Supplementary Information

for

Ho *et al.*, Comparative analysis of metazoan chromatin organization

Supplementary Content

Supplementary Methods

Preprocessing of ChIP-seq data

Raw sequences were aligned to their respective genomes (hg19 for human; dm3 for fly; and ce10 for worm) using bowtie[28] or BWA[29] following standard preprocessing and quality assessment procedures of ENCODE and modENCODE[6]. Validation results of the antibodies used in all ChIP experiments are available at the Antibody Validation Database[24] (http://compbio.med.harvard.edu/antibodies/). Most of the ChIP-seq datasets were generated by 36 bp (in human and worm), 42 bp (in worm), or 50 bp (in fly and worm) single-end sequencing using the Illumina HiSeq platform, with an average of ~20 million reads per sample replicate (at least two replicates for each sample). Quality of the ChIP-seq data was examined as follows. For all three organisms, cross-correlation analysis was performed, as described in the published modENCODE and ENCODE guidelines[6]. This analysis examines ChIP efficiency and signal-to-noise ratio, as well as verifying the size distribution of ChIP fragments. The results of this cross-correlation analysis for the more than 3000 modENCODE and ENCODE ChIP-seq data sets are described elsewhere[30]. In addition, to ensure consistency between replicates in the fly data, we further required at least 80% overlap of the top 40% of peaks in the two replicates (overlap is determined by number of bp for broad peaks, or by number of peaks for sharp peaks; peaks as determined by SPP[31] *etc*). Library complexity was checked for human. For worm, genome-wide correlation of fold enrichment values was computed for replicates and a minimum threshold of 0.4 was required. In all organisms, those replicate sets that do not meet these criteria were examined by manual inspection of browser profiles to ascertain the reasons for low quality and, whenever possible, experiments were repeated until sufficient quality and consistency were obtained. To enable the cross-species comparisons described in this paper, we have reprocessed all data using MACS[32]. (Due to the slight differences in the peak-calling and input normalization steps, there may be slight discrepancies between the fly profiles analyzed here and profiles available at the modENCODE data portal (http://data.modencode.org) or modMine (http://modmine.org, which redirects to http://intermine.modencode.org). For every pair of aligned ChIP and matching input-DNA data, we used MACS[33] version 2 to generate fold enrichment signal tracks for every position in a genome:

macs2 callpeak -t ChIP.bam -c Input.bam -B --nomodel --shiftsize 73 --SPMR -g hs -n ChIP

macs2 bdgcmp -t ChIP_treat_pileup.bdg -c ChIP_control_lambda.bdg -o ChIP_FE.bedgraph -m FE

Depending on analysis, we applied either log transformation or z-score transformation.

Preprocessing of ChIP-chip data

For the fly data, genomic DNA Tiling Arrays v2.0 (Affymetrix) were used to hybridize ChIP and input DNA. We obtained the log-intensity ratio values (M-values) for all perfect match (PM) probes: $M = \log_2(\text{ChIP intensity}) - \log_2(\text{input intensity})$, and performed a whole-genome baseline shift so that the mean of M in each microarray is equal to 0. The smoothed log intensity ratios were calculated using LOWESS with a smoothing span corresponding to 500 bp, combining normalized data from two replicate experiments. For the worm data, a custom Nimblegen two-channel whole genome microarray platform was used to hybridize both ChIP and input DNA. MA2C[34] was used to preprocess the data to obtain a normalized and median centered $\log_2$ ratio for each probe. All data are publicly accessible through the modENCODE data portal or modMine.

Preprocessing of GRO-seq data

Raw sequences of the fly S2 and human IMR90 datasets were downloaded from NCBI Gene Expression Omnibus (GEO) using accession numbers GSE25887[35] and GSE13518[36] respectively. The sequences were then aligned to the respective genome assembly (dm3 for fly and hg19 for human) using bowtie[28]. After checking for consistency based on correlation analysis and browser inspection, we merged the reads of the biological replicates before proceeding with downstream analyses. Treating the reads mapping to the positive and negative strands separately, we calculated minimally-smoothed signals (by a Gaussian kernel smoother with bandwidth of 10 bp in fly and 50 bp in human) along the genome in 10 bp (fly) or 50 bp (human) non-overlapping bins.

Preprocessing of DNase-seq data

Aligned DNase-seq data were downloaded from the modENCODE data portal and the ENCODE UCSC download page (http://encodeproject.org/ENCODE/). Additional *Drosophila*

embryo DNase-seq data were downloaded from[37]. After confirming consistency, reads from biological replicates were merged. We calculated minimally-smoothed signals (by a Gaussian kernel smoother with bandwidth of 10 bp in fly and 50 bp in human) along the genome in 10 bp (fly) or 50 bp (human) non-overlapping bins.

Preprocessing of MNase-seq data

The MNase-seq data were analyzed as described previously[38]. In brief, tags were mapped to the corresponding reference genome assemblies. The positions at which the number of mapped tags had a Z-score > 7 were considered anomalous due to potential amplification bias. The tags mapped to such positions were discarded. To compute profiles of nucleosomal frequency around TSS, the centers of the fragments were used in the case of paired-end data. In the case of single-end data, tag positions were shifted by the half of the estimated fragment size (estimated using cross-correlation analysis[39] toward the fragment 3'-ends and tags mapping to positive and negative DNA strands were combined). Loess smoothing in the 11-bp window, which does not affect positions of the major minima and maxima on the plots, was applied to reduce the high-frequency noise in the profiles.

GC-content and PhastCons conservation score

We downloaded the 5bp GC% data from the UCSC genome browser annotation download page (http://hgdownload.cse.ucsc.edu/downloads.html) for human (hg19), fly (dm3), and worm (ce10). Centering at every 5 bp bin, we calculated the running median of the GC% of the surrounding 100 bp (i.e., 105 bp in total).

PhastCons conservation score was obtained from the UCSC genome browser annotation download page. Specifically, we used the following score for each species.

Target species	PhastCons scores generated by multiple alignments with	URL
C. elegans (ce10)	6 Caenorhabditis nematode genomes	http://hgdownload.cse.ucsc.edu/goldenPath/ce10/phastCons7way/
D. melanogaster (dm3)	15 Drosophila and related fly genomes	http://hgdownload.cse.ucsc.edu/goldenPath/dm3/phastCons15way/
H. Sapiens (hg19)	45 vertebrate genomes	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/phastCons46way/vertebrate/

Both GC and phastCons scores were then binned into 10 bp (fly and worm) or 50 bp (human) non-overlapping bins.

Genomic sequence mappability tracks

We generated empirical genomic sequence mappability tracks using input-DNA sequencing data. After merging input reads up to 100M, reads were extended to 149 bp which corresponds to the shift of 74 bp in signal tracks. The union set of empirically mapped regions was obtained. They are available at the ENCODE-X Browser (http://encode-x.med.harvard.edu/data_sets/chromatin/).

Coordinates of unassembled genomic sequences

We downloaded the "Gap" table from UCSC genome browser download page (http://hgdownload.cse.ucsc.edu/downloads.html). The human genome contains 234 Mb of unassembled regions whereas fly contains 6.3 Mb of unassembled genome. There are no known unassembled (i.e., gap) regions in worm.

Gene annotation

We used human GENCODE version 10 (http://www.gencodegenes.org/releases/10.html) for human gene annotation[40]. For worm and fly, we used custom RNA-seq-based gene and transcript annotations generated by the modENCODE consortium (see Gerstein et al., Comparative Analysis of the Transcriptome across Distant Species).

Worm TSS definition based on capRNA-seq (capTSS)

We obtained worm TSS definition based on capRNA-seq from Chen *et al*.[23]. Briefly, short 5'-capped RNA from total nuclear RNA of mixed stage embryos were sequenced (i.e., capRNA-seq) by Illumina GAIIA (SE36) with two biological replicates. Reads from capRNA-seq were mapped to WS220 reference genome using BWA[29]. Transcription initiation regions (TICs) were identified by clustering of capRNA-seq reads. In this analysis we used TICs that overlap with wormbase TSSs within -199:100bp. We refer these capRNA-seq defined TSSs as capTSS in this study.

Gene expression data

Gene expression level estimates of various cell-lines, embryos or tissues were obtained from the modENCODE and ENCODE projects (see Gerstein et al., Comparative Analysis of the Transcriptome across Distant Species[7]). The expression of each gene is quantified in terms of RPKM (reads per million reads per kilobase). The distribution of gene expression in each cell line was assessed and a cut-off of RPKM=1 was determined to be generally a good threshold to separate active vs. inactive genes. This definition of active and inactive genes was used in the construction of meta-gene profiles.

Genomic coverage of histone modifications

To identify the significantly enriched regions, we used SPP R package (ver.1.10)[31]. The 5'end coordinate of every sequence read was shifted by half of the estimated average size of the fragments, as estimated by the cross-correlation profile analysis. The significance of enrichment was computed using a Poisson model with a 1 kb window. A position was considered significantly enriched if the number of IP read counts was significantly higher (Z-score > 3 for fly and worm, 2.5 for human) than the number of input read counts, after adjusting for the library sizes of IP and input, using SPP function *get.broad.enrichment.cluster*.

Genome coverage in each genome is then calculated as the total number of base pair covered by the enriched regions or one or more histone marks. It should be noted that genomic coverage reported in Supplementary Fig. 2 refers to percentage of histone mark coverage with respective to mappable region. A large portion (~20%) of human genome is not mappable based on our empirical criteria. These unmappable regions largely consist of unassembled regions, due to difficulties such as mapping of repeats. Furthermore, some unmappable regions may be a result of the relatively smaller sequencing depth compared to fly and worm samples. Therefore it is expected that empirically determined mappability is smaller in human compared to fly and worm.

Identification and analysis of enhancers

We used a supervised machine learning approach to identify putative enhancers among DNaseI hypersensitive sites (DHSs) and p300 or CBP-1 binding sites, hereafter referred collectively as "regulatory sites". The basic idea is to train a supervised classifier to identify H3K4me1/3 enrichment patterns that distinguish TSS distal regulatory sites (i.e., candidate enhancers) from proximal regulatory sites (i.e., candidate promoters). TSS-distal sites that carry these patterns are classified as putative enhancers.

Human DHS and p300 binding site coordinates were downloaded from the ENCODE UCSC download page (http://genome.ucsc.edu/ENCODE/downloads.html). When available, only peaks identified in both replicates were retained. DHSs and p300 peaks that were wider than 1 kb were removed. DHS positions in fly cell lines were defined as the 'high-magnitude' positions in DNase I hypersensitivity identified by Kharchenko *et al*[11]. We applied the same method to identify similar positions in DNase-seq data in fly embryonic stage 14 (ES14)[37], which roughly corresponds to LE stage. Worm MXEMB CBP-1 peaks were determined by SPP with default parameters. CBP-1 peaks that were identified within broad enrichment regions wider than 1 kb were removed. For fly and human cell lines, DHS and p300 data from matching cell types were used. For fly late embryos (14-16 h), the DHS data from embryonic stage 14 (10:20–11:20 h) were used. For worm EE and L3, CBP-1 data from mixed-embryos were used.

To define the TSS-proximal and TSS-distal sites, inclusive TSS lists were obtained by merging ensemble v66 TSSs with GENCODE version 10 for human, and modENCODE transcript annotations for fly and worm, including all alternate sites. Different machine learning algorithms were trained to classify genomic positions as a TSS-distal regulatory site, TSS-proximal regulatory site or neither, based on a pool of TSS-distal (>1 kb) and TSS–proximal (<250 bp) regulatory sites and a random set of positions from other places in the genome. The random set included twice as many positions as the TSS-distal site set for each cell type. Five features from each of the two marks, H3K4me1 and H3K4me3, were used for the classification: maximum fold-enrichment within +/-500 bp, and four average fold enrichment values in 250 bp bins within +/-500 bp. The pool of positions was split into two equal test and training sets. The performance of different classifier algorithms was compared using the area under Receiver Operator Characteristics (ROC) curves. For human and fly samples, the best performance was obtained using the Model-based boosting (mboost) algorithm[41], whereas for the worm data sets,

the Support Vector Machine (SVM) algorithm showed superior performance. TSS-distal sites that in turn get classified as "TSS-distal" make up our enhancer set. In worm, the learned model was used to classify sites within 500-1000 bp from the closest TSS, and those classified as TSS-distal were included in the final enhancer set to increase the number of identified sites. Our sets of putative enhancers (hereafter referred to as 'enhancers') include roughly 2000 sites in fly cell lines and fly embryos, 400 sites in worm embryos, and 50,000 sites in human cell lines.

It should be noted that while enhancers identified at DHSs (in human and fly) or CBP-1 binding sites (in worm) may represent different classes of enhancers, for the purpose of studying the major characteristics of enhancers, both definitions are a reasonable proxy for identifying enhancer-like regions. We repeated all human enhancer analysis with p300 sites (worm CBP-1 is an ortholog of p300 in human). Half of the p300-based enhancers overlap with DHS-based enhancers (Supplementary Table 3). In addition, all the observed patterns were consistent with the enhancers identified using DHSs (Supplementary Fig. 3), including the association of enhancer H3K27ac levels with gene expression (Supplementary Figs. 4-6), patterns of nucleosome turnover (Supplementary Figs. 7-9) and histone modifications and chromosomal proteins (Supplementary Fig. 10). The results based on DHS-based enhancers were validated by analyzing p300-based enhancers (Supplementary Fig. 11).

For Supplementary Fig. 3-6, the enrichment level of a histone mark around a site (DHS or CBP-1 enhancer) is calculated based on the maximum ChIP fold enrichment within +/- 500 bp region of the site. These values are also used to stratify enhancers based on the H3K27ac enrichment level. For Supplementary Fig. 10, we extracted histone modification signal +/- 2 kb around each enhancer site in 50 bp bins. ChIP fold enrichment is then averaged across all the enhancer sites in that category (high or low H3K27ac). These average signals across the entire sample (*e.g.*, human GM12878) are then subjected to Z-score transformation (mean = 0, standard deviation = 1). All z-scores above 4 or below -4 are set to 4 and -4 respectively.

In terms of analysis of average expression of genes that are proximal to a set of enhancers (Supplementary Fig. 5), we identify genes that are located within 5, 10, 25, 40, 50, 75, 125, 150, 175 and 200 kb away from the center of an enhancer in both directions, and take an average of the expression levels of all of the genes within this region.

Analysis of HiC-defined topological domains

We used the genomic coordinates of the topological domains defined in the original publication on fly late embryos[20], and human embryonic stem cell lines[19]. The human coordinates were originally in hg18. We used UCSC's liftOver tool (http://genome.ucsc.edu/cgi-bin/hgLiftOver) to convert the coordinates to hg19.

Analysis of chromatin states near topological domain boundaries

For each chromatin state, the number of domain boundaries where the given state is at a given distance to the boundary is counted. The random expected value of counts is calculated as the number of all domain boundaries times the normalized genomic coverage of the chromatin state. The ratio of observed to expected counts is presented as a function of the distance to domain boundaries.

Analysis of chromatin states within topological domains

In supplementary Fig. 39, the interior of topological domains is defined by removing 4 kb and 40 kb from the edges of each topological domain for fly and human Hi-C defined domains respectively. To access the chromatin state composition of each topological domain, the coverage of the domain interior by each chromatin state is calculated in bps and normalized to the domain size, yielding a measure between 0 and 1. Then the matrix of values corresponding to chromatin states in one dimension and topological domains in the second dimension is used to cluster the chromatin states hierarchically. Pearson correlation coefficients (1-r) between domain coverage values of different chromatin states are taken as the distance metric for the clustering. The clustering tree is cut as to obtain a small number of meaningful groups of highly juxtaposed chromatin states. The coverage of each chromatin state group is calculated by summing the coverage of states in the group. Each topological domain is assigned to the chromatin state group with maximum coverage in the domain interior.

Definition of lamina associated domains (LADs)

Genomic coordinates of LADs were directly obtained from their original publications, for worm[42], fly[43] and human[44]. We converted the genomic coordinates of LADs to ce10 (for worm),

dm3 (for fly) and hg19 (for human) using UCSC's liftOver tool with default parameters (http://genome.ucsc.edu/cgi-bin/hgLiftOver). For Supplementary Fig. 16b, the raw fly DamID ChIP values were used after converting the probe coordinates to dm3.

LAD chromatin context analysis

In Supplementary Fig. 15, scaled LAD plot, long and short LADs were defined by top 20% and bottom 20% of LAD sizes, respectively. For a fair comparison between human and worm LADs in the figure, a subset of human LADs (chromosomes 1 to 4, N = 391) was used, while for worm LADs from all chromosomes (N= 360) were used. 10 kb (human) or 2.5 kb (worm) upstream and downstream of LAD start sites and LAD ending sites are not scaled. Inside of LADs is scaled to 60 kb (human) or 15 kb (worm). Overlapping regions with adjacent LADs are removed.

To correlate H3K9me3, H3K27me3 and EZH2/EZ with LADs, the average profiles were obtained at the boundaries of LADs with a window size of 120 kb for human, 40 kb for fly and 10 kb for worm. The results at the right side of domain boundaries were flipped for Supplementary Fig. 16a.

LAD Replication Timing analysis

The repli-seq BAM alignment files for the IMR90 and BJ human cell lines were downloaded from the UCSC ENCODE website. Early and late RPKM signal was determined for non-overlapping 50 kb bins across the human genome, discarding bins with low mappability (i.e., bins containing less than 50% uniquely mappable positions). To better match the fly repli-seq data, the RPKM signal from the two early fractions (G1b and S1) and two late fractions (S4 and G2) were each averaged together. The fly Kc cell line replication-seq data was obtained from GEO. Reads were pooled together from two biological replicates (S1: GSM1015342 and GSM1015346; S4: GSM1015345 and GSM1015349), and aligned to the *Drosophila melanogaster* dm3 genome using Bowtie[28]. Early and late RPKM values were then calculated for each non-overlapping 10 kb bin, discarding low mappability bins as described above. To make RPKM values comparable between both species, the fly RPKM values were normalized to the human genome size. All replication timing bins within a LAD domain were included in the

analysis. An equivalent number of random bins were then selected, preserving the observed LAD domain chromosomal distribution.

Cell Type	Phase	Link
IMR90	G1b	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwRepliSeq/wgEncodeUwRepliSeqImr90G1bAlnRep1.bam
IMR90	S1	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwRepliSeq/wgEncodeUwRepliSeqImr90S1AlnRep1.bam
IMR90	S4	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwRepliSeq/wgEncodeUwRepliSeqImr90S4AlnRep1.bam
IMR90	G2	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwRepliSeq/wgEncodeUwRepliSeqImr90G2AlnRep1.bam
BJ	G1b	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwRepliSeq/wgEncodeUwRepliSeqBjG1bAlnRep2.bam
BJ	S1	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwRepliSeq/wgEncodeUwRepliSeqBjS1AlnRep2.bam
BJ	S4	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwRepliSeq/wgEncodeUwRepliSeqBjS4AlnRep2.bam
BJ	G2	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwRepliSeq/wgEncodeUwRepliSeqBjG2AlnRep2.bam

Analysis of DNA structure and nucleosome positioning

The ORChID2 algorithm was used to predict DNA shape and generate consensus profiles for paired-end MNase-seq fragments of size 146-148 bp as previously described[45]. Only 146-148 bp sequences were used in this analysis to minimize possible effect of over- and under-digestion in the MNase treatment. The ORChID2 algorithm provides a more general approach than often-used investigation of mono- or dinucleotide occurrences along nucleosomal DNA since it can capture even degenerate sequence signatures if they have pronounced structural features.

For individual sequence analyses, we used the consensus profile generated above and trimmed three bases from each end to eliminate edge effects of the prediction algorithm, and then scanned this consensus against each sequence of length 146-148 bp. We retained the maximum correlation value between the consensus and individual sequence, and compared this to shuffled versions of each sequence (Supplementary Figs. 19-20). To estimate the sequence effect on nucleosome positioning we calculated the area between the solid lines and normalized by the

area between the dashed lines (Supplementary Fig. 20a; upper panel) and reported this result in Supplementary Fig. 19b.

Construction of meta-gene profiles

We defined transcription start site (TSS) and transcription end site (TES) as the 5' most and 3' most position of a gene, respectively, based on the modENCODE/ENCODE transcription group's gene annotation (see Gerstein et al., Comparative Analysis of the Transcriptome across Distant Species[7]). To exclude short genes from this analysis, we only included genes with a minimum length of 1 kb (worm and fly) or 10 kb (human). To further alleviate confounding signals from nearby genes, we also excluded genes which have any neighboring genes within 1 kb upstream of its TSS or 1 kb downstream of its TES. The ChIP enrichment in the 1 kb region upstream of TSS or downstream of TES, as well as 500 bp downstream of TSS or upstream of TES, were not scaled. The ChIP-enrichment within the remaining gene body was scaled to 2 kb. The average ChIP fold enrichment signals were then plotted as a heat map or a line plot.

Analysis of broadly and specifically expressed genes

For each species, we obtained RNA-seq based gene expression estimates (in RPKM) of multiple cell lines or developmental stages from the modENCODE/ENCODE transcription groups (see Gerstein et al., Comparative Analysis of the Transcriptome across Distant Species[7]). Gene expression variability score of each gene was defined to be the ratio of standard deviation and mean of expression across multiple samples. For each species, we divide the genes into four quartiles based on this gene expression variability score. Genes within the lowest quartile of variability score with RPKM value greater than 1 is defined as "broadly expressed". Similarly, RPKM>1 genes within the highest quartile of variability score is defined as "specifically expressed". We further restricted our analysis to protein-coding genes that are between 1 and 10 kb (in worm and fly) or between 1 and 40 kb (in human) in length.

For ChIP-chip analysis of BG3, S2 and Kc cells, ChIP signal enrichment for each gene was calculated by averaging the smoothed log intensity ratios from probes that fall in the gene body. For all other cell types, ChIP-seq read coordinates were adjusted by shifting 73 bp along the read and the total number of ChIP and input fragments that fall in the gene body were counted.

Genes with low sequencing depth (as determined by having less than 4 input tags in the gene body) were discarded from the analysis. ChIP signal enrichment is obtained by dividing (library normalized) ChIP read counts to Input read count. The same procedure was applied to calculate enrichment near TSS of genes, by averaging signals from probes within 500 bp of TSSs for BG3 cells and using read counts within 500 bp of TSSs for ChIP-seq data.

Genome-wide correlation between histone modifications

In Extended Data Fig. 1c and Supplementary Fig. 24, eight histone modifications commonly profiled in human (H1-hESC,GM12878 and K562), fly (LE, L3 and AH), and worm (EE and L3), were used for pairwise genome-wide correlation at 5 kb bin resolution. Unmappable regions and regions that have fold enrichment values less than 1 for all 8 marks (low signal regions) were excluded from the analysis. To obtain a representative correlation value for each species, an average Pearson correlation coefficient for each pair of marks was computed over the different cell types and developmental stages of each species. The overall correlation (upper triangle of Extended Data Fig. 1c) was computed by averaging the three single-species correlation coefficients. Intra-species variance was computed as the average within-species variance of correlation coefficients. Inter-species variance was computed as the variance of the within-species average correlation coefficients. For the large correlation heatmaps in Supplementary Fig. 25, 10 kb (worm and fly) or 30 kb (human) bins were used with no filtering of low-signal regions.

Chromatin segmentation using hiHMM

We performed joint chromatin state segmentation of multiple species using a hierarchically linked infinite hidden Markov model (hiHMM). In a traditional HMM that relies on a fixed number of hidden states, it is not straightforward to determine the optimal number of hidden states. In contrast, a non-parametric Bayesian approach of an infinite HMM (iHMM) can handle an unbounded number of hidden states in a systematic way so that the number of states can be learned from the training data rather than be pre-specified by the user[46]. For joint analysis of multi-species data, the hiHMM model employs multiple, hierarchically linked, iHMMs over the same set of hidden states across multiple species - one iHMM per species. More specifically, within a hiHMM, each iHMM has its own species-specific parameters for both transition matrix

$\pi^{(c)}$ and emission probabilities $\mu^{(c)}$ for *c*={human, fly, worm}. Emission process was modeled as a multivariate Gaussian with a diagonal covariance matrix such that $y_t^{(c)} \mid s_t^{(c)} = k \sim N(\mu_k^{(c)}, \Sigma^{(c)})$ where $y_t^{(c)}$ represents *m*-dimensional vector for observed data from *m* chromatin marks of species *c* at genomic location *t*, and $s_t^{(c)}$ represents the corresponding hidden state at *t*. The parameters $\mu_k^{(c)}$ correspond to the mean signal values from state *k* in species *c*, and $\Sigma^{(c)}$ is the species-specific covariance matrix. To take into account the different self-transition probabilities in different species, we also incorporate an explicit parameter $p_0^{(c)}$ that controls the self-transition probability. In the resulting transition model, we have $p(s_t^{(c)} = k \mid s_{t-1}^{(c)} = j) = p_0^{(c)}\delta(k = j) + (1 - p_0^{(c)})\pi_{jk}^{(c)}$. Each row of the transition matrix $\pi^{(c)}$ across all the species follows the same prior distribution of the so-called Dirichlet process that allows the state space to be shared across species. Using this scheme, data from multiple species are weakly coupled only by a prior. Therefore hiHMM can capture the shared characteristics of multiple species data while still allowing unique features for each species. This hierarchically linked HMM has been first applied to the problem of local genetic ancestry from haplotype data[46] in which the same modeling scheme for the transition process but a different emission process has been adopted to deal with the SNP haplotype data.

This hierarchical approach is substantially different from the plain HMM that treats multi-species data as different samples from a homogeneous population. For example, different species data have different gene length and genome composition, so one transition event along a chromosome of one species does not equally correspond to one transition in another species. So if a model has just one set of transition probabilities for all species, it cannot reflect such difference in self-transition or between-state transition probabilities. Our model hiHMM can naturally handle this by assuming species-specific transition matrices. Note that since the state space is shared across all the populations, it is easy to interpret the recovered chromatin states.

Since hiHMM is a non-parametric Bayesian approach, we need Markov chain Monte-Carlo (MCMC) sampling steps to train a model. Instead of Gibbs sampling, we adopted a dynamic programming scheme called Beam sampling[47], which significantly improves the mixing and

convergence rate. Although it still requires longer computation time than parametric methods like a finite-state HMM, this training can be done once offline and then we can approximate the decoding step of the remaining sequences by Viterbi algorithm using the trained HMM parameters.

ChIP-seq data were further normalized before being analyzed by hiHMM. ChIP-seq normalized signals were averaged in 200 bp bins in all three species. MACS2 processed ChIP-seq fold change values were $\log_2$ transformed with a pseudocount of 0.5, i.e., $y=\log_2(x+0.5)$, followed by mean-centering and scaling to have standard deviation of 1. The transformed fold enrichment data better resemble a Gaussian distribution based on QQ-plot analysis.

