genes upregulated and downregulated in response to starvation when comparing expression observed between 6hr starved L1 larvae and 6hr fed L1 larvae. To compare gene expression associated with L1 condition-specific noncoding DHS with embryo condition-specific noncoding DHS, the L1 starved vs. embryo expression ratio of genes associated with each L1 or embryo condition-specific peak were tested with a two-sample, two-sided KS statistical test.

Additional L1 analysis

Statistical enrichment analysis of PHA-4 in L1 DHS was performed by testing 10,000 randomized permutations of L1 DHS (all), noncoding DHS, and for intersection with PHA-4 ChIP-seq peaks from L1 starved larvae stage (Zhong et al. 2010) and testing with the one-sample two-sided KS test. The following gene classes were identified within the L1 noncoding DHS by comparison with existing datasets: DAF-16 target genes (top 3000 DAF-16 target genes from list from Tepper et al. 2013), PHA-4 target genes (gene list of PHA-4 targets in L1 arrest from Zhong et al. 2010), and top quartile of genes upregulated in 6hr starved L1 larvae vs. embryo and top deciles of genes upregulated and downregulated in response to starvation between 6hr starved L1 larvae and 6hr fed L1 larvae (microarray expression data from Baugh et al. 2009). PHA-4 and DAF-16 motifs in specific gene loci were detected using FIMO (Grant et al. 2011) using *p* value thresholds of $5x10^{-5}$ and $1x10^{-4}$, respectively.

Resource and data access

The following data will be made available through WormBase (data files are listed in Appendix Table 2.4). DNase signal tracks from merged sample from either embryo or L1 arrest are shown (total, positive-strand, and negative-strand reads), with additional corresponding tracks for each biological replicate. 2) Tracks for All DHS (post-IDR filtering), Noncoding DHS and TF Footprint regions for embryo and L1 arrest DNase-seq. 3) Gene annotations for each noncoding DHS. 4) L1 condition-specific noncoding DHS and gene annotations. 5) Lists of novel motifs discovered (position-frequency matrices in MEME format. 6) Enriched Gene Ontology and anatomy terms and motif-associated genes for each motif. Read data will be deposited in the NCBI Short Read Archive (SRA).

Competing Interests

Authors declare no competing interests.

Authors' Contributions

Study conception and design MH PWS

Acquisition of data MH

Analysis and interpretation of data MH

Drafting of manuscript MH PWS

Critical revision MH PWS

Acknowledgements

We thank WormBase, especially Xiaodong Wang, Wen Chen, and Gary Williams for curating regulatory motifs and expression data from scientific literature. This research was supported by NIH grant GM084389 to PWS and the Howard Hughes Medical Institute, with which PWS is an investigator. MH was supported by an NSF GRFP predoctoral fellowship. We would like to thank Ali Mortazavi, Ricardo Ramirez, Igor Antoshechkin, Carmie Puckett-Robinson, James Done, John DeModena, Steven Kuntz, Erich Schwarz, Jim McGhee and Erin Osborne Nishimura for assistance and advice on experimental design, sequence library construction, and analysis and interpretation of data. We thank David Angeles for help performing anatomy enrichment analysis. We are grateful to Mihoko Kato, Hillel Schwartz, Pei-Yin Shih and Porfirio Quintero for helpful suggestions on the manuscript.

References

Allen, M.A., Hillier, L., Waterston, R.H. and Blumenthal, T. 2011. A global analysis of trans-splicing in *C. elegans*. Genome Res. *21*, 255-264.

Araya CL, Kawli T, Kundaje A, Jiang L, Wu B, Vafeados D, Terrell R, Weissdepp P, Gevirtzman L, Mace D, et al. 2014. Regulatory analysis of the *C. elegans* genome with spatiotemporal resolution. Nature 512:453-6.

Bailey TL. 2011. DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics 27:1653-9.

Bando T, Ikeda T, Kagawa H. 2005. The homeoproteins MAB-18 and CEH-14 insulate the dauer collagen gene col-43 from activation by the adjacent promoter of the Spermatheca gene sth-1 in *Caenorhabditis elegans*. J Mol Biol 348:101-12.

Banerji J, Rusconi S, Schaffner W. 1981. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. Cell. 27(2 Pt 1):299-308.

Baum PD, Guenther C, Frank CA, Pham BV, Garriga G. 1999. The *Caenorhabditis elegans* gene ham-2 links Hox patterning to migration of the HSN motor neuron. Genes Dev **13**:472-83.

Baugh LR, Hill AA, Slonim DK, Brown EL, Hunter CP. 2003. Composition and dynamics of the *Caenorhabditis elegans* early embryonic transcriptome. Development. 130:889-900.

Baugh LR, Sternberg PW, 2006. DAF-16/FOXO regulates transcription of cki-1/Cip/Kip and repression of lin-4 during *C. elegans* L1 arrest. Curr Biol. 16(8):780-5.

Baugh LR, Demodena J, Sternberg PW. 2009. RNA Pol II accumulates at promoters of growth genes during developmental arrest. Science 324:92-4.

Baugh LR. 2013. To grow or not to grow: nutritional control of development during *Caenorhabditis elegans* L1 arrest. Genetics. 194(3):539-55.

Boulin T, Hobert O. 2012. From genes to function: the *C. elegans* genetic toolbox. Wiley Interdiscip Rev Dev Biol 1:114-37.

Boyle AP, Song L, Lee BK, London D, Keefe D, Birney E, Iyer VR, Crawford GE, Furey TS. 2011. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. Genome Res 21:456-64.

Bulger M, Groudine M. 2010. Enhancers: the abundance and function of regulatory sequences beyond promoters. Dev Biol 339:250-7.

Burghoorn J, Piasecki BP, Crona F, Phirke P, Jeppsson KE, Swoboda P. 2012. The in vivo dissection of direct RFX-target gene promoters in *C. elegans* reveals a novel cis-regulatory element, the C-box. Dev Biol, 368, 415-26.

Chen RA, Down TA, Stempor P, Chen QB, Egelhofer TA, Hillier LW, Jeffers TE, Ahringer J. 2013. The landscape of RNA polymerase II transcription initiation in *C. elegans* reveals promoter and enhancer architectures. Genome Res 23:1339-47.

Chisholm AD, Horvitz HR. 1995. Patterning of the *Caenorhabditis elegans* head region by the Pax-6 family member vab-3. Nature 377:52-5.

Conradt B, Horvitz HR. 1999. The TRA-1A sex determination protein of *C. elegans* regulates sexually dimorphic cell deaths by repressing the egl-1 cell death activator gene. Cell 98:317-27.

Dale RK, Pedersen BS, Quinlan AR. 2011. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. Bioinformatics 27: 3423–3424.

Dupuy D, Bertin N, Hidalgo CA, Venkatesan K, Tu D, Lee D, Rosenberg J, Svrzikapa N, Blanc A, Carnec A, et al. 2007. Genome-scale analysis of in vivo spatiotemporal promoter activity in *Caenorhabditis elegans*. Nat Biotech 25:663-8.

Edgar LG. 1992. Genes controlling specific cell fates in *C. elegans* embryos. Bioessays 14:705-8.

Elemento O, Tavazoie S. 2005. Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. Genome Biol 6:R18

ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature. 489:57-74.

Ferreira HB, Zhang Y, Zhao C, Emmons SW. 1999. Patterning of *Caenorhabditis elegans* posterior structures by the Abdominal-B homolog, egl-5. Dev Biol 207:215-28.

Fox RM, Watson JD, Von Stetina SE, McDermott J, Brodigan TM, Fukushige T, Krause M, Miller DM 3rd. 2007. The embryonic muscle transcriptome of *Caenorhabditis elegans*. Genome Biol 8:R188.

Fraser P, Pruzina S, Antoniou M, Grosveld F. 1993. Each hypersensitive site of the human beta-globin locus control region confers a different developmental pattern of expression on the globin genes. Genes Dev **7**:106-13.

Fukushige T, Hawkins MG, McGhee JD. 1998. The GATA-factor elt-2 is essential for formation of the *Caenorhabditis elegans* intestine. Dev Biol 198:286-302.

Fukushige T, Hendzel MJ, Bazett-Jones DP, McGhee JD. 1999. Direct visualization of the elt-2 gut-specific GATA factor binding to a target promoter inside the living *Caenorhabditis elegans* embryo. Proc Natl Acad Sci 96:11883-8.

Gaudet J, Mango SE. 2002. Regulation of organogenesis by the *Caenorhabditis elegans* FoxA protein PHA-4. Science 295:821-5.

Gaudet J, Muttumu S, Horner M, Mango SE. 2004. Whole-genome analysis of temporal gene expression during foregut development. PLoS Biol 2:e352.

Gaudet J, McGhee JD. 2010. Recent advances in understanding the molecular mechanisms regulating *C. elegans* transcription. Dev Dyn 239:1388-404.

Gems D, Sutton AJ, Sundermeyer ML, Albert PS, King KV, Edgley ML, Larsen PL, Riddle DL. 1998. Two pleiotropic classes of daf-2 mutation affect larval arrest, adult behavior, reproduction and longevity in *Caenorhabditis elegans*. Genetics. 150(1):129-55.

Gene Ontology Consortium. 2000. Gene ontology: tool for the unification of biology. Nat Genet 25:25-9.

Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, et al. 2010. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. Science 330:1775-87.

Gissendanner CR, Kelley K, Nguyen TQ, Hoener MC, Sluder AE, Maina CV. 2008. The *Caenorhabditis elegans* NR4A nuclear receptor is required for spermatheca morphogenesis. Dev Biol 313:767-86.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: Scanning for occurrences of a given motif. Bioinformatics 27:1017–1018.

Grishkevich V, Hashimshony T, Yanai I. 2011. Core promoter T-blocks correlate with gene expression levels in *C. elegans*. Genome Res 21:707-17.

Gross DS, Garrard WT. 1988. Nuclease hypersensitive sites in chromatin. Annu Rev Biochem **57**:159-97.

Gupta S, Stamatoyannopolous JA, Bailey T, Noble WS. 2007. Quantifying similarity between motifs. Genome Biol 8:R24.

Harris TW, Baran J, Bieri T, Cabunoc A, Chan J, Chen WJ, Davis P, Done J, Grove C, Howe K, et al. 2014. WormBase 2014: new views of curated biology. Nucleic

Acids Res 42:D789-93.

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39:311-8.

Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, Thurman RE, Neph S, Kuehn MS, Noble WS, et al. 2009. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nat Methods 6:283-9.

Ihuegbu NE, Stormo GD, Buhler J. 2012. Fast, sensitive discovery of conserved genome-wide motifs. J Comput Biol 19:139–147.

John S, Sabo PJ, Thurman RE, Sung MH, Biddie SC , Johnson TA, Hager GL, Stamatoyannopoulos JA. 2011. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nat Genet 43:264-268.

Kagoshima H, Cassata G, Tong YG, Pujol N, Niklaus G, Bürglin TR. 2013. The LIM homeobox gene ceh-14 is required for phasmid function and neurite outgrowth. Dev Biol 380:314-23.

Kalb JM, Lau KK, Goszczynski B, Fukushige T, Moons D, Okkema PG, McGhee JD. 1998. pha-4 is Ce-fkh-1, a fork head/HNF-3alpha,beta,gamma homolog that functions in organogenesis of the *C. elegans* pharynx. Development 125:2171-80.

Kirienko NV, McEnerney JD, Fay DS. 2008. Coordinated regulation of intestinal functions in *C. elegans* by LIN-35/Rb and SLR-2. PLoS Genet 4:e1000059.

Kirienko NV, Fay DS. 2010. SLR-2 and JMJC-1 regulate an evolutionarily conserved stress-response network. EMBO J 29:727-39.

Kirouac M, Sternberg PW. 2003. cis-Regulatory control of three cell fate-specific genes in vulval organogenesis of *Caenorhabditis elegans* and C. briggsae. Dev Biol. 257(1):85-103.

Krause M, Harrison SW, Xu SQ, Chen L, Fire A. 1994. Elements regulating cell- and stage-specific expression of the *C. elegans* MyoD family homolog hlh-1. Dev Biol 166:133-48.

Krivega I, Dean A. 2012. Enhancer and promoter interactions— long distance calls. Curr Opin Genet Dev 22:79-85.

Kruesi WS, Core LJ, Waters CT, Lis JT, Meyer BJ. 2013. Condensin controls recruitment of RNA polymerase II to achieve nematode X-chromosome dosage compensation. Elife. 2:e00808.

Kuntz SG, Schwarz EM, DeModena JA, De Buysscher T, Trout D, Shizuya H, Sternberg PW, Wold BJ. 2008. Multigenome DNA sequence conservation identifies Hox cis-regulatory elements. Genome Res 18:1955-68.

Kuntz SG, Williams BA, Sternberg PW, Wold BJ. 2012. Transcription factor redundancy and tissue-specific regulation: evidence from functional and physical network connectivity. Genome Res 22:1907-19.

Lei H, Liu J, Fukushige T, Fire A, Krause M. 2009. Caudal-like PAL-1 directly activates the bodywall muscle module regulator hlh-1 in *C. elegans* to initiate the embryonic muscle gene regulatory network. Development 136:1241-9.

Landmann F, Quintin S, Labouesse M. 2004. Multiple regulatory elements with spatially and temporally distinct activities control the expression of the epithelial differentiation gene lin-26 in *C. elegans*. Dev Biol. 265:478-90.

Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, et al. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res 22: 1813–1831.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25.

Li, H. Handsaker, B. Wysoker, A. Fennell, T. Ruan, J. Homer, N. Marth, G. Abecasis, G. Durbin, R. and 1000 Genome Project Data Processing Subgroup. 2009. The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25:2078-9.

Li Q, Brown JB, Huang H, and Bickel PJ. 2011. Measuring reproducibility of high-throughput experiments. Ann Appl Stat 5, 1699-2264

Maxwell CS, Antoshechkin I, Kurhanewicz N, Belsky JA, Baugh LR, 2012. Nutritional control of mRNA isoform expression during developmental arrest and recovery in *C. elegans*. Genome Res. 22(10):1920-9.

Maxwell CS, Kruesi WS, Core LJ, Kurhanewicz N, Waters CT, Lewarch CL, Antoshechkin I, Lis JT, Meyer BJ, Baugh LR, 2014. Pol II docking and pausing at growth and stress genes in *C. elegans*. Cell Rep. 6(3):455-66.

McGhee JD, Sleumer MC, Bilenky M, Wong K, McKay SJ, Goszczynski B, Tian H, Krich ND, Khattra J, Holt RA, et al. 2007. The ELT-2 GATA-factor and the global regulation of transcription in the *C. elegans* intestine. Dev Biol 302:627-45.

McGhee JD, Fukushige T, Krause MW, Minnema SE, Goszczynski B, Gaudet J, Kohara Y, Bossinger O, Zhao Y, Khattra J, et al. 2009. ELT-2 is the predominant transcription factor controlling differentiation and function of the *C. elegans* intestine, from embryo to adult. Dev Biol 327:551-65.

Mercer TR, Edwards SL, Clark MB, Neph SJ, Wang H, Stergachis AB, John S, Sandstrom R, Li G, Sandhu KS, et al. 2013. DNaseI-hypersensitive exons colocalize with promoters and distal regulatory elements. Nat Genet 45:852-9.

Muñoz MJ, Riddle DL. 2003. Positive selection of *Caenorhabditis elegans* mutants with increased stress resistance and longevity. Genetics. 163(1):171-80.

Murphy CT, McCarroll SA, Bargmann CI, Fraser A, Kamath RS, Ahringer J, Li H, Kenyon C. 2003 Genes that act downstream of DAF-16 to influence the lifespan of *Caenorhabditis elegans*. Nature. 424(6946):277-83

Murray JI, Boyle TJ, Preston E, Vafeados D, Mericle B, Weisdepp P, Zhao Z, Bao Z, Boeck M, Waterston RH. 2012. Multidimensional regulation of gene expression in the *C. elegans* embryo. Genome Res 22:1282-94.

Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E, Maurano MT, Vierstra J, Thomas S, et al. 2012. BEDOPS: high-performance genomic feature operations. Bioinformatics 28:1919-20.

Niu W, Lu ZJ, Zhong M, Sarov M, Murray JI, Brdlik CM, Janette J, Chen C, Alves P, Preston E, et al. 2011. Diverse transcription factor binding features revealed by genome-wide ChIP-seq in *C. elegans*. Genome Res 21:245-54.

Noonan JP, McCallion AS. 2010. Genomics of long-range regulatory elements. Annu Rev Genomics Hum Genet 11:1-23.

Okkema PG, Harrison SW, Plunger V, Aryana A, Fire A. 1993. Sequence requirements for myosin gene expression and regulation in *Caenorhabditis elegans*. Genetics. 135:385-404.

Okkema PG, Fire A. 1994. The *Caenorhabditis elegans* NK-2 class homeoprotein CEH-22 is involved in combinatorial activation of gene expression in pharyngeal muscle. Development 120: 2175–2186.

Okkema PG, Krause M. 2005. Transcriptional regulation. WormBook. 1-40.

O'Meara MM, Bigelow H, Flibotte S, Etchberger JF, Moerman DG, Hobert O. 2009. Cis-regulatory mutations in the *Caenorhabditis elegans* homeobox gene locus cog-1 affect neuronal development. Genetics. 181:1679-86.

Page BD, Guedes S, Waring D, Priess JR. 2001. The *C. elegans* E2F- and DP-related proteins are required for embryonic asymmetry and negatively regulate Ras/MAPK signaling. Mol Cell **7**:451-60.

Piper J, Elze MC, Cauchy P, Cockerill PN, Bonifer C, Ott S. 2013. Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. Nucleic Acids Res 41:e201.

Puckett Robinson C, Schwarz EM, Sternberg PW. 2013. Identification of DVA interneuron regulatory sequences in *Caenorhabditis elegans*. PLoS One. 8(1):e54971.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841-2.

Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. 2014. deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res 42:W187-91

Ray P, Schnabel R, Okkema PG. 2008. Behavioral and synaptic defects in *C. elegans* lacking the NK-2 homeobox gene ceh-28. Dev Neurobiol 68:421-33.

Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, et al. 2000. Genome-wide location and function of DNA binding proteins. Science 290:2306-9.

Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, et al. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 4:651–657.

Ruvinsky I, Ohler U, Burge CB, Ruvkun G. 2007. Detection of broadly expressed neuronal genes in *C. elegans*. Dev Biol 302:617-26.

Sengupta AK, Kuhrs A, Müller J. 2004. General transcriptional silencing by a Polycomb response element in Drosophila. Development. 131(9):1959-65.

Shi B, Guo X, Wu T, Sheng S, Wang J, Skogerbø G, Zhu X, Chen R. 2009. Genome-scale identification of *Caenorhabditis elegans* regulatory elements by tiling-array mapping of DNase I hypersensitive sites. BMC Genomics 10:92.

Solari F, Bateman A, Ahringer J. 1999. The *Caenorhabditis elegans* genes egl-27 and egr-1 are similar to MTA1, a member of a chromatin regulatory complex, and are redundantly required for embryonic patterning. Development 126:2483-94.

Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee BK, Sheffield NC, Gräf S, Huss M, Keefe D, et al. 2011. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res 21:1757-67.

Spencer WC, Zeller G, Watson JD, Henz SR, Watkins KL, McWhirter RD, Petersen S, Sreedharan VT, Widmer C, Jo J, et al. 2011. A spatial and temporal map of *C. elegans* gene expression. Genome Res 21:325-41.

Steiner FA, Henikoff S. Cell type-specific affinity purification of nuclei for chromatin profiling in whole animals. Methods Mol Biol. 2015;1228:3-14.

Sullivan AM, Arsovski AA, Lempe J, Bubb KL, Weirauch MT, Sabo PJ, Sandstrom R, Thurman RE, Neph S, Reynolds AP, et al. 2014. Mapping and Dynamics of Regulatory DNA and Transcription Factor Networks in A. thaliana. Cell Rep 8:2015–2030.

Sulston JE, Schierenberg E, White JG, Thomson JN. 1983. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. Dev Biol 100:64-119.

Sung MH, Guertin MJ, Baek S, Hager GL. 2014. DNase footprint signatures are dictated by factor dynamics and DNA sequence. Mol Cell. 56(2):275-85.

Supek F, Bošnjak M, Škunca N, Šmuc T. 2011. REVIGO summarizes and visualizes long lists of Gene Ontology terms. PLoS ONE 6:e21800.

Tepper RG, Ashraf J, Kaletsky R, Kleemann G, Murphy CT, Bussemaker HJ, 2013. PQM-1 complements DAF-16 as a key transcriptional regulator of DAF-2-mediated development and longevity. Cell. 154(3):676-90.

Thomas S, Li XY, Sabo PJ, Sandstrom R, Thurman RE, Canfield TK, Giste E, Fisher W, Hammonds A, Celniker SE, et al. 2011. Dynamic reprogramming of chromatin accessibility during Drosophila embryo development. Genome Biol 12:R43.

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B et al. 2012. The accessible chromatin landscape of the human genome. Nature. 489(7414):75-82.

Tsuboi D, Qadota H, Kasuya K, Amano M, Kaibuchi, K. 2002. Isolation of the interacting molecules with GEX-3 by a novel functional screening. Biochem Biophys Res Commun 292:697-701.

Tuan D, Solomon W, Li Q, London IM. 1985. The "beta-like-globin" gene domain in human erythroid cells. Proc Natl Acad Sci U S A. 82(19):6384-8.

Van Auken K, Weaver D, Robertson B, Sundaram M, Saldi T, Edgar L, Elling U, Lee M, Boese Q, Wood WB. 2002. Roles of the Homothorax/Meis/Prep homolog UNC-62 and the Exd/Pbx homologs CEH-20 and CEH-40 in *C. elegans* embryogenesis. Development 129:5255-68.

Van Gilst MR., Hadjivassiliou H, Yamamoto KR. 2005. A *Caenorhabditis elegans* nutrient response system partially dependent on nuclear receptor NHR-49. Proc. Natl. Acad. Sci. USA 102: 13496–13501

Van Nostrand EL, Sánchez-Blanco A, Wu B, Nguyen A, Kim SK. 2013. Roles of the developmental regulator unc-62/Homothorax in limiting longevity in *Caenorhabditis elegans*. PLoS Genet. 9(2):e1003325.

Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, et al. 2009. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature 457:854-8.

Wenick AS, Hobert O. 2004. Genomic cis-regulatory architecture and trans-acting regulators of a single interneuron-specific gene battery in *C. elegans*. Dev Cell 6:757-70.

Zhao G, Schriefer LA, Stormo GD. 2007. Identification of muscle- specific regulatory modules in *Caenorhabditis elegans*. Genome Res 17:348-357.

Zhang CC, Bienz M. Segmental determination in Drosophila conferred by hunchback (hb), a repressor of the homeotic gene Ultrabithorax (Ubx). Proc Natl Acad Sci U S A. 1992 Aug 15;89(16):7511-5.

Zhong M, Niu W, Lu ZJ, Sarov M, Murray JI, Janette J, Raha D, Sheaffer KL, Lam HY, Preston E, et al. 2010. Genome-wide identification of binding sites defines distinct functions for *Caenorhabditis elegans* PHA-4/FOXA in development and environmental response. PLoS Genetics 6:e1000848.

Zhou HM, Walthall WW. 1998. UNC-55, an orphan nuclear hormone receptor, orchestrates synaptic specificity among two classes of motor neurons in *Caenorhabditis elegans*. J Neurosci 18:10438-44.

Chapter 3

High-throughput and massively parallel functional testing of *Caenorhabditis elegans* enhancers

Margaret C. W. Ho[♪], Hillel T. Schwartz[♪], Paul W. Sternberg[*]

[♪] This project and its experimental design were conceived by both authors equally. MH generated the DNase-seq dataset of putative *C. elegans* CRMs, wrote the design scripts and constructed the oligo libraries for enhancer testing. HS performed the reporter construct cloning and will also perform oligo library cloning, transgenic injections, and collection of DNA and poly-A RNA for sequencing. MH will computationally analyze all sequence data. MH, HS and PWS will write the paper.

Division of Biology and Biological Engineering, Howard Hughes Medical Institute, California Institute of Technology, Pasadena, CA

[*] Corresponding author: pws@caltech.edu

Introduction

Mass functional testing of enhancers in metazoans is a challenge, owing to the inefficiency of performing individual reporter gene assays. In this study I test methods for massively parallel reporter analysis of candidate *Caenorhabditis elegans* enhancers and see if we can functionally validate putative *cis*-regulatory modules (CRMs) from a systematic screen of DNase-hypersensitivity sites (DHS) in *Caenorhabditis elegans* embryos (Ho and Sternberg, submitted).

I have taken two approaches inspired by two previous studies, massively parallel reporter assay (MPRA) and self-transcribed active regulatory regions sequencing (STARR-seq), previously performed in mammalian cell lines and *Drosophila* S2 cells, respectively (Melnikov et al. 2012; Arnold et al. 2013). In these studies, large numbers of putative enhancer sequences are generated as oligonucleotide libraries and cloned into individual constructs. These constructs are then pooled and used to test the ability of these sequences to drive reporter gene expression in transiently transfected cell lines. DNA and poly-adenylated (poly-A) RNA are simultaneously extracted from transfected cells, fragmented, and sequenced via shotgun sequencing. Each candidate enhancer sequence can be identified in the RNA and in the DNA by its sequence (in the case of STARR) or associated unique barcode (in the case of MPRA). These studies thus utilize high throughput sequencing as a cheap and powerful readout of functional activity of enhancer-driven transcription.

