Comparative analysis of metazoan chromatin architecture

Joshua W. K. Ho[1,2*], Tao Liu[3,4*], Youngsook L. Jung[1,2*], Burak H. Alver[1^], Soohyun Lee[1^], Kohta Ikegami[5^], Kyung-Ah Sohn[6,7^], Aki Minoda[8,9^], Michael Y. Tolstorukov[1,2,10^], Alex Appert[11^], Stephen C. J. Parker[12,13^], Tingting Gu[14^], Anshul Kundaje[15,16^], Nicole C. Riddle[14^], Eric Bishop[1,17^], Thea A. Egelhofer[18^], Sheng'en Shawn Hu[19^], Artyom A. Alekseyenko[2,20^], Andreas Rechtsteiner[18^], Dalal Asker[21,22^], Jason A. Belsky[23], Sarah K. Bowman[10], Q. Brent Chen[5], Ron A-J Chen[11], Daniel S. Day[1,24], Yan Dong[11], Andréa C. Dosé[9], Xikun Duan[19], Charles B. Epstein[16], Sevinc Ercan[5,25], Elise A. Feingold[13], Jacob M. Garrigues[18], Nils Gehlenborg[1,16], Peter J. Good[13], Psalm Haseley[1,2], Daniel He[9], Moritz Herrmann[11], Michael M. Hoffman[26], Tess E. Jeffers[5], Peter V. Kharchenko[1], Paulina Kolasinska-Zwierz[11], Chitra V. Kotwaliwale[9,27], Nischay Kumar[15,16], Sasha A. Langley[8,9], Erica N. Larschan[28], Isabel Latorre[11], Max W. Libbrecht[26,29], Xueqiu Lin[19], Richard Park[1,17], Michael J. Pazin[13], Hoang N. Pham[8,9,27], Annette Plachetka[2,20], Bo Qin[19], Yuri B. Schwartz[21,30], Noam Shoresh[16], Przemyslaw Stempor[11], Anne Vielle[11], Chengyang Wang[19], Christina M. Whittle[9,27], Huiling Xue[1,2], Robert E. Kingston[10], Ju Han Kim[7,31], Bradley E. Bernstein[16,27], Abby F. Dernburg[8,9,27], Vincenzo Pirrotta[21], Mitzi I. Kuroda[2,20], William S. Noble[26,29], Thomas D. Tullius[17,32], Manolis Kellis[15,16], David M. MacAlpine[23#], Susan Strome[18#], Sarah C. R. Elgin[14#], Julie Ahringer[11#], Xiaole Shirley Liu[3,4,16#], Gary H. Karpen[8,9#], Jason D. Lieb[5#], Peter J. Park[1,2,33#]

1. Center for Biomedical Informatics, Harvard Medical School, Boston, MA, USA
2. Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
3. Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, MA 02215, USA
4. Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, 450 Brookline Ave, Boston, MA 02215, USA
5. Department of Biology and Carolina Center for Genome Sciences, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
6. Division of Information and Computer Engineering, Ajou University, Suwon 443749, Korea
7. Systems Biomedical Informatics Research Center, College of Medicine, Seoul National University, Seoul 110799, Korea
8. Department of Genome Dynamics, Life Sciences Division, Lawrence Berkeley National Lab, Berkeley, California, USA
9. Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, California 94720, USA
10. Department of Molecular Biology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA
11. The Gurdon Institute and Department of Genetics, University of Cambridge, Tennis Court Road, Cambridge CB3 0DH, UK
12. National Institute of General Medical Sciences, National Institutes of Health, Bethesda, MD, USA
13. National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA
14. Department of Biology, Washington University in St. Louis, St. Louis, MO 63130 USA

15. Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA
16. Broad Institute, Cambridge, MA, USA
17. Program in Bioinformatics, Boston University, Boston, MA, USA
18. Department of Molecular, Cell and Developmental Biology, University of California Santa Cruz, Santa Cruz CA 95064, USA
19. Department of Bioinformatics, School of Life Science and Technology, Tongji University, Shanghai, 200092, China
20. Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
21. Department of Molecular Biology and Biochemistry, Rutgers University, Piscataway, NJ 08854
22. Food Science and Technology Department, Faculty of Agriculture, Alexandria University, Alexandria, Egypt.
23. Department of Pharmacology and Cancer Biology, Duke University Medical Center, Durham, NC, USA
24. Harvard/MIT Division of Health Sciences and Technology, Cambridge, MA, USA
25. Department of Biology, Center for Genomics and Systems Biology, New York, NY, USA
26. Department of Genome Sciences, University of Washington, Seattle, WA, USA
27. Howard Hughes Medical Institute, Chevy Chase, MD 20815 USA
28. Department of Molecular Biology, Cellular Biology and Biochemistry, Brown University, Providence, RI
29. Department of Computer Science and Engineering, University of Washington, Seattle, WA, USA
30. Department of Molecular Biology, Umeå University, 901 87 Umeå, Sweden
31. Seoul National University Biomedical Informatics (SNUBI), Div. of Biomedical Informatics, College of Medicine, Seoul National University, Seoul 110799, Korea
32. Department of Chemistry, Boston University, Boston, MA 02215, USA
33. Informatics Program, Children's Hospital, Boston, MA, USA

* These authors contributed equally
^ These authors contributed equally
# Co-corresponding authors

Abstract

Utilization of genetic information in eukaryotic genomes is regulated by the dynamic chromatin environment. To understand which features of chromatin architecture are conserved and which differ across species, we carried out a comparative analysis of chromatin-related features using data from cell lines and developmental stages of *Homo sapien*s, *Drosophila melanogaster*, and *Caenorhabditis elegans* generated by the Encyclopedia of DNA Elements (ENCODE) and the model organism ENCODE (modENCODE) consortia. Our cross-species genome-wide chromatin analysis compared the placement and co-localization of histone modifications, composition of nuclear lamina-associated domains, organization of large-scale topological domains, chromatin environment at promoters and enhancers, and the role of sequence-dependent DNA shape in nucleosome positioning. We find that the overall chromatin organization is similar among the three organisms. However, significant differences also exist, most notably in the composition and configuration of repressive chromatin. These results provide insights into species differences and similarities in the establishment and function of epigenetic states.

Introduction

The *Homo sapiens* (human), *Drosophila melanogaster* (fly), and *Caenorhabditis elegans* (worm) genomes differ in size (~3.4×10^9, ~1.7×10^8, ~1.0×10^8 bp, respectively, per haploid complement, including unassembled regions), chromosome architecture, and gene organization. For instance, fly and human chromosomes contain single centromeres, but worm centromeres are distributed across the length of each chromosome. Many worm genes are co-transcribed in operons, and primary transcripts are often trans-spliced[1], features that are rare in human and fly. Nevertheless, many important developmental, signaling, and human disease–associated protein-coding genes are conserved[2,3]. As a result, comprehensive studies of worm and fly orthologs have contributed significantly to our understanding of genetic and molecular mechanisms of many important human genes and regulatory regions. However, genome sequence alone is insufficient to understand how genetic information is utilized and regulated in cellular and developmental contexts, as eukaryotic genomes are packaged into chromatin through interactions with histone proteins and other molecules. Thus, elucidating and comparing chromatin architecture is critical for a deeper understanding of gene regulation and other genome functions.

Chromatin immunoprecipitation experiments followed by DNA sequencing (ChIP-seq) or genome-wide microarray hybridization (ChIP-chip) have been used to determine the genomic distributions of many post-translational covalent histone modifications and histone variants in human[4–6], fly[7,8], worm[9,10], and other species. This has revealed conserved correlations at the gene level, *e.g.,* active promoters are enriched for H3K4me3 (histone H3 lysine 4 trimethylation), H3K4me1 is enriched at enhancers, and H3K36me3 is often associated with actively transcribed gene bodies. Subsequent analyses introduced the concept of chromatin states[8,9,11] – each state consisting of a combination of histone modifications – which in conjunction with transcriptome and transcription factor binding data have been used to infer putative biochemical functions for much of the large fraction of metazoan genomes that are not protein-coding. To date, these analyses have mostly focused on data from individual or closely related species, so conservation and divergence of chromatin organization among distantly related metazoans remain largely unexplored.

This study presents a systematic comparison of chromatin architecture across three representative and evolutionarily distant genomes – human, fly, and worm. Our results, which are summarized in Table 1, reveal both similarities and differences in the patterns of histone marks and higher order organization for chromatin states, topological domains, and heterochromatin organization, and their relationships with gene expression. Overall, our study provides an important resource for comparative animal genomics, enabling further investigations of chromatin structure as a key regulator in the flow of genetic information in eukaryotes.

Generation and analysis of chromatin data

The ENCODE (http://encodeproject.org) and modENCODE (http://modencode.org) consortia have annotated the genomes of human, fly and worm through the generation and analysis of diverse genomic datasets[12,13]. The chromatin data now consist of over 1,400 genome-wide profiles of core histones, histone variants, histone modifications, and chromatin-associated proteins (Supplementary Fig. 1; see Methods), more than doubling the number of profiles reported in the previous consortium publications[7,9,12]. ChIP targets and chromatin datasets profiled in at least two species are shown in Fig. 1a (see full dataset in Supplementary Fig. 2; Supplementary Tables 1, 2). We have developed a database and web application (http://encode-x.med.harvard.edu/data_sets/chromatin/) with faceted browsing that allows users to efficiently explore the data and choose tracks for visualization or download. The organism-specific data can also be accessed via modMine[14] or the ENCODE Data Coordination Center[15]. Our analysis also includes complimentary data from several published studies.

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 (Fig. 1b; 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.

We profiled two different types of source material for the three species: cultured cell lines for human, whole animals for worms, and both cell lines and whole animals for flies. Here, we largely focus on human cell lines (H1-hESC, GM12878 and K562); *Drosophila* late embryos (LE), third instar larvae (L3) and cell lines derived from embryos or L3 (S2, Kc, BG3); and *C. elegans* early embryos (EE) and stage 3 larvae (L3). Whole animals provide insight into developmental states and processes, but the mixture of cell types makes it difficult to determine whether two overlapping features occur in the same cell type. The fly and worm tissues analyzed here comprise mostly somatic cells, but still include diverse somatic lineages and stages. Thus, every effort has been made to support broad conclusions with data from both cell lines and developmental stages. Below, we discuss caveats raised by the use of mixed cell populations when relevant.

