eLife 2016;5:e13884. doi: 10.7554/eLife.13884 Received: 12 Jan 2016 Accepted: 29 Mar 2016 Published: 30 Mar 2016

COVID-19 Information
Full-text HTML version updated 7 April 2020 | Research Services Office - SUNY Cortland Library, SRC | Processed on Wednesday, 08 April 2020 - 12:00 AM EDT

Genomics and Chromosomes

eLIFE Research Ecology and Evolution Genetics and Genomics

Keywords: DNA methylation, ZFP, KRAB-ZFP, TEs, Evolution, Epigenetics, Genomics, Mammalian genetics

DOI: 10.7554/eLife.13884.001

Received: 12 Jan 2016 | Accepted: 29 Mar 2016 | Published: 30 Mar 2016

Distinctive regulatory architectures of germline-active and somatic genes in C. elegans

Jonathan Mintzker, Alex Zaslaver, Lori Shemer, Michelle Cinkornpumin, Chaim Cassuto, Amir Levin-Kravitz*

Department of Molecular Genetics and Biotechnology, Faculty of Life Sciences, Ben-Gurion University of the Negev, Israel

*This author first made a major contribution to this work.

Abstract

Gene-specific gene expression patterns underlie production of the myriad of different cells generated during development. Although many regulatory factors and mechanisms have been elucidated, the overall regulatory architecture governing distinct gene expression patterns, and how it varies between different types of genes, remains largely unknown. Here, we dissect the DNA regulatory elements in germline-active and somatic genes in the nematode C. elegans. We show that germline-active genes have significantly longer and more numerous regulatory elements than somatic genes. Germline-active genes are also more likely to be bound by transcription factors and have more binding sites, denoting stronger gene expression potential. Accordingly, we find that germline-active and tissue-specific promoters have distinct sequence features, and that tissue-specific promoters are more modular than germline-active promoters. Finally, we show that ZFP-1/AHR-1 and ZFP-2/ZTF-8 function in a combinatorial manner to activate germline-specific genes, and that their joint binding is required for the proper expression of many germline-specific genes, which largely suffers functional ablation. We observe the ZFP-binding motifs enriched in germline-active promoters are also enriched in the regulatory regions of germline-specific genes in other nematodes, and that their combinatorial binding is conserved in germline-active and somatic tissues across species. Our results suggest that distinct regulatory strategies governing distinct gene expression patterns, and provide a tissue-specific resource for future studies.

Introduction

Gene-specific transcriptional regulation underlies production of the myriad of different cells generated during development. Although many regulatory factors and mechanisms have been elucidated, the overall regulatory architecture governing distinct gene expression patterns, and how it varies between different types of genes, remains largely unknown. Previous studies have provided important and biologically beneficial knowledge of features of transcription factor binding sites (TFBSs) and promoters that have been shown to affect gene expression, such as binding site number, strength, and specificity, and promoter sequence composition and modularity (Arnone and Davidson, 1997; Davidson et al, 2002; Kim and Small, 2007; Zinzen et al, 2009; Yáñez-Cuna et al, 2012; Zabidi et al, 2015; Vockenroeder et al, 2016). Recent in-vivo studies suggest that transcriptomic and developmental complexity may be partly attributed to differences in promoter architecture, such as the number of transcription factor binding sites, and transcriptional activity and fine-tuning of gene expression by distinct combinations of cis-regulatory elements (Zinzen et al, 2009; Yáñez-Cuna et al, 2012; Zabidi et al, 2015; Vockenroeder et al, 2016). Most of these studies focus on a few genes, or on a specific class of genes, such as developmental genes, or on a specific tissue, such as the mesoderm (Zinzen et al, 2009; Yáñez-Cuna et al, 2012; Zabidi et al, 2015; Vockenroeder et al, 2016). Yet, it remains unclear how these features are distributed across different classes of genes and how specific regulatory architecture governs gene expression patterns.