To train the hiHMM, the following representative chromosomes were used:

- Worm (L3): chrII, chrIII, chrX
- Fly (LE and L3): chr2L, chr2LHet, chrX, chrXHet
- Human (H1-hESC and GM12878): chr1, chrX

It should be noted that H4K20me1 profile in worm EE is only available as ChIP-chip data. This is why worm EE was not used in the training phase. In the inference phase, we used the quantile-normalized signal values of the H4K20me1 EE ChIP-chip data.

One emission and one transition probability matrix was learned from each species. We also obtained the maximum a priori (MAP) estimate of the number of states, *K*. We then used Viterbi decoding algorithm to generate a chromatin state segmentation of the whole genome of worm (EE and L3), fly (LE and L3) and human (H1-hESC and GM12878). To avoid any bias introduced by unmappable regions, we removed the empirically determined unmappable regions before performing Viterbi decoding. These unmappable regions are assigned a separate "unmappable state" after the decoding.

The chromatin state definition can be accessed via the ENCODE-X Browser (http://encode-x.med.harvard.edu/data_sets/chromatin/).

Chromatin segmentation using Segway

We compared the hiHMM segmentation with a segmentation produced by Segway[48], an existing segmentation method. Segway uses a dynamic Bayesian network model, which includes explicit representations of missing data and segment lengths.

Segway models the emission of signal observations at a position using multivariate Gaussians. Each label *k* has a corresponding Gaussian characterized by a mean vector μ_k and a diagonal covariance matrix Σ. At locations where particular tracks have missing data, Segway excludes those tracks from its emission model. For each label, Segway also includes a parameter that models the probability of a change in label. If there is a change in label, a separate matrix of transition parameters models the probability of switching to every other label. Given these emission and transition parameters, Segway can calculate the likelihood of observed signal data. To facilitate modeling data from multiple experiments with a single set of parameters, we performed a separate quantile normalization on each signal track prior to Segway analysis. We took the initial unnormalized values from MACS2's log-likelihood-ratio estimates. We compared the value at each position to the values of the whole track, determining the fraction of the whole track with a smaller value. We then transformed this fraction, using it as the argument to the inverse cumulative distribution function of an exponential distribution with mean parameter $\lambda = 1$. We divided the genome into 100 bp non-overlapping bins, and took the mean of the transformed values within each bin. We then used these normalized and averaged values as observations for Segway in place of the initial MACS2 estimates.

We trained Segway using the Expectation-Maximization algorithm and data from all three species: a randomly-sampled 10% of the human genome (with data from H1-hESC and GM12878) and the entire fly (LE and L3) and worm (EE and L3) genomes. Using these data sets jointly, we trained 10 models from 10 random initializations. In every initialization, we set each mean parameter μ_{ik} for label *i* and track *k* by sampling from a uniform distribution defined in $[-0.2\sigma, 0.2\sigma]$, where σ is the empirical standard deviation of track *k*. We placed a Dirichlet prior on the self-transition model to make the expected segment length 100 kb. We always initialized transition probability parameters with an equal probability of switching from one label to any other label. While these parameters changed during training, we increased the likelihood of a flatter transition matrix by including a Dirichlet prior of 10 pseudocounts for

each ordered pair of labels. To increase the relative importance of the length components of the model, we exponentiated transition probabilities to the power of 3. After training converged, we selected the model with the highest likelihood. We then used the Viterbi algorithm to assign state labels to the genome in each cell type of each organism.

Chromatin segmentation using ChromHMM

We also compared hiHMM with another existing segmentation method called ChromHMM[49]. ChromHMM uses a hidden Markov model with multivariate binary emissions to capture and summarize the combinatorial interactions between different chromatin marks. ChromHMM was jointly trained in virtual concatenation mode using 8 binary histone modification ChIP-seq tracks (H3K4me3, H3K27ac, H3K4me1, H3K79me2, H4K20me1, H3K36me3, H3K27me3 and H3K9me3) from two developmental stages in worm (EE, L3), two developmental stages in fly (LE, L3) and two human cell-lines (GM12878 and H1-hESC). The individual histone modification ChIP-seq tracks were binarized in 200 bp non-overlapping, genome-wide, tiled windows by comparing the ChIP read counts (after shifting reads on both strands in the 5' to 3' direction by 100 bp) to read counts from a corresponding input-DNA control dataset based on a Poisson background model. A *p*-value threshold of 1e-3 was used to assign a presence/absence call to each window (0 indicating no significant enrichment and 1 indicating significant enrichment). Bins containing < 25% mappable bases were considered unreliable and marked as 'missing data' before training. In order to avoid a human-specific bias in training due to the significantly larger size of the human genome relative to the worm and fly genomes, the tracks for both the worm and fly stages were repeated 10 times each, effectively up-weighting the worm and fly genomes in order to approximately match the amount of training data from the human samples. ChromHMM was trained in virtual concatenation mode using expectation maximization to produce a 19 state model which was found to be an optimal trade-off between model complexity and interpretability. The 19 state model was used to compute a posterior probability distribution over the state of each 200 bp window using a forward-backward algorithm. Each bin was assigned the state with the maximum posterior probability.

The states were labeled by analyzing the state-specific enrichment of various genomic features (such as locations of genes, transcription start sites, transcription end sites, repeat regions etc.)

and functional datasets (such as transcription factor ChIP-seq peaks and gene expression). For any set of genomic coordinates representing a genomic feature and a given state, the fold enrichment of overlap was calculated as the ratio of the joint probability of a region belonging to the state and the feature to the product of independent marginal probability of observing the state in the genome and that of observing the feature. Similar to the observations of hiHMM states, there are 6 main groups of states: promoter, enhancer, transcription, polycomb repressed, heterochromatin, and low signal.

Heterochromatin region identification

To identify broad H3K9me3+ heterochromatin domains, we first identified broad H3K9me3 enrichment region using SPP[31], based on methods *get.broad.enrichment.cluster* with a 10 kb window for fly and worm and 100 kb for human . Then regions that are less than 10 kb of length were removed. The remaining regions were identified as the heterochromatin regions. The boundaries between pericentric heterochromatin and euchromatin on each fly chromosome are consistent with those from lower resolution studies using H3K9me2[14] (Supplementary Fig. 34).

Genome-wide correlation analysis for heterochromatin-related marks

For heterochromatin related marks in Fig. 3b, the pairwise genome-wide correlations were calculated with 5 kb bins using five marks in common in the similar way as described above. Unmappable regions or regions that have fold enrichment values < 0.75 for all five marks were excluded from the analysis.

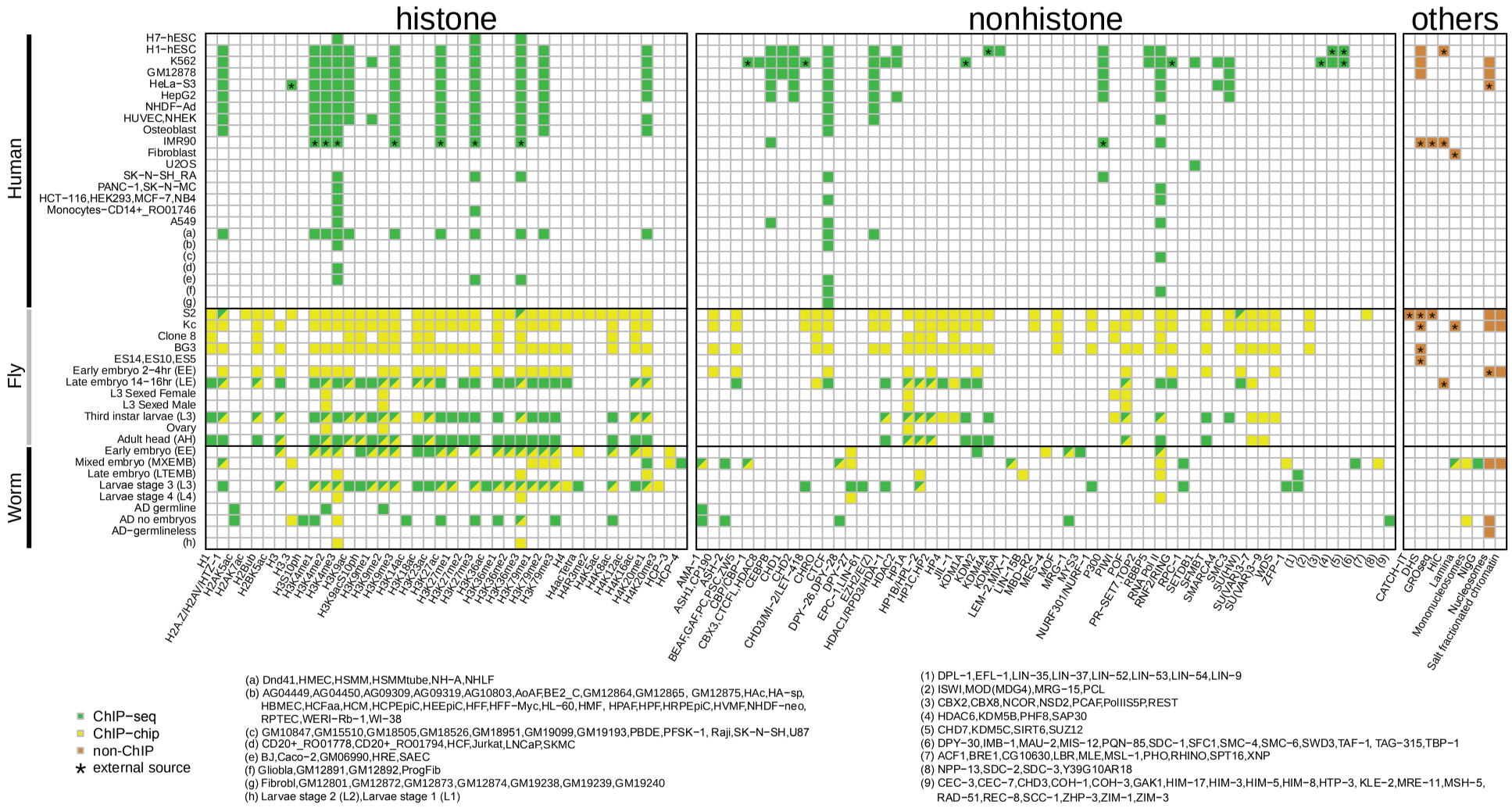
Chromatin-based topological domains based on Principal Component Analysis

We respectively partitioned the fly and worm genomes into 10 kb and 5 kb bins, and assign average ChIP fold enrichment of multiple histone modifications to each bin (See below for the list of histone modifications used). Aiming to reduce the redundancy induced by the strong correlation among multiple histone modifications, we projected histone modification data onto the principal components (PC) space. The first few PCs, which cumulatively accounted for at least 90% variance, were selected to generate a "reduced" chromatin modification profile of that bin. Typically 4-5 PCs were selected in the fly and worm analysis. Using this reduced chromatin modification profile, we could then calculate the Euclidean distance between every pair of bin

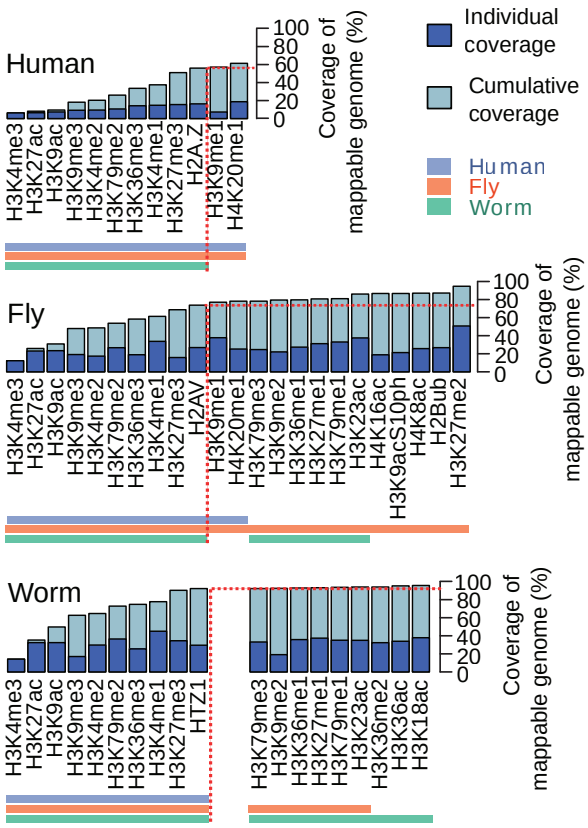
in the genome. In order to identify the boundaries and domains, we calculated a boundary score for each bin:

$$boundary\_score(k) = \frac{\sum_{i=-5, i\neq 0}^{i=5} d_{k+i,k}}{10}$$

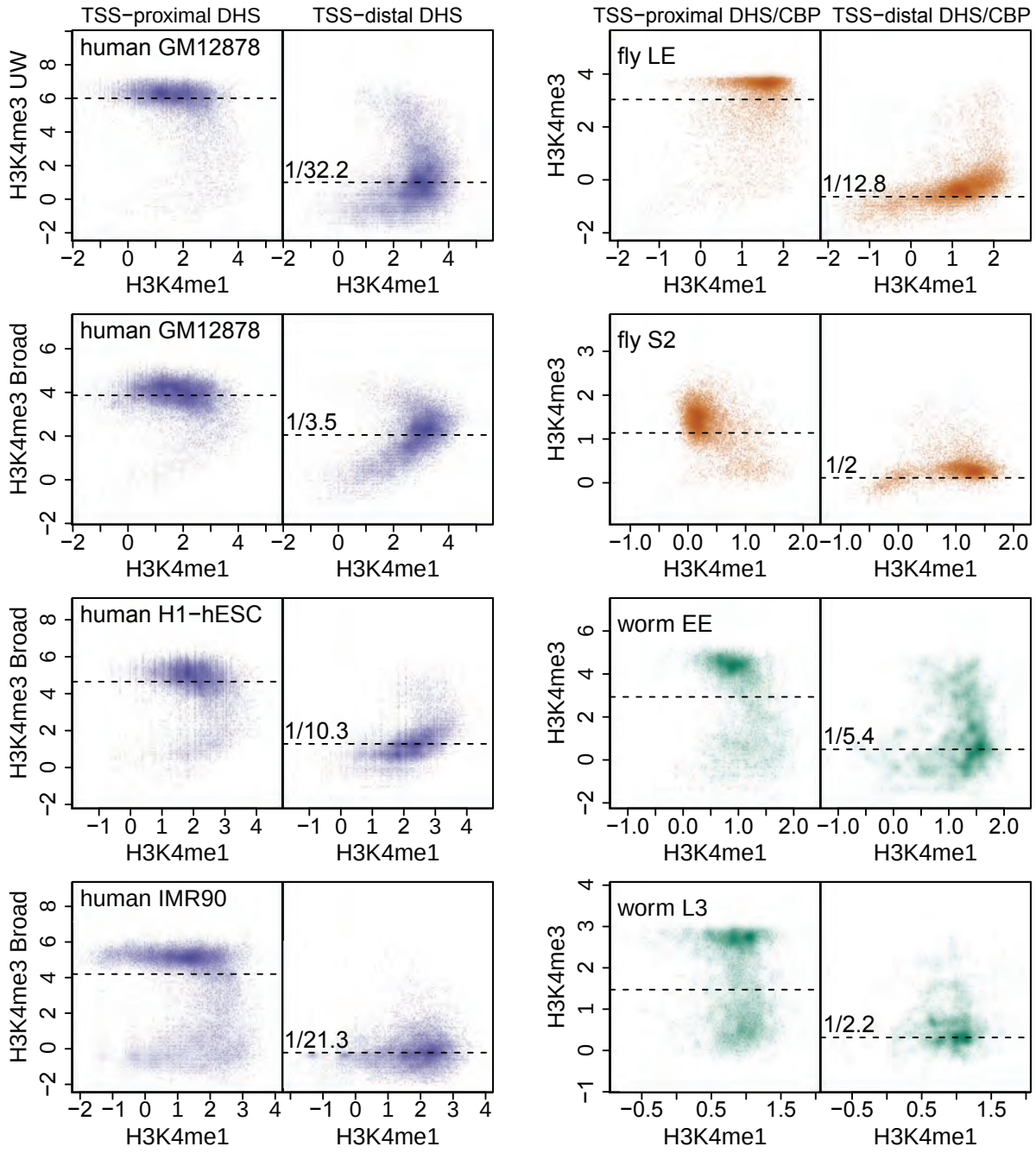
in which, $d_{k+i,k}$ is the Euclidean distance between the $k+i$ th bin and the k th bin. If a bin has larger distances between neighbors, in principle, it would have a higher boundary score and be recognized as a histone modification domain boundary. The boundary score cutoffs are set to be 7 for fly and worm. If the boundary scores of multiple continuous bins are higher than the cutoff, we picked the highest one as the boundary bin. The histone marks used are H3K27ac, H3K27me3, H3K36me1, H3K36me3, H3K4me1, H3K4me3, H3K79me1, H3K79me2, H3K9me2 and H3K9me3 for fly LE and L3, and H3K27ac, H3K27me3, H3K36me1, H3K36me3, H3K4me1, H3K4me3, H3K79me1, H3K79me2, H3K79me3, H3K9me2 and H3K9me3 for worm EE and L3.



Supplementary Fig. 1. A summary of the full dataset. This table shows all chromatin-associated data sets generated in the modENCODE/ENCODE project (the main Fig. 1 shows only those profiled in at least two organisms) as well as other published data sets used in this paper. Cell types and developmental stages that share the same pattern across the rows are merged for a more compact display; likewise, factors that share the same pattern across the columns were merged (comma separated). Orthologs with different mark or factor names among the three species are shown with the names separated by slash (/). Data that were not generated by the modENCODE/ENCODE consortium are marked by asterisks (*). Covalent Attachment of Tagged Histones to Capture and Identify Turnover (CATCH-IT) data[50] represent genome-wide measurement of nucleosome turnover performed using metabolic labeling followed by capture of newly synthesized histones.

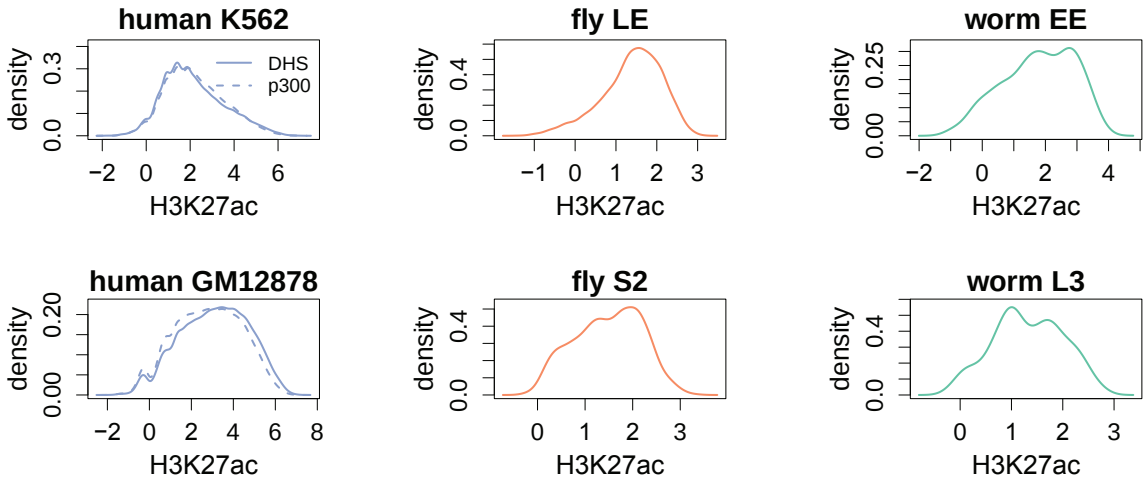


Supplementary Fig. 2. Genomic coverage of various histone modifications in the three species. Red lines indicate the ten marks analyzed in common in the three species and their cumulative coverage. The color bars underneath each plot indicate whether data is available for a given histone modification in that sample (K562 in human, L3 in fly and L3 in worm). In all three organisms, a large fraction of the assembled and mappable genome is occupied by at least one of the profiled histone modifications. For example, after excluding genomic regions that are unassembled or unmappable, the ten histone modifications profiled in at least one cell type or developmental stage of all three organisms display enrichments covering 56% of the human genome, 74% of the fly genome, and 92% of the worm genome (see Methods). The higher genomic coverage by histone modifications in worms and flies compared to humans is likely related to both the smaller genome size (which allows better sequencing coverage) and the higher proportion of protein-coding regions in the genomes of these organisms.

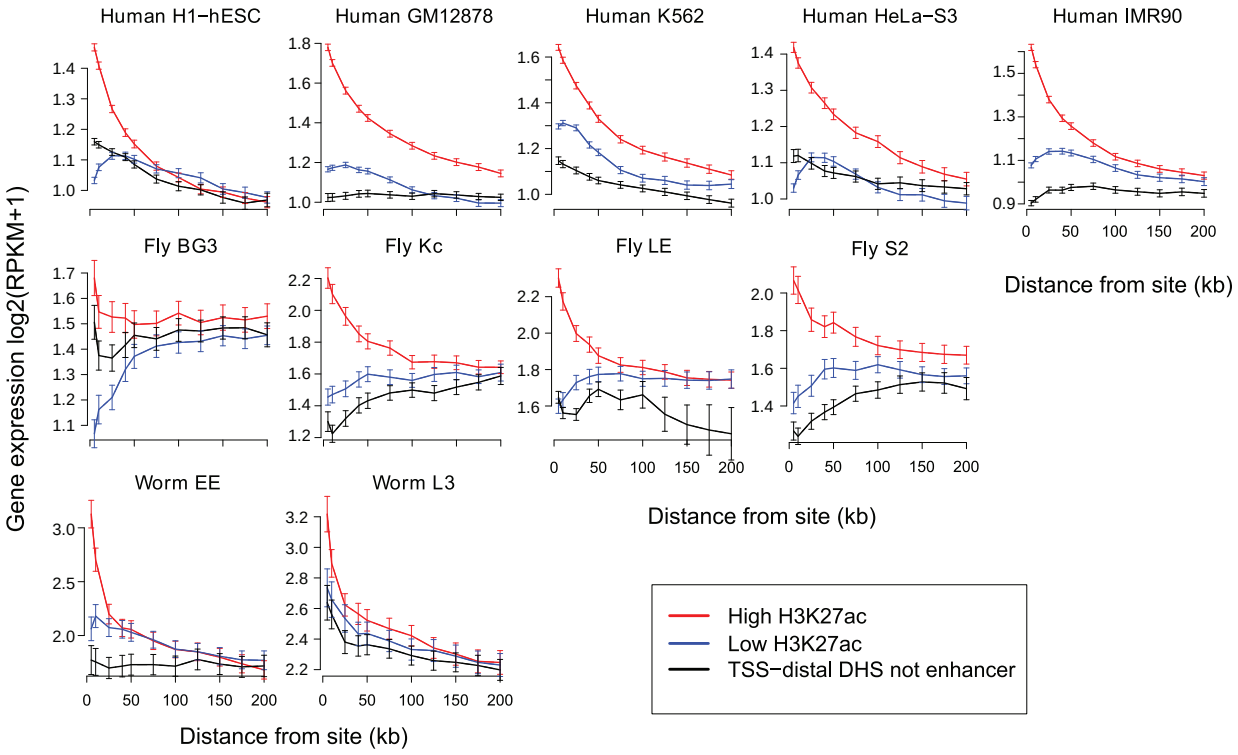


Supplementary Fig. 3. H3K4me1/3 enrichment patterns in regulatory elements defined by DNase I hypersensitive sites (DHS) or CBP-1 binding sites. ChIP signal enrichment (log2 scale) of H3K4me3 vs. H3K4me1 at TSS-proximal (<250 bp) and TSS-distal (>1 kb) DHSs (blue: human, orange: fly) or CBP-1 binding sites (green: worm). The labels “UW” and “Broad” denote, two ENCODE data generation centers: University of Washington and Broad Institute, respectively. The median H3K4me3 enrichment values are marked by horizontal dashed lines. The numbers (*e.g.*, 1/3.5) on the dashed lines denote the linear fold enrichment of H3K4me3 at the median TSS-distal site relative to the median TSS-proximal site (e.g., for UW GM12878 data, the median TSS-proximal DHS has 32.2 fold higher enrichment than the median TSS-distal DHS). To characterize the chromatin features that can distinguish enhancers from promoters, we

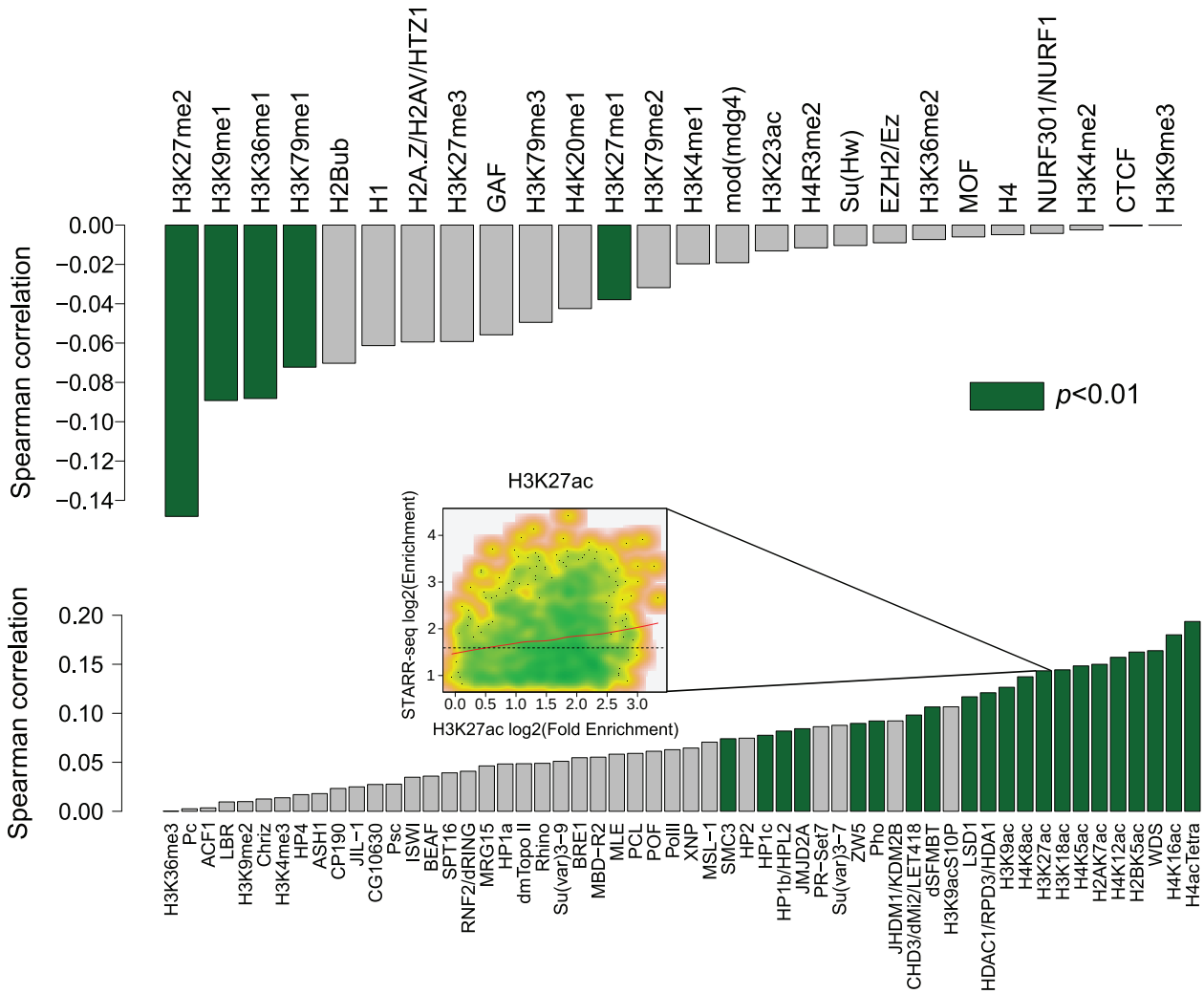
compared the enrichment patterns of H3K4me1 and H3K4me3 at TSS-proximal and TSS-distal DHSs in human and fly. Since DHS data were not available in worm, we examined the binding sites of CBP-1, the worm ortholog of human p300/CBP[51]. We observe that DHSs (or CBP-1 sites) generally fall into two clusters for all cell types: those proximal to TSSs constitute a cluster with stronger H3K4me3 signal (left column), while those distal to TSSs constitute a cluster showing stronger H3K4me1 signals (right column). Although the enrichment levels of H3K4me1/3 at these sites vary considerably between cell types, platforms (array vs. sequencing), and even different laboratories for the same cell type, these two marks clearly distinguish TSS-distal sites (enhancers) from TSS-proximal sites (promoters). Here, we define putative enhancer sites to be DHSs (or CBP-1 sites) with the H3K4me1/3 pattern that is characteristic of TSS-distal sites, as determined by a supervised machine learning approach (see Methods)



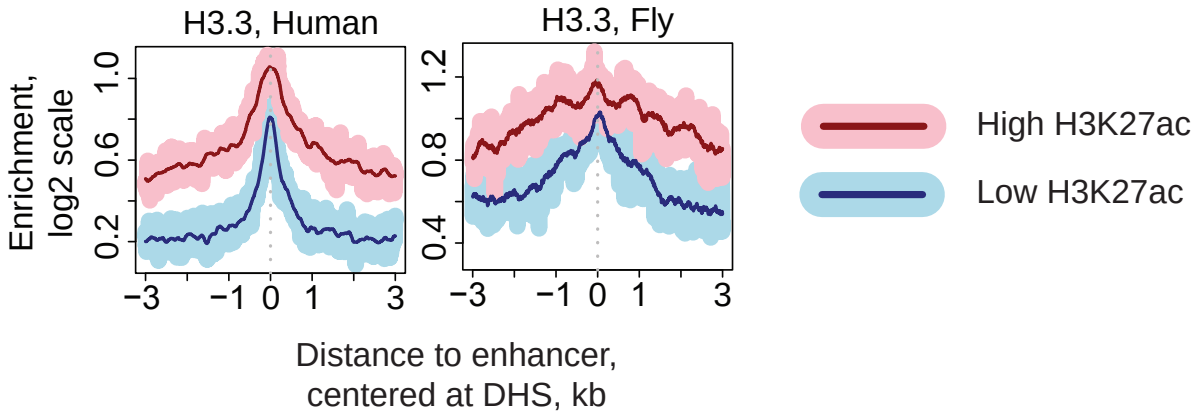
Supplementary Fig. 4. Distribution of H3K27ac enrichment levels at putative enhancers. One key observation is that H3K27ac density displays a wide range of enrichment levels at enhancers in all three species. This result in human cells is consistent whether using enhancers identified as DHSs (solid line) or as p300 binding sites (dashed line). X-axis is log2 ChIP fold enrichment of H3K27ac at +/-500 bp of enhancer sites.