Genome-wide correlations between histone modifications

We investigated whether genome-wide correlations of histone modifications are conserved in distant metazoan genomes, as shown recently within the mammalian lineage[16]. We computed pairwise correlations of enrichment signals for the eight histone marks that had been mapped in multiple samples in each of the three organisms. Modifications enriched within or near actively transcribed genes are consistently correlated with each other in all three organisms (Fig. 2a). In contrast, we found major differences in the relationships between two repressive marks associated with silenced genes: H3K27me3, which is associated with Polycomb (Pc)-mediated silencing, and H3K9me3, associated with heterochromatin. This is indicated by the high variance of correlation between these two marks among the different organisms (black cell in Fig. 2a). In worm, these two marks are strongly correlated at both developmental stages analyzed (r = 0.64 in EE and 0.51 in L3), whereas their correlation is low in human (r = -0.24 ~ -0.06) and fly (r = -0.03 ~ -0.1) (Supplementary Fig. 3). This overlap is not due to antibody cross-reactivity (Supplementary Fig. 4) and is consistent in multiple cell-types (Supplementary Fig. 5). Detailed investigation of this variation is described below.

We also found that the degree of correlation between specific histone marks varied between developmental stages or cell types in the same organism (Fig. 2a). For instance, co-enrichment of H3K27me3 and H3K4me1 is observed at enhancers in embryonic stem cells H1-hESC, but not in differentiated GM12878 cells (Supplementary Fig. 3). A similar co-enrichment is observed in fly LE but not in fly adult head (AH; Supplementary Fig. 3). These results are consistent with a higher prevalence of poised enhancers (defined by co-occurrence of H3K27me3 and H3K4me1[17,18]) in undifferentiated lineages.

Joint chromatin segmentation identifies shared and distinct chromatin states across species

Previous studies identified prevalent combinations of marks, or 'chromatin states' in human[11,19] and fly[7,8], which were found to correlate with functional features such as promoters, coding regions of active genes, enhancers, and heterochromatin. These 'chromatin state maps' provide systematic and automated cell type- or developmental stage-specific annotations of the genome, including protein-coding and non-protein-coding domains. Cross-species chromatin state mapping requires robust identification of both common and species-specific patterns of histone modifications, while adjusting for different genome sizes and variable dynamic ranges in ChIP signals. To address these challenges, we applied three state-of-the-art algorithms—ChromHMM[20], Segway[21], and a novel method called hierarchically-linked infinite hidden Markov model (hiHMM; see Methods)—to jointly generate a chromatin state map across worm (stages EE and L3), fly (stages LE and L3) and human (cell lines H1-hESC and GM12878). The three algorithms gave largely concordant results (Supplementary Figs. 6, 7), but we selected hiHMM segmentation for further analysis, as it was developed specifically for cross-species modeling (see Methods). The hiHMM model allows key features of the state definitions to be shared across multiple species and cell types within a species, while retaining the ability of each species to have its own chromatin state definition. The method can also learn the optimal number of states needed to capture the prevalent combinations directly from the data.

Using the eight histone marks mapped across all species, we generated chromatin state annotations with 16 states that capture the most prevalent combinations and features (Figs. 2b,c). Our chromatin state maps are in agreement with existing species-specific chromatin maps for human[11] and fly[8], even though they were generated using a different number of histone marks and different cell types than previous studies (Supplementary Fig. 8). Based on their associations with known genomic features, we categorized the 16 states into six groups: promoter (state 1), enhancer (states 2–3), gene bodies (states 4–9), Polycomb-repressed (states 10–11), heterochromatin (states 12–13) and weak or low signal (states 14–16). In general, similar combinations of histone marks are enriched in each state across the three species, and each state is enriched for similar genomic features (Supplementary Fig. 9), chromosomal proteins (Supplementary Fig. 10) and transcription factors (Supplementary Fig. 11).

The worm-specific high correlation between H3K9me3 and H3K27me27 mentioned above is captured in part by states 12 and 13, in which these two marks are both enriched in worm, but only H3K9me3 is present in human and fly. The chromatin state map also shows two distinct types of repressed regions: bivalent domains[22] (state 10) in which strong H3K27me3 is accompanied by marks for active genes or enhancers, and Polycomb-repressed domains with only H3K27me3 (state 11). An example of a developmental stage-specific state is seen in worm, where the H4K20me1-enriched state (state 6) has higher enrichment in chromosome X in L3 compared to EE (Supplementary Fig. 12). This is consistent with the chromosome-wide enrichment of this mark associated with dosage compensation[23].

Chromatin state similarity within genome-wide topological domains

Three-dimensional genome-wide chromatin conformation capture (Hi-C) assays have revealed a prominent topological domain structure in embryonic stem (ES) and differentiated cells in human and mouse[24] and in fly late embryos[25]. The physical domains defined by Hi-C overlap extensively with either active or repressive chromatin marks[25] and are generally bounded by insulator elements and active genes[24,25]. Joint chromatin segmentation in matching cell types allows a direct comparison of the role of topological domains in genome regulation in human and fly. Our analysis reveals similar features in the enrichment of chromatin states both inside and at the boundaries of the topological domains for the two species. In particular, the promoter state and active transcription states are enriched at domain boundaries defined by Hi-C (Fig. 2d; Supplementary Fig. 13). The interiors of individual domains are relatively uniform in both human and fly, with each domain consisting of chromatin states that belong to one of four common classes: active, Polycomb-repressed, heterochromatin and low signal (Supplementary Fig. 14). One notable exception is that the H4K20me1-enriched gene body state (state 6) is found in Polycomb-repressed domains in human, but it marks the introns of long active genes in fly

(Supplementary Fig. 14). In both species, roughly half of the active genes are found in small active physical domains, which cover about 15% of the mappable genome.

We next sought to define genome-wide topological domains based on chromatin state similarity between neighboring regions that may be predictive of three-dimensional chromatin interactions. For each pair of genomic locations, we defined a Euclidean distance metric that combines genomic proximity and similarity in chromatin state (see Methods), resulting in a genome-wide similarity map in each species (fly LE is shown in Fig. 2e as an example). Comparison of this map to Hi-C based topological domains in fly revealed a striking similarity between the two (Supplementary Fig. 15). It has previously been shown that genome-wide chromatin conformation capture assays reveal interactions between proximal genomic regions with coordinated epigenetic marks and gene expression properties[24,25]; our analysis suggests that topological domains can indeed be largely captured based on chromatin marks alone.

Organization and composition of transcriptionally 'silent' domains

Previous studies have revealed the existence of two distinct types of transcriptionally-repressed chromatin that are conserved across metazoans. Classical 'heterochromatin' is generally concentrated in pericentric and telomeric chromosomal regions, and enriched for H3K9me2/me3 and associated binding proteins and histone methyltransferases (HMTases); such packaging silences normally euchromatic genes (juxtaposed by rearrangement or transposition) by a stochastic process that gives a 'variegated' pattern of expression. In contrast, 'Polycomb-associated silenced domains' are found in many different chromosome regions, are enriched for H3K27me3 and a different set of chromatin binding proteins and HMTases, and have been implicated in cell-type-specific silencing of developmentally regulated genes[8,26].

Heterochromatin constitutes a distinct chromosomal and nuclear element that plays important roles in genome organization, genome stability, chromosome inheritance and gene regulation. We used genome-wide segmentation of H3K9me3 enrichment as a proxy for heterochromatin (Fig. 3a; see Methods) and identified heterochromatic domains in human, fly, and worm (Supplementary Fig. 16). The boundaries between pericentric heterochromatin and euchromatin on each fly chromosome are consistent with those from lower resolution studies using H3K9me2[26] (Supplementary Fig. 17). As expected, the majority of the H3K9me3-enriched domains in fly and human are concentrated in the pericentric regions (as well as other specific domains, such as the Y chromosome), whereas in worm they are distributed in subdomains throughout the chromosome arms[10] (Fig. 3a). H3K9me2, which is also associated with silent chromatin[26], shows a stronger correlation with H3K9me3 in fly than in worm (r = 0.89 vs. r = 0.40, respectively), whereas H3K9me2 is well correlated with H3K9me1 in worm but not in fly (r = 0.44 vs. r = -0.32, respectively; Fig. 3b). In all three organisms, H3K9me1 shows low correlation with H3K9me3. These findings suggest differences in heterochromatin states and highlight the diversity of H3K9 methylation patterns in human, fly, and worm.

To explore the relationship between heterochromatic domains and differentiation, we determined the proportion of the genome in heterochromatin in different cell types and developmental stages in all three species. More of the genome is covered by H3K9me3 in differentiated cells/tissues than in embryonic cells/tissues in all three species, sometimes by two-fold or greater[27] (Fig. 3c). Blocks of H3K9me3-associated chromatin were previously seen in the euchromatic arms of fly chromosomes in a cell type-specific pattern, presumably reflecting the silencing of genes during differentiation[8,26]. In human differentiated cell types, the euchromatic regions enriched with H3K9me3 often form large domains (up to ~11 Mb in size), far away from the centromere, and are cell-type-specific. Such domains, which are euchromatic in some cell types but enriched for H3K9me3 in other cell types, fit the definition of 'facultative' heterochromatin. This distinction cannot be made in worm, as all data sets are from samples with mixed cell types (embryos and larvae).

We next compared the patterns of histone modifications along expressed and silent genes in euchromatin and heterochromatin (Fig. 3d). We previously reported depletion of H3K9me3 at the transcription start site (TSS) of expressed genes located in fly heterochromatin[28], and now find a similar pattern in human (Fig. 3d; Supplementary Fig. 18). A different pattern is observed in worm heterochromatin, in which expressed genes have a lower enrichment of H3K9me3 across the gene body than silent genes do, with no systematic difference in promoter depletion (Fig. 3d; Supplementary Fig. 18). A conspicuous difference is in the pattern of H3K27me3: in euchromatic regions, silent genes in human and fly have a much higher level of H3K27me3 enrichment than in heterochromatic regions. The pattern is nearly opposite in worm, with silent worm genes having much stronger H3K27me3 enrichment in heterochromatic regions than in euchromatic regions (Fig. 3d; Supplementary Fig. 18). We also observe that expressed genes in fly heterochromatin have much higher levels of active marks than those in euchromatin, whereas the opposite is seen in worm (Supplementary Fig. 18).