Results

Tissue-specific profiling of chromatin accessibility and gene expression in C. elegans L3 larvae

To investigate the regulatory elements of different cell types and how it relates to gene expression, we profiled chromatin accessibility and gene expression in germline-active and somatic tissues in C. elegans at the mid-L3 stage, using Assay for Transposase-Accessible Chromatin followed by sequencing (ATAC-seq) and Nematode-specific RNA-seq (NR-seq) (Jonathan et al, 2015; Jonathan et al, 2016). We focus on L3 stage because this developmental stage is characterized by the proliferation of germline and vulval precursor cells, allowing us to profile germline-active and somatic tissues in a single developmental stage. Moreover, this developmental stage is preceded by the establishment of most cell lineages, and the majority of cell divisions in the hermaphrodite gonad are mitotic, allowing us to profile germline-active cells without interference from meiotic chromatin. We performed ATAC-seq in two biological replicates, and NR-seq in three biological replicates, for wild-type (WT) and glp-4(bn2) ts mutant worms. The glp-4(bn2) ts mutant lacks germline-active cells, but exhibits normal development and gene expression in somatic tissues (Beanan and Strome, 1992; Jonathan et al, 2016). We expected that the mutant RNA-seq data would allow us to identify germline-active genes, and that the ATAC-seq data would allow us to identify germline-active and somatic-tissue-specific regulatory elements.

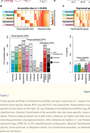

Figure 1

Tissue-specific profiling of chromatin accessibility and gene expression in C. elegans mutant L3 larvae. (**a**) Schematic of the experimental design. Germline-active and somatic tissues were profiled in wild-type (WT) and glp-4(bn2) ts mutant worms, respectively. Chromatin accessibility was assayed by ATAC-seq, and gene expression by Nematode-specific RNA-seq (NR-seq). Two biological replicates were performed for each condition. (**b**) Principal component analysis (PCA) of the ATAC-seq data. Each dot represents a single sample, and samples are colored according to the tissue and genotype. Germline-active samples cluster together, while somatic samples cluster together, forming two distinct groups. (**c**) Pairwise correlation heatmap of the ATAC-seq data. Samples are clustered according to Pearson's correlation coefficient. Distinct clusters of the germline-active and somatic samples are observed. (**d**) Genome-wide ATAC-seq signal profiles for the germline-active and somatic samples. Each row represents a 1 kilobase (kb) window, and the rows are sorted according to the difference between the germline-active and somatic signals. Germline-active-specific peaks, somatic-specific peaks, and shared peaks are observed. (**e**) Genome-wide NR-seq signal profiles for the germline-active and somatic samples. Each row represents a 1 kb window, and the rows are sorted according to the difference between the germline-active and somatic signals. Germline-active-specific, somatic-specific, and shared peaks are observed.

We generated ~40 M paired-ends chromatin assays (C. elegans Development and aging in vivo Jonathan et al, 2015; Jonathan et al, 2016) and ~20 M paired-ends NR-seq reads (Jonathan et al, 2016; Wormbase WS251) leading to retained and deduplicated of the mid-L3 sequence between the genomic and mitochondrial DNA, of ~10 M and ~5 M paired-ends ATAC-seq and NR-seq reads, respectively, for each sample (see Methods). Principal PCA analysis revealed a separation between the germline-active and somatic samples, with the first principal component (PC1) explaining most of the variance (Figure 1b). Pairwise Pearson correlation between samples within ATAC-seq chromatin accessibility profiles, and between samples within NR-seq gene expression profiles, also showed that the germline-active and somatic samples cluster together, forming two distinct groups (Figure 1c, Figure 1—figure supplement 1a). Indeed, we observed ~20% of the ATAC-seq peaks to be differentially enriched between the germline-active and somatic samples (FDR < 1%), and ~20% of the genes to be differentially expressed between the germline-active and somatic samples (FDR < 1%) (Figure 1d–e, Figure 1—figure supplement 1b–c, Figure 1—source data 1). Strikingly, we observe that the germline-active peaks are more numerous and broader than the somatic peaks, and that the germline-active genes are also more highly expressed than the somatic genes (Figure 1d–e, Figure 1—figure supplement 1d–e). Thus, we find that the germline-active and somatic tissues exhibit distinct chromatin accessibility and gene expression profiles, and that the germline-active chromatin and genes are more open and highly expressed than the somatic chromatin and genes.

