Title
Comprehensive analysis of the chromatin landscape in Drosophila melanogaster

Journal
Nature, 471(7339)

ISSN
0028-0836

Authors
Kharchenko, Peter V
Alekseyenko, Artyom A
Schwartz, Yuri B
et al.

Publication Date
2011-03-01

DOI
10.1038/nature09725

Peer reviewed

Published in final edited form as:

Nature. 2011 March 24; 471(7339): 480–485. doi:10.1038/nature09725.

Comprehensive analysis of the chromatin landscape in Drosophila

Peter V. Kharchenko[1,2], **Artyom A. Alekseyenko**[3,4], **Yuri B. Schwartz**[5,†], **Aki Minoda**[6], **Nicole C. Riddle**[7], **Jason Ernst**[8,9], **Peter J. Sabo**[10], **Erica Larschan**[3,4,11], **Andrey A. Gorchakov**[3,4], **Tingting Gu**[7], **Daniela Linder-Basso**[5,§], **Annette Plachetka**[3,4], **Gregory Shanower**[5,‡], **Michael Y. Tolstorukov**[1,2], **Lovelace J. Luquette**[1], **Ruibin Xi**[1], **Youngsook L. Jung**[1,3], **Richard W. Park**[1,12], **Eric P. Bishop**[1,12], **Theresa P. Canfield**[10], **Richard Sandstrom**[10], **Robert E. Thurman**[10], **David M. MacAlpine**[13], **John A. Stamatoyannopoulos**[10,14], **Manolis Kellis**[8,9], **Sarah C. R. Elgin**[7], **Mitzi I. Kuroda**[3,4], **Vincenzo Pirrotta**[5], **Gary H. Karpen**[6,*], and **Peter J. Park**[1,2,3,*]

[1]Center for Biomedical Informatics, Harvard Medical School, Boston, MA, USA

[2]Children's Hospital Informatics Program, Boston, MA USA

[3]Division of Genetics, Department of Medicine, Brigham & Women's Hospital, Boston, MA USA

[4]Department of Genetics, Harvard Medical School, Boston, MA, USA

[5]Department of Molecular Biology & Biochemistry, Rutgers University, Piscataway, NJ, USA

[6]Department of Molecular and Cell Biology, University of California at Berkeley, and Department of Genome Dynamics, Lawrence Berkeley National Lab, Berkeley, CA, USA

[7]Department of Biology, Washington University in St. Louis, St. Louis, MO, USA

[8]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge MA, USA

[9]Broad Institute of MIT and Harvard, Cambridge, MA, USA

[10]Department of Genome Sciences, University of Washington, Seattle, WA, USA

[11]Department of Molecular Biology, Cell Biology and Biochemistry, Brown University, Providence, RI, USA

[12]Graduate Program in Bioinformatics, Boston University, Boston, MA, USA

[*]Correspondence and requests for materials should be addressed to P.J.P (peter_park@harvard.edu) and G.H.K (karpen@fruitfly.org).
[†]Present address: Department of Molecular Biology, Umea University, 901 87, Umea, Sweden.
[§]Present address: Department of Plant Biology and Pathology, SEBS, Rutgers University
[‡]Present address: Department of Basic Sciences, The Commonwealth Medical College, Scranton, PA

Author contribution P.V.K. performed most bioinformatic analysis. A.A.A., Y.B.S., A.M., N.C.R., E.L., A.A.G., T.G., D.L., A.P., and G.S. generated data, directed by S.C.R.E., M.I.K., V.P., and G.H.K. The 30-state analysis was performed by J.E., and M.K., while M.Y.T., L.J.L., R.X., Y.L.J., R.P., and E.P.B. performed additional bioinformatic analysis. P.J.S., T.P.C., R.S., R.E.T., and J.A.S. generated and processed DHS data. D.M.M. helped with replication analysis. P.J.P. supervised all analysis. G.H.K. coordinated the entire project. P.V.K., G.H.K., and P.J.P. wrote the manuscript, with contributions from S.C.R.E., M.I.K., V.P., Y.B.S, N.C.R, A.A.A, and A.M.

Author information The data are available from modENCODE site: http://www.modencode.org. GRO-seq data is available from Gene Expression Omnibus (GEO GSE25321). The authors declare no competing financial interests.

[13]Department of Pharmacology and Cancer Biology, Duke University Medical Center, Durham, NC, USA

[14]Department of Medicine, University of Washington, Seattle, WA, USA

Summary

Chromatin is composed of DNA and a variety of modified histones and non-histone proteins, which impact cell differentiation, gene regulation and other key cellular processes. We present a genome-wide chromatin landscape for *Drosophila melanogaster* based on 18 histone modifications, summarized by 9 prevalent combinatorial patterns. Integrative analysis with other data (non-histone chromatin proteins, DNaseI hypersensitivity, GRO-seq reads produced by engaged polymerase, short/long RNA products) reveals discrete characteristics of chromosomes, genes, regulatory elements, and other functional domains. We find that active genes display distinct chromatin signatures that are correlated with disparate gene lengths, exon patterns, regulatory functions, and genomic contexts. We also demonstrate a diversity of signatures among Polycomb targets that include a subset with paused polymerase. This systematic profiling and integrative analysis of chromatin signatures provides insights into how genomic elements are regulated, and will serve as a resource for future experimental investigations of genome structure and function.

The model organism Encyclopedia of DNA Elements (modENCODE) project is generating a comprehensive map of chromatin components, transcription factors, transcripts, small RNAs, and origins of replication in *D. melanogaster* and *C. elegans*1,2. Drosophila has been used as a model system for over a century to study chromosome structure and function, gene regulation, development, and evolution. The availability of high-quality euchromatic and heterochromatic sequence assemblies3-5, extensive annotation of functional elements6, and a vast repertoire of experimental manipulations enhance the value of epigenomic studies in Drosophila.

Genome-wide profiling of chromatin components provides a rich annotation of the potential functions of the underlying DNA sequences. Previous work has identified patterns of post-translational histone modifications and non-histone proteins associated with specific elements (*e.g.* transcription start sites, enhancers), as well as delineating the transcriptional status of genes and large domains7,8. Here, we present a comprehensive picture of the chromatin landscape in a model eukaryotic genome. We define combinatorial chromatin ‘states’ at different levels of organization, from individual regulatory units to the chromosome level, and relate individual states to genome functions.

Combinatorial chromatin states

We performed chromatin immunoprecipitation (ChIP)-array analysis for numerous histone modifications and chromosomal proteins (Supp. Table 1), using antibodies tested for specificity and cross-reactivity9 (Supp. Figure 1). Here, we describe analyses of cell lines S2-DRSC (S2) and ML-DmBG3-c2 (BG3), derived from late male embryonic tissues (stages 16-17) and the central nervous system of male third instar larvae, respectively (see http://www.modencode.org for data from other cell lines and animal stages). Analysis

reveals groups of correlated features, including those associated with heterochromatic regions10, Polycomb-mediated repression11, and active transcription12 (Supp. Figure 2), similar to those observed in other organisms13,14. This suggests that specific histone modifications work together to achieve distinct chromatin “states”.