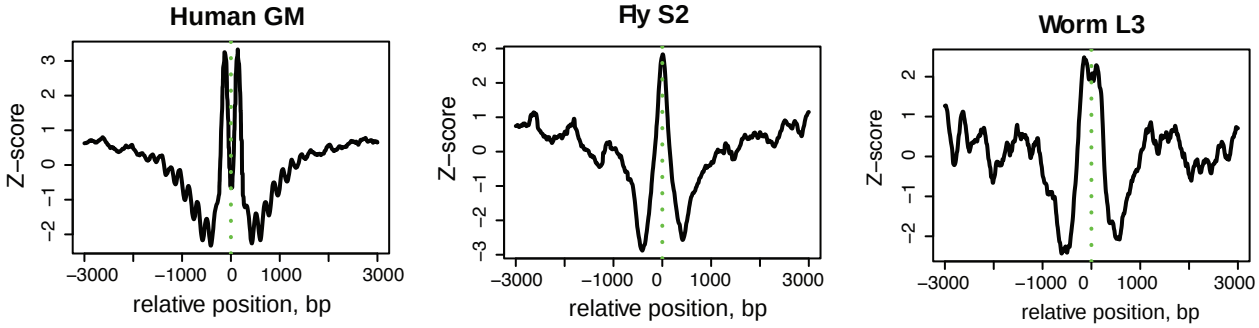
Supplementary Fig. 5. Relationship of enhancer H3K27ac levels with expression of nearby genes. Average expression of genes that are close (vary between 5 to 200 kb) to enhancers with high (top 40%; red line) or low (bottom 40%; blue line) levels of H3K27ac in various human, fly and worm samples. As a control, we analyzed TSS-distal DHSs (in human and fly) or CBP-1 sites (in worm) that are not classified as enhancers (dashed black). RPKM: reads per kilobase per million. Error bar: standard error of the mean. The proximity of genes to enhancers with higher H3K27ac levels is positively correlated with expression, in a distance-dependent manner. This observation is consistent across multiple cell-types and tissues in all three species.



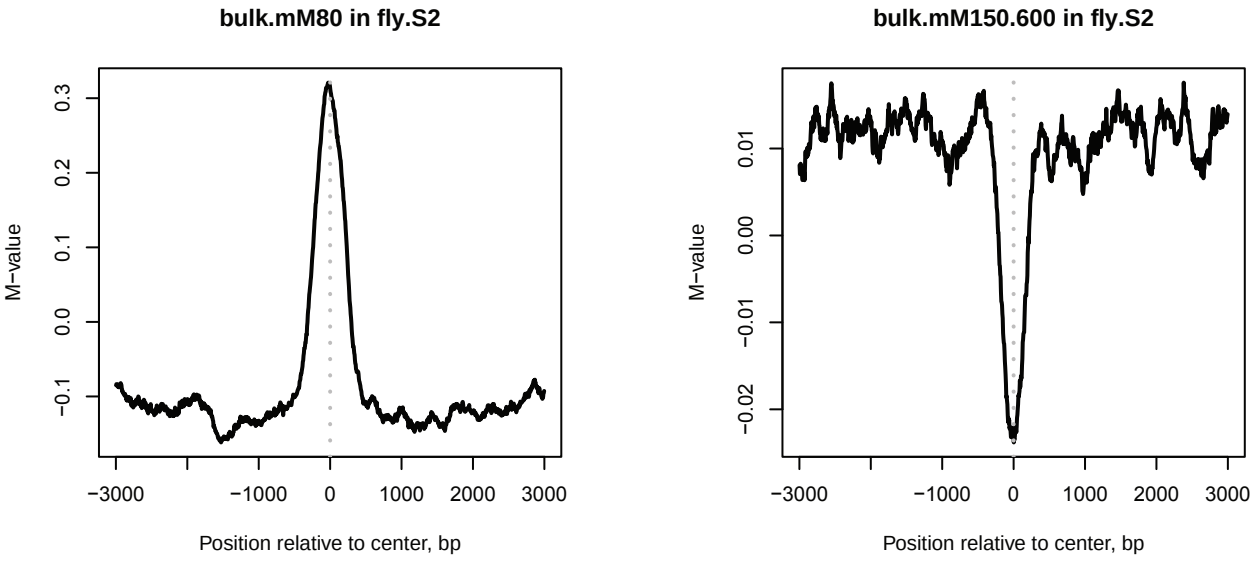
Supplementary Fig. 6. Correlation of enrichment of 82 histone marks or chromosomal proteins at enhancers with STARR-seq defined enhancer strength in fly S2 cells. Histone marks or chromosomal proteins whose enrichment is anti-correlated (top bar plot) or positively correlated (bottom bar plot) with STARR-seq enrichment level[52], which is a proxy for enhancer strength based on the ability of ~600 bp DNA fragments to stimulate transcription from an associated promoter. All histone lysine acetylation marks, including H3K27ac, show a moderate but significant positive correlation with enhancer activity (p<0.01).



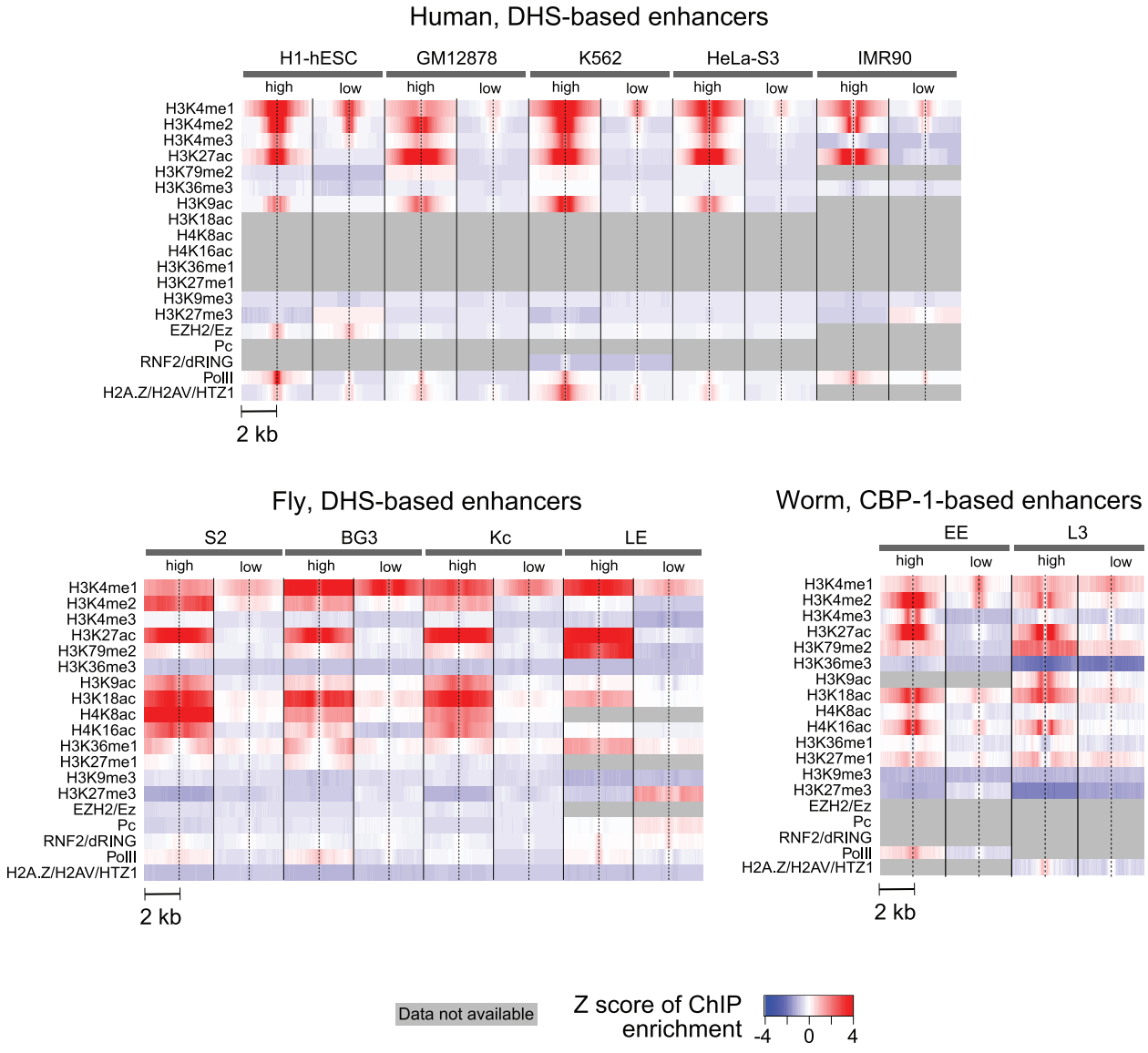
Supplementary Fig. 7. Nucleosome turnover at enhancers. ChIP signal enrichment ($\log_2$ scale) of H3.3 around enhancers in human Hela-S3 cells (ChIP-seq) and fly S2 cells (ChIP-chip), which is known to be present in regions with higher nucleosome turnover[50]. We found that the local increase in nucleosome occupancy (see Supplemental Fig. 8) indeed overlaps with the peak of H3.3 enrichment, and that the levels of H3.3 and H3K27ac enrichment are correlated. These findings, together with the specific patterns of nucleosome occupancy[53], indicate that increased nucleosome turnover is one of the major characteristics of chromatin at active enhancers.



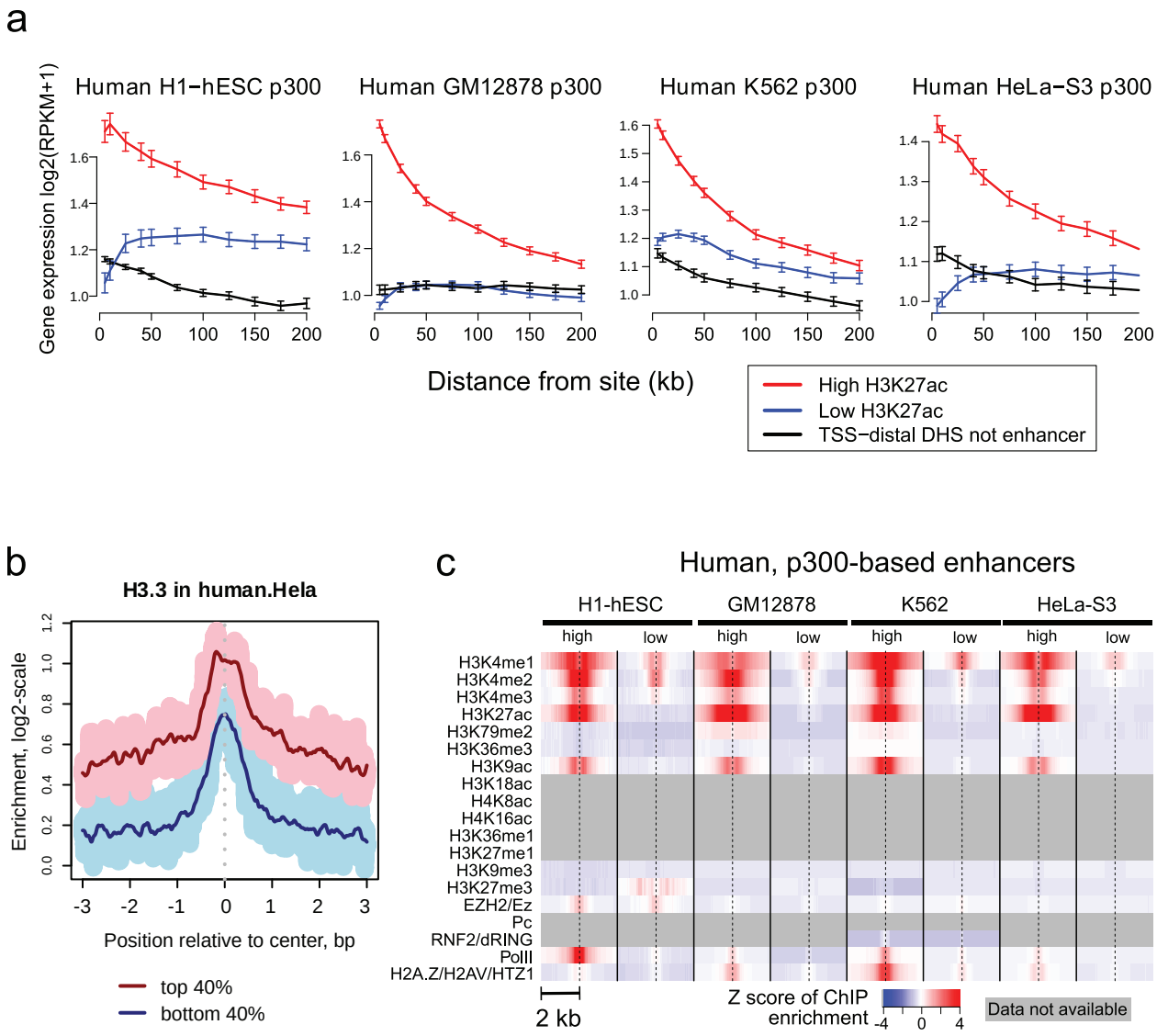
Supplementary Fig. 8. Nucleosome occupancy at enhancers. Average nucleosome density profiles were computed for DHS and CBP-1-identified enhancers in human GM12878 cells, fly S2 cells, and worm L3. In each case nucleosome occupancy was inferred from MNase-seq data obtained for the corresponding or similar cell types[54-56]. Green dashed lines indicate centers of the enhancer regions. In general, nucleosome occupancy is lower in the broad region around enhancers (roughly ±2 kb) but with a local (±400 bp) increase at the centers of the enhancers (defined by DHS and CBP-1 peaks). This pattern is similar to that reported for non-promoter regulatory sequences in the human genome[57]. In human, this increase is characterized by two well-positioned nucleosomes flanking the nucleosome-depleted region at the enhancer center, and this feature may be indicative of the presence of relatively unstable nucleosomes (this may be occluded by lower resolution in fly and worm).



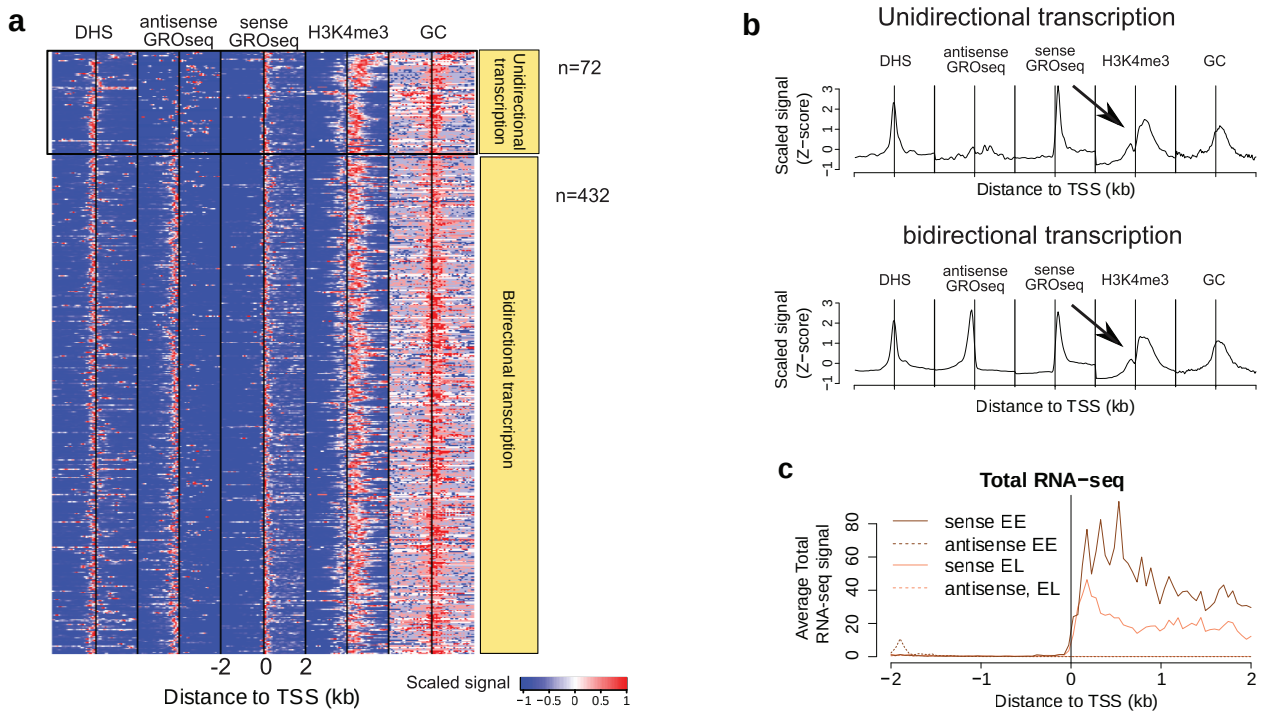
Supplementary Fig. 9. Salt extracted fractions of chromatin at enhancers. The average profiles are shown for the 80 mM (left) and 150-600 mM (right) salt fractions in fly S2 cells[58]. The 80 mM fraction is enriched with easily mobilized nucleosomes and preferentially represents accessible, “open” chromatin. The 150-600 mM fraction derives from a 600 mM extraction following a 150 mM extraction and therefore is depleted of such nucleosomes, representing more compacted, “closed” chromatin. We note that the peak in the 80 mM fraction at enhancers indicates that these loci are enriched in relatively unstable nucleosomes, which is in agreement with our observation of increased nucleosome turnover at these sites (see Supplementary Fig. 7).



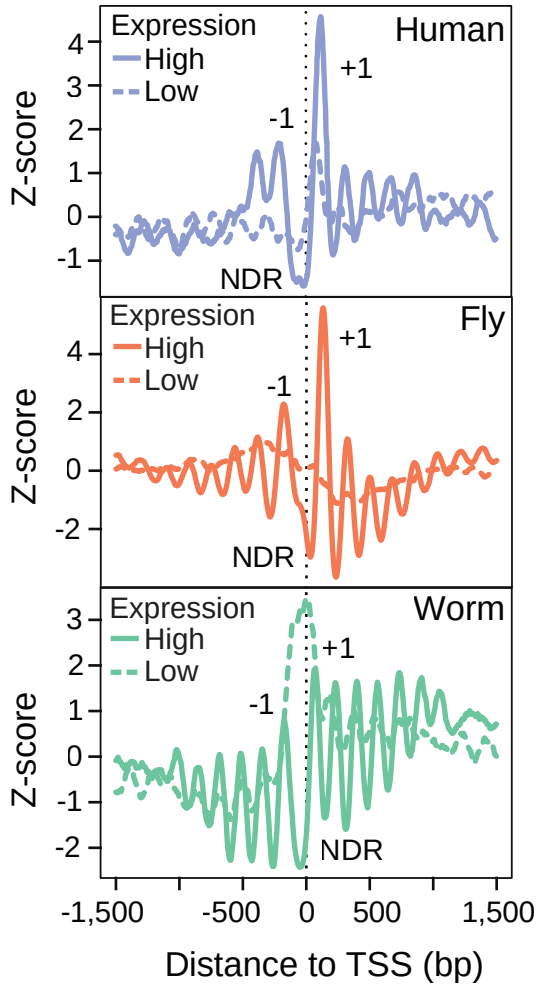
Supplementary Fig. 10. Chromatin environment described by histone modification and binding of chromosomal proteins at enhancers. Z-score of average ChIP fold enrichment of some key histone modifications and chromosomal proteins around +/-2 kb of the center of enhancers with high H3K27ac or low H3K27ac. Most active histone marks in addition to H3K4me1 show stronger enrichment at enhancers with high H3K27ac, including H3K4me2 and many H3 lysine acetylation marks. H3K27me3 is generally not enriched at enhancers except in embryonic stem cells such as human H1-hESC, where there is also enrichment of binding by the Polycomb protein EZH2. Enhancers with high H3K27ac have a higher prevalence of PolII binding in all three species, consistent with the elevated level of H3K4me3 at these sites compared to that in enhancers with low H3K27ac. H2A.Z is enriched in human enhancers, but the H2Av ortholog is not enriched in any fly samples. These configurations are likely to be correlated to the generation of short transcripts from these sites, as reported recently[59].



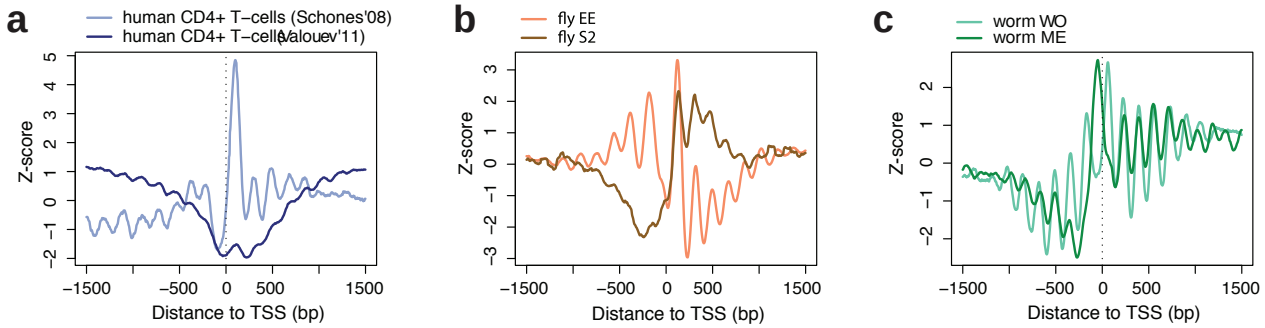
Supplementary Fig. 11. Analysis of p300-based enhancers from human. As an additional validation, we repeated all key analyses in human cell lines using the population of p300-based enhancers; the general trends remain the same as found for the DHS-based sites in corresponding human cell lines. **a,** Average expression of genes that are close to enhancers with high (top 40%; red line) or low (bottom 40%; blue line) levels of H3K27ac in human cell lines. As a control, we analyzed TSS-distal p300 binding sites that are not classified as enhancers (dashed black). RPKM: reads per kilobase per million. Error bar: standard error of the mean. **b,** ChIP signal enrichment (log2 scale) of H3.3 around p300-based enhancers in human Hela-S3 cells. **c,** Z-score of average ChIP fold enrichment of some key histone modifications and chromosomal proteins around +/-2 kb of the center of high H3K27ac or low H3K27ac enhancers in human cell lines. The observed patterns at human enhancers hold even if the putative enhancers were centered at p300 sites instead of DHSs.



