modENCODE and ENCODE resources for analysis of metazoan chromatin organization

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^], Yuri B. Schwartz[21,22^], Dalal Asker[21,23], Jason A. Belsky[24], Sarah K. Bowman[10], Q. Brent Chen[5], Ron A-J Chen[11], Daniel S. Day[1,25], Yan Dong[11], Andrea C. Dose[9], Xikun Duan[19], Charles B. Epstein[16], Sevinc Ercan[5,26], Elise A. Feingold[13], Francesco Ferrari[1], 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[27], Tess E. Jeffers[5], Peter V. Kharchenko[1], Paulina Kolasinska-Zwierz[11], Chitra V. Kotwaliwale[9,28], Nischay Kumar[15,16], Sasha A. Langley[8,9], Erica N. Larschan[29], Isabel Latorre[11], Max W. Libbrecht[27,30], Xueqiu Lin[19], Richard Park[1,17], Michael J. Pazin[13], Hoang N. Pham[8,9,28], Annette Plachetka[2,20], Bo Qin[19], Noam Shoresh[16], Przemyslaw Stempor[11], Anne Vielle[11], Chengyang Wang[19], Christina M. Whittle[9,28], Huiling Xue[1,2], Robert E. Kingston[10], Ju Han Kim[7,31], Bradley E. Bernstein[16,28], Abby F. Dernburg[8,9,28], Vincenzo Pirrotta[21], Mitzi I. Kuroda[2,20], William S. Noble[27,30], Thomas D. Tullius[17,32], Manolis Kellis[15,16], David M. MacAlpine[24#], Susan Strome[18#], Sarah C. R. Elgin[14#], Xiaole Shirley Liu[3,4,16#], Jason D. Lieb[5&#], Julie Ahringer[11#], Gary H. Karpen[8,9#], 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. Institute of Endemic Diseases, Medical Research Center, Seoul National University, Seoul 110799, 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.   Department of Molecular Biology, Umea University, 901 87 Umea, Sweden
23.   Food Science and Technology Department, Faculty of Agriculture, Alexandria University, Alexandria, Egypt.
24.   Department of Pharmacology and Cancer Biology, Duke University Medical Center, Durham, NC, USA
25.   Harvard/MIT Division of Health Sciences and Technology, Cambridge, MA, USA
26.   Department of Biology, Center for Genomics and Systems Biology, New York, NY, USA
27.   Department of Genome Sciences, University of Washington, Seattle, WA, USA
28.   Howard Hughes Medical Institute, Chevy Chase, MD 20815 USA
29.   Department of Molecular Biology, Cellular Biology and Biochemistry, Brown University, Providence, RI
30.   Department of Computer Science and Engineering, University of Washington, Seattle, WA, USA
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

[*] Co-first authors

[^] Co-second authors

[#] Co-corresponding authors

[+] Present Address: Victor Chang Cardiac Research Institute and The University of New South Wales, Sydney, Australia

[&] Present Address: Department of Molecular Biology and Lewis Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08540

Abstract

Chromatin influences nearly every aspect of eukaryotic genome function. To investigate chromatin organization and regulation across species, we generated a large collection of genome-wide chromatin datasets from cell lines and developmental stages of *Homo sapiens, Drosophila melanogaster* and *Caenorhabditis elegans.* Here, we present a resource of >800 new datasets generated through the ENCODE and modENCODE consortia, bringing the total to over 1400. Comparison of combinatorial patterns of histone modifications, nuclear lamina-associated domains, organization of large-scale topological domains, chromatin environment at promoters and enhancers, nucleosome positioning, and DNA replication reveals many conserved features of chromatin organization among the three organisms. We also find significant differences, most notably in the composition and chromosomal locations of repressive chromatin. These datasets and analyses provide a rich resource for comparative and species-specific investigations of chromatin composition, organization, and function.