We utilized a machine-learning approach to identify the prevalent combinatorial patterns of 18 histone modifications, capturing the overall complexity of chromatin profiles observed in S2 and BG3 genomes with 9 combinatorial states (Figure 1a, Methods). The model associates each genomic location with a particular state, generating a chromatin-centric annotation of the genome (Figure 1b). We examined each state for enrichment in non-histone proteins (Figure 1a, Supp. Figure 3) and gene elements, as well as distribution across the karyotype (Figure 1b, Supp. Figure 4) and finer-scale levels (Figure 1c-e).

Most distinct chromatin states are associated with transcriptionally active genes. Active promoter and transcription start site (TSS)-proximal regions are identified by state 1 (Figure 1; red), marked by prominent enrichment in H3K4me3/me2 and H3K9ac. The transcriptional elongation signature associated with H3K36me3 enrichment is captured by state 2 (purple), found preferentially over exonic regions of transcribed genes. State 3 (brown), typically found within intronic regions, is distinguished by high enrichment in H3K27ac, H3K4me1, and H3K18ac. A related chromatin signature is captured by state 4 (coral), distinguished by enrichment of H3K36me1, but notably lacking H3K27ac. The number of genes associated with each chromatin state and the distribution of states within genes are shown in Supp. Figure 5.

Several aspects of large-scale organization are revealed by the karyotype view (Figure 1b). Chromosome X is strikingly enriched for state 5 (green), distinguished by high levels of H4K16ac in combination with some enrichment in H3K36me3 and other marks of “elongation” state 2 (a pattern associated with dosage compensation in male cells15). Pericentromeric heterochromatin domains and chromosome 4 are characterized by high levels of H3K9me2/me3 (state 7, dark blue)10. Finally, the model distinguishes another set of heterochromatin-like regions containing moderate levels of H3K9me2/me3 (state 8, light blue, Figure 1e). Surprisingly, this state occupies extensive domains in autosomal euchromatic arms in BG3 cells, and in chromosome X in both cell lines16.

Further aspects of chromatin organization can be visualized by folding the chromosome using a Hilbert curve (Figure 2a)17, which maintains the spatial proximity of nearby elements. Thus, local patches of corresponding colors reveal the sizes and relative positions of domains associated with particular chromatin states (Figure 2b; Supp. Figures 6-9). For instance, specks of TSS-proximal regions (state 1) are typically contained within larger blocks of transcriptional elongation marks (state 2), which are in turn encompassed by extensive patches of H3K36me1-enriched domains (state 4) and variable-sized blocks of state 3. The clusters of open chromatin formed by these gene-centric patterns are separated by extensive silent domains (state 9) and regions of Polycomb-mediated repression (state 6). Factors responsible for domain boundaries were not identified in our analysis (Supp. Figure 10).

We also developed a multi-scale method to characterize chromatin organization at the spatial scale appropriate for the genome properties being investigated. For example, we observe that chromatin patterns most accurately reflect the replication timing of the S2 genome at scales of ~170kb (Supp. Section 1). This is consistent with size estimates of chromatin domains influencing replication timing18, and suggests that multiple replication origins are coordinately regulated by the local chromatin environment (each replicon is ~28-50kb 19).

To examine combinatorial patterns not distinguished by the simplified 9-state model, we also generated a 30-state combinatorial model that utilizes presence/absence probabilities of individual marks20 (Supp. Figure 11). The increased number of states may identify finer variations that are biologically significant, *e.g*., a signature corresponding to transcriptional elongation in heterochromatic regions16.

Chromatin state variation among genes

Active genes generally display enrichments or depletions of individual marks at specific gene segments (Figure 3a). When classified according to their chromatin signatures (Supp. Figure 12), active genes fall into subclasses correlated with expression magnitude (Supp. Section 2), gene structure, and genomic context (*e.g*. heterochromatic genes combine H3K9me2/me3 with some active marks). Of particular interest is one class of long expressed genes, many with regulatory functions, which are enriched for H3K36me1 (cluster 2, Supp. Figure 12; 131 genes in S2, 202 in BG3; Supp. Table 2).

To further examine the patterns associated with long genes, we clustered expressed autosomal genes ≥4kb based on blocks of enrichment for each chromatin mark (Figure 3b; 1055 genes). We observe that genes with large 5’-end introns (green subtree, Figure 3b; 552 genes) show extensive H3K27ac and H3K18ac enrichment, broader H3K9ac domains, and blocks of H3K36me1 enrichment (chromatin state 3, Figure 3b, last column). These genes are enriched for developmental and regulatory functions (Supp. Table 3), and are positioned within domains of Nipped-B21 (Figure 3b), a cohesin-complex loading protein previously associated with transcriptionally active regions21,22. In contrast, genes with more uniformly distributed coding regions (red subtree, Figure 3b) lack most state 3 marks, and H3K9ac enrichment is restricted to the 2kb downstream of the TSS. These differences are not explained by variation in histone density (Supp. Figure 13). Overall, the presence or absence of state 3 is the most common difference in the chromatin composition of expressed genes that are 1kb and longer (Supp. Figure 14), and the presence of state 3 consistently correlates with a reduced fraction of coding sequence in the gene body, mainly associated with the presence of a long first intron.

State 3 domains are highly enriched for specific chromatin remodeling factors (SPT16 and dMI-2; Supp. Figures 15,16), whereas state 1 regions around active TSSs are preferentially bound by NURF301 and MRG15. ISWI is enriched in both states 1 and 3 (Supp. Figures 16,17). State 3 domains also exhibit the highest levels of nucleosome turnover23, and show higher enrichment of the transcription-associated H3.3 histone variant24 than either the TSS- or elongation-associated states 1 and 2 (Supp. Figures 15,16). Consistent with earlier

analyses of cohesin-bound regions25, state 3 sequences tend to replicate early in G1 phase, and show abundance of early replicating origins (Supp. Figure 18). A regulatory role for state 3 domains is suggested by enrichment for a known enhancer binding protein (dCBP/p300 26) in adult flies, and for enhancers validated in transgene constructs27 (Supp Figure 19).

Modes of regulation in Polycomb domains

In Drosophila, loci repressed by Polycomb group (PcG) proteins are embedded in broad H3K27me3 domains that are regulated by Polycomb Response Elements (PREs) bound by E(Z), PSC, and dRING (Figure 1d)28,29. We find that regions of H3K4me1 enrichment surround all PREs, 90% of which also display narrower peaks of H3K4me2 enrichment (Supp. Figure 20). While this pattern is reminiscent of transcriptionally-active promoter regions, PREs lack H3K4me3, suggesting that a different mechanism of H3K4 methylation is employed, perhaps involving the Trithorax H3K4 histone methyltransferase (HMTase) found at all PREs29.

To examine chromatin states associated with PcG targets, we analyzed the chromatin and transcriptional signatures of TSSs in Polycomb-bound domains (Figure 4a, Supp. Figure 21). In addition to fully repressed TSSs (cluster 1, Figure 4a), we identify TSSs maintained in the “balanced” state29 (cluster 2, Figure 4a), distinguished by coexistence of Polycomb with active marks (including the HMTase ASH1) and production of full-length mRNA transcripts (*e.g. Psc* domain, Figure 1d).

