Reading the maps: Organization and function of chromatin types in Drosophila

Braunschweig, U.

Citation for published version (APA):
Chapter 4

Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells

Guillaume J. Filion1*, Joke G. van Bemmel1*, Ulrich Braunschweig1*, Wendy Talhout1, Jop Kind1, Wim Brugman2, Ines de Castro Genebra de Jesus1,3, Ron M. Kerkhoven2, Bas van Steensel1,5

*Accepted for publication in Cell*

---

1* Equal contribution

1Division of Gene Regulation and

2Central Microarray Facility,
Netherlands Cancer Institute, Plesmanlaan 121,
1066 CX Amsterdam, the Netherlands

3Present address: Genome Function Group,
MRC Clinical Sciences Centre, Imperial College School of Medicine, Hammersmith Hospital Campus, Du Cane Road,
London W12 0NN, United Kingdom

5Corresponding author: b.v.steensel@nki.nl
Abstract

The local protein composition of chromatin is important for the regulation of transcription and other functions, yet the diversity of chromatin composition and the distribution along chromosomes is still poorly characterized. By integrative analysis of genome-wide binding maps of 53 broadly selected chromatin components in Drosophila cells, we show that the genome is segmented into five principal chromatin types that are defined by unique, yet overlapping combinations of proteins, and form domains that can extend over >100 kb. We identify a novel silent chromatin type that covers about half of the genome and lacks classic heterochromatin markers. Furthermore, transcriptionally active euchromatin consists of two distinct types that differ in molecular organization and H3K36 methylation, and regulate distinct classes of genes. Finally, we find that the five chromatin types are ordered in a highly non-random fashion along the chromosome arms. These results provide a global view of chromatin diversity and domain organization in a metazoan cell.

Introduction

Chromatin consists of DNA and all associated proteins. The scaffold of chromatin is formed by nucleosomes, which are histone octamers in a tight complex with DNA. This scaffold serves as the docking platform for hundreds of structural and regulatory proteins. Furthermore, histones carry a variety of post-translational modifications that form recognition sites for specific proteins (Berger 2007; Rando and Chang 2009). The local composition of chromatin is a major determinant of the transcriptional activity of a gene; some chromatin proteins enhance transcription, while others have repressive effects.

Traditionally, chromatin is divided into heterochromatin and euchromatin. This division was originally based on differential staining by DNA-binding dyes, visualized by light microscopy. Extensive biochemical, molecular and genetic studies over the past two decades have indicated that a more refined classification is warranted. For example, in Drosophila at least two types of heterochromatin exist that have distinct regulatory functions and consist of different proteins.

The first type is marked by Polycomb Group (PcG) proteins and methylation of lysine 27 of histone H3 (H3K27). PcG chromatin forms large continuous domains (sometimes more than 100 kb in length) that encompass one or multiple genes; it is a repressive type of chromatin that primarily regulates genes with developmental functions (Sparmann and van Lohuizen 2006; Ringrose 2007). The second type is marked by Heterochromatin Protein 1 (HP1) and several associated proteins, combined with methylation of H3K9. This type of heterochromatin can also cover large genomic segments, particularly around centromeres. Reporter genes integrated in or near HP1 heterochromatin tend to be repressed, but paradoxically many genes that are naturally bound by HP1 are transcriptionally active (Hediger and Gasser 2006; Yasuhara and Wakimoto 2006). Direct comparison of genome-wide binding maps indicates that PcG and HP1 heterochromatin are non-overlapping (de Wit et al. 2007).

HP1 and PcG chromatin illustrate two important principles of chromatin organization: each type is marked by unique combinations of proteins, and can cover long stretches of DNA. But are there other major types of chromatin that follow these same principles? For example, is euchromatin also organized into domains with distinct protein compositions? Are there perhaps additional types of repressive chromatin that have remained unnoticed?

In order to construct a “big picture” of chromatin type diversity and domain organization, we conducted a systematic survey of 53 broadly selected chromatin proteins and four key histone modifications in Drosophila cells.
We generated genome-wide location maps of each component, providing a rich description of chromatin composition along the genome. By integrative computational analysis of this large dataset we identified, besides PcG and HP1 chromatin, three additional principal chromatin types, which are defined by unique combinations of proteins. One of these is a novel type of silent chromatin that covers ~50% of the genome. In addition, we identified two types of transcriptionally active euchromatin that are bound by different proteins and harbor distinct classes of genes.

**Results**

**Genome-wide location maps of 53 chromatin proteins**

We constructed a database of high-resolution binding profiles of 53 chromatin proteins in the *Drosophila melanogaster* cell line Kc167 (Fig. 1A; Supplementary Data File 1). This embryo-derived cell line has a gene expression signature that resembles that of embryos (Greil et al. 2003).

---

**Figure 1**

Overview of protein binding profiles and derivation of the 5-type chromatin segmentation. (A) Sample plot of all 53 DamID profiles (log$_2$ enrichment over Dam-only control). Positive values are plotted in black, negative values in grey for contrast. Below the profiles, genes on the top and bottom strand are depicted as lines with blocks indicating exons. (B) Two-dimensional projections of the data onto the first three principal components. Colored dots indicate the chromatin type of probed loci as inferred by a 5-state HMM. (C) Values of the first three principal components along the region shown in (A), with domains of the different chromatin types after segmentation by the 5-state HMM highlighted by the same colors as in (B).
For a representative cross-section of the chromatin proteome, we selected proteins from most known chromatin protein complexes, including a variety of histone-modifying enzymes, proteins that bind specific histone modifications, general transcription machinery components, nucleosome remodelers, insulator proteins, heterochromatin proteins, structural components of chromatin, and a selection of DNA binding factors (DBFs) (Supplementary Table 1).

For ~40 of these proteins, full-genome high-resolution binding maps have not previously been reported in any *Drosophila* cell type or tissue. While chromatin immunoprecipitation (ChIP) is widely used to map protein-genome interactions (Collas 2009), large-scale application of this method is hampered by the limited availability of highly specific antibodies. Moreover, at least for some chromatin proteins, ChIP results can greatly depend on the choice of crosslinking reagents (Wang et al. 2009) and can be unreliable for proteins with short residence times (Gelbart et al. 2005; Schmiedeberg et al. 2009).

