Chromatin accessibility dynamics reveal novel functional enhancers in *C. elegans*

Aaron C. Daugherty[1,7], Robin Yeo[1,7], Jason D. Buenrostro[2,3], William J. Greenleaf[1,4], Anshul Kundaje[1,5], Anne Brunet[1,6*]

[1] Department of Genetics, Stanford University, Stanford CA 94305, USA

[2] Broad Institute of MIT and Harvard, Harvard University, Cambridge, MA 02142, USA

[3] Harvard Society of Fellows, Harvard University, Cambridge, MA 02138, USA

[4] Department of Applied Physics, Stanford University, Stanford, CA 94305, USA

[5] Department of Computer Science, Stanford University, Stanford, CA 94305, USA

[6] Glenn Laboratories for the Biology of Aging, Stanford University, Stanford CA 94305, USA

[7] These authors contributed equally to this work

* Corresponding author. Email: anne.brunet@stanford.edu

Running title: Chromatin accessibility in *C. elegans*

Key words: Chromatin accessibility, development, enhancers, *C. elegans*

Abstract

Chromatin accessibility, a crucial component of genome regulation, has primarily been studied in homogenous and simple systems, such as isolated cell populations or early development models. Whether chromatin accessibility can be assessed in complex, dynamic systems *in vivo* with high sensitivity remains largely unexplored. In this study, we use ATAC-seq to identify chromatin accessibility changes in a whole animal, the model organism *C. elegans*, from embryogenesis to adulthood. Chromatin accessibility changes between developmental stages are highly reproducible, recapitulate histone modification changes, and reveal key regulatory aspects of the epigenomic landscape throughout organismal development. Importantly, our analysis of dynamic changes in chromatin accessibility within whole organisms sensitively identified novel cell-type- and temporal-specific enhancers, which we functionally validate *in vivo*. Furthermore, by integrating transcription factor binding motifs into a machine learning framework, we identify EOR-1 as a potential early regulator of chromatin accessibility changes. Our study provides a unique resource for *C. elegans*, a system in which the prevalence and importance of enhancers remains poorly characterized, and demonstrates the power of using whole organism chromatin accessibility to identify novel regulatory regions in complex systems.

Introduction

Chromatin accessibility represents an essential level of genome regulation and plays a pivotal role in many biological and pathological processes, including development, tissue regeneration, aging, and cancer (Simon et al. 2014; Stergachis et al. 2013; Tsompana et al. 2014). However, most genome-wide chromatin accessibility studies to date have been in relatively simple systems, including cultured or purified cells, and early embryos (Lara-Astiaso et al. 2014; Thomas et al. 2011; Wang et al. 2012; West et al. 2014). Assessing chromatin accessibility directly in complex systems composed of multiple cell types could allow for high-throughput discovery of regulatory regions whose activity are restricted to rare or undefined sub-populations of cells; this is particularly relevant for enhancers, which are thought to be highly cell-type- and temporally-specific (Ren et al. 2015).

The primary limitation for studying chromatin accessibility in complex systems is that most assays lack the sensitivity and precision to detect regions active only in rare sub-populations of cells, or require so many cells that precise temporal synchronization of samples is impractical. However, the Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) has been shown to assess native chromatin accessibility with high sensitivity and base pair resolution while requiring orders of magnitude less starting material than other assays (Buenrostro et al. 2013). This approach has been used successfully in cultured or purified cells even down to single cells, though such low input relies heavily on existing knowledge, and has only been demonstrated in homogenous cell types (Buenrostro et al. 2015) (Cusanovich et al. 2015). Rather than purifying specific cell types, we wondered whether ATAC-seq could be sensitive enough to detect

subtle changes in chromatin accessibility in complex mixtures of tissue, and in so doing uncover novel biological insights that would have otherwise been obscured.

The nematode *Caenorhabditis elegans* is a particularly powerful model to study chromatin accessibility in a complex system and potentially identify novel regulatory regions. *C. elegans* has highly synchronous life stages, as well as consistent and well characterized cellular composition throughout each life stage of development (Sulston et al. 1983). Rapid transgenesis and transparency (Mello et al. 1995) also make *C. elegans* an ideal system to efficiently validate genomic regions of functional importance and visualize tissue- or cell-specificity (Harfe et al. 1998; Jantsch-Plunger et al. 1994; Lei et al. 2009). While there exist some preliminary reports on chromatin states in *C. elegans* (Evans et al. 2016; Gerstein et al. 2010; Hsu et al. 2015; Liu et al. 2011; Valouev et al. 2008), high-resolution, genome-wide chromatin accessibility maps throughout development have not yet been reported. In this study, we show that studying high-resolution chromatin accessibility dynamics in synchronized *C. elegans* populations allows us to characterize highly reproducible changes in chromatin accessibility between developmental stages and to identify functional temporal- and tissue-specific novel enhancers *in vivo*. Our study provides a unique resource to define *C. elegans* regulatory regions as well as a guide for the interpretation of chromatin structure in complex multi-tissue systems *in vivo*.

Results

High-resolution chromatin accessibility profiles from three *C. elegans* life stages

To sensitively measure high-resolution chromatin accessibility at different life stages in *C. elegans,* we used the Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq). The low input requirements of ATAC-seq (several orders of magnitude less than standard ChIP-seq (Furey 2012)) allowed us to generate three independent biological replicates that are tightly synchronized at three key life stages: early embryo, larval stage 3 (L3), and young adults, thereby limiting variation within stages (see Methods) (Fig. 1A). We generated then sequenced these ATAC-seq libraries, as well as an input control, to a median depth of over 17 million unique, high-quality mapping reads per sample (Supplemental Table 1). We designed a computational framework to integrate the input control and emphasize single base pair resolution (Fig. 1B). The high correlation of ATAC-seq signal between each of the 3 biological replicates (Spearman's rho > 0.837) demonstrates the high reproducibility of this approach (Fig. 1C). The ability to cluster samples by their developmental stage demonstrates that chromatin accessibility is strikingly different between these three life stages (Fig. 1C). These differences between developmental stages are likely due to both changes in accessibility within cells, as well as the organisms' changing cellular composition throughout development. Together, these results indicate that reproducible high-resolution chromatin accessibility can be obtained from low amounts (at least an order of magnitude less than standard histone ChIP-seq or DNase-seq) of complex, multi-tissue samples.

To investigate the changes in chromatin accessibility between life stages, we identified and characterized the ATAC-seq peaks that significantly changed accessibility between early embryo and larval stage 3 (L3) (12,193 peaks, Supplemental Fig. 1C) and between L3 and young adult (783 peaks, Supplemental Fig. 1D) (FDR < 0.05) (see

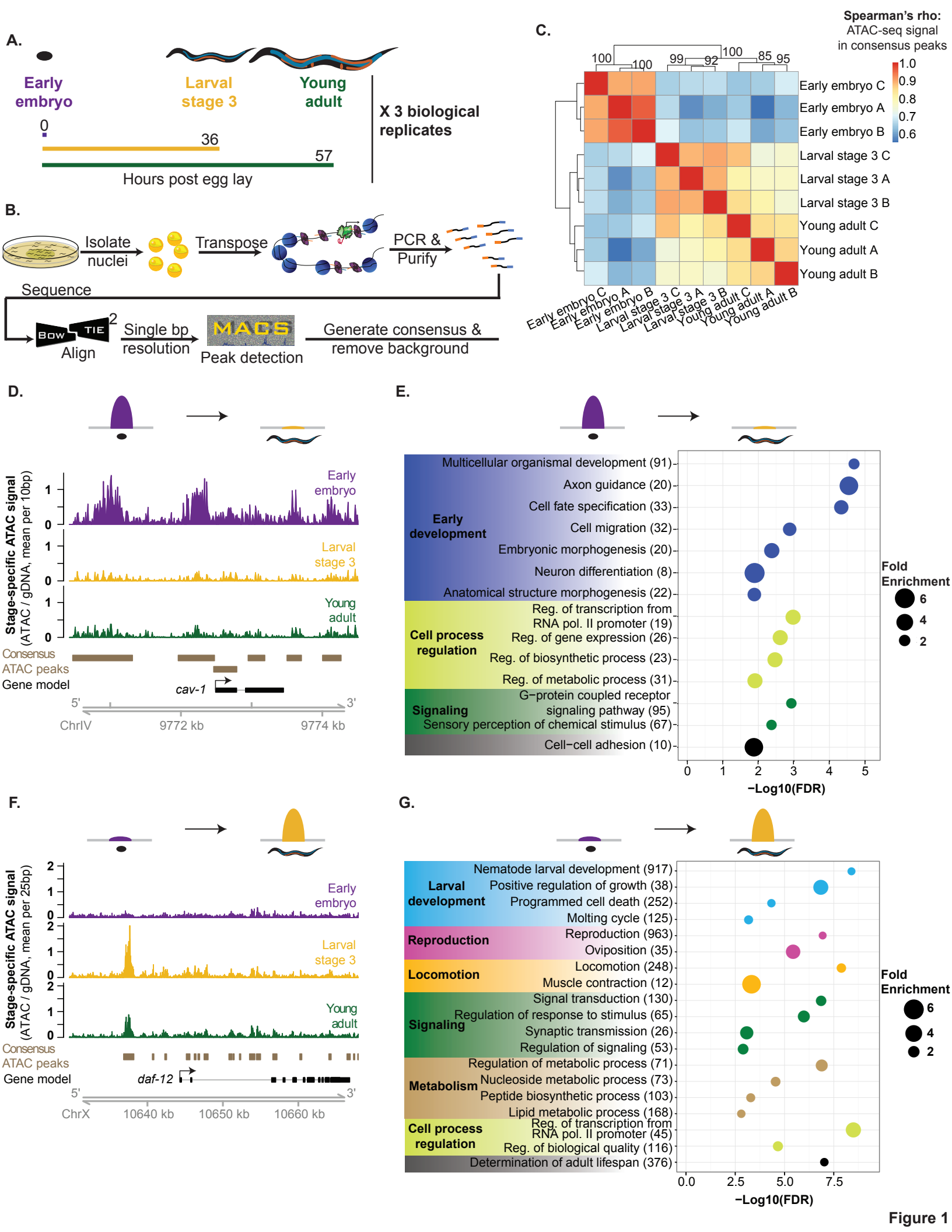
A.
Early embryo
Larval stage 3
Young adult
0
36
57
Hours post egg lay
X 3 biological replicates
B.
Isolate nuclei
Transpose
PCR & Purify
Sequence
Bow Tie 2
Align
Single bp resolution
MACS
Peak detection
Generate consensus & remove background
C.
Spearman's rho: ATAC-seq signal in consensus peaks
1.0
0.9
0.8
0.7
0.6
100
100
99
92
100
85
95
Early embryo C
Early embryo A
Early embryo B
Larval stage 3 C
Larval stage 3 A
Larval stage 3 B
Young adult C
Young adult A
Young adult B
D.
Stage-specific ATAC signal (ATAC / gDNA, mean per 10bp)
Early embryo
Larval stage 3
Young adult
Consensus ATAC peaks
Gene model
cav-1
5'
3'
ChrIV
9772 kb
9774 kb
E.
Early development
Multicellular organismal development (91)
Axon guidance (20)
Cell fate specification (33)
Cell migration (32)
Embryonic morphogenesis (20)
Neuron differentiation (8)
Anatomical structure morphogenesis (22)
Cell process regulation
Reg. of transcription from RNA pol. II promoter (19)
Reg. of gene expression (26)
Reg. of biosynthetic process (23)
Reg. of metabolic process (31)
Signaling
G−protein coupled receptor signaling pathway (95)
Sensory perception of chemical stimulus (67)
Cell−cell adhesion (10)
0
1
2
3
4
5
−Log10(FDR)
Fold Enrichment
6
4
2
F.
Stage-specific ATAC signal (ATAC / gDNA, mean per 25bp)
Early embryo
Larval stage 3
Young adult
Consensus ATAC peaks
Gene model
daf-12
5'
3'
ChrX
10640 kb
10650 kb
10660 kb
G.
Larval development
Nematode larval development (917)
Positive regulation of growth (38)
Programmed cell death (252)
Molting cycle (125)
Reproduction
Reproduction (963)
Oviposition (35)
Locomotion
Locomotion (248)
Muscle contraction (12)
Signaling
Signal transduction (130)
Regulation of response to stimulus (65)
Synaptic transmission (26)
Regulation of signaling (53)
Metabolism
Regulation of metabolic process (71)
Nucleoside metabolic process (73)
Peptide biosynthetic process (103)
Lipid metabolic process (168)
Cell process regulation
Reg. of transcription from RNA pol. II promoter (45)
Reg. of biological quality (116)
Determination of adult lifespan (376)
0.0
2.5
5.0
7.5
−Log10(FDR)
Fold Enrichment
6
4
2