TSSs in clusters 3 and 4 are distinguished by the presence of adjacent PREs (Figure 4a). Surprisingly, 53% of the PRE-proximal TSSs produce short RNA transcripts30 (cluster 3, Figure 4a), suggesting stalling of engaged RNA pol II 30. Using the global run-on sequencing (GRO-seq) assay to accurately assess engaged RNA polymerases31, we observe that cluster 3 TSSs produce short transcripts in the sense orientation. The level of GRO+ signal is similar to that found at fully-transcribed genes (Supp. Figure 22); thus, transcription initiates in cluster 3, but elongation fails. Interestingly, these genes are enriched for regulatory and developmental functions, even more than other genes within Polycomb domains (see Supp Tables 4,5). Genes without TSS-proximal PREs generally lack short transcript signatures (*e.g*. clusters 1 in Figure 4a; see Supp. Figure 21 for exceptions). Importantly, engaged polymerases and transcripts are not a general feature of PREs; TSS-distal PREs typically lack short RNA and GRO-seq signals (Figure 4b, Supp. Figure 22) despite being similarly enriched in H3K4me1/me2. The striking link between TSS-proximal PREs and the production of short RNAs suggests a potential mechanism for control of these developmental regulatory genes, whereby the same features that recruit H3K4 methyl marks to PREs also facilitate RNA pol II recruitment to nearby TSSs.

DHS plasticity and chromatin states

We utilized a DNase I hypersensitivity assay32,33 to examine the distributions of putative regulatory regions and their relationships with chromatin states. DHS mapping broadly identifies sites with low nucleosome density and regions bound by non-histone proteins34,35. Short-read sequencing identified 8616 high-magnitude DNase I

hypersensitive sites (DHSs) in S2 cells and 6354 in BG3 cells (and a comparable number of low-magnitude DHSs, Supp. Figure 23; see Methods). Approximately half of the high-magnitude DHSs are found at transcriptionally-active TSSs (Supp. Figure 24). Thus, the chromatin context of the TSS-proximal DHSs is dominated by the features expected for an active TSS, including RNA Pol II, H3K4me3 and other state 1 marks (clusters 1,2 Figure 5a, Supp. Figure 25).

Of the 36% TSS-distal DHSs, most (60%) are positioned within annotated expressed genes (Supp. Figure 24). These gene-body DHSs are distinguished from TSS-proximal DHSs by low H3K4me3, higher levels of H3K4me1, H3K27ac, and other marks linked to chromatin state 3 (clusters 3,4 Figure 5a, Supp. Figure 26). An additional 20% of the TSS-distal DHSs are outside of annotated genes, but show signatures associated with active transcription starts or elongation, suggesting new alternative promoters or unannotated genes (Supp. Figures 27,28). The remaining 20% of TSS-distal DHSs that appear intergenic (6% of all DHSs) are typically enriched for H3K4me1, but lack other active marks (cluster 5, Figure 5a).

Most DHS positions fall into the TSS-proximal state 1 or the intron-biased state 3 (Figure 5b). State 3 lacks H3K4me3 and is enriched for H3K4me1/H3K27ac/H3K18ac, similar to mammalian enhancer elements36,37. Many state 3 DHS positions bind regulatory proteins: GAGA factor binds to 49% of these DHSs in S2 cells, and developmental transcription factors bind to 44% of these DHSs in embryos38. Intriguingly, we find that TSS-distal DHSs in Drosophila exhibit low-level bi-directional transcripts (Figure 5a shortRNA panel, Supp Figures 29,30), analogous to the enhancer RNAs (eRNAs) characterized in mice39. Analysis of GRO-seq data (Figure 5e) suggests that eRNA-like transcripts are common to both intra- and inter-genic TSS-distal DHSs in Drosophila, a feature that is conserved with mammals.

The association of DHSs with chromatin states 1 and 3 (Figure 5c) persists even in chromosome 4 and pericentromeric heterochromatin, where such states are infrequent (Supp. Figure 31). This suggests that these chromatin states and associated remodeling factors (*e.g.* ISWI, SPT16) provide the context necessary for non-histone chromosomal protein binding at DHSs, or are the consequence of such binding events. To investigate this interdependency, we analyzed a high-confidence set of loci that exhibit DHSs in only one of the two examined cell lines (Supp. Figure 32). Surprisingly, although in general more DHSs are in state 1 regions, 91% of the cell type-specific DHSs are found within state 3 domains (14-fold increase compared to state 1 DHSs; Supp. Table 6, Figure 5d). Comparison with DHSs in an additional cell type (Kc167, Supp. Figure 33) confirms that DHSs displaying plasticity between cell types are mostly found in state 3. When DHSs are absent, the altered loci maintain chromatin state 3 in 23% of the cases (Figure 5d), indicating that the presence of state 3 is not always dependent on the DHS. More frequently, the altered loci transition to state 4 (43% of the cases), an open chromatin state that lacks many of the histone modifications and chromatin remodelers characteristic of state 3. While the less frequent transitions to the Polycomb state 6 (7%) or background state 9 (17%) typically coincide with gene silencing, most of the genes that maintain state 3 or transition to state 4 remain transcriptionally active (Supp. Figure 34). These observations provide further support for an

enhancer-like function for state 3 DHSs, and suggest a more subtle regulatory role than simple linkage to the presence or absence of gene expression.

Chromatin annotation of genome functions

The genomic chromatin state annotation and discovery of refined chromatin signatures for chromosomes, domains, and subsets of regulatory genes demonstrate the utility of a systematic, genome-wide profiling of an organism that is already understood in considerable detail. Clearly, the definition and functional annotation of chromatin patterns will be enhanced by incorporation of data for different types of components. Five 'colors' of chromatin were recently identified in Kc167 cells using chromosomal protein maps40. Comparison with our 9-state model shows similarities as well as differences in the ability to distinguish functional elements (Supp. Figure 35); thus, further integration of such data in the same cell type may resolve additional functional features. Our results illustrate the utility of integrating multiple data types (histone marks, non-histone proteins, chromatin accessibility, short RNAs, and transcriptional activity) for comprehensive characterization of functional chromatin states.

An important, repeated theme is that chromatin state analysis identifies unexpected distinctions between subsets of active genes. Besides the differences linked to genomic context (*e.g.*, male X chromosome, heterochromatin), the main source of variability is the presence of the acetylation-rich state 3 (Figure 6). Several lines of evidence suggest that the intronic positions marked by state 3 are important for gene regulation. State 3 regions show specific associations with known chromatin remodelers (SPT16, dMi-2 and ISWI) and gene regulatory proteins (*e.g.* GAF, dCBP/p300), and the highest rates of nucleosome turnover and transcription-dependent deposition of the H3.3 variant. State 3 genes are also bound by cohesin complex proteins, thought to associate with decondensed chromatin21 to promote looping interactions with promoter regions22.

A regulatory role for state 3 chromatin is further suggested by the high density of DHSs, comparable to that of active TSS state 1, and the fact that state 3 accounts for most of the DHS plasticity among cell types. The combinations of histone marks found in state 3 are similar to signatures of mammalian enhancers36, which also show high variability between cell types37. Furthermore, state 3 DHSs exhibit low levels of short, non-coding bidirectional transcripts reminiscent of eRNAs identified in mice39. Together, these findings suggest that state 3 regions contain enhancers or other regulatory elements, and that a combination of modifications can be used to identify new elements in the genome.