Consistent with the above findings, we observed a low correlation between H3K9me3 and H3K27me3 in human and fly, and a high correlation between H3K9me3 and H3K27me3 in worm (Fig. 3b; Supplementary Fig. 19). The high correlation in worm is driven by the overlapping distributions of H3K27me3 and H3K9me3 on the arms of worm chromosomes. The vast majority of H3K9me3 resides in the arms (Fig. 3a), and H3K27me3 is also arm-enriched. In the arms, virtually all H3K9me3 domains reside within H3K27me3 domains (Supplementary Fig. 20). H3K9me3 and H3K27me3 could reside on the same or adjacent nucleosomes in individual cells, as observed in plants[29]. Alternatively, the two marks may occur in different cell types in worm embryos and larvae. Additional experiments, such as sequential ChIP, will be needed to resolve this. H3K27me3 shows a high correlation with CENP-A (also known as CenH3 and HCP-3, an H3 variant associated with kinetochore function[30]) across the entire length of the chromosomes (Supplementary Figs. 20, 21). Thus in worm, H3K27me3 may be involved in several functions: Polycomb-type silencing as observed in fly and human, organization of heterochromatic domains on the chromosome arms, and holocentromere function. In addition, these results suggest that the characteristics of transcriptionally-repressed chromatin observed in human and fly are similar, but are distinct from worm (see Discussion; Supplementary Fig. 20).

Chromatin context of lamina-associated domains (LADs)

Lamina-associated domains (LADs) are regions associated with nuclear lamina proteins, including the B-type lamins in human and fly[31,32] and the integral nuclear membrane protein LEM-2 in worm[33]. They have been observed to correlate with transcriptionally silent domains and are altered during differentiation[34]. LADs are known to be enriched for H3K27me3 at their boundaries in human[31] and worm[33]; they also have been associated with weak enrichments of H3K9me2 in human[31] and H3K9me3 in worm[33]. To elucidate the chromatin environment in LADs across these three species, we investigated the association between the repressive histone modifications, H3K27me3 and H3K9me3, and LADs, using data from worm mixed-stage embryos[33], fly Kc cells[35] and human fibroblast Tig3 cells[31]. We find that LADs are enriched for

H3K27me3 and are often flanked by E(Z) in fly or its human ortholog EZH2, both H3K27 methyltransferases and members of Polycomb Repressive Complex 2 (Supplementary Fig. 22). In human, we also find that the LADs identified in fibroblasts (Tig3) coincide with the regions enriched for H3K9me3 in other fibroblast or fibroblast-like cell types (Fig. 3e).

Our examination also revealed 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 (Fig. 3f; Supplementary Fig. 23). Although LADs are generally smaller in worm, we observed a similar though weaker trend, with longer LADs more frequently enriched for H3K9me3 (Fig. 3f; Supplementary Fig. 23). No long LADs in the H3K9me3 heterochromatic regions were reported in fly data generated from Kc167 cells using DamID[36]; however, this may reflect the specific cellular origin (plasmatocyte) of Kc167 cells[37] (Supplementary Fig. 22), as well as the fact that these analyses do not include the simple tandem repeats that constitute the majority of fly heterochromatin. One consistent feature between fly and human, however, is the association of LADs with late replication, which suggests that they generally reside in (and may promote) a repressive chromatin environment that impacts both transcription and DNA replication (Supplementary Fig. 24).

Chromatin profiles at promoters and gene bodies

We investigated the chromatin environment at promoters and bodies of protein-coding genes, and found that histone modification patterns are similar in human, fly, and worm (Fig. 4a). As expected from previous studies, the 5’ ends of expressed genes show enrichment for H3K4me3 and other active histone marks, expressed gene bodies are enriched for H3K36me3 (peaking at the 3' end, except for worm EE[38]), and many repressed genes show H3K27me3 enrichment in all three species. However, we also found notable inter-species differences. For example, H3K23ac is enriched mostly at 5' ends of expressed genes in worm, but is enriched in both expressed and silent genes in fly. H4K20me1 is enriched in both expressed and silent genes in human but only in expressed genes in fly and worm.

Human promoters exhibit a bimodal enrichment for H3K4me3 and other active marks, immediately upstream and immediately downstream of the TSSs (Fig. 4b). In contrast, fly and worm promoters exhibit a unimodal distribution of active marks, downstream of the TSSs. Since genes that have a neighboring gene within 1 kb of a transcription start or end site were removed from this analysis, this bimodal histone modification pattern in human cannot be attributed to nearby genes. This difference is also not explained by chromatin accessibility determined by DNase I hypersensitivity (DHS), or by fluctuations in GC content around the TSSs (Fig. 4b), although the GC profiles are highly variable across species. Using Global Run On (GRO)-seq data, we found that bidirectional transcription is frequently observed at human promoters, while it is much less common at fly promoters, consistent with recent findings in fly[39]. In human, the bimodal enrichment of active marks, most notably H3K4me3, is present in TSSs regardless of whether antisense transcription is observed at those promoters (Supplementary Fig. 25).

Since nucleosome occupancy underlies chromatin structure, we also compared the nucleosomal profiles at TSSs in the three organisms. Such profiles, 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 (Supplementary Fig. 26)[40,41]. However, the main features of the 'classic' nucleosome occupancy profile[42], 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 (Fig. 4c). 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.

Previous studies have identified differences in chromatin structure of genes expressed in most stages and tissues ('broadly expressed genes') and genes expressed in only certain stages, tissues, or cell types ('specifically expressed genes'). In particular, in fly Kc cells a subset of highly expressed genes were found to lack H3K36me3[36], which is generally thought to be generated co-transcriptionally. We observe that specifically expressed genes indeed have lower average H3K36me3 enrichment relative to broadly expressed genes, after controlling for gene expression levels (Supplementary Figs. 27-29; see Methods). 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. 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 (Supplementary Fig. 27). It is possible that other modes of transcriptional regulation exist, *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[38]. Profiling of chromatin patterns and gene expression in individual cell types is needed to test whether cellular heterogeneity fully accounts for our observations.

Chromatin feature at enhancers

Enhancers are *cis*-acting elements that play a critical role in the regulation of gene expression. They usually fall within DNase I hypersensitive sites (DHSs), are bound by the transcriptional co-factor p300/CBP when active, and are associated with specific histone modifications, such as high enrichment of H3K4me1 and low enrichment of H3K4me3[43]. 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[44]. 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 of Fig. 5a), while those distal to TSSs constitute a cluster showing stronger H3K4me1 signals (right column of Fig. 5a). 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 (Supplementary Fig. 30), these two marks clearly distinguish TSS-distal sites (enhancers) from TSS-proximal sites (promoters). Here, we define putative enhancer sites (hereafter referred to as “enhancers”) 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).

In all three species, enhancers exhibit a wide range of enrichment for H3K27ac, reported to be a marker for enhancer activity[17,18] (Supplementary Fig. 31). We found that the proximity of genes to enhancers with higher H3K27ac levels is positively correlated with expression, in a distance-dependent manner (Fig. 5b). This observation is consistent across multiple cell-types and tissues in all three species (Supplementary Fig. 32). We note that H3K27ac and other H3 acetylation marks show a moderate but significant positive correlation with potential enhancer strength as determined by Self-Transcribing Active Regulatory Region Sequencing (STARR-seq)[45] in fly S2 cells (Supplementary Fig. 33).

We further investigated nucleosome occupancy and turnover around enhancers with respect to H3K27ac levels. In general, nucleosome occupancy is lower in the broad region around enhancers (roughly ±2 kb; Supplementary Fig. 34) 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[46]. In human, this increase is characterized by two well-positioned nucleosomes flanking the nucleosome-depleted region at the enhancer center (this feature may be occluded by lower resolution in fly and worm). Given the increased DNA accessibility at these sites, the local nucleosome occupancy peak (±400 bp) may represent relatively unstable nucleosomes, even at well-positioned sites (Supplementary Fig. 35). Using available data in human and fly, we next examined the enrichment levels of histone variant H3.3, which is known to be present in regions with higher nucleosome turnover[47]. We found that the local increase in nucleosome occupancy indeed overlaps with the peak of H3.3 enrichment, and that the levels of H3.3 and H3K27ac enrichment are correlated (Fig. 5c). These findings, together with the specific patterns of nucleosome occupancy[48], indicate that increased nucleosome turnover is one of the major characteristics of chromatin at active enhancers.

Both categories of enhancers (with high or low H3K27ac) have elevated DNA sequence conservation compared to surrounding regions (as measured by Phastcon score; see Methods), supporting their putative role as regulatory elements (Fig. 5d; Supplementary Fig. 36). When we examine the chromatin environment, 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 (Fig. 5d), 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 (Fig. 5d; some enrichment in worm L3, see Supplementary Fig. 36). These configurations are likely to be correlated to the generation of short transcripts from these sites, as reported recently[39]. The observed patterns at human enhancers hold even if the enhancers were centered at p300 sites instead of DHSs (Supplementary Fig. 37).

Nucleosome positioning across species

Sequence-dependent variation in DNA shape and deformability can influence nucleosome positioning, but the impact of this phenomenon in different genomes is currently under debate[49–54]. Here we tested the idea that structural properties of DNA, such as minor groove width, influence the phasing of nucleosomes in human, fly and worm genomes. We used the ORChID2 algorithm[55] to predict shape profiles of nucleosomal DNA fragments identified by paired-end MNase-seq experiments[56–58]. 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. We find that consensus shape profiles, obtained by averaging individual nucleosome-bound sequences aligned by the inferred dyad position, are highly similar across species (Fig. 6a). The shape profiles feature a 10-bp periodic signal, reminiscent of the periodic occurrence of short sequence motifs, *e.g.,* AA/TT or GG/CC, previously found in nucleosome-bound fragments[57,59–62]. To account for the different nucleotide compositions of the three genomes, we stratified nucleosomal fragments by GC content, and found that in lower GC content regions the shape profile is more pronounced within a species, and more similar among species (Supplementary Fig. 38). The observation of a common consensus shape profile suggests that the influence of DNA shape on nucleosome positioning is similar for the three species, and is reflected in the rotational setting of nucleosomes in their genomes.