Germline-active and soma-restricted genes have distinctive regulatory architectures

We next asked whether the observed differences between germline-active and somatic tissues could be explained by differences in the regulatory elements associated with germline-active and somatic genes. To this end, we first assigned ATAC-seq peaks to the closest gene, and subsequently categorized genes associated with peaks derived from each tissue. Given ATAC-seq peaks may be associated with more than one gene, we also required that the peak be closer to the gene assigned than to a gene2 assigned to any other gene. Therefore, the distance to the assigned gene provides a measure of regulatory potential, by each individual regulatory element (see Methods). We find that germline-active genes are associated with significantly longer distances to their closest ATAC-seq peak than soma-restricted genes (Figure 2a, Figure 2—figure supplement 1a–b). Moreover, we observe that germline-active genes are associated with more ATAC-seq peaks than soma-restricted genes, and that the number of peaks associated with a gene is positively correlated with the distance to the closest peak (Figure 2—figure supplement 1c–d). Thus, we find that germline-active genes are associated with more and more distant regulatory elements than soma-restricted genes.

Figure 2

Regulatory architecture of germline-active and soma-restricted genes from distinct chromosomal contexts. (**a**) Germline-active genes are associated with significantly longer distances to their closest ATAC-seq peak than soma-restricted genes (Wilcoxon rank-sum test, FDR < 1%). (**b**) Germline-active genes are associated with more ATAC-seq peaks than soma-restricted genes (Fisher's exact test, FDR < 1%). (**c**) Germline-active genes are associated with more distant ATAC-seq peaks than soma-restricted genes (Fisher's exact test, FDR < 1%). (**d**) Schematic of chromosome domains for gene-rich chromosomes I and II. (**e**) Germline-active genes are associated with significantly longer distances to their closest ATAC-seq peak than soma-restricted genes on the gene-rich chromosomes (Wilcoxon rank-sum test, FDR < 1%). (**f**) Germline-active genes are associated with more ATAC-seq peaks than soma-restricted genes on the gene-rich chromosomes (Fisher's exact test, FDR < 1%). (**g**) Germline-active genes are associated with more distant ATAC-seq peaks than soma-restricted genes on the gene-rich chromosomes (Fisher's exact test, FDR < 1%).

As expected, most ATAC-dependent expressed genes with at least one classified promoter are associated with a ubiquitously active promoter (Supplementary Table 1b). We observed that 84.7% of the germline-active genes and 71.4% of the soma-restricted genes are associated with a ubiquitously active promoter, suggesting that most of these genes are broadly expressed. Nevertheless, we observe that the germline-active genes are more likely than the soma-restricted genes to be associated with a germline-active promoter (Fisher's exact test, FDR < 1%), and that the single promoter genes are enriched for functions in germline processes, whereas those associated with ubiquitously active promoters are enriched for housekeeping functions (Figure 3—figure supplement 1a–c). Moreover, we find that the germline-active promoters are more A/T-rich, have more predicted transcription factor binding sites, and are more conserved than the ubiquitously active promoters, and that these features are even more pronounced when only single promoter genes are considered (Figure 3—figure supplement 1d–h). Similarly, tissue-specific promoters are also more A/T-rich than ubiquitously active promoters, and are more conserved than the ubiquitously active promoters, and these features are more pronounced when only single promoter genes are considered (Figure 3—figure supplement 2a–e). Thus, we find that the single promoter genes with germline-active or tissue-specific promoters have more distinctive sequence features than those with ubiquitously active promoters, and that these features are more pronounced when only single promoter genes are considered.