Introduction. Utilization of information contained in genome sequences is dynamically regulated by chromatin, which consists of DNA, histones, non-histone proteins, and RNA. Studies in *C. elegans* (worm) and *D. melanogaster* (fly) have contributed significantly to our understanding of genetic and molecular mechanisms of genome functions in humans, and have revealed that the components and mechanisms involved in chromatin regulation are often conserved. Nevertheless, the three organisms have prominent differences in genome size (human: $\sim3.4\times10^9$ bp, fly*: $\sim1.7\times10^8$ bp, worm*: $\sim1.0\times10^8$ bp), chromosome architecture, and gene organization. For instance, human protein-coding regions occupy only 3.0% of the assembled genome compared to 28% in fly and 34% in worm (see Gerstein *et al*., The Comparative ENCODE RNA Resource Reveals Conserved Principles of Transcription, co-submitted). Human and fly chromosomes have single centromeres flanked by extensive stretches of pericentric heterochromatin, whereas worm chromosomes have centromeres distributed along their length with dispersed heterochromatin-like regions enriched in the distal chromosomal 'arms'. Comparative studies among species are necessary to determine if global differences in chromosome organization reflect functional variation at the level of chromatin composition and structure. Such comparisons will also uncover chromatin features that are conserved among eukaryotes and potential species-specific mechanisms for regulation of genome functions (see Boyle *et al*., Comparative analysis of regulatory information and circuits across distant species, co-submitted).

A community resource of modENCODE and ENCODE chromatin data. Here we present 1453 chromatin datasets from the modENCODE and ENCODE consortia, of which 815 are new, including the majority of the sequencing-based datasets in fly and worm and key histone mark profiles (e.g., H3K9me3) in an extended set of human cell lines. These datasets were created to determine the genome-wide distributions of a large number of chromatin features in multiple cell types and developmental stages (Supplementary Table 1), in order to facilitate exploratory analyses and hypothesis generation by the research community.

We used chromatin immunoprecipitation followed by DNA sequencing (ChIP-seq) or microarray hybridization (ChIP-chip) to generate profiles of core histones, histone

variants, histone modifications, and chromatin-associated proteins (Fig. 1; Supplementary Fig 1, Supplementary Table 2). Additional data include DNase I hypersensitivity sites in fly and human cells, and nucleosome occupancy maps in all three organisms. Compared to the initial consortia publications[1-3], this represents a tripling of the number of fly and worm datasets and a substantial increase in human datasets (Fig. 1b,c). Uniform quality control standards for experimental protocols, antibody validation, and data processing were used throughout the projects[4] (see Methods). All data are freely available at modMine[5] (http://intermine.modencode.org) or the ENCODE Data Coordination Center[6] (http://genome.ucsc.edu/ENCODE/). We have also 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.

We used the human, fly, and worm chromatin data to perform a systematic comparison of chromatin composition and organization across these evolutionarily distant genomes, focusing largely on targets profiled in at least two organisms (Fig. 1) and from these sample types: human cell lines H1-hESC, GM12878 and K562; fly late embryos (LE), third instar larvae (L3) and cell lines derived from embryos or L3 (S2, Kc, BG3); and worm early embryos (EE) and stage 3 larvae (L3). Our analysis results, summarized in Table 1, reveal similarities and differences in chromatin composition and organization.

Most features of chromatin organization are conserved. Not surprisingly, the three species show many common chromatin features. Most of the genome in each species is covered by at least one histone modification (Supplementary Fig. 2). Consistent with the functional conservation of chromatin regulatory proteins, histone modifications in human, fly, and worm exhibit similar patterns around promoters, gene bodies, enhancers, and other chromosomal elements (Supplementary Figs. 3 –13). Nucleosome occupancy patterns around protein-coding genes and enhancers are also largely similar across species, although we observed subtle differences in H3K4me3 enrichment patterns around TSS across the three species (Supplementary Figs 12-15). The configuration and composition of large-scale features such as topological domains and lamina-associated domains are similar (Supplementary Figs. 16 –18). Lamina-associated domains in human

and fly are enriched for domains that replicate late in S-phase and for H3K27me3, suggesting that they may promote a repressive chromatin environment that impacts both DNA replication and transcription (Supplementary Fig. 19). Finally, DNA structural features associated with nucleosome positioning are strongly conserved across species (Supplementary Figs. 20, 21).

