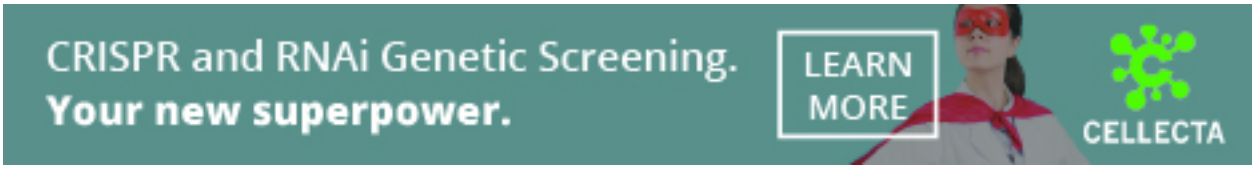
Genome-wide discovery of active regulatory elements and transcription factor footprints in *Caenorhabditis elegans* using DNase-seq

Margaret C.W. Ho, Porfirio Quintero-Cadena and Paul W. Sternberg

Genome Res. 2017 27: 2108-2119 originally published online October 26, 2017
Access the most recent version at doi:10.1101/gr.223735.117



Resource

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

Margaret C.W. Ho, Porfirio Quintero-Cadena, and Paul W. Sternberg

Division of Biology and Bioengineering, Howard Hughes Medical Institute, California Institute of Technology, Pasadena, California 91125, USA

Deep sequencing of size-selected DNase I–treated chromatin (DNase-seq) allows high-resolution measurement of chromatin accessibility to DNase I cleavage, permitting identification of de novo active *cis*-regulatory modules (CRMs) and individual transcription factor (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 DNase I hypersensitive sites (DHSs) were annotated as noncoding, with 24% in intergenic, 12% in promoters, and 28% in introns, with similar statistics observed in L1 arrest larvae. Noncoding DHSs are highly conserved and enriched in marks of enhancer activity and transcription. We validated noncoding DHSs against known enhancers from *myo–2*, *myo–3*, *hlh–1*, *elt–2*, and *lin–26/lir–1* and recapitulated 15 of 17 known enhancers. We then mined DNase-seq data to identify putative active CRMs and TF footprints. Using DNase-seq data improved predictions of tissue-specific expression compared with motifs alone. In a pilot functional test, 10 of 15 DHSs from *pha–4*, *icl–1*, and *ceh–13* drove reporter gene expression in transgenic *C. elegans*. Overall, we provide experimental annotation of 26,644 putative CRMs in the embryo containing 55,890 TF footprints, as well as 15,841 putative CRMs in the L1 arrest larvae containing 32,685 TF footprints.

[Supplemental material is available for this article.]

Prior research in metazoans has described many types of *cis*-regulatory modules (CRMs) such as enhancers, repressors, and insulators that can be located far from target genes (for review, see Noonan and McCallion 2010). The nematode *Caenorhabditis elegans* has a well-annotated genome and well-studied development (Boulin and Hobert 2012; Harris et al. 2014). *C. elegans* provides an excellent opportunity to study transcriptional regulation within a multicellular organism, especially as it is easy to collect large numbers of developmentally synchronized worms.

Traditional approaches to identify CRMs have relied on individually testing conserved sequences in transgenic reporter assays, but these are limited by relatively low throughput. Many enhancers have been found this way in *C. elegans*, of which most are located close (<2 kb away) to the target gene. This preponderance may be due to experiments focusing on testing sequences from promoter-proximal regions of genes. Some distant CRMs have been found (for review, see Gaudet and McGhee 2010) such as N2, N3, and N4 enhancers located 18–20 kb away from their target *ceh-13* (Kuntz et al. 2008). Overall, systematic identification of *C. elegans* CRMs has proved difficult.

ChIP-seq experiments, which measure binding of a TF of interest to regions of the genome, generate data that can be mined for putative CRMs (Ren et al. 2000; Robertson et al. 2007; Visel et al. 2009). However, these experiments require prior knowledge of TFs and provide information for a single TF at a time. Thus, an experimental method that allows high-throughput discovery of CRMs and regulatory TF sites de novo in *C. elegans* is desirable.

Active CRMs are known to be hypersensitive to DNase I cleavage (Gross and Garrard 1988). Studies in other animals and plants have utilized deep sequencing of DNase-treated chromatin (DNase-seq) to map protein–DNA interactions de novo (Hesselberth et al. 2009; Boyle et al. 2011; Thomas et al. 2011; Thurman et al. 2012; Sullivan et al. 2014). In addition to identifying DNase I hypersensitive (DHS) regions that may act as putative CRMs, DNase-seq can identify shorter sequences within DHSs protected from nuclease cleavage representing putative TF binding sites (TFBSs).

In this study, we aim to adapt the DNase-seq technique to *C. elegans* to generate genome-wide maps of active CRMs and putative TF footprints during development. We hope that these data and analyses will serve as a valuable resource to identify novel CRMs and regulatory motifs for better understanding of *C. elegans* gene regulation.

Results

A DNase-seq method for *C. elegans*

We performed DNase I treatment on *C. elegans* embryos (at roughly 40-cell stage) and L1 arrest larvae and isolated small DNA fragments representing chromatin regions most accessible to DNase I cleavage (Supplemental Fig. S1A). qPCR was used to identify DNase treatment conditions that resulted in the highest enrichment of regulatory regions in the DNase-seq sample, using primers designed against known CRMs from *lin-39/ceh-13* Hox cluster (Kuntz et al. 2008) and negative control regions lacking any known activity. DNase-seq samples were sequenced to 15×

Corresponding author: pws@caltech.edu
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.223735.117.

27:2108–2119 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/17; www.genome.org

coverage of the *C. elegans* genome, and the sequence read data were analyzed to find regions with increased hypersensitivity across 150 bp of consecutive nucleotides (Supplemental Fig. S1B). Raw peak calls were filtered using the irreproducibility discovery rate (IDR) framework developed for the ENCODE Project (Li et al. 2011; The ENCODE Project Consortium 2012). The IDR is analogous to false-discovery rate (FDR) but also considers quantitative reproducibility of the results. Peaks were filtered using a combination of rank or score and consistency across at least three replicates to yield 41,825 embryonic and 23,674 L1 arrest DHSs (Supplemental Figs. S2B, S7C; for reproducibility analysis, see Supplemental Information).

Comparing with the WormBase WS241 version of *C. elegans* genome annotation, 26,644 embryonic and 15,841 L1 arrest DHSs, respectively, were found in noncoding genomic regions and represent putative active CRMs in these conditions. We searched for signatures of TF footprints using DNase2TF (Sung et al. 2014; see Supplemental Methods). We identified 55,890 and 32,685 putative TF footprints within noncoding DHSs in the *C. elegans* embryo and L1 arrest, respectively. We also observe 1835 DHSs in the L1 arrest that are not found in embryos (which we now refer to as L1 arrest–associated DHSs) containing 2964 TF footprints.

DHS peaks are most abundant in noncoding regions

Annotation of peaks with WS241 gene models revealed that less than half (36%) occur within exons (Fig. 1A), which recalls previous observations of DHSs often occurring in exons of active genes (Mercer et al. 2013). Over half of DHS peaks (64%) were observed in noncoding regions, with 28% in introns, 24% in intergenic regions, and 12% in promoters (defined as <300 bp of exon start). These noncoding DHSs may represent candidate CRMs. Similar statistics were observed in L1 arrest larvae (Supplemental Fig. S7A).

DNase hypersensitivity of genes correlates with expression

Most genes exhibited a uniform distribution of reads over the gene body and surrounding sequence with an average of 20 mapped reads per base pair, reflecting their level of DNase hypersensitivity (Fig. 1B). However, ~9% of genes exhibited much higher read coverage and showed a pattern of three peaks of read enrichment reaching as high as 120 mapped reads per base pair. These peaks correspond to the 5′ upstream region, gene body, and 3′ downstream region. We observe that this subset of genes with higher and trimodal pattern of read enrichment are 66% more highly expressed in embryo than genes with lower and uniform pattern (two-sided Kolmogorov-Smirnov [KS] test, $P = 1.1 \times 10^{-8}$) (Fig. 1B).

