nature genetics

Differential chromatin marking of introns and expressed exons by H3K36me3

Paulina Kolasinska-Zwierz[1], Thomas Down[1], Isabel Latorre[1], Tao Liu[2], X Shirley Liu[2,3] & Julie Ahringer[1]

Variation in patterns of methylations of histone tails reflects and modulates chromatin structure and function[1]. To provide a framework for the analysis of chromatin function in *Caenorhabditis elegans*, we generated a genome-wide map of histone H3 tail methylations. We find that *C. elegans* genes show distributions of histone modifications that are similar to those of other organisms, with H3K4me3 near transcription start sites, H3K36me3 in the body of genes and H3K9me3 enriched on silent genes. We also observe a novel pattern: exons are preferentially marked with H3K36me3 relative to introns. H3K36me3 exon marking is dependent on transcription and is found at lower levels in alternatively spliced exons, supporting a splicing-related marking mechanism. We further show that the difference in H3K36me3 marking between exons and introns is evolutionarily conserved in human and mouse. We propose that H3K36me3 exon marking in chromatin provides a dynamic link between transcription and splicing.

Chromatin regulation has been studied in a variety of systems, but most extensively in unicellular yeasts and mammalian cells. *C. elegans* has many features that make it well suited as an alternative system for studies of chromatin regulation. Of particular note are its well-annotated genome, the ease of RNAi, and the rich resource of chromatin mutants for loss-of-function studies[2–4]. Importantly, *C. elegans* has a complement of chromatin factors very similar to that of humans, in contrast to yeast[5], and allows investigations of chromatin function in a multicellular organism[6,7]. Because modifications to histone tails are correlated with and can regulate chromatin structure[1], we decided to map their positions on a genome-wide scale, to provide a framework for chromatin studies in *C. elegans*.

To generate an initial map of the distributions of histone methylations across the *C. elegans* genome, we used chromatin immunoprecipitation (ChIP) followed by microarray hybridization to determine the genome-wide association of trimethylation of lysine 4, lysine 9 and lysine 36 of histone H3 (H3K4me3, H3K9me3 and H3K36me3). We prepared chromatin extracts from highly synchronized triplicate wild-type worms at the third larval stage and carried out chromatin immunoprecipitations using commercial antibodies (see Methods). Immunoprecipitated DNA was amplified and hybridized to 2.1 million feature full-genome tiling microarrays (Roche). Pairwise comparisons of same antibody ChIPs showed strong correlation between replicate data (**Supplementary Table 1** online), and the three replicates showed similar enrichment patterns across different genomic regions (**Fig. 1a**,**b**). To correct for differences in nucleosome occupancy, we subtracted the H3 mean ChIP signal from those of H3K4me3, H3K9me3 and H3K36me3 (see Methods). To investigate relationships between transcription and different histone modifications, we generated four sets of genes: (i) 'top10', those in the top 10% of expression level in our samples, determined by gene expression profiling; (ii) 'bottom10', those in the bottom 10% of expression level; (iii) 'ubiq', genes annotated or expected to be actively transcribed in all nuclei; and (iv) 'serp', serpentine receptor genes, most of which are thought to encode chemosensory receptors transcribed in only a few neurons and thus to be transcriptionally inactive in most nuclei[8].

To gain initial insight into gene regions enriched for different modifications, we plotted mean $\log_2$ ChIP signals across all genes. We aligned genes at the first and last nucleotides of the annotated transcripts and extended these regions 1 kb upstream and 1 kb downstream of genomic DNA (**Fig. 1c–e**). We call the first base of annotated transcript the TSS (transcript start site). As with other organisms[9–12], we observed a peak of H3K4me3 enrichment near the TSS that correlates with transcriptional activity (**Fig. 1d** and **Supplementary Fig. 1** online). Highly transcribed genes (ubiq and top10) showed strong 5′ enrichment of H3K4me3, but inactive genes (serp and bottom10) showed no enrichment.