Next, we asked whether the periodic signal we found in the consensus profile is pronounced enough in individual nucleosome-bound sequences to influence translational positioning of nucleosomes (Fig. 6b; Supplementary Fig. 38). For fly and human, this analysis revealed only modest enrichment in similarity to the consensus profile in individual nucleosomal sequences compared to randomly shuffled sequences (~1%). For worm, the enrichment is more pronounced (~3%). This result is consistent with the larger number of distinct dinucleotides that are periodically distributed in worm compared to the fly and human genomes (*e.g.*, the number of different dinucleotides displaying the 10.4-bp periodicity is 13, 4, and 1 for the worm, fly, and human genomes, respectively[63]). Our results indicate that subtle, periodic variation in DNA shape influences the rotational positioning of the histone octamer core, while other factors may have a greater effect on translational positioning.

Discussion

In metazoans, a single genome must generate numerous cell types to create the extraordinary diversity of body patterns, tissues and behaviors that characterize this kingdom. The regulation of chromatin at local and global scales is central to gene and genome functions, playing a key role in the determination and differentiation of distinct cell types during development. We have analyzed the largest collection of chromatin datasets heretofore considered across three representative metazoan species to determine if the patterns of chromatin organization and composition associated with functional genomic elements are conserved. Our comparative analysis reveals both shared and distinct principles of chromatin architecture among these organisms, which are summarized in Table 1.

We observe many commonalities among the three species. For example, the existence of similar chromatin states in human, fly, and worm suggests functional conservation of histone marking (Fig. 2b). Patterns of histone modifications and nucleosome occupancy around protein-coding genes and enhancers are largely similar across species (Figs. 4, 5). Similarities in the configuration and composition of topological domains, lamina-associated domains, and the borders and flanking regions of these domains also demonstrate common organizational principles (Figs. 2d, 2e, 3f). Furthermore, DNA structural features associated with nucleosome positioning are also strongly conserved across species, with evidence for a greater role in worm (Fig. 6).

Strikingly, however, our results suggest the existence of three distinct types of repressed chromatin in the three species (Table 1). The first type contains H3K27me3 but little or no H3K9me3 (states 10 and 11). This type defines developmentally regulated Polycomb-silenced domains in human and fly, and likely in worm as well. The second type is enriched for H3K9me3 but lacks H3K27me3 (represented by human and fly states 12 and 13). This type defines constitutive, predominantly pericentric heterochromatin in human and fly, and is essentially absent from the worm genome. We note that H3K9me3-only domains are also found in fly and human chromosome arms in cell-type-specific patterns, which may represent 'facultative heterochromatin', an alternative mechanism for cell type-specific silencing[26,64]. The third type contains both H3K9me3 and H3K27me3 and occurs predominantly in worms (represented by worm states 12 and 13). Although our experiments preclude us from determining conclusively whether these two marks co-exist on the same nucleosomes (see below), co-occurrence is supported by the observation that H3K9me3 and H3K27me3 are both required for silencing of heterochromatic transgenes in worms[65]. In addition, mass spectrometry analyses in human and sequential ChIP experiments in plants indicate that these two marks can be present on the same nucleosomes[29,66].

When evaluating inter-species distinctions in the distributions of chromatin marks and states, it is important to consider the impact of differences in global genome and gene organization. This includes inter-species differences in average gene size and density, the relative proportions of divergent and tandem genes, repeated DNA content and distributions, distance between promoters and enhancers, centromere function, and global domain organization. In particular, human and fly chromosomes have single centromeres and contiguous and large pericentric heterochromatin domains. In worm, the overlapping distributions of H3K9me3 and H3K27me3 and the interspersion and discontinuous nature of repressive domains could be a consequence of the holocentric organization of worm chromosomes. This in turn could reflect the distinct distributions of DNA repeats in these species, although cause and effect cannot yet be disentangled.

In addition, chromatin structure and histone modifications change dynamically during development and can vary substantially between different cell types. For fly, we used homogenous cell lines to validate observations made in whole organisms, which contain mixed populations of cell types. However, cell lines are not available for worm. Conversely, all human analysis was done exclusively on homogeneous cell lines and not on tissues or developmental stages. Future studies should include a broader range of specific cell types and developmental stages to understand the diversity of chromatin states across different conditions and the changes critical for cell type-specific gene expression and differentiation.

It is important to note that there can be a spectrum of chromatin features associated with a specific functional element within each organism, and the average profile is affected by the relative proportions of particular types of functional elements rather than absolute differences in regulation. For example, although the chromatin pattern shown in Fig. 4a is typical of a protein coding gene, the pattern can be variable among individual genes depending on gene structure[67], tissue-specificity (Supplementary Figs. 27-29), and whether they are located in heterochromatin (Fig. 3d). Furthermore, the bidirectional transcription observed at human promoters appears to be absent in fly when analyzing average patterns (Fig. 4b), even when there are clear examples of individual fly promoters that display these properties[68].

Both *C. elegans* and *Drosophila* have been used extensively in modern biological research for understanding human gene function, development, and disease. The analyses of chromatin architecture presented here provide a blueprint for interpreting experimental results in model systems and their relevance to human biology. More generally, the insights and the public resources generated by this project provide a deeper appreciation of the commonalities and differences in the chromatin architecture of diverse metazoan genomes, and form a foundation for understanding how genome functions are regulated in the context of development and disease**.**

Methods

For full details of Methods, see Supplementary Information.

References

1. Blumenthal, T. *et al.* A global analysis of Caenorhabditis elegans operons. *Nature* **417,** 851–854 (2002).
2. Rubin, G. M. *et al.* Comparative genomics of the eukaryotes. *Science* **287,** 2204–2215 (2000).
3. Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).
4. Barski, A. *et al.* High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* **129,** 823–837 (2007).
5. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448,** 553–560 (2007).
6. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459,** 108–112 (2009).
7. The modENCODE Consortium *et al.* Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE. *Science* **330,** 1787 –1797 (2010).
8. Kharchenko, P. V. *et al.* Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature* **471,** 480–485 (2011).
9. Gerstein, M. B. *et al.* Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* **330,** 1775–1787 (2010).
10. Liu, T. *et al.* Broad chromosomal domains of histone modification patterns in C. elegans. *Genome Res.* **21,** 227–236 (2011).
11. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473,** 43–49 (2011).
12. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
13. Celniker, S. E. *et al.* Unlocking the secrets of the genome. *Nature* **459,** 927–930 (2009).
14. Contrino, S. *et al.* modMine: flexible access to modENCODE data. *Nucleic Acids Res.* **40,** D1082–1088 (2012).
15. Rosenbloom, K. R. *et al.* ENCODE whole-genome data in the UCSC Genome Browser: update 2012. *Nucleic Acids Res.* **40,** D912–917 (2012).
16. Xiao, S. *et al.* Comparative epigenomic annotation of regulatory DNA. *Cell* **149,** 1381–1392 (2012).
17. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences* **107,** 21931 – 21936 (2010).
18. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470,** 279–283 (2011).

19. Hoffman, M. M. *et al.* Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.* **41,** 827–841 (2013).

20. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods* **9,** 215–216 (2012).

21. Hoffman, M. M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods* **9,** 473–476 (2012).

22. Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125,** 315–326 (2006).

23. Vielle, A. *et al.* H4K20me1 Contributes to Downregulation of X-Linked Genes for C. elegans Dosage Compensation. *PLoS Genet.* **8,** e1002933 (2012).

24. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485,** 376–380 (2012).

25. Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148,** 458–472 (2012).

26. Riddle, N. C. *et al.* Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin. *Genome Res.* **21,** 147–163 (2011).

27. Hawkins, R. D. *et al.* Distinct Epigenomic Landscapes of Pluripotent and Lineage-Committed Human Cells. *Cell Stem Cell* **6,** 479–491 (2010).

28. Riddle, N. C. *et al.* Enrichment of HP1a on Drosophila Chromosome 4 Genes Creates an Alternate Chromatin Structure Critical for Regulation in this Heterochromatic Domain. *PLoS Genet.* **8,** e1002954 (2012).

29. Lindroth, A. M. *et al.* Dual histone H3 methylation marks at lysines 9 and 27 required for interaction with CHROMOMETHYLASE3. *EMBO J* **23,** 4146–4155 (2004).

30. Monen, J., Maddox, P. S., Hyndman, F., Oegema, K. & Desai, A. Differential role of CENP-A in the segregation of holocentric C. elegans chromosomes during meiosis and mitosis. *Nat. Cell Biol.* **7,** 1248–1255 (2005).

31. Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453,** 948–951 (2008).

32. Pickersgill, H. *et al.* Characterization of the Drosophila melanogaster genome at the nuclear lamina. *Nat. Genet.* **38,** 1005–1014 (2006).

33. 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,** R120 (2010).

34. Peric-Hupkes, D. *et al.* Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. *Mol. Cell* **38,** 603–613 (2010).

35. Van Bemmel, J. G. *et al.* The insulator protein SU(HW) fine-tunes nuclear lamina interactions of the Drosophila genome. *PLoS ONE* **5,** e15013 (2010).

36. Filion, G. J. *et al.* Systematic Protein Location Mapping Reveals Five Principal Chromatin Types in Drosophila Cells. *Cell* **143,** 212–224 (2010).

37. Cherbas, L. *et al.* The transcriptional diversity of 25 Drosophila cell lines. *Genome Res.* **21,** 301–314 (2011).

38. 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,** e1001091 (2010).

39. Core, L. J. *et al.* Defining the Status of RNA Polymerase at Promoters. *Cell Reports* **2,** 1025–1035 (2012).

40. Jin, C. *et al.* H3.3/H2A.Z double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions. *Nat. Genet.* **41,** 941–945 (2009).

41. Henikoff, J. G., Belsky, J. A., Krassovsky, K., MacAlpine, D. M. & Henikoff, S. Epigenome characterization at single base-pair resolution. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 18318–18323 (2011).