Figure 1

Figure 1. ATAC-seq in whole *C. elegans* captures chromatin accessibility dynamics across three life-stages. (*A*) Three independent biological replicates each consisting of tightly temporally synchronized *C. elegans* were used for ATAC-seq; hours post egg-lay are at 20ºC. (*B*) *C. elegans* were flash frozen and nuclei were isolated before assaying accessible chromatin using transposons loaded with next-generation sequencing adaptors, allowing paired-end sequencing. A custom analysis pipeline emphasizing high-resolution signal and consistent peaks, as well as accommodating input control was developed to generate stage-specific as well as consensus (across all stages) ATAC-seq peaks. (*C*) ATAC-seq signal within consensus ATAC-seq peaks was compared between all samples using Spearman's rho to cluster samples. Replicate batches are noted as letters following the stage. (*D,E*) Comparison of ATAC-seq signal between all three stages at a region that decreases (D) or increases (E) in accessibility during development. (*F,G*) Genes that lose accessibility between embryo and larval stage 3 (L3) are enriched for early development functions (F), while genes that gain accessibility are enriched for larval-related functions (G); all calculations and genes lists are from GOrilla and the number of genes in each term are listed in parentheses.

Methods). For example, several ATAC-seq peaks decrease from embryo to L3 in the promoter region of the *cav-1* gene, which is expressed during embryogenesis but not larval development (Parker et al. 2009) (Fig. 1D). Conversely, several ATAC-seq peaks drastically increase from embryo to L3 in the promoter and regions upstream of the *daf-12* gene, which is known to be a key regulator of stage-specific developmental programs particularly at L3 (Antebi et al. 1998; Antebi et al. 2000) (Fig. 1F). Confirming these specific examples, the most enriched gene ontology (GO) terms for genes with decreased chromatin accessibility from embryo to L3 include early development terms such as embryonic morphogenesis and cell fate specification (Fig. 1E and Supplemental Table 4), while the most enriched terms for genes with increased chromatin accessibility from embryo to L3 include larval development and locomotion (Fig. 1G and Supplemental Table 5). Similarly, we observe strong enrichments of GO terms reflecting the major phenotypic changes occurring between L3 and adult, including terms like larval development and reproduction (Supplemental Fig. 1E,F). Together, these results indicate that ATAC-seq in whole organisms can identify changes in DNA accessibility that represent key biological differences between stages, regardless of whether these changes are due to activation/repression of specific regions within a cell type or to changes in cell-type composition.

ATAC-seq as a single assay describes the epigenome

Accessible chromatin encompasses several key features of the epigenome, including active and poised regulatory regions. To verify that our ATAC-seq data correctly identifies regulatory regions throughout the epigenome, we used multiple histone

modification ChIP-seq datasets and ChromHMM (Ernst et al. 2012) to build predictive models of the epigenome. ChromHMM is a Hidden Markov Model that classifies regions of the genome into chromatin states (e.g. heterochromatin) using the co-occurrence of multiple histone modifications from each life stage (in our case, ChIP-seq datasets characterizing a total of 21 histone modification) (Supplemental Table 6) (see Methods). We found that ATAC-seq peaks from all three stages are significantly enriched in active and poised regulatory chromatin states (e.g. promoter), as defined by this ChromHMM model, and significantly depleted in heterochromatic states (Fig. 2A). We also observed that ATAC-seq signal is correlated with individual active histone modifications at transcription start sites (Fig. 2B) and genome-wide (Supplemental Fig. 2A). Thus, ATAC-seq correctly identifies poised and active regulatory regions at both specific loci and genome-wide.

Beyond simply identifying regulatory regions important for individual life stages, chromatin accessibility dynamics should highlight regulatory regions critical for transitions from embryo to larval stages, and from larval stages to adulthood. We examined whether genomic regions that showed accessibility changes from one life stage to another were enriched for specific chromatin state transitions. Regions that lost chromatin accessibility from embryo to L3 or from L3 to adult were enriched for transitions from active regulatory chromatin states (especially predicted enhancers) to repressed or heterochromatic states (Fig. 2D, Supplemental Fig. 2E). Conversely, the regions that gained accessibility during development were enriched for transitions from inactive chromatin states to active regulatory states (again, especially predicted enhancers) (Fig. 2C,E, Supplemental Fig. 2F). Collectively these results show that

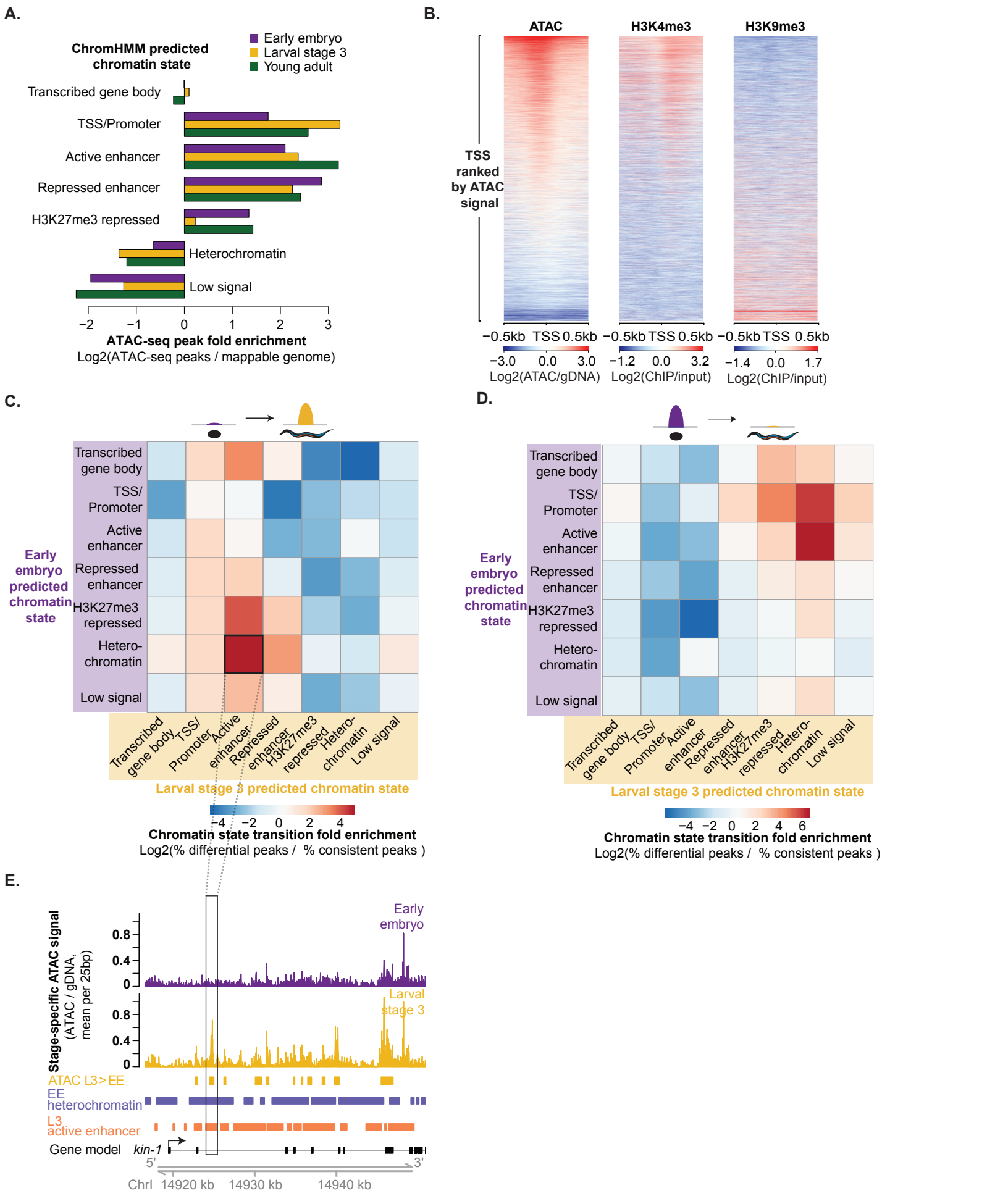
A.
ChromHMM predicted chromatin state
Early embryo
Larval stage 3
Young adult
Transcribed gene body
TSS/Promoter
Active enhancer
Repressed enhancer
H3K27me3 repressed
Heterochromatin
Low signal
−2 −1 0 1 2 3
ATAC-seq peak fold enrichment
Log2(ATAC-seq peaks / mappable genome)
B.
ATAC
H3K4me3
H3K9me3
TSS ranked by ATAC signal
−0.5kb TSS 0.5kb
−3.0 0.0 3.0
Log2(ATAC/gDNA)
−1.2 0.0 3.2
Log2(ChIP/input)
−1.4 0.0 1.7
Log2(ChIP/input)
C.
Early embryo predicted chromatin state
Transcribed gene body
TSS/ Promoter
Active enhancer
Repressed enhancer
H3K27me3 repressed
Hetero-chromatin
Low signal
Larval stage 3 predicted chromatin state
−4 −2 0 2 4
Chromatin state transition fold enrichment
Log2(% differential peaks / % consistent peaks)
D.
Early embryo predicted chromatin state
Transcribed gene body
TSS/ Promoter
Active enhancer
Repressed enhancer
H3K27me3 repressed
Hetero-chromatin
Low signal
Larval stage 3 predicted chromatin state
−4 −2 0 2 4 6
Chromatin state transition fold enrichment
Log2(% differential peaks / % consistent peaks)
E.
Stage-specific ATAC signal (ATAC / gDNA, mean per 25bp)
Early embryo
Larval stage 3
0.8
0.4
0
ATAC L3>EE
EE heterochromatin
L3 active enhancer
Gene model
kin-1
5'
3'
Chrl 14920 kb 14930 kb 14940 kb

Figure 2

Figure 2. ATAC-seq as a single assay describes the epigenome. (*A*) Enrichment of stage specific ATAC-seq peaks in ChromHMM-predicted chromatin states relative to values expected by chance; significance derived via 10,000 bootstrapping iterations (early embryo transcribed gene body $p = 0.633$, all others $p \leq 1e\text{-}4$). (*B*) Larval stage 3 (L3) ATAC-seq signal at 19,899 previously defined transcription start sites (TSS +/- 0.5kb) is correlated with active histone modifications (H3K4me3) and anti-correlated with heterochromatin (H3K9me3). (*C,D*) Regions that increase in accessibility in ATAC-seq peaks are enriched for transitions from inactive chromatin states to active regulatory states (C), while regions that decrease in accessibility are enriched for transitions from active regulatory states to inactive chromatin states (D). (*E*) An example of an increase in chromatin accessibility overlapping with a transition from heterochromatin to a predicted active enhancer chromatin state.

ATAC-seq performed on an entire organism, even a complex multi-tissue adult, is sensitive enough to detect important global changes in chromatin structure. Thus, ATAC-seq as a single assay constitutes an attractive alternative to performing multiple histone modification ChIP-seq experiments for identifying key regulatory regions, especially considering that ATAC-seq requires orders of magnitude less input than a single ChIP-seq.