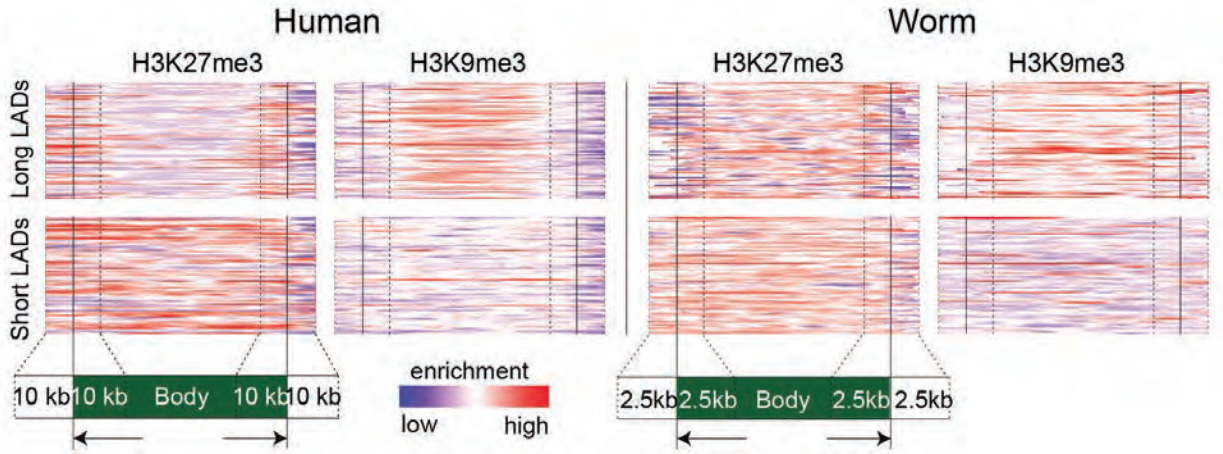
Supplementary Fig. 12. Relationship between sense-antisense bidirectional transcription and H3K4me3 at TSS. a, The majority of human expressed genes have sense-antisense bidirectional transcription at the TSS. Even in the small number of TSSs with unidirectional transcription, there is still a clear signal of bimodal H3K4me3 enrichment. The GC content pattern is the same as in expressed genes with unidirectional and bidirectional transcription. **b,** An average plot summarizing the results in panel A. **c,** Independently generated total RNA-seq data generated by modENCODE using fly early (2-4 hours) and late (14-16 hours) embryos support the observation made in fly S2 GRO-seq data that there is no evidence of strong antisense transcription at fly promoters.



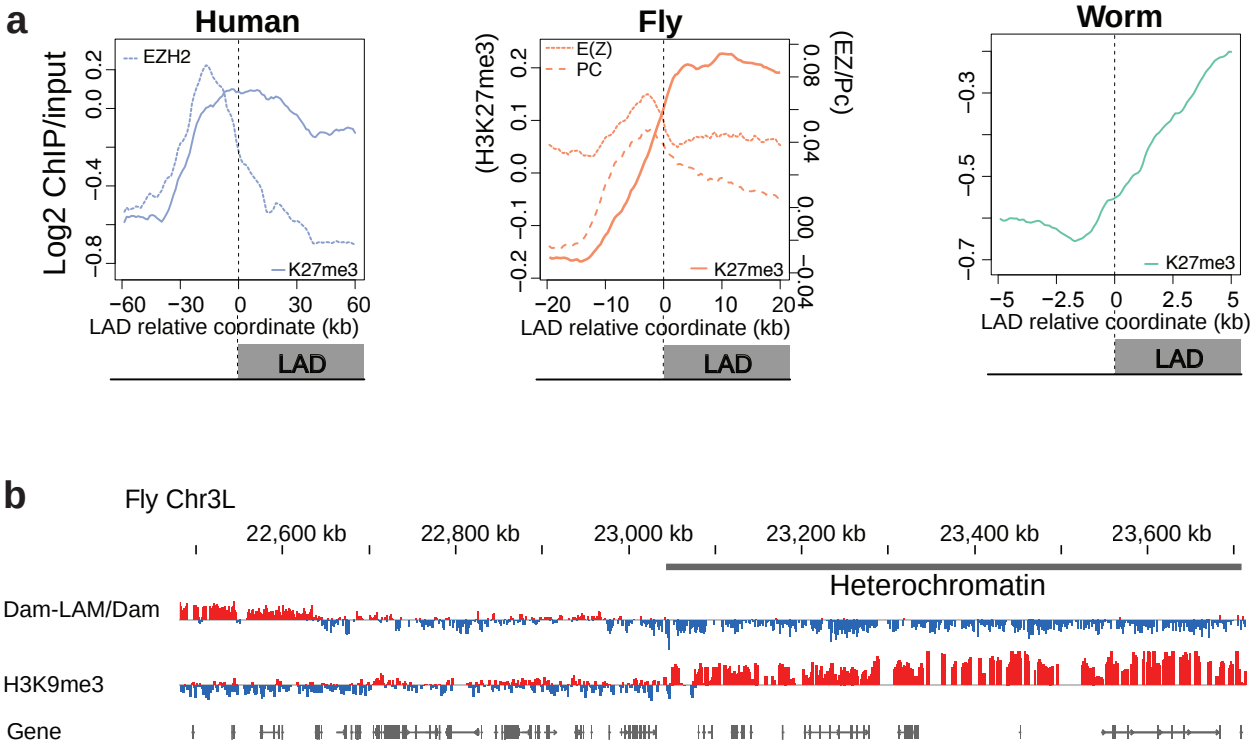
Supplementary Fig. 13. Profiles of the well positioned nucleosome at Transcription Start Sites (TSSs) of protein coding genes. Nucleosome frequency profiles (as represented as Z-scores) around TSSs for human CD4+ T cells, fly EE and worm adults. The profiles were computed for highly expressed (top 20% in all three species) and lowly expressed genes (bottom 20% for fly and human, and bottom 40% for worm; see Methods). The main features of the 'classic' nucleosome occupancy profile[60], comprising a nucleosome-depleted region at the TSS flanked by well-positioned nucleosomes ('-1', '+1', *etc.*) are observed in expressed genes for all three organisms. The similarity between the profiles, especially in the context of different nucleotide compositions of the TSS-proximal regions across the species, underscores the importance and conservation of specific nucleosome placement for gene regulation..



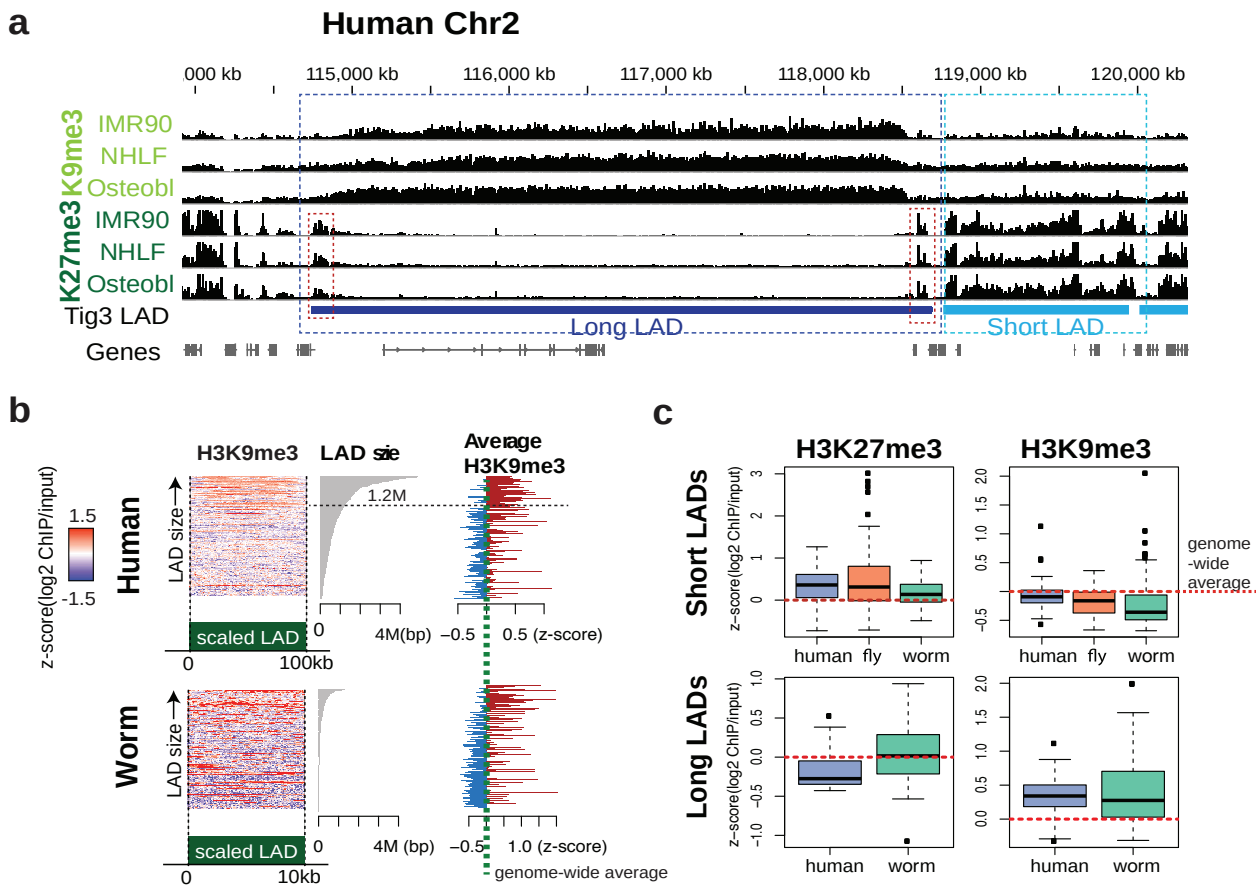
Supplementary Fig. 14. Nucleosome occupancy profile at TSS based on two MNase-seq datasets for each species. Comparison of the nucleosome occupancy profiles at TSS obtained in different studies. Two TSS-proximal profiles are plotted for each species: **a**, obtained for human CD4+ T-cells[61,62], **b**, obtained for fly embryos (this study) and S2 cells[63], and, **c**, obtained for worm embryos[64] and whole adult organisms[55]. All data were uniformly processed as described in Methods. The nucleosomal profiles at TSS, obtained under different biochemical conditions (*e.g.*, degree of chromatin digestion or salt concentration used to extract mono-nucleosomes), may vary substantially even for the same cell type, due to interplay between nucleosome stability and observed occupancy[65,66].



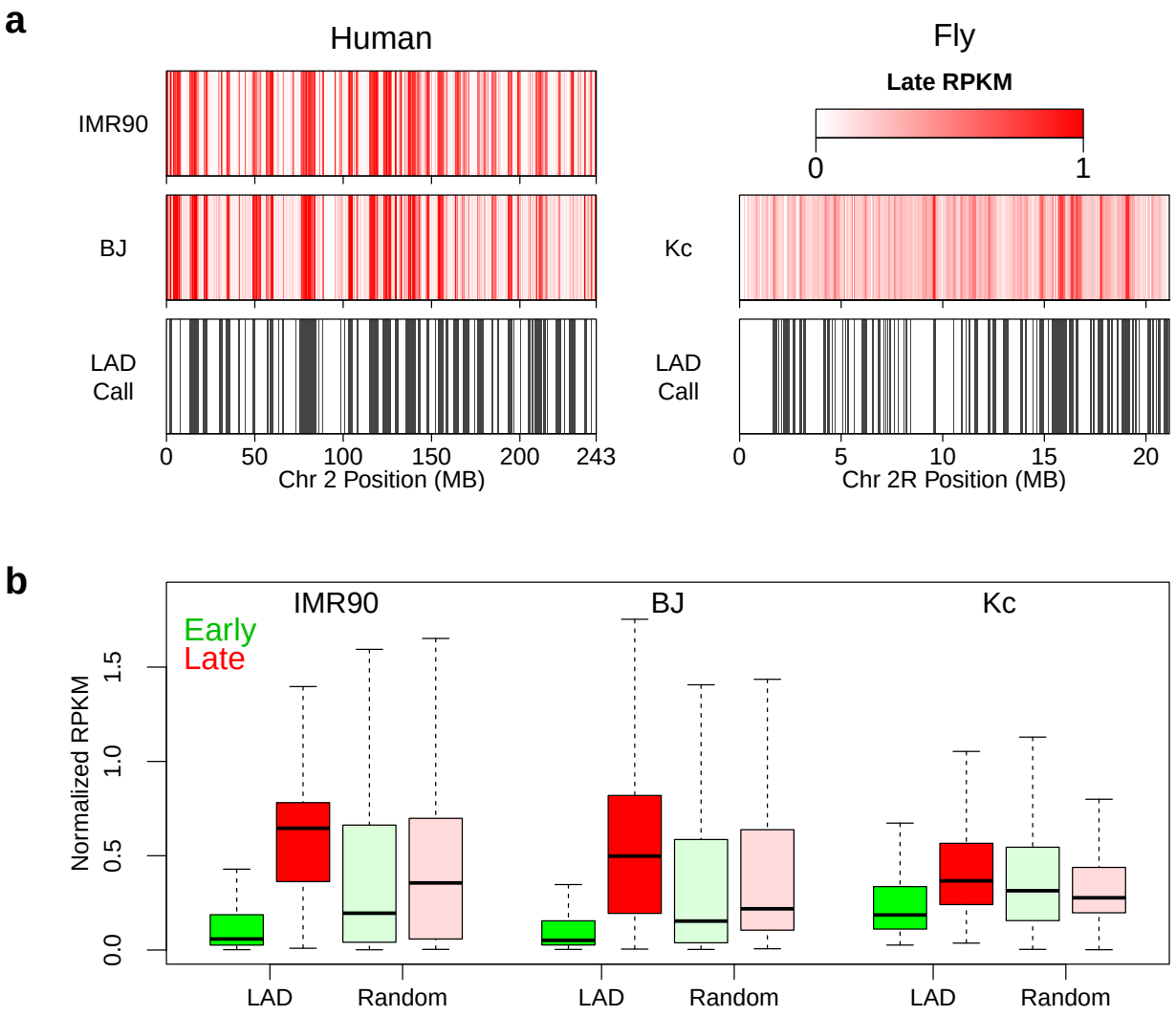
Supplementary Fig. 15. Association between repressive chromatin and lamina-associated domains (LADs). Heatmap of the enrichment of H3K9me3 and H3K27me3 in scaled LADs (upper panels: long LADs as defined as the 20% longest LADs; lower panel: short LADs as defined as the 20% shortest LADs). Each row represents H3K27me3 or H3K9me3 enrichment in each LAD. (H3K9me3 and H3K27me3 from IMR90, LADs from Tig3 for human; H3K9me3 and H3K27me3 from EE, LADs from MXEMB for worm). Our examination reveals a simple relationship that depends on LAD size. In human fibroblasts, long LADs (> 1 Mb) tend to be found in H3K9me3-enriched heterochromatic regions, with sharp enrichment of H3K27me3 at the LAD boundaries; in contrast, short LADs (< 1 Mb) are enriched for H3K27me3 across the domain with a low occupancy of H3K9me3. Although LADs are generally smaller in worm, we observe a similar though weaker trend, with longer LADs more frequently enriched for H3K9me3.



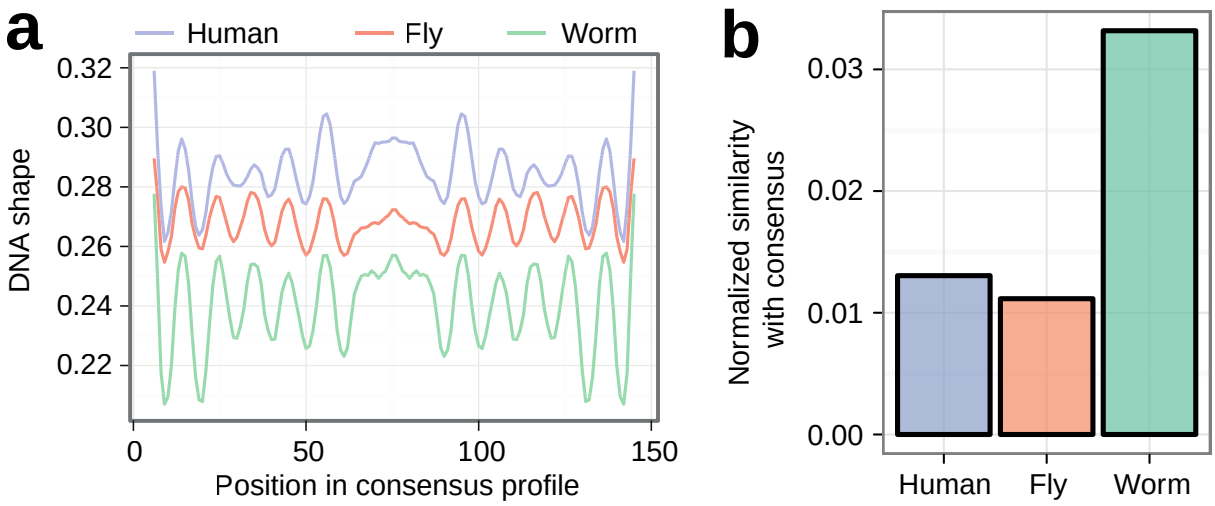
Supplementary Fig. 16. Chromatin context in lamina-associated domains. a, Average profiles of H3K27me3 (NHLF in human, Kc in fly, and EE in worm) and EZH2/E(Z) (NHLF in human and Kc in fly) at LAD boundaries. LADs are enriched for H3K27me3 and are often flanked by E(Z) in fly or EZH2 (the ortholog) in human, both H3K27 methyltransferases and members of Polycomb Repressive Complex 2. **b**, Genome browser shot of the profiles fly Kc Lam DamID in chromosome 2L. The levels of Lam (DamID) are negative in heterochromatin (gray block enriched with H3K9me3 in Kc). Y-axis: log2 enrichment of Lam (DamID) normalized by controls (first row); log2 ChIP/input (second row) in the range of -3 and 3.



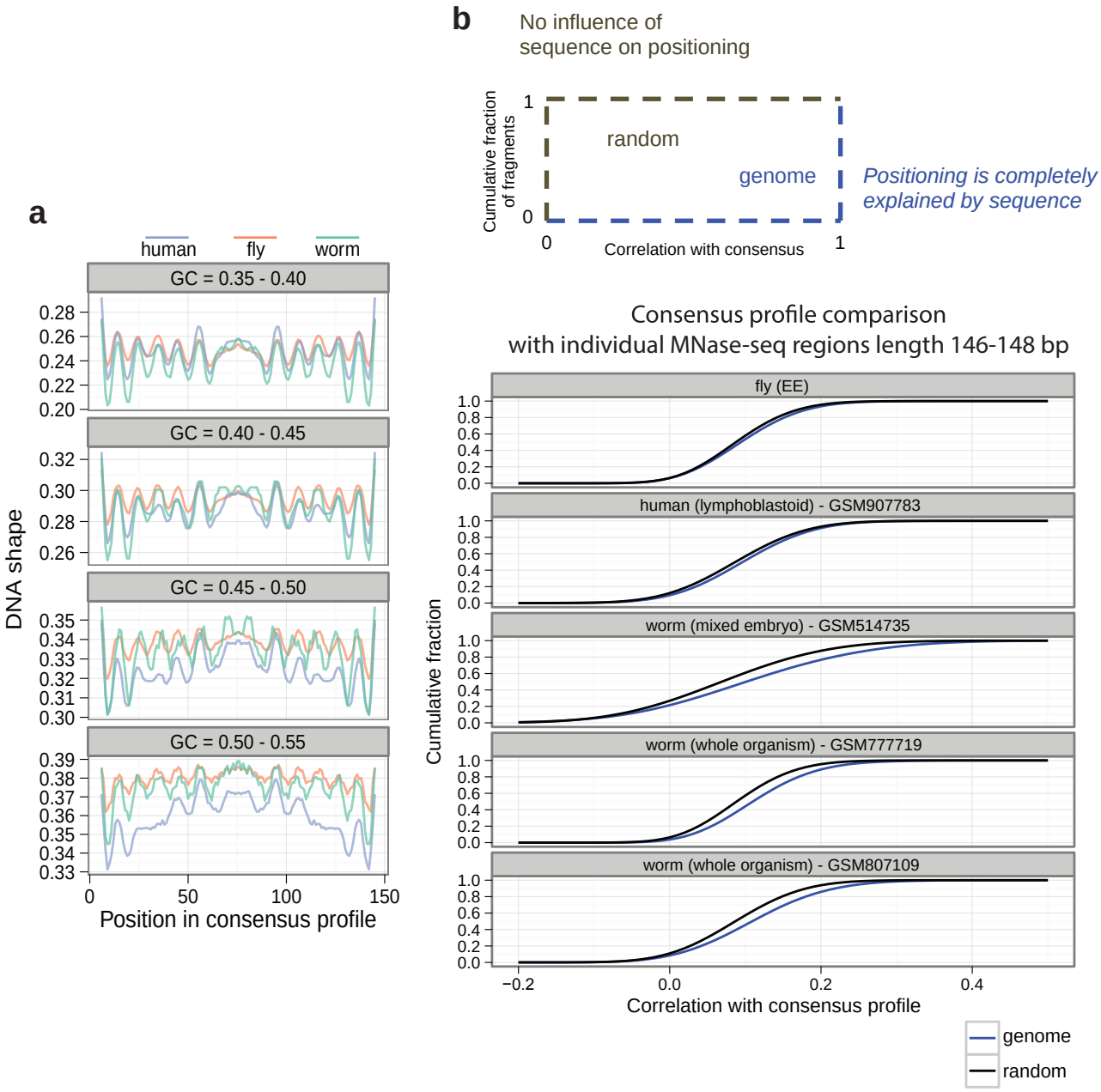
Supplementary Fig. 17. Chromatin context in short and long lamina-associated domains (LADs) in three organisms. In human, long LADs tend to be localized in H3K9me3-enriched heterochromatic regions, with sharp enrichment of H3K27me3 at the LAD boundaries; in contrast, short LADs are enriched for H3K27me3 across the domain body with a low occupancy of H3K9me3. **a,** Example of typical patterns of H3K9me3 and H3K27me3 profiles of fibroblast or fibroblast-like cell lines in human long LADs (dark blue) and short LADs (light blue) from fibroblast Tig3. Y-axis: fold enrichment of ChIP/input in the range of 0 and 2. The enrichment of H3K27me3 is observed at the boundaries of long LADs (red dashed). **b,** Relationship between the level of H3K9me3 enrichment in LADs and the size of LADs in human and worm. Longer LADs are more frequently enriched for H3K9me3. Left: heatmap of the enrichment in scaled LADs (upper: human; lower: worm). Each row represents H3K9me3 enrichment in each LAD, sorted by the size of LAD. Middle: LAD domain size. Right: average H3K9me3 values in each LAD. Genome-wide average values are indicated by the green dashed lines. In human, H3K9me3 is often associated with LADs of > ~1.2 Mb. (Human: IMR90 for H3K9me3 and Tig3 for LADs; worm: early embryos for H3K9me3 and mixed embryos for LADs). **c,** The average enrichments in long LADs (top 20% in LAD size) and short LADs (bottom 20% in LAD size). In short LADs, in all three species, the levels of H3K27me3 enrichment are higher than the genome-wide average, whereas the levels of H3K9me3 enrichment are low. In long LADs, the levels of H3K9me3 enrichment are higher than the genome-wide average.) No long LADs in the H3K9me3 hetero-chromatic regions were reported in fly data generated from Kc167 cells using DamID[43]; however, this may reflect the specific cellular origin (plasmatocyte) of Kc167 cells[67], as well as the fact that these analyses do not include the simple tandem repeats that constitute the majority of fly heterochromatin (Human: IMR90 for H3K9me3/H3K27me3 and Tig3 for LADs; fly data from Kc; worm: early embryos for H3K9me3/H3K27me3 and mixed embryos for LADs).



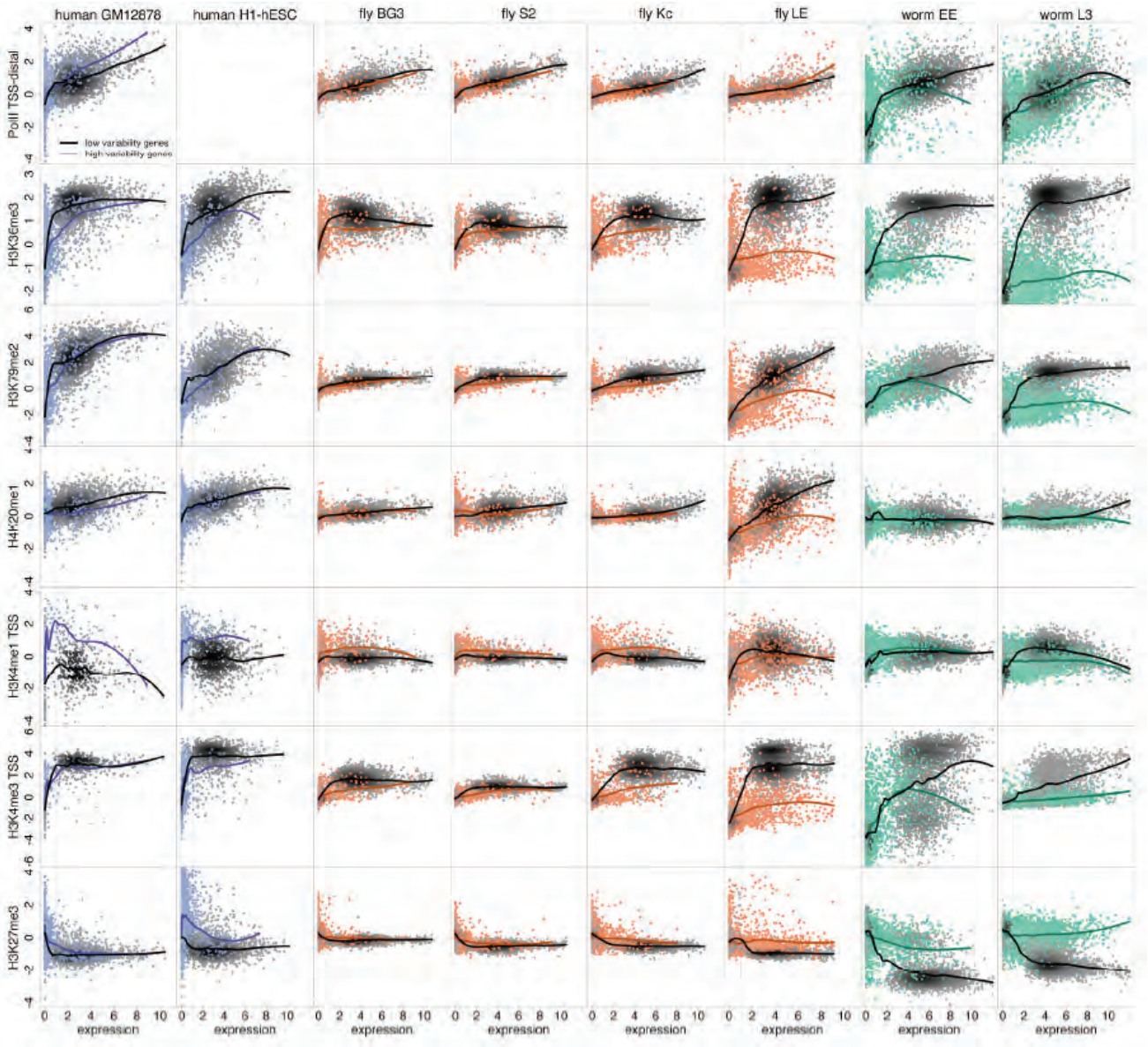
Supplementary Fig. 18. LAD domains are late replicating. a, Distribution of late replicating domains and LADs across human chromosome 2 and fly chromosome 2R. Late replicating domains (red) are shown in human and fly cell lines by plotting the relative RPKM of BrdU-enriched fractions from late S-phase binned across 50 kb (human) and 10 kb (fly) windows. LADs for human[44] and fly[43] are indicated in black. **b,** LLADs are enriched for late replicating sequences and depleted of early replicating sequences. Boxplots depicting the genome-wide distribution of early (green) and late (red) replicating sequences in LAD and random domains for human and fly cell lines. Thus one consistent feature between fly and human is the association of LADs with late replication, which suggests that LADs generally reside in (and may promote) a repressive chromatin environment that impacts both transcription and DNA replication.