In *C. elegans*, many genes are trans-spliced at their 5′ ends to a 21-bp leader sequence[13]. In these cases, the transcription start sites are not known because the 5′ end of the primary transcript is spliced off and degraded. In addition, some groups of genes are transcribed in operons, with trans-splicing separating transcripts from different genes. Spliced leader SL1 is found on genes adjacent to promoters and SL2 generally occurs on downstream operon genes not adjacent to promoters. To investigate the relationship between H3K4me3 and the transcription start site, we separated genes into SL1 genes and those not annotated to contain SL1 or SL2. We found a peak of H3K4me3 200 bp downstream of the presumed TSS (the first annotated base) for

[1]The Gurdon Institute and Department of Genetics, University of Cambridge, Tennis Court Road, Cambridge CB2 1QN, UK. [2]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, 44 Binney Street, Boston, Massachusetts 02115, USA. [3]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, USA. Correspondence should be addressed to J.A. (ja219@cam.ac.uk).

Received 24 October 2008; accepted 9 January 2009; published online 1 February 2009; doi:10.1038/ng.322

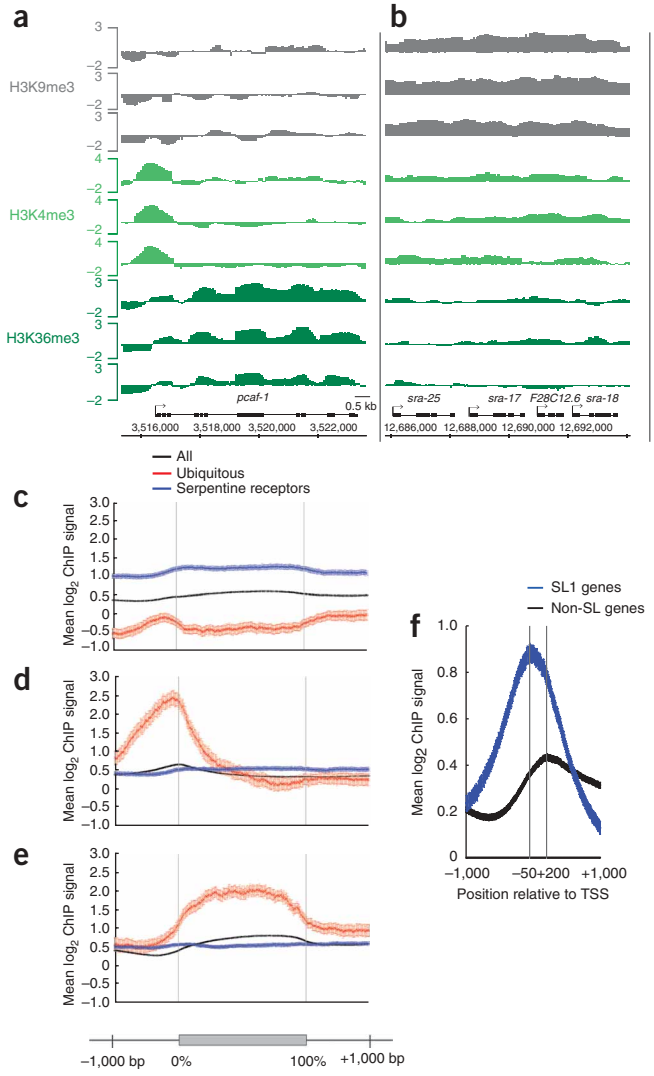

Figure 1 Patterns of histone methylations across *C. elegans* genes. (**a**,**b**) Mean $\log_2$ ChIP signal of normalized triplicate samples across a genomic region containing an actively transcribed gene (**a**) and a cluster of transcriptionally repressed genes (**b**). (**c**–**e**) H3K9me3 (**c**), H3K4me3 (**d**) and H3K36me3 (**e**) mean $\log_2$ ChIP signals across all genes (black), ubiq genes (red) and serp genes (blue) aligned at the first and last nucleotides (0% and 100%) and extended with 1 kb upstream and downstream sequence. (**f**) H3K4me3 mean $\log_2$ ChIP signals from −1000 to +1000 relative to TSSs (transcript start sites; defined as the first known nucleotide in the mature transcript, not including *trans*-spliced leaders). Blue, SL1 genes; black, genes not annotated to contain spliced leaders.

non-SL1 annotated genes (**Fig. 1f**). In contrast, the peak of H3K4me3 for SL1 genes occurs 50 bp upstream of the first annotated base. The H3K4me3 peak position suggests that the transcription start site for SL1 genes is on average 250 bp upstream of the trans-splice site. Peaks of H3K4me3 should prove a useful guide for identifying promoters of SL1 and non-SL1 genes.