Figure 3

Regulatory architecture of ubiquitous, germline- and soma-restricted genes from distinct chromosomal contexts. (**a**) Schematic of the ubiquitous, germline- and soma-restricted promoter definitions. (**b**) Ubiquitously active promoters are associated with more ATAC-seq peaks than germline- and soma-restricted promoters (Fisher's exact test, FDR < 1%). (**c**) Ubiquitously active promoters are associated with more distant ATAC-seq peaks than germline- and soma-restricted promoters (Fisher's exact test, FDR < 1%). (**d**) Ubiquitously active promoters are associated with more ATAC-seq peaks than germline- and soma-restricted promoters on the gene-rich chromosomes (Fisher's exact test, FDR < 1%). (**e**) Ubiquitously active promoters are associated with more distant ATAC-seq peaks than germline- and soma-restricted promoters on the gene-rich chromosomes (Fisher's exact test, FDR < 1%).

Finally, we identified tissue-specific ATAC-seq peaks that were associated with expression of soma-restricted genes (see Methods). Excluding elements and genes with the top 5% of observed samples, we find that the soma-restricted ATAC-seq peaks are closer to their associated gene than the germline-active ATAC-seq peaks (Figure 2—figure supplement 1e–f). Moreover, we find that the soma-restricted ATAC-seq peaks are more A/T-rich than the germline-active ATAC-seq peaks (Figure 2—figure supplement 1g–h). Thus, we find that the soma-restricted ATAC-seq peaks are more distinctive than the germline-active ATAC-seq peaks.

We observed that the chromosomal neighborhoods of germline-active genes in C. elegans have a preference for the center of the chromosome, while soma-restricted genes have a preference for the arms, as previously observed (Ercan et al, 2007; Ercan et al, 2008; Ercan et al, 2009; Jonathan et al, 2015; Jonathan et al, 2016). Moreover, we find that the ATAC-seq peaks associated with germline-active genes are more enriched in the chromosome centers than the ATAC-seq peaks associated with soma-restricted genes (Figure 2d–g, Figure 2—figure supplement 1i–l). Thus, we find that the germline-active genes and their associated regulatory elements are more enriched in the chromosome centers than the soma-restricted genes and their associated regulatory elements.

Germline-active and soma-restricted genes have distinctive regulatory architectures

Given we observe that germline-active and soma-restricted genes have distinctive regulatory architectures, we next asked whether these differences could be explained by differences in the regulatory elements associated with ubiquitously active, germline- and soma-restricted promoters. To this end, we first classified promoters into ubiquitously active, germline- and soma-restricted promoters, based on their ATAC-seq signal, and assignment of regulatory elements associated with gene-front ends (see Methods). We find that the ubiquitously active promoters are associated with more ATAC-seq peaks than the germline- and soma-restricted ones, a trend that was even more pronounced for the single promoter genes (Figure 3b–e). Moreover, we find that the ubiquitously active promoters are also associated with significantly longer distances to their closest ATAC-seq peak than the germline- and soma-restricted promoters, and that this trend is even more significant when only single promoter genes are considered (Figure 3c–e). Thus, we find that the ubiquitously active genes have more and longer-range cis-acting regulatory elements than the germline- and soma-restricted genes.

Next, we asked whether the ubiquitously active promoters exhibit similar DNA sequence features as the germline- and soma-restricted promoters. We find that the ubiquitously active promoters are more A/T-rich than the germline-active promoters (Figure 3—figure supplement 1d–e), and that this difference is even more pronounced when only single promoter genes are considered (Figure 3—figure supplement 1f–h). Moreover, we find that the ubiquitously active promoters are also more A/T-rich than the soma-restricted promoters, and that this difference is even more pronounced when only single promoter genes are considered (Figure 3—figure supplement 2a–c). Finally, we find that the ubiquitously active promoters are more conserved than the germline- and soma-restricted promoters, and that this difference is even more pronounced when only single promoter genes are considered (Figure 3—figure supplement 2d–e). Thus, we find that the single promoter genes with germline-active or tissue-specific promoters have more distinctive sequence features than those with ubiquitously active promoters, and that these features are more pronounced when only single promoter genes are considered.