ATAC-seq identifies new distal regulatory regions that serve as tissue- and stage-specific enhancers in *C. elegans*

Enhancers are key regulators of temporal- and tissue-specific gene expression that play important and conserved functions during development (Ren et al. 2015; Sun et al. 2015). However, the identification and characterization of novel enhancers remains a challenge because they can be specifically active in rare cell populations, some of which may not have even been characterized (Heinz et al. 2015). The extent and functional importance of distal regulatory regions in the *C. elegans* genome has been particularly underexplored (Reinke et al. 2013). While several methods to identify potential enhancers genome-wide have recently been developed, these methods suffer from notable drawbacks (Arnold et al. 2013; Giresi et al. 2007; He et al. 2010; Mito et al. 2007; Visel et al. 2009; Wang et al. 2008; Zhu et al. 2015; Zhu et al. 2013). For example, the co-occurrence of multiple histone modifications (e.g. H3K27ac, H3K4me1) or PolII through ChIP-seq experiments lacks the sensitivity to detect enhancers active only in rare sub-populations of cells and the resolution to precisely identify the active enhancer region (Furey 2012). We therefore investigated whether whole-organism ATAC-seq could overcome these challenges and

facilitate the identification of tissue- and stage-specific enhancers. Active and repressed enhancers were highly and significantly enriched in distal non-coding ATAC-seq peaks in all three stages (Supplemental Fig 3A). Furthermore, distal non-coding ATAC-seq peaks were significantly more conserved than expected by chance (Supplemental Fig. 3B, C), a defining feature of enhancers (Pennacchio et al. 2006). ATAC-seq peaks also exhibited significant overlap with RNA polymerase II transcription initiation complex binding at distal regions ($p < 1e-323$, one-sided Fisher's exact test), a feature used previously to predict enhancers at a single stage (embryos) (Chen et al. 2013). These analyses support the notion that distal non-coding ATAC-seq peaks include active enhancers. In *C. elegans*, a small number of functional enhancers have been experimentally mapped, including four enhancers in the upstream regulatory region of *hlh-1*, the *C. elegans* MyoD ortholog that regulates muscle development (Lei et al. 2009). ATAC-seq peaks overlap three of these four regions, and despite its lack of statistical significance, the fourth region still exhibits noticeable ATAC-seq signal in embryos (Supplemental Fig. 3D). Given that *hlh-1* is exclusively expressed in muscle (Krause et al. 1994), a tissue comprising less than 10% of *C. elegans'* cellular composition, these observations indicate that ATAC-seq performed on a whole organism is sensitive enough to identify functional tissue-specific enhancers.

We next sought to determine whether ATAC-seq dynamics could be leveraged to identify novel functional enhancers. Previous work has demonstrated that enhancers are precisely activated/inactivated at very specific times in development (Bonn et al. 2012; Kvon 2015; Sun et al. 2015). We hypothesized that the temporal resolution of our ATAC-seq data would capture these tightly orchestrated regulatory dynamics. To

experimentally test the functional enhancers predicted by our ATAC-seq data, we selected 13 distal non-coding ATAC-seq peaks that exhibit large changes in accessibility between stages. We generated multiple transgenic *C. elegans* strains with these 13 putative enhancer regions upstream of a minimal promoter (*pes-10*) driving expression of green fluorescent protein (GFP) containing a nucleolar localization signal (NLS) (Fig 3A, Table 1, Supplementary Table 7). To control for specificity, we also tested regions flanking 10 of the 13 the ATAC-seq peaks (ensuring these flanking regions were not themselves ATAC-seq peaks) (Table 1, Supplemental Table 7). By fluorescence microscopy, we examined the spatiotemporal GFP pattern in these transgenic strains to assess the temporal- and tissue-specificity of the putative enhancers. Using stringent criteria for defining enhancer activity (see Methods), we found that 6 of the 13 distal non-coding ATAC-seq peaks led to specific and consistent spatiotemporal GFP pattern, and for all but one of these, this pattern was observed regardless of the genomic orientation of the region – an important characteristic of enhancers (Ren and Yue, 2015). In contrast, 0 of the 10 flanking regions led to such a GFP pattern, indicating enrichment for functional enhancers in ATAC-seq peaks that change between stages ($p = 0.038$, one-sided Fisher's exact test) (Table 1, Fig. 3B-G, Supplemental Fig. 4A-E). Thus, dynamic chromatin accessibility is an excellent marker of functional enhancers and can be used to successfully identify novel distal regulatory regions.

Interestingly, the enhancers we experimentally identified display diverse spatiotemporal activity patterns. Three of the enhancer regions (putatively associated with *gei-13*, *mlt-8*, and *nhr-25*) are active in the head or tail hypodermis during development, while others are active specifically in the pharynx (C54G6.3) or muscle

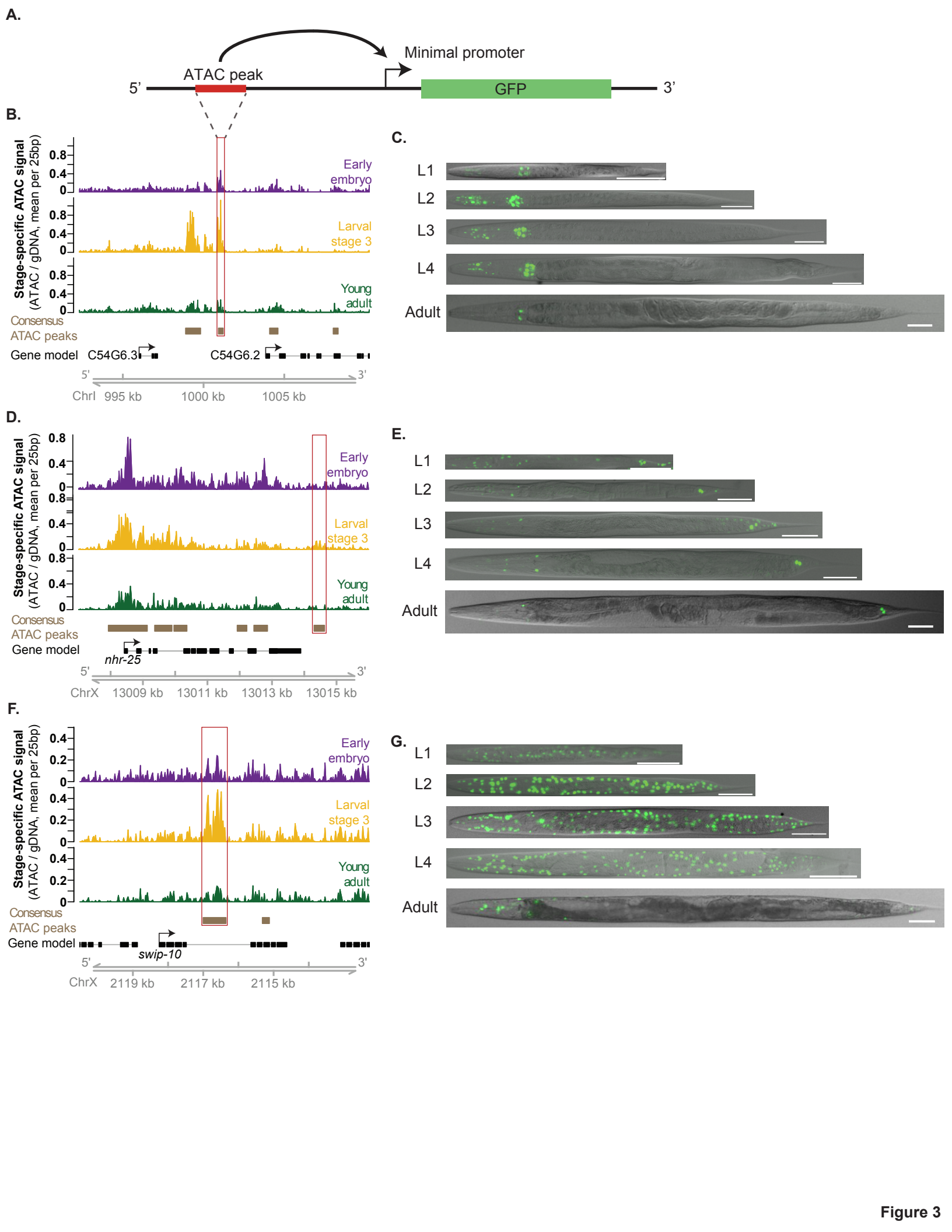
A.
Minimal promoter
ATAC peak
5'
GFP
3'
B.
Stage-specific ATAC signal
(ATAC / gDNA, mean per 25bp)
0.8
0.4
0
Early embryo
Larval stage 3
Young adult
Consensus ATAC peaks
Gene model
C54G6.3
C54G6.2
5'
3'
ChrI
995 kb
1000 kb
1005 kb
C.
L1
L2
L3
L4
Adult
D.
Stage-specific ATAC signal
(ATAC / gDNA, mean per 25bp)
0.8
0.4
0
Early embryo
Larval stage 3
Young adult
Consensus ATAC peaks
Gene model
nhr-25
5'
3'
ChrX
13009 kb
13011 kb
13013 kb
13015 kb
E.
L1
L2
L3
L4
Adult
F.
Stage-specific ATAC signal
(ATAC / gDNA, mean per 25bp)
0.4
0.2
0
Early embryo
Larval stage 3
Young adult
Consensus ATAC peaks
Gene model
swip-10
5'
3'
ChrX
2119 kb
2117 kb
2115 kb
G.
L1
L2
L3
L4
Adult

Figure 3

Figure 3. Dynamic ATAC-seq peaks identify functional enhancers with unique spatiotemporal specificity during *C. elegans* development. (A) Functional enhancer constructs used to generate *C. elegans* transgenic lines. Putative regulatory regions, or corresponding flanking regions for negative controls, are inserted upstream of a minimal promoter (*pes-10*) driving green fluorescent protein (GFP) localized to the nucleolus. (*B,D,F*) Distal (>1 kb from a TSS) non-coding ATAC-seq peaks with the largest fold change in ATAC-seq signal between any two stages were screened for potential enhancer activity. The approximate regions tested near *C54G6.2* (*B*), *nhr-25* (*D*), and *swip-10* (*F*) are boxed in red. (*C, E,* G) Specific patterns of spatiotemporal enhancer activity in transgenic lines. Representative images of GFP expression in staged *C. elegans* transgenic lines are presented with a 50 µm scale bar. All images were straightened with ImageJ and are grayscale images with florescence overlaid.

Table 1.	5’ → 3’ Orientation	3’ → 5’ Orientation	Negative Control (flanking)
Putative Gene	# GFP-positive lines	# GFP-positive lines	# GFP-positive lines
C54G6.3	8/8	2/2	0/3
nhr-25	8/8	3/3	0/8
swip-10	8/8	2/2	0/1
mlt-8	11/11	5/5	0/3
gei-13	10/11	6/6	0/1
No-insert control	0/3	-	-

Table 1. Transgenic reporter lines generated for validation of ATAC-seq peaks as functional enhancers. A summary of the number of independent transgenic strains generated to assess functional activity of putative enhancer regions identified by ATAC-seq dynamics. The number of independent strains that exhibited consistent spatiotemporal expression patterns of GFP with inserted ATAC-seq peaks in the native genomic orientation, reverse orientation, and with a nearby a flanking region, is reported here.

cells (*swip-10*). In addition, the enhancers are located at a wide diversity of genomic positions relative to their putatively associated gene: upstream of the TSS (at distances varying from 1-9kb), within introns, and downstream of the coding sequence. A particularly exciting example is the *nhr-25*-associated enhancer, which is downstream of the 3' UTR (more than 5kb downstream of the TSS). NHR-25 is a conserved nuclear receptor primarily expressed in the hypodermis (and somatic gonad) during larval development (Gissendanner et al. 2000). We find that this 3' *nhr-25* enhancer specifically drives GFP expression in approximately 20 hypodermal cells in the head and tail of the worm during larval development. The limited expression pattern of the 3' *nhr-25* enhancer (as well as the other enhancers we identified) shows that ATAC-seq performed on whole organisms is sensitive enough to identify regulatory regions active in specific cell types, though we cannot rule out the possibility that these regions are accessible, but their activity repressed (e.g. by H3K27me3) in other cells. Intriguingly, both human and mouse orthologs of *nhr-25* (NR5a in mammals) contain a binding site for CTCF in their 3' region. As CTCF is a well-established regulator of chromatin structure (Ong et al. 2014), this raises the exciting possibility that this region may be a conserved region of three-dimensional chromatin structure regulation.