We therefore used the DamID technology, which does not require crosslinking or antibodies. With DamID, DNA adenine methyltransferase (Dam) fused to a chromatin protein of interest deposits a stable adenine-methylation ‘footprint’ *in vivo* at the interaction sites of the chromatin protein (van Steensel et al. 2001; Greil et al. 2006), so that even transient interactions may be detected (Wolffe and Leblanc 2000; Ringrose and Paro 2007). Note that the fusion protein is expressed at very low levels, averting overexpression artifacts. The DamID profiles of all 53 proteins were generated in duplicate under standardized conditions and detected using oligonucleotide microarrays that query the entire fly genome at ~300 bp intervals.

Comparisons to published and new ChIP data confirm the overall reliability of the DamID data (Supplementary Figure 1A: ChIP-DamID comparisons; associated with Fig 1), which was also reported in previous comparative studies (Moorman et al. 2006; Negre et al. 2006).

For reference purposes, we also generated ChIP maps of histone H3 and the histone marks H3K4me2, H3K9me2, H3K27me3 and H3K79me3 on the same array platform.

**Most of the fly genome interacts with non-histone chromatin proteins**

Comparison of the DamID profiles for all 53 proteins shows a variety of binding patterns (Fig. 1A). Nevertheless, several sets of proteins exhibit profiles that are locally or globally highly similar. Some similarities were anticipated, such as for PC, PCL, SCE and E(Z), which are all PcG proteins (Sparmann and van Lohuizen 2006); and for HP1, SU(VAR)3-9, LHR and HP6, which are part of classic HP1-type heterochromatin (Greil et al. 2007).

We also observe extensive colocalization of Lamin (LAM), histone H1 (H1), Effete (EFF), Suppressor of Underreplication (SUUR) and the AT-hook protein D1, which have not been linked previously except for LAM and SUUR (Pindyurin et al. 2007). Prominent is the overlap in the binding patterns of a large set of approximately 30 proteins including histone modifying enzymes (e.g. RPD3 and SIR2), components of the basal transcription machinery (e.g., CDK7, TBP), and others detailed below.

In order to identify target and non-target loci for each protein, we applied a 2-state Hidden Markov Model (HMM) to each individual binding map (Supplementary Data File 2; Supplementary Methods). This method identifies the most likely segmentation into “bound” and “unbound” probed loci. According to the resulting binary classifications, the genome-wide occupancy by individual proteins varies broadly, ranging from about 2% (GRO) to 79% (IAL).

Interestingly, 99.99% of the probed loci is bound by at least one protein, and 99.6% by at least three proteins. This indicates that, at least at the resolution of our maps, essentially no part of the fly genome is permanently in a configuration that consists of nucleosomes only. Approximately 1% of the genome shows extremely high protein occupancy, being bound by 36 to 44 of the 53 mapped proteins. This observed overlap does not necessarily derive from the simultaneous presence of all overlapping proteins; it indicates that all overlapping proteins can at least occupy identical sites.
Principal chromatin types defined by combinations of proteins

Next, we used a computational classification strategy to identify the major types of chromatin, defined as distinct combinations of proteins that are recurrent throughout the genome. To identify such combinations, we initially performed Principal Component Analysis on the 53 quantitative DamID profiles to reduce the dimensionality of the data.

We then focused on the first three principal components, which together account for 57.7% of the total variance. By projecting the genomic sites on the principal components, we could distinguish five distinct lobes in the three-dimensional scatter plot (Fig 1B). No additional distinct lobes could be observed upon further inspection of higher-level principal components. Importantly, the five groups were also clearly separated when using the previously defined binary target definitions (Supplementary Figure 1B), showing that this result is robust to different quantification methods. We therefore concluded that five major types of chromatin can be distinguished, and that further sub-classification is not supported by the data structure.

Having established that classification into five types properly summarizes the data, we fitted a 5-state HMM onto the first three principal components. Thus, every probed sequence in the genome was assigned one of five exclusive chromatin types (Supplementary Methods).

To avoid semantic confusion, and in line with the Greek word *chroma* (which means color), we labeled each of the five protein signatures with a color (BLUE, GREEN, BLACK, RED and YELLOW).

The HMM classification produced a mosaic pattern of chromosomal domains that vary widely in length (Fig 1C). We emphasize that this segmentation is purely data-driven, without using any other knowledge besides the 53 DamID profiles.

The segmentation is generally robust: removal of any of the proteins except for PC still yields a 5-state classification that is on average 96.7% identical to the model obtained with all 53 proteins. A detailed analysis of the robustness is summarized in Supplementary Figure 1C.

Domain organization of chromatin types

The five types of chromatin differ substantially in their genome coverage, numbers of domains, and numbers of genes (Fig 2A). We identified a total of 8,428 domains that typically range from ~1 to 52 kb (5th-95th percentiles) with a median length of 6.5 kb, although the size distribution depends on the chromatin type (Fig 2B). 441 domains are larger than 50 kb, and 155 are larger than 100 kb, with the largest domain being 737 kb. Many individual domains include multiple neighboring genes (Fig 2C); the largest number of which within a single domain is 139 (for a centromere-proximal GREEN domain). Taken together, these data indicate that the fly genome is generally organized into large regions that are covered by specific combinations of proteins.

BLUE and GREEN chromatin correspond to known heterochromatin types

Visualization of the protein occupancy in each of the five chromatin types (Fig 3A) shows that most proteins are not confined to a single chromatin type. Rather, the five chromatin types are defined by unique combinations of proteins. Importantly, BLUE and GREEN chromatin closely resemble previously identified chromatin types. GREEN chromatin corresponds to classic heterochromatin that is marked by SU(VAR)3-9, HP1, and the HP1-interacting proteins LHR and HP6. As described previously (Ebert et al. 2006; Greil et al. 2007), this type of chromatin is prominent in pericentric regions and on chromosome 4 (Supplementary Figure 2A, associated with Fig 3).

To further validate this classification, we conducted genome-wide ChIP of H3K9me2, a histone mark that is predominantly generated by SU(VAR)3-9 and bound by HP1 (Schotta et al. 2003). Indeed, H3K9me2 is highly and specifically enriched in GREEN chromatin (Fig 3B).