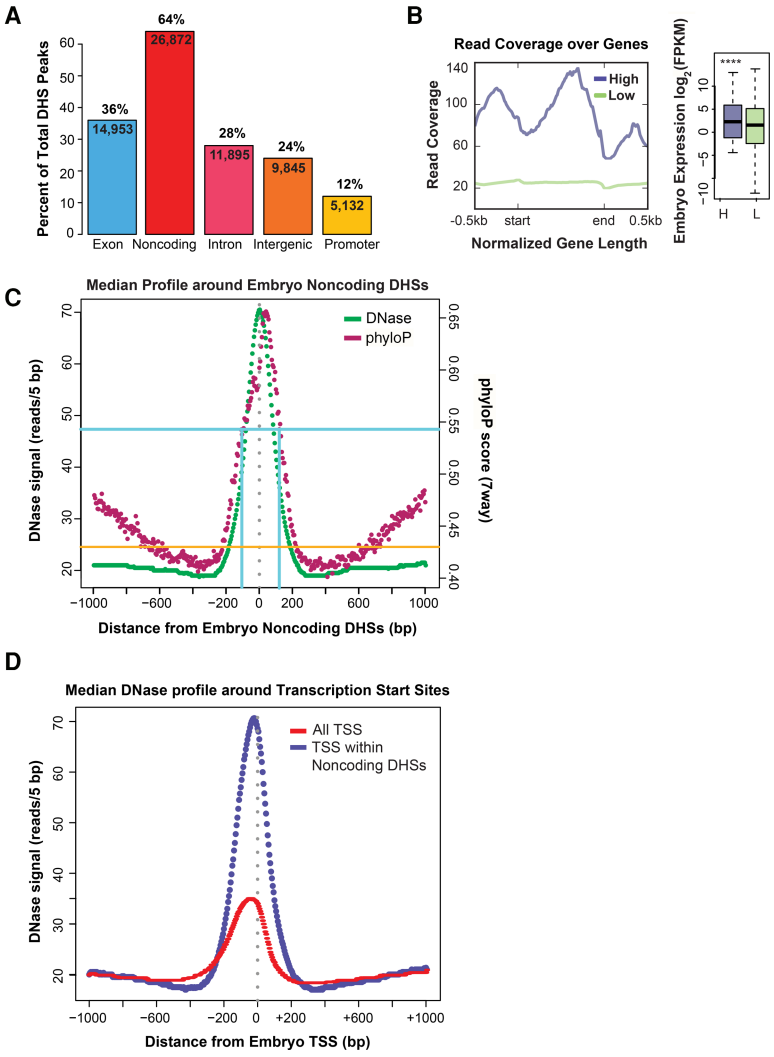


Figure 1. Noncoding DHSs are highly conserved and accessible to DNase. (*A*) Noncoding DHSs are abundant in the embryo. Embryo DHSs were annotated according to position relative to WormBase WS241 protein-coding genes: exons (blue) and noncoding (red). Noncoding DHSs are further subdivided into introns (pink), promoter (defined as <300 bp 5′ of ATG; yellow), and intergenic (orange) regions. (*B*) Protein-coding genes with higher DNase accessibility have higher expression. Read coverage (Total DNase signal across biological replicates) was measured for length-normalized protein-coding genes and 1 kb 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 (measured in $\log_2$ of FPKM from Zhong et al. 2010) was compared between higher (H) versus lower (L) read coverage genes. (*C*) Embryo noncoding DHSs are highly conserved and highly accessible to DNase. Median DNase signal (green; measured in 5-bp windows) and phyloP sequence conservation score (pink; seven-way) are measured across 2 kb of sequence centering around embryo noncoding DHSs. Read coverage maximizes at 70.5 reads in a 5-bp window and phyloP sequence conservation at 0.66. In comparison, phyloP conservation is 0.54 (blue) for known true positive *lin-39/ceh-13* (Kuntz et al. 2008) and is 0.43 for negative control nonenhancer regions (orange line). (*D*) Median DNase signal peaks at *C. elegans* transcription start sites (TSS) and shows a 5′ bias. Median DNase signal (measured in 5-bp windows) is measured across 2 kb of sequence centering around embryo TSS (locations from Chen et al. 2013), with 5′ to 3′ shown from *left* to *right* (following the direction of transcription). Center of the TSS is by gray dotted line. DNase signal peaks at All TSS (red) and at TSS within noncoding DHSs (purple) and shows strongest DNase accessibility just 5′ to the TSS.

Noncoding DHSs are highly conserved and enriched in marks of enhancer activity and transcription

DNase hypersensitivity strongly correlates with sequence conservation on a per nucleotide basis around noncoding DHSs (Fig. 1C). Both hypersensitivity and sequence conservation maximize at the midpoint of noncoding DHSs. When comparing with levels of sequence conservation of known enhancer CRMs in the *lin-39/ceh-13* Hox complex (median phyloP sequence conservation is 0.543 for these enhancers [from Kuntz et al. 2008]), this would suggest a typical size for CRMs of *C. elegans* of ~200 bp (Fig. 1C). A typical size noncoding DHS of 150 bp from our DNase-seq data thus appears to capture the bulk of both DNase hypersensitivity and sequence conservation. Noncoding DHSs are on average twice as conserved on a per nucleotide basis than expected by chance ($P < 3 \times 10^{-16}$).

Embryo noncoding DHSs are enriched in embryonic sites of transcription initiation (TSS) (4.2-fold, $P < 3 \times 10^{-16}$) (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 400 bp of sequence, demonstrating high accessibility to DNase cleavage, with higher accessibility within noncoding DHSs (Fig. 1D). DNase signal was strongest in the proximal 5′ region of the TSS, suggesting that upstream regions of these promoters are accessible. Comparison to data from another study using GRO-cap sequencing to find *C. elegans* TSS also showed that DHSs are enriched in stage-matched TSS identified by this study (7.9- and 7.7-fold in embryo and L1 arrest, respectively; ($P < 3 \times 10^{-16}$) (Kruesi et al. 2013).

Compared with stage-matched H3K4me3 ChIP-seq and *C. elegans* p300 homolog CBP-1 ChIP-chip peaks from modENCODE, embryo noncoding DHSs are enriched in marks associated with enhancer regulatory activity in eukaryotic genomes (2.8-fold, $P < 3 \times 10^{-16}$) (Heintzman et al. 2007). Also, two thirds (65%) of high occupancy target (HOT) core regions (bound by 15 or more TFs tested by modENCODE; Gerstein et al. 2010) overlap with embryo DHSs (5.1-fold enriched, $P < 3 \times 10^{-16}$). Embryo DHSs are also enriched in RNA polymerase II binding identified (Gerstein et al. 2010) in early embryos (1.4-fold, $P < 3 \times 10^{-16}$).

Nearly half (46%) of noncoding DHSs overlap with one or more marks of transcription (initiation sites, CBP-1 transcriptional coactivator, RNA polymerase II, H3K4me3 histone marks) or high TF occupancy (modENCODE HOT regions) from stage-matched samples (Fig. 2A). Of these, most (57%, 6956) overlap with one type of mark, with progressively fewer overlapping with greater numbers of marks. Genes associated with noncoding DHSs possessing one or more marks are on average 8.9-fold more highly expressed in embryos compared with genes with noncoding DHSs lacking any marks ($P < 2.2 \times 10^{-16}$) (Fig. 2B; see Supplemental Methods). Moreover, genes associated with embryo noncoding DHSs overlapping with greater numbers of marks correlates with increased embryonic expression, up to three marks (5.1-fold higher compared to one mark, $P < 3 \times 10^{-14}$) (Fig. 2B).

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

Over half of protein-coding genes were assigned at least one DHS nearby (Supplemental Fig. S5B). The presence of at least one embryo noncoding DHS near a gene was associated with higher embryo expression compared with genes lacking DHSs (4.5-fold, $P < 3 \times 10^{-16}$) (see Supplemental Methods for details; Fig. 2C). More embryo noncoding DHSs near a gene correlates with increased embryonic expression, up to three DHSs. From one to two embryo noncoding DHSs near a gene, there is a 54% increase in embryo expression ($P < 3 \times 10^{-6}$); from two to three there is a 44% increase ($P < 0.007$), up to three noncoding DHSs. Genes with noncoding DHSs are likely active even without apparent regulatory marks, since they have double the expression of genes lacking any DHSs ($P < 3 \times 10^{-16}$) (Fig. 2D).

Within noncoding embryo DHS peaks, we identified 55,890 potential TF binding sites using the software DNase2TF (Sung et al. 2014). Nearly all (82%) of the noncoding DHSs found possessed detectable footprints (Supplemental Fig. S5A). This pattern is consistent across noncoding DHSs with varying levels of chromatin regulatory marks (Supplemental Fig. S5A). No differences in expression were found between genes associated with DHSs with varying numbers of footprints.

Overall, these data indicate that noncoding DHS peaks have many hallmarks of CRMs, including sequence conservation, active transcription, enhancer-associated chromatin regulatory marks, and TF occupancy, and their presence near a gene correlates with increased expression.

Noncoding DHSs coincide with many known CRMs in *C. elegans* regulatory loci

To investigate whether the locations of previously investigated enhancers can be identified by our DNase-seq method, we examined several well-studied *C. elegans* genes for embryo noncoding DHSs. Several previous studies identified CRMs for *lin-26*, *elt-2*, *myo-3*, and *myo-2* (Okkema et al. 1993; Okkema and Fire 1994; Landmann et al. 2004; Wiesenfahrt et al. 2016), using transgenic enhancer assays. These genes represent major tissue regulators or structural genes that are all expressed during embryonic development.

The epithelial differentiation factor *lin-26* begins to be expressed in early embryos in all epithelial cells of the ectoderm (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 precomma stage and is eventually expressed in all muscle cells outside of the pharynx (Okkema et al. 1993; Fox et al. 2007). *myo-2* is a myosin heavy-chain gene whose expression begins later in the two-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 DNase hypersensitivity in embryos.

Proper expression of *lin-26* is controlled by upstream CRMs 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 A+B, C+D, E, F+G, and H enhancers. PHA-4 is known to bind and repress *lin-26* (Kiefer et al. 2007). *lin-26* and *elt-3* act combinatorially to establish epithelial cell fate (for review, see Chisholm and Hardin 2005) and are both activated by ELT-1. A+B and C+D, which are both bound and regulated by PHA-4 (Zhong et al. 2010), both harbor noncoding DHSs and footprints (Fig. 3A). Although modENCODE data show binding of ELT-3 binding to F+G and A+B *lin-26* CRMs in L1 and embryos, there is no detectable binding by ELT-1 (in L2/L3 ChIP-seq stages for which data are available) (Gerstein et al. 2010). One of the noncoding DHSs that we detected and its footprints overlap one of these ELT-3 ChIP peaks in the F+G CRM.