Supplementary Fig. 19. DNA shape conservation in nucleosome sequences. a, Consensus ORChID2 profiles as a measure of DNA shape (y-axis) in 146-148 bp nucleosome-associated DNA sequences identified by paired-end MNase-seq in human, fly and worm. A larger value of DNA shape (y-axis) corresponds to a wider minor groove and weaker negative charge. ORChID2 provides a quantitative measure of DNA backbone solvent accessibility, minor groove width, and minor groove electrostatic potential. DNA shape analysis can reveal structural features shared by different sequences that are not apparent in the typical approach of evaluating mono- or di-nucleotide frequencies along nucleosomal DNA, since it can capture structural features in regions with degenerate sequence signatures. Consensus shape profiles, obtained by averaging individual nucleosome-bound sequences aligned by the inferred dyad position, are highly similar across species. **b,** Normalized correlation (similarity) of ORChID2 profile of individual nucleosome-associated sequence with the consensus profile (see Methods and Supplementary Fig. 20). The result indicates that the proportion of sequences that are positively correlated with the consensus profile is higher than would be expected by random in all three species, and this proportion is higher in worm than in fly and human.

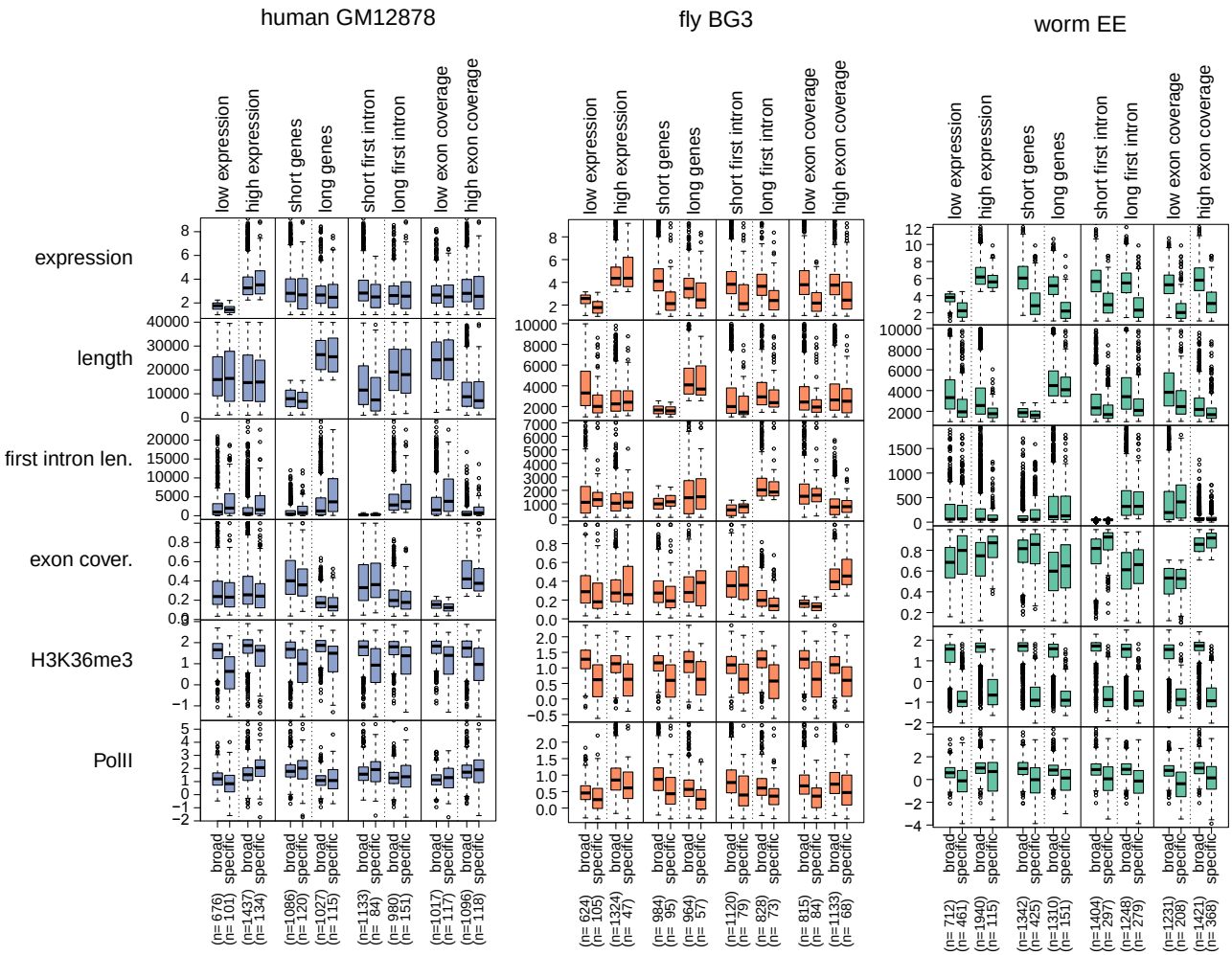


Supplementary Fig. 20. DNA shape in nucleosome sequences. a, Consensus ORChID2 profiles in 146-148 bp nucleosome-associated DNA sequences stratified by average GC content. The subset of GC content sequences used here (GC: 35 % to 55 %) represents 74.4%, 65.6%, and 64.6% of the human, fly, and worm reads, respectively. Note that the worm dataset used here (GSM807109) is a representative of three independent worm MNase-seq datasets. **b**, An outline of the analysis procedure used to evaluate individual sequence DNA shape similarity to the consensus (upper panel) and continuous distributions of similarity scores (lower panel).

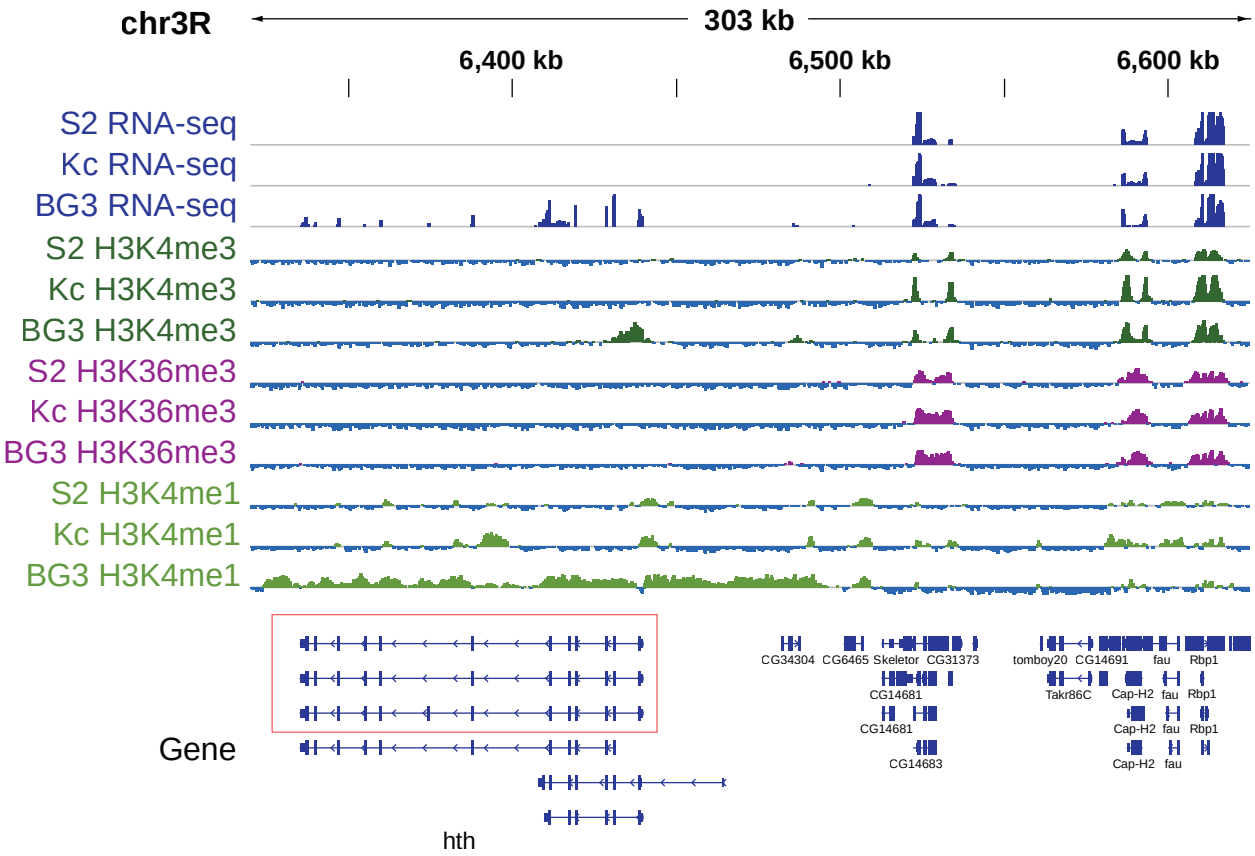


Supplementary Fig. 21. Chromatin context of broadly expressed and specifically expressed genes. ChIP signal enrichment (log_2 scale) of different marks is plotted against gene expression (log_2 scale) for protein coding genes with low expression variability (black points), and high expression variability (colored points), across cell types. ChIP signal enrichment is calculated over the whole gene body for H3K36me3, H3K79me2, H4K20me1 and H3K27me3, within 500 bp of the TSS for H3K4me1 and H3K4me3, and over the gene body excluding the first 500 bp at the 5' end for PolII. Different columns show different cell types as labeled. The expressed gene cut-off of RPKM=1 is denoted with vertical dashed lines. In fly LE and worm L3, most ChIP enrichment and depletion signals appear to be significantly lower in specifically expressed genes. This observation is understood to be due to the different sensitivities of RNA-seq and ChIP-seq protocols when examining samples with heterogeneous cell types. Genes expressed in only a sub-population of the cells can be identified as expressed in RNA-seq assays, but the chromatin signal from the sub-population of cells with these genes actively expressed is washed out by the signal from the remaining cells, where these genes are silent. In human and fly cell lines and worm early embryos, the majority of the marks show similar enrichment and depletion patterns

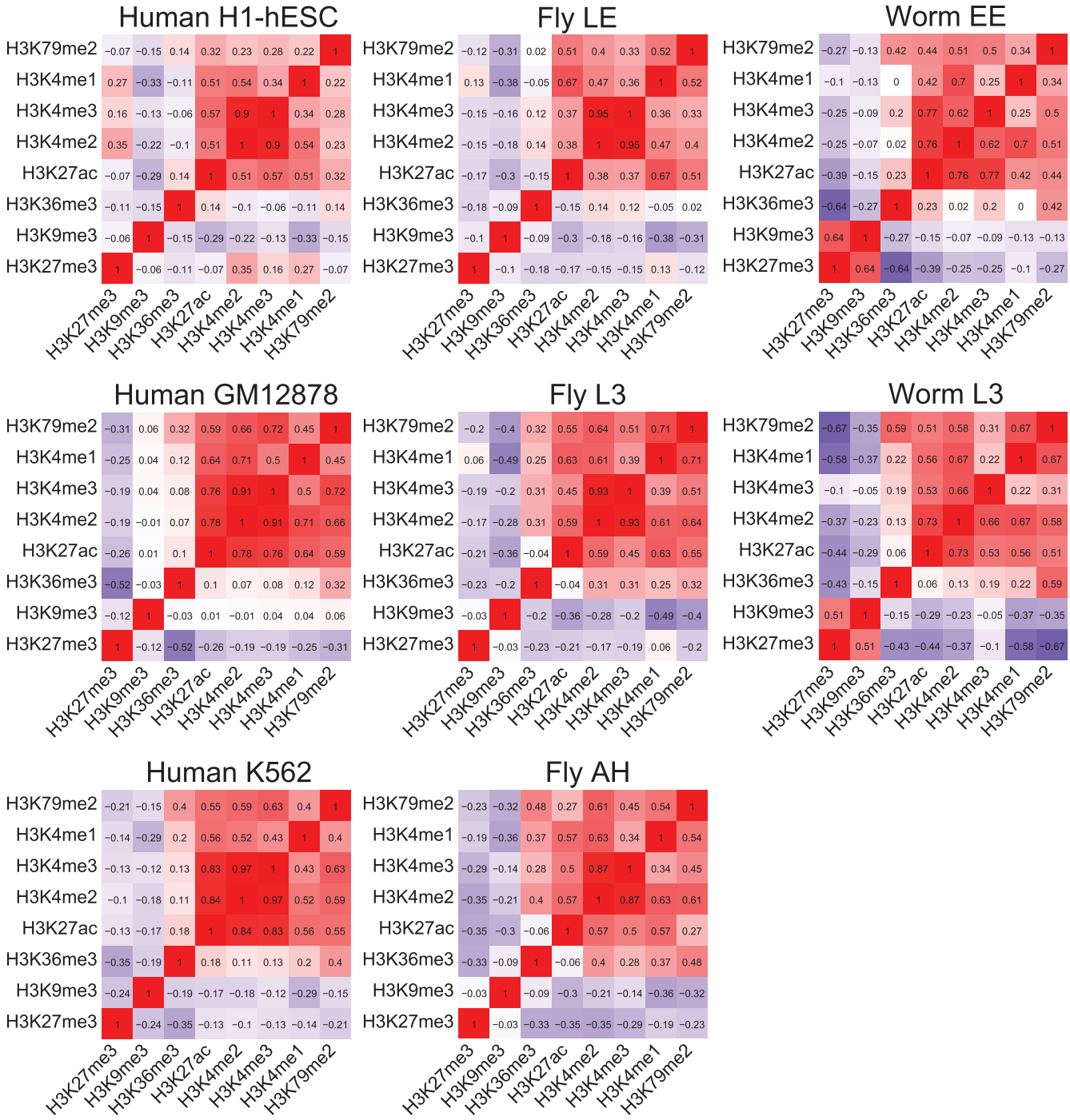
for broadly and specifically expressed genes. Two particular marks show consistent differences in these cell types: H3K4me1 levels are observed to be on average higher in specifically expressed genes relative to broadly expressed genes in both species, consistent with the role of H3K4me1 in marking cell-type specific regulatory regions. On the other hand, H3K36me3 levels are observed to be on average lower in specifically expressed genes relative to broadly expressed genes. This is consistent with previously reported results in fly Kc cells[68] and worm early embryos[69]. We verified that the difference in H3K36me3 levels is not due to differences in gene structure such as gene length, first intron length or exon coverage (Supplementary Fig. 22; See Supplementary Fig. 23 for an example.) However, the differences are much larger in whole animals than in cell lines, suggesting that the observation may be a consequence of sampling mixed cell types, where a large number of transcripts could come from genes enriched for H3K36me3 in only a small fraction of the cells in the population. Consistent with this hypothesis, chromatin signals associated with active gene expression are lower over specifically expressed genes compared to broadly expressed genes in these samples. It is possible that different modes of transcriptional regulation are being utilized, *e.g.*, it is hypothesized that in worm EE, H3K36me3 marking of germline- and broadly expressed genes is carried out by the HMT MES-4, providing epigenetic memory of germline transcription, whereas specifically expressed genes are marked co-transcriptionally by the HMT MET-1[69]. Profiling of chromatin patterns and gene expression in additional individual cell types is needed to test whether cellular heterogeneity fully accounts for our observations.



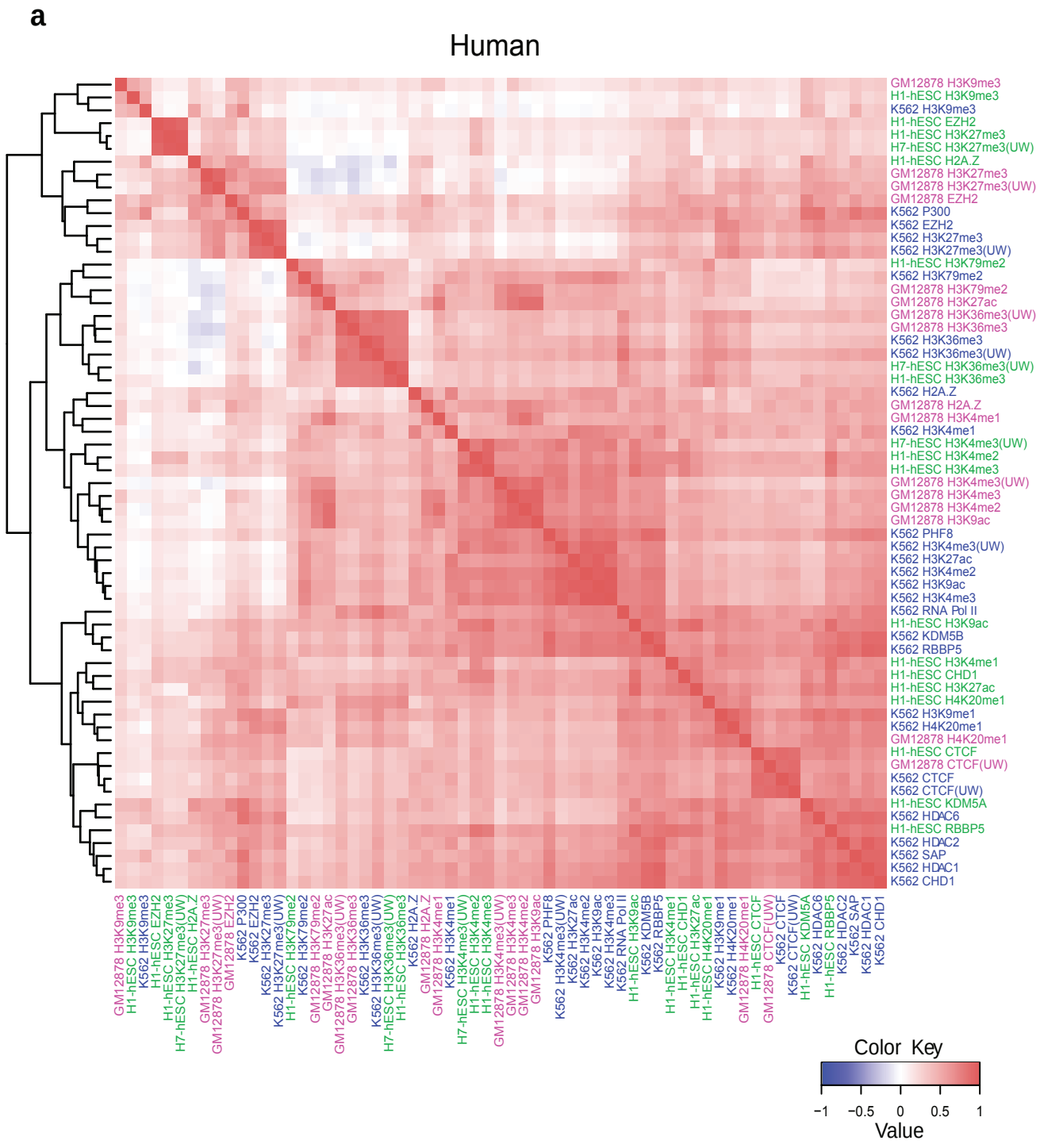
Supplementary Fig. 22. Structure and expression of broadly and specifically expressed genes. Boxplots show gene expression [$\log_2$(RPKM+1)], gene length, first intron length, exon coverage and H3K36me3 and PolII enrichment ($\log_2$) for broadly and specifically expressed genes. The genes are categorized according to expression, gene length, first intron length and exon coverage in the four vertical panels respectively.



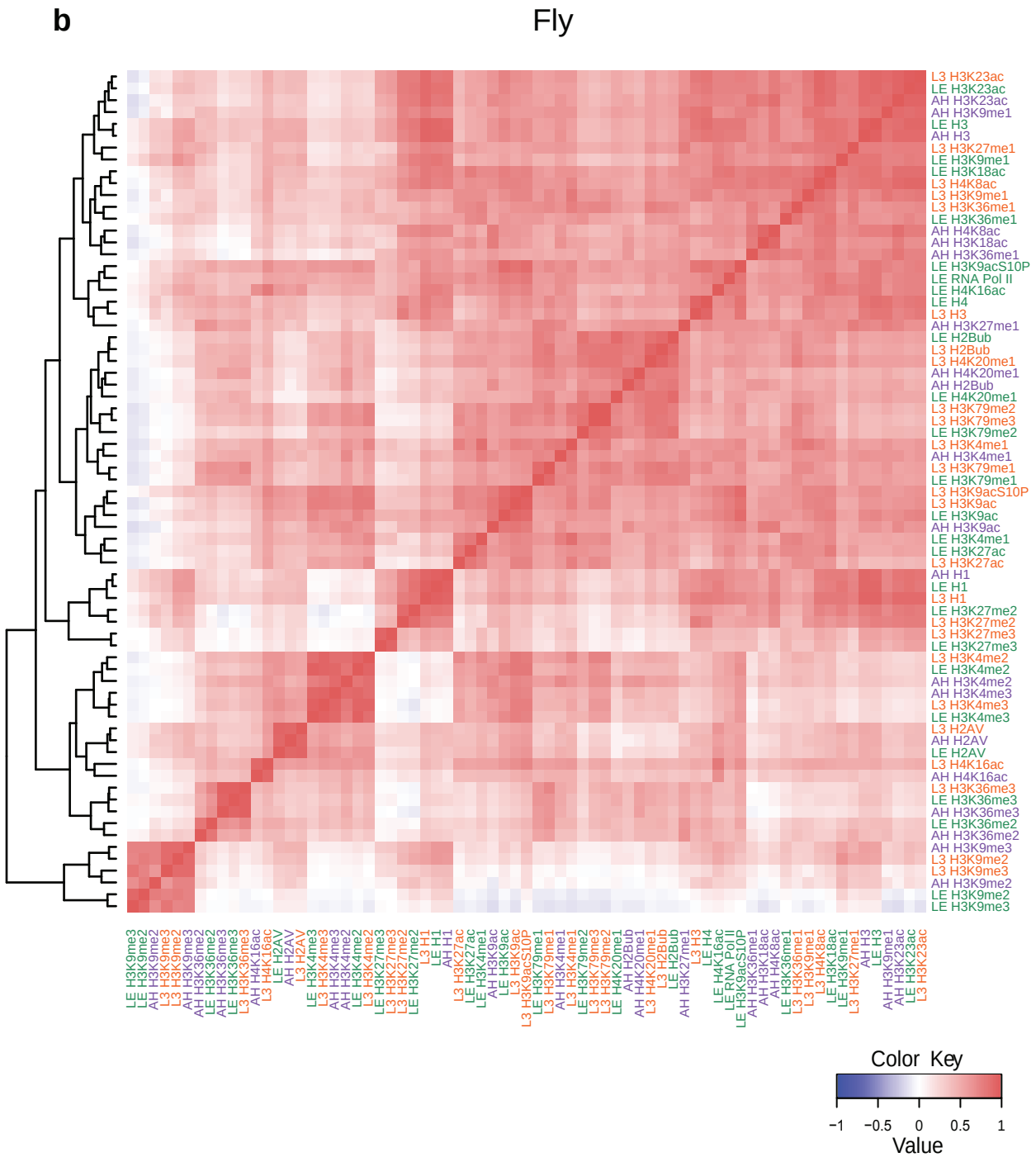
Supplementary Fig. 23. Example genome browser screenshot showing broadly and specifically expressed genes. The fly *hth* gene is specifically expressed in BG3 cells. Cell-type-specific enrichment of H3K4me3 at the TSS and of H3K4me1 over the gene body of *hth* is observed, whereas H3K36me3 remains low over the *hth* gene independent of its expression level.



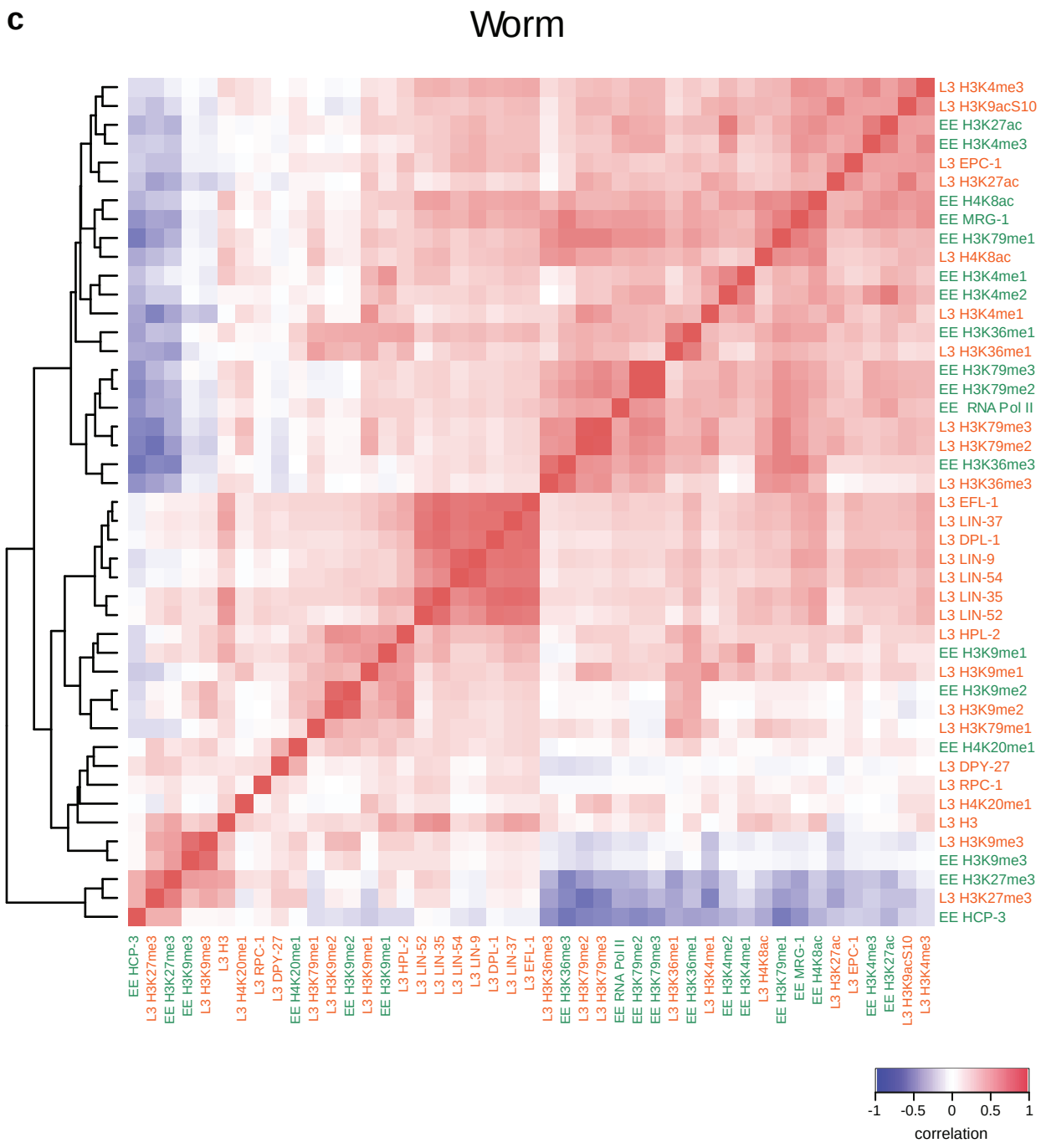
Supplementary Fig. 24. A Pearson correlation matrix of histone marks in each cell type or developmental stage. Each entry in the matrix is the pairwise Pearson correlation between marks across the genome, computed using 5 kb bins across the mappable regions excluding regions with no signal at all (ChIP fold enrichment over input <1 for all 8 marks). This numerical matrix shows the same difference in H3K9me3 and H3K27me3 as reported in Extended Data Fig. 1c. Note that embryonic cell/sample types (H1-hESC in human, LE in fly, and EE in worm) display a higher correlation between H3K4me1 and repressive mark H3K27me3, compared to cells or samples that are more differentiated (GM12878 and K562 in human; L3 and AH in fly; L3 in worm).