BLUE chromatin corresponds to PcG chromatin as shown by the extensive binding by the PcG proteins PC, E(Z), PCL and SCE. Indeed, well-known PcG target loci such as the Hox gene clusters are localized in BLUE domains (Supplementary Figure 2B). Furthermore, genome-wide ChIP of H3K27me3, the histone mark that is generated by E(Z) and recognized by PC (Sparmann and
van Lohuizen 2006) is highly enriched in BLUE chromatin (Fig 3B). We emphasize that these histone modification profiles serve as independent validation because they were not used in the 5-state HMM classification. The fact that two major well-known chromatin types were faithfully recovered indicates that our chromatin classification strategy is biologically meaningful.

Interestingly, we identified several additional proteins that mark BLUE or GREEN chromatin, or both. For example, moderate degrees of occupancy of the histone deacetylase (HDAC) RPD3 occur in both BLUE and GREEN chromatin, in accordance with known biochemical and genetic interactions of RPD3 with PcG proteins as well as SU(VAR)3-9 (Czermin et al. 2001; Tie et al. 2003).

The presence of EFF in BLUE chromatin is consistent with a reported role of this protein in PcG-mediated silencing (Fauvarque et al. 2001). Furthermore, LAM interacts extensively with

**Figure 2**
Characteristics of the five chromatin types. (A) Coverage and gene content of chromatin domains of each type. The chromatin type of a gene is defined as the chromatin type at its transcription start site (TSS). Grey sectors correspond to genes whose TSS maps at the transition between two chromatin types. Silent genes have an average RNA tag count below 1 per million total tags (see (D)). (B) Length distribution of chromatin domains, i.e. genomic segments covered contiguously by one chromatin type. (C) Distribution of the number of genes per chromatin domain. Because some genes overlap with more than one domain, genes are assigned to a chromatin type based on the type at the transcription start site. (D) Histogram of mRNA expression determined by RNA tag profiling. Data are represented as log_{10} (tags per million total tags). Dashed vertical lines in (B)-(D) indicate medians.
BLUE chromatin (consistent with the reported overlap of Lamin binding and H3K27me2 in human cells (Guelen et al. 2008)), but not with GREEN chromatin.

BLACK chromatin is the prevalent type of silent chromatin

BLACK chromatin covers 48% of the probed genome and is thus by far the most abundant type (Fig 2A). With a median size of 17 kb and with 134 domains larger than 100 kb, BLACK chromatin domains tend to be longer than domains of the four other types (Fig 2B). BLACK chromatin is overall relatively gene-poor (Fig 2A; compare genome coverage and number of genes), but it nevertheless harbors 4,162 genes.

By mRNA high-throughput sequencing we detected no transcriptional activity (<1 mRNA molecule per 10 million) for 66% of the genes in BLACK chromatin, while the remaining 34% have very low
activity (Fig 2D). This is in agreement with the low coverage of BLACK chromatin by RPII18, a subunit shared by all three RNA polymerases (Fig 3A) and a lack of the active histone marks H3K4me2 and H3K79me3 as detected by ChIP (Fig 3B).

We note that the majority of silent genes in the genome is located in BLACK chromatin (Fig 2A). Thus, BLACK chromatin is a distinctively silent type of chromatin that covers a large part of the genome.

BLACK chromatin is almost universally marked by four of the 53 mapped proteins: histone H1, D1, IAL and SUUR, while SU(HW), LAM and EFF are also frequently present (Fig 3A). We analyzed the fine distribution of these proteins in more detail. Close-up views show that H1, D1, IAL, SUUR and LAM have a broad distribution within BLACK domains, while SU(HW) exhibits a distinct, more focal pattern (Fig 4A).

Averaged profiles centered around the 5’ ends of genes show that within BLACK chromatin these proteins display an overall enrichment without a clear preference for upstream regions, promoters, or transcription units (Fig 4B). In the combined other chromatin types these proteins are depleted from promoter regions and to a lesser extent from gene bodies.

This is less pronounced for IAL, which may be related to the fact that this protein binds primarily to mitotic chromosomes (Giet and Glover 2001). In comparison, TBP, a protein that is not enriched in BLACK chromatin, is depleted at genes within BLACK chromatin and exhibits local enrichment at promoter regions (Fig 4B). Together, these results indicate that BLACK domains consist of relatively homogeneous stretches of chromatin bound by H1, D1, LAM, IAL, SUUR and EFF, interspersed by focal sites occupied by SU(HW).

Given that genes in BLACK chromatin in embryonic Kc167 cells are expressed at very low levels, we investigated whether these genes are perhaps activated later in development. Indeed, a survey of expression profiling data (Stolc et al. 2004) indicates that most genes in BLACK chromatin become active at a later stage (Fig 4C). This suggests that BLACK chromatin is under strong developmental control.

Consistent with this notion, we found that BLACK chromatin is particularly rich in highly conserved non-coding elements (HCNEs) (Fig 4D), which are thought to mediate gene regulation (Engstrom et al. 2007). The density of HCNEs in BLACK chromatin is comparable to that in BLUE chromatin, which harbors many developmentally regulated genes (Tolhuis et al. 2006), and is much higher than in the other three chromatin types. Together, this suggests that BLACK chromatin is under strong developmental control.

YELLOW and RED chromatin are two distinct types of euchromatin

In contrast to BLACK and BLUE chromatin, RED and YELLOW chromatin have hallmarks of transcriptionally active euchromatin: Most genes in these two chromatin types produce substantial amounts of mRNA (Fig 2D), and levels of RNA polymerase (Fig 3A), H3K4me2 and H3K79me3 are typically high, whereas levels of H3K9me2 and H3K27me3 are low (Fig 3B).

RED and YELLOW chromatin share various chromatin proteins (Fig 3A). Among these are the HDACs RPD3 and SIR2, as well as the RPD3-interacting protein SIN3A. HDACs have recently also been found in transcriptionally active chromatin in human cells (Wang et al. 2009).

Other proteins that are highly abundant in both RED and YELLOW chromatin include DF31, a little-studied protein that drives chromatin decondensation in vitro (Crevel et al. 2001); ASH2, a homolog of a subunit of a H3K4 methyltransferase complex in yeast and vertebrate cells (Nagy et al. 2002) and MAX, a DBF that is part of the MYC network of regulators of growth and proliferation (Orian et al. 2003).