42. Yuan, G.-C. *et al.* Genome-scale identification of nucleosome positions in S. cerevisiae. *Science* **309,** 626–630 (2005).

43. Bulger, M. & Groudine, M. Functional and mechanistic diversity of distal transcription enhancers. *Cell* **144,** 327–339 (2011).

44. Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457,** 854–858 (2009).

45. Arnold, C. D. *et al.* Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science* (2013). doi:10.1126/science.1232542

46. Tillo, D. *et al.* High nucleosome occupancy is encoded at human regulatory sequences. *PLoS ONE* **5,** e9129 (2010).

47. Deal, R. B., Henikoff, J. G. & Henikoff, S. Genome-Wide Kinetics of Nucleosome Turnover Determined by Metabolic Labeling of Histones. *Science* **328,** 1161–1164 (2010).

48. He, H. H. *et al.* Nucleosome dynamics define transcriptional enhancers. *Nat Genet* **42,** 343–347 (2010).

49. Peckham, H. E. *et al.* Nucleosome positioning signals in genomic DNA. *Genome Res.* **17,** 1170–1177 (2007).

50. Gupta, S. *et al.* Predicting human nucleosome occupancy from primary sequence. *PLoS Comput. Biol.* **4,** e1000134 (2008).

51. Kaplan, N. *et al.* The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458,** 362–366 (2009).

52. Zhang, Y. *et al.* Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. *Nat. Struct. Mol. Biol.* **16,** 847–852 (2009).

53. Tolstorukov, M. Y., Volfovsky, N., Stephens, R. M. & Park, P. J. Impact of chromatin structure on sequence variability in the human genome. *Nat. Struct. Mol. Biol.* **18,** 510–515 (2011).

54. Chang, G. S. *et al.* Unusual combinatorial involvement of poly-A/T tracts in organizing genes and chromatin in Dictyostelium. *Genome Res.* **22,** 1098–1106 (2012).

55. Bishop, E. P. *et al.* A map of minor groove shape and electrostatic potential from hydroxyl radical cleavage patterns of DNA. *ACS Chem. Biol.* **6,** 1314–1320 (2011).

56. Gilchrist, D. A. *et al.* Pausing of RNA Polymerase II Disrupts DNA-Specified Nucleosome Organization to Enable Precise Gene Regulation. *Cell* **143,** 540–551 (2010).

57. Gaffney, D. J. *et al.* Controls of nucleosome positioning in the human genome. *PLoS Genet.* **8,** e1003036 (2012).

58. 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 Res.* **22,** 766–777 (2012).

59. Johnson, S. M., Tan, F. J., McCullough, H. L., Riordan, D. P. & Fire, A. Z. Flexibility and constraint in the nucleosome core landscape of Caenorhabditis elegans chromatin. *Genome Res.* **16,** 1505–1516 (2006).

60. Mavrich, T. N. *et al.* Nucleosome organization in the Drosophila genome. *Nature* **453,** 358–362 (2008).

61. Trifonov, E. N. & Sussman, J. L. The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proc. Natl. Acad. Sci. U.S.A.* **77,** 3816–3820 (1980).

62. Satchwell, S. C., Drew, H. R. & Travers, A. A. Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* **191,** 659–675 (1986).

63. Bettecken, T. & Trifonov, E. N. Repertoires of the nucleosome-positioning dinucleotides. *PLoS ONE* **4,** e7654 (2009).

64. Zhu, J. *et al.* Genome-wide Chromatin State Transitions Associated with Developmental and Environmental Cues. *Cell* **152,** 642–654 (2013).

65. Towbin, B. D. *et al.* Step-wise methylation of histone H3K9 positions heterochromatin at the nuclear periphery. *Cell* **150,** 934–947 (2012).

66. Voigt, P. *et al.* Asymmetrically modified nucleosomes. *Cell* **151,** 181–193 (2012).

67. Huff, J. T., Plocik, A. M., Guthrie, C. & Yamamoto, K. R. Reciprocal intronic and exonic histone modification regions in humans. *Nat Struct Mol Biol* **17,** 1495–1499 (2010).

68. Gerstein, M. B. *et al.* An integrative comparison of metazoan transcriptome. *Submitted*

Acknowledgements

This project is mainly funded by NHGRI U01HG004258 (GHK, SCRE, MIK, PJP, VP), U01HG004270 (JDL, JA, AFD, XSL, SS), U01HG004279 (DMM), U54HG004570 (BEB) and U01HG004695 (WSN). It is also supported by NHBIB 5RL9EB008539 (JWKH), NHGRI K99HG006259 (MMH), NIGMS fellowships (SCJP, ENL), NIH U54CA121852 (TDT), NSF 1122374 (DSD), National Natural Science Foundation of China 31028011 (XSL), MEST Korea NRF-2010-0028631 (JHK), NRF-2012-0000994 (K-AS), and Wellcome Trust 54523 (JA). We thank David Acevedo and Cameron Kennedy for technical assistance.

Author Contributions

Lead data analysis team: JWKH, TL, YLJ, BHA, SL, K-AS, MYT, SCJP, AK, EB, SSH, AR. **Lead data production team:** KI, AM, AA, TG, NCR, TAE, AAA, DA. **(Ordered alphabetically) Data analysis team:** DSD, XD, NG, PH, MMH, PVK, NK, ENL, MWL, RP, NS, CW, HX; **Data production team:** JAB, SKB, QBC, RA-JC, YD, ACD, CBE, SE, JMG, DH, MH, TEJ, PK-Z, CVK, SAL, IL, XL, HNP, AP, BQ, PS, YBS, AV, CMW. **NIH scientific project management:** EAF, PJG, MJP. The role of the NIH Project Management Group in the preparation of this paper was limited to coordination and scientific management of the modENCODE and ENCODE consortia. **Paper writing:** JWKH, TL, YLJ, BHA, SL, K-AS, MYT, SCJP, SSH, AR, KI, TDT, MK, DMD, SS, SCRE, JA, XSL, GHK, JDL, and PJP. **Group leaders for data analysis or production:** REK, JHK, BEB, AFD, VP, MIK, WSN, TDT, MK, DMD, SS, SCRE, JA, XSL, GHK, JDL, and PJP. **Overall project management, and corresponding authors:** DMD, SS, SCRE, JA, XSL, GHK, JDL, and PJP. Larger efforts in analysis and data production are ascribed to the joint first authors.

Completing Financial Interests

The authors declare no competing financial interests.

Supplementary Information (see attached)

Table

Table 1. Summary of key findings in this study.

Topic	Findings	Human	Fly	Worm	Fig.
Genome-wide correlation	Correlation between H3K27me3 and H3K9me3	Low	Low	High	2a,b
Chromatin state maps	Similar histone marks and genomic features at each state	Yes	Yes	Yes	2b
Topological domains	Active promoters enriched at boundaries	Yes	Yes	ND	2d
	Similar chromatin states are enriched in each domain	Yes	Yes	ND	2e, S14
Silent domains: constitutive heterochromatin	Composition	H3K9me3	H3K9me3	H3K9me3+H3K27me3	2b, 3d
	Predominant location	Pericentric+Y	Pericentric+chr4+Y	Arms	S20
	Depletion of H3K9me3 at TSS of expressed genes	Yes	Yes	Weak	3d
Silent domains: Polycomb-associated	Composition	H3K27me3	H3K27me3	H3K27me3	2b, 3d
	Predominant location	Arms	Arms+Chr4	Arms+Centers	S20
LADs	Short LADs	H3K27me3	H3K27me3	H3K27me3	3f, S23
	Long LADs	H3K9me3 internal, H3K27me3 borders	ND	H3K9me3+H3K27me3	3f
Promoters	5' H3K4me3 enrichment	Bimodal peak around TSS	Single peak downstream of TSS	Single peak downstream of TSS	4a,b
	Well positioned +1 nucleosome at expressed genes	Yes	Yes	Yes	4c
Gene bodies	Lower H3K36me3 in specifically expressed genes	Yes	Yes	Yes	S27
Enhancers	High H3K27ac sites are more active	Yes	Yes	Yes	5b
	High H3K27ac sites have higher nucleosome turnover	Yes	Yes	ND	5c
Nucleosome positioning	10-bp periodicity profile	Yes	Yes	Yes	6a
	Positioning signal in genome	Weak	Weak	Less weak	6b

ND: No Data

Figure legends

Fig. 1. Dataset overview. a, All ChIP-based histone, non-histone chromosomal proteins, and other non-ChIP-based genome-wide profiles that were mapped in at least two species. Cell types or developmental stages are shown on the left (see Supplementary Table 1 for detailed description); those that share the same profiles are merged and separated by a comma. Orthologs with different protein names in the three species are represented with all of the names separated by slash (/) (see Supplementary Table 2 for detailed description). Data generated outside the consortium are marked by asterisks (*). A full dataset listing is in Supplementary Fig. 2. **b,** Genomic coverage of various histone modifications in the three species. Red lines indicate the ten marks common to the three samples 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 worm).

Fig. 2. Shared and organism-specific chromatin states. a, Genome-wide correlations between histone modifications show intra- and inter- species similarities and differences. Upper left half: pairwise correlations between marks in each genome, averaged across all three species. Lower right half: pairwise correlation, averaged over cell types and developmental stages, within each species (pie chart), inter-species variance (grey-scale background) and intra-species variance (grey-scaled small rectangles) of correlation coefficients. h,f,w indicate human, fly and worm, respectively. **b,** 16 chromatin states derived by joint segmentation using hiHMM based on genome-wide enrichment patterns of the 8 histone marks in each state. The genomic coverage of each state in each cell-type or developmental stage is also shown (see Supplementary Figs. 8-12 for detailed analysis of the states). States are named by putative functional characteristics. **c,** 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 have enrichment of H3K4me3 at their promoters and a transcription state in their gene body. While H3K9me3 is a hallmark of heterochromatin in all three species, H3K27me3 is also enriched in worm heterochromatin. **d,** Occurrences of three active chromatin states near Hi-C-defined topological domain boundaries, normalized to random expectation (see also Supplementary Fig. 13). **e,** Comparison of Hi-C-based and chromatin-based topological domains in fly LE. Local histone modification similarity (Euclidian distance) and Hi-C interaction frequencies are presented as a

juxtaposed heatmap of correlation matrices. Red indicates higher similarity and more interactions. Chromatin-defined boundary scores and domains are compared to several insulator proteins and histone marks in the same chromosomal regions (see also Supplementary Fig. 15).