We next looked at the genome-wide distribution of H3K9me3. This modification is generally associated with repressed chromatin[1]. In mammalian cells, H3K9me3 is enriched in repressed constitutive heterochromatin, repetitive DNA, DNA transposons and other repetitive elements[14,15]. Studies on small gene sets also detected H3K9me3 in the bodies of actively transcribed genes[16,17], but this does not seem to be a general property, judging from genome-wide studies[18,19]. In *C. elegans* chromatin, we found that H3K9me3 is highly enriched across inactive genes, covering promoters, transcribed regions and 3′ regions (blue line in **Fig. 1c** and **Supplementary Fig. 1**). In contrast, active genes show very low H3K9me3 signals (red line in **Fig. 1c** and **Supplementary Fig. 1**). Regions with clustered inactive genes often showed continuous H3K9me3 enrichment across and between genes (**Fig. 1b**).

In yeast and mammalian chromatin, there is a well-documented association of H3K36me3 with transcribed regions[1]. The Set2 histone methytransferase that catalyzes this modification is associated with elongating RNA polymerase II, and the modification is made co-transcriptionally[20–22]. There is evidence that one function of H3K36me3 in the gene body is to prevent aberrant transcription initiation[23,24]. We found that *C. elegans* genes also show high levels of H3K36me3 in gene bodies. The level of H3K36me3 is low at the 5′ end, increases to a plateau and then decreases at the 3′ end (**Fig. 1e** and **Supplementary Fig. 1**).

We observed that H3K36me3 signals often showed discrete peaks and troughs in the gene bodies, with peaks correlating with exonic regions (**Fig. 1a**). To explore whether this was a genome-wide phenomenon, we plotted H3K36me3 signals across aligned intron–exon and exon–intron boundaries and observed a prominent enrichment of H3K36me3 in exon regions compared to introns (**Fig. 2a–c** and **Supplementary Figs. 2a** and **3** online). In contrast, neither H3K4me3 nor H3K9me3 showed exon enrichments (**Fig. 2d–i** and **Supplementary Figs. 2b,c** and **3**). H3K36me3 exon enrichment is not due to GC bias, as exon signals are higher than those of introns across the whole range of percentage GC content (**Fig. 2j**). We observed high and level H3K36me3 signals across exons of different lengths and lower signal across introns (**Fig. 2b,c** and **Supplementary Fig. 3**).

We next asked whether H3K36me3 exon marking was dependent on transcription or was instead a constitutive feature of exons. We found that the highly expressed ubiq and top10 genes show a higher exon marking relative to all genes, whereas bottom10 and serp genes show low or no marking, respectively (**Fig. 2a** and **Supplementary Fig. 2a**). We conclude that exon marking is associated with transcription.

Because chromatin marking of exonic sequence with H3K36me3 depends on transcription and transcribed exons are spliced into mature transcripts, we wondered whether marking was related to the process of splicing. If so, then the chromatin of exons that are constitutively included in transcripts would be expected to have a higher level of H3K36me3 than alternatively spliced exons. To address this possibility, we assembled a set of exon trios where an alternative exon was flanked by two constitutive exons and compared H3K36me3 levels in the three exons (**Fig. 3**). The alternative and constitutive exons had similar GC contents (**Fig. 3g**). We also compared these trios to a control set of length-matched trios where all three exons were constitutively included. We found that alternative exons have significantly reduced H3K36me3 exon signals relative to their constitutive neighbors and to the matched control exons (**Fig. 3a**). In contrast, there was no difference between the sets of trios in levels of H3K4me3 or H3K9me3 (**Fig. 3b,c**). The reduction in H3K36me3 signal in alternative exons indicates that exon marking is related to splicing.