Besides these similarities, RED and YELLOW chromatin display striking differences. RED chromatin is abundantly marked by several proteins that are mostly absent from the four other chromatin types (Fig 3A). Among these are the nucleosome remodeling protein Brahma (BRM); the regulator of chromosome structure SU(VAR)2-10; the Mediator subunit MED31; the 55 kDa subunit of CAF1, which participates in various histone-modifying complexes (Martinez-Balbas et al. 1998; Tie et al. 2001); and several DBFs including the ecdysone receptor (ECR), GAGA factor (GAF), and Jun-related antigen (JRA).
Figure 4

BLACK chromatin is pervasively bound by distinctive proteins and contains silent, but regulatable genes. (A) Sample plots of binding profiles of the six proteins that are the most prevalent in BLACK chromatin. Genes on both strands as well as chromatin types are depicted below the profiles. (B) Average binding of the 6 proteins present in BLACK chromatin and TBP around transcription start sites of genes lying in BLACK chromatin (black line) and other genes (grey line). Only genes lying entirely within one type have been considered. (C) Average expression of BLACK genes (black line) and other genes (grey) during fly development (Stolc et al. 2004). For each time point the values are normalized to the mean of all genes. (D) Density of highly conserved non-coding elements (HCNEs) per chromatin type.

These differences in protein composition prompted us to investigate the timing of DNA replication during S-phase, which is known to differ between chromatin types (Gilbert 2002). Analysis of a genome-wide replication timing map from Kc167 cells (Schwaiger et al. 2009) shows that DNA in RED and YELLOW chromatin is generally replicated early in S-phase, as may be expected for euchromatin. However, RED chromatin tends to be replicated even earlier than YELLOW chromatin (Fig 5A).
This coincides with a strong enrichment of origin recognition complex (ORC) binding in RED chromatin as mapped by ChIP (MacAlpine et al. 2010) (Fig 5B), suggesting that DNA replication is often initiated in RED chromatin.

These observations further underscore that RED and YELLOW chromatin are distinct types of euchromatin.
Active genes in YELLOW but not RED chromatin carry H3K36me3

Only one protein of the dataset is abundant in YELLOW but not in RED chromatin: MRG15, which is a chromodomain-containing protein (Leung et al. 2001). Because human MRG15 has previously been reported to bind H3K36me3 (Zhang et al. 2006), we compared the fine distribution of MRG15 and H3K36me3 along genes within the two chromatin types (Bell et al. 2010). Indeed, both are highly enriched along genes in YELLOW chromatin, but nearly absent from RED chromatin (Fig 5C,D).

These data are consistent with binding of MRG15 to H3K36me3 in vivo. Interestingly, H3K36me3 was previously thought to be a universal marker of elongating transcription units (Lee and Shilatifard 2007; Rando and Chang 2009). Our analysis reveals that, at least in Drosophila Kc167 cells, this histone mark is mostly absent from genes lying in RED chromatin, even though these genes are expressed at similar levels as genes in YELLOW chromatin (Fig 2D).

RED and YELLOW chromatin mark different types of genes

The substantial differences between RED and YELLOW chromatin suggested that the genes they harbor may be regulated by two globally distinct pathways. We therefore investigated whether genes located in RED and YELLOW chromatin have different characteristics. We began by comparing the embryonic tissue expression patterns of genes in the two chromatin types. Strikingly, genes with a broad expression pattern over many embryonic stages and tissues (Tomancak et al. 2007) are highly enriched in YELLOW chromatin, while genes with more restricted expression patterns are depleted (Fig 6A).

Consistently, Gene Ontology (GO) analysis revealed that annotation terms related to universal cellular functions such as “ribosome”, “DNA repair” and “nucleic acid metabolic process” are almost exclusively found in YELLOW chromatin (Fig 6B), while genes in RED chromatin are linked to more specific processes such as “receptor binding”, “defense response”, “transcription factor activity” and “signal transduction” (Fig 6C). Taken together, these results indicate that genes in YELLOW chromatin tend to be expressed in most cell types, while the expression of genes in RED chromatin tends to be restricted to certain cell types only.

Features of RED chromatin point to complex gene regulation

Several other properties of RED chromatin suggest extensive gene regulation in these regions. First, intergenic regions in RED domains contain about two-fold more HCNEs than YELLOW chromatin (Fig 4D), although not as much as BLACK and BLUE chromatin. Furthermore, genome-wide formaldehyde-assisted identification of regulatory elements (FAIRE) (Giresi et al. 2007; Braunschweig et al. 2009) points to a high density of regulatory chromatin complexes in RED chromatin (Fig 6D).

We find that individual RED chromatin loci can be locally occupied by as many as 44 of the 53 tested proteins, with an overall median occupancy of 30 proteins (Supplementary Figure 3, associated with Fig 6). These proteins belong to a wide variety of functional and structural categories (e.g., homeobox, leucine zipper, chromodomain, bromodomain, helix-loop-helix, zinc fingers, BTB/POZ proteins). In contrast, YELLOW chromatin has a less complex composition with a median occupancy of 12 proteins.

We considered that the high occupancy in RED chromatin is due to intrinsic properties such as ‘openness’ or nucleosome remodeling activity that would locally facilitate protein binding. Alternatively, proteins might be individually targeted to RED regions through protein-protein interactions or other specific mechanisms.

We reasoned that specific targeting mechanisms should not impact the binding of a non Drosophila DBF to the genome; accordingly, enrichment in RED chromatin should be observed only under the first hypothesis. To test this, we generated a DamID profile for the DNA-binding domain (DBD) of yeast Gal4. The recognition motif for this DBD is present at roughly equal densities among the five chromatin types (data not shown). We observed no enrichment of Gal4-DBD in RED chromatin (Figure 6E). This suggests that RED chromatin does not intrinsically promote protein-DNA interactions, and that high protein occupancy is more likely due to specific targeting mechanisms.
Global rules of domain organization.