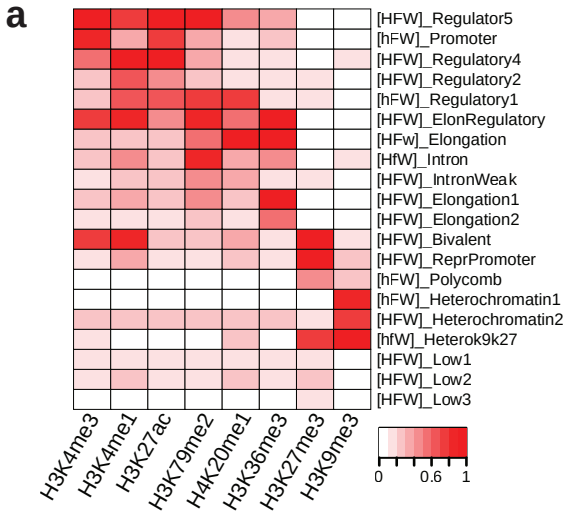
Supplementary Fig. 25. a. (See below for legend)



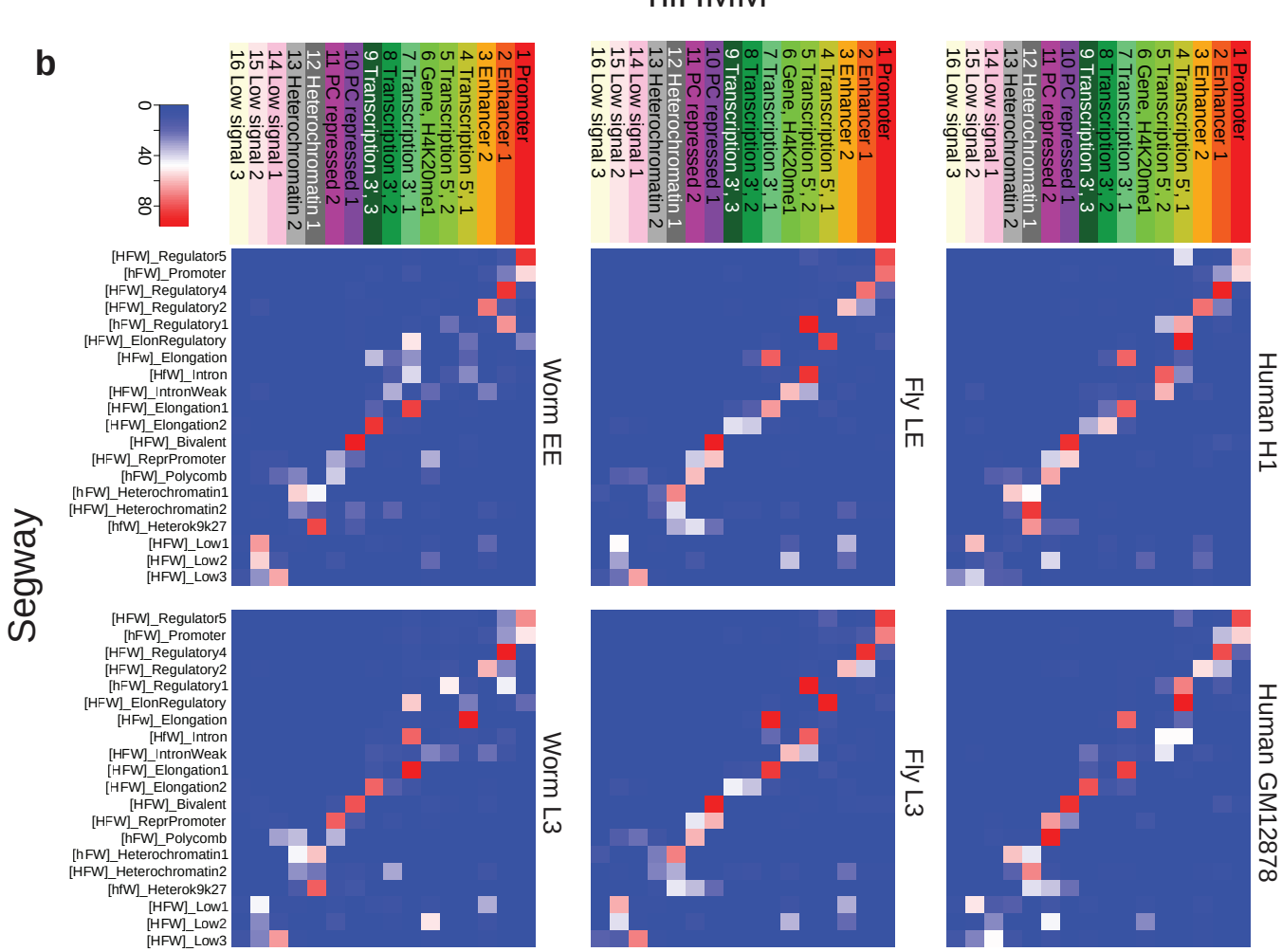
Supplementary Fig. 25. b. (See below for legend)

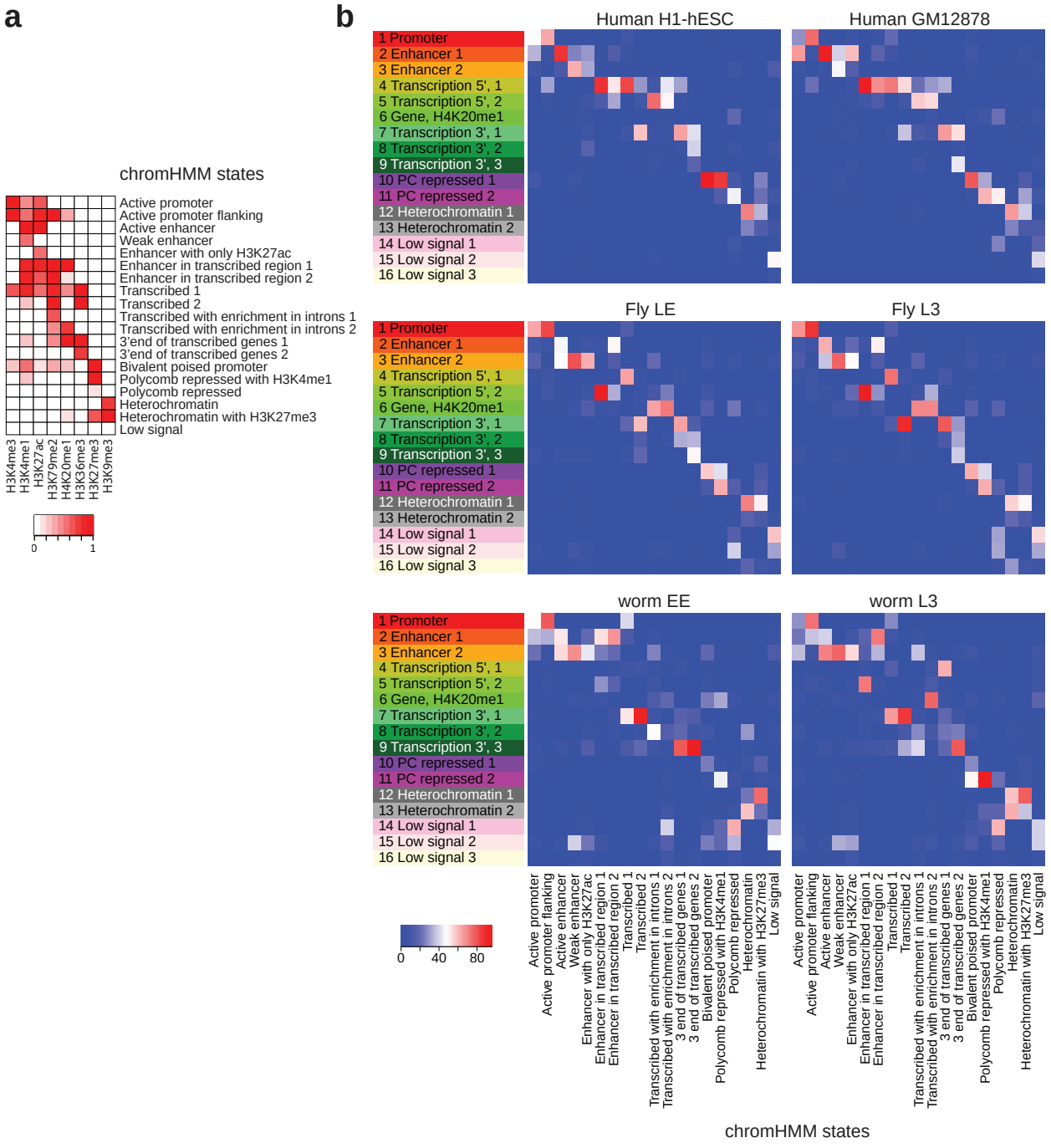


Supplementary Fig. 25. Genome-wide correlation of ChIP-seq datasets for human, fly and worm. a, The genome-wide correlations between chromatin marks and factors for human. Each entry in the heatmap shows the Pearson correlation coefficient between a pair of marks/factors, computed using 30-kb bins across the whole genome. The dendrogram shows the hierarchical clustering result based on Pearson correlation coefficients. Datasets marked with 'UW' were generated at the University of Washington; the rest were generated at the Broad Institute. **b,c,** The same correlation matrices for fly and worm, respectively. The resolution for calculating correlation is 10 kb.

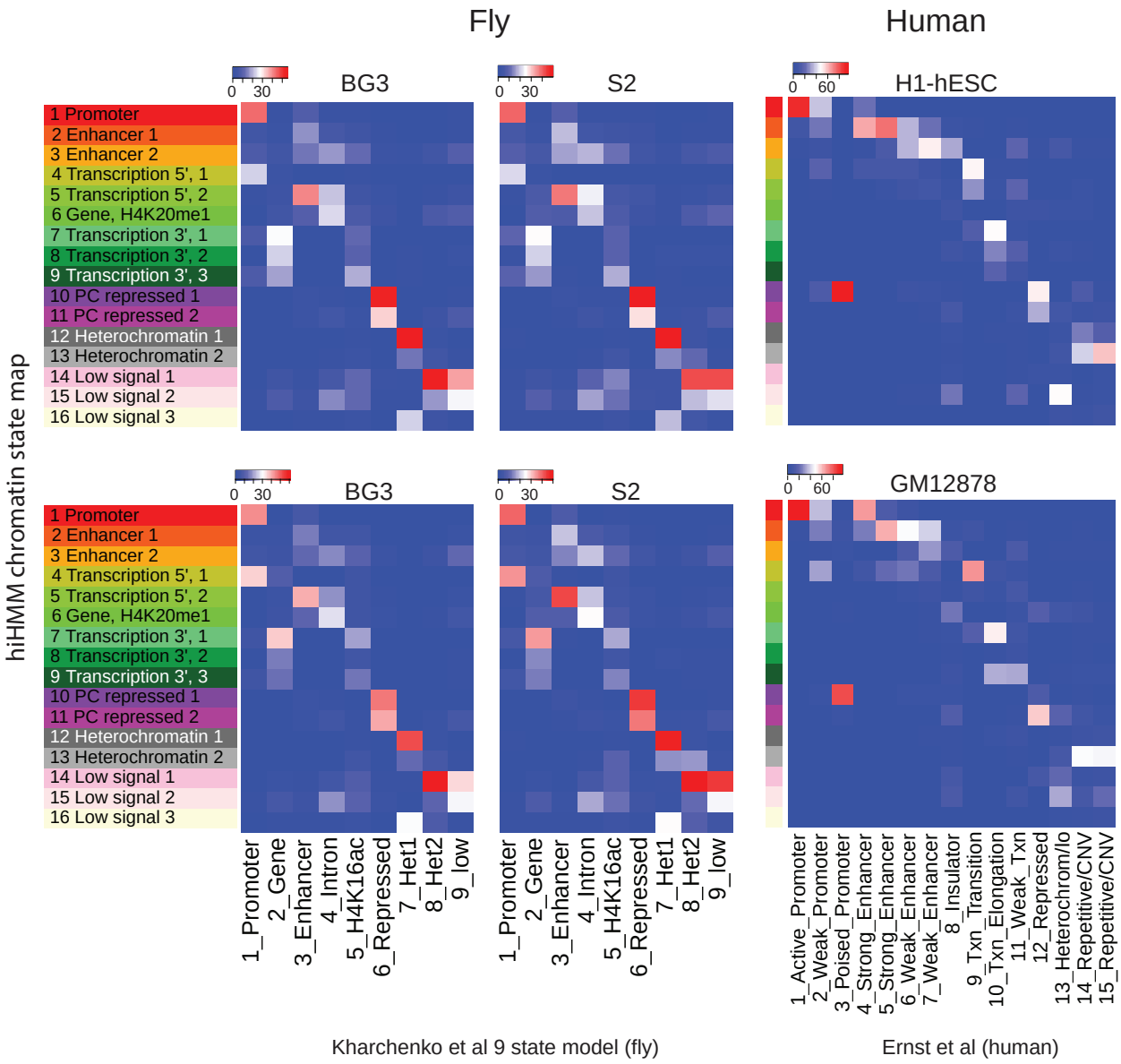


Supplementary Fig. 26. Comparison of chromatin state maps generated by hiHMM and Segway. a, Histone modification enrichment in each of the 20 chromatin states (*i.e.*, emission matrix) identified by Segway. Color represents relative enrichment of a histone mark (scaled between 0 and 1). Letters in brackets preceding each state name indicate coverage within each species (H: human, F: fly, W: worm; upper case: high coverage, lower case: low coverage). In general, Segway identified similar types of states as hiHMM (Fig. 2). **b,** This figure shows the percentage of each hiHMM region that is occupied by each Segway state. The blue-red color bar shows percentage of overlap. There is a strong overlap between certain hiHMM-states and certain Segway states, indicating the identification of analogous states. For example, the Promoter state in hiHMM (state 1) strongly overlaps with Segway state "[hFW]_Promoter". Thus the two algorithms give largely concordant results.

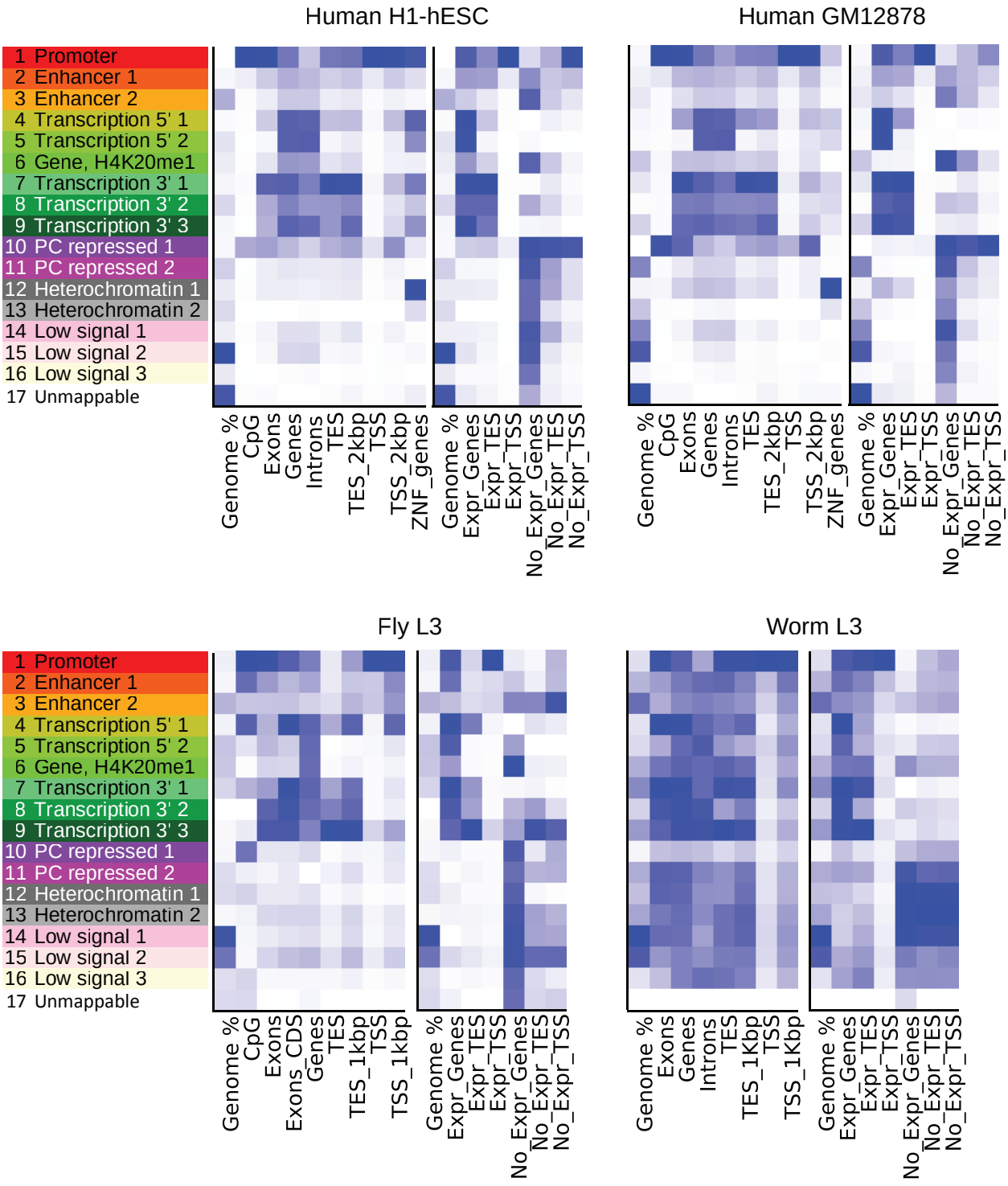




Supplementary Fig. 27. Comparison of chromatin state maps generated by hiHMM and ChromHMM. a, Histone modification enrichment in each of the 19 chromatin states (*i.e.*, emission matrix) identified by ChromHMM. Color represents emission probability given the enrichment of a histone mark. In general, ChromHMM identified similar types of states as did hiHMM (Fig. 2). **b,** This figure shows the percentage of each hiHMM region that is occupied in each ChromHMM state. The blue-red color bar shows percentage of overlap. There is a strong overlap between certain hiHMM-states and certain ChromHMM states, indicating analogous states. For example, the Promoter state in hiHMM (state 1) strongly overlap with ChromHMM states "Active promoter" and more weakly with "Active promoter flanking". Thus the two algorithms give largely concordant results.

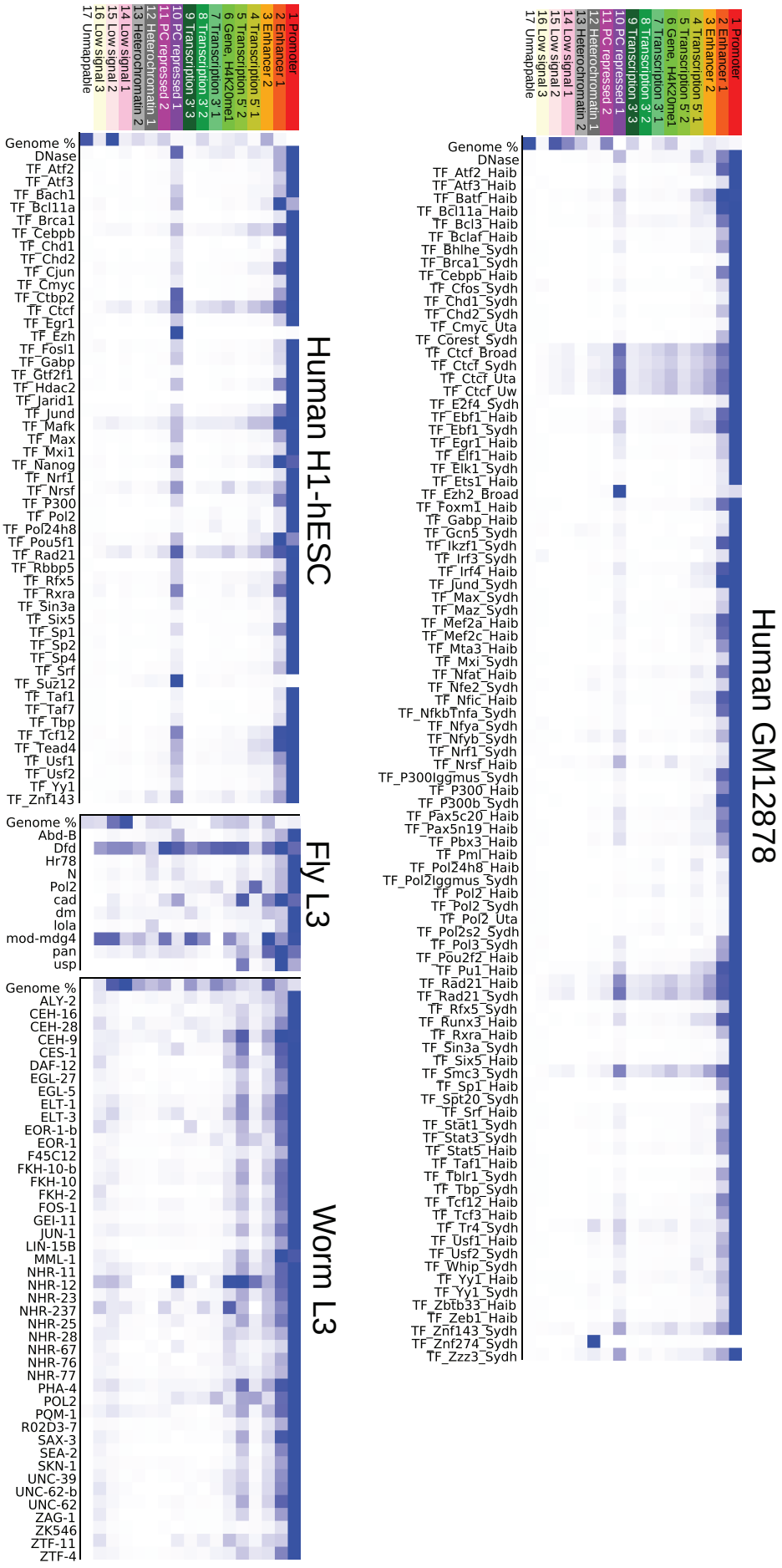


Supplementary Fig. 28. Comparison of hiHMM-based chromatin state model with species-specific models. The fly hiHMM segmentations for LE and L3 were compared to the 9-state model established for fly S2 and BG3 cell lines by Kharchenko[11] *et al*. The human hiHMM segmentations for GM12878 and H1-hESC were compared to their respective chromatin maps produced by ChromHMM[49]. The color bar shows the percentage of a given hiHMM states that is occupied by each species-specific model state. Our chromatin state maps are in agreement with existing species-specific chromatin maps for human[49] and fly[11], even though these state maps were generated using a different number of histone marks and different cell types.

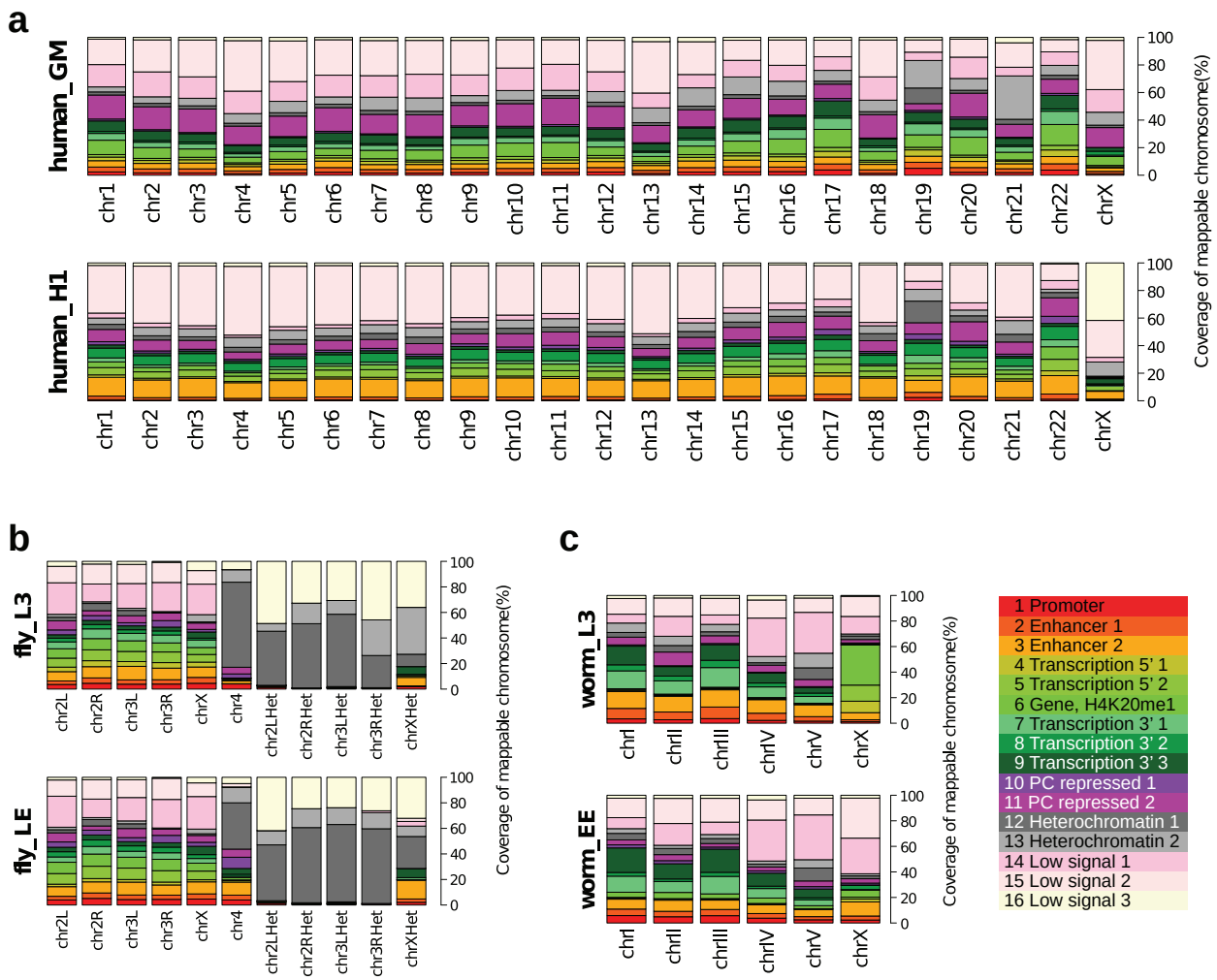


Supplementary Fig. 29. Distribution of genomic features in each hiHMM-based chromatin state. Each entry in the heatmap represents the relative enrichment of that state in a given genomic feature. The scale was normalized between 0 to 1 for each column. In general, similar combinations of histone marks are enriched in each state across the three species (Fig. 2), and each state is enriched for similar genomic features (above).