The MPRA approach used a mixture of custom single strand (ss) DNA oligos synthesized in parallel on oligo arrays as a source of sequences for the enhancer libraries, which is versatile and relatively cheap (Melnikov et al., 2012). While these custom oligos are currently limited to only a few hundred base pairs in length, oligos around 200bp in size should be sufficient for testing the typical size of *Caenorhabditis elegans* enhancers. Using custom oligos also allows the variation of sequence within the library to test point mutations or larger changes in sequence and observe the effect on function. The oligos are then designed to include unique barcodes associated with each enhancer sequence, such that the barcode is encoded downstream of the reporter gene and will be present in the resulting mRNA transcript. The presence of the barcode thus indicates that the associated enhancer was able to drive reporter gene expression. Furthermore, designing custom oligos with sequence tags flanking the sequence enables multiplexed synthesis of separate libraries of oligos that can be individually amplified from the mixture using primers designed against the sequence tags.

In the STARR-seq study by Arnold et al. (2013) the putative enhancer sequence is cloned downstream of the reporter gene, so that if it is able to activate transcription of the reporter gene, the enhancer itself is transcribed as well. This assay design allows direct detection of the enhancer sequence in the RNA-seq data, and removes the need for unique barcodes for each enhancer sequence. The published STARR-seq research used fragmentation of genomic DNA as the source of putative enhancer sequences. While this is convenient, cheap and relatively unbiased as a method of obtaining sequences, it does not allow control of the exact sequences to be tested and does not enable selective mutational analysis of the sequences. However, it does allow longer sequences to be tested in the assay and the results are especially useful for defining the boundaries of enhancer activity since the read profile around STARR-seq peaks should reflect the boundaries of sequences that are able to drive expression.

There are a few potential caveats in doing these types of parallel assays. It is still unknown whether, when testing these sequences in parallel, there is complete independence between individual reporters on a transgene array or if there are any potential interactions between reporters. This is important to consider when applying these approaches to *C. elegans* since injecting transgenic reporters into *C. elegans* generates extrachromosomal arrays with complex and heritable structure with rearrangement and recombination (Mello et al. 1991). Another important consideration is whether these assays are able to give quantitative information as well as qualitative information about enhancer activity. In the MPRA study in mammalian cell lines, Melnikov and colleagues (2012) were able to systematically dissect a synthetic cAMP-regulated enhancer (CRE) and a virus-inducible enhancer of human interferon-β (IFNB) using scanning mutagenesis across the sequences of these enhancers and testing effects of these mutations on enhancer activity. Enhancer activity was measured by the abundance of barcodes in RNA-seq data and was normalized to the representation of the barcodes in DNA. In STARR-seq, Arnold and colleagues compared the activity of different enhancers using similar methods of normalization to DNA. Both of these studies were performed in cell lines whose uniformity of cell type facilitated gathering of quantitative information. Application of these assays to a multicellular system containing many different tissues and cell types, as in the case of transgenic *Caenorhabditis elegans,* will make quantitative comparison challenging, since sequences are likely to function as enhancers in some tissues but not others. However, even qualitative information from MPRA and STARR in *C. elegans* as to whether test sequences are able to function as enhancers will be very useful, since it will allow high throughput parallel screening. Enhancers that show significant activity above threshold in MPRA or STARR can be isolated for individual characterization.

Our assay design in *C. elegans* makes use of 200bp single stranded custom DNA oligos synthesized on microarrays by Agilent. I designed our oligo order to contain 27,000 oligo sequences with approximately 10^8 copies of each individual oligo. I used 7bp unique barcodes, each present in two copies as a tandem repeat (we henceforth refer to this as the 2x7bp barcode). I disallow A at positions 2 and 5 within the barcode so as to avoid a polyA signal in the barcode, giving a maximum of 9216 barcodes possible within each library. The purpose of providing two tandem copies of the 7bp barcode is to guard against potential sequencing errors.

Each *C. elegans* MPRA oligo contains 120bp of test enhancer sequence followed by a spacer and paired barcodes, flanked by library amplification tags (Figure 3.1). The final reporter construct positions the 120bp test enhancer sequence upstream of minimal Δ*pes-10* promoter sequence followed by *gfp* reporter gene, the unique 2x7bp barcode, and the 3' UTR sequence from the *C. elegans* gene *rla-1*, which encodes a ribosomal subunit essential for animal viability likely to be expressed in all cell types (Sönnichsen et al. 2005). The Δ*pes-10* promoter is a deleted promoter that is sensitive to enhancer activities commonly used in *C. elegans* enhancer assays (1995 Fire Vector Kit[3]). If the test enhancer is able to drive reporter gene expression, its associated barcode will also be transcribed and detectable in the RNA-seq data.

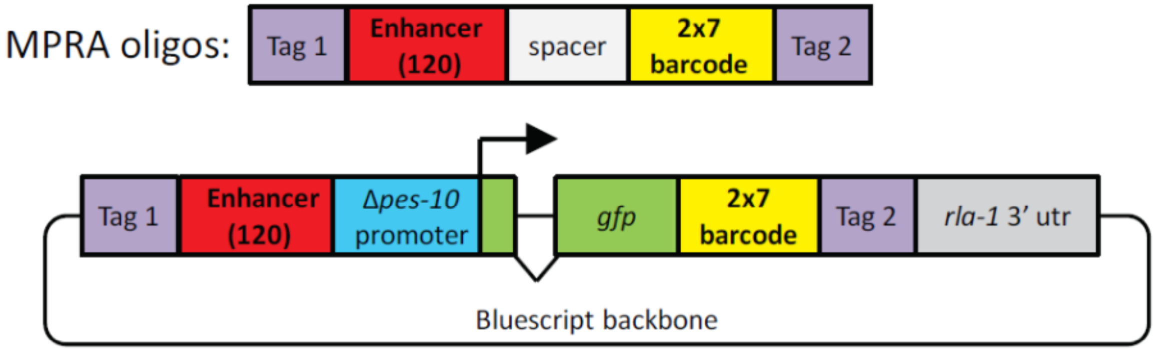


Figure 3.1 Custom oligo design and reporter construct for *C. elegans* MPRA scheme, based on Melnikov et al. (2012)

C. elegans MPRA oligos are designed to test 120bp of sequence for enhancer activity. Sequence tags are 19bp apiece, with a 12bp spacer. The *rla-1* 3' UTR includes some genomic sequence and is 67bp in length.

[3] Dr. Andrew Fire, https://www.addgene.org/firelab/

Each *C. elegans* STARR oligo contains 100bp of test enhancer sequence, along with a uniquely associated 2x7 bp barcode and the 3' UTR from *rla-1*, flanked by library amplification tags (Figure 3.2). The final reporter construct positions the unique 2x7bp barcode, the *rla-1* 3' UTR, and the 100bp test enhancer downstream of the Δ*pes-10* minimal promoter and *gfp* reporter gene, so that the barcode is transcribed and detectable in RNA-seq data.

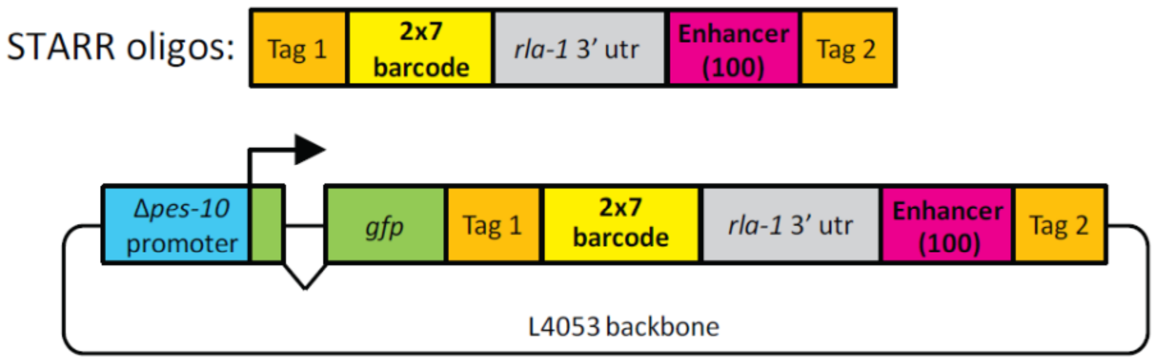


Figure 3.2 Custom oligo design and reporter construct for *C. elegans* STARR scheme, based on Arnold et al. (2013)

C. elegans STARR oligos are designed to test 100bp of sequence for enhancer activity. Sequence tags are 19bp apiece and the *rla-1* 3' UTR is 48bp in length.

My experimental libraries were designed to test putative enhancer sequences from DNase hypersensitivity sites previously found in *C. elegans* embryos (Ho and Sternberg, submitted). I designed a pilot set of 3,056 oligos each for the MPRA and STARR schemes containing individual libraries for noncoding DHS near hypoxia-regulated and uniquely expressed gut genes and separate individual libraries for versions of these noncoding DHS with mutations in predicted HIF-1 binding sites or motifs of intestinally-regulated TFs (ELT-2, MAB-3, DAF-16, SLR-2, SKN-1, TRA-1) (Table 3.1). HIF-1 is an ortholog of mammalian hypoxia-induced factor HIF-1, and is required for survival in low-oxygen environments (Jiang et al. 2001). In addition, I also designed two additional libraries apiece for MPRA and for STARR in order to test 38% of the 26,644 putative CRMs that I identified in *C. elegans* embryos (see Chapter 2). I chose to use hypoxia genes and uniquely expressed gut genes in order to facilitate comparison of the results of MPRA or STARR assays in transgenic worms in normal conditions or with changes to their environment. For example, we could expose worms to hypoxic conditions (24 hours in low oxygen conditions such as 1% O_2) or changes in diet, such as feeding them DA1877 strain of *Comamonas*, a bacterial food source which has been shown to induce many diet-induced phenotypic effects in *C. elegans* compared to feeding worms the OP50 strain of *E. coli* (MacNeil et al. 2013 a,b).

Materials and Methods

For both MPRA and STARR approaches in *C. elegans*, oligo libraries are PCR amplified using primers designed against library tags (Figure 3.3). Then they are cloned into base reporter vectors (see Figures 3.4 and 3.5 for details). Plasmid DNA is then prepared from pooled transformants. A Δ*pes-10::gfp* reporter cassette is then cloned into the relevant restriction sites of the collected plasmids containing MPRA oligos, and plasmid is prepared from the resulting pooled transformants. Having now completed the STARR and MPRA libraries, each is linearized by *Asi*SI digest. Each wild-type (WT) or mutant oligo library of hypoxia and gut enhancers was synthesized in separate libraries, so that WT and mutated versions of candidate enhancers could be separated into different libraries and thus limit the chances for mispriming and template switching during PCR amplification. Once reporter assembly is complete, these paired libraries will be combined for injection.

Libraries of linearized constructs will be injected into the gonads of young adult *C. elegans pha-1 (ts)* hermaphrodites at a concentration of roughly 20-40 ng/uL linearized transgene with approximately 20 ng/uL of *pha-1* rescue construct and 60 ng/uL of carrier DNA. The carrier DNA is used to minimize the incidence of individual reporters directly abutting one another and help maintain the extrachromosomal array (Evans 2006) and will be composed of DNA ladder or possibly digested *C. elegans* genomic DNA with *Asi*SI-compatible ends. Depending on success of microinjection and transgene maintenance Hillel may adjust the concentrations of the transgene injection mixture. The temperature-sensitive mutant *pha-1(e2123)* is viable at 15°C but is inviable at 25°C, allowing us to use it as a selective marker (Granato et al. 1994). Injected worms are grown at 25°C; transgenic F_1 animals rescued for the *pha-1* phenotype are picked and used to establish stable transgenic lines, which are tracked to ensure that they arise from different injected P_0s and represent independent transgenesis events. Transgenic worms are pooled from individual transgenic lines and mixed stages are isolated from these pools and simultaneous extraction of DNA and polyA RNA is performed. DNA and polyA-RNA samples are fragmented. In the case of polyA RNA, priming with universal primers to generate cDNA will be performed. Then fragmented DNA and cDNA samples will be prepared into libraries and multiplex sequenced on a HiSeq II (Illumina, San Diego, CA) using single-read shotgun sequencing.

Reads below quality threshold Q20 using FastQC[4] will be removed for quality control. I will then align reads to a database of our linearized reporter constructs using Bowtie to identify reads that align to our database, and remove any unaligned reads (which may come from genomic DNA, endogenous RNA transcripts, or contaminant RNA/DNA).

For MPRA and STARR reads, I will count the representation of barcodes in the DNA and RNA-seq reads and we will normalize the count detected for each barcode in the RNA-seq data to the relative abundance of that barcode in DNA reads. I will use DNA read alignment to the database of linearized reporter constructs to estimate representation of test enhancers in the extrachromosomal

[4] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

array. I will consider enhancers and barcodes that are detected in the DNA-seq data but whose barcodes do not appear in the RNA-seq data above the threshold set by our negative control (non-enhancer test sequences; see Table 3.2) to be inactive under the conditions tested. Those enhancers and barcodes that appear in the DNA-seq data and whose barcodes appear in the RNA-seq data above the threshold set by non-enhancer negative control sequences will be considered enhancers. I will compare WT and mutant libraries to determine if there are differences as a result of the mutations in HIF-1 or intestinal TF motifs in test enhancers.

Custom scripts to assemble the unique 7bp barcode sets and to construct oligo sequences were written in Ruby with biopieces[5] and Bash with BEDTOOLS (Quinlan and Hall 2010). Since our cloning scheme for the oligos makes use of 8-cutter REs *Pac*I, *Not*I, *Asc*I, *Asi*SI, all sequence tags, spacers, 3' UTRs and plasmid backbones were chosen to be free of these RE sites and enhancer sequences were cleared of these RE sites with 1-2bp mutations.

[5] https://code.google.com/p/biopieces/

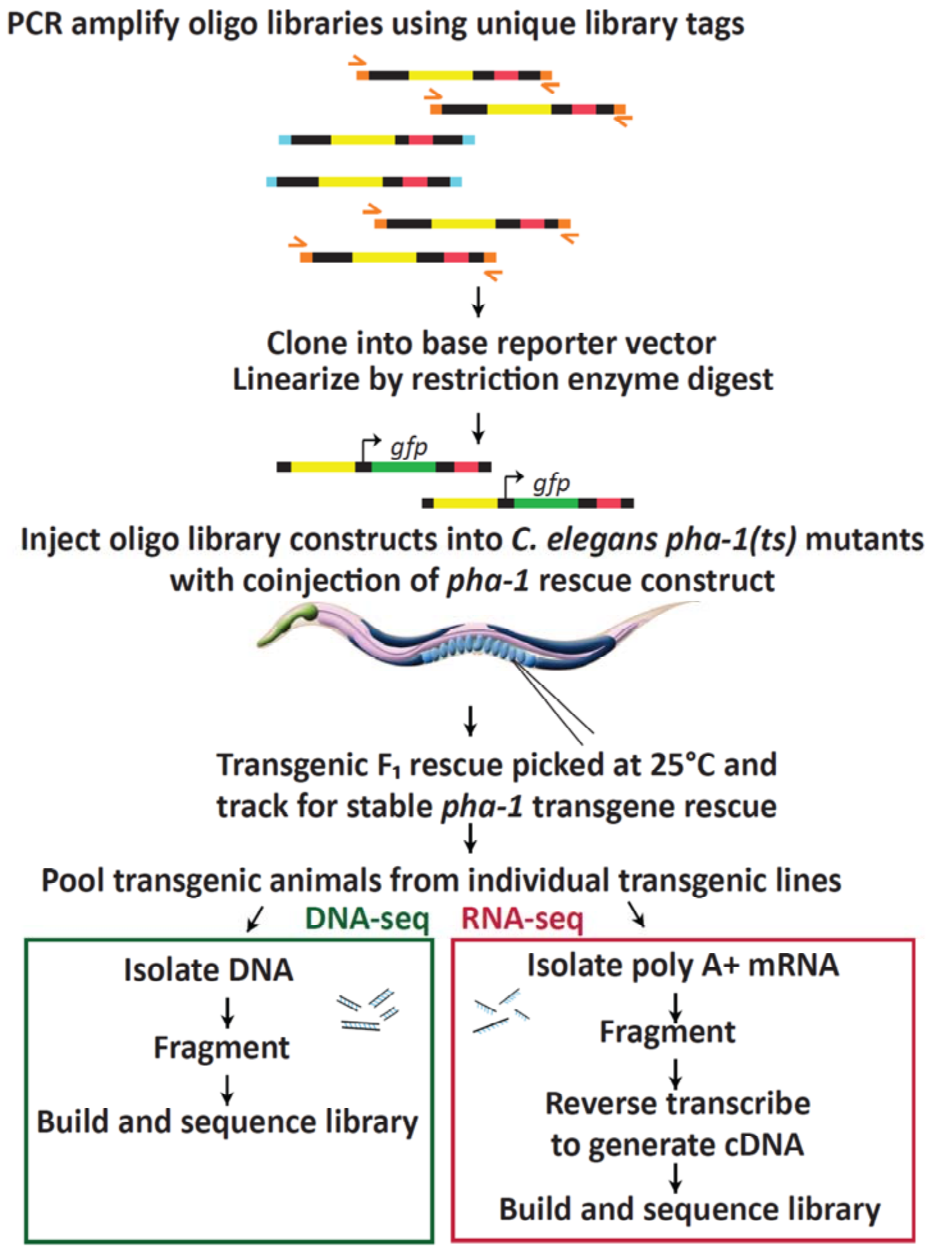

PCR amplify oligo libraries using unique library tags
Clone into base reporter vector
Linearize by restriction enzyme digest
gfp
gfp
Inject oligo library constructs into *C. elegans pha-1(ts)* mutants
with coinjection of *pha-1* rescue construct
Transgenic F_1 rescue picked at 25°C and
track for stable *pha-1* transgene rescue
Pool transgenic animals from individual transgenic lines
DNA-seq
RNA-seq
Isolate DNA
Fragment
Build and sequence library
Isolate poly A+ mRNA
Fragment
Reverse transcribe
to generate cDNA
Build and sequence library

Figure 3.3 Experimental scheme for MPRA and STARR in *C. elegans*

Primers corresponding to sequence tags for each library are used to amplify the library (orange library being amplified, green library is not amplified in this example). Test enhancer sequences on each oligo are shown in yellow with a uniquely associated barcode in red.

I designed our oligos to test enhancers by taking the central 100bp (in the case of STARR) and 120 bp (in the case of MPRA) of noncoding DNaseI hypersensitive sites (DHS) from *C. elegans* DNase-seq data from embryos (Ho and Sternberg, submitted). DNase-seq uses high throughput sequencing of DNaseI-treated chromatin to identify noncoding regions of a few hundred base pairs in length in the *C. elegans* genome that are accessible to DNaseI cleavage and which represent putative *cis*-regulatory modules. To obtain test enhancers for the gut and hypoxia oligo libraries, I took lists of genes expressed in the intestine (McGhee et al. 2007), filtered to remove genes in muscle and neuronal tissue datasets (muscle and neural SAGE from Meissner et al. (2009)), and genes regulated by hypoxia either in a HIF-1 dependent or independent manner (Shen et al. 2005). I then selected embryo noncoding DHS located in the 2.5kb region surrounding these genes (Ho and Sternberg, submitted) for test enhancers in wild type (WT) libraries (Table 3.1).

For mutant libraries, 8bp mutations were made in the sequences of any ELT-2 (McGhee et al. 2009), MAB-3 (Yi and Zarkower 1999), DAF-16 (Furuyama et al. 2000), SLR-2 (Kirienko and Fay 2010), SKN-1 (Blackwell 1994), or TRA-1 (Zarkower and Hodgkin 1993) motifs found in gut noncoding DHS and any HIF-1 motifs found in HIF-1 dependent hypoxia noncoding DHS (Table 3.1).

For the oligo libraries to test the majority of noncoding DHS found in *C. elegans* embryos, I removed the lowest 10% scoring noncoding DHS. From the remaining 90% noncoding DHS, I randomly selected 10,196 noncoding DHS, representing 38% of the total embryo noncoding DHS. 8,000 of these noncoding

DHS were used to design embryo DHS library A and 2,196 were used to design embryo DHS library B (Table 3.1).

Library	Number of regions	Total oligos in library
MPRA – Hypoxia – WT	364†	538‡
MPRA – Hypoxia – mutant	53	177
MPRA – Gut – WT	1,908	2,082*
MPRA – Gut – mutant	135	259
STARR – Hypoxia – WT	364†	538‡
STARR – Hypoxia – mutant	53	177
STARR – Gut – WT	1,908	2,082*
STARR – Gut – mutant	135	259
MPRA2 – embryo DHS A	8,000	8,124
MPRA2 – embryo DHS B	2,196	2,320
STARR – embryo DHS A	8,000	8,124
STARR – embryo DHS B	2,196	2,320
Total	25,312	27,000

Table 3.1 List of MPRA and STARR oligo libraries.

For every oligo library there are 124 control sequence oligos (see Table 3.2). WT refers to wild-type.

† includes 179 HIF-1 dependent and 185 independent noncoding DHS sequences

‡ includes 50 duplicate Hypoxia WT (HIF-1 dependent) noncoding DHS sequences with unique barcodes, see explanation below

* includes 50 duplicate Gut WT noncoding DHS sequences with unique barcodes, see explanation below

I also selected regions to use as controls in our oligo set. These included 79 embryo noncoding DHS near additional gut-expressed genes not otherwise represented in the library, and four negative control sequences within the N5 and N6 non-enhancer regions tested by Kuntz et al. (2008). I also included 41 embryo noncoding DHS within known enhancers from *ceh-13, lin-39, myo-2, hlh-1*, and *lin-26* loci (Table 3.2).

I also duplicated 50 of the test enhancer sequences in each of the gut WT and hypoxia WT libraries (Table 3.1) with additional unique barcodes as a control for reproducibility of measured enhancer activity between constructs using different barcodes.

Control regions	*Number of regions*
Positive Controls (DHS of additional gut-expressed genes)	79
Negative Controls (Kuntz *et al.* 2008 negative regions)	4
Other Controls (*ceh-13/lin-39* Hox enhancers, *myo-3*, *myo-2*, *hlh-1, lin-26* regulatory regions)	41

Table 3.2 List of control sequence oligos added to every library
A total of 124 control sequence oligos are added to every MPRA or STARR library with unique 2x7bp barcodes. Four negative control regions within N5 and N6 non-enhancers found by Kuntz et al. (2008) were included, along with 79 positive controls of noncoding DHS of additional gut expressed genes not included in the gut gene set and 41 other controls of noncoding DHS corresponding mapping to regulatory regions of well-studied *C. elegans* genes.

The set of unique 7bp barcodes (repeated in tandem in order to guard against sequencing errors, referred to as 2x7bp) were designed with no adenine (A) allowed in positions 2 and 5 within the barcode to prevent any poly-A signal in the barcode. This resulted in 9216 possible unique barcodes in a given library. Each oligo library used unique barcodes within this set, but we re-used barcodes between different oligo libraries, although not between hypoxia and gut or between wild-type and mutant iterations of the same library. Each oligo library is amplified from the mixture of 27,000 custom oligos using unique forward and reverse primers corresponding to Tag 1 and reverse complement of Tag 2 flanking oligo sequences (Table 3.3).

Primer Set	*Forward (Sense of Tag1)*	*Reverse (RC of Tag2)*
MPRA – Hypoxia – WT	ACTGACCCTGACCCTGACC	TCCTGTGCCTGTGCCTGTG
MPRA – Hypoxia – mutant	ACCAGGACCAGGACCAGAC	AGGAGCAGTAGCAGGAGCC
MPRA – Gut – WT	ACACAGCCACAGCCACAGC	CAGACGGAGACGGAGACGG
MPRA – Gut – mutant	TTGGTCCTGGTCTTGATCG	TCCGACTCTGGCTCTGTCG
STARR – Hypoxia – WT	TCTCTGCCTCTGCCTCTGC	TCAGTCCCAGTCCCAGTCC
STARR – Hypoxia – mutant	ACGGTCACGGTCACAGTTC	AGCCAGAGCAAGAGCCAAG
STARR – Gut – WT	AGGACACGGACACGGACAC	TACACCGACACCGACACCG
STARR – Gut – mutant	CGTCCTCGACCTCGTAATG	TGACCTGGACCTTGACCTC
MPRA2 – embryo DHS A	GAAGGGCTGGGAAGACACC	TCCCATCGGTAGCGTGGAG
MPRA2 – embryo DHS B	GCTGGCTTGGCGAATGTGC	CGGTTCGGATCGAGGCTTC
STARR – embryo DHS A	CCGACCACGACTCAACTGG	GGACCGGAGTGCTGTCTAC
STARR – embryo DHS B	GCCGCACTCTCACCTACTC	GAGGCAGGCACTTCGGTTG

Table 3.3 Sequence Tags for oligo libraries

Each sequence tag is 19bp and amplification primers are designed

Preliminary Results

I prepared the noncoding DHS sequences and mutations, wrote scripts, and constructed oligo libraries for manufacture by Agilent. We have received the custom oligos and are presently in the cloning stage and sample preparation of the project, which will be largely handled by my co-author Dr. Hillel Schwartz. I will analyze all resulting DNA and RNA-seq sequence data and write a computational pipeline for all subsequent analyses.