Finally, to obtain more insight into the relationships between the different chromatin types, we asked whether the chromatin domains are ordered in a particular pattern along the chromosome arms. Specifically, we investigated whether certain chromatin types have preferential adjacent types. We calculated for each possible pair of types how often they are directly adjacent, and compared this to the frequency that may be expected by random chance. The resulting observed/expected ratios reveal a remarkable non-random ordering of chromatin types (Figure 7). In particular, several combinations are strongly underrepresented as neighbors, such as BLUE/GREEN, RED/GREEN, RED/BLACK, and to a lesser extent YELLOW/BLUE. This suggests that these pairs of chromatin types are not compatible as neighbors. Other pairs show significant preferential adjacency, such as YELLOW/GREEN, BLUE/BLACK and BLUE/RED. Thus, the linear ordering of chromatin domains along chromosomes is highly non-random.
Discussion

Here, we report detailed genome-wide binding maps of a broad set of 53 chromatin components, as well as four histone modifications. Systematic and unbiased integration of the 53 protein maps indicates that the Drosophila genome is packaged into a mosaic of five principal chromatin types, each defined by a unique combination of proteins. Extensive evidence demonstrates that the five types differ in a wide range of characteristics besides protein composition, such as biochemical properties, transcriptional activity, histone modifications, replication timing, as well as sequence properties and functions of the embedded genes. This validates our classification by independent means and provides important insights into the functional properties of the five chromatin types.

The number of chromatin states

We emphasize that the five chromatin types should be regarded as the major types. Some may be further divided into sub-types, depending on how fine-grained one wishes the classification to be. For example, within each of the transcriptionally active chromatin types, promoters and 3’ ends of genes exhibit (mostly quantitative) differences in their protein composition (data not shown) and thus could be regarded as distinct sub-types. However, these local differences are minor relative to the differences between the five principal types that we describe here. The five types are robust to the choice of mapped proteins (Supplementary Fig 1C), and establish distinctions that were not clearly made previously: First, BLACK chromatin is a silent compartment that is very different from PcG (BLUE) and HP1 (GREEN) chromatin; second, YELLOW and RED chromatin are two distinct types of euchromatin.
We cannot exclude that the accumulation of binding profiles of additional proteins would reveal other novel chromatin types.

We anticipate that the pattern of chromatin types along the genome will vary between cell types. For example, many of the inactive genes that are embedded in BLACK chromatin in embryonic Kc167 cells, are active in larvae and adult flies (Fig 4C). Thus, the chromatin of these genes is likely to switch to an active type during development.

While the integration of data for 53 proteins provides substantial robustness to the classification of chromatin along the genome, a subset of only five marker proteins (histone H1, PC, HP1, MRG15 and BRM), which together occupy 97.6% of the genome, can recapitulate this classification with 85.5% agreement (Supplementary Fig 1D). Assuming that no unknown additional principal chromatin types exist in some cell types, DamID or ChIP of this small set of markers may thus provide an efficient means to examine the distribution of the five chromatin types in various cells and tissues, with acceptable accuracy. The correspondence between the chromatin states and histone modifications (Fig 3, 5) suggests that one would obtain an approximate picture by mapping H3K9me2, H3K27me3, H3K36me3 and H3K4me2, but this set lacks exclusive markers for RED and BLACK chromatin.

BLACK chromatin: a novel type of silent chromatin

About half of the *Drosophila* genome is packaged by BLACK chromatin, which consists of a previously unknown combination of proteins. Essentially all genes in BLACK chromatin exhibit extremely low expression levels, indicating that BLACK chromatin either completely lacks activating factors or constitutes a strongly repressive environment.

The abundant presence of LAM and histone H1, two proteins previously linked to chromatin compaction and gene repression (Pickersgill et al. 2006; Reddy et al. 2008; Sancho et al. 2008; Braunschweig et al. 2009) supports the latter notion. Importantly, BLACK chromatin is strongly depleted of PcG proteins, HP1, SU(VAR)3-9 and associated proteins, indicating that BLACK chromatin is different from previously characterized types of heterochromatin (which we identified de novo here as BLUE and GREEN chromatin).

The proteins that mark BLACK domains provide important clues to the molecular biology of this type of chromatin. Three of these proteins are essential: loss of LAM, EFF or histone H1 causes lethality during *Drosophila* development (Cenci et al. 1997; Lenz-Bohme et al. 1997; Lu et al. 2009), suggesting an important role for BLACK chromatin. The enrichment of LAM points to a role of the nuclear lamina in gene regulation in BLACK chromatin (Pickersgill et al. 2006).

Indeed, loci that interact with LAM were previously shown to be preferentially located near the nuclear lamina (Pickersgill et al. 2006; Shevelov et al. 2009). Furthermore, depletion of LAM causes derepression of several LAM-associated genes (Shevelov et al. 2009). Recent evidence indicates that the insulator protein SU(HW) is a regulator of LAM - genome interactions (J.v.B., U.B, B.v.S, manuscript submitted).

D1 is a little-studied protein with 11 AT-hook domains that may contribute to its targeting to the relatively AT-rich BLACK chromatin (data not shown). Interestingly, overexpression of D1 causes ectopic pairing of intercalary heterochromatin (Smith and Weiler 2010), suggesting a role in the regulation of higher-order chromatin structure.

SUUR specifically regulates late replication on polytene chromosomes (Zhimulev et al. 2003) EFF is highly similar to the yeast and mammalian ubiquitin ligase Ubc4 that mediates ubiquitination of histone H3 (Liu et al. 2005; Singh et al. 2009), raising the possibility that nucleosomes in BLACK chromatin may carry specific ubiquitin marks. These insights provide important leads for further study of this previously unknown yet prevalent type of silent chromatin.

RED and YELLOW: distinct types of euchromatin

In RED and YELLOW chromatin most genes are active, and the overall expression levels are similar between these two chromatin types. However, RED and YELLOW chromatin differ in many respects. One of the conspicuous distinctions is the disparate levels of H3K36me3 at active transcription units.

This histone mark is thought to be laid down in the course of transcription elongation and may block the activity of cryptic promoters inside the transcription unit (Li et al. 2007). The absence of
H3K36me3 from genes in RED chromatin is all the more surprising that these genes are on average 2.4 times larger than genes in YELLOW chromatin (data not shown). Why active genes in RED chromatin lack H3K36me3 remains to be elucidated.

The genomic organization of the YELLOW and RED chromatin types is also different. YELLOW domains typically contain a cluster of genes (Fig 2C) whereas most RED domains contain rarely more than two genes.