Genes within repressive Polycomb domains also display several distinct combinatorial chromatin patterns (Figure 4a), which likely represent a range of functional states: repressed, paused, or expressed genes in either balanced29 or fully activated states. Alternatively, distinct signatures might mark subsets of regulatory genes that require either long-term repression or the ability to reverse functional states, depending on environmental or developmental cues. The PRE-proximal paused TSSs have some similarity to the "bivalent" genes in mammalian cells, which also display transcriptional pausing of key regulatory and developmental genes41,42. However, the mammalian "bivalent state" is characterized by the

simultaneous presence of PcG proteins, H3K27me3 and H3K4me3, which in Drosophila is found only in the fully-elongating "balanced" state29,43.

In summary, comprehensive analysis of chromatin signatures has enormous potential for annotating functional elements in both well-studied and new genomes. Going forward, our systematic characterization of the epigenomic and transcriptional properties of Drosophila cells should spur in-depth experimental analyses of the relationship between chromatin states and genome functions, ranging from whole chromosomes down to individual regulatory elements and circuits.

Methods Summary

Histone modification and chromosomal protein antibodies were characterized for cross-reactivity. ChIP-chip was performed in duplicate, using Affymetrix Drosophila Tiling 2.0R Arrays. Digital DNaseI-seq assays were performed as described previously44, and Global Run-On library (GRO-seq) data was generated as described in Core *et al*31. Short RNA data was generated by Nechaev *et al*30, and RNA-seq data was generated by Graveley *et al*.45. The chromatin state models were generated as Hidden Markov Models of different histone marks. DHSs were identified as read density peaks significantly enriched relative to the genomic DNA control. Clustering of chromatin signatures was determined using the PAM algorithm.

Methods

Growth conditions

ML-DmBG3-c2 cells were obtained from DGRC (https://dgrc.cgb.indiana.edu/), and S2-DRSC cells were from the DRSC (http://www.flyrnai.org/). All cell lines were grown to a density of ~5×10^6 cells/ml in Schneider's media (Gibco) supplemented with 10% FCS (HyClone). 10 μg/ml insulin was added to the ML-DmBG3-c2 media.

Antibodies

Antibodies are listed in Supplemental Table 1. Commercial antibodies against modified histones were tested by Western-blot for the lack of cross-reactivity with the corresponding recombinant histone produced in *E.coli* and non-histone proteins from embryonic nuclear extracts. Antibody specificity was further assayed by Western dot/slot blot against a panel of synthetic modified histone peptides. Only antibodies that showed <50% of total signal associated with non-histone proteins, and more than 5-fold higher affinity for the corresponding histone peptide, were used in ChIP experiments.

The specificity of antibodies against chromosomal proteins was tested by Western blots with nuclear extracts prepared from mutant flies or S2 cells subjected to RNAi knockdown.46. An antibody was considered specific if it recognized a major band of expected mobility that was absent in the sample prepared from mutant flies, or diminished 2-fold or more after RNAi depletion. When possible, distributions of a chromosomal protein were mapped with two antibodies generated against different epitopes (see Supp. Figure 17). Data from chromatin proteins for which only one antibody was available was validated by comparison

with published genomic distributions for a different component of the same complex, or to published genomic distributions generated with a different antibody.

ChIP and microarray hybridization

Crosslinked chromatin from cultured cells was prepared as described in Schwartz *et al*.28 with the following modifications. Prior to ultrasound shearing, cells were permeabilized with 1% SDS, and shearing was done in TE-PMSF (0.1% SDS, 10mM Tris-HCl pH8.0, 1mM EDTA pH8.0, 1mM PMSF) using a Bioruptor (Diagenode) (2 × 10 min, 1 × 5 min; 30sec on, 30 sec off; high power setting).

ChIP was performed as in Schwartz *et al*.28 and IP'd DNA was amplified using the whole genome amplification kit (WGA2, Sigma) according to the manufacturer's instructions (chemical fragmentation step was omitted). The amplified material was labeled and hybridized to Drosophila Tiling Arrays v2.0 (Affymetrix) as in Schwartz *et al*.28.

Processing of ChIP data

At least two independent biological replicates were assessed for each ChIP profile. The log2 intensity ratios (M values) were calculated for each replicate. The profiles were smoothed using local regression (lowess) with 500bp bandwidth, and the genome-wide mean was subtracted. The regions of significant enrichment were determined as clusters of at least 1kb in length, with gaps no more than 100bp where M value exceeds a statistically significant (0.1% FDR) enrichment threshold. The set of biological replicates was deemed consistent if the enriched regions from individual experiments had a 75% reciprocal overlap, or if at least 80% of the top 40% of the regions identified in each experiment were identified in the other replicate (before comparison the replicates were size-equalized by increasing the significance threshold for a replicate with more enriched sequence). The data from individual replicates were then combined using local regression smoothing, and used for all of the presented analysis, unless noted otherwise.

DNaseI hypersensitivity

Digital DNaseI-seq assays were performed as described previously44. The sequenced reads were aligned to the dm3 genome assembly, recording only uniquely mappable reads. To detect DNase I hypersensitive sites, hotspot positions were identified based on a 300bp scanning window statistic (Poisson model relative to 50kb background density, Z-score threshold of 2), and peaks of read density were selected within the hotspots using randomization-based thresholding at 0.1% FDR. The set of high-magnitude DHSs analyzed here (except for Supp. Figure 23) was identified as a subset of all peaks that show statistically significant enrichment over the normalized genomic DNA read density profile (using a 300bp window centered around the peak, binomial model, with Z-score threshold of 3). This method controls for copy number variation and sequencing/mapping biases, however it may also reduce the sensitivity of DHS detection. In the DHS chromatin profile clustering analysis (Figure 5a, relevant supplementary figures), DHSs found within 1kb of another DHS were excluded if their enrichment magnitude (relative to genomic background) was lower (to avoid showing the same region more than once).

RNA sequencing

The preparation of RNA-seq libraries and sequencing is described in Graveley *et al*.45. The sequenced reads were aligned to the dm3 genome assembly and annotated exon junctions, recording only uniquely mappable reads. The RPKM (reads per kilobase of exonic sequence per million reads mapped) was estimated for each exon. The total transcriptional output of each annotated gene was estimated based on the maximum of all exons within the gene. The presented analysis uses $\log_{10}(RPKM+1)$ values unless otherwise noted.

GRO sequencing

Global Run-On library was prepared from S2 cells and sequenced as described in Core *et al*31. The reads were aligned to the dm3 genome assembly, recording only uniquely mappable reads. The smoothed profiles of reads mapping to each strand were calculated using Gaussian smoothing (σ=100bp). The analysis uses $\log_{10}(d+1)$, where d is the smoothed density value.

Short RNA data processing

The short RNA data for S2 cells was generated by Nechaev *et al*30, and was aligned and processed in the same way as the GRO-seq data.