The detailed cloning scheme for *C. elegans* MPRA is as follows (Figure 3.4). Hillel has generated the base MPRA vector by cloning the *rla-1* 3'UTR into Bluescript using *Not*I and *Bam*HI sites. He has PCR amplified MPRA oligo libraries, adding *Not*I and *Asi*SI sites and will clone the library into the prepared base MPRA vector which has been digested with *Not*I and *Pac*I. This is possible because *Asi*SI ligates to *Pac*I and results in a site that is not recleavable by either *Pac*I or *Asi*SI. He has PCR amplified Δ*pes-10::gfp* from L4053 and will clone the reporter gene cassette into the cloned oligo MPRA library.

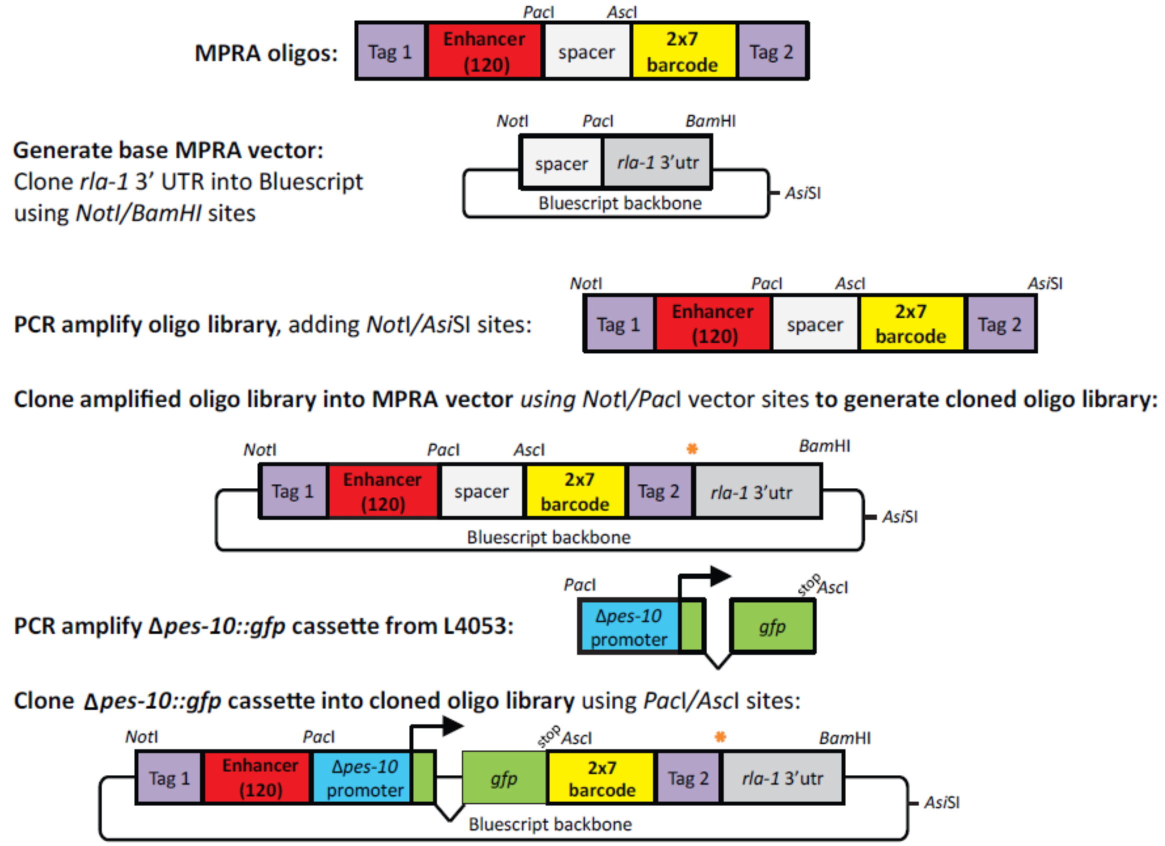


Figure 3.4 Cloning scheme for *C. elegans* MPRA oligos

The orange asterisk (*) refers to the site which remains after *Asi*SI ligates to *Pac*I and which is not recleavable by either enzyme. The *Asi*SI site present in the base MPRA vector and resulting reporter construct is roughly 200bp downstream from the *rla-1* 3' UTR and is used to linearize the construct prior to injection.

The detailed cloning scheme for *C. elegans* STARR is as follows (Figure 3.5). Hillel has generated the base STARR vector by cloning a *Not*I and *Asi*SI cassette into L4053, replacing the *unc-54* 3'UTR. He has PCR amplified the STARR oligo library, adding *Not*I and *Asi*SI sites. He will clone the library into the STARR vector using *Not*I and *Pac*I sites in the vector, which is possible because *Asi*SI ligates to *Pac*I and results in a site that is not recleavable by either *Pac*I or *Asi*SI. The reporter construct libraries are then linearized by digest with *Asi*SI to a site that is around 200bp downstream of Tag 2.

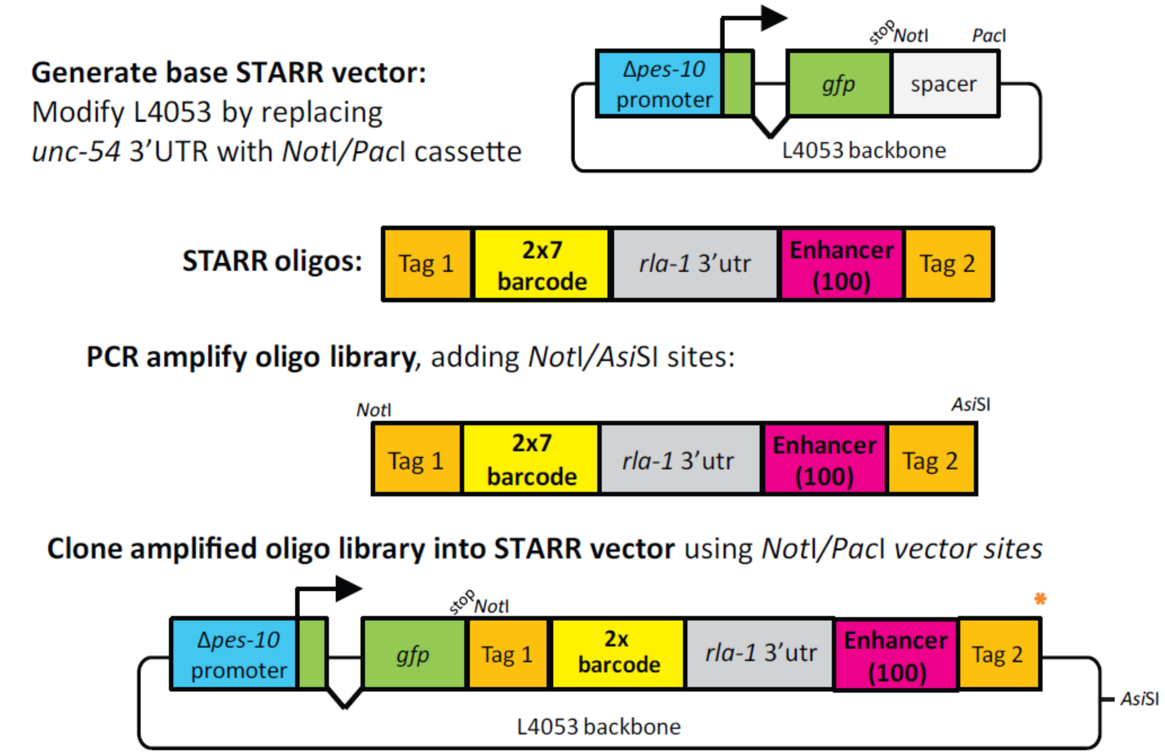


Figure 3.5 Cloning scheme for *C. elegans* STARR oligos

The orange asterisk (*) refers to the site remaining after *Asi*SI ligates to *Pac*I and which is not recleavable by either enzyme. The *Asi*SI site present in the final STARR reporter construct is roughly 200bp downstream from Tag 2 and is used to linearize the construct prior to injection.

Our findings from cloning a previous oligo set found that oligos that truncate during synthesis and that remain in the mixture can generate polymerization products that amplify off of other templates. Thus our current experimental design has been revised to clean up the oligo mixture and minimize any unnecessary PCR amplification. We linearize the plasmid using restriction enzyme digest instead of PCR amplification. We are also size-selecting the oligo library to remove some truncated oligos using an SPRI-style method. In the presence of a “crowding agent” polyethylene glycol and NaCl will allow negatively charged DNA (in this case, ssDNA) to bind to carboxyl groups on a paramagnetic bead surface. We use a 150bp size selection kit from NVIGEN, but it is equivalent to Solid Phase Reversible Immobilisation (SPRI) and Ampure (Beckman Coulter) kit protocols (DeAngelis et al. 1995). We have altered our oligo design such that any homologous sequence or fixed region among oligos is on the 3’ end. We have designed our mutant sequences to be in separate oligo libraries from wild-type sequences to prevent any potential mixing up of barcodes within the library. Finally, we use an excess of primers in PCR to shift the equilibrium towards DNA synthesis from free primer and template instead of annealing of partial products.

References

Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science. 339(6123):1074-7.

Arnold CD, Gerlach D, Spies D, Matts JA, Sytnikova YA, Pagani M, Lau NC, Stark A. 2014. Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during *cis*-regulatory evolution. Nat Genet. 46(7):685-92

Blackwell TK, Bowerman B, Priess JR, Weintraub H. 1994. Formation of a monomeric DNA binding domain by Skn-1 bZIP and homeodomain elements. Science. 266(5185):621-8.

DeAngelis MM, Wang DG, Hawkins TL. 1995. Solid-phase reversible immobilization for the isolation of PCR products. Nucleic Acids Res. 23(22):4742-3.

Evans, T. C., ed. Transformation and microinjection. 2006. *WormBook*, ed. The *C. elegans* Research Community, WormBook.

Furuyama T, Nakazawa T, Nakano I, Mori N. 2000. Identification of the differential distribution patterns of mRNAs and consensus binding sequences for mouse DAF-16 homologues. Biochem J. 349(Pt 2):629-34.

Granato M, Schnabel H, Schnabel R. 1994. *pha-1*, a selectable marker for gene transfer in *Caenorhabditis elegans*. Nucleic Acids Res. 122(9):1762-3.

Ho MC and Sternberg PW. Submitted. Genome-wide discovery of active regulatory elements and transcription factor footprints in *Caenorhabditis elegans* embryos and L1 arrest larvae using deep sequencing of DNaseI hypersensitivity regions.

Jiang H, Guo R, Powell-Coffman JA. 2001. The *Caenorhabditis elegans hif-1* gene encodes a bHLH-PAS protein that is required for adaptation to hypoxia. Proc Natl

Acad Sci U S A, 98, 7916-21.

Kirienko NV, Fay DS. 2010. SLR-2 and JMJC-1 regulate an evolutionarily conserved stress-response network. EMBO J 29:727-39.

Kuntz SG, Schwarz EM, DeModena JA, De Buysscher T, Trout D, Shizuya H, Sternberg PW, Wold BJ. 2008. Multigenome DNA sequence conservation identifies Hox cis-regulatory elements. Genome Res 18:1955-68.

Macneil LT, Walhout AJ. 2013a. Food, pathogen, signal: The multifaceted nature of a bacterial diet. Worm. 2(4):e26454.

MacNeil LT, Watson E, Arda HE, Zhu LJ, Walhout AJ. 2013b. Diet-induced developmental acceleration independent of TOR and insulin in *Caenorhabditis elegans*. Cell. 153(1):240-52.

McGhee JD, Fukushige T, Krause MW, Minnema SE, Goszczynski B, Gaudet J, Kohara Y, Bossinger O, Zhao Y, Khattra J, et al. 2009. ELT-2 is the predominant transcription factor controlling differentiation and function of the *Caenorhabditis elegans* intestine, from embryo to adult. Dev Biol 327:551-65.

Meissner B, Warner A, Wong K, Dube N, Lorch A, McKay SJ, Khattra J, Rogalski T, Somasiri A, Chaudhry I, et al. 2009. An integrated strategy to study muscle development and myofilament structure in *Caenorhabditis elegans*. PLoS Genet. 5(6):e1000537.

Mello CC, Kramer JM, Stinchcomb D, Ambros V. 1991. Efficient gene transfer in *Caenorhabditis elegans*: extrachromosomal maintenance and integration of transforming sequences. EMBO J. 10(12):3959-70.

Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG Jr, Kinney JB, et al. 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat Biotechnol. 30(3):271-7.

Shen C, Nettleton D, Jiang M, Kim SK, Powell-Coffman JA. 2005. Roles of the HIF-1 hypoxia-inducible factor during hypoxia response in *Caenorhabditis elegans*. J Biol Chem. 280(21):20580-8.

Sönnichsen B, Koski LB, Walsh A, Marschall P, Neumann B, Brehm M, Alleaume AM, Artelt J, Bettencourt P, Cassin E, et al. 2005. Full-genome RNAi profiling of early embryogenesis in *Caenorhabditis elegans*. Nature. 434(7032):462-9.

Yi W, Zarkower D. 1999. Similarity of DNA binding and transcriptional regulation by *Caenorhabditis elegans* MAB-3 and Drosophila melanogaster DSX suggests conservation of sex determining mechanisms. Development. 126(5):873-81.

Zarkower D, Hodgkin J. 1993. Zinc fingers in sex determination: only one of the two *Caenorhabditis elegans* Tra-1 proteins binds DNA in vitro. Nucleic Acids Res. 21(16):3691-8.

Chapter 4

Application of DNase-seq to an entomopathogenic nematode, *Steinernema carpocapsae*

Introduction

Having demonstrated proof-of-principle DNase-seq in *C. elegans*, I am applying the technique to study *Steinernema carpocapsae*, a distant nematode relative. *S. carpocapsae* and related nematodes of the *Steinernema* genus are a class of parasitic nematodes called entomopathogenic nematodes (EPNs) that have evolved an association with insect-pathogenic bacteria (reviewed in Dillman and Sternberg, 2012). Together, the nematode, acting as a vector, and its mutualistic bacterial pathogen, are able to rapidly kill their insect hosts. This distinctive association between nematode and bacterial pathogen is found among many nematode species, but are best studied in two genera, *Heterorhabditis* and *Steinernema*. EPNs have a lifecycle in which infective juvenile (IJ) stage individuals seek out and infect an insect host and release their payload of insect-pathogenic bacteria into the nutrient-rich internal environment (Figure 4.1). The bacteria proliferate and rapidly kill the host, creating an ideal environment for the nematodes to develop and reproduce. When all resources in the insect host have been consumed, the new generation of IJs is able to escape from the dead host and seek out the next insect host. The EPN lifestyle appears in several multiple distantly related genera including *Steinernema*, *Heterorhabditis* and reportedly *Oscheius* as well (reviewed in Dillman and Sternberg 2012 and Dillman et al. 2012c).

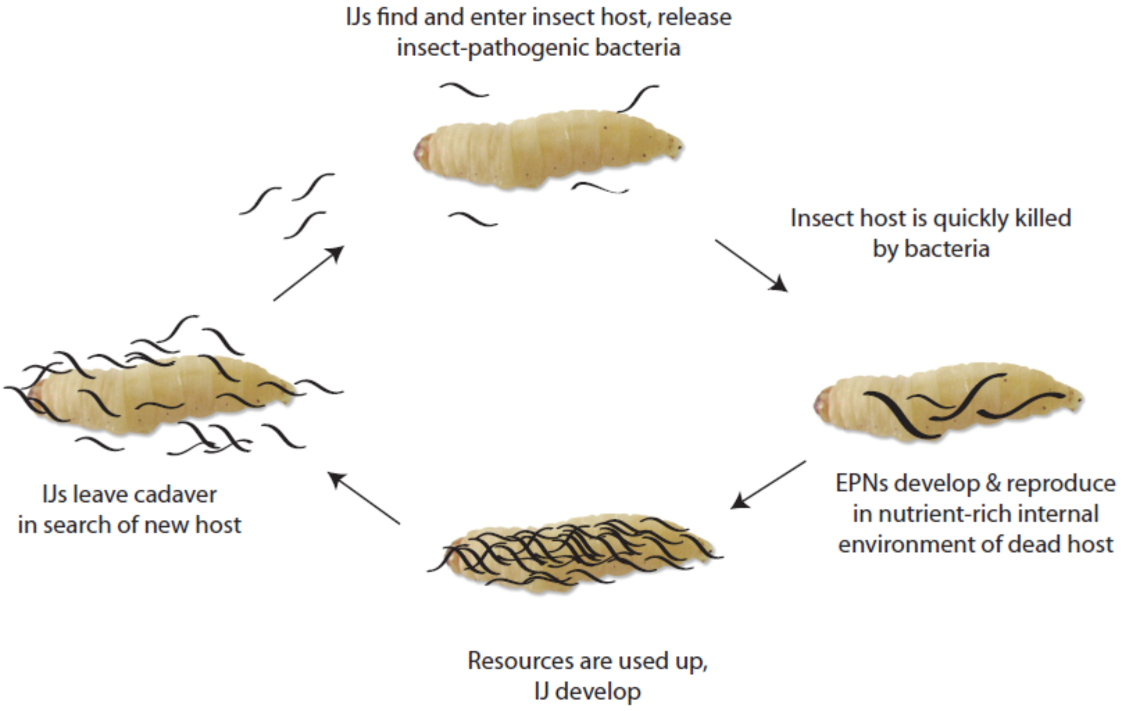

IJs find and enter insect host, release insect-pathogenic bacteria
Insect host is quickly killed by bacteria
EPNs develop & reproduce in nutrient-rich internal environment of dead host
Resources are used up, IJ develop
IJs leave cadaver in search of new host

Figure 4.1 Lifecycle of entomopathogenic nematodes (EPNs)
The infective juvenile (IJ stage) is a developmentally arrested third-larval stage of EPNs that is stress-resistant and the only stage that is free-living. IJs seek out and infect insect hosts, releasing insect-pathogenic bacteria once inside. The insect pathogenic bacteria rapidly proliferates and kills the host. In the host, IJs develop into adults and will reproduce. When all resources are used, a new generation of IJs will develop and emerge from the cadaver to seek new hosts. Figure adapted from Dillman and Sternberg 2012.

Nematode species in the *Steinernema* genus are members of Clade IV and share with *Caenorhabditis elegans* a common ancestor that lived several hundred million years ago (Figure 4.2; Dillman, Macchietto, submitted). While most *C. elegans* are hermaphrodites, *Steinernema* species are gonochoristic. *Steinernema* nematodes are a fascinating model for insect parasitism as well as for bacteria-host associations. Studies in *S. carpocapsae* and other EPNs have shed light on olfaction and host-seeking behavior of parasitic nematodes (Hallem et al. 2011; Dillman et al. 2012a).

Five *Steinernema* species (*S. carpocapsae*, *S. feltiae*, *S. glaseri*, *S. monticolum*, *S. scapterisci*), of which are all EPNs, have had their genomes and transcriptomes sequenced (Dillman, Macchietto et al. Submitted). These data allow evolutionary comparisons to be made among *Steinernema* species to locate protein-coding genes that may facilitate parasitism within this group, mechanisms that facilitate partnership between mutualistic *Xenorhabdus* bacteria and the *Steinernema* host nematode, and differences among *Steinernema* species in their host range and responses to different host odors (Dillman et al 2012a). Furthermore, important comparisons can be made to the best studied nematode species, *C. elegans*. An example of this is comparison of the Hox genes, which are an important class of transcription factors that regulate development in metazoans. Nematodes have lost many Hox genes compared to metazoan (Aboobaker and Blaxter a,b) but between *C. elegans*, *Panagrellus redivivus*, and the five *Steinernema* species, there appears to be good conservation among five of the six *C. elegans* Hox genes (*ceh-13, lin-39, mab-5, egl-5,* and *php-3*), whereas *nob-1* appears to have been lost in these species (Figure 4.3; Dillman, Macchietto et al. Submitted).

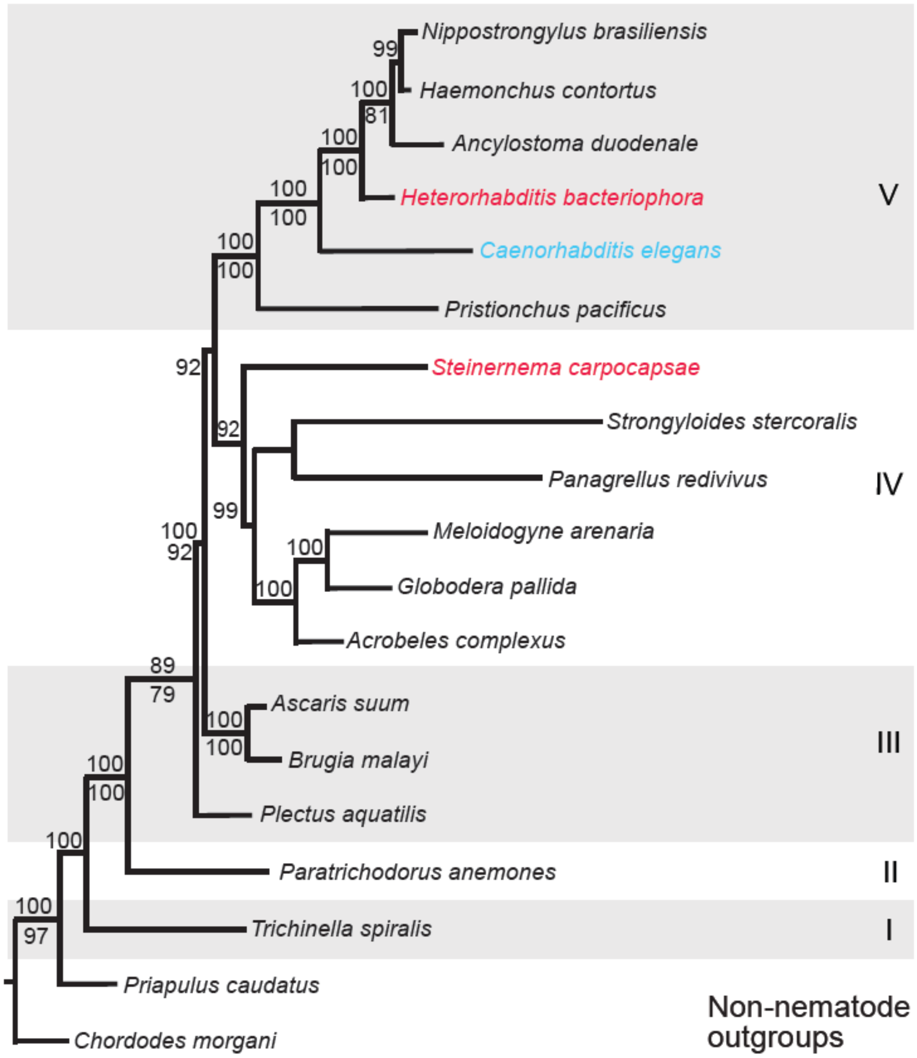
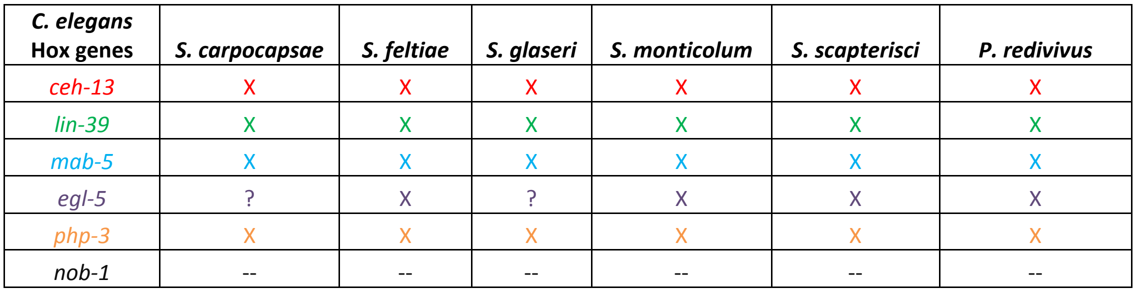


Figure 4.2 Phylogenetic position of *Steinernema carpocapsae* in Nematoda. EPNs are shown in red. *Steinernema carpocapsae* is in Clade IV, while *Heterorhabditis bacteriophora* and *Caenorhabditis elegans* (shown in blue) are in Clade V. Other nematodes in Clades I through V are shown in black, as well as non-nematode outgroup species. Figure adapted from Hallem et al. 2011.

Studies of conservation of noncoding DNA between *C. elegans and Steinernema* gene orthologs have elucidated many novel conserved noncoding regulatory motifs (Dillman, Macchietto et al. Submitted). There is otherwise little known about *cis*-regulatory sequences in *Steinernema* nematodes. Returning to the example of the Hox genes, the 22kb intergenic region between *lin-39* and *ceh-13* Hox genes in *C. elegans* has been well-studied by Kuntz and colleagues (2008) who characterized 11 enhancers (N1-N4, N7-N11, I4 and I8) that were able to drive gene expression in transgenic reporter assays. In *S. carpocapsae*, the *lin-39/ceh-13* intergenic region is much larger at around 35kb, and harbors multiple additional protein-coding genes which do not appear to be related in function to the Hox genes (Figure 4.4). It would be very interesting to examine whether these *lin-39/ceh-13* enhancers found by Kuntz and colleagues (2008) possess orthologs in *S. carpocapsae* and are functionally conserved in their ability to drive either *lin-39* or *ceh-13* expression.