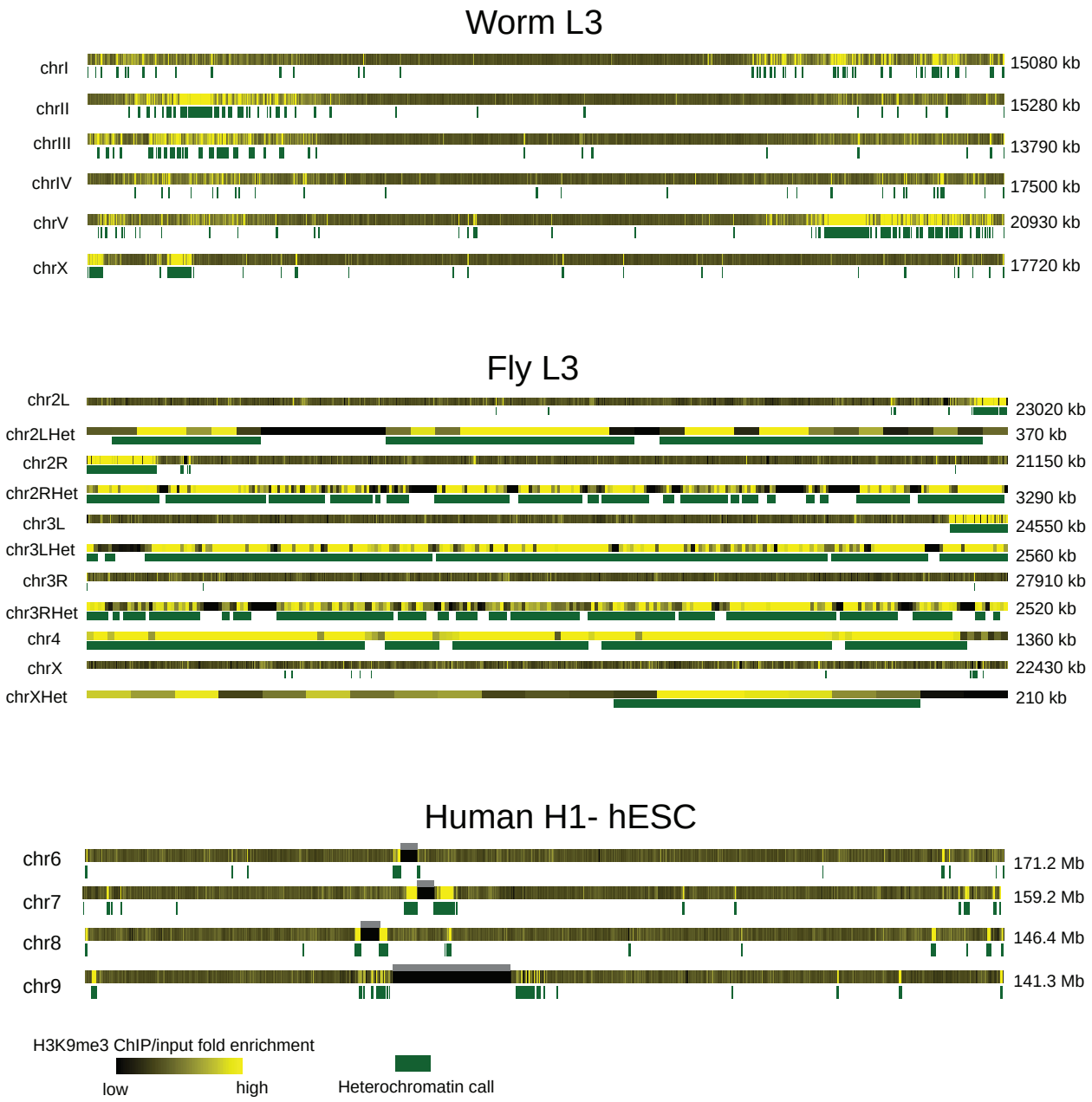
Supplementary Fig. 30. Enrichment of chromosomal proteins in individual chromatin states generated by hiHMM. ChIP enrichment Z-scores computed based on genome-wide mean and standard deviation, for individual non-histone proteins in human H1-hESC and GM12878, fly late embryo (LE) and third instar larvae (L3) and worm early embryo (EE), mixed embryo (MxE) and larvae stage 3 (L3).



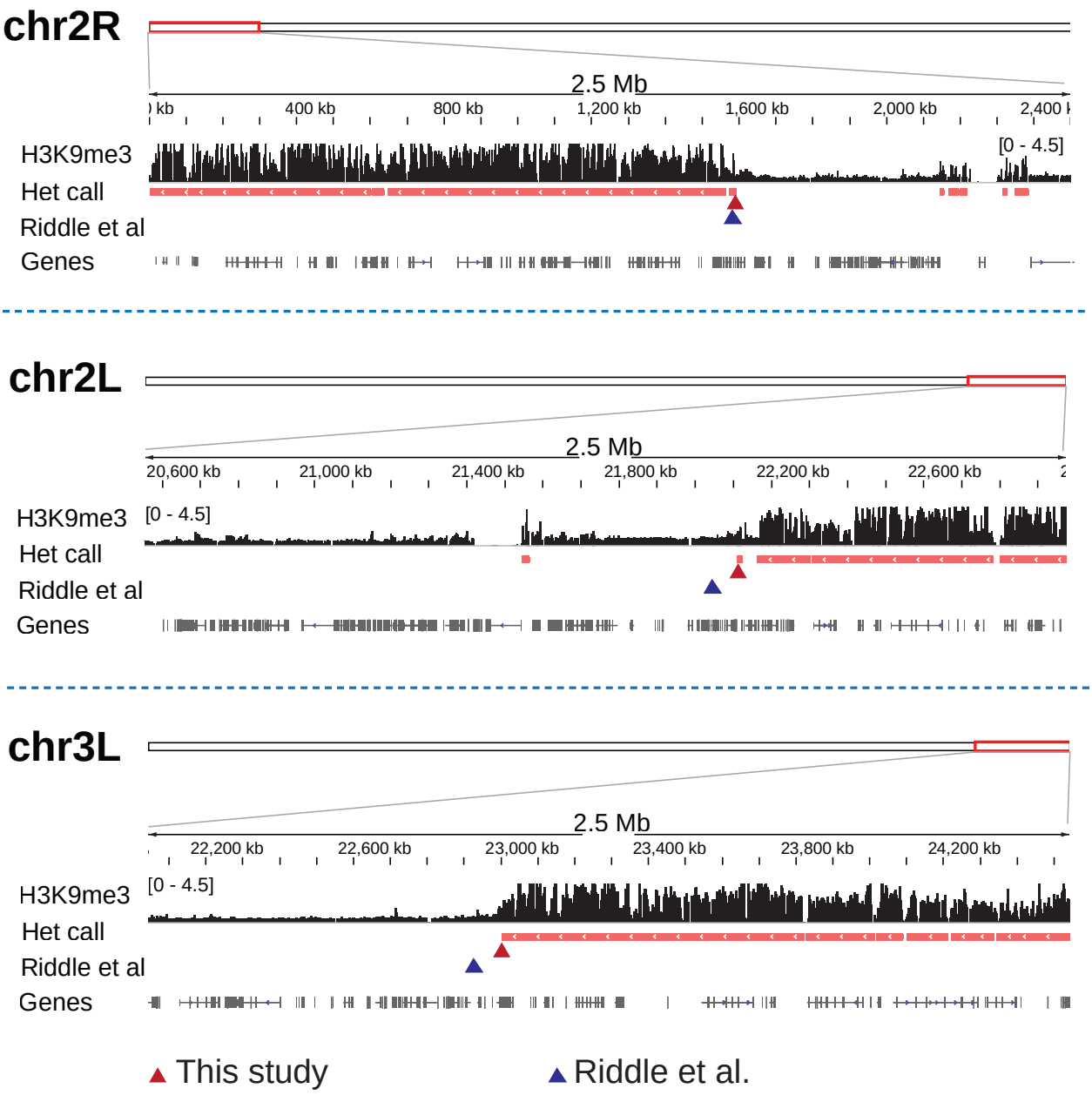
Supplementary Fig. 31. Enrichment of transcription factor binding sites in individual chromatin states generated by hiHMM. Each entry in the heatmap represents the relative enrichment of that state in the entire set of TF binding sites. ChIP enrichment scaled from 0 to 1, for individual transcription factors in human H1-hESC and GM12878, fly third instar larvae (L3) and worm larvae stage 3 (L3). Most transcription factors are associated either with the promoter states or with Pc repression 1, but some are broadly distributed. Note that many transcription factors are associated with state 10, PC Repressed 1, and relatively few with state 11, PC Repressed 2. This result supports that there are two distinct types of Polycomb-associated repressed regions: strong H3K27me3 accompanied by marks for active genes or enhancers (state 10) and strong H3K27me3 without active marks (state 11) (see Fig. 2).



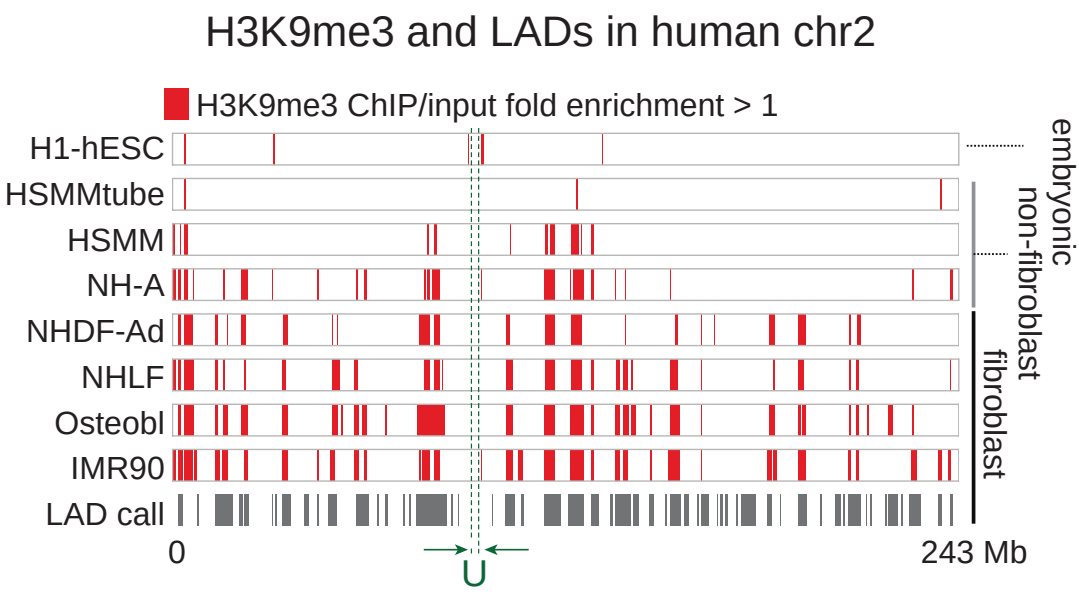
Supplementary Fig. 32. Coverage by hiHMM-based states in mappable regions of individual chromosomes in human (a), fly (b), and worm (c). Fly annotated heterochromatic arms (chr2LHet, chr2RHet, chr3LHet, chr3RHet and chrXHet) are disproportionally enriched for heterochromatin states and "Low signal 3" state, which is consistent with our understanding of the marks enriched in these regions. Similarly, chromosome four is enriched in heterochromatin. Furthermore, in worm, a higher proportion of chrX is covered by the H4K20me1-enriched state 6 in L3 compared to EE, which is consistent with the role of H4K20me1 in worm chrX dosage compensation at L3.



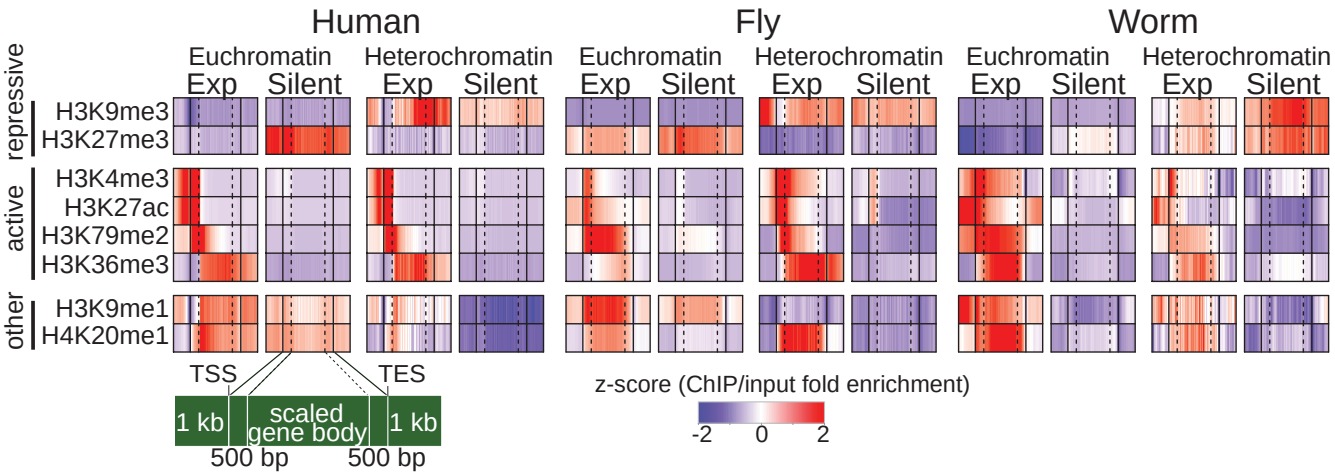
Supplementary Fig. 33. Heterochromatin domains defined based on H3K9me3-enrichment in worm, fly and human. H3K9me3 profiles from worm L3 (upper), fly L3 (middle) and human H1-hESC (bottom) in heatmaps and the identified heterochromatic regions (enriched for H3K9me3; green) are shown. For human, examples are shown for selected chromosomes. Note centromeric regions of human chromosomes are poorly assembled (regions marked with grey above H3K9me3 enrichment heatmap). Significantly enriched regions are determined using a Poison model for ChIP and input tag distributions with a window size of 10 kb (fly and worm) or 100 kb (human), using SPP[31] (see Methods). The majority of the H3K9me3-enriched domains in fly L3 and human H1-hESC are concentrated in the pericentric regions, while in worm L3 they are distributed in subdomains throughout the chromosome arms.



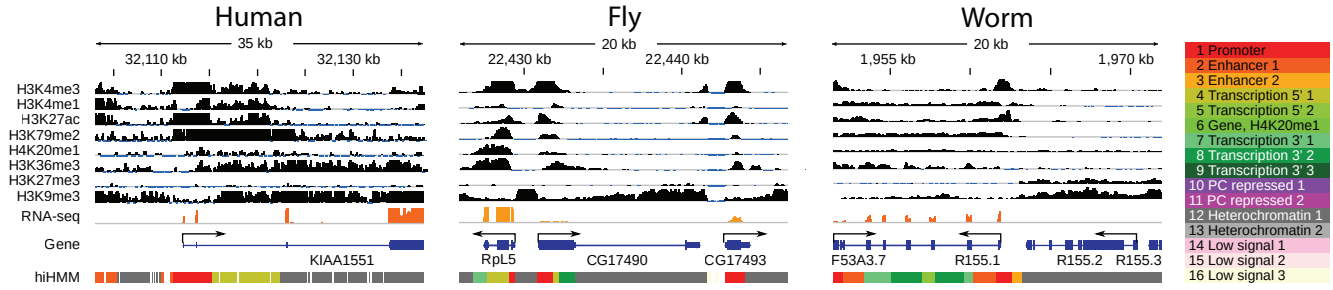
Supplementary Fig. 34. Borders between pericentric heterochromatin and euchromatin in Fly L3 from this study compared to those based on H3K9me2 ChIP-chip data. The screen-shots are from near the pericentric heterochromatic regions in fly chr2R (upper), chr2L (middle) and chr3L (bottom). H3K9me3 ChIP-seq profiles (ChIP/input fold enrichment) are shown in the top rows. Heterochromatin (Het) calls, the regions identified as significantly enriched for H3K9me3 with a 10 kb window, are shown in the middle (see Methods). The end or start sites of continuous H3K9me3 enrichment regions are marked with red triangles. Blue triangles indicate identified borders between pericentric heterochromatin and euchromatin from Riddle *et al.*[14] based on H3K9me2 ChIP-chip profiles. The boundaries between pericentric heterochromatin and euchromatin on each fly chromosome are consistent with those from the lower resolution studies using H3K9me2.



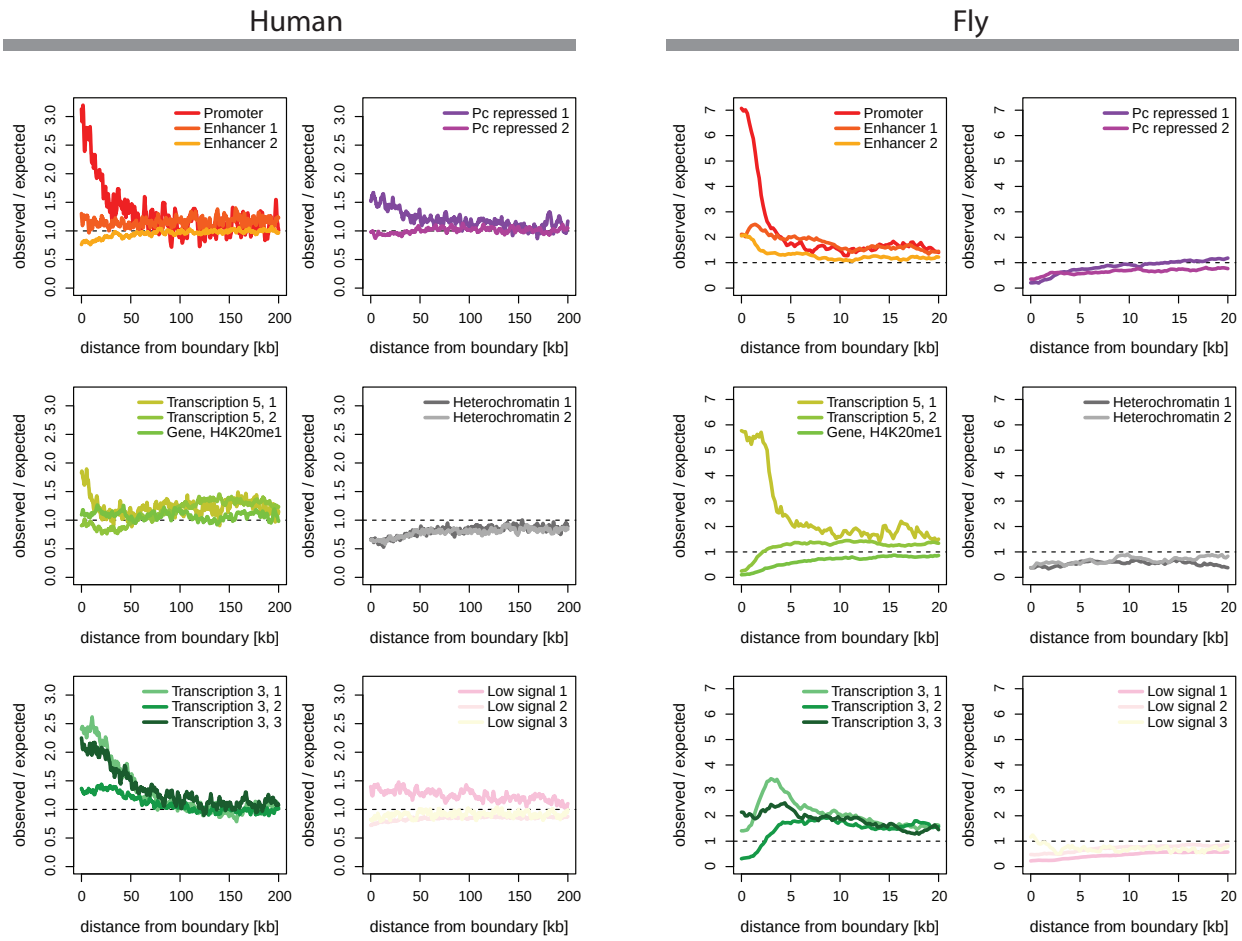
Supplementary Fig. 35. Distribution of H3K9me3 in different cells in human chr2 as an example. U indicates an unassembled region in a centromere. In human differentiated cell types, the regions enriched with H3K9me3 (red blocks) often form large domains (up to ~ 11 Mb in size) distant from the centromere and are cell type-specific. The LADs identified in fibroblasts (Tig3) correlate with the regions enriched for H3K9me3 in other fibroblast or fibroblast-like cell types.



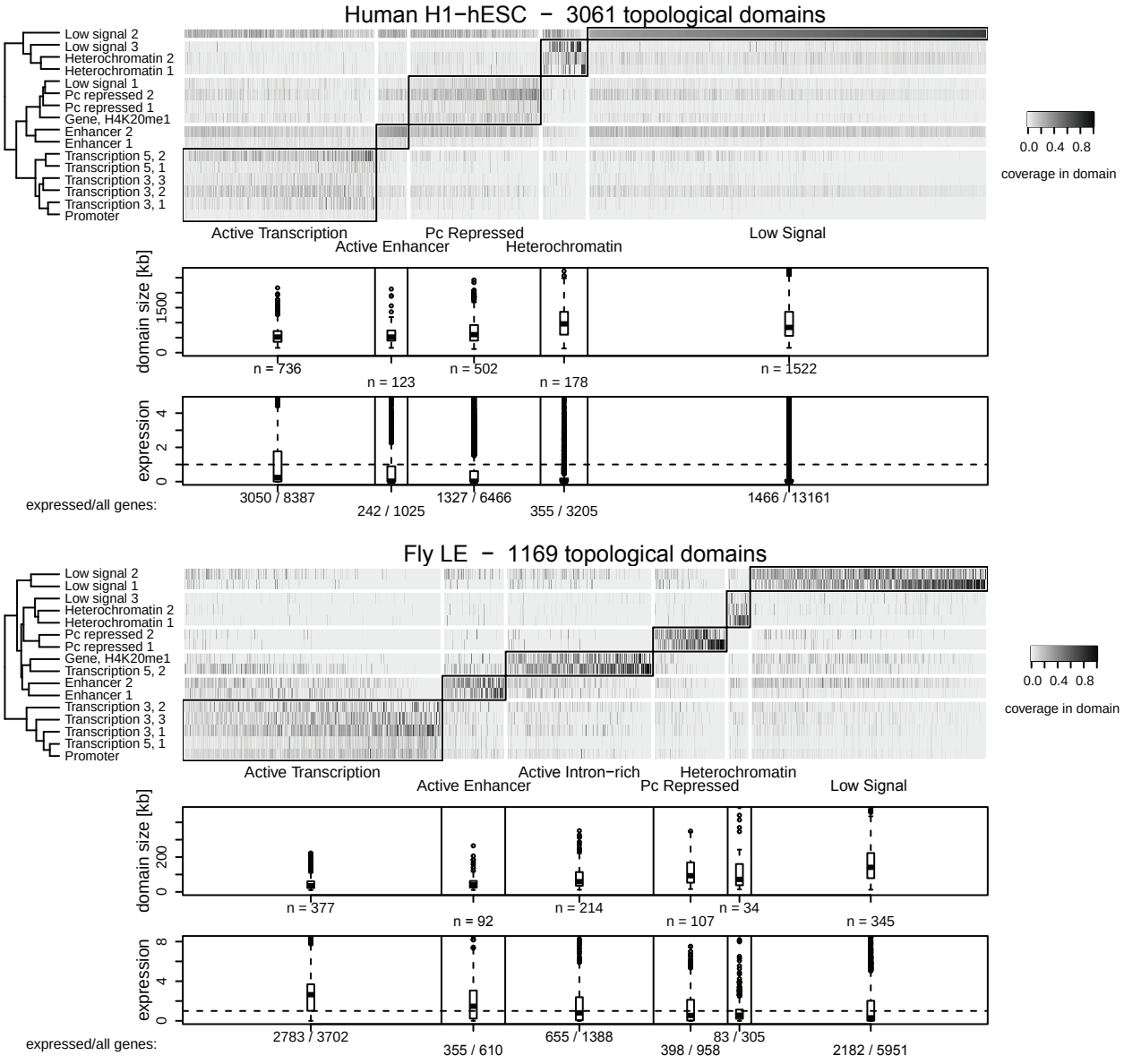
Supplementary Fig. 36. Gene body plots of several histone modifications for euchromatic and heterochromatic genes. Heterochromatic genes are defined as genes in H3K9me3-enrichment regions detected with 10 kb (fly and worm) or 100 kb (human) window (see Methods). Expressed or silent genes are defined using RNA-seq data (see Methods; human K562, fly L3 and worm L3). For human and fly, H3K9me3 is depleted at the TSS of expressed genes in the heterochromatic regions. In worm, H3K9me3 is predominantly confined to gene bodies, with overall lower levels in promoter regions.