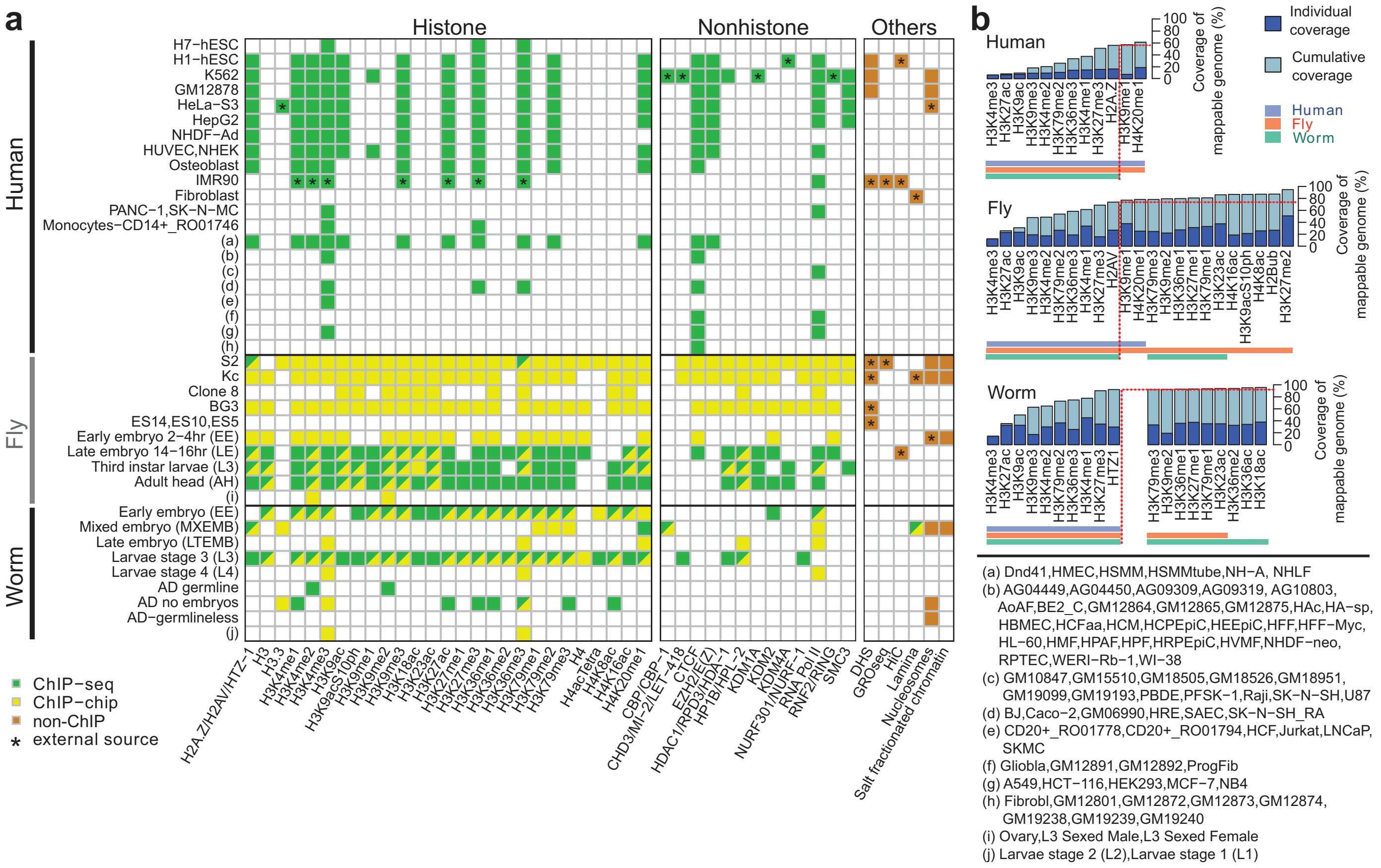
Fig. 3. Genome-wide organization of heterochromatin and lamina-associated domains. a, Enrichment profile of H3K9me1/me2/me3 and identification of heterochromatin domains in all three species based on H3K9me3 enrichment (illustrated for human H1-hESC, fly L3, and worm L3). In fly chromosome 2L, 2LHet, 2RHet and 2R are concatenated (dashed lines between them); C indicates a centromere. **b,** Genome-wide correlation among H3K9me1/me2/me3, H3K27me3, and H3K36me3 (K562 in human, L3 in fly and worm; no H3K9me2 profile is available for human). **c,** Genomic coverage of H3K9me3 in multiple cell types and developmental stages. Embryonic cell lines/stages are marked with an asterisk and a black bar. **d,** Average gene body profiles of expressed (Exp) and silent genes in euchromatin and heterochromatin in all three species (K562 in human, L3 in fly and worm). **e,** Distributions of H3K9me3 and lamina-associated domains (LADs) in human chr2. LADs were profiled in Tig3 fibroblast. **f,** 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).

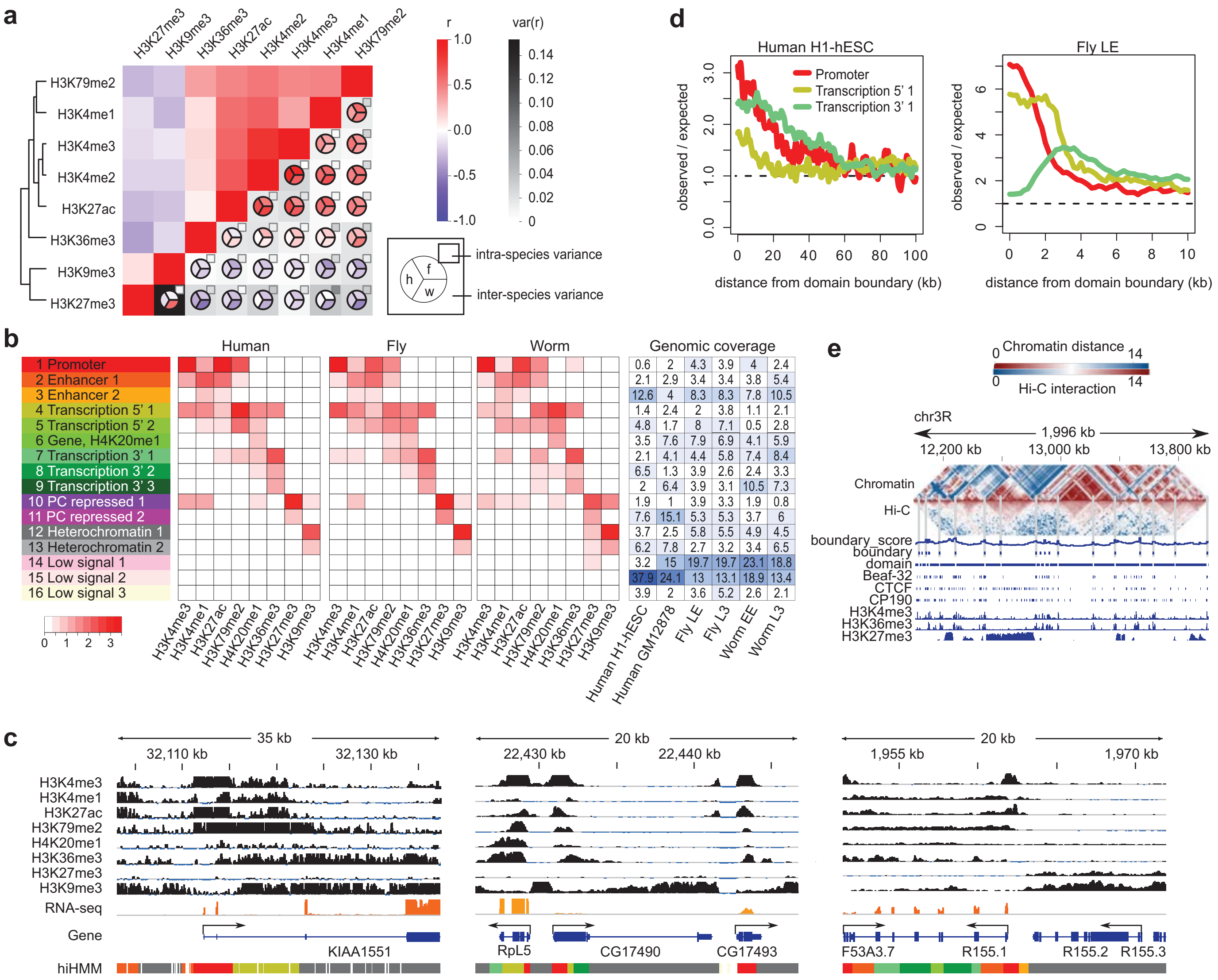
Fig. 4. Chromatin environment of protein-coding genes. a, Average gene body profiles of histone modifications on protein coding genes in human GM12878, fly L3 and worm L3. **b,** Comparative analysis of promoter architecture as shown by average profiles of H3K4me3 (human GM12878, fly L3 and worm L3), DNase hypersensitivity sites (DHS), GC content and nascent transcription (GRO-seq, in human IMR90 and fly S2) over all TSSs. **c,** Nucleosome frequency profiles (as represented as Z-scores) around TSSs for human CD4+ T cells, fly EE and worm adults. The profiles we computed for highly expressed (top 20%) and lowly expressed genes (bottom 20% for fly and human and 40% for worm; see Methods).