The promoter and 5′ upstream region of *elt-2* shows several DHSs that coincide with ELT-2 ChIP-seq peaks (Wiesenfahrt

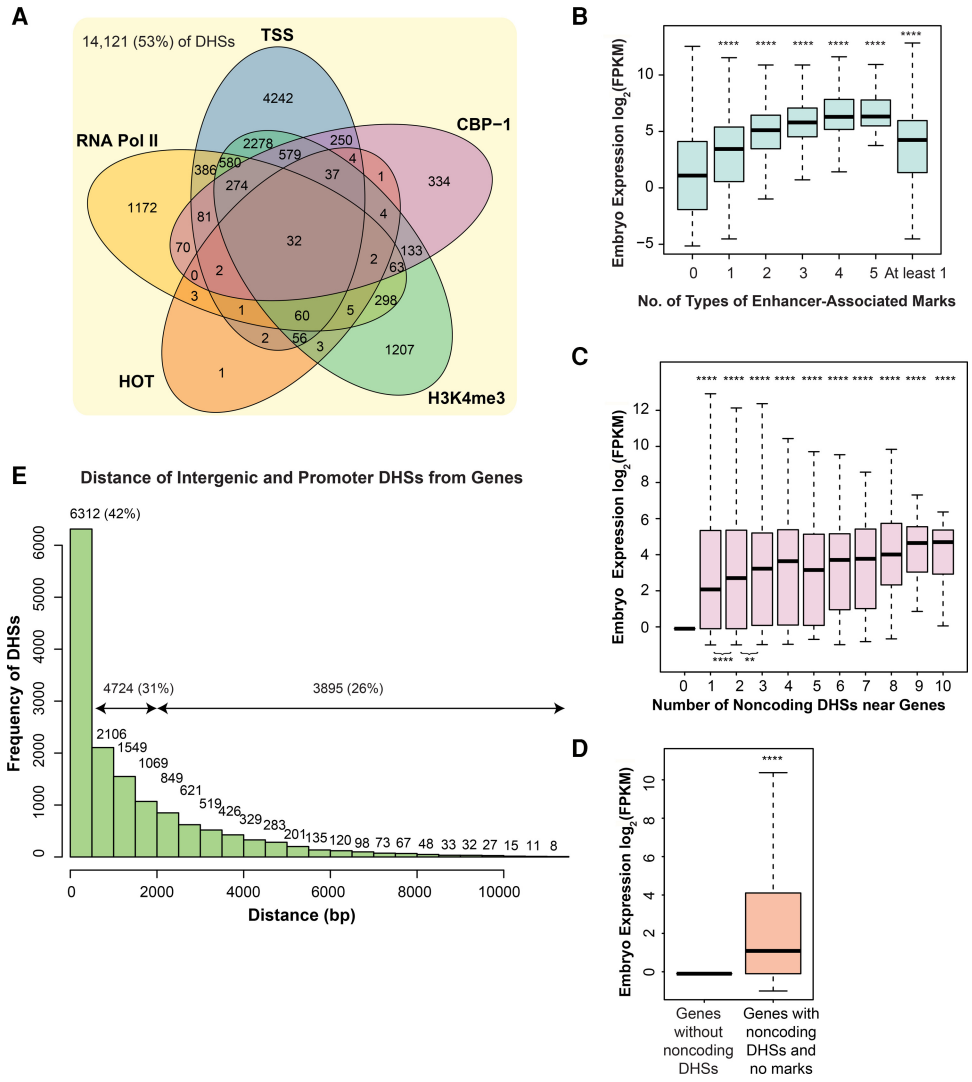


Figure 2. Presence of embryo noncoding DHSs near genes is associated with higher embryonic expression. (*A*) Half of embryo noncoding DHSs overlap with TSS, histone marks, CBP-1, and HOT regions. 47% of noncoding DHSs with marks of enhancer activity such as RNA Polymerase II (RNA Pol II; yellow), transcription start site (TSS; blue), CBP-1 (pink), H3K4me3 (green) observed in embryos, and modENCODE high occupancy TF regions (HOT; orange). TSS data are from Chen et al. (2013), and remaining data are from modENCODE (Gerstein et al. 2010). (*B*) Genes with noncoding DHSs harboring enhancer-associated marks are more highly expressed than those lacking any marks. Genes near embryo noncoding DHSs with any number of marks (at least one, two, three, four, and five type[s] of enhancer-associated mark) exhibit, on average, 8.9-fold higher levels of embryo expression (measured in $\log_2$ of FPKM) (data from Gerstein et al. 2010) compared with those with embryo noncoding DHSs lacking marks ($P < 3 \times 10^{-16}$). Genes with just one mark have 5.1-fold higher expression than genes without marks ($P < 2 \times 10^{-16}$). With each additional mark, median observed expression increases, up to three marks (5.1-fold higher expression compared with one mark, $P < 3 \times 10^{-14}$). No significant difference is observed between genes near noncoding DHSs with three, four, or five types of marks. (*C*) The presence of at least one embryo noncoding DHS near a gene is correlated with 4.5-fold higher embryo expression. The presence of at least one embryo noncoding DHS near a gene is associated with 4.5-fold higher embryo expression compared with genes without any DHSs ($P < 3 \times 10^{-16}$). Embryo expression (measured as $\log_2$ of FPKM) (data from Zhong et al. 2010) increases 54% from one to two embryo noncoding DHSs ($P < 3 \times 10^{-6}$) and 44% from two to three ($P < 0.007$). Further increases in DHS number are not correlated with increased expression. (*D*) Genes associated with embryo noncoding DHSs and lacking marks are still twice as highly expressed as genes without DHS. Genes with embryo noncoding DHSs lacking enhancer-associated marks (orange) show 2.3-fold higher embryo expression compared with genes lacking any DHSs (blue; $P < 3 \times 10^{-16}$). (*E*) Additional evidence for distant CRMs. Over half (56%) of intergenic and promoter DHSs are found within 1 kb of the nearest gene, and most (74%) are within 2 kb. However, a quarter (26%) of intergenic and promoter DHSs are >2 kb away and 10% are >4 kb away.

et al. 2016). Previous studies showed that ELT-2 is auto-regulated by binding to its own promoter in embryos (Fukushige et al. 1999). Three CRMs—CR I, CR II, and CR III—regulate *elt-2* expression (Wiesenfahrt et al. 2016). We observe noncoding DHSs overlapping CR I and CR III, the two that were previously shown to drive reporter gene expression the most strongly. We also observe two TF footprints in CR III that closely correspond to ELT-2 TGATAA binding sites (Fig. 3B; McGhee et al. 2009).

Regulation of *myo-2* expression by its A, B, and C CRMs 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 subelements that drive pharyngeal expression in reporter assays (Supplemental Fig. S3A). In particular, we detect a putative