Consistent with previous studies, we find that in all three species, expressed genes show enrichment for H3K4me3 and other 'active' marks at the 5' ends, and H3K36me3 on gene bodies (peaking at the 3' end except for worm EE, as noted previously[7]), while repressed genes are enriched for H3K27me3 (Fig. 2a). The level of H3K36me3 enrichment in genes expressed with stage- or tissue-specificity is lower than on those expressed broadly, possibly because profiling was done on mixed tissues. (Supplementary Figs. 22–24; see Methods). However, we also observe notable differences. For example, H3K23ac is enriched at promoters of expressed genes in worm, but is enriched across gene bodies of both expressed and silent genes in fly. H4K20me1 is enriched on both expressed and silent human genes but only on expressed genes in fly and worm (Fig. 2a). We further explored genome-wide co-occurrence of pairs of histone modifications. While most pairwise co-occurrence patterns are similar across the three species, there are clearly some species-specific patterns (Supplementary Figs. 25–27).

Joint chromatin segmentation identifies shared and distinct chromatin states across species. Previous studies identified prevalent combinations of marks, or 'chromatin states' in human[8,9] and fly[1,10], which correlate with functional features such as promoters, enhancers, coding regions of active genes, Polycomb-associated silencing, and heterochromatin. Compared to individual marks, such 'chromatin state maps' provide a more concise and systematic cell type- or developmental stage-specific annotation of the genome. To compare chromatin states across the three organisms, we developed and applied a novel hierarchical non-parametric machine learning method called hiHMM (see Methods) to jointly generate chromatin state maps from eight histone marks mapped in common; the results were also confirmed using published methods (Fig. 2b; Supplementary Figs. 28–30).

Similar combinations of histone marks are enriched in each state across the three species, indicating that combinatorial patterns of histone modifications are conserved. Based on 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). The association of these chromatin states with gene regions, chromosomal proteins, and transcription factors are highly similar in the three organisms (Supplementary Figs. 31–34).

Heterochromatin is more prevalent in differentiated cells relative to embryonic or stem cells. Heterochromatin is a classically defined and distinct chromosomal state that plays important roles in genome organization, genome stability, chromosome inheritance, and gene regulation. It is typically enriched for H3K9me3[11], which we used as a proxy for identifying heterochromatic domains in human, fly, and worm (Fig. 3a, Supplementary Figs. 35, 36; see Methods). As expected, the majority of the H3K9me3-enriched domains in human and fly are concentrated in the pericentromeric regions (as well as other specific domains, such as the Y chromosome and fly 4th chromosome), whereas in worm they are distributed throughout the distal chromosomal 'arms'[10,12,13] (Fig. 3a). In human, H3K9me3 is associated with more of the genome in differentiated cells than in stem cells[14] (Fig. 3b). Similarly, in fly and worm, we find that more of the genome contains H3K9me3 in differentiated cells/tissues compared to embryonic cells/tissues (Fig. 3b). We also observe large cell-type-specific blocks of H3K9me3 in human and fly[10,13,14] (Supplementary Fig.37). These results suggest a molecular basis for the classical concept of "facultative heterochromatin" formation to silence blocks of genes as cells specialize.

Organization and composition of transcriptionally 'silent' domains differ across species. Two distinct types of transcriptionally-repressed chromatin have been described. As illustrated above, classical 'heterochromatin' is generally concentrated in pericentromeric and telomeric chromosomal regions, and enriched for H3K9me3 and also H3K9me2[11]. In contrast, 'Polycomb-associated silenced domains' are scattered across the

genome, and are enriched for H3K27me3. These domains have been implicated in cell-type-specific silencing of developmentally regulated genes[10,13].

Our analyses identified several noteworthy features of silent chromatin. First, human, fly, and worm display significant differences in H3K9 methylation patterns. H3K9me2 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. 3c). The differences in H3K9 methylation patterns suggest potential differences in heterochromatin in the three organisms, which we explore further below. Second, the chromatin state maps reveal two distinct types of Polycomb-associated repressed regions: strong H3K27me3 accompanied by marks for active genes or enhancers (Fig. 2b, state 10; potentially due to mixed tissues for fly and worm) and strong H3K27me3 without active marks (state 11) (see also Supplementary Fig. 33). Third, we observe a worm-specific association of H3K9me3 and H3K27me3. These two marks are enriched together in states 12 and 13 in worm but not in human and fly.