Collectively, these findings demonstrate the power of using ATAC-seq dynamics as an unbiased approach to identify functional and conserved enhancers active in a small subset of cells. When applied to whole organisms or complex samples composed of diverse cellular populations, this approach is capable of capturing regulatory regions that may have been missed in studies of isolated cell populations.

Specific motifs for transcription factors, including a potential pioneer factor, predict changes in chromatin accessibility

To explore the regulatory underpinnings of chromatin accessibility dynamics, including that of enhancers, we examined the occurrence of experimentally defined *C. elegans* transcription factor (TF) binding motifs (Narasimhan et al. 2015) in the dynamic chromatin accessibility regions identified by ATAC-seq. Remarkably, in regions that change chromatin accessibility between early embryo and L3 or between L3 and young adult, we observe significant enrichment of motifs associated with TF homologs and orthologs that have previously been connected to chromatin accessibility dynamics (Fig. 4A and Supplemental Fig. 5A). For example, the motif for BLMP-1, the *C. elegans* ortholog of human BLIMP-1/PRDM1 (a TF recently shown to regulate target genes by recruiting chromatin-remodeling complexes in human B cells (Minnich et al. 2016)), is enriched in peaks that are more accessible in L3 when compared to embryos or adults. Furthermore, the motif for ELT-3, a GATA TF, is enriched in peaks more accessible in L3 than embryos (the GATA family was recently demonstrated to be an important regulator of chromatin accessibility during human hematopoiesis (Corces et al. 2016)).

Intriguingly, the DNA binding motif for EOR-1, which resembles a dimeric version of the canonical GAGA-motif, was significantly enriched in distal non-coding ATAC-seq peaks (Supplemental Fig. 5C) and in ATAC-seq peaks that gained in accessibility in L3 versus embryo (Fig. 4A). This GAGA/EOR-1 motif was also present in 2 of the 5 functional enhancers that we experimentally validated, including the *nhr-25* associated enhancer described above. TFs binding GAGA-motifs have been identified as modulators of chromatin structure dynamics from plants to humans (Hecker et al. 2015;

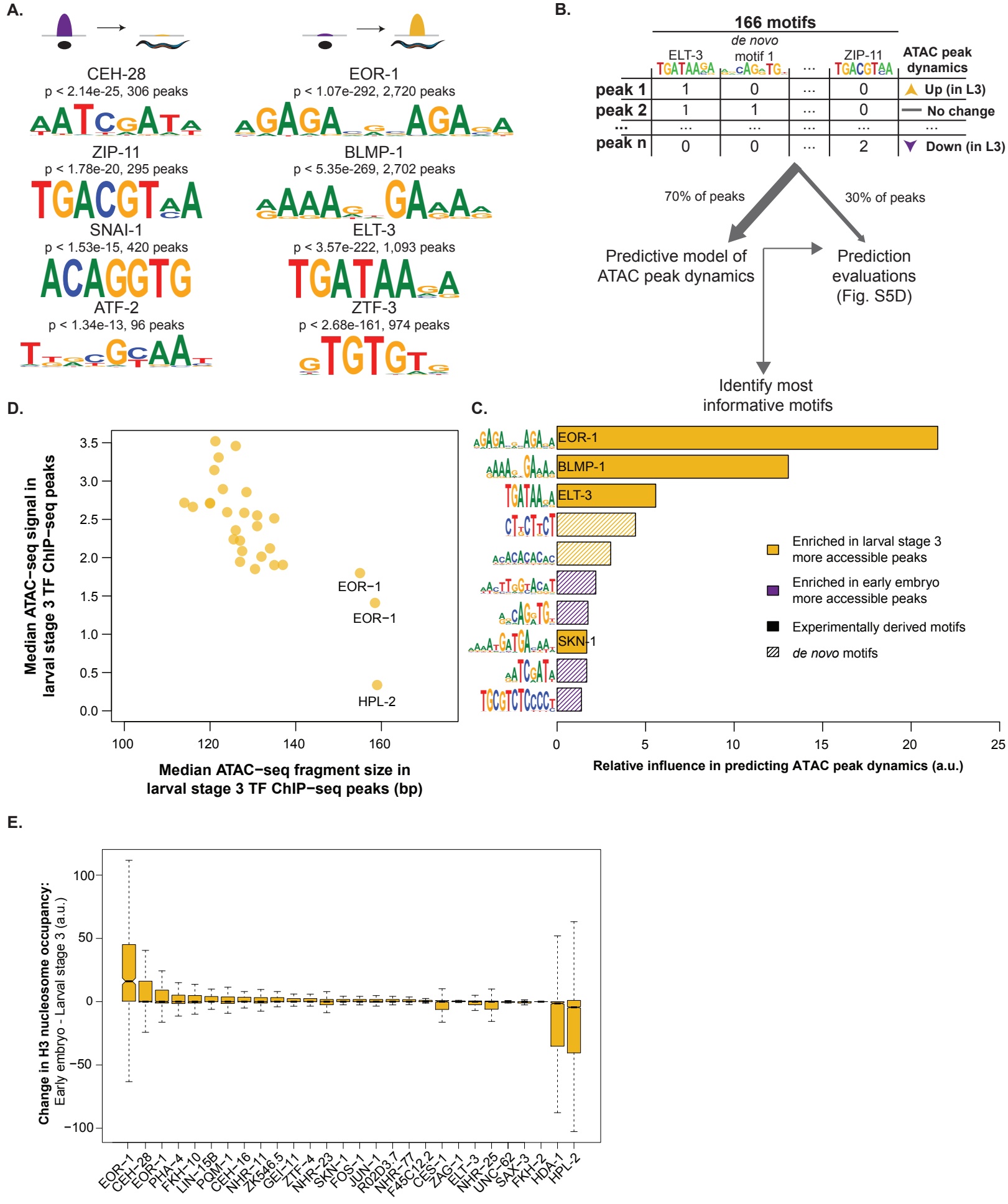
A.
CEH-28
p < 2.14e-25, 306 peaks
EOR-1
p < 1.07e-292, 2,720 peaks
ZIP-11
p < 1.78e-20, 295 peaks
BLMP-1
p < 5.35e-269, 2,702 peaks
SNAI-1
p < 1.53e-15, 420 peaks
ELT-3
p < 3.57e-222, 1,093 peaks
ATF-2
p < 1.34e-13, 96 peaks
ZTF-3
p < 2.68e-161, 974 peaks
B.
166 motifs
ELT-3
de novo motif 1
ZIP-11
ATAC peak dynamics
peak 1
peak 2
peak n
Up (in L3)
No change
Down (in L3)
70% of peaks
30% of peaks
Predictive model of ATAC peak dynamics
Prediction evaluations (Fig. S5D)
Identify most informative motifs
C.
EOR-1
BLMP-1
ELT-3
SKN-1
Enriched in larval stage 3 more accessible peaks
Enriched in early embryo more accessible peaks
Experimentally derived motifs
de novo motifs
Relative influence in predicting ATAC peak dynamics (a.u.)
D.
Median ATAC−seq signal in larval stage 3 TF ChIP−seq peaks
Median ATAC−seq fragment size in larval stage 3 TF ChIP−seq peaks (bp)
EOR−1
EOR−1
HPL−2
E.
Change in H3 nucleosome occupancy: Early embryo - Larval stage 3 (a.u.)
EOR−1
CEH−28
EOR−1
PHA−4
FKH−10
LIN−15B
POU−1
CEH−16
NHR−11
ZK546.5
GEI−11
ZTF−4
NHR−23
SKN−1
FOS−1
JUN−1
R02D3.7
NHR−77
F45C12.2
CES−1
ZAG−1
ELT−3
NHR−25
UNC−62
SAX−3
FKH−2
HDA−1
HPL−2

Figure 4

Figure 4. Motifs associated with large increases in chromatin accessibility during development reveal key transcription factors and potential pioneer factors. (*A*) ATAC-seq peaks which decreased (left) or increased (right) accessibility between early embryo and L3 are enriched for previously identified transcription factor binding motifs; p-values are Benjamini-Hochberg corrected for multiple hypothesis testing. (*B*) The number of instances of previously identified as well as *de novo* motifs (see Fig. S5B) in each consensus ATAC-seq peak were used as features in a machine learning model to predict how the ATAC-seq peaks changed between early embryo and L3 (increasing, decreasing, or no change). A training set (70% of all ATAC-seq peaks) was used to build the model, while the remaining held-out testing set was used to assess model quality (see Fig. S5D). (*C*) The relative influence of every motif from the machine learning model in Fig. 4B was quantified. Solid bars are previously defined motifs, while hashed bars are *de novo* identified motifs in dynamic ATAC-seq peaks. (*D*) The median ATAC-seq signal and fragment size at the midpoint (+/- 50 bp) of L3 TF ChIP-seq peaks; box plots of the same data are in Fig. S4GH. EOR-1 ChIP-seq data is noted for reference. (*E*) The change in H3 nucleosome occupancy between early embryo and larval stage 3 (L3) at the midpoint of each L3 transcription factor ChIP-seq peak was calculated using Danpos and publicly available H3 ChIP-seq.

Lund et al. 2013; Srivastava et al. 2013) and the motif itself has been shown to be required for full functionality in two previously defined *C. elegans* enhancers (Harfe et al. 1998; Jantsch-Plunger et al. 1994). Finally, GAGA/EOR-1 factors might play an important role in the regulation of chromatin accessibility as EOR-1 has been shown to genetically interact with components of two nucleosome remodeling complexes, SWI/SNF and RSC (Lehner et al. 2006), and a dimeric GAGA-motif nearly identical to the EOR-1 motif was found to be highly enriched in ChIP-seq peaks for two separate SWI/SNF components in *C. elegans* (Riedel et al. 2013).

To independently verify the importance of the GAGA/EOR-1 motif, we employed machine learning to identify TF binding motifs that are predictive of regions that gained or lost accessibility between early embryo and larval stage 3. Importantly, this method is unbiased and allowed the use of both *de novo* and experimentally derived DNA binding motifs. As input features for the machine learning model, we used the largest available set of previously defined *C. elegans* TF binding motifs (107 high-confidence motifs from 98 TFs out of more than 750 TFs in the *C. elegans* genome (Narasimhan et al. 2015)) as well as 59 motifs discovered *de novo* in regions that changed in accessibility between early embryo and L3 (Fig. 4B, Supplemental Fig. 5B, 5D) (see Methods). Using this machine learning model, we were able to identify the motifs that are the most predictive of ATAC-seq dynamics; interestingly, the top three most informative motifs in our model are the GAGA-motif that we identified above and motifs for known chromatin regulators (BLMP-1 and ELT-3) (Fig. 4C). Thus, machine learning independently validates the GAGA/EOR-1 motif as a potential important regulator of chromatin accessibility.