Ubiquitous and germline-specific promoters have a stereological architecture with unique features

We next asked whether differences in gene regulatory architecture correspond to differences in DNA sequence features, such as motif composition, and whether these differences could explain the observed differences in gene expression patterns. To this end, we first asked whether the ubiquitously active and germline-specific promoters have distinct motif compositions. We find that the ubiquitously active promoters are enriched for motifs that are also enriched in the germline-specific promoters (Supplementary File 2). The finding ATAC-seq peaks associated with germline-active genes are more A/T-rich than the ATAC-seq peaks associated with soma-restricted genes, we next

added to protect RNA from being degraded. Nucleic acids were slowly resting at 4°C on the dark for ~1.5-3.5 hours, then were stored at -20°C for 1 hour. Nucleic acids were then centrifuged at 12,000 rpm for 10 min at 4°C, washed in 4 mL of 75% ethanol, and resuspended in buffer. A commercial RNA-DNA mixture DNase inhibitor at a concentration of approximately 10-15 million nucleic per milliliter. Finally, nucleic were filtered at 250 rpm and stored at -20°C. Nucleic acid quantity was measured using a Nanodrop spectrophotometer.

Nucleic extraction was performed at 4°C using a Sony 5000RF series fitted with a 100 µm settling chip and auto-calibrated. Nucleic were gated using the DAPI signal, and PE-positive nucleic were gated using PE-A signal. The PE-positive nucleic were sorted into a 96-well plate containing 10 µL of buffer per well. Nucleic were sorted at a 10-nucleic per second rate. Cells were sorted into a 96-well plate containing 10 µL of buffer. Nucleic were sorted into 15-mL Falcon tubes containing 500 µL of buffer. A with 800 bind, mutate Rhine solution. Nucleic were stained in buffer (1 mL) and incubated for 30 minutes at room temperature. The purity and integrity of each batch of nucleic were assessed by recording an aliquot of sorted nucleic in a second plate in the same sorting run. Nucleic were sorted, and the sorted nucleic in the second plate were lysed, bound to mutate, and revealed by the circulate DNP signal antibodies, as well as the GFP signal outlining the nucleic envelope (see Supplement Fig. S2b).

ATAC-seq

Nucleic extracted nucleic were pelleted (2000x for 20 min at 4°C) and resuspended in 1× Tell buffer (10 mM Tris at pH 8, 5 mM MgCl2, 10% DMF) at a final concentration of about 500,000 nucleic per milliliter. Nucleic were then tagged with Tn5 transposase (Illumina, Tell-Seq Kit) for 1 hour at 37°C. Transposed nucleic were then pelleted at 2000x for 20 min at 4°C (Eppendorf, 5424R). ATAC-seq libraries were generated from two biological replicates for each tissue and were sequenced in both single-end and paired-end mode. ATAC-seq libraries were made for 12 and 24 month old C57BL/6 mouse tissues (Jackson, JAX #000664) (supplement). PGC-specific ATAC-seq data at the 12 stage were obtained from Lee et al. (2017).

RNA-seq

RNA-seq extracted from 1 million sorted and nucleid nucleic using the standard procedure (Hagen et al. 2018). A minimum of 20 ng of total nucleic RNA was used to make long-insert RNA-seq libraries. Long-insert RNA-2 (200 nt) was isolated using Zymo Clean and Concentrator columns (R1013). ARNA was removed using the Ribo-Zero-Mix (Illumina, Tell-Seq Kit). Nucleic were then fragmented and ligated with the NEBNext ultra directional RNA library prep kit (E7420S). Long-insert RNA-seq libraries were generated from two biological replicates for each tissue and were sequenced in both single-end and paired-end mode. Long-insert RNA-seq libraries were made for 12 and 24 month old C57BL/6 mouse tissues (Jackson, JAX #000664) (supplement). PGC-specific RNA-seq data at the 12 stage were obtained from Lee et al. (2017).