Chromatin state models

To derive a nine-state joint chromatin state model for S2 and BG3 cells (Figure 1a), the genome was first divided into 200bp bins, and the average enrichment level was calculated within each bin based on unsmoothed log2 intensity ratio values taking into account individual replicates, using all histone enrichment profiles and PC to discount the genome-wide difference in S2 H3K27me3 profiles. The bin-average values of each mark were shifted by the genome-wide mean, scaled by the genome-wide variance, and quantile-normalized between the two cells. The HMM with multivariate normal emission distributions was then determined from the Baum-Welch algorithm using data from both cell types, and 30 seeding configurations determined with K-means clustering. States with minor intensity variations (Euclidian distance of mean emission values < 0.15) were merged. Larger models (up to 30 states) were examined, and the final number of states was chosen for optimal interpretability.

An extensive discrete chromatin state model (Supp. Figure 11) was calculated as described in Ernst *et al*.20. The model was trained using 200bp grid with binary calls (enriched/not enriched). The binary calls were made based on a 5% FDR threshold determined from 10 genome-wide randomizations for each mark. For H1, H4 and H3K23ac regions of significant depletion rather than enrichment were called.

Regions of enrichment for individual marks (Figure 3)

To determine contiguous regions of enrichment for individual marks, a three-state HMM was used, with states corresponding to enriched, neutral, and depleted profiles (normally-distributed emission parameters: (μ=[-0.5 0 0.5], σ^2=0.3). The enriched regions were determined from the Viterbi path. The HMM segmentation was applied to unsmoothed M

value data taking into account individual biological replicates. The genes were clustered based on the combinatorial pattern of occurrence of enriched regions (coding exons and state panels were not used for clustering).

Classification of enrichment profiles (Figures 4,5)

Clustering of chromatin signatures around TSSs (Figure 4a), PREs (Figure 4b), and DHSs (Figure 5a, relevant supplements) was determined using the PAM algorithm. For clustering, each profile was summarized with average values within bins spanning ±2kb regions. 100bp bins were used for the central ±500bp region, 300bp bins outside.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank our technicians David Acevedo, Sarah Gadel, Cameron Kennedy, Ok-Kyung Lee, and Sarah Marchetti and Rutgers BRTC. We also thank our colleagues who donated antibodies: J. Kadonaga (H1), A.L. Greenleaf (RNA pol II), G. Reuter (SU(VAR)3-9), G. Cavalli (GAF), and I.F. Zhimulev/H. Saumweber (Chromator). The major support for this work came from the modENCODE grant U01HG004258 to G.H.K (PI) and S.C.R.E., M.I.K., P.J.P., and V.P. (co-PIs). Additional funding came from RC2 HG005639, U01 HG004279, R01 GM082798, R37 GM45744, RC1 HG005334, and NSF 0905968.

References

1. modENCODE. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science. in press.
2. Gerstein MB, et al. Integrative Analysis of the Caenorhabditis elegans Genome by the modENCODE Project. Science. in press.
3. Adams MD, et al. The genome sequence of Drosophila melanogaster. Science. 2000; 287:2185–2195. [PubMed: 10731132]
4. Clark AG, et al. Evolution of genes and genomes on the Drosophila phylogeny. Nature. 2007; 450:203–218. [PubMed: 17994087]
5. Hoskins RA, et al. Sequence finishing and mapping of Drosophila melanogaster heterochromatin. Science. 2007; 316:1625–1628. [PubMed: 17569867]
6. Tweedie S, et al. FlyBase: enhancing Drosophila Gene Ontology annotations. Nucleic Acids Res. 2009; 37:D555–559. [PubMed: 18948289]
7. Felsenfeld G, Groudine M. Controlling the double helix. Nature. 2003; 421:448–453. [PubMed: 12540921]
8. Mendenhall EM, Bernstein BE. Chromatin state maps: new technologies, new insights. Curr Opin Genet Dev. 2008; 18:109–115. [PubMed: 18339538]
9. Egelhofer TA, et al. An assessment of histone-modification antibody quality. Nat Mol Struct Biol. in press.
10. Eissenberg JC, Reuter G. Cellular mechanism for targeting heterochromatin formation in Drosophila. Int Rev Cell Mol Biol. 2009; 273:1–47. [PubMed: 19215901]
11. Schwartz YB, Pirrotta V. Polycomb complexes and epigenetic states. Current Opinion in Cell Biology. 2008; 20:266–273. [PubMed: 18439810]
12. Li B, Carey M, Workman JL. The role of chromatin during transcription. Cell. 2007; 128:707–719. [PubMed: 17320508]
13. Liu CL, et al. PLoS Biol. 2005; 3:e328. [PubMed: 16122352]
14. Barski A, et al. Cell. 2007; 129:823–837. [PubMed: 17512414]

15. Larschan E, et al. MSL complex is attracted to genes marked by H3K36 trimethylation using a sequence-independent mechanism. Mol Cell. 2007; 28:121–133. [PubMed: 17936709]
16. Riddle NC, et al. Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin. submitted.
17. Anders S. Visualization of genomic data with the Hilbert curve. Bioinformatics. 2009; 25:1231–1235. [PubMed: 19297348]
18. MacAlpine DM, Rodriguez HK, Bell SP. Coordination of replication and transcription along a Drosophila chromosome. Genes Dev. 2004; 18:3094–3105. [PubMed: 15601823]
19. Blumenthal AB, Kriegstein HJ, Hogness DS. The units of DNA replication in Drosophila melanogaster chromosomes. Cold Spring Harb Symp Quant Biol. 1974; 38:205–223. [PubMed: 4208784]
20. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol. 2010; 28:817–825. [PubMed: 20657582]
21. Misulovin Z, et al. Association of cohesin and Nipped-B with transcriptionally active regions of the Drosophila melanogaster genome. Chromosoma. 2008; 117:89–102. [PubMed: 17965872]
22. Kagey MH, et al. Mediator and cohesin connect gene expression and chromatin architecture. Nature. 2010; 467:430–435. [PubMed: 20720539]
23. Deal RB, Henikoff JG, Henikoff S. Genome-wide kinetics of nucleosome turnover determined by metabolic labeling of histones. Science. 2010; 328:1161–1164. [PubMed: 20508129]
24. Henikoff S, Henikoff JG, Sakai A, Loeb GB, Ahmad K. Genome-wide profiling of salt fractions maps physical properties of chromatin. Genome Res. 2009; 19:460–469. [PubMed: 19088306]
25. MacAlpine HK, Gordan R, Powell SK, Hartemink AJ, MacAlpine DM. Drosophila ORC localizes to open chromatin and marks sites of cohesin complex loading. Genome Res. 2010; 20:201–211. [PubMed: 19996087]
26. Visel A, et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature. 2009; 457:854–858. [PubMed: 19212405]
27. Zinzen RP, Girardot C, Gagneur J, Braun M, Furlong EE. Combinatorial binding predicts spatio-temporal cis-regulatory activity. Nature. 2009; 462:65–70. [PubMed: 19890324]
28. Schwartz YB, et al. Genome-wide analysis of Polycomb targets in Drosophila melanogaster. Nat Genet. 2006; 38:700–705. [PubMed: 16732288]
29. Schwartz YB, et al. Alternative Epigenetic Chromatin States of Polycomb Target Genes. PLoS Genet. 2010; 6:e1000805. [PubMed: 20062800]
30. Nechaev S, et al. Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. Science. 2010; 327:335–338. [PubMed: 20007866]
31. Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science. 2008; 322:1845–1848. [PubMed: 19056941]
32. Wu C. The 5’ ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I. Nature. 1980; 286:854–860. [PubMed: 6774262]
33. Wu C, Bingham PM, Livak KJ, Holmgren R, Elgin SC. The chromatin structure of specific genes: I. Evidence for higher order domains of defined DNA sequence. Cell. 1979; 16:797–806. [PubMed: 455449]
34. Elgin SC. The formation and function of DNase I hypersensitive sites in the process of gene activation. J Biol Chem. 1988; 263:19259–19262. [PubMed: 3198625]
35. 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. 2009; 41:941–945. [PubMed: 19633671]
36. Heintzman ND, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet. 2007; 39:311–318. [PubMed: 17277777]
37. Heintzman ND, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature. 2009; 459:108–112. [PubMed: 19295514]
38. MacArthur S, et al. Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. Genome Biol. 2009; 10:R80. [PubMed: 19627575]