C. elegans Hox genes	*S. carpocapsae*	*S. feltiae*	*S. glaseri*	*S. monticolum*	*S. scapterisci*	*P. redivivus*
ceh-13	X	X	X	X	X	X
lin-39	X	X	X	X	X	X
mab-5	X	X	X	X	X	X
egl-5	?	X	?	X	X	X
php-3	X	X	X	X	X	X
nob-1	--	--	--	--	--	--

Figure 4.3. Conservation of Hox genes in *Steinernema* nematodes

Five out of six *C. elegans* Hox genes appear to be conserved in Clade IV *Steinernema* and *Panagrellus* nematodes. Hox gene *nob-1* appears to have been lost. Adapted from Dillman, Macchietto et al. (submitted).

DNase-seq, a method that uses high-throughput sequencing of DNaseI-treated chromatin, is able to identify regions of a few hundred base pairs long as putative *cis*-regulatory sequences and has been successfully tested in *C. elegans* (Ho and Sternberg, submitted). I was able to identify seven (N1-N4, N8, N10, N11) out of nine conserved enhancers using DNaseI hypersensitive sites (DHS) I observed in *C. elegans* embryos, in addition to two additional enhancers (I4, I8) that were also identified by Kuntz and colleagues (2008) but which were not highly conserved on the sequence level. I have been applying DNase-seq to *S. carpocapsae* IJs in order to identify and study *cis-regulatory* sequences in this distantly related nematode which is new to the study of functional regulatory genomics. The IJ stage is the most easily collected stage of *S. carpocapsae* and is of particular interest because of its well-characterized host-seeking behavior.

Materials and Methods

Steinernema carpocapsae nematodes (strain All) were grown and maintained using standard culture methods (White 1927). In this culture method, five last-instar larvae of the waxmoth (*Galleria mellonella)* were placed on top of a disc of 55 mm Whatman 1 filter paper to serve as a pseudo-soil substrate in a 5 cm Petri dish. 300μl containing 500-1000 *S. carpocapsae* IJs suspended in water was evenly distributed on the filter paper to infect the waxmoth larvae. After 7-10 days the insect cadavers were transferred to White traps, in which IJs would emerge after 3-5 days (White 1927). Emerging IJs were collected and washed for 30 minutes in 0.4% Hyamine 1622 solution (Fluka), rinsed three times with water, and then once with 1X PBS. IJs were then frozen at -80°C. Ten to fifteen plates of *S. carpocapsae* were prepared at a time, yielding roughly 1.5 to 2 mL of packed IJs.

S. carpocapsae IJs were thawed and ground to fine powder with mortar and pestle over dry ice to break IJs open and isolate nuclei. Samples were reconstituted in nuclei purification buffer (0.1% Triton-X, spermine, spermidine, and protease inhibitor) and dounced for 30 strokes with a tight-fitting pestle on ice (nuclei isolation protocol from INTACT method; Steiner and Henikoff et al. 2015). Samples were spun at 0.1 *g* for 10 minutes to separate from debris, and purified further by spinning 10 minutes at 1000 *g* over a cushion of Optiprep (60% iodixanol) at 4°C. Isolated nuclei were visualized using DAPI staining.

Equal aliquots of *S. carpocapsae* IJ nuclei were treated with 0, 20, 40, 80, 120, 160 U/mL DNaseI in 1X DNaseI digestion buffer (containing $CaCl_2$, spermine, spermidine, protease inhibitor) for 3 minutes at 37°C. DNaseI digestion conditions follow the Stamatoyannopoulos lab protocol (Thurman et al. 2012). DNaseI digestion was quenched by adding STOP buffer containing 20mg/mL Proteinase K and incubating 55°C overnight. The samples were then treated with 45ug/mL boiled RNase A for 30 minutes. DNA was purified and concentrated using column purification and run on 1% agarose gel stained with Sybr Gold. The gel piece containing DNA fragments less than 500bp in size was purified. DNA yield was measured using a Qubit fluorometer. See Appendix 1 for adapted DNaseI protocol.

QPCR primers (Table 4.1) were designed against *S. carpocapsae* sequences from *lin-39/ceh-13* Hox intergenic regions showing conservation in 23 out of 30 consecutive nucleotides using multi-species sequence analysis using ungapped transitive alignments (MUSSA) analysis between *S. carpocapsae* and *C. elegans* (Figure 4.5). Six "negative" control regions from the *lin-39/ceh-13* Hox complex were chosen that were not conserved between *S. carpocapsae* and *C. elegans* and between *S. carpocapsae* and *S. feltiae.* In addition, QPCR primers were also designed against 100bp upstream noncoding "promoter" regions of predicted *S. carpocapsae* FAR (fatty acid- and retinol-binding protein) genes that are known to be highly expressed in IJ stage: g24938 and g8883 (Dillman, Macchietto et al., submitted). QPCR amplicon sizes ranged from 70bp to 97bp.

Primer	Forward	Reverse	Amplicon Size (bp)
g24938	TCGCTTTTGTGTTTCTCTAATTGAA	TGGTTGTAAAGAAGACGGTTGG	75
g8883	TGGATTCGGAACAGGAAAAA	AGTTCACGACCGCTGCTAGT	70
N3_3	GTGACTACCCGTTGACACCTG	GGAAGTTTCAGAAAACGATGGA	77
N3_3 *lin-39* proximal	GTAGTCCGAGGACGGGTTAAG	AGTCTCTCTTCTCGCCTGAATCT	89
N7_1	CAGAGAACGCGTGATTGTTG	GTTCCAAGCCACCTTTCCTT	83
I8	AGCAATCCTATGGAATTCTCCAC	AGCGTTACAAAAATTGCCAAAA	90
N9 3'	GGCTTCAAAGCAAGAAATATCAAT	CAGCAGCCCGAATTTTCATA	80
N10_1	GGGTGACCTGTAGCCGTTTT	CGAACTCCGTCCGTATCACT	83
N10_2	GAGGGAGCGGAGATAACGAT	TGTAAATGCGCCTCCTTACC	75
N11	TCGATCGCAAAAGAAGAGTTG	CTCCCATCAGAGTTCCAACAA	77
Neg1	AGGCGATCGAGGAAGAAGAG	TGAATCCGTTTTCCTCCAAG	97
Neg4	ATGGCGCAAGGATTTGAGTA	GTGCAGGCGACTTGCAGAT	94
Neg7A	ACGTCGTCTGGTTAGGATGTG	TGTTCAGAACGCCATCTTTGT	90
Neg8A	AGCTGGACGATTGTTTGAGG	GACGCGATGCACTTCGTATT	76
Neg9A	TGGTATCAAGATCTCCGTGTGA	CAGGCGTTGATGGATGTTCT	78
Neg9B	TCGACGCCCATTAATTAGATCA	TGATACCAGTGTTGGTTAACATGC	79

Table 4.1 QPCR primers for *S. carpocapsae* DNase-seq

QPCR primers are designed against the promoters of two FAR genes, eight conserved noncoding regions in *lin-39/ceh-13* intergenic region, and six non-conserved regions (as negative control).

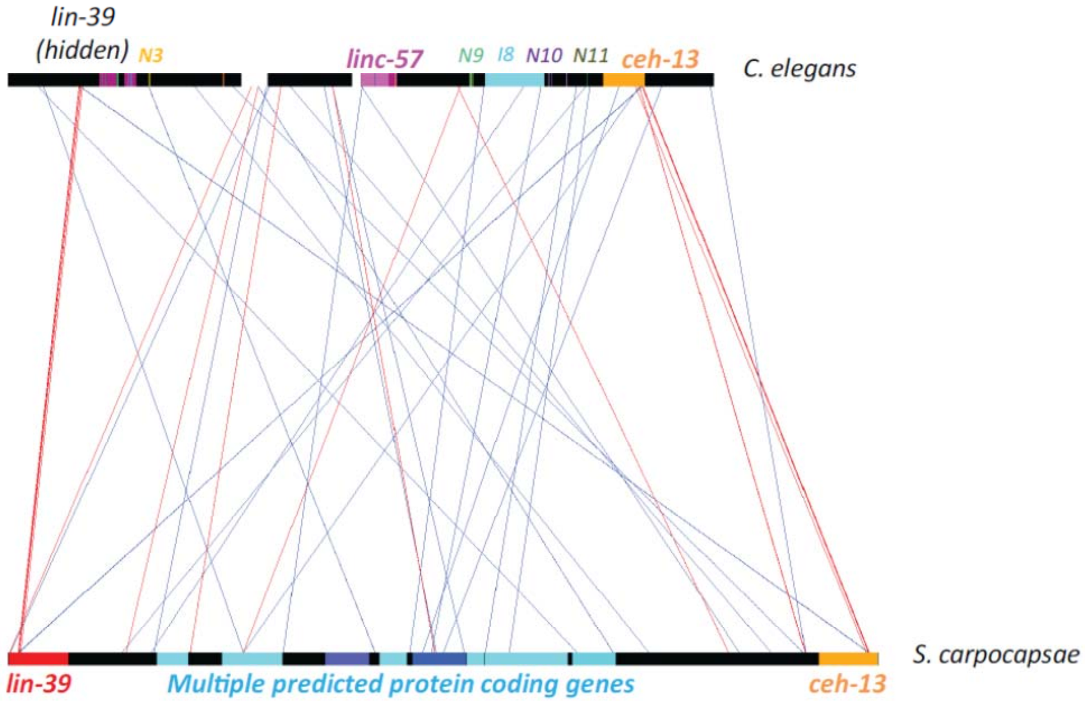


Figure 4.4 Multi-species sequence analysis using ungapped transitive alignments (MUSSA) between *ceh-13*/*lin-39* Hox gene clusters in *C. elegans* and *S. carpocapsae*

Alignment between *lin-39* (red) and *ceh-13* (orange) intergenic region in *C. elegans* and *S. carpocapsae* at a threshold of 23 conserved nucleotides out of 30 consecutive nucleotides. Long intergenic noncoding RNA *linc-57* (purple) in *C. elegans* is shown in purple, as well as the locations of known enhancers (N3, N9, I8, N10, and N11). Multiple predicted protein coding genes are shown in blue and dark blue in *S. carpocapsae*. Red lines in MUSSA analysis (Kuntz et al. 2008) indicate sense alignment between sequences. Blue lines indicate reverse complement alignment between sequences.

Preliminary Results

S. carpocapsae IJs were collected (see Materials and Methods) and frozen at -80°C, yielding enough material (roughly 12 mL of packed IJs, with around 3-4 mL packed IJs or 1.5 million worms in each experiment) for three to four DNase-seq biological replicates. One biological replicate (Replicate 1) has been treated with DNaseI and fully processed to the point of being ready for sequencing library preparation, with three other biological replicates in progress.

QPCR is performed using duplicate genomic DNA standards and absolute derivative measurement of C_p. Relative fold enrichment of regulatory regions was measured in samples by normalizing the observed concentration of each region by the mean of negative control regions (Neg1, Neg4, Neg7, Neg8, Neg 9A and B).

QPCR verification was performed on Replicate 1 *S. carpocapsae* IJ DNase seq samples (Figure 4.5). The sample showing relatively consistent levels of higher regulatory enrichment is the one treated with 160U/mL DNase-seq, with a final yield of 290 ng when measured using Qubit. Regulatory enrichment is highest for subregion showing homology to the N10 enhancer of *C. elegans*, with enrichment also observed for regions with homology to N3, N7, and N11 enhancers of *C. elegans lin-39/ceh-13* Hox genes. We do not observe regulatory enrichment for the two FAR genes in this sample. Regulatory enrichment as measured by QPCR is a proxy for the relative level of DNase hypersensitivity observed in the experiment and it is possible the region that we chose in the FAR promoter is not highly accessible, or that our negative control regions (which we presume are non-enhancers) are still relatively DNaseI-accessible.

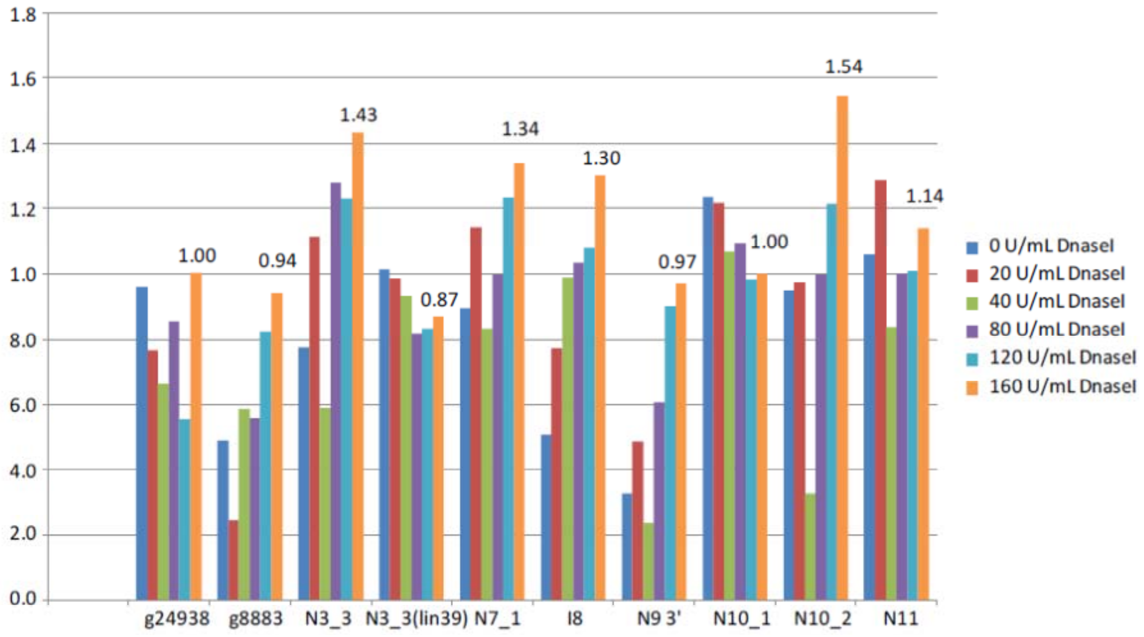


Figure 4.5. QPCR results from *S. carpocapsae* IJ Replicate 1

QPCR analysis shows a maximum regulatory enrichment of around 1.5 fold for the 160U/mL DNaseI sample. The highest enrichment is seen for a subregion showing homology to the N10 enhancer of *C. elegans*. Similar levels of enrichment are seen for other conserved regions with homology to the N3, N7, and N11 enhancers.

References

Aboobaker AA, Blaxter ML. 2003a. Hox Gene Loss during Dynamic Evolution of the Nematode Cluster. Curr Biol. 13(1):37-40.

Aboobaker A, Blaxter M. 2003b. Hox gene evolution in nematodes: novelty conserved. Curr Opin Genet Dev. 13(6):593-8.

Dillman AR, Macchietto M, Porter CF, Rogers A, William BA, Antoshechkin I, Lee M, Goodwin Z, Lu X, Lewis EE, Goodrich-Blair H, et al. Comparative genomics of Steinernema reveals deeply conserved regulatory networks in nematodes. Submitted.

Dillman AR, Sternberg PW. 2012. Entomopathogenic nematodes. Curr Biol. 22(11):R430-1.

Dillman AR, Guillermin ML, Lee JH, Kim B, Sternberg PW, Hallem EA. 2012a. Olfaction shapes host-parasite interactions in parasitic nematodes. Proc Natl Acad Sci U S A. 109(35):E2324-33.

Dillman AR, Mortazavi A, Sternberg PW. 2012b. Incorporating genomics into the toolkit of nematology. J Nematol. 44(2):191-205.

Dillman AR, Chaston JM, Adams BJ, Ciche TA, Goodrich-Blair H, Stock SP, Sternberg PW. 2012c. An entomopathogenic nematode by any other name. PLoS Pathog. 2012;8(3):e1002527.

Hallem EA, Dillman AR, Hong AV, Zhang Y, Yano JM, DeMarco SF, Sternberg PW. 2011. A sensory code for host seeking in parasitic nematodes. Curr Biol. 21(5):377-83.

Kuntz SG, Schwarz EM, DeModena JA, De Buysscher T, Trout D, Shizuya H, Sternberg PW, Wold BJ. 2008. Multigenome DNA sequence conservation identifies Hox cis-regulatory elements. Genome Res 18:1955-68.

White GF. 1927. A Method for Obtaining Infective Nematode Larvae from Cultures. Science. 66:302-303.

Chapter 5

Conclusions

C. elegans and other nematodes provide a fertile system to investigate *cis*-regulatory control of gene expression during development and its evolution across different species. This is particularly true at present since recent methods are now making it possible to systematically interrogate *cis*-regulatory function across the entire genomes of nematodes. Methods to discover *cis*-regulatory modules (CRMs) have been developed to be higher throughput and also transcription factor (TF)-agnostic, allowing additional CRMs regulated by TFs beyond those that are well-known to be studied. At the same time, sequencing data is high resolution, allowing the pinpointing of sites of potential TF binding. These findings contribute to our understanding of *C. elegans* transcriptional regulation at a genome-wide level by providing a resource that maps the sites of action by TFs and *cis*-regulatory modules (CRMs) in the embryo and L1 arrest and in genes that are regulated during these stages of development.

Even with recent work in this field, there is still room for improvement for TF discovery algorithms in DNase-seq data. One issue is that with current algorithms, lowering the statistical threshold for the identification of TF footprints in DNase-seq data does not eventually lead to all TF footprints from ChIP-seq data being found (Sung et al. 2014). That said, not all TF sites found in ChIP-seq are functional or drive enhancer activity. It thus remains to be seen what the differences are between TF sites of a given factor that are able to be found by DNase-seq and ChIP-seq. It is possible for example, that the capture of TF binding sites by DNase-seq and ChIP-seq reflects different kinetics of TF binding. The data from Sung et al. (2014) at least, seem to point in that direction, since within DNase-seq the depth of TF footprinting is correlated with the residence time of TFs to DNA. We also do not know exactly how DNase-seq and ATAC-seq compare in this respect.

We also face some challenges in understanding the spatial representation of DNase-seq data that arises from performing the assays in a multicellular organism. There are varying abundances of tissue and cell types in the *C. elegans* embryo or L1 larvae depending on developmental stage. One of the major issues that we face is attribution of DHS that we discover to their tissues of origin. We can, however, attempt to address some of these issues by considering that the cell lineage of *C. elegans* is invariant (Sulston et al. 1983).

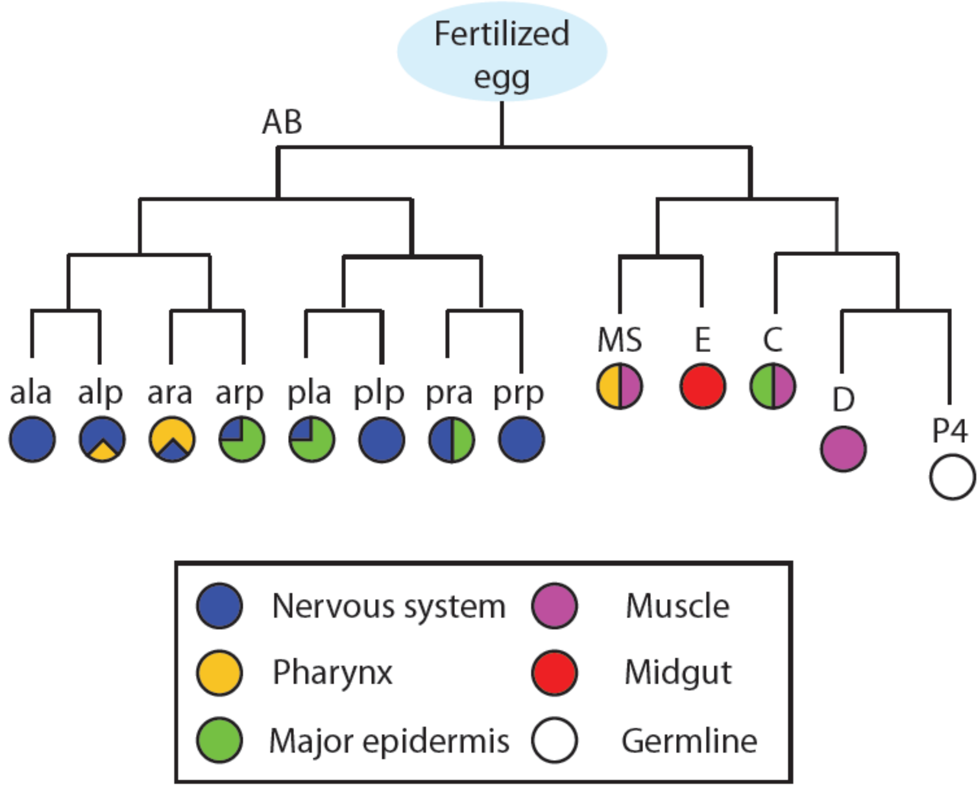


Figure 5.1 Origin of major tissues in the cell lineage of the *C. elegans* embryo
Figure adapted from Labouesse and Mango, 1999.

Different *C. elegans* tissues arise from different founder cell lineages that are created as a result of asymmetric cell divisions in the early embryo (Figure 5.2; reviewed in Labouesse and Mango, 1999 and Maduro 2010). Neuronal tissue arises from cells within the AB lineage, muscle tissue arises from the MS, C and D cell lineages, pharyngeal tissue arises from AB and MS cell lineages, intestinal (also known as mid-gut) tissue arises from the E cell lineage and epidermal tissue arises largely from the AB and C cell lineages (Figure 5.1). The germline is descended from the P_4 cell.

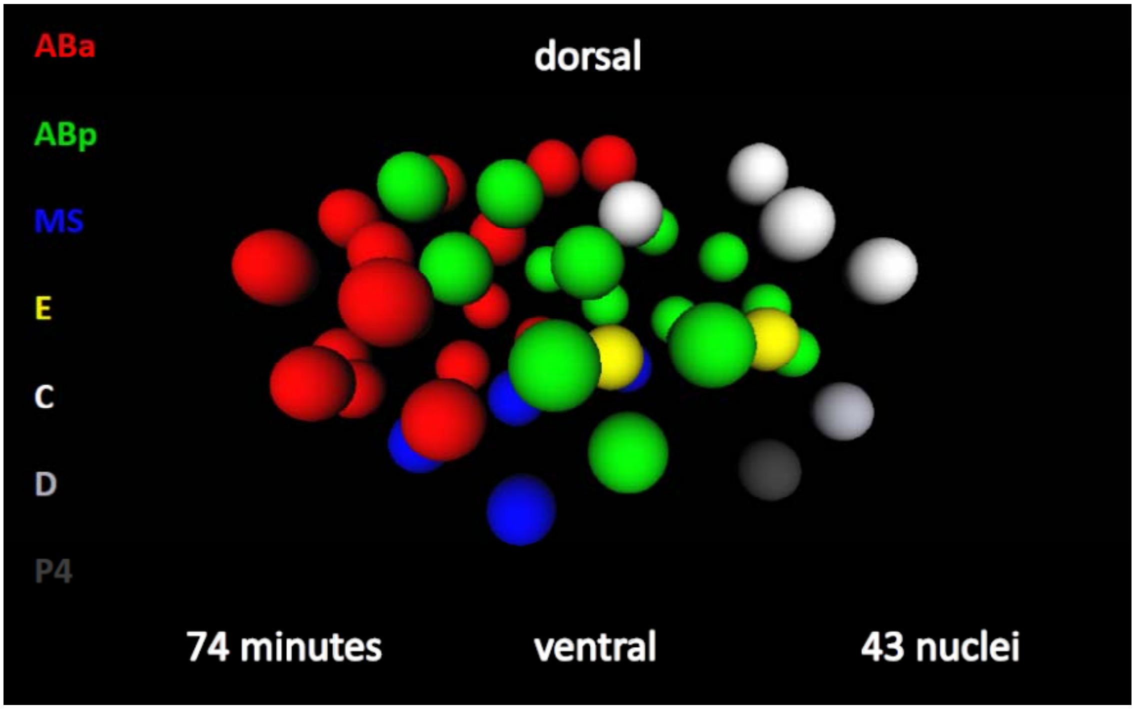


Figure 5.2 Cells in the *C. elegans* embryo at the 43-cell stage
Cell diagram was generated using WormGuides[6]. Cells are color-coded according to the major lineage group, with ABa in red, ABp in green, MS in blue, E in yellow, C in white, D in light gray, and P_4 in dark grey. Embryo is shown with dorsal side up and ventral side down. Anterior is to the left and posterior to the right.