We next determined whether EOR-1 plays a role in chromatin accessibility changes. Two EOR-1 ChIP-seq datasets at L3 (Araya et al. 2014; Boyle et al. 2014) revealed enrichment for a dimeric GAGA-motif (Supplemental Fig. 5E) and were significantly enriched for regions that overlapped ATAC-seq peaks that gain accessibility in L3 compared to early embryo (Supplemental Fig. 5F), indicating that EOR-1 is indeed bound to the GAGA-motif and that EOR-1 may be involved in regulation of chromatin accessibility. To examine if EOR-1 can bind to closed chromatin, we quantified ATAC-seq signal and fragment size, which correlates with nucleosome occupancy and chromatin state (Buenrostro et al. 2013). EOR-1 ChIP-seq peak summits have significantly less ATAC-seq signal and significantly larger fragment sizes (indicative of less accessible chromatin regions) than all 26 other high quality L3 TF Chip-seqs, except for the heterochromatin-associated protein HPL-2 (FDR < 0.05) (Fig. 4D, Supplemental Fig. 5GH), suggesting that EOR-1 can bind closed chromatin. We also quantified nucleosome occupancy dynamics in the summits of transcription factor ChIP-seq peaks using a publicly available histone H3 ChIP-seq datatset (Contrino et al. 2012; Gerstein et al. 2010). This revealed that EOR-1 summits have a large decrease in nucleosome occupancy, which should increase chromatin accessibility, from embryo to L3 (Fig. 4E). We note that in both ATAC-seq metrics and nucleosome occupancy meausres, EOR-1 seems to behave similarly to PHA-4 (Fig. 4E, Supplemental Fig. 5GH), a known pioneer factor (Hsu et al. 2015). These data may then suggest that EOR-1 could bind to less accessible regions of the genome than other TFs, and that this is correlated with chromatin opening during development. Given the prevalence of EOR-1 binding sites at

enhancers, these results are also consistent with a model in which EOR-1 acts as a pioneer factor regulating several types of genomic regions, including enhancers.

Discussion

Here we show for the first time that measuring chromatin accessibility in a whole organism is sensitive enough to detect dynamic changes throughout development and even to identify novel functional enhancers active in only a small subset of the whole organism. This approach, developed here for *C. elegans,* should be readily applicable to other complex samples *in vivo* such as whole mammalian organs or tumor samples.

In *C. elegans*, enhancer identification has historically been limited, with most previous studies employing a single gene approach and focusing on promoter-proximal regions (Harfe et al. 1998; Jantsch-Plunger et al. 1994; Lei et al. 2009). Others have attempted to identify enhancers genome-wide, but have not functionally validated predicted enhancer activity (Chen et al. 2013; Vavouri et al. 2007). In this study, we identify and functionally characterize distal-ATAC-seq peaks as novel, active enhancers. These enhancers have a range of spatiotemporal activity patterns that are orientation-independent, and are found throughout the genome, suggesting that enhancers may be more prevalent in *C. elegans* gene regulation than previously appreciated. Of particular interest is the enhancer downstream of conserved transcription factor *nhr-25*. We have shown that this putative-*nhr-25* enhancer acts in an orientation-independent manner, and given that it is more than 5kb downstream of the closest TSS (*nhr-25*), this enhancer may be looping to the *nhr-25* TSS to enhance transcription. This three-dimensional chromatin architecture is in contrast to the promoter-proximal model most often thought to regulate

gene expression in *C. elegans*, but is in agreement with recent studies that identified insulator-like loci throughout the *C. elegans* genome (Crane et al. 2015). Insulators are important regulators of three-dimensional chromatin architecture (Schoborg et al. 2014). Of note, NHR-25 orthologs in fly (Ftz-F1), mouse (NR5A), and human (NR5A), exhibit a consistent signature of 3-dimensional chromatin architecture downstream of the 3' UTR for each gene (insulator class I in flies, and CTCF-binding sites in mice and humans (Attrill et al. 2016; Rosenbloom et al. 2013; Sharov et al. 2006)). This raises the tantalizing possibility that the *nhr-25* enhancer we have identified is evolutionarily conserved.

Furthermore, we have uncovered a potential role for a likely GAGA factor in *C. elegans*, EOR-1. The EOR-1 motif we initially detected closely resembles a dimeric version of the canonical GAGA-motif bound by Trl/GAGA-Associated Factor (GAF) in *Drosophila*. GAF is a multi-faceted transcription factor that can associate with heterochromatin (Raff et al. 1994), remodel chromatin in concert with nucleosome remodelers (Okada et al. 1998), and acts as a transcriptional activator in part due to its ability to increase chromatin accessibility (Adkins et al. 2006). Similar to *Drosophila* GAF, *C. elegans* EOR-1 is a transcriptional activator in the Ras/ERK signaling pathway, and both EOR-1 and GAF proteins have a BTB/POZ domain on their N-terminal as well as C2H2 zinc-fingers and polyQ domains on their C-terminals (Howard et al. 2002). In *C. elegans* EOR-1 genetically interacts with at least two chromatin remodeling complexes (SWI/SNF and RSC) (Lehner et al. 2006). This is interesting given our findings that EOR-1 binds to less accessible and potentially even nucleosome-occupied regions of the genome, and that the EOR-1 motif is predictive of increased accessibility in development.

GAGA-factors are conserved regulators of gene expression (Mahmoudi et al. 2002; Ohtsuki et al. 1998; Petrascheck et al. 2005; Srivastava et al. 2013), and a GAGA motif has been noted to be necessary for enhancer functionality in *C. elegans* (Harfe et al. 1998). Indeed, 2 out of 5 of the novel validated enhancers identified in this study (*nhr-25,* and *swip*-10) contain an EOR-1/GAGA motif. Collectively, these data suggest that EOR-1 is a particularly important regulator of chromatin accessibility at enhancers in *C. elegans* and potentially other species.

Using *C. elegans* as a paradigm, we have shown that ATAC-seq performed on complex, heterogeneous samples can reveal novel, spatiotemporally-specific genetic regulators, and that measuring chromatin accessibility across a time course is capable of identifying important dynamic regions. We have highlighted important applications of this approach: discovering functional distal regulatory regions active in only a small subset of the total sample, and identifying candidate regulators of genome-wide chromatin dynamics. Using our approach as a blueprint in complex, multi-tissue samples could yield vital insights into complex and heterogeneous biological processes, which are not always amenable to single cell approaches, including development, cancer and aging.

Methods

Maintenance and strains

All *C. elegans* strains were maintained on NGM at 20°C using *Escherichia coli* OP50.1 as a food source. Wild-type (N2) worms were provided by Dr. Man-Wah Tan.

Plasmids for enhancer screen

The minimal *Ppes-10*::4xSV40NLS::GFP::*let-858* 3-UTR plasmid (pL4051) used for enhancer screening was a gift from Andrew Fire (Addgene plasmid #1629). The pRF4 co-injection marker plasmid *Prol-6::rol-6(su1006)* was a gift from Stuart Kim's lab.

Enhancer GFP plasmid construction

For the enhancer screen, each putative regulatory region was cloned in the pL4051 plasmid, upstream of a minimal promoter (*pes-10*) driving expression of a *C. elegans* intron- and photo-stability-optimized GFP containing an N-terminal nucleolar localization signal (NLS).

Putative regulatory regions were chosen by selecting ATAC-seq peaks that exhibited the largest differential accessibility between two stages and that were at least 1kb from a transcription start site. Flanking negative control regions were chosen by selecting regions within 2kb of the putative regulatory regions that were not in peaks of accessibility. Primers were designed to amplify each region as well as 50-500bp flanking either side (Supplemental Table 8). The fragments were amplified from genomic DNA extracted from N2 worms using NEBNext High-Fidelity and cloned into the pL4051 plasmid. Cloned PCR fragments were sequence-verified.

Transgenesis

Wildtype N2 day 1 adults were microinjected in following the standard protocol (Mello et al. 1995) with a mix containing 75 ng/µl of the respective enhancer-GFP reporter constructs and 75 ng/µl of pRF4, a *rol-6* co-injection marker.

Enhancer screen in *C. elegans*

Stable extrachromosomal transgenic lines for putative enhancer regions determined by ATAC-seq peaks, negative control regions (regions flanking ATAC-seq peaks), or the no-insert control were generated. For each transgenic line, mixed-staged worms were screened for GFP signal distinct from the no-insert control background signal, which is 1-2 nuclei near the pharynx in all larval and adult stages (Supplemental Fig. 4E). To quantify the consistency of GFP expression pattern, all lines (including the negative control lines) (Table 1) were scored for GFP expression pattern in a blinded manner.

For those regulatory regions that displayed a consistent GFP expression pattern in transgenic worms, we also generated 2-5 stable transgenic lines in which the region was inverted from its endogenous orientation relative to the putative TSS. These transgenic lines were assessed in the same manner as described above. In most cases, the closest downstream TSS was assigned as the putative TSS. However, in cases where a TSS was not present on either side of the region (*nhr-25, swip*-10)), we preferentially selected the closest TSS while also considering the GFP spatiotemporal activity we observed and the canonical functions of the neighboring genes.

Fluorescence microscopy

Transgenic worms were synchronized via a timed egg lay and prepared for live imaging at the indicated stages. Larval and adult worms were washed off plates in 100 mM levamisole (MP Biomedicals), resuspended in M9, allowed to settle, and washed two additional times with levamisole. Larval and adult worms were washed for a minimum of 30 minutes to clear bacteria from the gut. Worms were transferred to 2% agarose pads and imaged using a Zeiss AxioSkop 2 Plus at 20x magnification for larvae or 10x

magnification for adults. Images were acquired using an AxioCam MRc camera with AxioVision 4.7 software. Exposure times were kept constant for all stages of each strain. DIC and GFP images were merged and worm bodies were straightened in FIJI. One representative image is shown per stage with 50 μm scale bars.

Collection of timed samples for ATAC-seq

For ATAC-seq, three sets of completely independent biological replicates (i.e. performed at different times) were prepared in the following manner. To synchronize the parents of each sample, well-fed mixed stage N2 worms from one 10cm or three 6cm plate(s) (approximately 500 worms) were treated with 10% bleach for 4 min to isolate early embryos (as described in Wormbook). The resulting embryos (approximately 800) were grown to adulthood on 2-3 10cm plates. As soon as the worms reached adulthood, they were placed on fresh 10cm plates at a density of approximately 100 worms per plate for synced egg-laying. After 45 min, the adults were removed, and the plates were returned to 20°C to allow the embryos to grow to the desired stages.

To collect the embryos, the egg-laying adults were collected, and approximately 50 μl of packed adults were washed once with M9 medium and bleached to yield early embryos as described in Wormbook. Following the last wash after bleaching, the embryos were left in 100μl of M9 medium, and the tubes flash frozen in liquid nitrogen and kept at -80°C. The parents were also flash frozen and used for genomic DNA controls.

To collect larval stage 3 (L3) worms, plates were repeatedly checked after egg-laying to ensure no parents had remained. 36 hr after the start of the egg lay, L3 animals

were collected from 2 10cm plates. Finally, 57 hr post egg lay one of the remaining plates was checked every 10 min to ensure that the majority of the plate was composed of young adult animals with no more than 1-2 eggs. At that point young adult animals were harvested. Immediately after collection, all worm samples were washed twice with M9 before being flash frozen and stored at -80°C.

Nuclei purification and ATAC-seq of samples

Nuclei were purified from frozen samples as described (Haenni et al. 2012). Briefly, samples were thawed on ice, and then mixed with 150 μl of 2X nuclei purification buffer (20mM Hepes pH 7.6, 20mM KCl, 3mM $MgCl_2$, 2mM EGTA, 0.5 M Sucrose, 0.05% Triton, 1mM DTT, 0.05M NaF, 40mM β-glycerophosphate, 2mM Na_3VO_4). The samples were then transferred to a Wheaton stainless-steel homogenizer and were homogeneized with 3 plunger strokes. The resulting samples were centrifuged at 200 g for 1 min to remove worm debris. The supernatant was transferred to a fresh tube, and the remaining pellet was resuspended in 150 μl of 2X nuclei purification buffer and homogenized as described above. The process was repeated until no visible pieces of worm remained (approximately 5 times). The pooled supernatants were centrifuged at 200 g for 1 min to remove worm debris. Finally, the nuclei were pelleted at 1000 g for 10 min. All steps were completed at 4°C.