39. Kim TK, et al. Widespread transcription at neuronal activity-regulated enhancers. Nature. 2010; 465:182–187. [PubMed: 20393465]

40. Filion GJ, et al. Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. Cell. 2010; 143:212–224. [PubMed: 20888037]

41. Bernstein BE, et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell. 2006; 125:315–326. [PubMed: 16630819]

42. Kanhere A, et al. Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2. Mol Cell. 2010; 38:675–688. [PubMed: 20542000]

43. Schuettengruber B, et al. Functional anatomy of polycomb and trithorax chromatin landscapes in Drosophila embryos. PLoS Biol. 2009; 7:e13. [PubMed: 19143474]

44. Sekimata M, et al. CCCTC-binding factor and the transcription factor T-bet orchestrate T helper 1 cell-specific structure and function at the interferon-gamma locus. Immunity. 2009; 31:551–564. [PubMed: 19818655]

45. Gravely B, et al. Celniker SE. Characterization of transcriptional activity in Drosophila melanogaster. in press.

46. Clemens JC, et al. Use of double-stranded RNA interference in Drosophila cell lines to dissect signal transduction pathways. Proc Natl Acad Sci U S A. 2000; 97:6499–6503. [PubMed: 10823906]

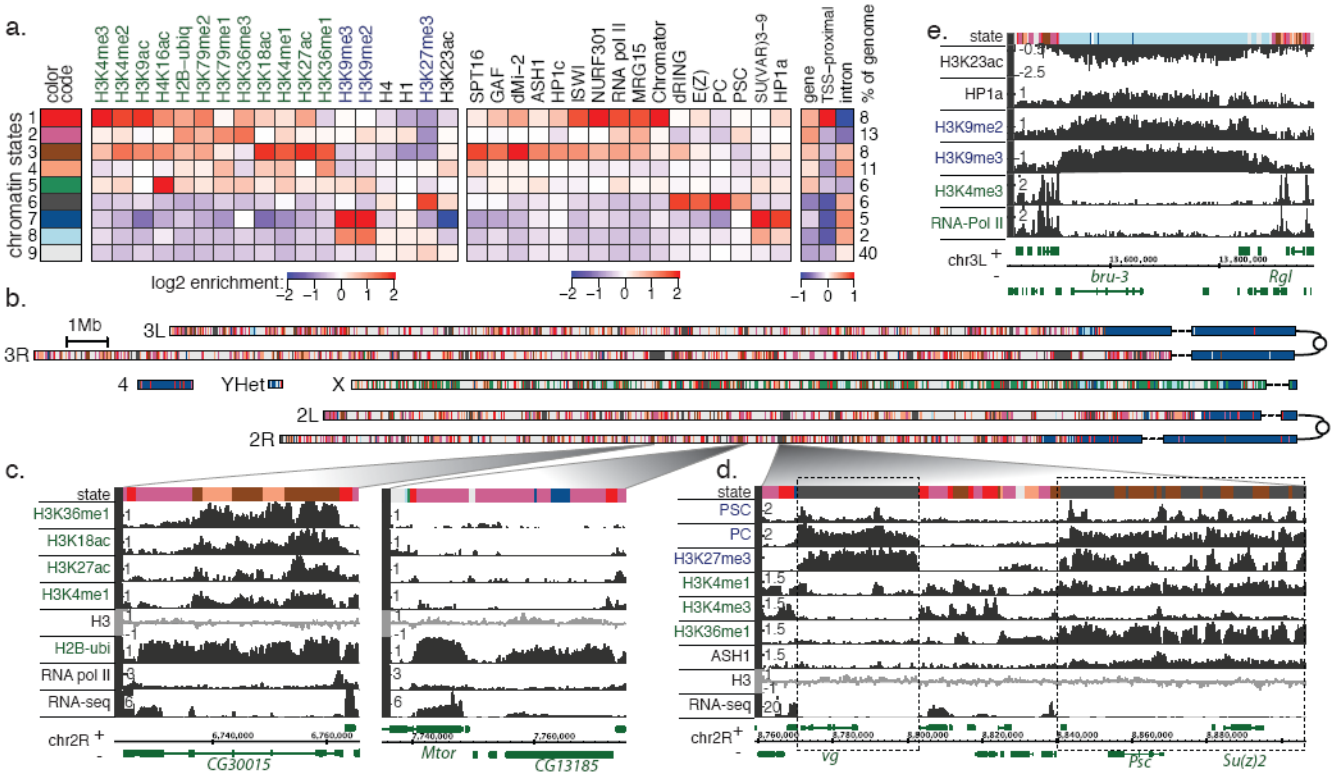

Figure 1. Chromatin annotation of the *Drosophila melanogaster* genome

a. A 9-state model of prevalent chromatin states found in S2 and BG3 cells. Each chromatin state (row) is defined by a combinatorial pattern of enrichment (red) or depletion (blue) for specific chromatin marks (first panel, columns). For instance, state 1 is distinguished by enrichment in H3K4me2/me3 and H3K9ac, typical of transcription start sites (TSS) in expressed genes. The enrichments/depletions are shown relative to chromatin input S2 data shown, see (Supp. Figure 3 for BG3 data and histone density normalization). The second panel shows average enrichment of chromosomal proteins. The third panel shows fold over/under-representation of genic and TSS-proximal (±1kb) regions relative to the entire tiled genome. The enrichment of intronic regions is relative to genic regions associated with each state.

b. A genome-wide karyotype view of the domains defined by the 9-state model in S2 cells. Centromeres are shown as open circles, and dashed lines span gaps in the genome assembly. Several prominent chromatin organization features are illustrated (color code in a), including the extent of pericentromeric heterochromatin (state 7), and the H4K16ac-driven signature of the dosage-compensated male X chromosome (state 5). (BG3 genome in Supp. Figure 4.)

c-e. Examples of chromatin annotation at specific loci. **c**. Two distinct chromatin signatures of transcriptionally active genes: one (left) is associated with enrichment in marks of states 3 and 4, while the other (right) is limited to states 1 and 2, recapitulating well-established TSS and elongation signatures (note: small patches of state 7 in CG13185 illustrate H3K9me2 found at some expressed genes in S2 cells16). **d**. A locus containing two Polycomb-associated domains, silent (left) and balanced (right). **e**. A large state 8 domain located within euchromatic sequence in BG3 cells, enriched for chromatin marks typically

associated with heterochromatic regions, but at lower levels than in pericentromeric heterochromatin (state 7).