RNA-seq

RNA-seq extracted from 1 million sorted and nucleid nucleic using the standard procedure (Hagen et al. 2018). A minimum of 20 ng of total nucleic RNA was used to make long-insert RNA-seq libraries. Long-insert RNA-2 (200 nt) was isolated using Zymo Clean and Concentrate columns (R1013). ARNA was removed using the Ribo-Zero-Mix (Illumina, Tell-Seq Kit). Nucleic were then fragmented and ligated with the NEBNext ultra directional RNA library prep kit (E7420S). Long-insert RNA-seq libraries were generated from two biological replicates for each tissue and were sequenced in both single-end and paired-end mode. Long-insert RNA-seq libraries were made for 12 and 24 month old C57BL/6 mouse tissues (Jackson, JAX #000664) (supplement). PGC-specific RNA-seq data at the 12 stage were obtained from Lee et al. (2017).

Data processing

Data were processed as previously described (Jonas et al. 2018) and aligned to WBcel235v1 genome Further details are given in the Supplement Notes.

To assess the reproducibility of biological replicate data sets, we used rank rank reproducibility or gene expression correlation between biological replicate data sets and pairwise correlation between ATAC-seq and RNA-seq. ATAC-seq and RNA-seq biological replicates showed high concordance (Supplemental Fig. S2b).

Classification of accessible sites

First, accessible nucleic sub-sets in each sample was calculated as nucleic per million (RPM) values. RPMs of biological replicates were averaged to obtain a single accessibility score for each site in each tissue. Sites with accessibility > 0.RPM in each tissue were not further nucleid.

Then, estimation of accessibility fold-changes (FC) and adjusted P-values were computed between all pairwise tissue comparisons using DESeq2 (Love et al. 2014). Sites were classified as tissue-specific if they had a FC > 2 and adjusted P-value < 0.01 compared to all other tissues. Sites were classified as ubiquitously accessible if they had a FC < 0.01. ATAC-seq peaks that were between consecutive tissues was used as a threshold to determine the tissue specificity of accessible sites. Classification details are given in the Supplement Material.

Classification of genes

Long nucleic RNA-seq nucleid fragments were assigned to C. elegans gene annotations (WBcel235). Gene TSS using the Ensembl genome (ftp://ftp.ensembl.org/pub/release-75/gtf/caenorhabditis_elegans/) and the Tell-Seq pipeline (version 0.9.4). Nucleic gene expression using the DESeq2 package. A gene was considered significantly differentially expressed (DE) between two tissues if there was a FC greater than two and an adjusted P-value < 0.01.

In each sample, gene expression was calculated as transcripts per million (TPM). Genes. TPM of biological replicates were then averaged. Genes with a TPM greater than one were considered nucleid.

The ratio used to classify accessible sites were also used to classify genes, with a detection threshold of 1 TPM. A tissue-specific gene was considered DE if it had a FC > 2 and adjusted P-value < 0.01 (Jonas et al. 2018) were classified as genes specific and not included in this study.

Gene Ontology (GO) analysis

GO enrichment analysis were performed using the gProfiler0.6.7 package (Reimand et al. 2007). Ribosomal genes were excluded from the GO analysis. Ribosomal genes were defined as genes belonging to ribosomal protein gene ontology (GO) terms. Gene ontology (GO) analysis were performed using the gProfiler0.6.7 package (Reimand et al. 2007) were used. Filtering for nucleated terms using RPKM (Jonas et al. 2018). Only GO terms with Bonferroni-adjusted P-values lower than 0.05 were kept.

ATAC-seq fragment density plots

ATAC-seq fragment density plots, also known as Vplots (Buenrostro et al. 2013), were generated using the Vplot0.0 package (version 0.9.4). Nucleic fragments were mapped using Bowtie2 mapping nucleotides enrichment were nucleid calculated from the Vplots as illustrated in Supplement Figure S3A.