The purified nuclei were immediately used for the ATAC-seq protocol (Buenrostro et al. 2013). Briefly, the nuclei were resuspended in 47.5 μl of Nextera Tagmentation buffer (Nextera DNA Sample Preparation Kit) and incubated with 2.5 μl of the Tn5 transposase at 37°C for 30 min. Resulting DNA fragments were purified using a

miniElute column (Qiagen) and amplified by NEBNext High-Fidelity PCR Master Mix in a total volume of 50 μl. The thermocycling protocol for this reaction was 72°C for 5 min, 98°C for 30 s and 5 cycles of 98°C for 10 s, 63°C for 30 s and 72°C for 1 min. Every sample shared the Adapter1 primer, and had a unique barcoded Adapter2 primer (Supplemental Table 8). To ensure over-amplification did not occur, after the initial 5 cycles, the number of remaining cycles required was estimated for each sample using qPCR (BioRad CFX96 Real-Time System). To do so, a similar reaction to the above was set up (NEBNext, Adapter1 and a single Adapter2), with the addition of SYBRGreen, and using 5 μl of the previous PCR as template. The final volume was 15 μl, and the thermocyling was as above except that the initial incubation step was 98°C for 30 seconds, and 40 cycles were performed. The number of additional cycles was determined to be the number it took for the qPCR to reach one-third maximal fluorescence. The original PCR was then resumed and each sample cycled as necessary. Following amplification, the samples were purified using QIAquick columns (Qiagen), and library quality was verified on an Agilent bioanalyzer. Sequencing was performed using 101 bp paired-end sequencing on an Illumina Hi-seq 2000.

Isolation and ATAC of gDNA controls

To generate an input control for ATAC-seq, approximately 100 μl of packed adults was thawed and embryos were collected as described above to avoid any bacterial contamination. Next, genomic DNA was isolated by proteinase K treatment (0.2 mg/ml, 55°C for 3 hr before heat killing the protein at 95°C for 25 min), then RNase treatment (1.25 mg/ml, 37°C for 30 min). gDNA was purified by double

phenol/chloroform/isoamyl alcohol (25:24:1, pH 8.0) extraction, with an additional chloroform extraction, and precipitated with ethanol and NaOAc (3mM, pH 5.5), and resuspended in 50 μl TE (10mM Tris-HCl pH7.5, 1mM EDTA). The gDNA was quantified using a Qubit, and 10ng used for a standard ATAC-seq protocol as described above.

Genomic analysis

For all analyses the ce10/WS220 version of the *C. elegans* genome (Rosenbloom et al. 2015). Base version of Perl (v5.16.3), Python (v2.7.9) and R (R Core Team 2013) as well as Samtools (v0.1.19-44428cd) and Bedtools (v 2.21.0) (Quinlan et al. 2010) were used throughout, unless noted otherwise.

Gene definitions

Refseq gene definitions were downloaded from UCSC Genome Browser, and were supplemented by replacing transcription start sites (TSSs) with experimentally defined (Chen et al. 2013) TSSs where possible.

Mappable regions

The mappable genome was defined as those regions that single end 100bp reads could be mapped. This was accomplished using HotSpot (John et al. 2011).

ATAC-seq read alignment and quality filtering

The nine experimental ATAC-seq libraries, as well as an input control (ATAC-seq on purified genomic DNA) were sequenced to a median depth of over 17 million unique, high-quality mapping reads per sample (Supplemental Table 1). Sequencing adaptors were trimmed using a custom script that aligns the 5' ends of the forward read and reverse complement of the reverse read, and then removes any aligning sequence (minimum length 3). Subsequently, all reads were aligned to the ce10 version of the *C. elegans* genome, including mitochondrial sequence, with bowtie2 (Langmead et al. 2012) (v2.1.0) with the following settings: --end-to-end, --no-mixed, and -X 2000. Next, all samples were filtered for mapping quality (MAPQ >= 30) and PCR duplicates were marked using Picard tools (Broad Institute 2014), and subsequently removed. Every read was then adjusted for the binding footprint of Tn5 (9 bp) as previously described (Buenrostro et al. 2013); specifically reads were shifted 4 (positive strand) or 5 (negative strand) bp 5' relative to the reference genome. Finally, to give optimal single-base resolution, reads were trimmed to the single 5'-most-base. Then as a means of assuring library quality, we examined the fragment size distribution of each sample, and observed approximately 147bp periodicity corresponding to mono-, di-, tri-, etc., nucleosomes, an important indicator of technical sample quality (Buenrostro et al. 2013) (Supplemental Fig. 1A).

ATAC-seq peak calling

For every replicate, prior to calling peaks with MACS (Zhang et al. 2008) (v2.1), single-base reads were shifted 75bp 5' to mimic read distributions of a 150 bp fragment of ChIP-seq, thereby allowing use of MACS. The following settings were used for MACS: -

g 9e7, -q 5e-2, --nomodel, --extsize 150, -B, --keep-dup all, and --call-summits. The resulting peaks included a portion that had multiple summits; in order to maximize our resolution these summits were subsequently separated into individual peaks by treating the midpoint between the two adjacent summits as the 5' and 3' end of each individual peak, respectively.

To identify only the most high-confidence set of peaks, peaks were called for each individual experimental replicate, as well as for pooled replicates from each stage (i.e. all single base pair reads for that stage, regardless of biological replicate), and for two pseudoreplicates which were generated by randomly splitting the pooled samples in half. Consensus peaks for each stage were then classified as any peaks called in the pooled-replicate sample that were at least 50% overlapping A) all biological replicates or B) both pseudoreplicates and at least 2 of the 3 biological replicates (Supplemental Table 2). We employed this conservative approach for all our peak-calling analyses except for the transcription factor ChIP-seq intersection analysis in which we used the standard approach of peak calling from pooled replicates.

To account for any unknown issues arising from either biological biases (e.g. Tn5 motif preferences, or undocumented repeats in the *C. elegans* genome), or biases within our analysis pipeline we treated our input control, purified *C. elegans* genomic DNA which was transposed with Tn5, in a manner identical to experimental samples through the stage of calling peaks. We expected that reads from the gDNA controls would be uniformly distributed across the genome, and in large part that is what we observed, but unexpectedly we identified regions that had significant pileups of reads. These gDNA pileups highly overlapped annotated repeats (Supplemental Fig. 1B), and reads within the

pileups tended to have consistent single nucleotide polymophisms that were not observed in experimental samples (data not shown). These two pieces of evidence suggest that these peaks arise from PCR artifacts or mapping errors. We took the conservative approach of eliminating consensus peaks that overlapped the gDNA pileups by 20% or more. In addition, we also removed any consensus ATAC-seq peaks that overlapped a blacklist region (a set of regions identified by modENCODE that have "anomalous, unstructured, high signal/read counts in next gen sequencing experiments independent of cell line and type of experiment" (Boyle et al. 2014)) by a single base pair.

A set of 30,832 meta-peaks was generated by first pooling all experimental reads regardless of stage and then calling and masking peaks as described above. This set of peaks was then combined with all stage-specific consensus peaks. To avoid duplicate peaks, overlapping peaks were merged into a single peak, if their summits were within 300 bp. The resulting peak set should include all stage-specific peaks as well as peaks which were not accessible enough to be detected in a single stage, but consistent enough to be detected with all stage-pooled samples (Supplemental Table 3).

Generation of ATAC-seq enrichment scores for mapping

The ATAC-seq single base pair reads for each developmental stage were pooled and quantified at every base in the genome using Bedtools (coverageBed –d); the end result being counts of single base pair ATAC-seq reads at every base. This was repeated for the gDNA input control as well. Each sample was then normalized for total sequencing depth to generate the number of reads per million mapped (RPMM) at every base. Finally, enrichment over background was calculated for each developmental stage by

taking the log2 of the experimental RPMM divided by the input control RPMM plus 0.1 to avoid 0 at every base in the genome. For display purposes, these enrichment values were binned into 10, 25, or 50bp non-overlapping windows, the mean enrichment score for that region reported, and the value returned to linear scaling; all values below 0 (i.e. more input signal than experimental signal) were trimmed to 0. All signal sample plots were generated using the R package gViz (Hahne et al. 2016).

Differential accessibility and batch effect removal

To identify regions of dynamic accessibility during development, DiffBind (Stark 2011) was performed using the consensus ATAC-seq peaks, along with single base pair reads for each biological replicate. In addition, the gDNA control single base pair reads were included along with the score setting of DBA_SCORE_RPKM_FOLD during the counting step, thereby calculating the fold change between sample and control, after normalizing for sequencing depth. Other non-default settings at this step included, setting bRemoveDuplicates to false, as we had already removed duplicates, setting the fragment size to 1 to avoid any read shifting, and bScaleControl to true. Next, to remove technical variation between biological replicates, we extracted the score calculated by DiffBind and used ComBat (Leek JT 2015). The batch-corrected scores were then returned to DiffBind and differential peak calling completed using the following non-default settings for dba.analyze: bTagwise=FALSE, bFullLibrarySize=TRUE, method=DBA_EDGER, bSubControl=TRUE, bReduceObjects=FALSE. A final FDR of 0.05 was used to call significantly different peaks.

Single replicate ATAC-seq signal correlation

The ComBat normalized signal data for each replicate, which was total enrichment over input in all consensus ATAC-seq peaks, was clustered with the pvclust (Suzuki 2014), using Spearman correlation and the hclust.method = 'complete'. The same Spearman rho values were plotted using the R package pheatmap (Kolde 2013).

ATAC peak enrichment

To assess enrichment of a set of peaks in chromatin states, the portion of peaks overlapping the loci of interest (e.g. promoters) by at least 50% was compared to the median portion of the null distribution meeting the same requirements; the same process was used for transcription factor ChIP-seq peaks, but only a single base pair of overlap was required as not all TF bind in the center of accessibility (Buenrostro et al. 2013). In each case, the null distribution was generated by shuffling the ATAC-seq peaks across the mappable genome (described above) masked for blacklisted regions (described above) and gDNA peaks (see ATAC-seq peak calling), 10,000 times using Bedtools. The log2 fold enrichment is a comparison between the portion of experimental peaks overlapping features of interest versus the median portion of null distribution peaks overlapping the same features. Significance was assessed by calculating an empirical cumulative distribution function for the null distribution values, and finding the quantile for the experimental peak portion.

ChIP-seq analysis

Histone modifications: The ce10 alignment files for early embryo and larval stage 3 were downloaded from modEncode (Ho et al. 2014). Reads for young adult samples were downloaded from the modEncode DCC and aligned to ce10 with BWA (Li et al. 2009) (v0.7.9a-r786) default settings to maintain consistency with the larval and embryo samples. Subsequently, all embryo, larval and young adult samples were filtered for mapping quality (MAPQ >= 30) and duplicates were marked with Picard and removed. Biological replicates were pooled for each histone mark and input, and psuedoreplicates created as described above for ATAC-seq samples. Next, fragment size was estimated via SPP (Kharchenko et al. 2008) for each biological replicate as well as the pooled samples and provided to MACS as "-shiftsize" for peak calling. In addition, the following settings were used for peak calling: -g 9e7, -p 1e-2, --nomodel, -B, --SPMR. Following that, consensus peaks were determined as above for ATAC-seq peaks, but requiring overlap with either A) both biological replicates or B) both pseudoreplicates and at least 1 of 2 biological replicates.

Transcription Factors: For early embryo, larval stage 3, and young adult peaks of transcription factor binding were taken from Araya, *et al.* 2014(Araya et al. 2014). Only the most high-confidence ChIP-seqs were used. A complete list of all TFs used is included in Supplemental Table 9.