Although profiles of H3K36me3 have been extensively mapped in other organisms, exon marking has not been observed before. To ask whether this phenomenon is specific to *C. elegans* or alternatively might be widespread, we analyzed genome-wide data for mapping of H3K36me3 in mouse and human chromatin[18,19] (**Fig. 4**). These mapping data were generated by massively parallel sequencing rather

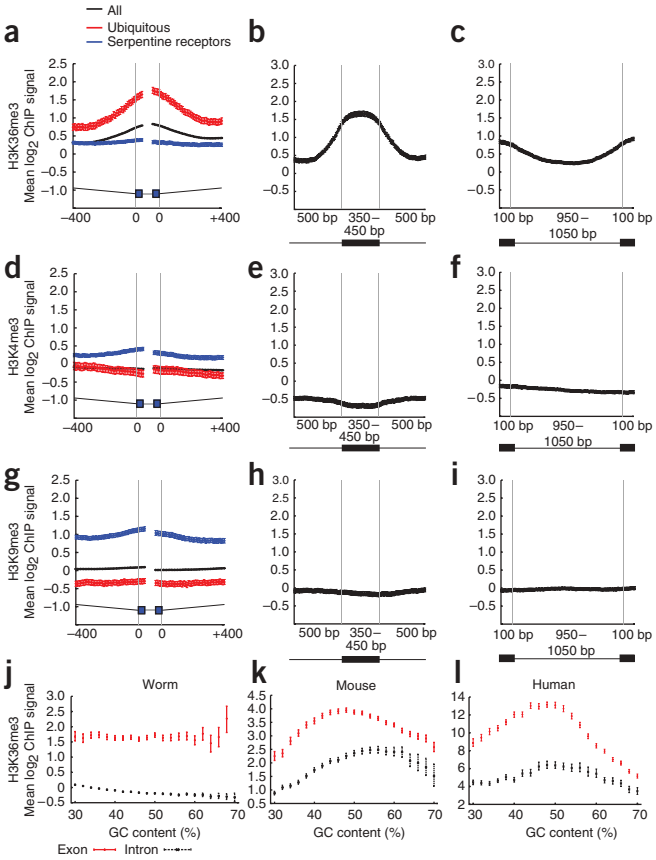

Figure 2 H3K36me3 is enriched across *C. elegans* exonic chromatin. (**a**–**i**) Mean $\log_2$ ChIP signals for H3K36me3 (**a**–**c**), H3K4me3 (**d**–**f**) and H3K9me3 (**g**–**i**). (**a**,**d**,**g**) Plot signals across intron–exon and exon–intron boundaries where introns are at least 400 bp and exons at least 80 bp, excluding first and last exons. Black, exons of all genes; red, ubiq genes; blue, serp genes. In the gene model, thin gray lines represent 400 bp of intron sequence, black boxes 40 bp of exon sequence and thick gray lines the centers of exons that vary in length. (**b**,**e**,**h**) Mean $\log_2$ ChIP signals across exons of length 350–450 bp flanked on both sides by introns of ≥500 bp. (**c**,**f**,**i**) Mean $\log_2$ ChIP signals across introns of length 950–1,050 bp. In diagrams below plots, black boxes represent exons and black lines introns. Exon enrichment of H3K36me3 relative to a neighboring intron was confirmed by qPCR of nonamplified ChIP material in 8 of 8 cases tested (not shown). (**j**–**l**) H3K36me3 signals for exons and introns according to GC content. In *C. elegans* (**j**), mouse (**k**) and human (**l**), exon signals are higher than those of introns at every percentage GC content, implicating that enrichment of H3K36me3 on exons is not due to GC bias.

than microarrays, providing a platform control. As in the *C. elegans* data, we found a strong enrichment of H3K36me3 in both mouse and human exons relative to introns (**Fig. 4a–c, g–i**). In contrast, we found essentially level signals for H3K4me3 (**Fig. 4d–f, j–l**). As a further control, we examined H3K27me1, found across active gene bodies like H3K36me3 (ref. 18), and found similar signals in exons and introns. (**Fig. 4m–o**). As in *C. elegans*, H3K36me3 exon enrichment is not due to GC bias (**Fig. 2k,l**). The H3K36me3 signal in long exons increases to a plateau, similar to the pattern in *C. elegans* exons (**Supplementary Fig. 3**). Across shorter exons more typical of human genes, H3K36me3 signal increases from 5′ to 3′ ends, resulting in an apparent peak near the 5′ splice site of the next intron (**Fig. 4b**). The lower H3K36me3 signals in introns increase near both the 5′ and 3′ splice sites (**Fig. 4c,i**).