Fig. 5. Chromatin features and physical properties of enhancers. a, ChIP signal enrichment ($\log_2$ scale) of H3K4me3 vs. H3K4me1 at TSS-proximal (<250 bp) and TSS-distal (>1 kb) DHSs (blue: human GM12878, orange: fly S2) or CBP-1 binding sites (green: worm EE). **b,** Average

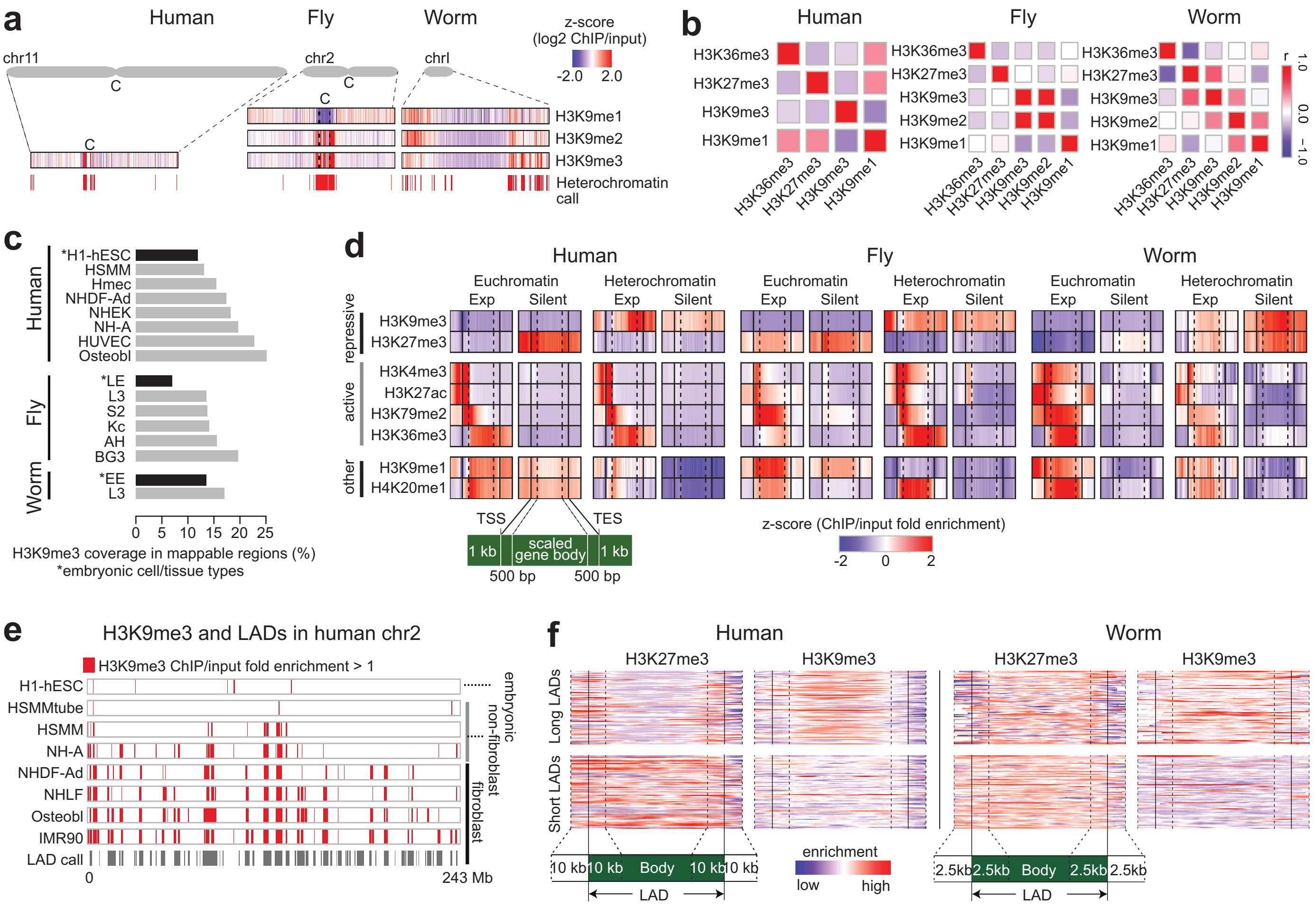
expression of genes that are close to enhancers with high (top 40%; red line) or low (bottom 40%; blue line) levels of H3K27ac in human GM12878, fly S2 and worm EE. As a control, we analyzed TSS-distal DHS (human and fly) or CBP-1 sites (worm) that are not classified as enhancers (dashed black). RPKM: reads per kilobase per million. Error bar: standard error of the mean. **c,** ChIP signal enrichment ($\log_2$ scale) of H3.3 around enhancers in human Hela-S3 cells (ChIP-seq) and fly S2 cells (ChIP-chip). **d,** z-score of average ChIP fold enrichment of key histone modifications and chromosomal proteins ±2 kb around the center of high H3K27ac and low H3K27ac enhancers. Grey bars indicate cases where data are not available. The centers of enhancers have higher average score for the DNA sequence conservation.

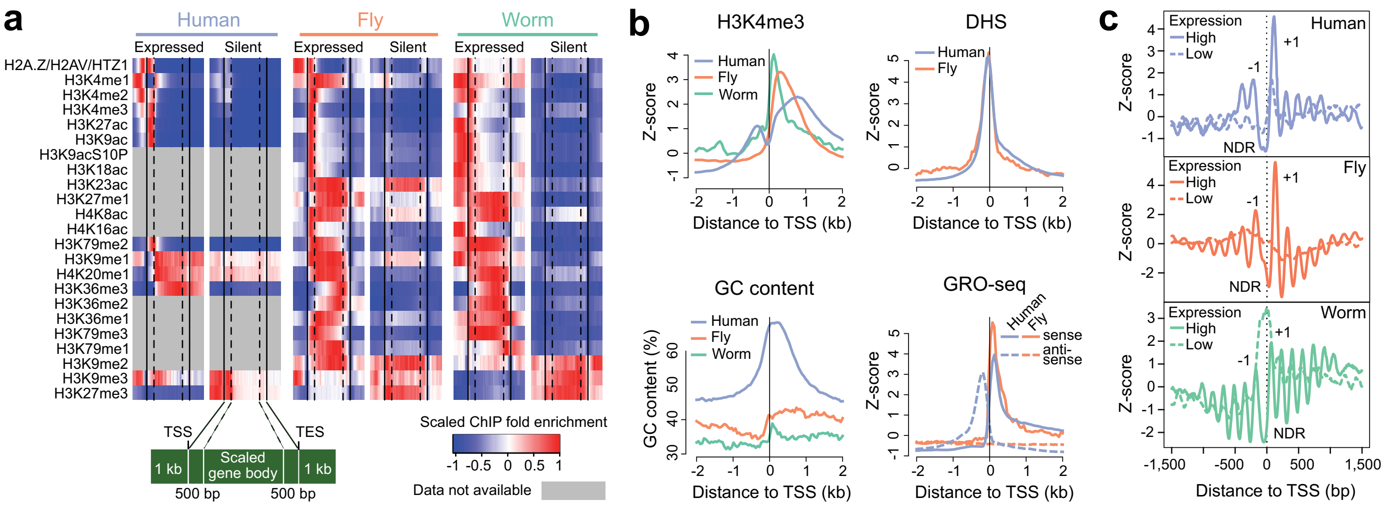
Fig. 6. DNA shape conservation in nucleosome sequences. a, Consensus ORChID2 profiles as a measure of DNA shape (y-axis) in 146-148 bp nucleosome-associated DNA sequences as identified by paired-end MNase-seq in human, fly and worm. A larger value of DNA shape (y-axis) corresponds to a wider minor groove and weaker negative charge. **b,** Normalized correlation (similarity) of ORChID2 profile of individual nucleosome-associated sequence with the consensus profile (see Methods and Supplementary Fig. 38). 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.