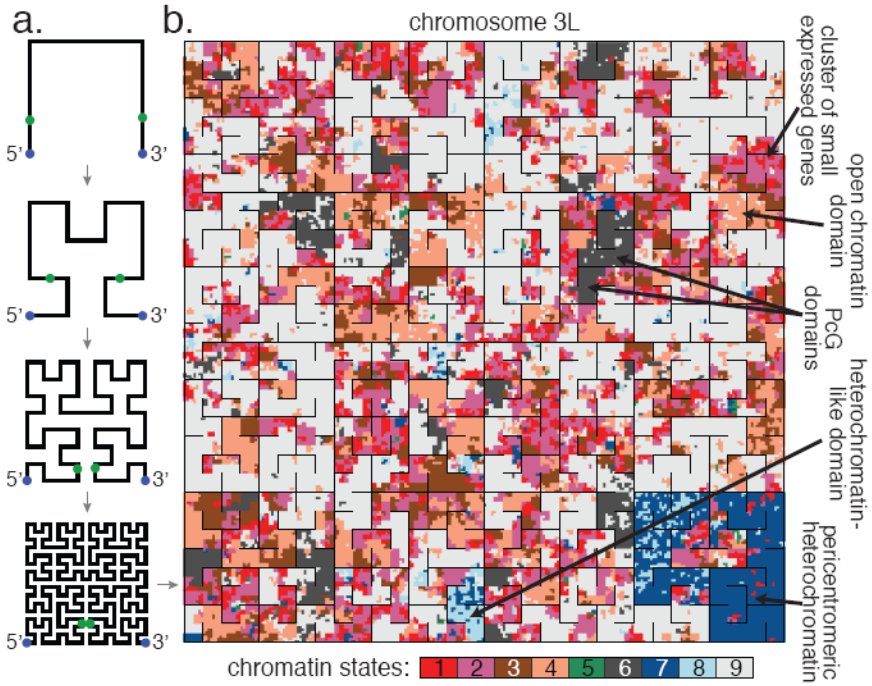

Figure 2. Visualization of spatial scales and organization using compact folding

a. The chromosome is folded using a geometric pattern (Hilbert space-filling curve) that maintains spatial proximity of nearby regions. An illustration of the first four folding steps is shown. Note that while this compact curve is optimal for preserving proximity relationships, some distal sites appear adjacent along the fold axis (green dots).

b. Chromosome 3L in S2 cells. A domain of a given chromatin state appears as a patch of uniform color of corresponding size. Thin black lines are used to separate regions that are distant on the chromosome. The folded view illustrates chromatin organization features that are not easily discerned from a linear view: active TSSs (state 1) appear as small specks surrounded by elongation state 2, commonly next to larger regions marked by H3K36me1-driven state 4, which also contains patches of intron-associated state 3. These open chromatin regions are separated by extensive domains of state 9. See Supp. Figures 6,7 for other chromosomes and BG3 data. The folded views can be browsed alongside the linear annotations and other relevant data online: http://compbio.med.harvard.edu/flychromatin.

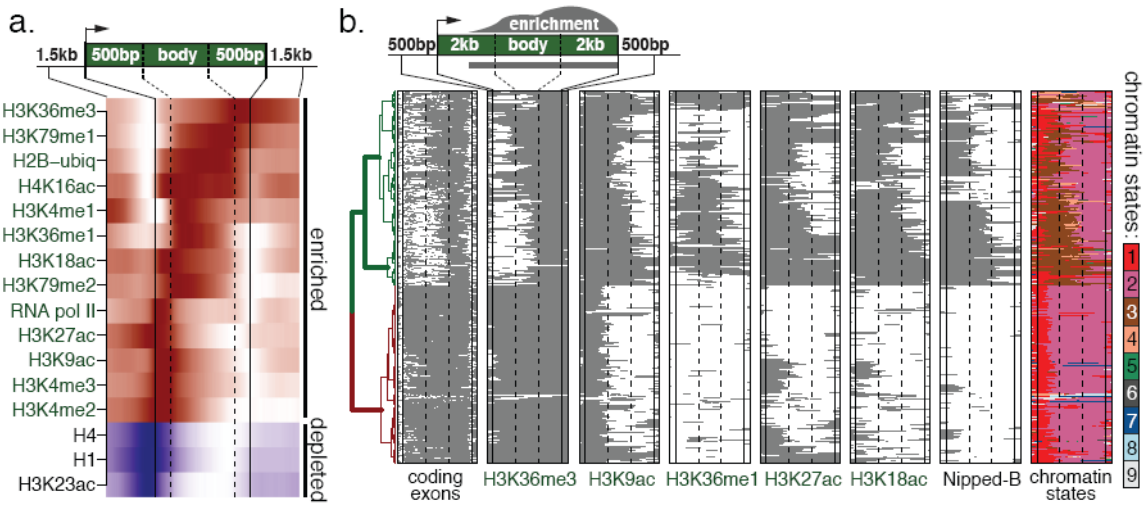

Figure 3. Chromatin patterns associated with transcriptionally active genes

a. Location and extent of chromatin features relative to boundaries of expressed genes (>1kb) in BG3 cells. The color intensity indicates the relative frequency of enrichment/ depletion of a given mark within the gene (normalized independently for each mark).

b. Regions enriched for ‘active’ chromatin marks in long transcribed genes. The plot shows the extent of regions enriched for various active marks at transcriptionally-active genes (>4kb) on BG3 autosomes. Each row represents a scaled gene. The first column illustrates coding exons; the last column shows chromatin state annotation. The clustering of the genes according to the spatial patterns of chromatin marks separates genes with a high fraction of coding sequence (red subtree, bottom) from genes containing long introns (green subtrees, top), which are associated with chromatin state 3 (last column) and binding of specific chromosomal proteins, such as Nipped-B21 (also see Supp. Figure 13).