The above analysis demonstrated that H3K36me3 exon marking is conserved in human and mouse. To explore whether marking in mammalian chromatin is likely to be related to splicing as it is in *C. elegans*, we used the mouse data[19] to ask whether alternative exons show reduced H3K36me3 signals relative to constitutive exons. Indeed, we found that mouse alternative exons have significantly lower H3K36me3 signals but no difference in levels of H3K4me3 (**Fig. 3d,e**). The GC contents of the alternative exons are also similar to those of the constitutive exons (**Fig. 3h**). We conclude that H3K36me3 marking of expressed exons is conserved.

What could be the function of H3K36me3 exon marking? Because constitutively expressed exons have higher marking than alternatively included ones, marking has a relationship with *cis*-splicing. There is increasing evidence that a significant amount of splicing occurs co-transcriptionally rather than post-transcriptionally, making interactions between chromatin and the splicing machinery plausible[25]. Indeed, although to our knowledge marking of exons in chromatin has not been observed previously, there are recent reports of chromatin factors having roles in splicing. For example, the H3K4me3 binding

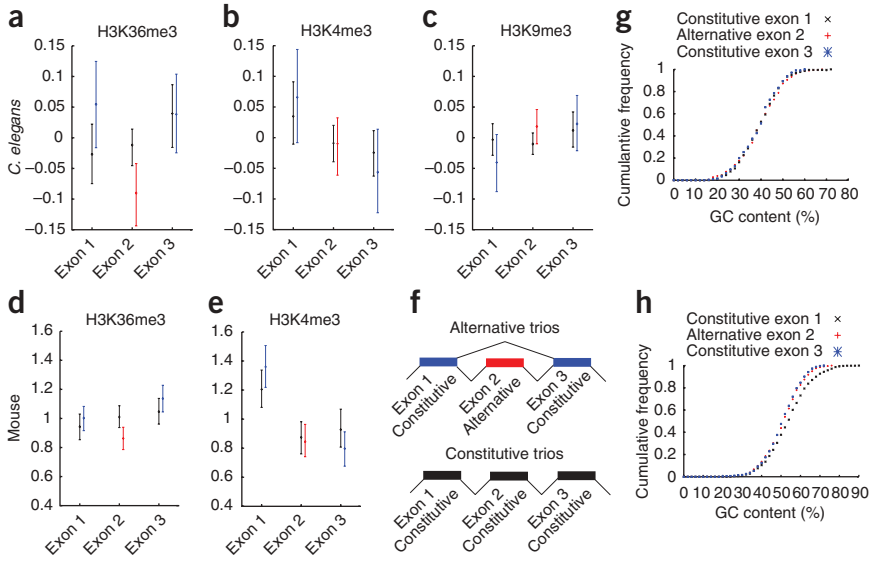

Figure 3 Alternative exons have lower H3K36me3 signal than constitutive exons. (**a**–**e**) Mean $\log_2$ ChIP signal of each *C. elegans* trio exon (**a**–**c**) or the average tag count for each mouse trio exon (**d**,**e**) is shown for the indicated histone modification. (**f**) Cartoon of alternative and constitutive trios of exons used in the analysis; colors correspond to data bars in **a**–**e**. Exon sizes in alternative and constitutive sets were length matched. In **a** and **d**, the alternative central exons (red) have significantly lower H3K36me3 signal than their matched constitutive central exons or than their constitutive neighbors ($P < 0.01$). The matched central exons do not show reduced H3K36me3 relative to their neighbors. There is also no significant difference in signals for the alternative exons compared to the matched exons for other histone modifications. The higher signals for exon 1 in **b** and **e** are due to closer proximity to the TSS. For *C. elegans* trios, $n = 54$ and for mouse, $n = 190$. Bars are 95% confidence intervals. (**g**,**h**) Cumulative GC content of constitutive and alternative exons in the alternative trios. (**g**) *C. elegans*. (**h**) Mouse. Alternative exons are similar in GC content to constitutive exons.