Embryos for our DNase-seq were collected from developmental stages that range from roughly 43-cell stage (this timepoint can also be described as the 2E stage, since there are two E or endodermal cells) and onwards. At the 43-cell stage,

[6] http://www.wormguides.org/

the *C. elegans* embryo has initiated gastrulation and contains four cells from the MS lineage, four cells from the C lineage, sixteen cells from the ABp lineage, sixteen cells from the ABa lineage, two cells from the E lineage, one cell from the D lineage, and one cell of the P_4 lineage (Figure 5.2; WormGuides; Sulston et al. 1983). By 350 minutes (shortly before bean stage), cell divisions are complete and this later stage *C. elegans* embryo has close to 560 cells (Figure 5.3; Sulston et al. 1983). The L1 larva has 558 cells upon hatching, of which most (389) cells come from the AB lineage (reviewed in Riddle et al. 1997). The large number of cells present in the L1 may partially explain our lower numbers (around 16,000) of noncoding DHS detected in L1 arrest larvae compared to embryos (around 26,000), since there is more cell heterogeneity and DNase-seq signal coming from any particular cell is likely to be more diluted.

Lineage-specific expression data has been generated for many important embryo differentiation genes by Murray et al. (2012), who used cell lineage tracing methods with fluorescent reporter genes to quantitatively measure expression in developing *C. elegans* embryos through the 350-cell stage. Many of these genes could be useful markers in our DNase-seq data to give a sense of how sensitive our method is, if we are able to detect DHS near these actively transcribed genes from different *C. elegans* lineages (Table 5.1). A few of these genes have known CRMs, such as *elt-2* and *hlh-1* (which we have discussed in Chapter 2), but others have not had their regulatory regions dissected.

Gene	Lineages in which expressed
elt-2	E
elt-7	E
end-1	E
end-3	E
ges-1	E
hlh-1	C and MS
lin-1	Mostly ABp
lin-32	Mostly ABa
nhr-69	C and E
pal-1	C and D
pgp-2	E
ref-2	MS and ABa
pha-4	E, select MS and ABa
tbx-35	MS
tbx-37	ABa
tbx-38	ABa
vab-7	C and ABa

Table 5.1. Some *C. elegans* differentiation genes and cell lineages in which they are expressed in embryos before the 350-cell stage.
Data summarized from EPIC (Expression Patterns in *Caenorhabditis*[7]; Murray et al. 2012)

Among the genes listed in Table 5.1 which I selected on the basis of lineage-restricted expression observed in the Murray et al. 2012 study, all possess detectable noncoding DHS in upstream regions except for *end-3*, *ges-1*, *tbx-35* and *tbx-38*. Those genes exclusively expressed in the E lineage (*end-1*, *end-3*, *elt-2*, *elt-7*, *pgp-2*, and *ges-1*) are a good test case to consider, since a single tissue, the intestine, arises from the E blastomere and its TF regulatory cascade has been well-studied (reviewed

[7] http://epic.gs.washington.edu/

in McGhee et al. 2007 and also described later in Murray et al. 2012). The GATA TFs END-1 and END-3 are expressed in the E cell lineages of the early embryo to specify the endoderm (reviewed in McGhee, 2007). Studies by Zhu et al. (1997) and Baugh et al. (2005) have shown that transcription of *end-1* and *end-3* is transient, with transcripts detectable in 1E-stage but gone by the 8E-stage. However, data from the reporter gene analysis by Murray et al (2012) shows that large "promoters" from *end-1* and *end-3* are able to drive reporter gene expression at least until the 350-cell stage. The expression driven by the *end-3* promoter is also weak. In our embryo DNase-seq data, we observe a noncoding DHS in upstream region of *end-1,* but not *end-3*. This suggests that at least this case for the E lineage we are able to detect activation of *end-1* which is highly expressed, but not *end-3* which is more weakly expressed.

Transcription of the GATA factor ELT-2 is activated by END-1 and END-3, is expressed from the 2E-stage (reviewed in McGhee et al. 2007). ELT-2 is the predominant factor expressed in the intestine after early endoderm specification. The study by Murray et al. (2012) showed the upstream "promoter" of *elt-2* strongly drives reporter expression in the E lineage at least until the 350-cell stage. As was discussed in detail in Chapter 2, we detect many of the noncoding DHS of *elt-2* which overlap ELT-2 binding sites that mediate *elt-2* autoregulation. Another GATA factor which is partially redundant with ELT-2 to specify the intestine is ELT-7 (reviewed in McGhee et al. 2007). The study by Murray et al. (2012) showed that the *elt-7* upstream "promoter" drove strong reporter expression in E lineage at least until

350-cell stage, and we detect noncoding DHS near this gene in our embryo DNase-seq data. Around this time, an ABC transporter *pgp-2* is also expressed in the E lineage starting from the 2E stage (Murray et al. 2012; Schroeder et al. 2007) and we are able to detect noncoding DHS upstream of the gene in embryo DNase-seq.

The intestinal differentiation gene, *ges-1*, is activated by ELT-2 and expressed in late embryogenesis, at around 250 minutes. Based on our embryo DNase-seq data, which does not show DHS for *ges-1*, it is possible that our developmental time window of embryo collection could be too early to detect any *ges-1* DHS. Another possibility is that if our sampling did include embryos collected at this stage of development (which would be around the 200-cell stage), in terms of the number of cells, any signal from the E lineage would be diluted by the more proliferated AB lineages. Thus, using these genes as marker genes for developmental timing and assuming that the DHS do indeed reflect CRM activity at this time, we can conclude our DNase-seq signal and sample is sufficient to detect early endoderm and intestine genes in E lineages in the early stages of embryonic development, but the signal from promoter of later stage gene, *ges-1*, are not present 1) due to errors in detection 2) the embryos collected do not include this later stage of development, or 3) if there are some late-stage embryos included in the collection, the DNase-seq signal originating from in the E lineage is diluted because of the large number of AB lineage cells dominating these later stage embryos.

We can also consider another tissue, such as the pharynx, which arises from AB and MS lineages. Specification of the pharynx tissues (including many muscle cells) is dependent on PHA-4/FOXA and T-box transcription factors TBX-2, TBX-35, TBX-37, and TBX-38 (reviewed in Mango et al. 2007). Expression of the redundant pair of TFs TBX-37 and TBX-38 is initiated in the ABa lineage at the 24-cell stage (Good et al. 2004). TBX-35 is expressed in the MS lineage (Murray et al. 2012). I observe noncoding DHS upstream of *tbx-37* in the embryo DNase-seq data but not near *tbx-38* and *tbx-35*. One explanation could that *tbx-37* is more highly expressed and in more total cells in the embryo; *tbx-37* is both highly expressed in all of the ABa lineages, whereas *tbx-38* expression is more restricted within the ABa lineage, mostly descendents of ABala, and *tbx-35* is moderately expressed in the MS lineage which does not contain as many cells as the ABa lineage. That said, we were able to detect highly expressed intestinal genes in E lineage cells in the previous case, so the number of cells is probably not the only limiting factor -- it is possible that lower expression of the gene may also impact DHS detection. The organ selector gene PHA-4, which is required for pharynx development, is expressed beginning in the 4E-stage (50-100 cells; Horner et al. 1998). Its expression was detected in the E cell lineage and selected cells in the ABa and MS lineages by Murray et al. (2012). As was described in Chapter 2, we are able to detect several noncoding DHS for *pha-4* including sites of autoregulation.

The remaining genes that I investigated from Table 5.1 all possessed upstream noncoding DHS in the embryo DNase-seq data: *hlh-1* (bHLH TF specifying body wall muscle in the C and MS lineages), *lin-1* (an Ets TF expressed in mostly ABp cell lineages), *lin-32* (bHLH TF expressed in mostly ABa cell lineages), *nhr-69*

(NR2 family receptor expressed in C and E lineages), *pal-1* (homeodomain TF expressed in C and D lineages and important for body wall muscle development), *ref-2* (expressed in neural and hypodermal precursors in the MS and ABa lineages), and *vab-7* (Homeodomain TF expressed in posterior tissues in cells of ABa and C lineages). I also looked at some genes in Murray et al. (2012) dataset with many fewer cells: *mnm-2* (TF expressed in select descendants of ABa and ABp), *nhr-67* (ortholog of *Drosophila* and mammalian *tailless* that is expressed in select descendants of MS and ABp), and *ttx-3* (LIM homeodomain TF expressed in select descendants of ABa and ABp) and these contained many noncoding DHS upstream or in the introns of genes, suggesting that we are able to detect putative CRMs near these lineage-restricted genes and that our embryonic timepoints.

Thus far, our DNase-seq analyses have largely focused on CRMs that promote gene transcription such as enhancers and promoters, and our results have shown positive correlations between the number of DHS and gene expression levels. One explanation for this is the nature of eukaryotic gene regulation, which relies on a complex chromatin structure that acts as an intrinsic barrier to transcription. Eukaryotic gene regulation features a transcriptionally restrictive ground state, requiring context-specific activators to direct transcription (reviewed in Struhl 1999), which is a fundamentally different gene regulatory logic than prokaryotic gene expression. Thus eukaryotic genomes might have a bias towards CRMs that function to activate transcription of genes. Of course, CRMs function by binding both activator and repressor TFs to drive expression in specific spatiotemporal patterns.

Transcriptional repression is thus just as important as activation to properly control the expression of genes (reviewed in Payankaulam et al. 2010).

However, I think a subset of the noncoding DHS that we find may in fact harbor negative regulatory activity and could potentially act as silencers. Previous studies have shown that silencers are identifiable by DNaseI hypersensitivity assays, such as silencers of mouse interleukin 4 and CD4 genes (Siu et al. 1994; Ansel et al. 2004) and a constitutive autonomous silencer element recently found in human erythroid K562 cells (Qi et al. 2015). There are a few cases in our data that suggest that negative regulation may be occurring. In the case of *hlh-1*, a noncoding DHS was detected in the first intron harbored binding sites for PHA-4/FOXA (see Chapter 2 for details). We suspect that this may be a negative regulatory CRM or potential silencer of *hlh-1* expression in the pharynx. PHA-4 is known to be able to work as activator or repressor in different gene loci. For example, while PHA-4 activates many genes to promote pharyngeal differentiation, it also acts to repress ectodermal cell fate in the pharynx (Kiefer et al. 2007). We also found embryo noncoding DHS overlapping homeodomain binding sites for MAB-18 and CEH-14 which prevent activation of dauer collagen *col-43* by the promoter of *sth-1* that drive expression in the spermatheca (see Chapter 2 for details). Unfortunately, the numbers of identified silencers is lower than that for enhancers, despite the fact that there is precedent for some sequences to act as either an enhancer or silencer depending on the context. One such example is the neuron-restrictive silencer element (NRSE), which can act as an enhancer in neuronal cells but as a silencer in non-neuronal cells (Bessis et al.

1997). With improved methods and approaches in the future, perhaps we will able to gain a better understanding of these negatively regulating or silencer elements. One could imagine, for example, a massively parallel reporter assay designed to test some of the DHS that we have found as silencer elements.

But first, it is important to find what percentage of CRMs predicted by DNase-seq data are able to act as enhancers and drive transgene reporter activity. I believe that our experiments to perform massively parallel testing of ten thousand *C. elegans* enhancers will help answer that question. This proof-of-principle of DNase-seq in nematodes also opens the door to asking similar questions of more distantly related nematodes with highly varied lifestyles and developmental biology. I think that the data from *Steinernema carpocapsae* will be valuable in beginning to address questions of conservation of *cis*-regulatory sequences in *Steinernema*.

Other challenges remain, such as the need to identify cell-type specific enhancers from specific cell or tissue populations and robust and high throughput ways to directly detect the identity of specific TFs that bind enhancers and other *cis*-regulatory sequences. The identification of many thousands of CRMs and TF binding sites acting at different developmental stages of *C. elegans* is only one step on the path towards trying to understand the detailed mechanisms of *cis*-regulation in nematodes.

References

Ansel KM, Greenwald RJ, Agarwal S, Bassing CH, Monticelli S, Interlandi J, Djuretic IM, Lee DU, Sharpe AH, Alt FW, Rao A. 2004. Deletion of a conserved Il4 silencer impairs T helper type 1-mediated immunity. Nat Immunol. 5(12):1251-9.

Baugh LR, Hill AA, Claggett JM, Hill-Harfe K, Wen JC, Slonim DK, Brown EL, Hunter, CP. 2005. The homeodomain protein PAL-1 specifies a lineage-specific regulatory network in the *C. elegans* embryo. Development *132*, 1843–1854

Bessis A, Champtiaux N, Chatelin L, Changeux JP. 1997. The neuron-restrictive silencer element: a dual enhancer/silencer crucial for patterned expression of a nicotinic receptor gene in the brain. Proc Natl Acad Sci U S A. 94(11):5906-11.

Good K, Ciosk R, Nance J, Neves A, Hill RJ, and Priess JR. 2004. The T-box transcription factors TBX-37 and TBX-38 link GLP-1/Notch signaling to mesoderm induction in C. *elegans* embryos. Development 131, 1967–1978.

Horner MA, Quintin S, Domeier ME, Kimble J, Labouesse M, Mango SE. 1998. pha-4, an HNF-3 homolog, specifies pharyngeal organ identity in *Caenorhabditis elegans*. Genes Dev. 12(13):1947-52.

Kiefer JC, Smith PA, Mango SE. 2007. PHA-4/FoxA cooperates with TAM-1/TRIM to regulate cell fate restriction in the C. *elegans* foregut. Dev Biol. 303(2):611-24.

Labouesse M, Mango SE. 1999. Patterning the C. *elegans* embryo: moving beyond the cell lineage. Trends Genet. 15(8):307-13.

Maduro MF. 2010. Cell fate specification in the C. *elegans* embryo. Dev Dyn. 239(5):1315-29.

Mango SE. 2007. The C. *elegans* pharynx: a model for organogenesis. WormBook. 1-26.

McGhee JD. 2007. The C. *elegans* intestine. WormBook. 1-36.

Murray JI, Boyle TJ, Preston E, Vafeados D, Mericle B, Weisdepp P, Zhao Z, Bao Z, Boeck M, Waterston RH. 2012. Multidimensional regulation of gene expression in the C. *elegans* embryo. Genome Res. 22(7):1282-94.

Payankaulam S, Li LM, Arnosti DN. 2010. Transcriptional repression: conserved and evolved features. Curr Biol. 20(17):R764-71.

Riddle DL, Blumenthal T, Meyer BJ, et al., editors. 1997. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press.

Schroeder LK, Kremer S, Kramer MJ, Currie E, Kwan E, Watts JL, Lawrenson AL, Hermann GJ. 2007. Function of the *Caenorhabditis elegans* ABC transporter PGP-2 in the biogenesis of a lysosome-related fat storage organelle. Mol Biol Cell. 18(3):995-1008.

Siu G, Wurster AL, Duncan DD, Soliman TM, Hedrick SM. 1994. A transcriptional silencer controls the developmental expression of the CD4 gene. EMBO J. 13(15):3570-9.

Struhl K. 1999. Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. Cell. 98:1–4

Sulston JE, Schierenberg E, White JG, Thomson JN. 1983. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. Dev Biol 100:64-119.

Qi H, Liu M, Emery DW, Stamatoyannopoulos G. 2015. Functional validation of a constitutive autonomous silencer element. PLoS One. 10(4):e0124588

Zhu J, Hill RJ, Heid PJ, Fukuyama, M Sugimoto A, Priess, JR, Rothman JH. 1997. *end-1* encodes an apparent GATA factor that specifies the endoderm precursor in *Caenorhabditis elegans* embryos. Genes Dev. 11, 2883–2896.

Appendix I

DNaseI-seq protocol for nematodes

Protocol is adapted from Stam Lab (Thurman et al. 2012) for DNase-seq and INTACT protocol for *C. elegans* nuclei isolation (Florian Steiner)

Make Stam lab Buffer A and Tris NPB the week of experiment
Make 1x DNase digestion buffer on the day of experiment

Stock Reagents:

Unless otherwise noted, all buffers & stock solutions should be chilled to 4°C (on ice) prior to use.

0.5M Spermine

Dissolve 5 grams Spermine Free Base in 49.43mL final volume sterile dH20.
Store in convenient aliquots at -20°C.

0.5M Spermidine

Dissolve 1 gram Spermidine Free Base in 13.77mL final volume sterile dH20.
Store at 4°C.

DNaseI 10X Digestion Buffer (per 50mL)

Final concentration	*Stock concentration*	*Amount used from stock*
60mM CaCl2	1M CaCl2	3mL
750mM NaCl	5M NaCl	7.5mL

Combine stock solutions and 39.5mL sterile dH20.
Can be stored at room temperature up to 1 year.

Stock DNaseI

Solubilize on ice **with no vortexing** entire bottle of DNaseI Type II in following storage buffer at a final concentration of 10U/µL:
20mM Tris-HCl, pH 7.6
50mM NaCl
2mM MgCl2
2mM CaCl2
1mM Dithioerythritol
0.1 mg/mL Pefabloc SC
50% Glycerol
Store in 250 µL aliquots at -20°C.

Stam Lab Buffer A (per Liter)

Final Concentration	*Stock concentration*	*Amount used from stock*
Sterile MilliQ Water		918mL
15mM Tris-HCl, pH 8.0	1M Tris-HCl, pH 8.0	15mL
15mM NaCl	5M NaCl	3mL
60mM KCl	1M KCl	60mL
1mM EDTA, pH 8.0	0.5M EDTA, pH 8.0	2mL
0.5mM EGTA, pH 8.0	0.5M EGTA, pH 8.0	1mL
0.5mM Spermidine	0.5M Spermidine Free Base	1mL

Combine indicated amounts of stock solutions and sterile dH2O to final volume of 1 L. Store at 4°C. Use within 1 week.

1X DNaseI Digestion Buffer
Make day of use.
For 50mL: add 5mL 10X DNaseI Digestion Buffer to 45mL Buffer A.
Allow to equilibrate to 37°C for 60 minutes prior to use.

Stop Buffer (per Liter)

Final concentration	*Stock concentration*	*Amount used from stock*
50mM Tris-HCl, pH 8.0	1.0M Tris-HCl, pH 8.0	50mL
100mM NaCl	5.0M NaCl	20mL
0.10% SDS	10% SDS	10mL
100mM EDTA, pH 8.0	0.5M EDTA, pH 8.0	200mL
Molecular Biology Grade sterile H2O		720mL

Combine stock solutions and add sterile dH2O to a final volume of 1 L. Dispense into 25mL aliquots and store at 4°C. (SDS will precipitate upon storage at 4°C but will go back into solution upon warming to 37°C).

On day of use, add the following to a 25mL aliquot:
50 µL 0.5M Spermidine Free Base (final concentration: 1mM)
15 µL 0.5M Spermine Free Base (final concentration: 0.3mM)

NPB (Nuclei Purification Buffer):
10mM Tris pH7.5
40mM NaCl
90mM KCl
2mM EDTA
0.5mM EGTA
0.5mM Spermidine – add right before using
0.2mM Spermine – easily oxidized, add right before using
0.2mM DTT – easily oxidized, add right before using
0.1% Triton X-100
Store 4C, use within one week

Nuclei Isolation:

1. Grind worm pellets to fine powder under liquid nitrogen using liquid nitrogen cooled mortar and pestle
2. Bring volume to 7 mL with NPB
3. Pre-cool centrifuge to 4°C. All centrifugations should be done at 4°C.
4. Transfer to Dounce homogenizer with pipet
5. Homogenize with Dounce homogenizer 30 strokes with tight fitting pestle
6. Spin at 0.1 x *g* to pellet debris
7. Collect nuclei containing supernatant and pool in new 50 mL tube on ice.
8. For each 3 mL of supernatant, prepare 3 mL Optiprep (Sigma) cushion at bottom of 15mL tubes. Apply supernatant on top.
9. Spin nuclei down on cushion at 1000 x *g*
10. Collect nuclei in a 15mL conical tube, these are input nuclei
11. Proceed immediate to DNaseI treatment.
    Before DNaseI treatment, stain with DAPI and visualize using 100X lens on DIC. Use DAPI filter cartridge. Start with 20X magnification, using visual spectrum light, focus. Focus and close condenser to fine point on debris, then switch to higher magnification 100X using oil and open UV light source.

DNaseI Treatment of Nuclei

Work quickly using reagents maintained at appropriate temperatures.

1. Pre-cool centrifuge to 4°C. All centrifugations should be done at 4°C.
2. Add protease inhibitor tablet to Stam Lab Buffer A (1 tablet per 50mL solution) and solubilize. Keep on ice.
3. Prepare fresh 1X DNaseI Digestion Buffer (Dilute 10X DNaseI Digestion Buffer 1:9 with Stam Lab Buffer A).
4. Warm Stop Buffer and 1X DNaseI Digestion Buffer (minus DNaseI) in 37°C temperature bath. Allow solutions to equilibrate for 60 minutes prior to use.
5. Aliquot into equal volume tubes for DNaseI treatment.
6. Centrifuge for 5 minutes at 500 x *g* at 4°C. Remove supernatant from all nuclei pellets.
7. Add spermine free base and spermidine free base to Stop Buffer. (If SDS has precipitated out of solution, warm to 37°C to resuspend SDS **prior** to adding supplements).
8. Aliquot 1X DNaseI Digestion Buffer: In 15mL conical tubes, 1-5mL 1X DNaseI Digestion Buffer (1mL per 10 million expected nuclei); number of tubes is determined by number of DNaseI treatments to be done.
9. Just prior to starting DNaseI reaction with the nuclei pellet, add 5 µL **Proteinase K** per mL Stop Buffer.
10. Also just prior to starting DNaseI I reaction with the nuclei pellet, **add the appropriate amount of DNaseI enzyme to the 1X DNaseI Digestion Buffer aliquot.** Mix thoroughly but gently by pipeting (**DO NOT VORTEX**) as the enzyme denatures easily with aeration.
    For 10 U/mL digestion, add 4 µL of 10U/µL stock DNaseI to 4mL of 1X DNaseI

Digestion Buffer.
For 20 U/mL digestion, add 8 µL of 10U/µL stock DNaseI to 4mL of 1X DNaseI Digestion Buffer
For 40 U/mL digestion, add 16µL of 10U/µL stock DNaseI to 4mL of 1X DNaseI Digestion Buffer

Remaining steps should be timed carefully:

1. Gently tap nuclei pellets a few times on the side of the ice bucket to loosen. Place tubes with loose nuclei pellets in 37°C temperature bath and allow the temperature to equilibrate for 1 minute.
2. Gently resuspend nuclei with 1X DNaseI Digestion Buffer plus enzyme.
3. Pipet several times gently using wide-bore tips to ensure homogenous suspension.
4. Incubate for 3 minutes at 37°C in temperature bath.
5. Add equal volume of Stop Buffer to DNaseI reaction tube and mix by inverting tube several times.
6. Digest sample overnight in the 55°C temperature bath.
7. Store treated samples at 4°C.
8. Prior to gel electrophoresis and QPCR, incubate the samples at 37°C for 30 minutes with 1.5 µL 30 mg/mL RNaseA per mL of DNase-seq sample.
9. Proceed to DNA purification, gel extraction, Qubit and PCR.

Appendix II

Supplementary Information for Chapter 2

Appendix Table 2.1. Sequenced DNase-seq samples.

A. Sample yield and regulatory enrichment by QPCR. Four biological replicates of embryo (A-D) and five of L1 arrest (V-Z) DNase-seq were performed. The DNA yield of each sample was measured using Qubit fluorescence. The DNaseI treatment level that exhibited the highest fold QPCR regulatory enrichment (comparing *lin-39/ceh-13* Hox conserved enhancer regions vs. non-enhancer sequences from Kuntz et al. 2008; see Methods) was sequenced. **B. Read mapping to *C. elegans* genome with Bowtie 1.0.0.** Reads were mapped to the ce10/WS220 genome and alignment statistics reported by Bowtie are shown for each biological replicate: Number of 1) Reads processed by Bowtie after Q20 filtering and trimming (Reads Processed) 2) Reads with at least one reported alignment 3) Reads that failed to align 4) Reads with alignments suppressed due to multi-mapping to more than two unique genomic locations. Percentages are shown in parentheses. Uniquely mapping reads ranged between 38% and 76% in these samples result in slightly above 15X coverage in each sample. Out of four embryo samples, replicates A-C showed more ideal alignment statistics, reflecting DNA yield of biological replicates in (A).