Chromatin state prediction

Chromatin state predictions were generated using ChromHMM (Ernst et al. 2012) (v1.10). The model was built using H3K27ac, H3K27me3, H3K4me1 and H3K36me3 from early embryo, larval stage 3, and young adult samples as well as H3K4me3, H3K9me3, H3K79me2, and H4K20me1 from early embryo and larval stage 3. To

maximize the amount of data used to train the model, we also used H3K79me3 in lieu of H3K79me2 in young adult, where H3K79me2 was not available. During the binerization of the genome, settings included: bin = 100bp, p = 1e-3. For prediction of states, a 19 state model was empirically determined to most closely represent the expected states; for example, promoter-like states found near TSSs and transcription-like states found within gene bodies. The emission heatmap created by ChromHMM can be seen in Supplemental Fig. 2B. For ease of interpretation, similar states were subsequently merged, resulting in 7 main states: transcribed, promoter, enhancer, repressed enhancer, repressed, heterochromatin, and low-signal.

In the process of completing this study a similar model using a subset of this data, a different program, and data from other organisms was published (Ho et al. 2014). Our models are quite similar for both early embryo and larval stage 3 (Supplemental Fig. 2CD). Because our model used all of the chromosomes, only *C. elegans* data, and included young adults, we have used our model exclusively, though given the similarity between the models we would expect similar results with other the published model.

Motif discovery in ATAC-seq peaks of differential accessibility

de novo motifs were identified using the findMotifsGenome command in Homer (Heinz et al. 2010) (4.7.2). The background for this analysis was all consensus ATAC-seq peaks as these were the all regions considered when calling peaks of differential accessibility. Motif sizes were limited to 6, 8, 10, or 12bp via "-len", and *de novo* motifs compared to all motifs in the Homer database. Other non-default settings were: -size given -bits. By

default, Homer compares the *de novo* motifs to known TF motifs; we report only the TF family as almost all of the known motifs were outside of *C. elegans.*

To identify known *C. elegans* motifs, all experimentally-derived motifs from cisBP (v1.02) (Ray et al. 2013) were provided as known motifs (via "-mknown") to findMotifsGenome in Homer, using a stringent log odds score of 9 for every motif. This score controls the stringency with which motif matches are made, a lower score allows for degenerate motifs to count as matches, and we found 9 to empirically balance accuracy and stringency.

Predicting accessibility changes with motifs

To predict changes in accessibility between early embryo and larval stage 3, as determined with DiffBind (see above), the number of each mapped *C. elegans* motifs from cisBP (see above) was counted to create a matrix of 166 motif counts and 30,832 ATAC-seq peaks. The ATAC-seq peaks were then split into two groups: a training set (70%) which the subsequent model was built upon, and a testing-set (30% of all ATAC-seq peaks) which was used to verify the accuracy of the model. We next tried several different classification models to prediction whether peaks A) lost accessibility between early embryo and L3, B) stayed consistent between the stages, or C) gained accessibility during development. We found that a generalized boosting model (GBM) (Ridgeway 2015) performed the best, while still allowing for interpretation of which motifs were the most informative. Given the unbalanced classification problem (~60% of peaks were unchanged, and approximately 20% gained and lost accessibility, respectively) we used balanced accuracy (average of sensitivity and specificity, or the average accuracy of

predicting dynamic peaks and static peaks) as our primary metric of classification success. We accurately predicted more than 41.6% of the peaks that increased in accessibility between early embryo and L3. This conservative, yet accurate approach resulted in a balanced accuracy of approximately 0.7 (balanced accuracy is more appropriate for the unbalanced nature of these samples, and has an expected value of 0.5) (Supplemental Fig. 5D). Parameters for the GBM were optimized using the R package, caret (Kuhn 2015), to run 10-fold cross validation; the following parameters were optimized: interaction depth (how many levels there are in the individual trees): 2, 5, 8 or 11 and the number of trees to use per model: 6,000, 10,000, or 14,000. Otherwise default settings were used. Using the fit or prediction from this model every peak in each set, training and testing, respectively, was split into 3 classes, decreased accessibility during development, consistent, or increased accessibility.

An important aspect of the model selected is that it allows for interpretation of which motifs were the most influential or important in predicting the changes in accessibility. These values are arbitrary, but relative within the model (i.e. a motif with a score of 10 was twice as informative in predicting chromatin accessibility changes as a motif with a score of 5).

Conservation calculations

As a measure of conservation, we downloaded the PhastCons 7-way track from the UCSC Genome browser (Rosenbloom et al. 2015). This track assigns a score at the single base pair level using the conservation between seven nematodes. The score ranges from 0 to 1, where a higher score indicates better conservation. To avoid biases, unmappable

regions (described above), all blacklist regions (described above), and all ATAC-seq gDNA peaks were excluded from further analysis. In addition, Ensembl protein-coding exons (Rosenbloom et al. 2015) were excluded from further analysisto focus on non-coding conservation, a hallmark of regulatory regions. In order to focus on distal regulatory regions, we also excluded all regions 1 kb upstream and 0.5 kb downstream of TSSs. The remaining portion of the genome we refer to as the distal non-coding genome.

To calculate a conservation score for a set of regions (e.g. ATAC-seq peaks), the median single base pair conservation score was calculated for every region in which at least half of the region was included in the distal non-coding genome described above. To assess the significance of any such result, we generated a null distribution for every set of regions. To do this, we selected only those regions at least half within the distal non-coding genome, and randomized the location of the peaks requiring that they were at least half within the distal non-coding genome. We repeated this 10,000 times and calculated the single base pair median phastCons score for each region every time.

Chromatin state changes between stages

To identify which transitions in chromHMM-predicted chromatin states that our dynamic ATAC-seq peaks were enriched for, we split consensus ATAC peaks into 3 classes: more accessible in EE, more accessible in L3 and unchanged (from above). For each set, we annotated each peaks' early embryo chromHMM predicted chromatin state by requiring at least 50% of the ATAC-seq peak to overlap with the state. This split each of the 3 sets of ATAC-seq peaks into 7 classes (one for each chromatin state). Then for each of those classes in each set, we annotated the L3 chromHMM-predicted chromatin state as above.

The end result of this process was a 7x7 matrix corresponding to the 7 chromatin states for each set of ATAC-seq peaks. These were converted to portions in a row-wise manner (i.e. each row summed to 1), and then log2 enrichments calculated for the dynamic ATAC-seq peaks versus the consistent ATAC-seq peaks.

GO term enrichment

To identify the underlying biological process marked by chromatin accessibility dynamics, every ATAC-seq peak was associated with the nearest TSS, up to 10kb away. Genes were then ordered by the total change in ATAC-seq signal in their associated ATAC-seq peaks between the two stages being compare (e.g. early embryo versus larval stage 3). This approach thus takes into account both the number of ATAC-seq peaks and the intensity of signal in the peaks associated with genes. Gene Ontology enrichments were then calculated with the online GO program, GOrilla (Eden et al. 2009), using the single ranked list setting, and the fast mode was not used to optimize accuracy.

Heatmap plots

All read-based heatmaps were generated with NGS Plot (Shen et al. 2014) with default settings except for the color distribution (-CD), which was set to 1, thereby centering the color scale on 0. For ATAC-seq, pooled-replicate BAM (alignment) files of single base pair insert sites for each experimental stage versus the input control were used; fragment size was empirically set at 25bp. For histone modification ChIP-seq, pooled-replicate BAM files of the reads and input controls were used.

Insert size calculation

Calculations and plots were generated with Picard Tools from the quality-filtered aligned ATAC-seq reads.

Calculation of genomic loci overlap

The Jaccard index was calculated using Bedtools Jaccard; specifically: (length of intersection of 2 sets of genomic loci (bp)) / ((length of union) – (length of intersection)).

Data Access

Raw reads as well as stage-specific peaks can be found on the Gene Expression Omnibus website using accession GSE89608.

Acknowledgements

We thank Dr. Elizabeth Noblin and Max Lenail for their help imaging enhancer reporter strains. We thank Dr. Elizabeth Noblin, Dr. Lauren Booth, and Dr. Bérénice Benayoun for critically reading the manuscript. We thank Dr. Bérénice Benayoun for feedback on computational pipeline and analysis, and Dr. Andy Fire and Dr. Joanna Wysocka for suggestions on the project. Supported by NIH DP1 AG044848 (A.B.), a seed grant from Stanford School of Medicine (A.B.), and NSF graduate fellowship (A.C.D.).

Author contributions

A.C.D. and A.B. planned the study. A.C.D. performed the ATAC-seq experiments, designed the analytical pipeline, analyzed and interpreted the data, and wrote the

manuscript with help from A.B.. R.Y. generated the transgenic enhancer reporter lines, imaged and analyzed GFP expression, and contributed to paper writing. A.K. provided intellectual guidance with ATAC-seq analysis and machine learning. J.D.B. and W.J.G. provided early access to ATAC-seq protocols and feedback on the design. All authors discussed the results and commented on the manuscript.

Disclosure Declaration

The authors declare no conflict of interest.

References

Adkins NL, Hagerman TA, Georgel P. 2006. GAGA protein: a multi-faceted transcription factor. *Biochemistry and cell biology = Biochimie et biologie cellulaire* **84**(4): 559-567.

Antebi A, Culotti JG, Hedgecock EM. 1998. daf-12 regulates developmental age and the dauer alternative in Caenorhabditis elegans. *Development* **125**(7): 1191-1205.

Antebi A, Yeh WH, Tait D, Hedgecock EM, Riddle DL. 2000. daf-12 encodes a nuclear receptor that regulates the dauer diapause and developmental age in C. elegans. *Genes & development* **14**(12): 1512-1527.

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**(7515): 400-405.

Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**(6123): 1074-1077.

Attrill H, Falls K, Goodman JL, Millburn GH, Antonazzo G, Rey AJ, Marygold SJ, FlyBase C. 2016. FlyBase: establishing a Gene Group resource for Drosophila melanogaster. *Nucleic acids research* **44**(D1): D786-792.

Bonn S, Zinzen RP, Girardot C, Gustafson EH, Perez-Gonzalez A, Delhomme N, Ghavi-Helm Y, Wilczynski B, Riddell A, Furlong EE. 2012. Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nature genetics* **44**(2): 148-156.

Boyle AP, Araya CL, Brdlik C, Cayting P, Cheng C, Cheng Y, Gardner K, Hillier LW, Janette J, Jiang L et al. 2014. Comparative analysis of regulatory information and circuits across distant species. *Nature* **512**(7515): 453-456.

Broad Institute. 2014. Picard Tools.