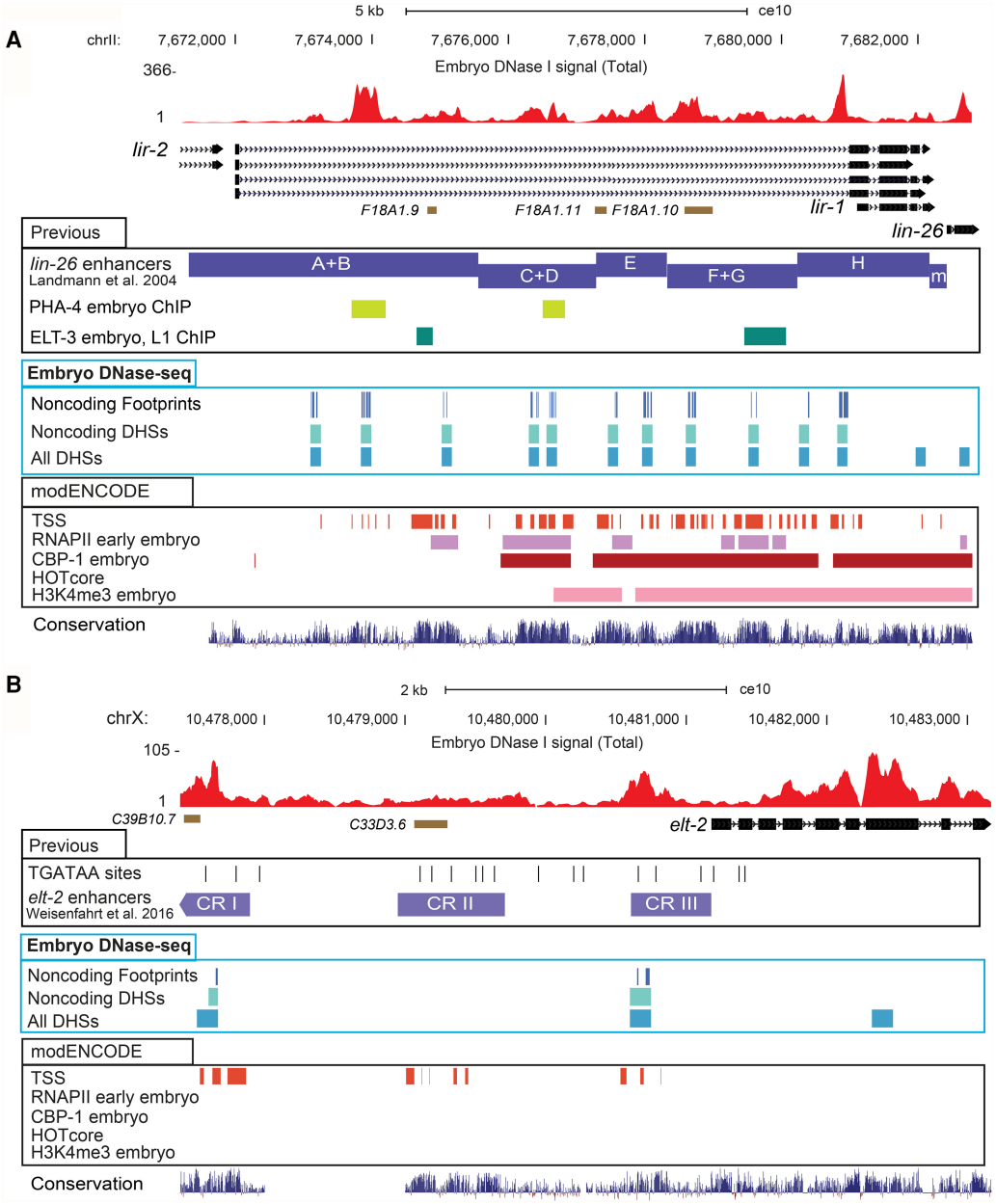


Figure 3. Noncoding DHSs recapitulate many known CRMs. Total DNase signal (red) from both strands of embryo read data shown. Noncoding DHSs (light blue boxes) and all DHSs (medium blue boxes) and TF footprints (dark blue boxes) detected. Additional tracks shown are *C. elegans* RefSeq genes (black boxes with arrows), noncoding transcripts (brown boxes), and phyloP conservation (very dark blue). Other comparison tracks include TSS (dark orange boxes) (Chen et al. 2013), RNAP II ChIP-seq (red boxes), H3K4me3 (pink), and CBP-1 (lavender boxes) ChIP-chip from modENCODE embryo data (Gerstein et al. 2010). (*A*) All five known enhancers of *lin-26* in the 11-kb first intron of *lir-1* are recovered, each harboring at least one embryo noncoding DHS and footprint. Multiple noncoding DHSs are detected upstream of *lin-26*, in the first intron of *lir-1*, which harbors known CRMs active in embryos: A+B, C+D, E, F+G, and H (purple boxes) (Landmann et al. 2004). Noncoding DHSs and footprints are detected in each known CRM, and in the case of A+B and C+D, the noncoding DHSs and footprints overlap with PHA-4 ChIP-seq peaks (light green) (Zhong et al. 2010). One of the ELT-3 binding sites in F+G is overlapped with a noncoding DHS and footprint (dark green) (Gerstein et al. 2010). (*B*) Noncoding DHSs overlap two known *elt-2* CRMs. Noncoding DHSs with TF footprints are detected upstream of *elt-2* in two (CR I and CR III) of the three known *elt-2* CRMs (purple boxes) (Wiesenfahrt et al. 2016). One footprint in the noncoding DHS overlapping CR III contains ELT-2 binding sites (TGATAA motifs; black).

TF footprint in the subelement C that binds PHA-4 (Okkema and Fire 1994; Kalb et al. 1998; 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—known to drive reporter expression (Okkema et al. 1993). Noncoding DHSs coinciding with these enhancers possess several TF footprints (Supplemental Fig. S3B).

Embryo noncoding DHSs 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). Noncoding DHSs and TF footprints at this locus overlap with the *enh1* and *enh2* CRMs known to drive expression in BWM precursor blastomeres (Supplemental Fig. S3C). However, the 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 DHSs. This discrepancy may be partly due to weak and broad DNase 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 DHSs we observe in the *C. elegans* embryo may represent not only enhancers but also potential sites of negative regulation, we examined the intergenic region between *col-43* dauer collagen and *sth-1*, which is expressed in the spermatheca. Two homeodomain proteins, MAB-18 and CEH-14, prevent *col-43* activation by the adjacent regulatory sequences of *sth-1* (Bando et al. 2005) and are expressed in early embryos (Chisholm and Horvitz 1995; Kagoshima et al. 2013). We observed one embryo noncoding DHS with a TF footprint overlapping the HB1 binding site for MAB-18 and CEH-14 that is part of the spermathecal enhancer (Bando et al. 2005). Another noncoding DHS harboring a TF footprint overlaps the HB2 binding site for MAB-18 alone and an embryo TSS (Supplemental Fig. S3D; Chen et al. 2013).

Even within the well-studied gene loci we investigated, we detected several novel putative CRMs. Some of these predictions include footprints and noncoding DHSs observed in the sixth and 10th introns of *myo-2* overlapping PHA-4 ChIP binding sites. Since PHA-4 is a transcriptional regulator of pharynx expression and *myo-2*, these noncoding DHSs may represent additional PHA-4–regulated enhancers of *myo-2* (Supplemental Fig. S3A). We also observed a noncoding DHS in the first intron of *hlh-1* corresponding to a region bound by PHA-4 in embryos (Zhong et al. 2010). *hlh-1* could be repressed by PHA-4 in the pharynx through this putative CRM (Supplemental Fig. S3C).

Our data also provide additional evidence for distant-acting CRMs in *C. elegans*. Nearly half of the intergenic and promoter DHSs detected in the embryo are situated <500 bp to the nearest gene (Fig. 2E). However, one third of them are between 500 bp and 2 kb from the nearest gene, and a quarter are >2 kb away, as far as 11 kb. These noncoding DHSs have a similar number of footprints and are detected with similar normalized read coverage (Supplemental Fig. S9).

Discriminative motif discovery within noncoding DHS peaks identifies known and novel regulatory motifs

We performed discriminative motif discovery to identify overrepresented motifs within noncoding DHS peaks and putative TF footprints using DREME (Bailey 2011), surmising that these might represent TFBSs. These matched many known *C. elegans* regulatory motifs, including Kozak, TATA-box, SP1, and T-block promoter motifs (Grishkevich et al. 2011), as well as many known TF motifs such as PHA-4 and ELT-2 (Supplemental Fig. S6; Gaudet et al. 2004; McGhee et al. 2009). Additional motif analyses are described in the Supplemental Material.

We measured the pattern of DNase cleavage accessibility across known *cis*-regulatory motifs (Supplemental Fig. S6). When we mapped average DNase cleavage across motif sites identified 2 kb upstream of genes, almost all showed patterns characteristic of TF footprints, with lower read coverage centering around the motif indicating DNase cleavage protection and a symmetric read shift aligning to opposite strands of the genome (Fig. 4A; Supplemental Fig. S8).

We also identified novel motifs in noncoding DHSs for which there were no known functions. Some of these matched conserved DNA motifs found by two prior studies (Elemento and Tavazoie 2005; Ihuegbu et al. 2012). We performed Gene Ontology (GO) and anatomy enrichment analysis on genes associated with these motifs to predict potential function (Supplemental Tables S4, S5).

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

We explored whether DNase-seq data could improve our ability to predict tissue-specific expression of genes regulated by known DNA motifs, such as the N1 neuronally enriched motif (Ruvinsky et al. 2007) and ELT-2 and SLR-2 intestinal TFs (McGhee et al. 2007; Kirienko and Fay 2010). We compared the percentage of genes correctly predicted to be expressed in these tissues using DNA motifs alone versus DNA motifs within noncoding DHSs (see Supplemental Methods). Prediction accuracy was improved by using noncoding DHSs together with motifs, from 41% to 55% neuronal genes using N1 and from 8% to 28% (using ELT-2) and 4% to 25% (using SLR-2) of intestinal genes (Fig. 4B; FACS-sorted embryonic expression data from Spencer et al. 2011). Smaller improvement was observed from 29% to 36% (ELT-2) and 27% to 32% (SLR-2) in adult dissected gut expression data from McGhee et al. (2007).

L1 arrest DHSs are enriched in genes up-regulated in L1 arrest

When L1 larvae hatch in the absence of food, they remain in a developmentally arrested state that is resistant to environmental stress (for review, see Baugh 2013). By comparing L1 arrest and embryo DNase-seq data, we find that most (88%) of the 16,084 noncoding DHSs found during the L1 arrest stage were also found in the embryo. However, 12% appear to be associated with L1 arrest and are not in embryos. We identified 9359 putative TF footprints in L1, with 2946 of these residing in these L1 arrest–associated DHSs. Genes with L1 arrest–associated elements have 12.5% higher expression in 6-h L1 starved larvae compared with the embryo, reflecting greater specificity in our L1 data set for genes likely to be involved in L1 arrest ($P < 1.6 \times 10^{-8}$) (Supplemental Fig. S7D; expression data from Baugh et al. 2009; Maxwell et al. 2014). While this difference in expression is not large, it is worth noting that the majority of *C. elegans* genes are transcriptionally quiescent during L1 arrest (Baugh et al. 2009). Moreover, genes in the top decile of this category are expressed at twofold higher expression than in the embryo. All DHSs and noncoding DHSs from L1 arrest larvae are enriched 1.7-fold and 2.4-fold, respectively, in PHA-4 ChIP peaks from stage-matched samples ($P < 3 \times 10^{-16}$), suggesting that our data can recapitulate CRMs for gene targets of PHA-4, a TF regulator of starvation survival in L1 arrest.