YELLOW chromatin is highly enriched in genes with a broad expression pattern, i.e., constitutively expressed genes. Possibly, the aggregation of these genes into domains of YELLOW chromatin helps to ensure stable expression of these genes. A similar domain organization of chromatin may explain the reported clustering of housekeeping genes in the human genome (Lercher et al. 2002).

A striking feature of RED chromatin is the large number of proteins that is present, suggesting that RED chromatin domains are “hubs” of regulatory activity. This is likely to be related to the predominantly tissue-specific expression of genes in RED chromatin, which presumably requires an intricate interplay of many regulatory proteins. Five lines of evidence led us to reject the possibility that the high protein occupancy in RED chromatin may originate from an artifact of DamID, e.g. caused by a high accessibility of RED chromatin. First, all DamID data are corrected for accessibility using parallel Dam-only measurements. Second, several proteins, such as EFF, SU(VAR)3-9 and histone H1 exhibit lower occupancies in RED than in any other chromatin type. Third, FAIRE, which is an independent technique to identify active regulatory elements, reports an enrichment exclusively in RED chromatin. Fourth, the ORC binding data also show a specific enrichment in RED chromatin, even though these data were acquired by ChIP, by another laboratory and on another detection platform (MacAlpine et al. 2010). Fifth, DamID of Gal4-DBD does not show any enrichment in RED chromatin.

RED chromatin, with its high occupancy of a broad variety of proteins, resembles DBF binding hotspots that were previously discovered in a smaller-scale study in Drosophila cells (Moorman et al. 2006). Discrete genomic regions targeted by many DBFs have recently also been found in mouse ES cells (Chen et al. 2008), hence it is tempting to speculate that an equivalent of RED chromatin may also exist in mammalian cells.

### Similarities between GREEN and YELLOW chromatin

GREEN chromatin corresponds to the previously well-characterized HP1-marked chromatin. Many genes in GREEN chromatin are expressed (Fig 2D). This is generally consistent with previous reports (de Wit et al. 2007; Johansson et al. 2007), despite the paradoxical fact that HP1-marked chromatin can cause position effect variegation, i.e., silencing of certain, susceptible genes when brought in close proximity of HP1-marked chromatin by chromosomal translocations or transgene insertion (Girton and Johansen 2008).

It is noteworthy that GREEN chromatin resembles YELLOW chromatin in several aspects. Aside from active transcription, both types of chromatin share the presence of MRG15. Furthermore, both types lack RED-specific proteins such as BRM, SU(VAR)2-10 and MED31 (data not shown). These data indicate that the process of transcription associated with the YELLOW signature can take place in the presence of the HP1/H3K9me2 complex in GREEN chromatin, whereas RED chromatin may be incompatible with the features of GREEN chromatin.

### Non-random ordering may reflect compatibility of chromatin types

A striking observation is that the five chromatin types are ordered in a highly non-random manner along the chromosome arms (Figure 7). Some pairs of types are rarely found as neighbors, while other pairs appear neutral or even show a slight preference to be neighbors. Interestingly, the neighbor preferences show striking parallels with the overall similarities in protein composition between chromatin types. For example, GREEN and YELLOW chromatin domains, which have overlapping protein compositions (as discussed above) are also preferential neighbors. Likewise, BLACK and BLUE chromatin domains share several proteins (Figure 3A) and are frequently adjacent to each other. Finally, the preferred adjacency of RED and BLUE chromatin may be linked to shared proteins, because it has been observed that GAF, DSP1 and PHO (which are most abundant in RED
chromatin) also have important functions in PcG (BLUE) chromatin (Dejardin et al. 2005). Conversely, pairs of chromatin types that differ substantially in composition, such as RED/BLACK, RED/GREEN and GREEN/BLUE show strong tendencies not to be neighbors.

These observations suggest that chromosomes may have evolved to minimize the occurrence of adjoining chromatin domains with dissimilar protein compositions. Proximity of such ‘incompatible’ domains might destabilize one or both domains and result in loss of robust regulation of the embedded genes, which could manifest itself as position effect variegation. Although additional studies will be needed to further test this model, our identification of five principal types of chromatin provides a firm basis for future dissection of the roles of global chromatin organization in gene regulation.

Material & Methods

Constructs

DamID constructs used for this study are listed in Supplementary Table 1. Newly generated constructs were cloned by using TOPO cloning and GATEWAY recombination as described (Braunschweig et al. 2009) or by Cre-mediated recombination.

We constructed an acceptor vector containing the Hsp70 promoter upstream of Dam by the Creator Acceptor Vector Construction Kit (Clontech, 631618). Donor clones in pDNR-Dual vectors containing the cDNA of interest were obtained from the Drosophila Genomics Recource Center, Bloomington. DamID vectors expressing the Myc-tagged Dam fusion proteins were obtained by Cre-mediated recombination according to the Creator™ DNA Cloning Kits User Manual (Clontech PT3460-1).

Nuclear localization was checked for all Dam-fusion proteins by immuno-fluorescence microscopy with the 9E10 anti-Myc antibody (Santa Cruz Biotechnology) after heat-shock induced expression as described previously (van Steensel and Henikoff 2000). Only MNT, GRO and IAL gave weak signals in interphase nuclei but were not discarded because MNT and GRO were successfully mapped by DamID in previous studies (Orian et al. 2003; Bianchi-Frias et al. 2004) and IAL binds metaphase chromosomes (Giet and Glover 2001).

DamID and Microarrays

DamID assays were carried out as described previously (Moorman et al. 2006) with a minor modification: proteins were grouped in sets sharing the same Dam controls for hybridization purposes. For each group, 3-5 DamID assays on Dam alone were carried out in parallel, the product of which was pooled before labeling. Each profile was the result of two independent experiments between which the dye orientation was inverted to minimize dye bias effects.

Fluorescent labeling was done with Klenow polymerase according to the NimbleGen array user’s guide, version 4.0. 13 μg labeled DNA per channel, 2.4 μl alignment oligo, 24 μl NimbleGen hybridization component A and 1x hybridization buffer in a total volume of 120 μl were heated to 98°C for 5 min and hybridized in a TECAN hybridization station for 16 hrs. at 42°C. Microarrays were NimbleGen oligo arrays with 385.000 probes (Choksi et al. 2006). Slides were washed sequentially with NimbleGen wash buffers I, II and III with 0.1 mM DTT at 42°C, 25°C, and 23°C, respectively. Slides were scanned at 5 μm resolution, and raw data extracted using NimbleScan software.