Appendix Table 2.1. Sequenced DNA samples

A. Sample yield and regulatory enrichment by QPCR

Replicate ID	Flowcell Sample ID	Strain & Stage	DNaseI treatment level w/ most enrichment	Highest Fold enrichment vs. bkgrd (N5, N6)	DNA yield
A	14140	N2 Embryo	160 U/mL	6.4	19 ng
B	13583	N2 Embryo	80 U/mL	6.3	50 ng
C	13578	N2 Embryo	120 U/mL	3.9	39 ng
D	13577	N2 Embryo	160 U/mL	5.3	3 ng
Z	13576	N2 L1 arrest	80 U/mL	4.7	336 ng
Y	13579	N2 L1 arrest	20 U/mL	5.8	8ng
X	13582	N2 L1 arrest	160 U/mL	5.5	17ng
W	14138	N2 L1 arrest	80 U/mL	1.4	27ng
V	14139	N2 L1 arrest	160 U/mL	5.7	25ng

B. Read mapping to C. elegans genome (ce10/WS220) with Bowtie 1.0.0

Replicate ID	Flowcell Sample ID	Strain & Stage	Reads Processed (Q20 filter+trim)	Reads with at least one reported alignment	Reads that failed to align	Reads w/ alignments suppressed due to multimapping
A	14140	N2 Embryo	39,673,047	28040916 (71%)	6059497 (15%)	5572634 (14.%)
B	13583	N2 Embryo	21,165,105	16086392 (76%)	657736 (3.1%)	4420977 (21%)
C	13578	N2 Embryo	38,482,313	24084424 (63%)	1021020 (2.7%)	13376869 (35%)
D	13577	N2 Embryo	18,523,832	7096637 (38%)	10334476 (56%)	1092719 (5.9%)
Z	13576	N2 L1 arrest	42,554,211	24045456 (57%)	863912 (2.0%)	17644843 (41%)
Y	13579	N2 L1 arrest	16,074,836	10010938 (62%)	1491568 (9.3%)	4572330 (28%)
X	13582	N2 L1 arrest	11,397,805	7679786 (67%)	699720 (6.1%)	3018299 (26%)
W	14138	N2 L1 arrest	30,376,192	20166440 (66%)	3950146 (13%)	6259606 (21%)
V	14139	N2 L1 arrest	32,487,179	15981175 (49%)	12066750 (37%)	4439254 (14%)

Appendix Table 2.2. QPCR validation

QPCR primers were designed to amplify MUSSA conserved regions from “true positive” enhancers of the *lin-39/ceh-13* Hox genes (Kuntz et al. 2008), conserved regions from the enh2 and enh4 enhancers of *hlh-1,* and intergenic and promoter regions of *unc-54*, *ceh-22*, *let-70*, and *cct-8* genes. QPCR primers were also designed to amplify subregions of negative control non-enhancer sequences, N5 and N6, previously described by Kuntz et al. (2008).

Appendix Table 2.2. QPCR primers and amplicons used to measure regulatory enrichment

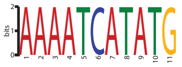
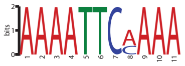
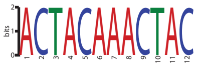
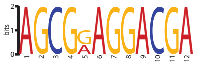
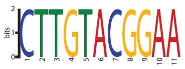
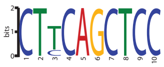
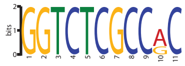
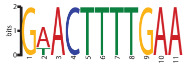
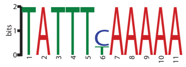
QPCR Region Label	Forward Primer	Reverse Primer	Amplicon Coordinates (ce10/WS220)	Corresponds to Regulatory Region
SK_N1_2	CAAAGTGCACAATGCTGTCC	CCGCAGCGGTATCTCTCTTA	chrIII:7,531,492-7,531,564	N1
SK_N2_1	TTGGGCTTGAAGTGGTTAGG	GTCGCGAGCCCATTTATCT	chrIII:7,532,042-7,532,129	N2
SK-N2_3	TCGCCTTCTTCCTTATGCTTC	AGGAAGCTACAGTACTCCCCTTCT	chrIII:7,532,219-7,532,291	N2
SK-N3_1	GAGACAAACAGCGGGAACAA	CGCAGTGAGGGAAAATGAAA	chrIII:7,533,122-7,533,211	N3
SK-N4_2	GATGGACATGGGGTGAGAAC	CGGCAACTTAAAAGCGAAAA	chrIII:7536786-7536879	N4
SK-N5_1	CCTTAACGCGACCAAGGTTA	ACTCCAAAATTGGCCCAAAA	chrIII:7,538,661-7,538,735	N5 (negative ctrl)
SK-N5_3	GGTCTTCCAATCTAGTGCAAACA	TCCCTCTTTTTCTCGTCATTTG	chrIII:7,538,116-7,538,200	N5 (negative ctrl)
SK-N6-1	ACGCCTTTCGAGAAGTCTATTGT	AATTTGTTGCAGGCCACATC	chrIII:7,543,275-7,543,364	N6 (negative ctrl)
SK-N7-1	AATGGCACCCATAAATCTCAAC	TCTCATCCTCTTCCTCTCTCCA	chrIII:7,544,309-7,544,395	N7
SK-N8-2	TGCCAAGGATCTAGAGGGTGT	CAATCCGACAACACCAATCA	chrIII:7,545,257-7,545,329	N8
SK-N9_1	TACAAGCCCACGACCATTCT	CCACAGAGAGACATGGGAACA	chrIII:7,548,980-7,549,053	N9
SK-N9_2	CGGTGCATTTTGGAAGAAGT	TCGGAACAGTTGGTAAGTTGC	chrIII:7,549,080-7,549,153	N9
SK_N11-1	CTCCTTCTTTTCCCCGTGTC	GAGAGAGACACCATCCGATCA	chrIII:7,554,774-7,554,850	N11
unc-54	TAAAGCTGTGTGCGGCAGCGGCA	ACTACGCGTAGGCGTCTCTCGC	chrI:14,863,598-14,863,685	unc-54 upstream
let-70	AAAATGAGCGACGGGGTGAG	GTACCCTCTTACGTTTCCTGTGTT	chrIV:11,082,900-11,082,975	let-70/klc-1
cct-8	GAGATGTGGGGTACGGTGGA	ATGACACCGAACTTGACGCG	chrIV:1,094,354-1,094,416	cct-8 upstream
ceh-22	CGGTTGTCAATTGCACTCGAG	GATAGAAGGCGTCGCTGCTG	chrV:10,672,580-10,672,654	ceh-22 distal promoter
hlh-1_enh2	AAGGTGTCGGTTGTAGCAGC	AGAGTTGAGCCGAGAGTTGC	chrII:4,517,444-4,517,507	hlh-1 enhancer 2
hlh-1_enh4	GCCTCCATCAACGTCTTAACGGC	CTCTCTTGCTTCCCGAGAAGCTACC	chrII:4,520,326-4,520,394	hlh-1 enhancer 4

Appendix Table 2.3. Predicted novel regulatory motifs

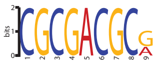
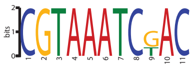
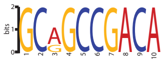
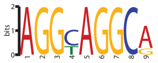
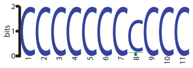
Novel regulatory motifs (shown in IUPAC and logos) were predicted as well as the Gene Ontology and anatomy enrichment of motif-associated genes. Motifs were predicted by DREME in different categories of noncoding DHS (left border). *P*-values (p-val) and erased E-value (E-val) are shown. In many cases, motifs matched a previously identified Stormo or Elemento motif (Ihuegbu et al. 2012; Elemento et al. 2005; Prior Match?). Motif-associated genes were selected by FIMO using a *P*-value cutoff (Threshold) to identify the presence of motifs within noncoding DHS. Number of motif-associated genes (#Genes) used in the analysis of GO enrichment using AmiGO is shown. Top enriched GO terms are shown (Gene Ontology Enrichment) and related GO terms were highlighted (blue background). If present, enriched anatomy terms (Anatomy Enrichment) are also shown.

Appendix Table 2.3. Predicted functions for novel regulatory motifs

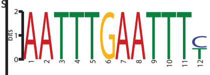
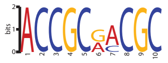
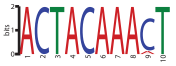
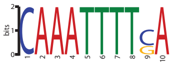
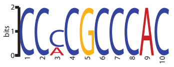
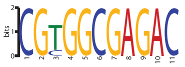
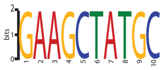
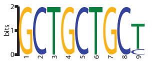
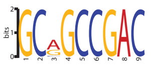
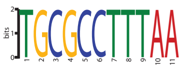
A Novel Intergenic motifs

IUPAC Motif	p-val	E-val	Prior Match?	Threshold	#	Gene Ontology Enrichment	Anatomy Enrichment	Motif Logo
AAAATCATATG	1.50E-08	0.029	None	0.05	19	peptidyl amino acid modification actin filament based movement G protein-coupled acetylcholine receptor pathway	—	
AAAATTCMAAA	1.10E-08	0.022	Stormo F01D5.10.8	0.05	77	response to chemical stimulus cellular metabolism	head neurons	
ACTACAAACTAC	3.70E-09	0.0073	Stormo C39D10.7.2	0.025	24	response to chemical stimulus steroid hormone mediated signaling response to Ca2+ ion organic substance metabolism dauer entry	seam cell spermatheca excretory cell	
AGCGRAGGACGA	2.30E-10	0.00045	None	0.025	39	regulation of cellular metabolism phosphorus metabolism phosphate metabolism purine ribonucleotide catabolism anatomical structure development	tail neurons head neurons coelomocyte	
CTTGTACGGAA	1.50E-08	0.029	None	0.05	18	ion membrane transport actin myosin filament sliding	—	
CTYCAGCTCC	2.50E-09	0.0049	None	0.05	27	establishment of localization ion transport vesicle mediated transport transmembrane transport signaling synaptic transmission cell-cell signaling cell communication	—	
GGTCTCGCCRC	6.30E-18	1.30E-11	None	0.025	63	nucleobase-containing compound metabolism cellular macromolecule biosynthesis heterocyclic metabolism	pharyngeal muscle	
GWACTTTTGAA	7.20E-06	7.20E-06	None	0.05	8	phosphorylation protein phosphorylation phosphate containing compound metabolism phosphorus metabolism peptidyl threonine phosphorylation cell fate commitment	—	
TATTTYAAAAA	4.00E-04	8.70E-02	None	0.05	12	signal transduction cell response to stimulus signaling cell communication activation of RasGTPase activity	—	

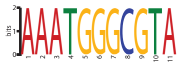
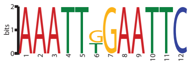
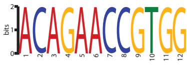
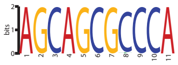
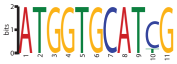
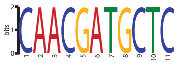
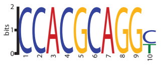
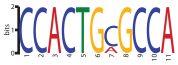
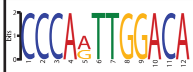
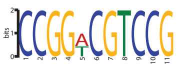
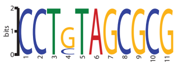
B Novel Promoter motifs

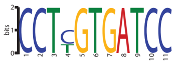
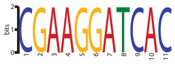
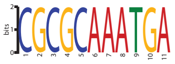
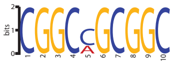
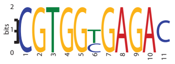
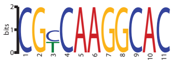
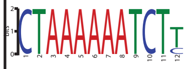
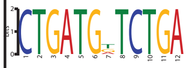
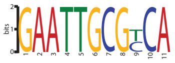
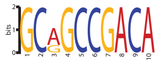
IUPAC Motif	p-val	E-val	Prior Match?	Threshold	#	Gene Ontology Enrichment	Anatomy Enrichment	Motif Logo
CGCGACGCR	7.50E-13	9.10E-07	None	0.05	50	reproduction organ development reproductive structure development developmental process involved in reproduction aging nucleobase-containing compound metabolism	head tail spermatheca	
CGTAAATCKAC	3.60E-09	0.0043	None	0.05	34	cellular metabolism translation cellular protein metabolism cellular macromolecule biosynthesis protein metabolism cellular macromolecule metabolism gene expression embryo development cellular localization transport	pharynx tail vulva	
GCRGCCGACA	1.10E-10	0.00013	None	0.05	33	transport establishment of localization macromolecule localization maintenance of location lipid localization localization lipid storage regulation of biological quality vesicle mediated transport membrane organization	intestine vulval muscle BWM anal dep. muscle	
AGGYAGGCR	5.10E-24	6.50E-18	None	0.05	67	establishment of localization developmental process	nerve ring	
CCCCCCCYCCC	3.90E-16	4.90E-10	Stormo 6R55.1a.3	0.025	129	cellular macromolecule metabolism small molecule metabolism organic substance metabolism nitrogen compound metabolism phosphate-containing compound metabolism establishment of localization transport	tail neurons	

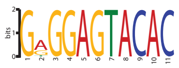
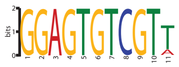
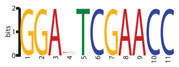
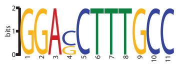
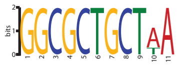
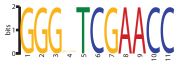
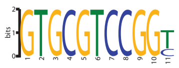
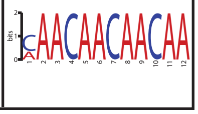
C Novel Intron motifs

IUPAC Motif	p-val	E-val	Prior Match?	Threshold	#	Gene Ontology Enrichment	Anatomy Enrichment	Motif Logo
AATTTGAATTTY	1.50E-15	2.80E-09	None	0.025	53	cellular component organization or biogenesis	ventral cord neurons tail tail neurons nerve ring head neurons	
ACCGCRMCGC	1.40E-28	2.70E-22	None	0.025	71	regulation of signal transduction	—	
ACTACAAAMT	3.00E-53	6.30E-47	Stormo C39D10.7.2	0.025	123	transport establishment of localization endocytosis receptor-mediated-endocytosis vesicle mediated transport localization embryo development heterocycle metabolism	coelomocyte	
CAAATTTTSA	1.50E-08	0.027	None	0.05	70	establishment of localization transport receptor-mediated endocytosis reproduction	germline	
CCMCGCCCAC	9.50E-09	0.017	None	0.05	56	cellular macromolecule metabolism protein metabolism organic substance metabolism macromolecule metabolism cellular component organization or biogenesis	ventral cord neurons ventral nerve cord dorsal nerve cord nerve ring	
CGYGGCGAGAC	2.00E-32	4.00E-26	None	0.025	103	RNA metabolism cellular nitrogen compound metabolism nucleobase-containing compound metabolism cellular protein metabolism cellular macromolecule metabolism protein metabolism gene expression nitrogen compound metabolism carboxylic acid metabolism small molecule metabolism organic acid metabolism establishment of localization transport	tail neurons BWM anal dep. muscle head neurons	
GAAGCTATGC	3.40E-15	6.50E-09	None	0.05	31	glucose transport positive regulation of barrier septum assembly chemical homeostasis	—	
GCTGCTGCY	2.00E-19	4.00E-13	Elemento Motif 151	0.05	93	cellular response to stimulus signal transduction regulation of response to stimulus signaling cell communication response to stimulus metabolism protein metabolism organic substance metabolism catabolism regulation of metabolism	—	
GCVGCCGAC	3.70E-41	7.60E-35	None	0.05	165	response to stimulus	hypodermis vulval muscle BWM	
TGCGCCTTTAA	1.50E-08	0.027	None	0.025	25	establishment of localization RNA splicing	—	

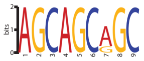
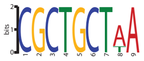
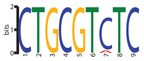
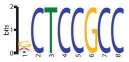
D Novel Noncoding motifs

IUPAC Motif	p-val	E-val	Prior Match?	Threshold	#	Gene Ontology Enrichment	Anatomy Enrichment	Motif Logo
AAATGGGCGTA	7.40E-09	0.024	None	0.05	29	nucleobase-containing compound metabolism cellular nitrogen compound metabolism cellular aromatic compound metabolism heterocycle metabolism reproduction	—	
AAATTKGAATTC	3.70E-11	0.00012	None	0.025	48	ribosome biogenesis cellular component biogenesis rRNA metabolism protein metabolism cellular macromolecule metabolism cellular protein metabolism protein glycosylation glycosylation organic substance metabolism ion transport transmembrane transport metal ion transport ion transmembrane transport	muscle cell coelomocyte	
ACAGAACCGTGG	4.60E-10	0.0015	F45F2.11.3 Stormo	0.025	32	cellular component organization apoptotic process cellular component organization or biogenesis aging anatomical structure development reproduction	excretory cell coelomocyte germline	
AGCAGCGYCCA	7.20E-31	2.50E-24	None	0.025	54	positive regulation of biological process regulation of biological process regulation of multicellular organismal process reproduction vesicle-mediated transport transport endocytosis RNA metabolism RNA processing nucleic acid metabolism cellular macromolecule metabolism gene expression	hypodermis	
ATGGTGCATYG	1.10E-13	3.70E-07	None	0.05	39	biological procellular metabolism phosphorus metabolism phosphate-containing compound metabolism	—	
CAACGATGCTC	4.60E-10	0.0015	Stormo F55A3.1.4	0.05	29	reproduction	—	
CCACGCAGGY	5.80E-11	0.00019	None	0.05	32	phosphate-containing compound metabolism phosphorylation dephosphorylation phosphorus metabolism cellular protein metabolism cellular macromolecule metabolism protein metabolism insulin receptor signaling pathway dauer larval development determination of adult lifespan aging dauer entry	—	
CCACTGMGCCA	3.60E-12	0.000012	None	0.025	57	cellular metabolism cell communication cellular response to stimulus	ventral cord neurons	
CCCARTTGGACA	2.20E-13	7.30E-07	None	0.025	40	cell death death reproduction cell metabolism	coelomocyte germline	
CCGGWCGTCCG	2.90E-11	0.000094	None	0.025	32	cellular amino acid metabolism regulation of actin cytoskeleton organization glutamine family amino acid metabolism	—	
CCTSTAGCGCG	9.20E-10	0.003	None	0.05	18	nematode larval development post embryonic development post-embryonic organ development cellular protein metabolism protein metabolism ncRNA processing apoptotic cell clearance mitoch. respiratory chain complex I biogenesis mitoch. respiratory chain complex I assembly NADH dehydrogenase complex assembly	—	

IUPAC Motif	p-val	E-val	Prior Match?	Threshold	#	Gene Ontology Enrichment	Anatomy Enrichment	Motif Logo
CCTYGTGATCC	1.80E-09	0.0059	None	0.05	16	anatomical structure development embryo development post-embryonic development nematode larval development reproduction genitalia development establishment of localization transport receptor-mediated endocytosis	—	
CGAAGGATCAC	7.20E-11	0.00023	None	0.05	36	anatomical structure development embryo development post-embryonic development nematode larval development embryo devt ending in birth or egg hatching genitalia development developmental process involved in reproduction reproduction	nerve ring excretory cell	
CGCGCAAATGA	7.40E-09	0.024	Elemento Motif 95	0.05	26	localization embryo development embryo devt ending in birth or egg hatching glycoprotein metabolism protein glycosylation	—	
CGGCMGCGGC	1.40E-09	0.0045	Stormo W04C9.6.7	0.05	190	signal transduction cellular response to stimulus regulation of cellular process regulation of biological process response to chemical stimulus response to stimulus signalling localization establishment of localization transport macromolecule localization vesicle-mediated transport cell communication organic substance metabolism cellular nitrogen compound metabolism cellular metabolism macromolecule metabolism	—	
CGTGGYGAGAC	1.30E-17	4.50E-11	None	0.025	199	cellular protein metabolism macromolecule biosynthesis gene expression heterocycle metabolism cellular nitrogen compound metabolism nucleobase-containing compound metabolism cellular aromatic compound metabolism macromolecule metabolism biosynthesis transport receptor-mediated endocytosis vesicle-mediated transport localization	BWM germline	
CGYCAAGGCAC	1.10E-16	3.60E-10	Stormo T03F7.5.4	0.025	32	protein dephosphorylation regulation of vesicle-mediated transport	—	
CTAAAAAATCTY	5.60E-11	0.00018	None	0.025	28	regulation of cellular process regulation of response to stimulus regulation of signal transduction regulation of phosphorus metabolism regulation of phosphate metabolism small molecule metabolism	—	
CTGATGDTCTGA	3.60E-12	0.000012	None	0.025	29	transport cation transport ion transport ATP-hydrolysis coupled proton transport hydrogen transport vesicle mediated transport receptor-mediated endocytosis localization reproduction multicellular organismal development developmental process involved in reproduction molting cycle molting cycle, collagen, and cuticulin-based cuticle genitalia development hermaphrodite genitalia development	tail dorsal nerve cord ventral nerve cord nerve ring	
GAATTGCGYCA	7.20E-11	0.00023	None	0.05	31	phosphate-containing compound metabolism dephosphorylation peptidyl-tyrosine dephosphorylation phosphorus metabolism cellular metabolism	—	
GCRGCCGACA	8.90E-59	3.20E-52	None	0.025	202	regulation of biological process cellular process response to stimulus biological regulation	intestine vulval muscle BWM anal dep. muscle	

IUPAC Motif	p-val	E-val	Prior Match?	Threshold	#	Gene Ontology Enrichment	Anatomy Enrichment	Motif Logo
GDGGAGTACAC	4.80E-33	1.70E-26	Elemento Motif 89	0.05	108	RNA metabolism nucleobase-containing compound metabolism cellular macromolecule biosynthesis cellular nitrogen compound metabolism gene expression biosynthesis protein metabolism cellular aromatic compound metabolism heterocycle metabolism nitrogen compound metabolism cellular amino acid metabolism transport localization biosynthesis	spermatheca	
GGAGTGTCGTW	4.50E-13	1.50E-6	None	0.05	33	metabolism organic substance metabolism organophosphate metabolism phosphate-containing compound metabolism organophosphate biosynthesis phosphorus metabolism nucleoside metabolism ATP synthesis coupled proton transport	—	
GGANTCGAACC	3.30E-17	1.10E-10	None	0.05	44	localization metal ion transport	pharyngeal muscle tail neurons excretory cell head neurons	
GGASCTTTGCC	7.20E-11	0.00023	None	0.05	22	regulation of signal transduction --> positive regulation of response to stimulus positive regulation of cellular process positive reg. of metaphase/anaphase transition regulation of signalling regulation of cell communication regulation of response to stimulus aging determination of adult lifespan regulation of cellular process regulation of biological process	—	
GGCGCTGCTWA	5.40E-17	1.80E-10	None	0.025	40	regulation of precursor metabolites, energy regulation of cellular respiration	pharyngeal muscle anal dep. muscle	
GGGNTCGAACC	6.90E-37	2.40E-30	None	0.025	99	establishment of localization transmembrane transport cell communication establishment of localization in cell localization phosphorus metabolism phosphorate-containing compound metabolism response to stimulus developmental process involved in reproduction system development regulation of dauer larval development establishment or maintenance of cell polarity establishment of cell polarity	—	
GTGCGTCCGGY	9.20E-10	0.003	None	0.05	29	regulation of actin cytoskeleton organization glutamine family amino acid metabolism	—	
MAACAACAACAA	3.70E-11	0.00012	Stormo F58H7.3.1	0.025	42	reproduction response to external stimulus regulation of cellular process positive regulation of locomotion regulation of locomotion positive regulation of biological process	ventral cord neurons	

E Novel Footprint motifs

IUPAC Motif	p-val	E-val	Prior Match?	Threshold	#	Gene Ontology Enrichment	Anatomy Enrichment	Motif Logo
AGCAGCRGC	6.40E-07	8.90E-08	None	0.1	91	organic substance metabolism cellular metabolism	—	
CGCTGCTWA	8.50E-03	5.30E-16	None	0.1	37	nitrogen compound metabolism cellular aromatic compound metabolism heterocycle metabolism cellular nitrogen compound metabolism nucleobase-containing compound metabolism	—	
CTGCGTMTC	7.70E-13	3.00E-14	None	0.1	83	phosphate-containing compound metabolism organic substance metabolism nitrogen compound metabolism cellular biosynthesis	intestine pharynx excretory cell	
DCTCCGCC	2.60E-09	1.90E-09	None	0.1	51	organic substance metabolism macromolecule metabolism primary metabolism localization	—	

Appendix Table 2.4. DNase-seq data files

List of data files and sequence tracks to be made available for download and viewing through WormBase. Read data will be deposited in the NCBI Short Read Archive (SRA).