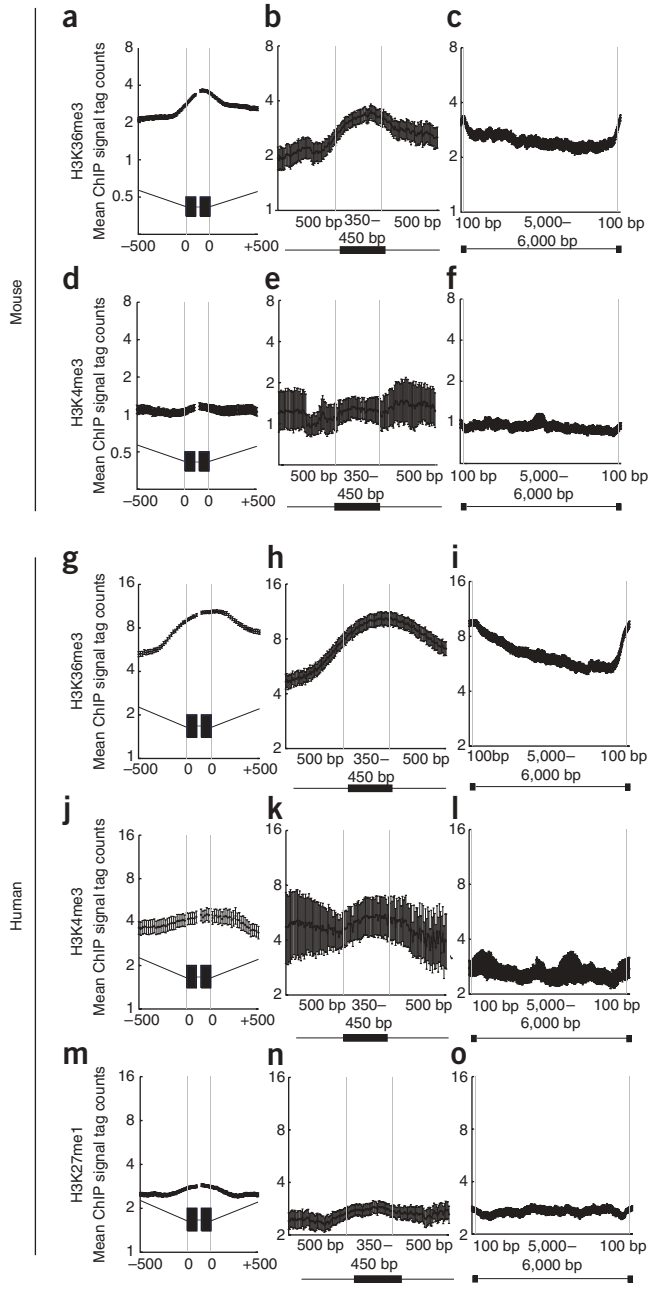

Figure 4 H3K36me3 is enriched across human and mouse exonic chromatin. (**a**–**o**) Average tag counts after ChIP from mouse MEF chromatin[19] (**a**–**f**) or human CD4$^+$ T cell chromatin[18] (**g**–**o**). (**a**–**c**, **g**–**i**) H3K36me3. (**d**–**f**, **j**–**l**) H3K4me3. (**m**–**o**) H3K27me1. First column, signals across intron–exon and exon–intron boundaries where introns are at least 500 bp and exons at least 200 bp. In the gene model, thin gray lines represent 500-bp intron sequence, black boxes 100-bp exon sequence and thick gray lines the centres of exons that vary in length. Middle column, mean ChIP tag counts across exons. Last column, mean ChIP tag counts across introns. Exons of length 350–450 bp and introns of length 5–6 kb were used, excluding the first 2 kb of the gene to eliminate promoter specific signals. In diagrams below plots, black boxes represent exons and black lines introns.

protein CHD1 is associated with the splicesome and required for high splicing efficiency[26]. In addition, splicing factors have been reported to associate with both chromatin and the RNA polymerase II complex[25] (Pol II). An attractive possibility is that marked exons in chromatin provide a mechanism to facilitate efficient splicing. For example, marked exons might aid recruitment of splicing factors to chromatin.