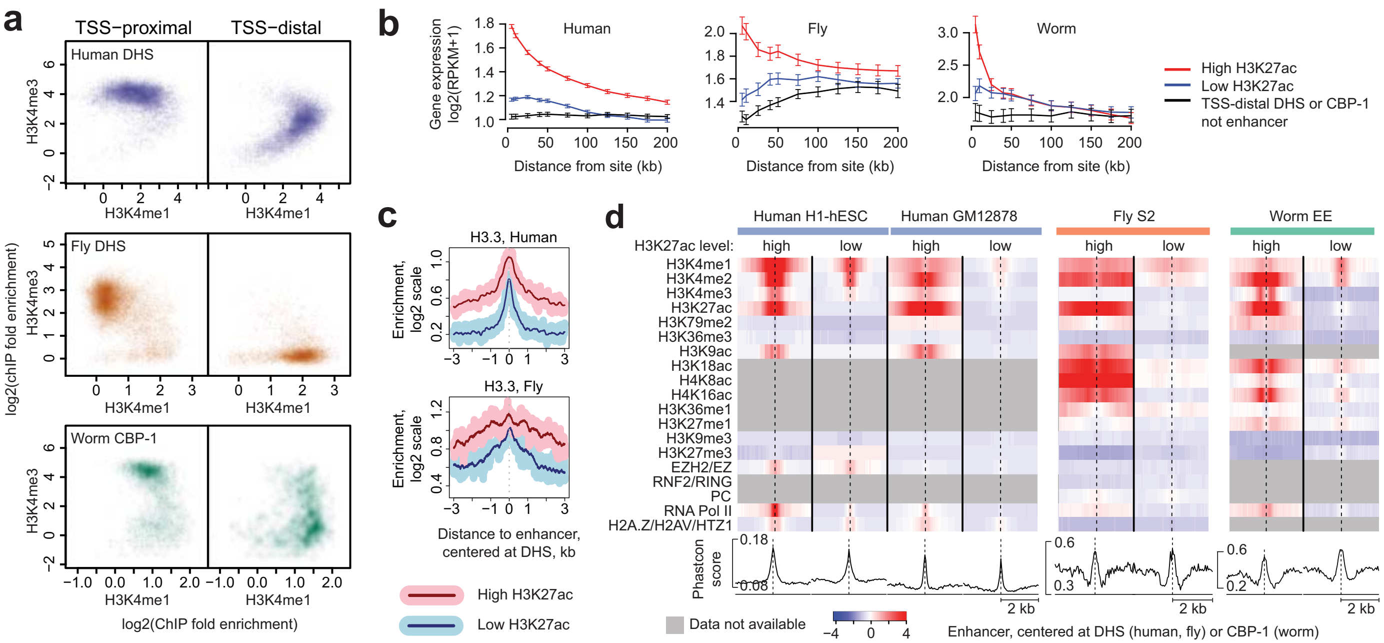

a
Histone
Nonhistone
Others
Human
Fly
Worm
H7−hESC
H1−hESC
K562
GM12878
HeLa−S3
HepG2
NHDF−Ad
HUVEC,NHEK
Osteoblast
IMR90
Fibroblast
PANC−1,SK−N−MC
Monocytes−CD14+_RO01746
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
S2
Kc
Clone 8
BG3
ES14,ES10,ES5
Early embryo 2−4hr (EE)
Late embryo 14−16hr (LE)
Third instar larvae (L3)
Adult head (AH)
(i)
Early embryo (EE)
Mixed embryo (MXEMB)
Late embryo (LTEMB)
Larvae stage 3 (L3)
Larvae stage 4 (L4)
AD germline
AD no embryos
AD−germlineless
(j)
H2A.Z/H2AV/HTZ−1
H3
H3.3
H3K4me1
H3K4me2
H3K4me3
H3K9ac
H3K9acS10ph
H3K9me1
H3K9me2
H3K9me3
H3K18ac
H3K23ac
H3K27ac
H3K27me1
H3K27me3
H3K36me1
H3K36me2
H3K36me3
H3K79me1
H3K79me2
H3K79me3
H4
H4acTetra
H4K8ac
H4K16ac
H4K20me1
CBP/CBP−1
CHD3/MI−2/LET−418
CTCF
EZH2/E(Z)
HDAC1/RPD3/HDA−1
HP1B/HPL−2
KDM1A
KDM2
KDM4A
NURF301/NURF−1
RNA Pol II
RNF2/RING
SMC3
DHS
GROseq
HiC
Lamina
Nucleosomes
Salt fractionated chromatin
🟩 ChIP−seq
🟨 ChIP−chip
🟧 non−ChIP
* external source
b
Human
Fly
Worm
Individual coverage
Cumulative coverage
Human
Fly
Worm
Coverage of mappable genome (%)
0
20
40
60
80
100
H3K4me3
H3K27ac
H3K9ac
H3K9me3
H3K4me2
H3K79me2
H3K36me3
H3K4me1
H3K27me3
H2A.Z
H3K9me1
H4K20me1
H3K4me3
H3K27ac
H3K9ac
H3K9me3
H3K4me2
H3K79me2
H3K36me3
H3K4me1
H3K27me3
H2AV
H3K9me1
H4K20me1
H3K79me3
H3K9me2
H3K36me1
H3K27me1
H3K79me1
H3K23ac
H4K16ac
H3K9acS10ph
H4K8ac
H2Bub
H3K27me2
H3K4me3
H3K27ac
H3K9ac
H3K9me3
H3K4me2
H3K79me2
H3K36me3
H3K4me1
H3K27me3
HTZ1
H3K79me3
H3K9me2
H3K36me1
H3K27me1
H3K79me1
H3K23ac
H3K36me2
H3K36ac
H3K18ac
(a) Dnd41,HMEC,HSMM,HSMMtube,NH−A, NHLF
(b) AG04449,AG04450,AG09309,AG09319, AG10803, AoAF,BE2_C,GM12864,GM12865,GM12875,HAc,HA−sp, HBMEC,HCFaa,HCM,HCPEpiC,HEEpiC,HFF,HFF−Myc, HL−60,HMF,HPAF,HPF,HRPEpiC,HVMF,NHDF−neo, RPTEC,WERI−Rb−1,WI−38
(c) GM10847,GM15510,GM18505,GM18526,GM18951, GM19099,GM19193,PBDE,PFSK−1,Raji,SK−N−SH,U87
(d) BJ,Caco−2,GM06990,HRE,SAEC,SK−N−SH_RA
(e) CD20+_RO01778,CD20+_RO01794,HCF,Jurkat,LNCaP, SKMC
(f) Gliobla,GM12891,GM12892,ProgFib
(g) A549,HCT−116,HEK293,MCF−7,NB4
(h) Fibrobl,GM12801,GM12872,GM12873,GM12874, GM19238,GM19239,GM19240
(i) Ovary,L3 Sexed Male,L3 Sexed Female
(j) Larvae stage 2 (L2),Larvae stage 1 (L1)

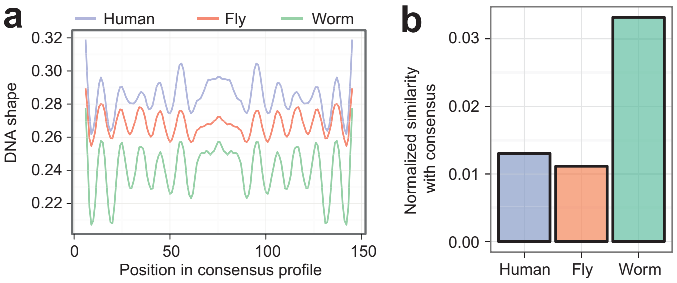


	Human H1-hESC	Human GM12878	Fly LE	Fly L3	Worm EE	Worm L3
1 Promoter	0.6	2	4.3	3.9	4	2.4
2 Enhancer 1	2.1	2.9	3.4	3.4	3.8	5.4
3 Enhancer 2	12.6	4	8.3	8.3	7.8	10.5
4 Transcription 5' 1	1.4	2.4	2	3.8	1.1	2.1
5 Transcription 5' 2	4.8	1.7	8	7.1	0.5	2.8
6 Gene, H4K20me1	3.5	7.6	7.9	6.9	4.1	5.9
7 Transcription 3' 1	2.1	4.1	4.4	5.8	7.4	8.4
8 Transcription 3' 2	6.5	1.3	3.9	2.6	2.4	3.3
9 Transcription 3' 3	2	6.4	3.9	3.1	10.5	7.3
10 PC repressed 1	1.9	1	3.9	3.3	1.9	0.8
11 PC repressed 2	7.6	15.1	5.3	5.3	3.7	6
12 Heterochromatin 1	3.7	2.5	5.8	5.5	4.9	4.5
13 Heterochromatin 2	6.2	7.8	2.7	3.2	3.4	6.5
14 Low signal 1	3.2	15	19.7	19.7	23.1	18.8
15 Low signal 2	37.9	24.1	13	13.1	18.9	13.4
16 Low signal 3	3.9	2	3.6	5.2	2.6	2.1


a
Human
Fly
Worm
z-score
(log2 ChIP/input)
-2.0
2.0
chr11
chr2
chrI
C
H3K9me1
H3K9me2
H3K9me3
Heterochromatin call
b
Human
Fly
Worm
H3K36me3
H3K27me3
H3K9me3
H3K9me2
H3K9me1
r
1.0
0.0
-1.0
c
Human
*H1-hESC
HSMM
Hmec
NHDF-Ad
NHEK
NH-A
HUVEC
Osteobl
Fly
*LE
L3
S2
Kc
AH
BG3
Worm
*EE
L3
0
5
10
15
20
25
H3K9me3 coverage in mappable regions (%)
*embryonic cell/tissue types
d
Human
Fly
Worm
Euchromatin
Heterochromatin
Exp
Silent
repressive
active
other
H3K9me3
H3K27me3
H3K4me3
H3K27ac
H3K79me2
H3K36me3
H3K9me1
H4K20me1
TSS
TES
1 kb
scaled gene body
1 kb
500 bp
500 bp
z-score (ChIP/input fold enrichment)
-2
0
2
e
H3K9me3 and LADs in human chr2
H3K9me3 ChIP/input fold enrichment > 1
H1-hESC
HSMMtube
HSMM
NH-A
NHDF-Ad
NHLF
Osteobl
IMR90
LAD call
0
243 Mb
embryonic
non-fibroblast
fibroblast
f
Human
Worm
H3K27me3
H3K9me3
Long LADs
Short LADs
10 kb
10 kb
Body
10 kb
10 kb
LAD
enrichment
low
high
2.5kb
2.5kb
Body
2.5kb
2.5kb
LAD


a
Human
Fly
Worm
Expressed
Silent
H2A.Z/H2AV/HTZ1
H3K4me1
H3K4me2
H3K4me3
H3K27ac
H3K9ac
H3K9acS10P
H3K18ac
H3K23ac
H3K27me1
H4K8ac
H4K16ac
H3K79me2
H3K9me1
H4K20me1
H3K36me3
H3K36me2
H3K36me1
H3K79me3
H3K79me1
H3K9me2
H3K9me3
H3K27me3
TSS
TES
1 kb
Scaled gene body
1 kb
500 bp
500 bp
Scaled ChIP fold enrichment
-1 -0.5 0 0.5 1
Data not available
b
H3K4me3
Human
Fly
Worm
Z-score
Distance to TSS (kb)
DHS
Human
Fly
GC content
GC content (%)
GRO-seq
Human
Fly
sense
anti-sense
c
Expression
High
Low
Human
Fly
Worm
-1
+1
NDR
Z-score
-1,500
-500
0
500
1,500
Distance to TSS (bp)


a
TSS−proximal
TSS−distal
Human DHS
H3K4me3
H3K4me1
Fly DHS
Worm CBP-1
log2(chIP fold enrichment)
log2(ChIP fold enrichment)
b
Gene expression
log2(RPKM+1)
Human
Fly
Worm
Distance from site (kb)
High H3K27ac
Low H3K27ac
TSS-distal DHS or CBP-1
not enhancer
c
H3.3, Human
H3.3, Fly
Enrichment,
log2 scale
Distance to enhancer,
centered at DHS, kb
High H3K27ac
Low H3K27ac
d
Human H1-hESC
Human GM12878
Fly S2
Worm EE
H3K27ac level:
high
low
H3K4me1
H3K4me2
H3K4me3
H3K27ac
H3K79me2
H3K36me3
H3K9ac
H3K18ac
H4K8ac
H4K16ac
H3K36me1
H3K27me1
H3K9me3
H3K27me3
EZH2/EZ
RNF2/RING
PC
RNA Pol II
H2A.Z/H2AV/HTZ1
Phastcon
score
0.18
0.08
0.6
0.3
0.6
0.2
2 kb
Data not available
−4
0
4
Enhancer, centered at DHS (human, fly) or CBP-1 (worm)


a
Human
Fly
Worm
0.32
0.30
0.28
0.26
0.24
0.22
DNA shape
0
50
100
150
Position in consensus profile
b
0.03
0.02
0.01
0.00
Normalized similarity with consensus
Human
Fly
Worm