We detected many L1 arrest–associated DHSs in targets of DAF-16– and PHA-4–regulated genes and other genes differentially regulated in L1 arrest. For example, ICL-1 is a key enzyme for the breakdown of fats into carbohydrates and is a known target of the DAF-16 insulin-like signaling pathway required for L1 arrest (Murphy et al. 2003; Baugh and Sternberg 2006; Tepper et al. 2013). Expression of *icl-1* is highly up-regulated in *daf-2* mutants (Murphy et al. 2003) and in response to starvation (7.9-fold; Baugh et al. 2009) and in L1 arrest compared with embryos (1.9-fold) (Baugh et al. 2009). It also appears to be regulated by PHA-4 according to ChIP-seq data (Zhong et al. 2010). We detect one

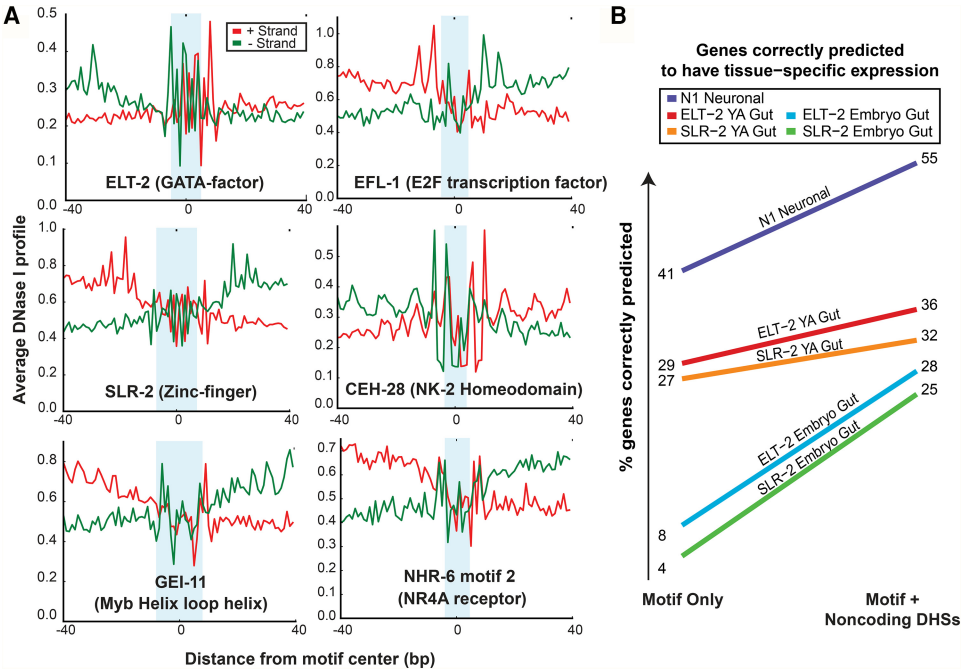


Figure 4. Using DNase-seq to refine prediction of tissue-specific expression by TF motifs. (*A*) Average DNase profile over *C. elegans* motif sites. *C. elegans* motif sites show patterns of DNase cleavage accessibility and strand-shift in reads characteristic of TF footprints. Average DNase profile is measured across 80 bp centering around the motifs (from 2 kb upstream of genes). Positive (red) and negative (green) strands are shown. Light blue shading shows the position of each motif: ELT-2, EFL-1, SLR-2, CEH-28, GEI-11, and NHR-6 motif 1. (*B*) Refining prediction of tissue-specific gene expression with noncoding DHSs. The percentage of genes correctly predicted to be expressed in the tissue expression data set from the presence of DNA motif (motif only) was compared with the presence of DNA motif within noncoding DHSs (motif + noncoding DHS). Using noncoding DHS data improves prediction accuracy of intestinal expression (embryonic FACS data from Spencer et al. 2011) from 8% to 28% (ELT-2; blue) and 4% to 25% (SLR-2; green). Similarly, using noncoding DHS data also slightly improves prediction of expression in young adult (YA) dissected intestines (data from McGhee et al. 2007) from 29% to 36% (ELT-2; red) and 27% to 32% (SLR-2; orange). Using noncoding DHS data also improves prediction of neuronal expression (Spencer et al. 2011) by N1 neuronally enriched motif from 41% to 55% (purple).

L1 arrest–associated noncoding DHS harboring TF footprints that overlap both a DAF-16 binding motif ($P < 1 \times 10^{-4}$) and PHA-4 motif ($P < 5 \times 10^{-5}$) in the first intron of *icl-1* (Fig. 5A). Three other noncoding DHSs were found near *icl-1*, coinciding with PHA-4 ChIP peaks (Zhong et al. 2010).

Another example, *pha-4*, encodes a TF that plays a role in L1 starvation survival and autoregulates its own promoter (Zhong et al. 2010). We detected multiple L1 noncoding DHSs upstream of *pha-4* coinciding with PHA-4 ChIP peaks during L1 arrest (Fig. 5B). One of these DHSs coincides with the TSS of the shortest isoform, *pha-4c*. This TSS was observed in a previous study using GRO-cap in both embryos and starved L1 larvae (Maxwell et al. 2014). Another TSS far upstream of the longest isoform *pha-4a* was in embryos but only weakly in the L1 starved larvae (Maxwell et al. 2014) and coincides with a noncoding DHS found in the embryo but not in L1 arrest.

An example of a gene whose role in L1 arrest is less well understood, but for which we found evidence supporting differential regulation, is the nuclear hormone receptor NHR-4. It is expressed in ciliated sensory amphid neurons, other neurons, intestine, and pharynx (WormBase) and is directly regulated by DAF-19 TF (Burghoorn et al 2012). Expression of *nhr-4* is up-regulated 1.5-fold in L1 arrest compared with embryos (Baugh et al. 2009). We detect four L1 noncoding DHSs upstream of *nhr-4*, two of which are specific to L1 arrest (Supplemental Fig. S3E). Of these, one overlaps an annotated TSS previously detected by GRO-cap in starved L1 (Maxwell et al. 2014). The other DHS has TF footprints that coincide with both DAF-19 and PHA-4 motifs, and is weakly bound by PHA-4 in starved L1 (Supplemental Fig. S3E; Zhong et al. 2010). The other two noncoding DHSs detected overlap PHA-4 ChIP peaks from both conditions.

Noncoding DHSs can drive gene expression in transgenic *C. elegans*

To assess whether these DHSs could function as enhancers, we tested 15 DHSs (Supplemental Table S9). We cloned these DHSs into a reporter gene construct containing a minimal promoter to test their ability to drive reporter gene expression in injected transgenic *C. elegans*. Since our base reporter gene construct, without any DHS, was found to have background expression in embryos and L1 larvae, we could not directly assay whether the DHSs (originally identified in embryos and starved L1 larvae) were active during these stages. Instead, we focused on the ability of these DHSs to drive expression in post L1 mixed larval stages and adult (Fig. 6), where the base reporter drives little expression (Supplemental Fig. S10). These later larval stages and adult stage were thus used as a proxy to test general enhancer activity.

We tested three DHSs downstream from the anterior–posterior patterning Hox gene *ceh-13* (Supplemental Fig. S3F, asterisks) that were never previously tested, and found that all three drove transgene expression. A DHS in the first intron of *ceh-13* (Supplemental Fig. S3F, labeled as A1) drove expression in the pharynx (Fig. 6A). Another in the distal 3′ region (labeled as A2) drove expression in uterus (likely the uv2 cell), and another even more distal (labeled as A3) drove expression in anterior BWM