A second possibility is that the splicing machinery could regulate directly or indirectly K36 methyltransferases on the traveling RNA polymerase complex, such as Set2. If so, the composition of the traveling RNA polymerase complex might differ in exonic and intronic regions. For example, engagement in splicing reactions might reduce binding of splicing factors to Pol II. If these factors compete with or inhibit the H3K36me3 methyltransferase, this could result in regional differences in H3K36me3 on chromatin. In addition to Set2, other H3K36 methyltransferases exist and could potentially be involved in exon marking.

It is also known that the rate of RNA polymerase procession can vary over the gene and that changes in processivity can affect inclusion of alternative exons[27,28]. It would be interesting to investigate whether H3K36me3 affects processivity, which in turn could affect splicing. H3K36me3 is known to prevent spurious transcription initiation[23,24], so it could have a general inhibitory influence on Pol II complex activity.

METHODS

Extract preparation. Triplicate samples of synchronized worms at the third larval stage were prepared by growing starved first-larval-stage worms in liquid culture at 20 °C. Larvae were cleaned by sucrose flotation and flash frozen in liquid nitrogen. Frozen worms were ground to a fine powder and fixed in 1% formaldehyde/PBS for 10 min, quenched with 0.125 M glycine, and then washed 3× in PBS with protease inhibitors. The pellet was resuspended in 1ml of FA buffer (50 mM HEPES/KOH pH 7.5, 1 mM EDTA, 1% Triton X-100, 0.1% sodium deoxyholate, 150 mM NaCl with protease inhibitors) per 4 ml of ground worm powder. Extract was sonicated to a size range of 200–1,000 bp using a Diagenode Bioruptor at high setting for 14 pulses, each lasting 30 s followed by a 1 min pause. The extract was spun for 10 min at 16,000*g* at 4 °C, and the soluble fraction was flash frozen in liquid nitrogen and stored at −80 °C until use.

Chromatin immunoprecipitation and expression profiling. Each ChIP was prepared in 500 µl of FA buffer containing 1% sarkosyl. The following antibodies and extract amounts were used: anti-H3 (3 µg abcam 1791 with 300 µg extract); anti-H3K4me3 (5 µl Active Motif AR0169 serum with 300 µg extract); anti-H3K36me3 (3 µg abcam ab9050 with 1 mg extract); anti-H3K9me3 (3 µg Upstate 07-442 with 1 mg extract). Additionally, 10% of extract was saved as a reference. After overnight rotation at 4 °C, 40 µl of blocked and washed magnetic protein A dynabeads (Invitrogen) were added, and the incubation continued for 2 additional hours. Beads were washed at room temperature two times for 5 min in FA buffer, once in FA with 500 mM NaCl for 10 min, once in FA with 1 M NaCl for 5 min, once in TEL buffer (0.25 M LiCl, 1% NP-40, 1% sodium deoxyholate, 1 mM EDTA, 10 mM Tris-HCl, pH 8.0) for 10 min and two times in TE pH 8.0 for 5 min. DNA was eluted twice with 57 µl elution buffer (1% SDS in TE with 250 mM NaCl) at 65 °C, 15 min each time. Eluted DNA was incubated with 20 µg of RNase for 30 min at 37 °C and then with 20 µg of Proteinase K for 1 h at 55 °C. Input DNA was also diluted in 114 µl elution buffer and treated with ChIP samples. Crosslinks were reversed overnight at 65 °C. DNA was purified on Qiagen PCR purification columns and one-third was used for LM-PCR amplification[29]. Two rounds of 20 cycle amplification were carried out; 100 ng of the first round was used for the second round. 6.5 µg of each amplified DNA was used for hybridization to NimbleGen 2.1 million feature full-genome tiled *C. elegans* microarrays (Roche). MA2C software[30] was used to normalize chromatin immunoprecipitation microarray data and average replicates using the robust mean variance method where C = 2. We assessed concordance between ChIP replicates by calculating an overall Pearson correlation coefficient for every probe in each pairwise combination of same antibody ChIP (**Supplementary Table 1**). To correct for differences in nucleosome density, H3-normalized log ratios were subtracted from those of the modified histones (H3K4me3,