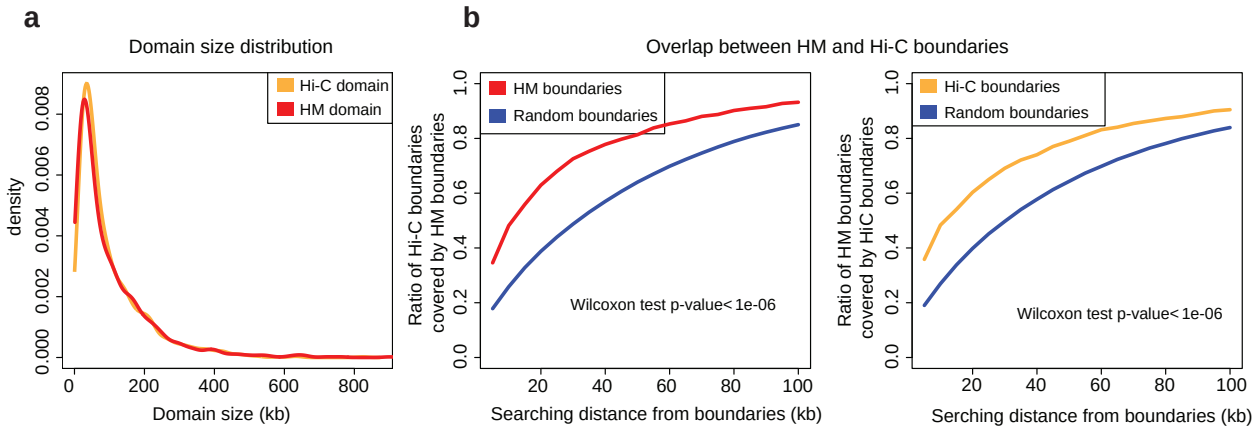
Supplementary Fig. 37. The chromatin state map around three examples of expressed genes in or near heterochromatic regions in human GM12878 cells, fly L3, and worm L3. Expressed genes are enriched for H3K4me3 (state 1, red) at their promoters and exhibit a transcription state across the body of the gene. While H3K9me3 is a hallmark of heterochromatin in all three species (resulting in a heterochromatin state across the body of the gene in this domain; states 12-13, grey), H3K27me3 is also enriched in worm heterochromatin.



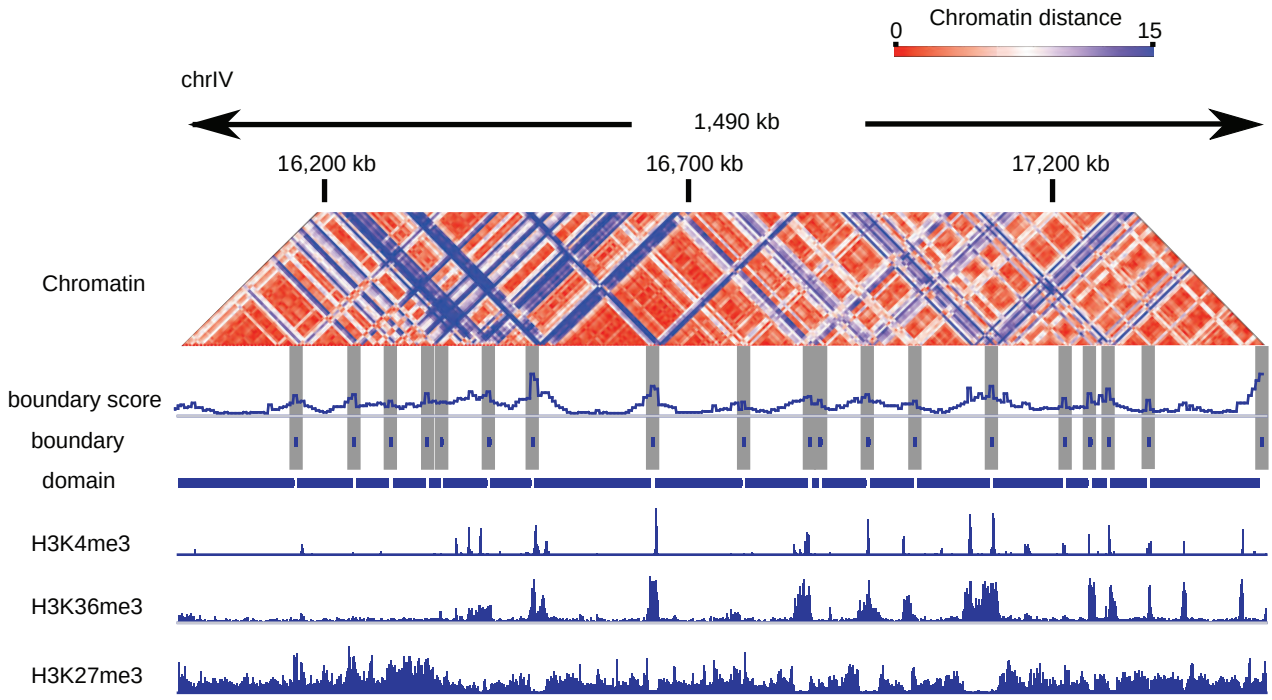
Supplementary Fig. 38. Chromatin context of topological domain boundaries. Observed occurrences of chromatin states near Hi-C defined topological domain boundaries normalized to random expectation. The two species generally show similar enrichment of active states near domain boundary and depletion of low signal and heterochromatin states. In human H1-hESC, the Pc repressed 1 state, which largely marks bivalent regions, is also observed to be enriched near domain boundaries.



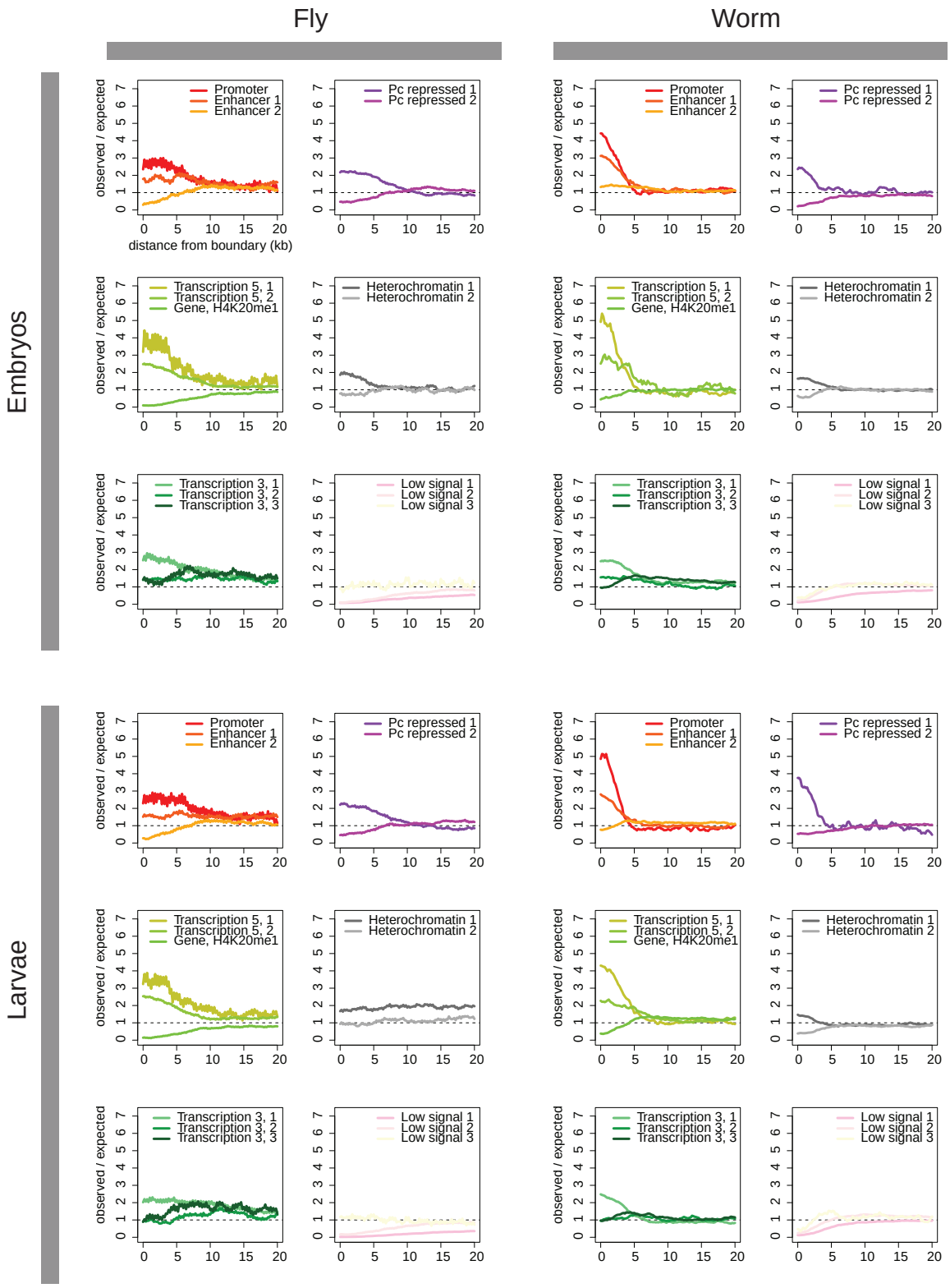
Supplementary Fig. 39. Classification of topological domains based on chromatin states. Coverage of chromatin states (rows) in individual topological domains (columns) is shown as a heatmap for fly late embryos and human H1-hES cells. Chromatin states are clustered according to their co-occurrence correlations in topological domains to identify the labeled meta-chromatin states. (Active Transcription, Active Enhancer, Active Intron-rich, Pc Repressed, Heterochromatin, and Low Signal domains). The topological domains are classified according to the dominant meta state in the domain. The clustering of chromatin states is observed to be generally similar in the two species. One notable exception is that the H4K20me1-enriched state 6 is found in Polycomb repressed domains in human, whereas the same state is enriched in introns of long active genes in fly. These long genes are observed to define a relatively distinct group of topological domains. The distributions of domain sizes and expression levels of genes for the different topological domain classes are also presented as boxplots. Active domains are observed to be smaller in size in both species: in fly LE, 377 (32%) domains are identified as active covering 15% of the fly genome and containing 43% of all active genes (RPKM>1). In human H1-hESC, 736 (24%) domains are identified as active covering 16% of the human genome and contain 47% of all active genes (RPKM>1).



Supplementary Fig. 40. Similarity between fly histone modification domains/boundaries and Hi-C. a, We used the fly histone-modification-defined (HM) boundaries to divide the fly genome into HM domains. We defined fly HM domain as the genomic region in between middle points of two nearby HM boundaries. Fly HM domains (red line) have the same size distribution as Hi-C domains (orange line), with a peak at about 50 kb and a long right tail. **b,** In order to show the significant overlap between boundaries defined from HM and Hi-C, we generated random boundaries through random shuffling while keeping the same domain size distribution for each chromosome. We generated random boundaries for 100 times. We then searched for Hi-C boundaries around HM and around random boundaries (left), as well as for HM boundaries around Hi-C and around random boundaries (right). Blue line is plotted as the average of 100 random boundary sets. Significant overlap between HM and Hi-C boundaries compared to random background is supported by a Wilcoxon test with p-value less than 10^{-6}.



Supplementary Fig. 41. A chromosomal view of the chromatin-based topological domains in worm early embryos at chromosome IV. Similar to Fig 3c, we generated chromatin-based topological domains in worm early embryos. Local histone modification similarity (Euclidian distance) is presented as a heatmap. Red indicates higher similarity and blue indicates lower similarity. Chromatin-defined boundary scores, boundaries and domains locations are compared to histone marks in the same chromosomal regions. There is no available Hi-C data in worm that could be used for direct comparison. Nonetheless, the enrichment of H3K4me3 and H3K36me3 at the predicted boundaries is reminiscent of what is observed in Hi-C topological boundaries in human and fly (Supplementary Figs. 38, 39). This analysis supports the idea that domain prediction based on genome-wide similarities of histone modifications may be used to discover large-scale topological domains without the need for HiC data.



Supplementary Fig. 42. Chromatin context of chromatin-based topological domain boundaries. Observed occurrences of chromatin states near histone marks defined domain boundaries normalized to random expectation. Fly late embryos, worm early embryos, fly and worm L3 larvae show enrichment of promoter and active transcription states near domain boundaries and depletion of low signal and heterochromatin states, similar to the observation at Hi-C boundaries (see Supplementary Fig. 38).

Supplementary Table 1. Abbreviation used key cell types and developmental stages described in this study.

Species	Abbreviation	Description
C. elegans	EE	Early embryos
	MXEMB	Mixed embryos
	LTEMB	Late embryos
	L3	Stage 3 larvae
	L4	Stage 4 larvae
	AD no embryos	Feminized adults that produce oocytes but no sperm, and therefore do not contain embryos (*fem-2(b245ts*) strain)
	AD germline	Purified germline nuclei from wildtype hermaphrodites (*ojIs9* strain carrying *zyg-12::gfp* transgene)
	AD-germlineless	AD without germline (*glp-4(bn2ts)* strain)
D. melanogaster	EE	Early embryos (2-4hr)
	LE	Late embryos (14-16hr)
	L3	Third instar larvae
	AH	Adult heads
	ES5,ES10,ES14	Embryonic stages 5, 10, and 14, respectively
	S2	S2-DRSC cell line: derived from late embryonic stage
	Kc	Kc157 cell line: dorsal closure stage
	BG3	ML-DmBG3-c2 cell line: central nervous system, derived from L3
	Clone 8	CME W1 Cl.8+ cell line: dorsal mesothoracic disc
H. sapiens	H1-hESC	Embryonic stem cells
	GM12878	B-lymphocytes
	K562	Myelogenous leukemia cell line
	A549	Epithelial cell line derived from a lung carcinoma tissue
	HeLa-S3	Cervical carcinoma cell line
	HepG2	Hepatocellular carcinoma
	HSMM	Skeletal muscle myoblasts
	HSMMtube	Skeletal muscle myotubes differentiated from the HSMM cell line
	HUVEC	Human umbilical vein endothelial cells
	IMR90	Fetal lung fibroblasts
	NH-A	Astrocytes
	NHDF-Ad	Adult dermal fibroblasts
	NHEK	Epidermal keratinocytes
	NHLF	Lung fibroblasts
	Osteobl	Osteoblasts (NHOst)

More information on the cell types and stages can be found at the project websites:

Details of *D. melanogaster* cell lines*:* https://dgrc.cgb.indiana.edu/project/index.html

Details of *H. sapiens*: http://encodeproject.org/ENCODE/cellTypes.html

Supplementary Table 2. List of protein names used in this study.

Name used	Official name			Name used	Official name			Name used	Official name		
	human	fly	worm		human	fly	worm		human	fly	worm
CHD1	CHD1			CBX2	CBX2			AMA-1			AMA-1
CHD2	CHD2			CBX3	CBX3			ASH-2			ASH-2
CHD3/MI-2/LET-418	CHD3	MI-2	LET-418	CBX8	CBX8			CEC-3			CEC-3
CBP/CBP-1	CREBBP	CBP	CBP-1	CEBPB	CEBPB			CEC-7			CEC-7
CTCF	CTCF	CTCF		CHD7	CHD7			COH-1			COH-1
EZH2/E(Z)	EZH2	E(Z)		CTCFL	CTCFL			COH-3			COH-3
HDAC1/RPD3/HAD-1	HDAC1	RPD3	HDA-1	REST	REST			DPL-1			DPL-1
HP1A		SU(VAR)205		SAP30	SAP30			EFL-1			EFL-1
HP1B/HPL-2		HP1B	HPL-2	SIRT6	SIRT6			EPC-1			EPC-1
HP1C		HP1C		NCOR	NCOR1			HCP-3			HCP-3
HP2		HP2		NSD2	WHSC1			HCP-4			HCP-4
HP4		HP4		P300	EP300			HIM-17			HIM-17
KDM1A	KDM1A	SU(VAR)3-3		PCAF	KAT2B			HIM-3			HIM-3
KDM2		KDM2	T26A5.5	PHF8	PHF8			HIM-5			HIM-5
KDM4A	KDM4A	KDM4A		RBBP5	RBBP5			HIM-8			HIM-8
KDM5A	KDM5A			ACF1		ACF1		HTP-3			HTP-3
KDM5B	KDM5B			ASH1		ASH1		HTZ-1			HTZ-1
KDM5C	KDM5C			BEAF		BEAF-32		IMB-1			IMB-1
RNF2/RING	RNF2	SCE		CG10630		BLANKS		KLE-2			KLE-2
HDAC11		HDACX		BRE1		BRE1		LEM-2			LEM-2
HDAC2	HDAC2			CHRO		CHRO		LIN-35			LIN-35
HDAC3		HDAC3		CP190		CP190		LIN-37			LIN-37
HDAC4a		HDAC4		GAF		GAF		LIN-52			LIN-52
HDAC6	HDAC6	HDAC6		ISWI		ISWI		LIN-53			LIN-53
HDAC8	HDAC8			JIL-1		JIL-1		LIN-54			LIN-54
MOD(MDG4)		MOD(MDG4)		LBR		LBR		LIN-61			LIN-61
NURF301/NURF-1		E(BX)	NURF-1	MBD-R2		MBD-R2		LIN-9			LIN-9
PR-SET7		PR-SET7		MLE		MLE		MAU-2			MAU-2
SMC3	SMC3	CAP		MOF		MOF		MES-4			MES-4
SMARCA4	SMARCA4			MRG15		MRG15		MRE-11			MRE-11
SU(HW)		SU(HW)		MSL-1		MSL-1		MIS-12			MIS-12
SU(VAR)3-7		SU(VAR)3-7		PC		PC		MIX-1			MIX-1
SU(VAR)3-9		SU(VAR)3-9		PCL		PCL		MRG-1			MRG-1
SUZ12	SUZ12			PHO		PHO		MSH-5			MSH-5
SETDB1	SETDB1			PIWI		PIWI		REC-8			REC-8
DPY-26			DPY-26	POF		POF		RPC-1			RPC-1
DPY-27			DPY-27	PSC		PSC		SCC-1			SCC-1
DPY-28			DPY-28	RHINO		RHINO		SDC-1			SDC-1
DPY-30			DPY-30	SFMBT		SFMBT		SDC-2			SDC-2
LIN-15B			LIN-15B	SPT16		DRE4		SDC-3			SDC-3
TAG-315			TAG-315	TOP2		TOP2		SMC-4			SMC-4
MYS3			LSY-12	WDS		WDS		SMC-6			SMC-6
NPP-13			NPP-13	XNP		XNP		ZIM-1			ZIM-1
PQN-85			PQN-85	ZW5		ZW5		ZIM-3			ZIM-3
RAD-51			RAD-51	TAF-1			TAF-1	ZFP-1			ZFP-1
				TBP-1			TBP-1	ZHP-3			ZHP-3

Some of the names used in this study (highlighted in red) are different from their official names.

Supplementary Table 3. Overlap of DHS-based and p300-peak-based enhancers in human cell lines. The analysis generally identifies more DHS-based enhancers than p300-based enhancers. Between 60% and 90% of the p300-enhancers are within 500 bp of a DHS-based enhancer, suggesting that p300 binding sites are generally in DHSs.

	DHS enhancer	w/ p300 enhancer within 100bp	w/ p300 enhancer within 500bp	p300 enhancer	w/ DHS enhancer within 100bp	w/ DHS enhancer within 500bp
GM12878	40531	14094 (35%)	19190 (47%)	29108	14067 (48%)	18391 (63%)
H1-hESC	73496	1973 (3%)	3404 (5%)	3986	2007 (50%)	3139 (79%)
K562	69865	27521 (39%)	34714 (50%)	43659	27489 (63%)	33449 (77%)
HeLa-S3	63189	16784 (27%)	21515 (34%)	22861	16762 (73%)	20217 (88%)

Supplementary references

28 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**, doi:10.1186/gb-2009-10-3-r25 (2009).

29 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).

30 Kundaje, A. *et al.* Adaptive calibrated measures for rapid automated quality control of massive collections of ChIP-seq experiments. *Submitted* (2013).

31 Kharchenko, P. V., Tolstorukov, M. Y. & Park, P. J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotech* **26**, 1351-1359, doi:10.1038/nbt.1508 (2008).

32 Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**, doi:10.1186/gb-2008-9-9-r137 (2008).

33 Adams, M. D. *et al.* The genome sequence of Drosophila melanogaster. *Science (New York, N.Y.)* **287**, 2185-2195 (2000).

34 Song, J. *et al.* Model-based analysis of two-color arrays (MA2C). *Genome Biology* **8**, doi:10.1186/gb-2007-8-8-r178 (2007).

35 Larschan, E. *et al.* X chromosome dosage compensation via enhanced transcriptional elongation in Drosophila. *Nature* **471**, 115-118, doi:10.1038/nature09757 (2011).

36 Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science (New York, N.Y.)* **322**, 1845-1848, doi:10.1126/science.1162228 (2008).

37 Thomas, S. *et al.* Dynamic reprogramming of chromatin accessibility during Drosophila embryo development. *Genome Biology* **12**, doi:10.1186/gb-2011-12-5-r43 (2011).

38 Tolstorukov, M. Y. *et al.* Histone variant H2A.Bbd is associated with active transcription and mRNA processing in human cells. *Molecular cell* **47**, 596-607, doi:10.1016/j.molcel.2012.06.011 (2012).

39 Tolstorukov, M. Y., Kharchenko, P. V., Goldman, J. A., Kingston, R. E. & Park, P. J. Comparative analysis of H2A.Z nucleosome organization in the human and yeast genomes. *Genome Research* **19**, 967-977, doi:10.1101/gr.084830.108 (2009).

40 Harrow, J. *et al.* GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research* **22**, 1760-1774, doi:10.1101/gr.135350.111 (2012).

41 Hothorn, T., Buehlmann, P., Kneib, T., Schmid, M. & Hofner, B. mboost: Model-Based Boosting, R package version 2. *Journal of Machine Learning Research* **11**, 2109-2113 (2010).

42 Ikegami, K., Egelhofer, T. A., Strome, S. & Lieb, J. D. Caenorhabditis elegans chromosome arms are anchored to the nuclear membrane via discontinuous association with LEM-2. *Genome Biology* **11**, doi:10.1186/gb-2010-11-12-r120 (2010).

43 van Bemmel, J. G. *et al.* The insulator protein SU(HW) fine-tunes nuclear lamina interactions of the Drosophila genome. *PloS one* **5**, doi:10.1371/journal.pone.0015013 (2010).

44 Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948-951, doi:10.1038/nature06947 (2008).

45 Bishop, E. P. *et al.* A map of minor groove shape and electrostatic potential from hydroxyl radical cleavage patterns of DNA. *ACS chemical biology* **6**, 1314-1320, doi:10.1021/cb200155t (2011).

46 Sohn, K.-A., Ghahramani, Z. & Xing, E. P. Robust Estimation of Local Genetic Ancestry in Admixed Populations using a Non-parametric Bayesian Approach. *Genetics* **191**, 1295-1308 (2012).

47 Van Gael, J., Saatci, Y., Teh, Y. W. & Ghahramani, Z. in *Proc. Int. Conf. on Machine Learning.* 1088-1095.
48 Hoffman, M. M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature methods* **9**, 473-476, doi:10.1038/nmeth.1937 (2012).
49 Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods* **9**, 215-216, doi:10.1038/nmeth.1906 (2012).
50 Deal, R. B., Henikoff, J. G. & Henikoff, S. Genome-wide kinetics of nucleosome turnover determined by metabolic labeling of histones. *Science* **328**, 1161-1164, doi:10.1126/science.1186777 (2010).
51 Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854-858, doi:10.1038/nature07730 (2009).
52 Arnold, C. D. *et al.* Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science (New York, N.Y.)*, doi:10.1126/science.1232542 (2013).
53 He, H. H. *et al.* Nucleosome dynamics define transcriptional enhancers. *Nat Genet* **42**, 343-347, doi:10.1038/ng.545 (2010).
54 Gaffney, D. J. *et al.* Controls of nucleosome positioning in the human genome. *PLoS genetics* **8**, doi:10.1371/journal.pgen.1003036 (2012).
55 Steiner, F. A., Talbert, P. B., Kasinathan, S., Deal, R. B. & Henikoff, S. Cell-type-specific nuclei purification from whole animals for genome-wide expression and chromatin profiling. *Genome research* **22**, 766-777, doi:10.1101/gr.131748.111 (2012).
56 Teves, S. S. & Henikoff, S. Heat shock reduces stalled RNA polymerase II and nucleosome turnover genome-wide. *Genes & development* **25**, 2387-2397, doi:10.1101/gad.177675.111 (2011).
57 Tillo, D. *et al.* High nucleosome occupancy is encoded at human regulatory sequences. *PloS one* **5**, doi:10.1371/journal.pone.0009129 (2010).
58 Henikoff, S., Henikoff, J. G., Sakai, A., Loeb, G. B. & Ahmad, K. Genome-wide profiling of salt fractions maps physical properties of chromatin. *Genome research* **19**, 460-469, doi:10.1101/gr.087619.108 (2009).
59 Core, L. J. *et al.* Defining the Status of RNA Polymerase at Promoters. *Cell Reports* **2**, 1025-1035, doi:10.1016/j.celrep.2012.08.034 (2012).
60 Yuan, G.-C. *et al.* Genome-scale identification of nucleosome positions in S. cerevisiae. *Science (New York, N.Y.)* **309**, 626-630, doi:10.1126/science.1112178 (2005).
61 Valouev, A. *et al.* Determinants of nucleosome organization in primary human cells. *Nature* **474**, 516-520, doi:10.1038/nature10002 (2011).
62 Schones, D. *et al.* Dynamic Regulation of Nucleosome Positioning in the Human Genome. *Cell* **132**, 887-898, doi:10.1016/j.cell.2008.02.022 (2008).
63 Gilchrist, D. A. *et al.* Pausing of RNA Polymerase II Disrupts DNA-Specified Nucleosome Organization to Enable Precise Gene Regulation. *Cell* **143**, 540-551, doi:10.1016/j.cell.2010.10.004 (2010).
64 Ercan, S., Lubling, Y., Segal, E. & Lieb, J. D. High nucleosome occupancy is encoded at X-linked gene promoters in C. elegans. *Genome research* **21**, 237-244, doi:10.1101/gr.115931.110 (2011).
65 Jin, C. *et al.* H3.3/H2A.Z double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions. *Nature genetics* **41**, 941-945, doi:10.1038/ng.409 (2009).
66 Henikoff, J. G., Belsky, J. A., Krassovsky, K., MacAlpine, D. M. & Henikoff, S. Epigenome characterization at single base-pair resolution. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 18318-18323, doi:10.1073/pnas.1110731108 (2011).

67 Cherbas, L. *et al.* The transcriptional diversity of 25 Drosophila cell lines. *Genome research* **21**, 301-314, doi:10.1101/gr.112961.110 (2011).
68 Filion, G. J. *et al.* Systematic Protein Location Mapping Reveals Five Principal Chromatin Types in Drosophila Cells. *Cell* **143**, 212-224, doi:10.1016/j.cell.2010.09.009 (2010).
69 Rechtsteiner, A. *et al.* The Histone H3K36 Methyltransferase MES-4 Acts Epigenetically to Transmit the Memory of Germline Gene Expression to Progeny. *PLoS Genet* **6**, doi:10.1371/journal.pgen.1001091 (2010).