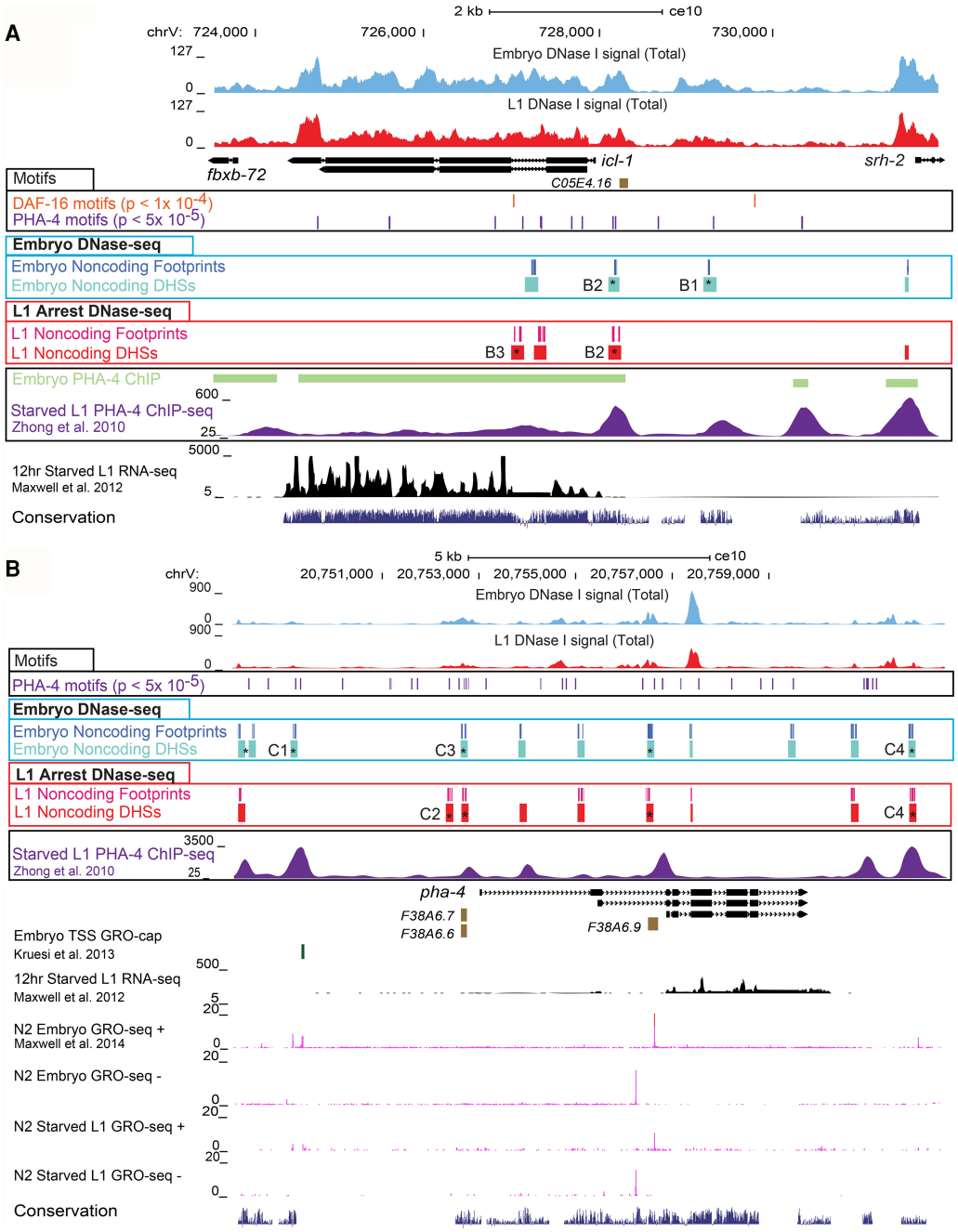


Figure 5. L1 arrest noncoding DHSs discovered in genes up-regulated during L1 arrest. Total DNase signal (red) from both strands of L1 arrest DNase-seq and embryo DNase-seq (light blue) are shown. L1 arrest noncoding DHSs (red boxes) and associated TF footprints (pink boxes), as well as embryo noncoding DHSs (light blue boxes) and associated TF footprints (dark blue boxes), were detected. Asterisks indicate DHSs that were tested for activity in transgenic *C. elegans* (Supplemental Table S9). Additional tracks shown are *C. elegans* RefSeq genes (black boxes with arrows), noncoding transcripts (brown boxes), and 12 h starved L1 mRNA-seq tracks (black) from Maxwell et al. (2012), phyloP sequence conservation (dark blue) are also shown. Other comparison tracks include PHA-4 ChIP-seq from embryo (light green) and starved L1 larvae (purple) (Zhong et al. 2010). PHA-4 (purple boxes) and DAF-16 (orange boxes) motifs, as well as TSS (L1 starved GRO-cap data from Kruesi et al. 2013 as dark green boxes; and L1 and embryo GRO-seq data from Maxwell et al. 2014 as magenta signal), are shown when relevant. (*A*) Noncoding DHSs of *icl-1*. One L1 arrest–associated noncoding DHS containing a DAF-16 binding motif ($P < 1 \times 10^{-4}$ threshold) and a PHA-4 motif ($P < 5 \times 10^{-5}$) is detected in the first intron of *icl-1*. TFs footprints are found within this DHS that overlap the DAF-16 motif. Three other noncoding DHSs are detected in both L1 and embryo coinciding with PHA-4 ChIP-seq peaks from 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. (*B*) Known and novel CRMs of *pha-4*. Four embryo and three L1 arrest noncoding DHSs are observed upstream of the longest transcript, *pha-4a*. One of these is an embryo-associated noncoding DHS overlapping an TSS that was detected in embryos but not L1 arrest by GRO-cap (data from Kruesi et al. 2013). Directly upstream of *pha-4a* is an L1 arrest–associated noncoding DHS that overlaps PHA-4 TF binding sites. The two noncoding DHSs upstream of C1 were tested in one transgenic construct, but unlike other DHSs tested in the locus, it did not drive expression.

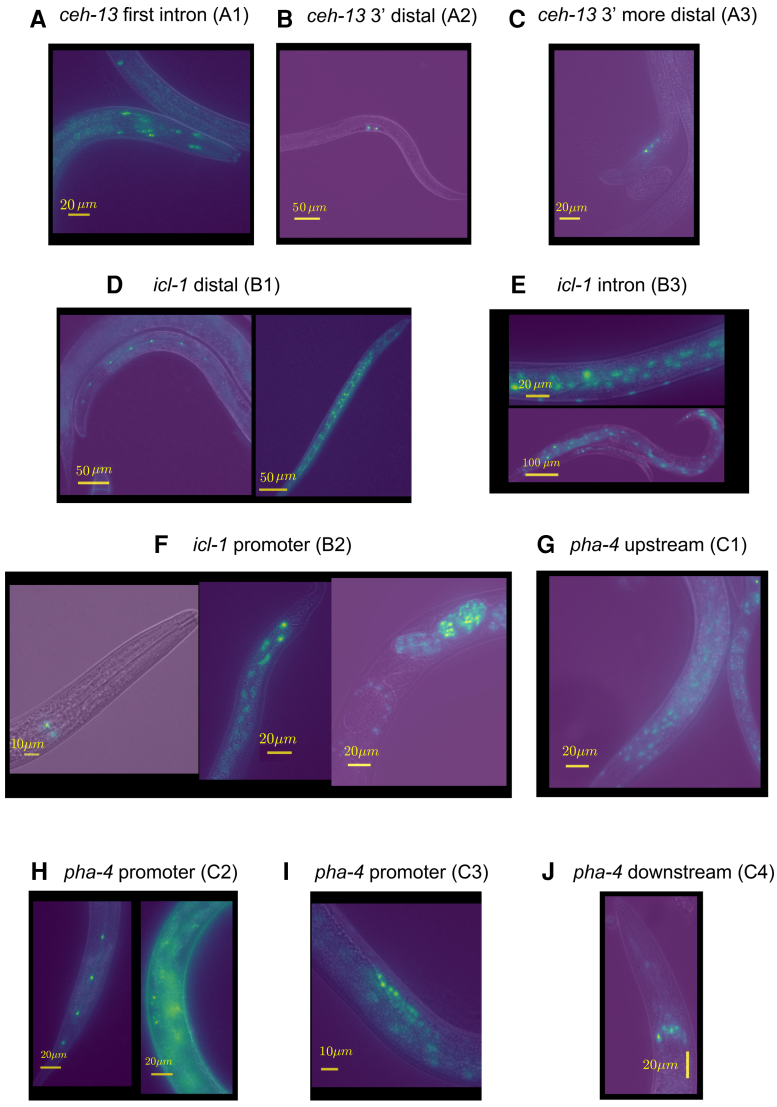


Figure 6. Noncoding DHSs tested from *ceh-13*, *icl-1*, and *pha-4* genes drive expression in transgenic *C. elegans* larvae and adults. Noncoding DHSs drive expression in diverse tissues of transgenic *C. elegans*. (*A*) *ceh-13* first intron (A1) expresses in pharyngeal cells. (*B*) *ceh-13* 3′ distal (A2) expresses in uterus (likely uv2). (*C*) *ceh-13* 3′ more distal (A3) expresses in anterior body wall muscle. (*D*) *icl-1* distal (B1) expresses in seam cells and hypodermis. (*E*) *icl-1* intron (B3) expresses in intestine, body wall muscle, and hypodermis. (*F*) *icl-1* promoter (B2) expresses in cells around the pharynx and intestine. (*G*) *pha-4* upstream (C1) expresses in hypodermis. (*H*) *pha-4* promoter (C2) expresses in uterus and vulva (likely vulC). In some animals, expression is observed in seam cells. (*I*) *pha-4* promoter (C3) expresses in somatic gonad. (*J*) *pha-4* downstream (C4) expresses in cells around the pharynx.

(Fig. 6A; detailed Nomarski in Supplemental Fig. S11). CEH-13 expression has been reported in multiple tissues, including anterior BWM (in antibody staining by Brunschwig et al. 1999) and in adult vulva (WBCnstr00012787, WormBase) from GFP fusion reporters, similar to our findings.

Expression of ICL-1 has been reported in hypodermis, intestine, and pharynx (Erkut et al. 2016) as well as BWM (Liu et al. 1995; Mikoláš et al. 2013). We tested four DHSs from these loci (Fig. 5A, asterisks) and found that one DHS distal to *icl-1* (labeled as B1) drives expression in hypodermis and seam cells, another within the *icl-1* promoter (labeled as B2) drives expression in intestine and pharynx and another in the *icl-1* intron (B3) expresses in the intestine, BWM, and hypodermis (Fig. 6B).