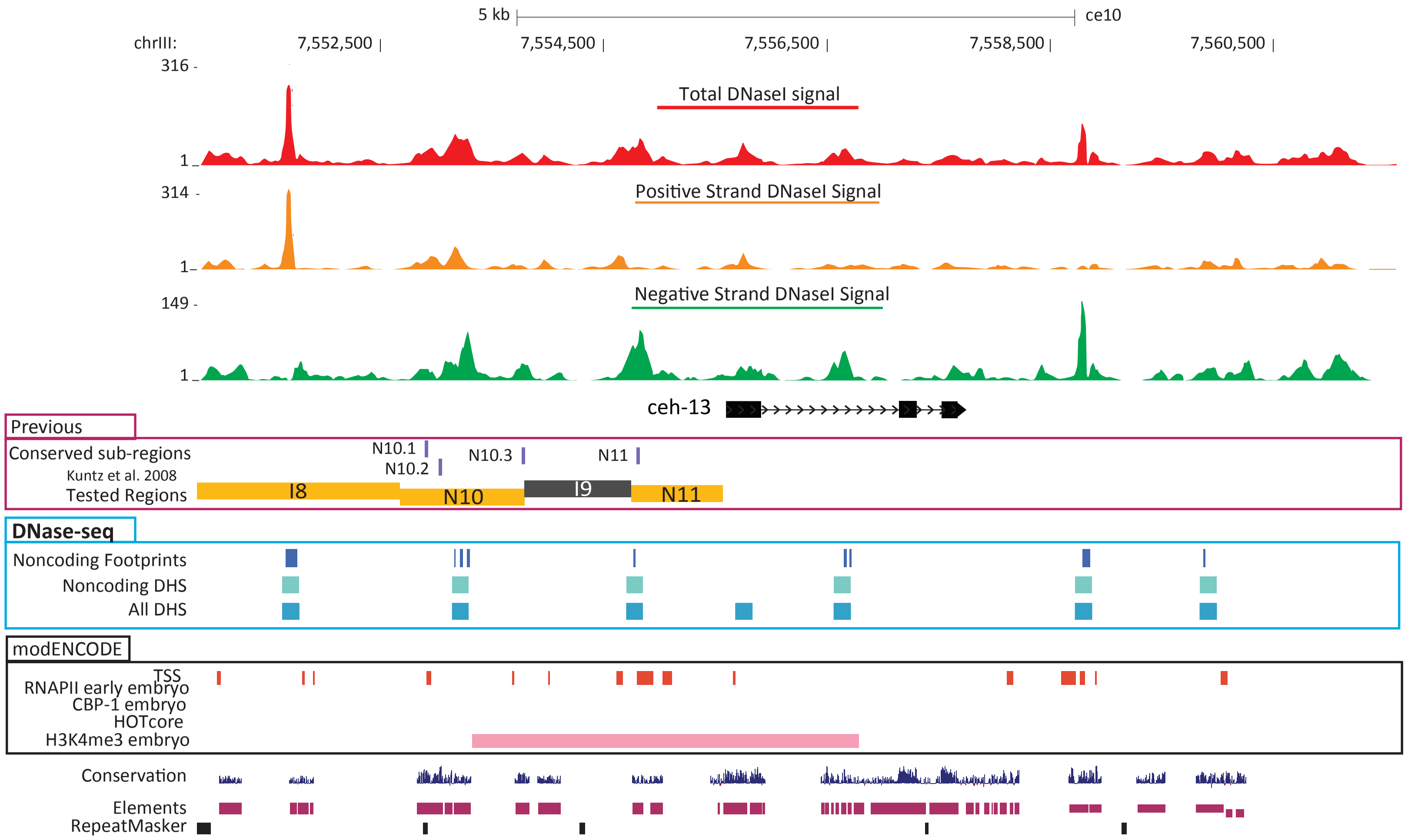
Appendix Table 2.4. DNaseI-seq data files

Filetype	File Name	Description
BigWig	merged.embryo.ce10.total.bw	Merged Embryo DNaseI signal (total)
BigWig	merged.embryo.ce10.positive.bw	Merged Embryo DNaseI signal (positive strand)
BigWig	merged.embryo.ce10.negative.bw	Merged Embryo DNaseI signal (negative strand)
BigWig	merged.L1.ce10.total.normalized.bw	Merged L1 DNaseI signal (total)
BigWig	merged.L1.ce10.positive.normalized.bw	Merged L1 DNaseI signal (positive strand)
BigWig	merged.L1.ce10.negative.normalized.bw	Merged L1 DNaseI signal (negative strand)
BigWig	merged.embryo.ce10.A.total.bw merged.embryo.ce10.B.total.bw merged.embryo.ce10.D.total.bw merged.embryo.ce10.C.total.bw	Embryo DNaseI signal (total) for each replicate A, B, C, D
BigWig	merged.embryo.ce10.A.positive.bw merged.embryo.ce10.B.positive.bw merged.embryo.ce10.C.positive.bw merged.embryo.ce10.D.positive.bw	Embryo DNaseI signal (positive strand) for each replicate A, B, C, D
BigWig	merged.embryo.ce10.A.negative.bw merged.embryo.ce10.B.negative.bw merged.embryo.ce10.C.negative.bw merged.embryo.ce10.D.negative.bw	Embryo DNaseI signal (negative strand) for each replicate A, B, C, D
BigWig	merged.L1.ce10.Z.total.bw merged.L1.ce10.Y.total.bw merged.L1.ce10.X.total.bw merged.L1.ce10.W.total.bw merged.L1.ce10.V.total.bw	L1 DNaseI signal (total) for each replicate Z, Y, X, W, V
BigWig	merged.L1.ce10.Z.positive.bw merged.L1.ce10.Y.positive.bw merged.L1.ce10.X.positive.bw merged.L1.ce10.W.positive.bw merged.L1.ce10.V.positive.bw	L1 DNaseI signal (positive strand) for each replicate Z, Y, X, W, V
BigWig	merged.L1.ce10.Z.negative.bw merged.L1.ce10.Y.negative.bw merged.L1.ce10.X.negative.bw merged.L1.ce10.W.negative.bw merged.L1.ce10.V.negative.bw	L1 DNaseI signal (negative strand) for each replicate Z, Y, X, W, V
BED	embryo.ce10.allDHS.bed	All DHS (post-IDR, filtered)
BED TXT	embryo.ce10.noncodingDHS.bed embryo.ce10.noncodingDHS_geneannot.txt	Noncoding DHS + gene annotation
BED	embryo.ce10.DHSfootprints.FDR0.05.bed	Embryo TF Footprints
BED	L1.ce10.allDHS.bed	All DHS (post-IDR, filtered)
BED TXT	L1.ce10.noncodingDHS.bed L1.ce10.noncodingDHS_geneannot.txt	Noncoding DHS + gene annotation
BED	L1.ce10.DHSfootprints.FDR0.05.bed	L1 Arrest TF Footprints
BED TXT	L1arrestspecific_noncodingDHS.bed L1arrestspecific_ncDHS_annot.txt	L1 arrest-specific noncoding DHS + gene annotation
TXT	embryo.DNaseI.novelmotifs.txt	Novel motifs (in MEME format)
TXT	embryo.DNaseI.motifassocgenes.txt	List of motif-associated genes for each motif
TXT	embryo.DNaseI.motifGO.txt embryo.DNaseI.motifanatomy.txt	Enriched Gene Ontology and anatomy terms for each motif

Appendix Figure 2.1. Additional known and novel enhancer CRMs

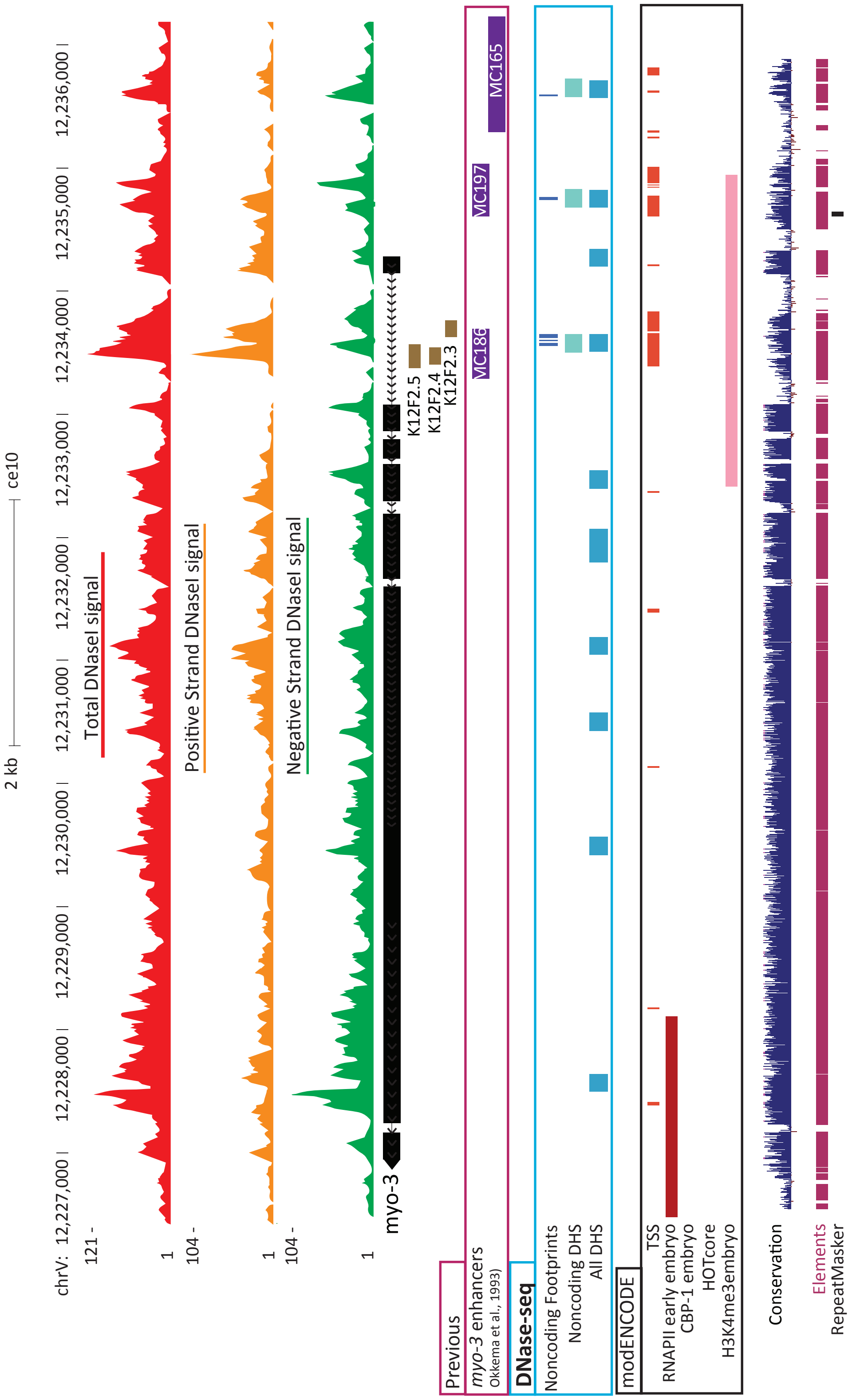
(A) I8 ("False Negative" in the Kuntz et al. study) detected, as well as N10 and N11 enhancers of *ceh-13*. A noncoding DHS containing TF footprints is detected in an evolutionarily conserved part of I8 region (reported as "false negative" in Kuntz et al. 2008) able to drive reporter expression. A second noncoding DHS containing TF footprints is also detected in the known highly conserved N10 enhancer. A third noncoding DHS harbors a TF footprint that overlaps with N11 enhancer and conserved MUSSA sub-region. Three other noncoding DHS containing TF footprints are detected in conserved regions downstream of *ceh-13* and in its first intron. These noncoding DHS are in regions of the *lin-39/ceh-13* Hox cluster not tested in the Kuntz et al. (2008) study but which are transcribed in embryos (Chen et al. 2013).

Appendix Figure 2.1A



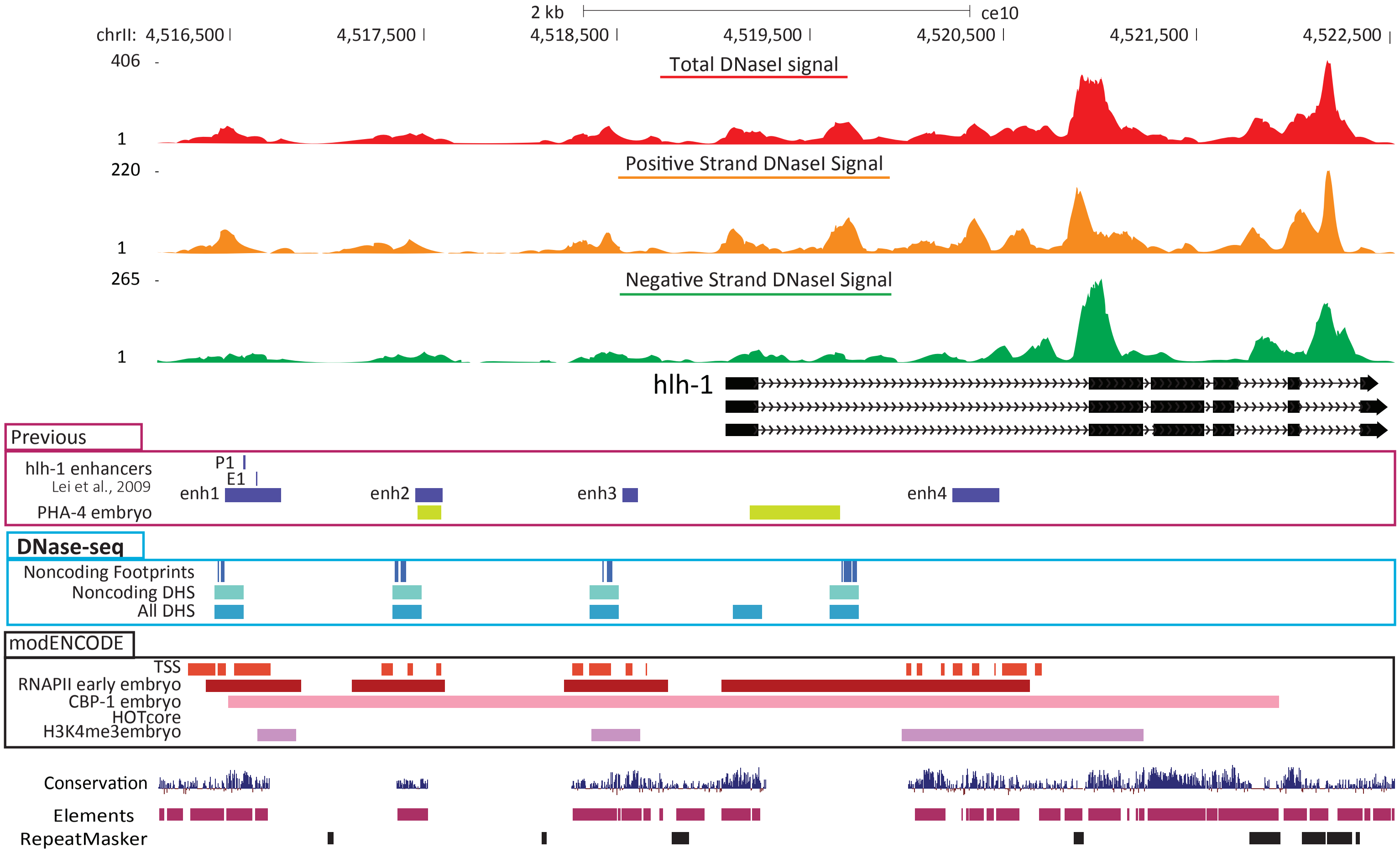
(B) Embryo noncoding DHS and footprints recapitulate known *myo-3* enhancers in 5' region and first intron. Two embryo noncoding DHS containing TF footprints are detected in 2kb region upstream of *myo-3* and overlap with MC197 and MC165 enhancers (purple). Another noncoding DHS is detected in the first intron which also harbors TF footprints and overlaps with MC186 enhancer (purple) and three ncRNA transcripts K12F2.5, K12F2.4, and K12F2.3. These noncoding DHS overlap with multiple TSS and MULTIZ conserved elements.

Appendix Figure 2.1B



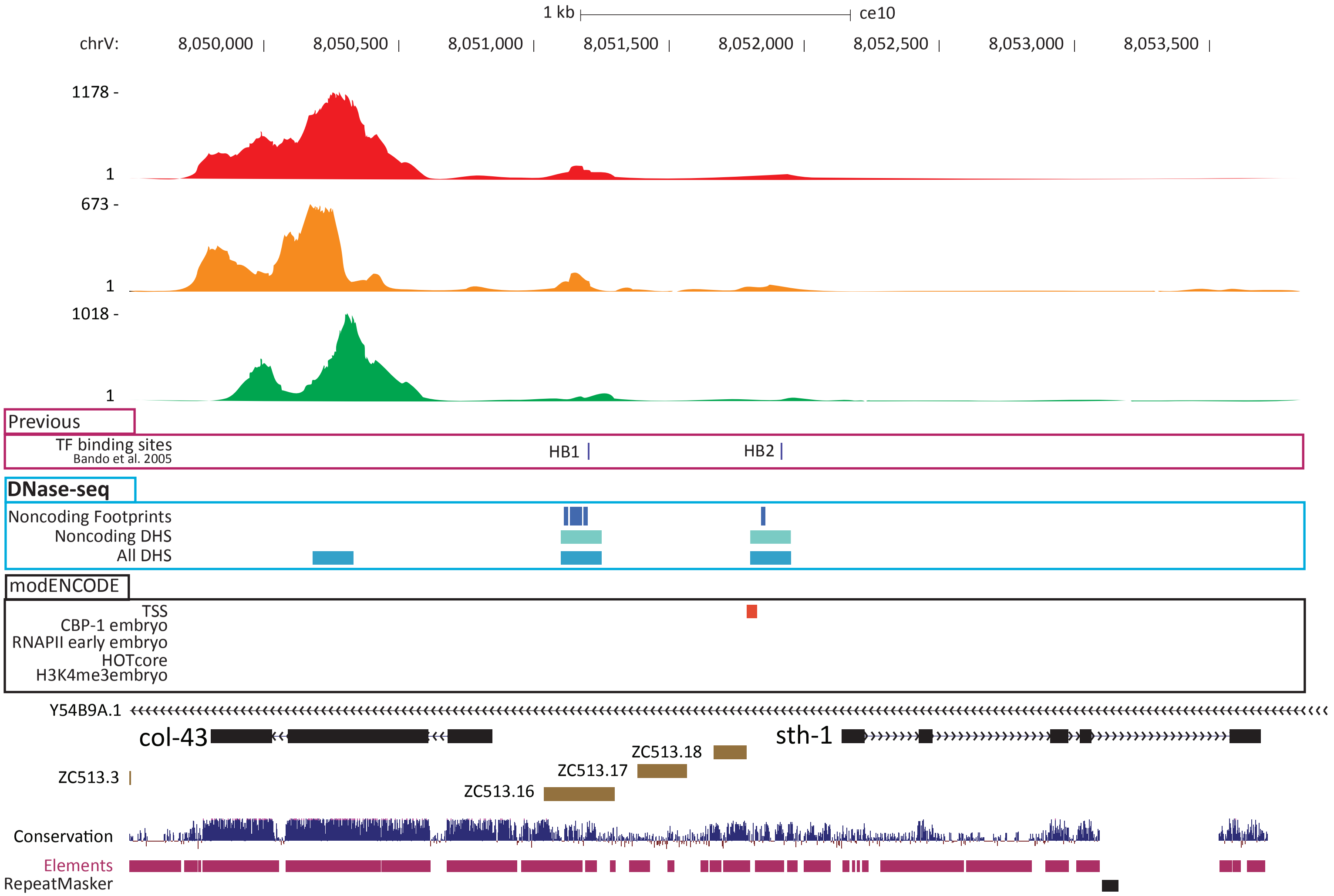
(C) Two known enhancers of *hlh-1* detected and additional intronic PHA-4 binding site. Three noncoding DHS harboring TF footprints are detected in 3kb region upstream of *hlh-1*, including the promoter, two of which overlap with known enh1 and enh2 enhancers (purple). These noncoding DHS overlap with conserved MULTIZ elements and marks of enhancer activity, such as RNAPII, CBP-1, TSS, and H3K4me3. Another noncoding DHS is detected in the first intron, and contains TF footprints which may correspond to regions of PHA-4 binding (Zhong et al. 2010).

Appendix Figure 2.1C



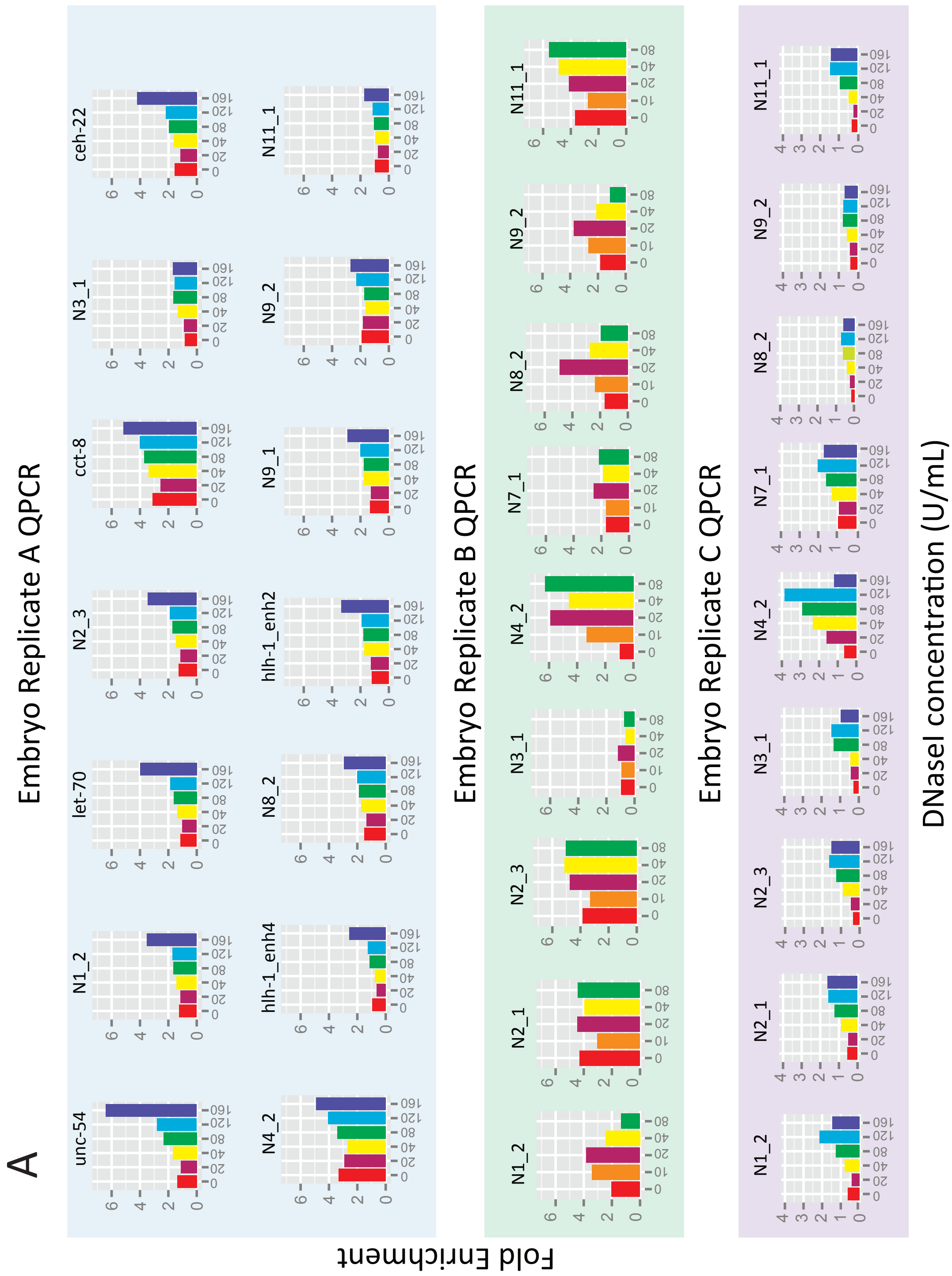
(D) Embryo noncoding DHS detected between *col-43* and *sth-1 and* overlap with homeodomain binding sites required for enhancer-blocking. Two noncoding DHS (light blue) harboring TF footprints are detected in intergenic region between *col-43* and *sth-1*. Of these, one overlaps with HB1 homeodomain site bound by MAB-18 and CEH-14 TFs as well as noncoding transcript ZC513.16 (Bando et al. 2005). Another overlaps with HB2 homeodomain site known to bind MAB-18 and a TSS (Chen et al. 2013). Homeodomain binding sites HB1 and HB2 shown in purple.