The unexpected strong association between H3K9me3 and H3K27me3 in worm, which was observed with several validated antibodies (Supplementary Fig. 38), suggests a species-specific difference in the organization of silent chromatin. To explore this further, we compared the patterns of histone modifications on expressed and silent genes in euchromatin and heterochromatin (Fig 3d; see Supplementary Fig. 39 for other marks). We previously reported prominent depletion of H3K9me3 at the transcription start site (TSS) and high levels of H3K9me3 in the gene body of expressed genes located in fly heterochromatin[13], and now find a similar pattern in human (Fig. 3d; Supplementary Fig. 39). In these two species, H3K9me3 is highly enriched in the body of both expressed and silent heterochromatic genes. 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 (Fig. 3d and Supplementary Figs. 39, 40). There are also conspicuous differences in the patterns of H3K27me3 in the three organisms. For example, H3K27me3 is highly associated with developmentally-silenced genes in euchromatic regions of human and fly, but not with silent genes in heterochromatic regions. In

contrast, consistent with the worm-specific association between H3K27me3 and H3K9me3, we observe high levels of H3K27me3 on silent genes in worm heterochromatin, while silent euchromatic genes show modest enrichment of H3K27me3 (Fig. 3d and Supplementary Fig. 39).

Our results suggest the existence of three distinct types of repressed chromatin (Supplementary Figs. 41–42). The first type contains H3K27me3 but little or no H3K9me3 (represented by human and fly states 10 and 11 and worm state 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 and 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. The third type contains both H3K9me3 and H3K27me3 and occurs predominantly in worm (represented by worm states 10, 12, and 13). Co-occurrence of these marks is consistent with the previous observation that H3K9me3 and H3K27me3 are both required for silencing of heterochromatic transgenes in worms[15]. H3K9me3 and H3K27me3 may reside on the same or adjacent nucleosomes in individual cells[16,17], or alternatively the two marks may occur in different cell types in the embryos and larvae analyzed here. Future studies will be needed to resolve this and determine the functional consequences of the overlapping distributions of H3K9me3 and H3K27me3 observed in worm.

Chromatin states and topological domains. Genome-wide chromatin conformation capture (Hi-C) assays have revealed prominent topological domain structures in human[18] and fly[19,20]. The physical interaction domains defined by Hi-C often have boundaries that are enriched for insulator elements and active genes[18,19] (Supplementary Fig. 43). As has been recently observed in human[21], the interiors of individual Hi-C domains in both human and fly often contain a relatively uniform chromatin state which belongs to one of four common classes: active, Polycomb-repressed, heterochromatin, or low signal (Supplementary Fig. 44). In both species, roughly half of the active genes are found in small active physical domains, which cover about 15% of each genome.

We also generated a genome-wide similarity map for chromatin marks (see Fig. 3e and Methods). In fly, we find that chromatin state similarity between neighboring regions is predictive of three-dimensional chromatin interaction domains defined by Hi-C (Fig. 3e and Supplementary Fig. 45), indicating that topological domains can be largely recapitulated based on chromatin marks alone. This suggests that chromatin-based domain boundaries in worm or potentially other species can be used as a substitute for Hi-C data if such data are not available (Supplementary Figs. 46, 47).

Discussion. We have generated the largest collection of chromatin datasets to date across three representative metazoan species in different cell lines and developmental stages. These high-quality datasets will serve as a resource to enable future investigations of chromatin as a key regulator of genetic information in eukaryotes. Our cross-species analysis revealed both shared and distinct features of chromatin architecture among these organisms (Table 1). The strongest difference appears to be in the regulation of gene silencing, where different patterns of repressive histone modifications are observed (Figs. 2, 3).

Both *Caenorhabditis elegans* and *Drosophila melanogaster* 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 these model systems, extending their relevance to human biology. 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. More generally, the extensive public resources generated by this project provide a foundation for researchers to investigate how diverse genome functions are regulated in the context of chromatin structure.

Methods

For full details of Methods, see Supplementary Information.

References

1   The modENCODE Consortium *et al.* Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE. *Science* **330**, 1787-1797, doi:10.1126/science.1198374 (2010).

2   Gerstein, M. B. *et al.* Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science* **330**, 1775-1787, doi:10.1126/science.1196914 (2010).

3   Bernstein, B. E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).

4   Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research* **22**, 1813-1831, doi:10.1101/gr.136184.111 (2012).