The identity of the hybridized material was tracked by the presence of unique oligonucleotide spikes in each sample. Furthermore, because the Dam-fusion expression vectors are produced in Dam-positive bacteria, small amounts of the transfected plasmids are co-amplified in the methylation-specific amplification protocol. This leads to a strong signal in the open reading frame of the mapped protein, which allows us to verify the identity of the used vector from the microarray data alone.

Chromatin Immunoprecipitation

All ChIP experiments and the subsequent linear amplification reactions were performed as described previously (Kind et al. 2008). Labeling and hybridization were performed as for the DamID samples. Chromatin was immunoprecipitated using anti-H3K27me3 (07-449) and anti-H3K4me2
(07-030) from Upstate Biotechnology and anti-H3K9me2 (1220), and anti-H3 (1791) antibodies from Abcam. Histone H1 was immunoprecipitated with affinity-purified anti-H1 serum (Braunschweig et al. 2009). The H3K79me3 antibody was kindly provided by Fred van Leeuwen and has been described in (Schubeler et al. 2004). The H3K27me3 antibody has been described in (Peters et al. 2003) to be highly specific, with minor cross-reactivity for H3K27me2. The specificity of the H3K9me2 and H3K4me2 antibodies have been tested by the manufacturers and are reported to be highly specific for the respective histone modifications.

**Digital gene expression**

Total RNA was isolated from growing Kc cells using TriZOL (Invitrogen), and remaining DNA was degraded by shearing and DNaseI digestion. Poly(A) selection, reverse transcription and tag sequencing was carried out on an Illumina Solexa GAII and using the tag profiling kit with DpnII. Two RNA samples were sequenced, yielding a total of 7.4 and 9.0 million reads, respectively. Mapping was carried out by BLAST, requiring at most 2 mismatches and 11 consecutively matching bases.

Only the tags mapping to the last GATC of a transcript were counted and represented 70.3% and 69.4% of the total number of reads, respectively. Counts were normalized to the total number of reads and replicates were averaged. Gene mapping informations were taken from *D. melanogaster* FlyBase release 5.8.

**Data analysis**

Microarray data normalization and analysis were performed with R (R Development Core Team 2009). All DamID data were subjected to loess normalization.

The median correlation between independent replicate experiments was 0.72. Because Dam can only methylate adenines in the sequence GATC, genomic fragments between GATCs (typically 200-300bp in size) constitute the minimal unit of measurement. These fragments are referred to as “probed loci” in the text.

Accordingly, binding log-ratios were determined by averaging the loess-normalized M values (log ratio of the channels) of all the microarray probes mapping to the same probed locus in FlyBase release 5. The chromatin type of each probed locus was determined as described in the Supplementary Methods.

Custom R scripts were used to calculate average binding profiles around 5’ and 3’ ends of genes. Genomic locations are converted to coordinates relative to the nearest 5’ (respectively 3’) end of a gene, before applying a running median with a window covering 2% of the plotted data. To ensure that points are aligned only once, windows around the end of a gene range from the midpoint of the gene to the mid distance to the next gene.

HCNEs are defined as sequences of at least 50 bp having at least 98% identity with another species (Engstrom et al. 2007). Accordingly we mapped HCNEs by aligning the introns and intergenic regions of *D. melanogaster* (FlyBase release 5.17) to the genome of *D. mojavensis* (FlyBase release 1) by exonerate (Slater and Birney 2005).

For GO analysis, the *D. melanogaster* GO Slim terms were downloaded from [http://www.geneontology.org/GO_slims/archive_GO_slims/goslim_Drosophila.0200](http://www.geneontology.org/GO_slims/archive_GO_slims/goslim_Drosophila.0200). The obsolete terms were updated according to the data obtained from [http://www.geneontology.org/ontology/obo_format_1_2/gene_ontology_ext.obo](http://www.geneontology.org/ontology/obo_format_1_2/gene_ontology_ext.obo) (date: 09:11:2009 14:57). Gene associations were downloaded from [http://www.geneontology.org/cgi-bin/downloadGOGA.pl/gene_association.fb.gz](http://www.geneontology.org/cgi-bin/downloadGOGA.pl/gene_association.fb.gz) (CVS version 1.159).

Genes were further associated with all the terms higher in the GO hierarchy as indicated by the “is_a” and “part_of” keywords. Enrichment or depletion for GO terms were tested against the hypergeometric distribution, at alpha level 0.01 (correcting for multiple testing by the Bonferroni method).

Expected frequencies of neighborhood of domains of one type to the four other types were calculated as relative frequencies of domains of the other types. To derive significance levels, domains of one type were retained while randomly permuting type assignments of all other domains 20,000 times. p-values were estimated from the obtained frequency distributions and corrected for multiple testing by the Bonferroni method.
Hidden Markov Models

M values of biological replicates were averaged to give probe-wise M values, and values from array probes overlapping with one DpnII restriction fragment were averaged to a single score. In our experience, after loess normalization, the Student’s t represents an excellent approximation of the score distribution.

To identify the binding sites of a protein, we applied a two-state Hidden Markov Model (HMM) with emissions following a Student’s t distribution. The parameters were fitted by iteratively maximizing the expected value of the log-likelihood of the observations with a modified Baum-Welch procedure.

Instead of the expectation-maximization algorithm, we used an adaptation of the ECME algorithm that is substantially faster (Liu and Rubin 1994). In short, each cycle consists of two subcycles. In both subcycles, the Forward-Backward smoothing estimates are computed, and alternatively the means and variances or the degrees of freedom are updated.

Gaps in the array design were filled by $L/(D-1)$ virtual positions for which emissions are not available (NA), where $L$ is the gap size and $D$ the mean distance between consecutive GATC sequences. When not available, the emission probability was set to the same value for every state in order not to discriminate between them. The most likely sequence of states was computed through the Viterbi algorithm.

Definition of chromatin types

By compiling the DamID scores of the 53 profiles, the probed genome can be viewed as a cloud of points in 53-dimensional space. In this space, two DpnII restriction fragments are close to each other if they are bound by the same proteins irrespective of their genomic coordinates. In this representation recurrent protein binding signatures separate as distinct lobes of the cloud.