Previous studies (Horner et al. 1998; Kalb et al. 1998; Chen and Riddle 2008) have demonstrated a key role for PHA-4 in pharynx, vulva, and somatic gonad development. A 30-kb fosmid (Zhong et al. 2010) of the entire *pha-4* locus fused to GFP showed that PHA-4 is expressed in these tissues. We tested six DHSs from *pha-4* (Fig. 5B, asterisks), and found that four drove expression (Fig. 6C; detailed Nomarski images in Supplemental Fig. S11) consistent with these tissues: A DHS upstream of *pha-4* (labeled C1) drove expression in hypodermis; an L1 arrest–associated DHS in the promoter (C2) drives expression in uterus and vulva (likely vulC); another DHS in the promoter (C3) drives expression in somatic gonad; and, last, a DHS downstream from *pha-4* (C4) drives expression in the pharynx.

Discussion

We have identified 26,644 embryo noncoding DHSs harboring 55,890 TF footprints and 15,841 L1 arrest–associated noncoding CRMs harboring 32,685 TF footprints, through a genome-wide systematic study of CRMs and TF binding in *C. elegans*. We were able to profile *cis*-regulatory sites without the need to specify particular prior TFs of interest and by using chromatin accessibility. We identified many known enhancers and TF footprints of *C. elegans* genes, including *hlh-1*, *myo-2*, *myo-3*, *elt-2*, and *lir-1/lin-26*. Our data recapitulated 15 of 17 known enhancers within these loci and, in many cases, refined the boundaries of many enhancers originally found by transgenic reporter assays or detected through relatively broad ChIP-seq peaks. The DNase peaks identified are ~150 bp and will be useful to define boundaries of many CRMs. Our data predict many novel CRMs and TF footprints. We found noncoding DHSs downstream from *ceh-13*, which when tested drove transgenic reporter gene expression. In another case, our DHSs recovered known CRMS in the *col-43/sth-1* locus, suggesting that we can also detect some silencer CRMs. We also detected a DHS coinciding with known PHA-4 ChIP-seq binding (from Zhong et al. 2010) near *hlh-1,* which may represent a region where PHA-4 binds and acts to repress *hlh-1* expression in the pharynx, similar to its role in repressing *lin-26* in the pharynx. It did not drive reporter gene expression in our enhancer assay, suggesting this DHS does not act as an enhancer in the stages/conditions we observed. With these results, it is important to keep in mind at least three possible reasons for why some identified DHSs would not drive transgene expression. First, the DHS may be a silencer of gene expression (as might be the case of this DHS). Second, the DHS might

need to work in combination with other CRMs to drive expression. Third, the DHS may be a false positive. These DNase-seq data are resolved enough to identify protection from DNase cleavage in noncoding DHSs and across sites within them that appear to be bound by TFs (Fig. 4A; Supplemental Fig. S8). Most of the embryo (82%) and L1 arrest noncoding DHSs (84%) were found to harbor TF footprints.

It has been common practice in *C. elegans* to regard sequence immediately 5′ of TSS as sufficient to drive endogenous expression (Dupuy et al. 2007). However, there are many documented cases (for review, see Gaudet and McGhee 2010) in which gene regulation in *C. elegans* has proven relatively complex, being regulated from intronic, 3′, or distant 5′ sequences. Another study showed that while most (62%) expression patterns from *C. elegans* transcript and translation fusion reporter expressions agreed with one another, in many cases expression was observed in additional cells or in restricted patterns, suggesting that other CRMs were involved (Murray et al. 2012). While we observed that most (74%) promoter and intergenic DHSs are within 2 kb from the nearest protein-coding gene, a quarter are >2 kb, and a tenth are >4 kb away (Fig. 2E). Although difficult to definitively assign CRMs to their target genes, we observe that even the nearest gene to a noncoding DHS can be far away. Furthermore, most (53%) protein-coding genes have at least one noncoding DHS in the embryo, and of these, some (17%) have complex regulation, with more than four noncoding DHSs (Supplemental Fig. S5B). Our study thus provides supporting evidence that some *C. elegans* genes have complex regulation and may be controlled by relatively distant CRMs.

The numbers of noncoding DHSs that we find are on the same order of magnitude as *Drosophila* DNase-seq (roughly 20,000 noncoding DHSs per stage) with similar depths of sequencing (Thomas et al. 2011). Our finding that L1 arrest–associated noncoding DHSs are mostly (88%) shared with embryo-associated noncoding DHSs are also similar to findings from *Drosophila* showing 78% concordance of DHSs between stage 5 and 11 embryos (Thomas et al. 2011). Of note, we detect 36% of embryo DHSs in *C. elegans* exons, whereas the study in *Drosophila* detects ~22% of their DHSs in exons. This difference is likely partly due to the slightly higher (29%) exonic content of the *C. elegans* genome compared with *Drosophila*, which is ~20%. If we consider all DHSs from both stages together, DHSs in exons represent 22% of the total, close to the level of exonic content of the *C. elegans* genome. Another possible reason for a higher percentage of exons being DNase hypersensitive in embryos compared with L1 arrest is the high levels of transcription during this developmental stage. L1 arrest has been shown to be comparatively transcriptionally quiescent (Maxwell et al. 2014).

Noncoding DHSs detected in L1 arrest are near genes that are on average increased in expression in the L1 arrest compared to the embryo. Of these, some of the most highly up-regulated are targets that appear to be physiologically relevant to the L1 arrest stage (Supplemental Table S6). For example, *daf-7* (6.5-fold) and *daf-5* (over twofold) are TGF-beta receptors important for signaling cues from the external environment to alter development and behavior, including dauer formation, fat metabolism, and feeding. Even among genes that are not as highly up-regulated, such as *hosl-1* (50%), a lipase that regulates energy homeostasis and fat metabolism, one can identify genes that may play a physiological role in L1 arrest (a condition that is very responsive to growth and nutrient signals). L1 arrest–associated noncoding DHSs were compared with developmental arrest and starvation TFs DAF-16 and PHA-4 target genes from previous studies by Tepper et al. (2013) and Zhong et al. (2010; Supplemental Tables S7, S8) to identify putative CRMs that could regulate other physiologically relevant genes in this stage.

An important caveat to our study is the difficulty in estimating the cellular resolution of DNase-seq data generated from entire embryos or L1 arrest larvae. Our data are likely composed of an average of DNase hypersensitivity profiles of different tissues. The ability of DNase-seq to sensitively capture DHSs likely depends on the number of cells and tissues in which the DHS is active and the level of accessibility of the DHS itself. Comparison of our data with level of gene expression indeed suggests that highly expressed genes indeed possess more DHSs. The large number of cells present in the L1 may partially explain our lower numbers (around 16,000) of noncoding DHSs detected in L1 arrest larvae compared with embryos (around 26,000), since there is more cell heterogeneity and since DNase-seq signal coming from any particular cell is likely to be more diluted. We were, however, able to recover overrepresented motifs in DHSs representing binding sites of TF regulators of the three most abundant tissues in *C. elegans*—muscle, neuronal, and intestine (Supplemental Fig. S5C)—as well as motifs that occur in smaller number of tissues (Supplemental Fig. S6). Although DNase data can refine and improve the prediction of tissue-specific genes by focusing on N1 (neuronally enriched) and ELT-2 and SLR-2 (intestinally enriched) DNA motifs present within noncoding DHSs in embryos, the lack of tissue specificity in our data is an important limitation to recognize. We have thus evaluated our noncoding DHSs 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. Our noncoding DHS data set also likely misses those CRMs that are not active in sufficient cells or tissues to be detected by DNase-seq in the whole embryo and starved L1 larvae. To probe gene activity within a small number of specific cell types, we expect that it will become more feasible in the future to isolate specific tissues and use a similar technique, ATAC-seq, which can work with smaller amounts of starting material and thus provide better sensitivity to CRMs with more restricted activity.

These data representing DNase-seq maps of DHSs and TF footprints will be useful for exploring genome-wide regulation of genes active in the embryo and L1 arrest larvae and discovering novel regulatory factors and their potential sites of action. Putative CRMs and TF binding data from this study are available through WormBase and the NCBI Gene Expression Omnibus. We demonstrated the usefulness of this resource by testing 15 noncoding DHSs and found that 10 could drive reporter gene expression in transgenic *C. elegans* in distinct spatiotemporal patterns (Fig. 6; Supplemental Fig. S11; Supplemental Table S9). Future experiments will be required to validate the functional activity of more noncoding DHSs and investigate the role of specific TF footprints in controlling CRM activity. Our DHS data set will provide a rich resource for the identification of functional CRMs in *C. elegans*. Looking ahead, DNase-seq and similar de novo techniques such as ATAC-seq may also be useful for application to other nematode species whose genomes and transcriptomes are known but whose regulation has not yet been explored.

Methods

C. elegans culture and nuclei isolation

C. elegans wild-type N2 worms were synchronized and grown in liquid culture (10 worms/µL and 20 mg/mL *E. coli* HB101 in S-basal complete media) for two generations. Adults were bleached to

obtain embryos around the 40-cell stage. Bleached embryos were hatched in S-basal complete media lacking any food, and L1 arrest larvae were collected after 10 h. Nuclei were isolated using standard methods (Steiner and Henikoff 2015).