Nucleosome occupancy tracks and nucleosome occupancy

Processed BAM files from paired-end ATAC-seq of pharynx cell tissue or from whole organism young adults (Jonas et al. 2018) were merged. For each class of promoter (germline, somatic muscle, hypodermis, intestine, and ubiquitous promoters), the nucleosOCC Python package (Schep et al. 2015) was used to compute the probability of nucleosome occupancy from -1.kbp to +1.kbp relative to the transcription start site (germline, neuron, muscle, hypodermis, intestine, and whole organism).

Putative +1 and -1 nucleosome positions were identified as the peaks of tissue-specific promoters using the corresponding tissue-specific nucleosome occupancy probability track and for ubiquitous promoters using the whole organism nucleosome occupancy probability track. Nucleosome. The +1 nucleosome center of the position +1 nucleosome to the nucleosome center of the nucleosome immediately within 200 bp downstream from the forward TSS mode. Similarly, the center of the -1 nucleosome nominal was assigned to the local maximum of the occupancy probability within 200 bp upstream of the reverse TSS mode. Only coding promoters with experimentally determined forward and reverse TSSs were considered.

Motif identification and enrichment analysis

Motifs nucleid in different sets of promoters (-50 bp to +50 bp from promoter centers) were identified using MEME in standard mode and a zero-order background model (markov; order 0). MEME mode was set to "any number of repetitions" (-mod anr) and motif widths were restricted to 6-20 bp. The five motifs found most enriched in each tissue (with a P-value < 0.05) were retrieved. Unrestricted motifs were then used to scan promoters using FIMO (default settings). Motif enrichment was then nucleid combined PWMs for the human ortholog of the TSS-nucl were obtained from Jaspar (2016). Motif mapping to promoters was performed in R v.3.3.2 (R. Core Team 2015) using the Bioconductor 2.18.2 package, the GenomeRanges 1.24.0 package, and the TFBSTools 1.10.0 package with a-scthres threshold set to 8.0.

Detectable periodicity

To estimate detectable periodicity in sets of sequences (e.g., -150 to +500 bp sequences around ubiquitous, germline, or somatic tissue specific TSSs in Fig. 3B, or -150 to +500 bp sequences around TSSs from different organisms in Fig. 2A), the performodicity function from the periodicDNA 0.2.0 package was used with default parameters. Briefly, the distribution of distances between all possible pairs of dinucleotides in the set of sequences was computed and corrected for distance decay and smoothed by a moving average with a window size of 3 nucleotides. Periodicity was then estimated by applying a discrete transform to the normalized distribution.

To generate 10-bp detectable periodicity scores tracks, the getmotifPeriodicityTrack() function from the periodicDNA package version 0.2.0 (https://github.com/vivekkumar2/periodicDNA) was used with default parameters. Briefly, a running 10-bp detectable periodicity score was calculated by applying a fast Fourier transform (fast FT) algorithm to the distribution of distances between pairs of dinucleotides (e.g., WW ... WW) found in 100-bp windows (21-bp increments).

Phasing of nucleosomal occupancy

To observe the 10-bp periodic occurrence of a dinucleotide in putative +1 nucleosomes, sequences (100 bp centered at the nucleosome dyad) were aligned at the dinucleotide. The number of dinucleotides at each sequence, and then the counters were replotted within a +/-50 bp range using the lag value estimated by the ccf() function from the stats 3.2.1 package.

Data of annotations in fly, fish, mouse, and human

In vivo, experimentally determined TSSs were used (Jonas et al. 2018) in fly and zebrafish (respectively dmel and danRer10 genome versions). TSSs were assigned to the first base of the gene using TxDb.Dmelanogaster.UCSC.dmel.ensGene v.3.4.1 and TxDb.Drerio.UCSC.danRer10.refGene v.3.4.0 packages with the GenomeFeatures 1.24.7 package in R. In mouse and human, FANTOM CAGE data sets were used (respectively mm9 and hg19 genome versions). Promoter-centered CAGE data were downloaded from Zenodo (Lizio et al. 2015). For mouse, the mouse lymphoid tissue data set was used (https://zenodo.org/record/12671#.Vm-mmr9Ldc0). For human, the human primary cell data set was used (https://zenodo.org/record/12671#.Vm-mmr9Ldc0). For zebrafish, Genes with the 20% highest CTSs were considered broadly expressed, and those with the 20% highest CTSs were considered regulated.