5   Contrino, S. *et al.* modMine: flexible access to modENCODE data. *Nucleic acids research* **40**, D1082-1088, doi:10.1093/nar/gkr921 (2012).

6   Rosenbloom, K. R. *et al.* ENCODE whole-genome data in the UCSC Genome Browser: update 2012. *Nucleic acids research* **40**, D912-917, doi:10.1093/nar/gkr1012 (2012).

7   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).

8   Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43-49, doi:10.1038/nature09906 (2011).

9   Hoffman, M. M. *et al.* Integrative annotation of chromatin elements from ENCODE data. *Nucleic acids research* **41**, 827-841, doi:10.1093/nar/gks1284 (2013).

10   Kharchenko, P. V. *et al.* Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature* **471**, 480-485, doi:10.1038/nature09725 (2011).

11   Elgin, S. C. & Reuter, G. Position-effect variegation, heterochromatin formation, and gene silencing in Drosophila. *Cold Spring Harb Perspect Biol* **5**, a017780, doi:10.1101/cshperspect.a017780 (2013).

12   Liu, T. *et al.* Broad Chromosomal Domains of Histone Modification Patterns in C. Elegans. *Genome Research* **21**, 227-236, doi:10.1101/gr.115519.110 (2011).

13   Riddle, N. C. *et al.* Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin. *Genome research* **21**, 147-163, doi:10.1101/gr.110098.110 (2011).

14   Hawkins, R. D. *et al.* Distinct Epigenomic Landscapes of Pluripotent and Lineage-Committed Human Cells. *Cell Stem Cell* **6**, 479-491, doi:10.1016/j.stem.2010.03.018 (2010).

15   Towbin, B. D. *et al.* Step-wise methylation of histone H3K9 positions heterochromatin at the nuclear periphery. *Cell* **150**, 934-947, doi:10.1016/j.cell.2012.06.051 (2012).

16   Lindroth, A. M. *et al.* Dual histone H3 methylation marks at lysines 9 and 27 required for interaction with CHROMOMETHYLASE3. *The EMBO Journal* **23**, 4146-4155, doi:10.1038/sj.emboj.7600428 (2004).

17   Voigt, P. *et al.* Asymmetrically modified nucleosomes. *Cell* **151**, 181-193, doi:10.1016/j.cell.2012.09.002 (2012).

18   Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380, doi:10.1038/nature11082 (2012).

19   Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**, 458-472, doi:10.1016/j.cell.2012.01.010 (2012).

20   Hou, C., Li, L., Qin, Z. S. & Corces, V. G. Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. *Mol Cell* **48**, 471-484, doi:10.1016/j.molcel.2012.08.031 (2012).

21   Zhu, J. *et al.* Genome-wide chromatin state transitions associated with developmental and environmental cues. *Cell* **152**, 642-654, doi:10.1016/j.cell.2012.12.033 (2013).

Acknowledgement

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:** JAB, DSD, XD, FF, NG, PH, MMH, PVK, NK, ENL, MWL, RP, NS, CW, HX; **Data production team:** 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, DMM, SS, SCRE, JA, XSL, GHK, JDL, and PJP. **Group leaders for data analysis or production:** REK, JHK, BEB, AFD, VP, MIK, WSN, TDT, MK, DMM, SS, SCRE, JA, XSL, GHK, JDL, and PJP. **Overall project management, and corresponding authors:** DMM, SS, SCRE, XSL, JDL, JA, GHK, and PJP.

Completing Financial Interests

The authors declare no competing financial interests.

Supplementary Information (see attached)

Table 1. Summary of key features analyzed by cross-species comparisons.