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. *Nature methods* **10**(12): 1213-1218.
Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, Chang HY, Greenleaf WJ. 2015. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**(7561): 486-490.
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 research* **23**(8): 1339-1347.
Contrino S, Smith RN, Butano D, Carr A, Hu F, Lyne R, Rutherford K, Kalderimis A, Sullivan J, Carbon S et al. 2012. modMine: flexible access to modENCODE data. *Nucleic acids research* **40**(Database issue): D1082-1088.
Corces MR, Buenrostro JD, Wu B, Greenside PG, Chan SM, Koenig JL, Snyder MP, Pritchard JK, Kundaje A, Greenleaf WJ et al. 2016. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature genetics*.
Crane E, Bian Q, McCord RP, Lajoie BR, Wheeler BS, Ralston EJ, Uzawa S, Dekker J, Meyer BJ. 2015. Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature* **523**(7559): 240-244.
Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, Steemers FJ, Trapnell C, Shendure J. 2015. Epigenetics. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**(6237): 910-914.
Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. 2009. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC bioinformatics* **10**: 48.
Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nature methods* **9**(3): 215-216.
Evans KJ, Huang N, Stempor P, Chesney MA, Down TA, Ahringer J. 2016. Stable Caenorhabditis elegans chromatin domains separate broadly expressed and developmentally regulated genes. *Proceedings of the National Academy of Sciences of the United States of America*.
Furey TS. 2012. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nature reviews Genetics* **13**(12): 840-852.
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**(6012): 1775-1787.
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 research* **17**(6): 877-885.
Gissendanner CR, Sluder AE. 2000. nhr-25, the Caenorhabditis elegans ortholog of ftz-f1, is required for epidermal and somatic gonad development. *Developmental biology* **221**(1): 259-272.
Haenni S, Ji Z, Hoque M, Rust N, Sharpe H, Eberhard R, Browne C, Hengartner MO, Mellor J, Tian B et al. 2012. Analysis of C. elegans intestinal gene expression and

polyadenylation by fluorescence-activated nuclei sorting and 3'-end-seq. *Nucleic acids research* **40**(13): 6304-6318.
Hahne F, Ivanek R. 2016. Visualizing Genomic Data Using Gviz and Bioconductor. *Methods in molecular biology* **1418**: 335-351.
Harfe BD, Vaz Gomes A, Kenyon C, Liu J, Krause M, Fire A. 1998. Analysis of a Caenorhabditis elegans Twist homolog identifies conserved and divergent aspects of mesodermal patterning. *Genes & development* **12**(16): 2623-2635.
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. *Nature genetics* **42**(4): 343-347.
Hecker A, Brand LH, Peter S, Simoncello N, Kilian J, Harter K, Gaudin V, Wanke D. 2015. The Arabidopsis GAGA-Binding Factor BASIC PENTACYSTEINE6 Recruits the POLYCOMB-REPRESSIVE COMPLEX1 Component LIKE HETEROCHROMATIN PROTEIN1 to GAGA DNA Motifs. *Plant physiology* **168**(3): 1013-1024.
Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell* **38**(4): 576-589.
Heinz S, Romanoski CE, Benner C, Glass CK. 2015. The selection and function of cell type-specific enhancers. *Nature reviews Molecular cell biology* **16**(3): 144-154.
Ho JW, Jung YL, Liu T, Alver BH, Lee S, Ikegami K, Sohn KA, Minoda A, Tolstorukov MY, Appert A et al. 2014. Comparative analysis of metazoan chromatin organization. *Nature* **512**(7515): 449-452.
Howard RM, Sundaram MV. 2002. C. elegans EOR-1/PLZF and EOR-2 positively regulate Ras and Wnt signaling and function redundantly with LIN-25 and the SUR-2 Mediator component. *Genes & development* **16**(14): 1815-1827.
Hsu HT, Chen HM, Yang Z, Wang J, Lee NK, Burger A, Zaret K, Liu T, Levine E, Mango SE. 2015. TRANSCRIPTION. Recruitment of RNA polymerase II by the pioneer transcription factor PHA-4. *Science* **348**(6241): 1372-1376.
Jantsch-Plunger V, Fire A. 1994. Combinatorial structure of a body muscle-specific transcriptional enhancer in Caenorhabditis elegans. *The Journal of biological chemistry* **269**(43): 27021-27028.
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. *Nature genetics* **43**(3): 264-268.
Kharchenko PV, Tolstorukov MY, Park PJ. 2008. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature biotechnology* **26**(12): 1351-1359.
Kolde R. 2013. pheatmap: Pretty Heatmaps.
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. *Developmental biology* **166**(1): 133-148.
Kuhn M, Contributions from Jed Wing and Steve Weston and Andre Williams and Chris Keefer and Allan Engelhardt and Tony Cooper and Zachary Mayer and Brenton Kenkel and the R Core Team and Michael Benesty and Reynald Lescarbeau and

Andrew Ziem and Luca Scrucca,. 2015. caret: Classification and Regression Training.
Kvon EZ. 2015. Using transgenic reporter assays to functionally characterize enhancers in animals. *Genomics* **106**(3): 185-192.
Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**(4): 357-359.
Lara-Astiaso D, Weiner A, Lorenzo-Vivas E, Zaretsky I, Jaitin DA, David E, Keren-Shaul H, Mildner A, Winter D, Jung S et al. 2014. Immunogenetics. Chromatin state dynamics during blood formation. *Science* **345**(6199): 943-949.
Leek JT JW, Parker HS, Fertig EJ, Jaffe AE and Storey JD. 2015. sva: Surrogate Variable Analysis.
Lehner B, Crombie C, Tischler J, Fortunato A, Fraser AG. 2006. Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways. *Nature genetics* **38**(8): 896-903.
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**(8): 1241-1249.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14): 1754-1760.
Liu T, Rechtsteiner A, Egelhofer TA, Vielle A, Latorre I, Cheung MS, Ercan S, Ikegami K, Jensen M, Kolasinska-Zwierz P et al. 2011. Broad chromosomal domains of histone modification patterns in C. elegans. *Genome research* **21**(2): 227-236.
Lund E, Oldenburg AR, Delbarre E, Freberg CT, Duband-Goulet I, Eskeland R, Buendia B, Collas P. 2013. Lamin A/C-promoter interactions specify chromatin state-dependent transcription outcomes. *Genome research* **23**(10): 1580-1589.
Mahmoudi T, Katsani KR, Verrijzer CP. 2002. GAGA can mediate enhancer function in trans by linking two separate DNA molecules. *The EMBO journal* **21**(7): 1775-1781.
Mello C, Fire A. 1995. DNA transformation. *Methods in cell biology* **48**: 451-482.
Minnich M, Tagoh H, Bonelt P, Axelsson E, Fischer M, Cebolla B, Tarakhovsky A, Nutt SL, Jaritz M, Busslinger M. 2016. Multifunctional role of the transcription factor Blimp-1 in coordinating plasma cell differentiation. *Nature immunology* **17**(3): 331-343.
Mito Y, Henikoff JG, Henikoff S. 2007. Histone replacement marks the boundaries of cis-regulatory domains. *Science* **315**(5817): 1408-1411.
Narasimhan K, Lambert SA, Yang AW, Riddell J, Mnaimneh S, Zheng H, Albu M, Najafabadi HS, Reece-Hoyes JS, Fuxman Bass JI et al. 2015. Mapping and analysis of Caenorhabditis elegans transcription factor sequence specificities. *eLife* **4**.
Ohtsuki S, Levine M. 1998. GAGA mediates the enhancer blocking activity of the eve promoter in the Drosophila embryo. *Genes & development* **12**(21): 3325-3330.
Okada M, Hirose S. 1998. Chromatin remodeling mediated by Drosophila GAGA factor and ISWI activates fushi tarazu gene transcription in vitro. *Molecular and cellular biology* **18**(5): 2455-2461.
Ong CT, Corces VG. 2014. CTCF: an architectural protein bridging genome topology and function. *Nature reviews Genetics* **15**(4): 234-246.

Parker S, Baylis HA. 2009. Overexpression of caveolins in Caenorhabditis elegans induces changes in egg-laying and fecundity. *Communicative & integrative biology* **2**(5): 382-384.
Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**(7118): 499-502.
Petrascheck M, Escher D, Mahmoudi T, Verrijzer CP, Schaffner W, Barberis A. 2005. DNA looping induced by a transcriptional enhancer in vivo. *Nucleic acids research* **33**(12): 3743-3750.
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**(6): 841-842.
R Core Team. 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Raff JW, Kellum R, Alberts B. 1994. The Drosophila GAGA transcription factor is associated with specific regions of heterochromatin throughout the cell cycle. *The EMBO journal* **13**(24): 5977-5983.
Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A et al. 2013. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**(7457): 172-177.
Reinke V, Krause M, Okkema P. 2013. Transcriptional regulation of gene expression in C. elegans. *WormBook : the online review of C elegans biology*: 1-34.
Ren B, Yue F. 2015. Transcriptional Enhancers: Bridging the Genome and Phenome. *Cold Spring Harbor symposia on quantitative biology*.
Ridgeway G. 2015. gbm: Generalized Boosted Regression Models.
Riedel CG, Dowen RH, Lourenco GF, Kirienko NV, Heimbucher T, West JA, Bowman SK, Kingston RE, Dillin A, Asara JM et al. 2013. DAF-16 employs the chromatin remodeller SWI/SNF to promote stress resistance and longevity. *Nature cell biology* **15**(5): 491-501.
Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M et al. 2015. The UCSC Genome Browser database: 2015 update. *Nucleic acids research* **43**(Database issue): D670-681.
Rosenbloom KR, Sloan CA, Malladi VS, Dreszer TR, Learned K, Kirkup VM, Wong MC, Maddren M, Fang R, Heitner SG et al. 2013. ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic acids research* **41**(Database issue): D56-63.
Schoborg T, Labrador M. 2014. Expanding the roles of chromatin insulators in nuclear architecture, chromatin organization and genome function. *Cellular and molecular life sciences : CMLS* **71**(21): 4089-4113.
Sharov AA, Dudekula DB, Ko MS. 2006. CisView: a browser and database of cis-regulatory modules predicted in the mouse genome. *DNA research : an international journal for rapid publication of reports on genes and genomes* **13**(3): 123-134.
Shen L, Shao N, Liu X, Nestler E. 2014. ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC genomics* **15**: 284.

Simon JM, Hacker KE, Singh D, Brannon AR, Parker JS, Weiser M, Ho TH, Kuan PF, Jonasch E, Furey TS et al. 2014. Variation in chromatin accessibility in human kidney cancer links H3K36 methyltransferase loss with widespread RNA processing defects. *Genome research* **24**(2): 241-250.

Srivastava S, Puri D, Garapati HS, Dhawan J, Mishra RK. 2013. Vertebrate GAGA factor associated insulator elements demarcate homeotic genes in the HOX clusters. *Epigenetics & chromatin* **6**(1): 8.

Stark R, Brown, G 2011. DiffBind: differential binding analysis of ChIP-Seq peak data.

Stergachis AB, Neph S, Reynolds A, Humbert R, Miller B, Paige SL, Vernot B, Cheng JB, Thurman RE, Sandstrom R et al. 2013. Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. *Cell* **154**(4): 888-903.

Sulston JE, Schierenberg E, White JG, Thomson JN. 1983. The embryonic cell lineage of the nematode Caenorhabditis elegans. *Developmental biology* **100**(1): 64-119.

Sun Y, Nien CY, Chen K, Liu HY, Johnston J, Zeitlinger J, Rushlow C. 2015. Zelda overcomes the high intrinsic nucleosome barrier at enhancers during Drosophila zygotic genome activation. *Genome research* **25**(11): 1703-1714.

Suzuki R, Shimodaira, Hidetoshi 2014. pvclust: Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling.

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 biology* **12**(5): R43.

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

Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, Zeng K, Malek JA, Costa G, McKernan K et al. 2008. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome research* **18**(7): 1051-1063.

Vavouri T, Walter K, Gilks WR, Lehner B, Elgar G. 2007. Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans. *Genome biology* **8**(2): R15.

Visel A, Rubin EM, Pennacchio LA. 2009. Genomic views of distant-acting enhancers. *Nature* **461**(7261): 199-205.

Wang YM, Zhou P, Wang LY, Li ZH, Zhang YN, Zhang YX. 2012. Correlation between DNase I hypersensitive site distribution and gene expression in HeLa S3 cells. *PloS one* **7**(8): e42414.

Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh TY, Peng W, Zhang MQ et al. 2008. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature genetics* **40**(7): 897-903.

West JA, Cook A, Alver BH, Stadtfeld M, Deaton AM, Hochedlinger K, Park PJ, Tolstorukov MY, Kingston RE. 2014. Nucleosomal occupancy changes locally over key regulatory regions during cell differentiation and reprogramming. *Nature communications* **5**: 4719.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome biology* **9**(9): R137.

Zhu B, Zhang W, Zhang T, Liu B, Jiang J. 2015. Genome-Wide Prediction and Validation of Intergenic Enhancers in Arabidopsis Using Open Chromatin Signatures. *The Plant cell* **27**(9): 2415-2426.

Zhu Y, Sun L, Chen Z, Whitaker JW, Wang T, Wang W. 2013. Predicting enhancer transcription and activity from chromatin modifications. *Nucleic acids research* **41**(22): 10032-10043.