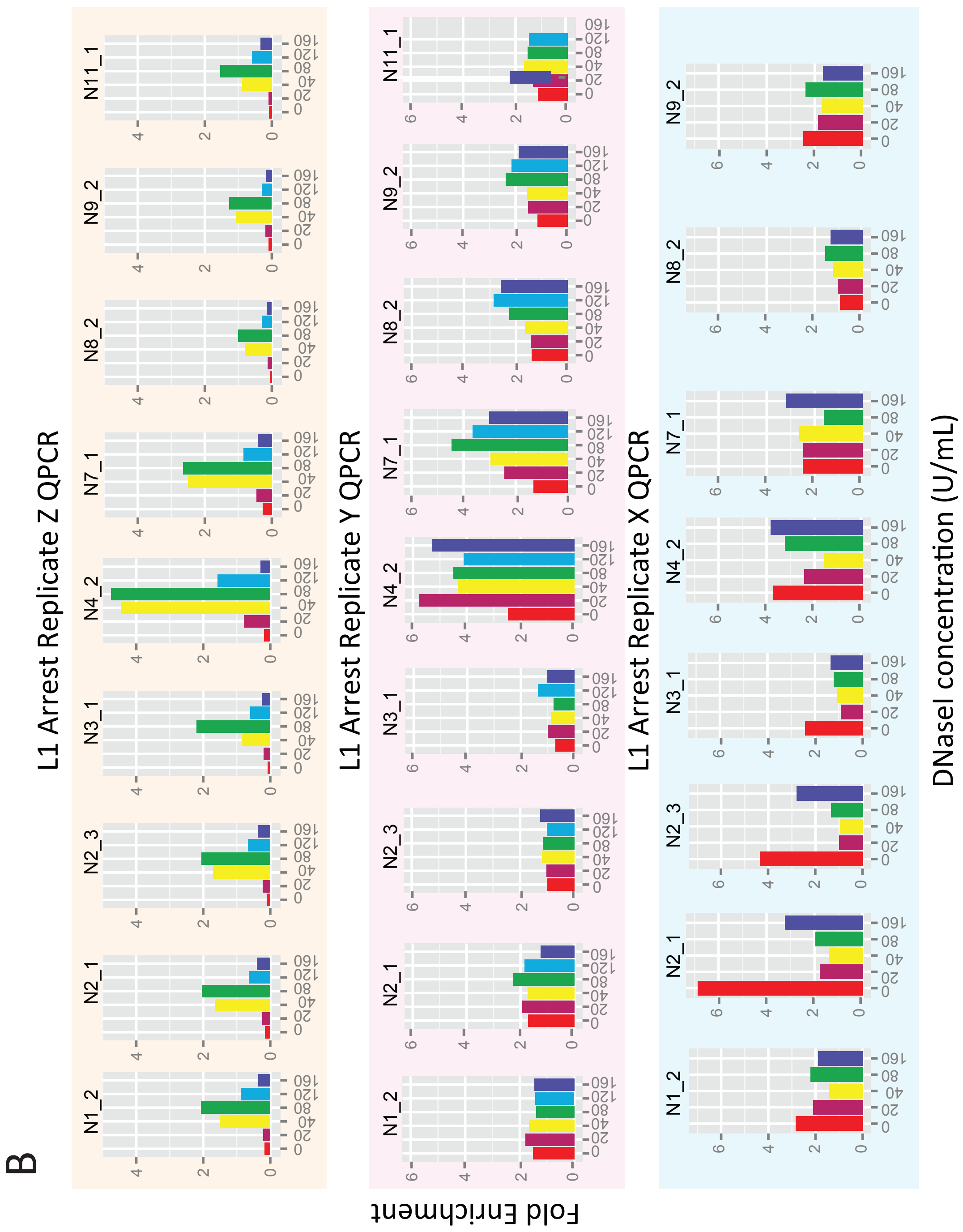
Appendix Figure 2.1D

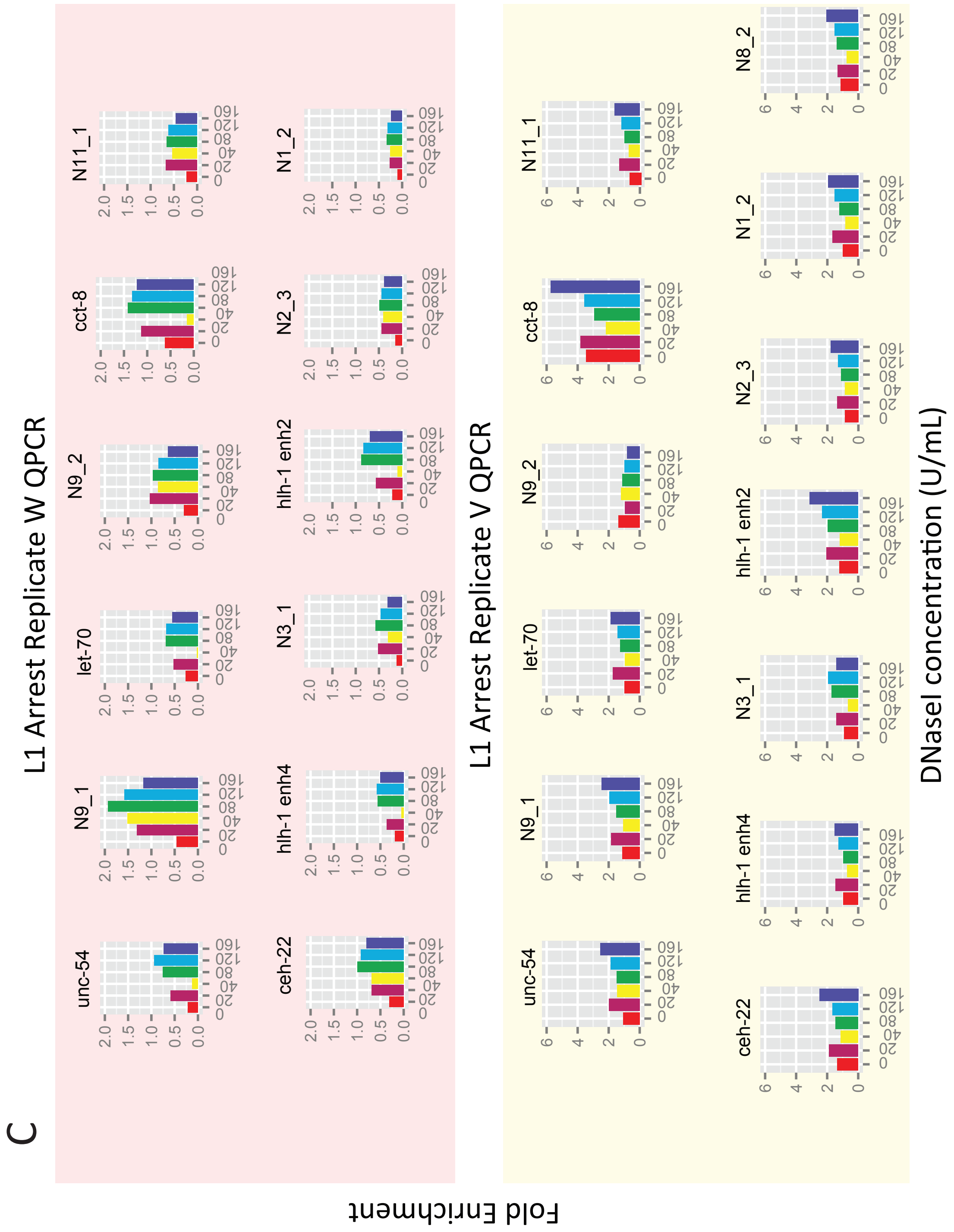


Appendix Figure 2.2. Regulatory enrichment by QPCR

QPCR was performed on DNaseI-treated DNA using primers designed to amplify conserved parts of known enhancers and negative control regions N5, N6 (see Methods and Table S2). Fold enrichment is measured by normalizing measured QPCR concentration by the average concentration of negative control regions. A range of DNaseI concentrations from 0 (red), 10 (orange), 20 (magenta), 40(yellow), 80 (green), 120 (blue), and 160 (purple) U/mL were used to treat each sample. The sample with DNaseI concentration exhibiting the highest relative fold regulatory enrichment was sequenced. In the cases of embryo replicates A-C, these were 160 U/mL, 80 U/mL, and 120 U/mL, respectively. In the cases of L1 arrest replicates X-Z, these were 80 U/mL, 20 U/mL, 160 U/mL, 80 U/mL and 160 U/mL, respectively.


A
Embryo Replicate A QPCR
unc-54
N1_2
let-70
N2_3
cct-8
N3_1
ceh-22
N4_2
hlh-1_enh4
N8_2
hlh-1_enh2
N9_1
N9_2
N11_1
Embryo Replicate B QPCR
N1_2
N2_1
N2_3
N3_1
N4_2
N7_1
N8_2
N9_2
N11_1
Embryo Replicate C QPCR
N1_2
N2_1
N2_3
N3_1
N4_2
N7_1
N8_2
N9_2
N11_1
0
2
4
6
0
1
2
3
4
0
10
20
40
80
120
160
Fold Enrichment
DNaseI concentration (U/mL)

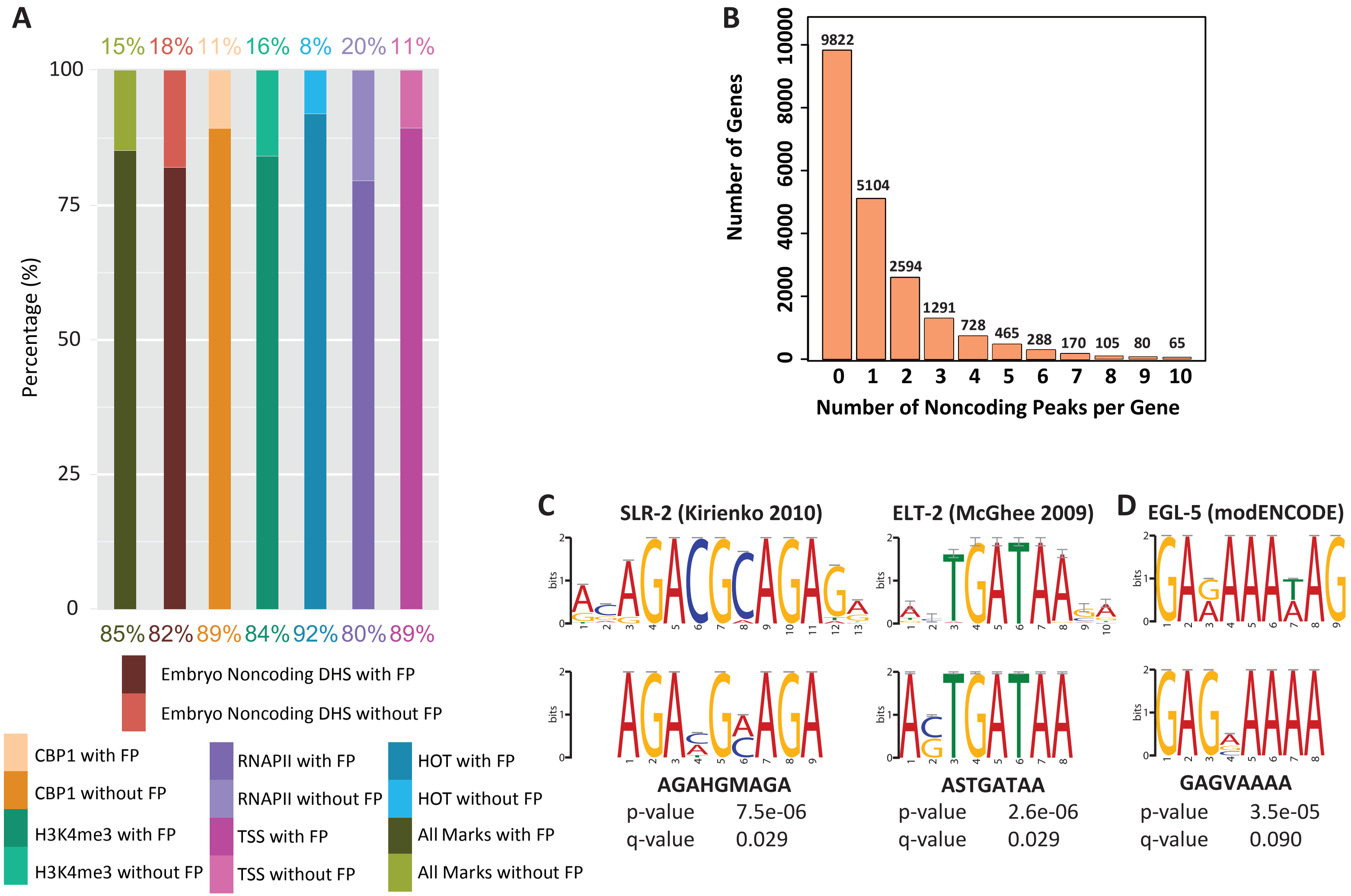

B
L1 Arrest Replicate Z QPCR
L1 Arrest Replicate Y QPCR
L1 Arrest Replicate X QPCR
N1_2
N2_1
N2_3
N3_1
N4_2
N7_1
N8_2
N9_2
N11_1
Fold Enrichment
DNaseI concentration (U/mL)
0
20
40
80
120
160
0
2
4
6


C
L1 Arrest Replicate W QPCR
unc-54
N9_1
let-70
N9_2
cct-8
N11_1
ceh-22
hlh-1 enh4
N3_1
hlh-1 enh2
N2_3
N1_2
0.0
0.5
1.0
1.5
2.0
0
20
40
80
120
160
L1 Arrest Replicate V QPCR
unc-54
N9_1
let-70
N9_2
cct-8
N11_1
ceh-22
hlh-1 enh4
N3_1
hlh-1 enh2
N2_3
N1_2
N8_2
0
2
4
6
Fold Enrichment
DNaseI concentration (U/mL)

Appendix Figure 2.3. Frequency of TF footprints, Noncoding DHS, and genes and motifs predicted from noncoding DHS for tissue-specific gene sets

(A) Percentage of embryo noncoding DHS containing footprints for different promoter/enhancer-associated marks. The number of noncoding DHS that were observed with footprints (darker shading) or without footprints (lighter shading) are shown for each type of enhancer-associated mark: TSS (pink), H3K4me3 (emerald green), RNAPII (purple), CBP-1 (orange), HOT (blue) or All Marks (lime green), and for the noncoding DHS as a whole (red). (**B) Number of embryo noncoding DHS per gene.** Distribution of noncoding DHS overlapping near protein-coding genes shows that 53% (10,890) of protein-coding genes were assigned at least one embryo noncoding DHS nearby, according to annotation that assigned the nearest gene to each noncoding DHS. 9822 (47%) of genes were not annotated with nearby embryo noncoding DHS. 17% (1,901) were annotated with more than four embryo noncoding DHS. **(C) Known gut motifs identified.** Two motifs identified in our analysis of overrepresented motifs in noncoding DHS of gut-expressed genes (genes identified in SAGE of dissected adult *C. elegans* intestine by McGhee et al. 2007) match known binding motifs of two gut TFs, SLR-2 and ELT-2. Shown are the motif comparisons between the identified motifs from DREME and the consensus motifs (Kirienko and Fay 2010; McGhee et al. 2009) and their associated *p* and *q*-value measured by TOMTOM. **(D) Known neuronal motif identified.** One motif identified in our analysis of overrepresented motifs in genes expressed in neurons (genes identified in SAGE of FACS-sorted neurons by Spencer et al. 2011) matches known binding motif of one neuronal TF, EGL-5. Shown is motif comparison between identified motif from DREME and consensus motif from Gerstein et al. (2010) and the associated *p* and *q*-value measured by TOMTOM.

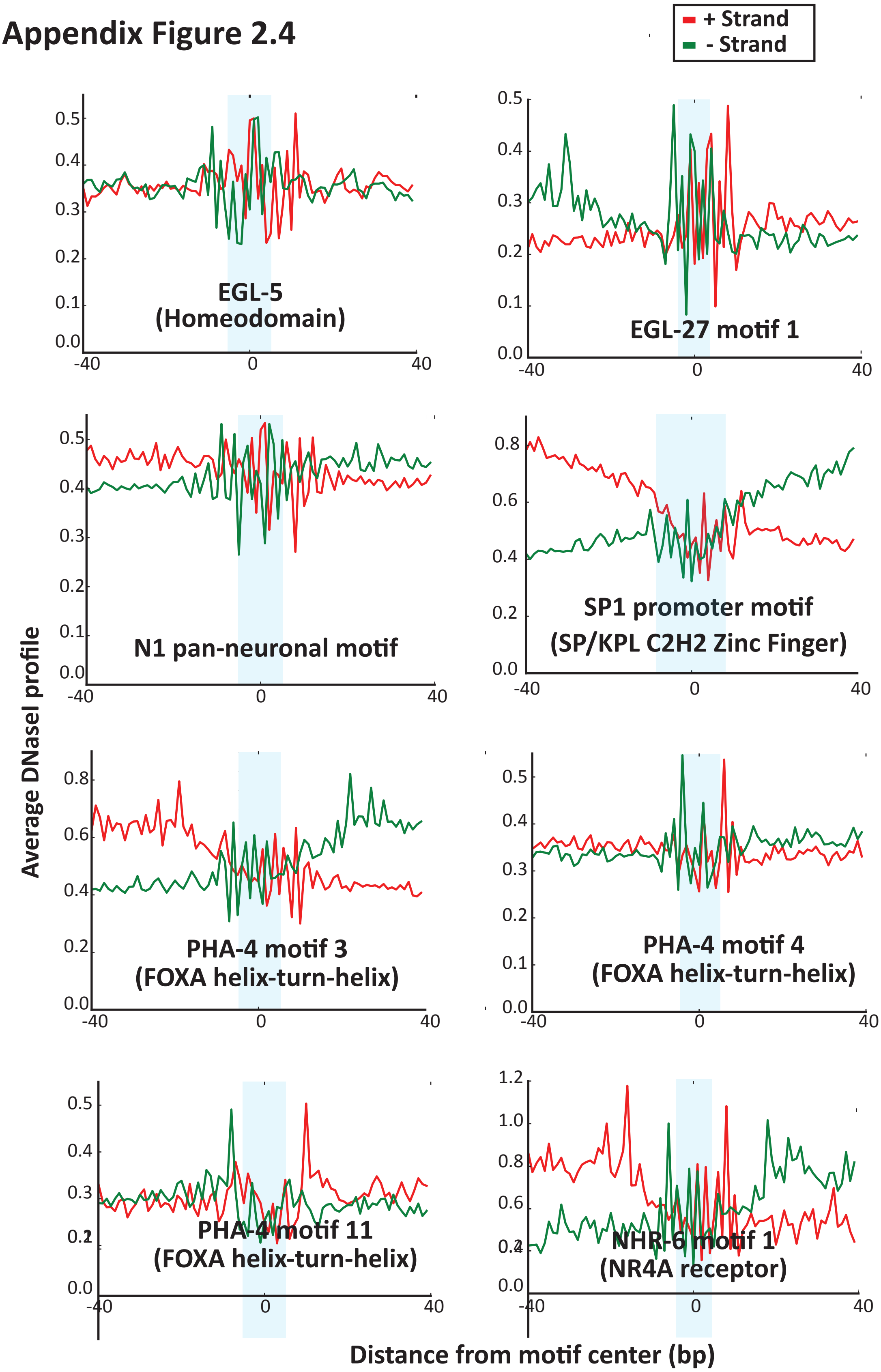
Appendix Figure 2.3



Appendix Figure 2.4. Average DNaseI profile over *C. elegans* motif sites

Known *C. elegans* regulatory motif sites show characteristic patterns of accessibility to DNaseI cleavage and demonstrate strand-shift in reads that is indicative of TF footprints. Average DNaseI profile is calculated over thousands of predicted motif sites within the 2 kb upstream region of genes using the start sites of reads across 80bp region surrounding the motif. Positive strand is shown in red and negative strand is shown in green. Light blue indicates the base pair position of the motif site: EGL-5 (9bp), EGL-27 (8bp), N1 (10bp), SP1 (15bp), PHA-4 motifs 3 (10bp), 4, 11 (9bp), and NHR-6 (7bp).

Appendix Figure 2.4

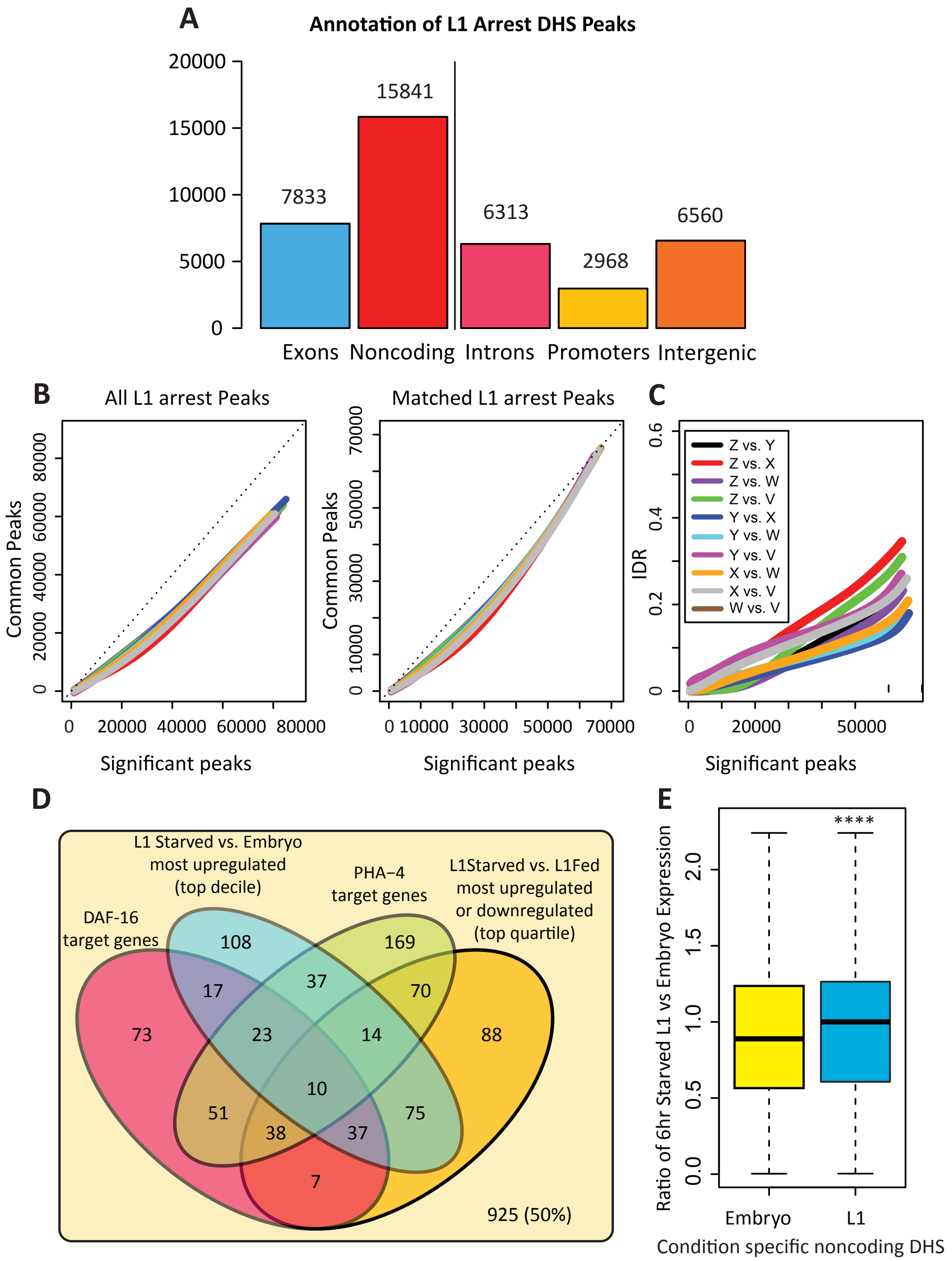


Appendix Figure 2.5. L1 stage specific DHS are more highly expressed in L1 arrest compared to the embryo and are found in genes that are targets of DAF-16 and PHA-4 and whose expression is affected by starvation

(A) Genomic location of L1 arrest DHS shows abundance of noncoding DHS. L1 arrest DHS were annotated according to position relative to WormBase WS241 protein-coding genes: exons (blue) and noncoding (red). Noncoding DHS are further subdivided into introns (pink), promoter (defined as less than 300bp 5' of ATG; yellow) and intergenic (orange) regions. 67% of L1 arrest DHS were annotated in noncoding regions, with 33% annotated in exons. Within L1 arrest noncoding DHS, 27%, 13%, and 28% were annotated in introns, promoters, and intergenic regions, respectively. **(B) L1 Arrest biological replicates show reproducibility of matched peaks.** Comparison between number of common peaks and significant peaks in pairs of L1 arrest biological replicates when all raw peaks are assessed together (All Peaks) or peaks matching in replicates (Matched Peaks). Pairwise comparisons of L1 arrest biological replicates: A vs. Z vs. Y (black), Z vs. X (red), Z vs. W (purple), Z vs. V (green), Y vs. X (blue), Y vs. W (light blue), Y vs. V (violet), X vs. W (orange), X vs. V (grey), W vs. V (brown) are shown. **(C) Observed relationship between irreproducible discovery rate (IDR) threshold and number of significant peaks called in biological replicates.** 49,882 reproducible L1 arrest DHS peaks remained after IDR filtering using threshold 0.1. Filtering for ce10 blacklist regions and repeat regions resulted in 23,670 L1 arrest DHS peaks. **(D) Genes associated with L1 condition-specific noncoding DHS include many DAF-16 and PHA-4 targets, starvation responsive genes, and genes upregulated in the 6hr starved larvae compared to the embryo.** Venn diagram showing number of genes associated with L1 condition-specific regulatory elements that are DAF16 target genes (pink;

from Tepper et al. 2013), PHA-4 target genes in L1 Starved (green; from Zhong et al, 2010) and starvation responsive genes (yellow; significant expression difference in 6hr starved versus 6hr fed L1 larvae, Baugh et al. 2009) and that are most upregulated in L1 starved larvae compared to the embryo (blue; data from Baugh et al. 2009). **(E) Expression ratio of genes possessing L1 or embryo condition-specific noncoding DHS** Boxplot showing the ratio of expression of genes possessing L1 (blue) or embryo (yellow) condition-specific noncoding DHS. Ratio is measured by dividing the expression observed in 6hr L1 starved larvae by embryo expression (data from Baugh et al. 2009).

Appendix Figure 2.5


A
Annotation of L1 Arrest DHS Peaks
Exons
Noncoding
Introns
Promoters
Intergenic
B
All L1 arrest Peaks
Common Peaks
Significant peaks
Matched L1 arrest Peaks
Common Peaks
Significant peaks
C
IDR
Significant peaks
Z vs. Y
Z vs. X
Z vs. W
Z vs. V
Y vs. X
Y vs. W
Y vs. V
X vs. W
X vs. V
W vs. V
D
L1 Starved vs. Embryo most upregulated (top decile)
PHA−4 target genes
L1Starved vs. L1Fed most upregulated or downregulated (top quartile)
DAF-16 target genes
925 (50%)
E
Ratio of 6hr Starved L1 vs Embryo Expression
****
Embryo
L1
Condition specific noncoding DHS