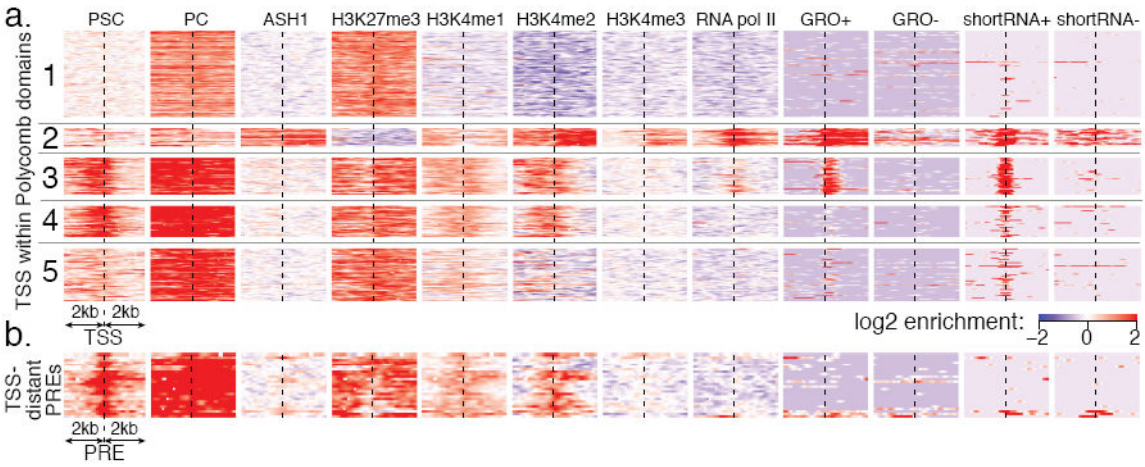

Figure 4. Signatures of TSSs within domains of Polycomb-mediated repression

a. Distinct classes of TSSs in S2 cell Polycomb domains. Each row represents a TSS. Clusters 1-5 illustrate distinct TSS states (see Supp. Figure 21 for complete set of clusters). Cluster 1 shows fully repressed TSSs with the expected pattern of PC and H3K27me3 enrichment; cluster 2 shows 21 TSSs found within ASH1 domains, maintained in a "balanced" state. Clusters 3 and 4 distinguish TSSs located in the immediate proximity of Polycomb response elements (PREs), showing the symmetric H3K4me1/me2 enrichment typical of all PREs. Many such TSSs (cluster 3, 42 TSSs) produce short, non-polyadenylated transcripts along the sense strand (GRO+/shortRNA+ columns), indicating the presence of paused polymerase.

b. PRE positions distant from annotated TSSs. TSS-distal PREs exhibit enrichment for H3K4me1/me2, but are not associated with GRO or shortRNA signatures.

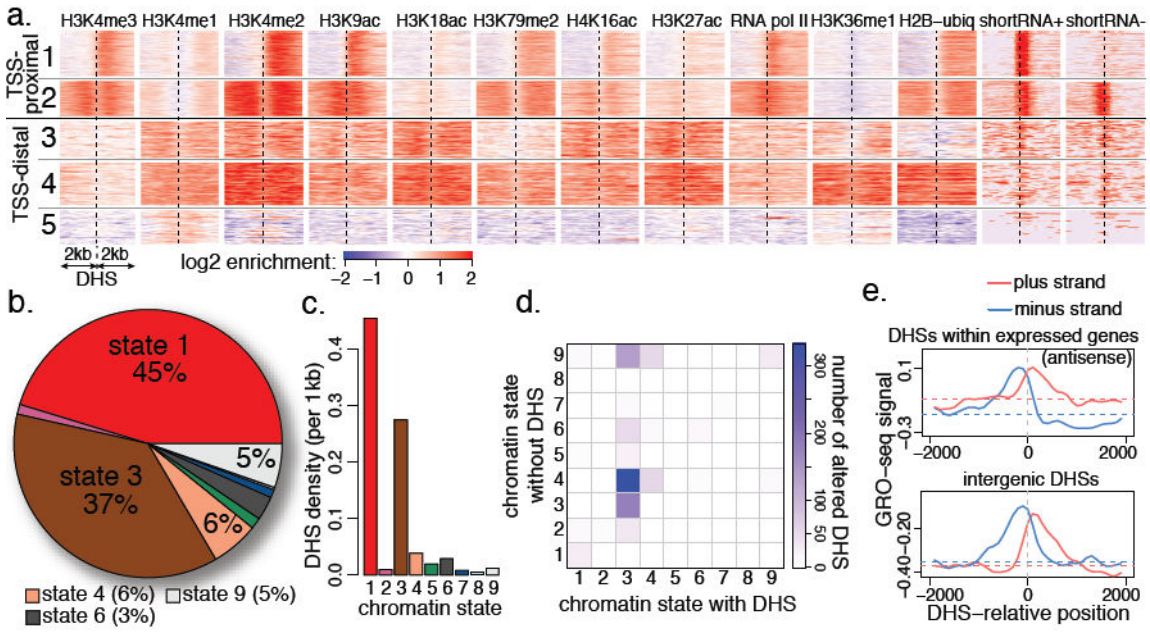

Figure 5. Chromatin signatures of regulatory elements identified by DNaseI hypersensitivity

a. Representative classes of high-magnitude DNaseI hypersensitive sites (DHSs) and chromatin signatures in S2 cells. TSS-proximal (within 2kb) DHSs show chromatin signatures expected of expressed gene promoters : high H3K4me3 and RNA pol II signal extending in the direction of transcription (left to right; cluster 2 groups bidirectional promoters). TSS-distal DHSs are associated with high H3K4me1 and low H3K4me3 levels. Most TSS-distal DHSs found within the bodies of expressed genes (clusters 3, 4) are associated with chromatin state 3. A cluster of rare intergenic DHSs (cluster 5) is associated with localized peaks of H3K4me1/2 (complete sets of clusters in Supp. Figures 25,26,28).

b. Distribution of DHS positions among chromatin states. The vast majority of DHSs are found within the TSS-proximal state 1 or enhancer-like state 3 regions.

c. States 1 and 3 exhibit the highest density of DHSs.

d. Cell line-specific DHSs are positioned predominantly within the enhancer-like state 3. The transition matrix shows the chromatin state of loci containing DHSs in one cell line (x-axis), and the state of the same locus in the other cell line where the DHS is absent (y-axis). Most of the DHSs that differ between cell lines originate from state 3. When DHSs are absent, the loci typically transition to an open chromatin state 4 (43%), or maintain state 3 (23%). In both scenarios, most of the associated genes remain transcriptionally active (see Supp. Figure 34).

e. Low levels of engaged RNA polymerase are associated with TSS-distal DHSs. The top plot shows the local increase in the antisense GRO-seq signal for DHSs located within transcribed genes; dashed lines show median levels. Intergenic DHS positions (bottom plot) also show bi-directional GRO-seq signal of comparable magnitude. See Supp. Figures 29,27,30.

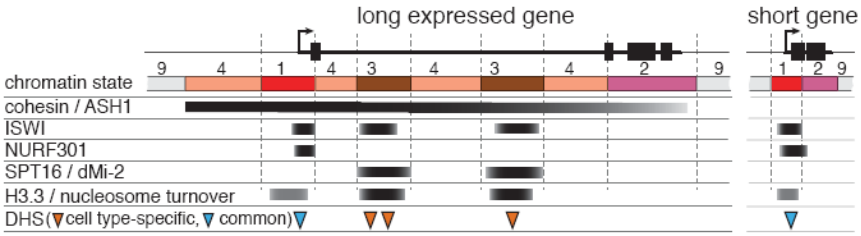

Figure 6. Spatial arrangements of chromatin states associated with active transcription

Unlike short or exon-rich expressed genes, expressed genes with long intronic regions commonly contain one or more regions of enhancer-like state 3, associated with specific chromosomal proteins, high nucleosome turnover and DHSs displaying cell-type plasticity.