Topic	Findings	Human	Fly	Worm	Fig.
Promoters	5' H3K4me3 enrichment	Bimodal peak around TSS	Single peak downstream of TSS	Weak bimodal peak around TSS	2a, S12-13
	Well positioned +1 nucleosome at expressed genes	Yes	Yes	Yes	S14
Gene bodies	Lower H3K36me3 in specifically expressed genes	Yes	Yes	Yes	S22-S24
Enhancers	High H3K27ac sites are more active	Yes	Yes	Yes	S5-6
	High H3K27ac sites have higher nucleosome turnover	Yes	Yes	ND	S7
Nucleosome positioning	10-bp periodicity profile	Yes	Yes	Yes	S20a
	Positioning signal in genome	Weak	Weak	Less weak	S20b
LADs	Short LADs	H3K27me3	H3K27me3	H3K27me3	S18
	Long LADs	H3K9me3 internal, H3K27me3 borders	ND	H3K9me3+H3K27me3	S16
	Late replication in S-phase	Yes	Yes	ND	S19
Genome-wide correlation	Correlation between H3K27me3 and H3K9me3	Low	Low	High	S25,41
Chromatin state maps	Similar histone marks and genomic features at each state	Yes	Yes	Yes	2b, S31-34
Silent domains: constitutive heterochromatin	Composition	H3K9me3	H3K9me3	H3K9me3+H3K27me3	2b
	Predominant location	Pericentric+Y	Pericentric+chr4+Y	Arms	3a, S42
	Depletion of H3K9me3 at TSS of expressed genes	Yes	Yes	Weak	3d
Silent domains: Polycomb-associated	Composition	H3K27me3	H3K27me3	H3K27me3	2b
	Predominant location	Arms	Arms+Chr4	Arms+Centers	3a,S42
Topological domains	Active promoters enriched at boundaries	Yes	Yes	ND	S43
	Similar chromatin states are enriched in each domain	Yes	Yes	ND	S44

ND: No Data

Figure legends

Fig. 1. Dataset overview. a, Histone modification, chromosomal protein, and other profiles that were mapped in at least two species; a full dataset is shown in Supplementary Fig. 1. 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 (*). **b,** Number of all datasets generated by this (New; red) and the previous consortium-wide publications[1-3] (Old; pink). Each dataset corresponds to a replicate-merged normalized profile of a histone, histone variant, histone modification, non-histone chromosomal protein, nucleosome, or salt-fractionated nucleosome. **c,** Number of unique histone marks or non-histone chromosomal proteins that have been profiled to date by the consortia.

Fig. 2. Shared and organism-specific chromatin states. a, Average gene body profiles of histone modifications on protein coding genes in human GM12878, fly L3 and worm L3. **b,** 16 chromatin states derived by joint segmentation using hiHMM (hierarchical HMM; see Methods) 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. 28–34 for detailed analysis of the states). States are named by putative functional characteristics.

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

H3K9me3 and H3K27me3 of expressed and silent genes in euchromatin and heterochromatin in the three species (K562 cells in human, L3 in fly and worm). **e,** Comparison of Hi-C-based and chromatin-based topological domains in fly LE. Local histone modification similarity (Euclidian distance; see Methods) 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. 45).

(a) Ovary, L3 sexed male, L3 sexed female
(b) Larvae stage 2 (L2), larvae stage 1 (L1)

a

Human | Fly | Worm
Expressed | Silent | Expressed | Silent | Expressed | Silent

Row labels (top to bottom):
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 | Scaled gene body | TES
1 kb | 500 bp | | 500 bp | 1 kb

Scaled ChIP fold enrichment
−1 0 1
Data not available

b

Human | Fly | Worm
H3K4me3, H3K4me1, H3K27ac, H3K79me2, H4K20me1, H3K36me3, H3K27me3, H3K9me3

0 1 2 3

Genomic coverage (%)

	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

chr11 | chr2 | chrI

z-score
(log2 ChIP/input)
−2.0   2.0

C

H3K9me1
H3K9me2
H3K9me3
Heterochromatin call
H3K27me3

b

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

c

Human
H3K36me3
H3K27me3
H3K9me3
H3K9me1

Fly
H3K36me3
H3K27me3
H3K9me3
H3K9me2
H3K9me1

Worm
H3K36me3
H3K27me3
H3K9me3
H3K9me2
H3K9me1

r
1
0
−1

d

Human | Fly | Worm

H3K9me3 (ChIP/input)

H3K27me3 z-score

TSS   TES
1 kb scaled gene body 1 kb
500 bp   500 bp

Heterochromatin
— Expressed
— Silent

Euchromatin
— Expressed
— Silent

e

Chromatin distance
0   15
0   Hi-C interaction   14

chr3R
12,200 kb 12,600 kb 13,000 kb 13,400 kb 13,800 kb
2,000 kb

Chromatin
Hi-C
boundary score
boundary
domain
Beaf-32
CTCF
CP190
H3K4me3
H3K36me3
H3K27me3