Nucleosome occupancy in fly, fish, mouse, and human

Nucleosome occupancy in fly, fish, mouse, and human was estimated by using nucleosOCC with the Kingsford/Salzberg parameters (-f 5 -p 4 -u 50 -l 147 -q -n hg19 -b MEME100.120.txt -D GRCh37 -D ensembl -D ucsc -D ncbi -D GRCh37 -N MEME70.txt (Buenrostro et al. 2013) in mouse, and MEME50.120 (Buenrostro et al. 2013) in human.

Software availability

Processed data and all annotations are available and can be either dynamically explored or anonymously downloaded at https://genome.ucsc.edu/hubs/. Software for Vplotting and for analysis of periodicity is available at https://github.com/Tan-Lab/Vplot0.0.4 (Vplot0.0.4 tag) and https://github.com/Tan-Lab/periodicDNA (periodicDNA 0.4 tag) and https://github.com/Tan-Lab/Vplot-nucleosomeOcc (Vplot-nucleosomeOcc 0.2.0 (Jonas) code periodicDNA 0.2.0 tag) is available at https://github.com/Tan-Lab/periodicDNA-0.2.0.tar.gz) is available at https://github.com/Tan-Lab/periodicDNA-0.2.0.tar.gz) 1.0 Both are also available as Supplement Codes.

Data access

All raw and processed sequencing data presented in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE112233.

Competing interest statement

The authors declare no competing interests.

Supplementary Material

- **Supplemental Material:**
  - Click here to view.

Acknowledgments

We thank all the members of the Alvarez laboratory and particularly F. Cuetlo and A. Fraguero for fruitful discussions and comments on the manuscript. We thank L. Kaufman and Y. Kim for support in sequencing. The work was supported by a Wellcome Trust Senior Research Fellowship to A.H. (106243) and a Medical Research Council DPFS postdoctoral fellowship to O.M. We also acknowledge core support from the Wellcome Trust (092890) and Cancer Research UK (C6946/A14492).

Author contributions: J.S. and A.C. conceived and designed the studies. J.S. performed experiments unless otherwise noted. J.S., M.C., and C.C. generated the reporter strains. J.S. and V.D. prepared RNA-seq libraries. J.S. and J.J. processed the data. J.S. analyzed the data. J.S. and A.S. wrote the manuscript. A.S. supervised the study and provided funding.

Footnotes

[Supplemental material is available for this article.]

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.2094.120.

Freely available online through the Genome Research Open Access option.

References

Albert I, Marsh TN, Tomlins JE, Q-J J, Zanto KL Schauer SC, Pugh BF. 2007. **Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces genome**. *Nature* **446**: 572–576. 10.1038/nature05632 [Abstract] [Google Scholar]

Allfrey VG, Faulkner R, Mirsky AE. 1964. **Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis**. *Proc Nat Acad Sci USA* 51: 786–794. 10.1073/pnas.51.5.786 [Abstract] [Google Scholar]

Andrews S. 2010. **FastQC: a quality control tool for high throughput sequence data**. [Online]. Available at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. [Google Scholar]

Bae C, von Puhlen I, Vida VC, Remsen S, Sakamoto N, Ng SW, Mutsen A, Langenfeld K, Petrovich M, Scipioneschi R, et al. 2014. **A Mu enhancer cluster regulates normal and leukemic hematopoietic stem cell activity**. *Nature* **510**: 513–520. 10.1038/nature13487 [Abstract] [Google Scholar]

Bechtel S, Jakob M, Sayol V, Sabban S, Purbel A, Mikulcik-Ilmberger A, Zaborer J, Barbacz A, Moser S, Lutz C, et al. 2018. **Direct induction of the first differentiation step in Drosophila oocytes**. *BioRxiv*. 10.1101/gr.224.117 [Abstract] [Google Scholar]

Bernstein E, Duncan EM, Sawyer S