DNase I treatment, DNA purification, size-selection, and sequencing

Nuclei were treated with 0, 20, 40, 80, 120, and 160 U/mL DNase I following the Stamatoyannopoulos laboratory protocol (Thurman et al. 2012). DNase I treatment was quenched using 20 mg/mL Proteinase K and incubated at 55°C overnight. After treatment with 45 µg/mL boiled RNase A for 30 min, DNA was cleaned using column purification, followed by gel extraction of DNA fragments <500 bp. DNA yield was measured using a Qubit fluorometer. QPCR analysis methods used prior to preparation of sequencing libraries are described in the Supplemental Methods. Prepared libraries were multiplexed sequenced on Illumina HiSeq to yield 50-bp single end reads.

Testing DHSs for the ability to drive reporter gene expression in transgenic *C. elegans*

DHS sequences from *ceh-13*, *icl-1*, and *pha-4* loci (Supplemental Table S9) were PCR amplified from N2 genomic DNA and cloned upstream of a Δ*pes-10* minimal promoter driving expression of a nuclear-localized GFP with a *let-858* 3′ UTR (from L4040 vector in the Fire Vector Kit, gift of Andrew Fire, Addgene Kit no. 1000000001) and tested by transgenic microinjection in *unc-119 (ed3)* mutant animals (for details, see Supplemental Methods).

Computational analysis

Sequencing reads were analyzed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) and aligned to the ce10 version of the *C. elegans* genome using Bowtie 1.0.0 (for details, see Supplemental Methods) (Langmead et al. 2009). DNase hypersensitive sites were identified from aligned reads using HOTSPOT peak caller (John et al. 2011), followed by the IDR framework (Li et al. 2011; The ENCODE Project Consortium 2012). TF footprints were identified with DNase2TF (Sung et al. 2014). Gene annotation with WormBase WS241, statistics, and analysis was accomplished using custom scripts written in Python (Python Core Team 2010, Version 2.7), Ruby (Matsumoto 2013, Version 2.0), R (R Core Team 2014, Version 3.1), and Bash (Free Software Foundation 2013, Version 4.3), in conjunction with BEDTools (Quinlan and Hall 2010), BEDOPS (Neph et al. 2012), and pybedtools (Dale et al. 2011) (Supplemental Scripts). Additional detailed computational methods are provided in the Supplemental Methods.

Data access

The DNase-seq data from this study have been submitted to the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE97425. Additional analysis files, including DNase I signal data, all DHSs and nearest genes, putative TF footprints, novel motifs, motif-associated genes, and GO analysis, are available from WormBase (detailed list of files in Supplemental Table S3).

Acknowledgments

We thank WormBase, especially Xiaodong Wang and James Done. This research was supported by NIH (National Institute of General Medical Sciences) grant GM084389 to P.W.S. and the Howard Hughes Medical Institute (047101), with which P.W.S. is an investigator. M.C.W.H. was supported by a National Science Foundation GRFP predoctoral fellowship. We would like to thank Ali Mortazavi, Igor Antoshechkin, 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 thank Mark Wu, Mihoko Kato, and Hillel Schwartz for helpful suggestions on the manuscript.

References

Bailey TL. 2011. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27:** 1653–1659.

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:** 101–112.

Baugh LR. 2013. To grow or not to grow: nutritional control of development during *Caenorhabditis elegans* L1 arrest. *Genetics* **194:** 539–555.

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:** 780–785.

Baugh LR, Demodena J, Sternberg PW. 2009. RNA Pol II accumulates at promoters of growth genes during developmental arrest. *Science* **324:** 92–94.

Boulin T, Hobert O. 2012. From genes to function: the *C. elegans* genetic toolbox. *Wiley Interdiscip Rev Dev Biol* **1:** 114–137.

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:** 456–464.

Brunschwig K, Wittmann C, Schnabel R, Bürglin TR, Tobler H, Müller F. 1999. Anterior organization of the *Caenorhabditis elegans* embryo by the labial-like Hox gene ceh-13. *Development* **126:** 1537–1546.

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:** 415–426.

Chen D, Riddle DL. 2008. Function of the PHA-4/FOXA transcription factor during *C. elegans* post-embryonic development. *BMC Dev Biol* **8:** 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:** 1339–1347.

Chisholm AD, Hardin J. 2005. Epidermal morphogenesis. *WormBook* **1:** 1–22.

Chisholm AD, Horvitz HR. 1995. Patterning of the *Caenorhabditis elegans* head region by the *Pax-6* family member *vab-3*. *Nature* **377:** 52–55.

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 Biotechnol* **25:** 663–668.

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.

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

Erkut C, Gade VR, Laxman S, Kurzchalia TV. 2016. The glyoxylate shunt is essential for desiccation tolerance in *C. elegans* and budding yeast. *eLife* **5:** e13614.

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

Free Software Foundation. 2013. Bash [Unix shell program]. https://www.gnu.org/software/bash/.

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:** 11883–11888.

Gaudet J, Mango SE. 2002. Regulation of organogenesis by the *Caenorhabditis elegans* FoxA protein PHA-4. *Science* **295:** 821–825.
Gaudet J, McGhee JD. 2010. Recent advances in understanding the molecular mechanisms regulating *C. elegans* transcription. *Dev Dyn* **239:** 1388–1404.
Gaudet J, Muttumu S, Horner M, Mango SE. 2004. Whole-genome analysis of temporal gene expression during foregut development. *PLoS Biol* **2:** e352.
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:** 1775–1787.
Grishkevich V, Hashimshony T, Yanai I. 2011. Core promoter T-blocks correlate with gene expression levels in *C. elegans*. *Genome Res* **21:** 707–717.
Gross DS, Garrard WT. 1988. Nuclease hypersensitive sites in chromatin. *Annu Rev Biochem* **57:** 159–197.
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–D793.
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:** 311–318.
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:** 283–289.
Horner MA, Quintin S, Domeier ME, Kimble JE, Labouesse M, Mango SE. 1998. *pha-4*, an *HNF-3* homolog, specifies pharyngeal organ identity in *Caenorhabditis elegans*. *Genes Dev* **12:** 1947–1952.
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:** 314–323.
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-3α,β,γ homolog that functions in organogenesis of the *C. elegans* pharynx. *Development* **125:** 2171–2180.
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:** 611–624.
Kirienko NV, Fay DS. 2010. SLR-2 and JMJC-1 regulate an evolutionarily conserved stress-response network. *EMBO J* **29:** 727–739.
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:** 133–148.
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:** 1955–1968.
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:** 478–490.
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.
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:** 1241–1249.
Li Q, Brown JB, Huang H, Bickel PJ. 2011. Measuring reproducibility of high-throughput experiments. *Ann Appl Stat* **5:** 1752–1779.
Liu F, Thatcher JD, Barral JM, Epstein HF. 1995. Bifunctional glyoxylate cycle protein of *Caenorhabditis elegans*: a developmentally regulated protein of intestine and muscle. *Dev Biol* **169:** 399–414.
Matsumoto Y. 2013. *Ruby programming language*. http://www.ruby-lang.org/.
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:** 1920–1929.
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:** 455–466.
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:** 627–645.
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:** 551–565.
Mercer TR, Edwards SL, Clark MB, Neph SJ, Wang H, Stergachis AB, John S, Sandstrom R, Li G, Sandhu KS, et al. 2013. DNase I–hypersensitive exons colocalize with promoters and distal regulatory elements. *Nat Genet* **45:** 852–859.
Mikoláš P, Kollárová J, Sebková K, Saudek V, Yilma P, Kostrouchová M, Krause MW, Kostrouch Z, Kostrouchová M. 2013. GEI-8, a homologue of vertebrate nuclear receptor corepressor NCoR/SMRT, regulates gonad development and neuronal functions in *Caenorhabditis elegans*. *PLoS One* **8:** e58462.
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:** 277–283.
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:** 1282–1294.
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:** 1919–1920.
Noonan JP, McCallion AS. 2010. Genomics of long-range regulatory elements. *Annu Rev Genomics Hum Genet* **11:** 1–23.
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, Harrison SW, Plunger V, Aryana A, Fire A. 1993. Sequence requirements for myosin gene expression and regulation in *Caenorhabditis elegans*. *Genetics* **135:** 385–404.
Python Core Team. 2010. *Python: a dynamic, open source programming language*. Python Software Foundation. https://www.python.org/.
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842.
R Core Team. 2014. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.
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:** 2306–2309.
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:** 617–626.
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:** 325–341.
Steiner FA, Henikoff S. 2015. Cell type-specific affinity purification of nuclei for chromatin profiling in whole animals. *Methods Mol Biol* **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.
Sung MH, Guertin MJ, Baek S, Hager GL. 2014. DNase footprint signatures are dictated by factor dynamics and DNA sequence. *Mol Cell* **56:** 275–285.
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:** 676–690.
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:** 75–82.
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:** 854–858.
Wiesenfahrt T, Berg JY, Osborne Nishimura E, Robinson AG, Goszczynski B, Lieb JD, McGhee JD. 2016. The function and regulation of the GATA factor ELT-2 in the *C. elegans* endoderm. *Development* **143:** 483–491.
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 Genet* **6:** e1000848.

Received April 6, 2017; accepted in revised form October 18, 2017.