We performed Principal Component Analysis (PCA) to obtain the best representation of the cloud in lower dimensions. It immediately appeared that at least three principal components are required to give a qualitative representation of the data, because a distinct sub-cloud separates in the third component.

The projection on the first 3 principal components revealed 5 distinct lobes, and none of the 5 lobes forked on the fourth or higher principal components. Therefore, we reduced the data to its first 3 principal components. The total variance captured hereby was 57.8%. By comparison, the median score of variance captured by applying a two-state HMM to a single DamID profile was 39.4%, suggesting that the procedure efficiently removes technical variation from the dataset without causing strong loss of information.

Mapping of chromatin types

Fitting an HMM to the original dataset would involve estimating 7,441 parameters (5x4 transition parameters, 5x53 means, 5x53 variances, 5x1374 covariances and 1 degree of freedom). We therefore replaced the dataset with its projections on the first 3 principal components to reduce the number of parameters to 66 (5x4 transition parameters, 5x3 means, 5x3 variance, 5x3 covariances and 1 degree of freedom), representing a substantial gain in robustness.

To map the chromatin types in the fly genome, we applied our adaptation of the Baum-Welch algorithm to the three-dimensional projection of the data. The initial degree of freedom was set to 6, the initial means of the modified Baum-Welch algorithm were determined visually. The initial covariance matrices were set to the covariance matrix of the 3-dimensional cloud.

Note that the projected DamID values can be expected to be distributed as a Student’s t variable because they are weighted averages of the DamID binding scores.
The output of the modified Baum-Welch algorithm showed some sensitivity to initial parameter values, mainly to the means. For some initial values, the algorithm failed to identify 5 clearly separated states. However, there was only one fitted model where the states were clearly separated, showing that the segmentation is robust to initial conditions. A detailed description of the HMM will be provided as supplementary online files. An implementation of the procedure as an R package is available upon request.

**Data availability**

DamiID, ChIP and expression data, as well as a list of the coordinates of all identified chromatin domains are available from NCBI’s Gene Expression Omnibus, accession number GSE22069.

**Acknowledgments**

We thank Francesco Russo for help with vector cloning; Marja Nieuwland and Arno Velds for help with RNA tag sequencing; Dirk Schübeler’s laboratory for sharing H3K36 methylation data prior to publication; Luke Ward for data analysis; Reuven Agami, Fred van Leeuwen, Harmen Bussemaker, Wouter Meuleman and Ludo Pagie for helpful suggestions. Supported by an EMBO Long-term Fellowship to J.K., and grants from the Netherlands Genomics Initiative and an EURYI Award to B.v.S.
References


References


References


R Development Core Team. 2009. R: A Language and Environment for Statistical Computing. [www.r-project.org](http://www.r-project.org)


## Supplementary Material

### Supplementary Table 1: Overview of chromatin components that were used in this study

<table>
<thead>
<tr>
<th>Name</th>
<th>Flybase Protein Accession</th>
<th>Plasmid Backbone</th>
<th>Transcription Factor</th>
<th>Transcription Machinery</th>
<th>Polycistrionic Group</th>
<th>Nuclear Envelope</th>
<th>Histone Modification</th>
<th>Cofactor</th>
<th>Chromatin Assembly/Remodeling</th>
<th>Trafficking Group</th>
<th>Structural Protein</th>
<th>Insulator</th>
<th>Regulation</th>
<th>Plasmid Published</th>
<th>Profile published before</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT5C</td>
<td>FBpp0070787</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASF1</td>
<td>FBpp0074715</td>
<td>pCreDamMyc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASH2</td>
<td>FBpp0084600</td>
<td>pCreDamMyc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BCD</td>
<td>FBpp0081805</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BEAF32b</td>
<td>FBpp0086572</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BBM</td>
<td>FBpp0075279</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAF1</td>
<td>FBpp0082511</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDK7</td>
<td>FBpp0070723</td>
<td>pCreDamMyc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG31367</td>
<td>FBpp0292134</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CBP</td>
<td>not specified</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CTCF</td>
<td>FBpp0076586</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D1</td>
<td>FBpp0081668</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DSF1</td>
<td>FBpp0085273</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DSP1</td>
<td>FBpp0088319</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DWG</td>
<td>FBpp0070465</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E(Z)</td>
<td>FBpp0076008</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ECR</td>
<td>not specified</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAF</td>
<td>FBpp0086978</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRo</td>
<td>FBpp0084337</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H1</td>
<td>FBpp0082448</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HP1</td>
<td>FBpp0079251</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPTc</td>
<td>FBpp0083702</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HP6</td>
<td>FBpp0077354</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IAL</td>
<td>FBpp0079058</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IRA</td>
<td>FBpp0086979</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAM</td>
<td>FBpp0078733</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LHR</td>
<td>FBpp0086073</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOLAL</td>
<td>FBpp0085956</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAX</td>
<td>FBpp0074785</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MED31</td>
<td>FBpp0079496</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MINt</td>
<td>FBpp0075954</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRGB18</td>
<td>FBpp0082767</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MUS5209</td>
<td>FBpp0078406</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PC</td>
<td>FBpp00748059</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCAF</td>
<td>FBpp0075701</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCL</td>
<td>FBpp0085914</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHO</td>
<td>FBpp0082628</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHOL</td>
<td>FBpp0082630</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PROD</td>
<td>FBpp0067165</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPD3</td>
<td>FBpp0071673</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPD918</td>
<td>FBpp0078433</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCE</td>
<td>FBpp0084614</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIN3A</td>
<td>FBpp0087003</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIR2</td>
<td>FBpp0080015</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SUHH4</td>
<td>FBpp0072400</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SuzVAR2-12-10</td>
<td>FBpp0078785</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SuzVAR3-7</td>
<td>FBpp0082094</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SuzVAR3-9</td>
<td>FBpp0082583</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SUUR</td>
<td>FBpp0079533</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TBP</td>
<td>FBpp0071596</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TIP60</td>
<td>FBpp0070636</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOP1</td>
<td>FBpp0073822</td>
<td>pGWMycDam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a) Moorman et al., 2006  
b) van Steensel et al., 2009  
c) Bianchi-Frias et al., 2004  
d) van Bemmel et al., submitted, 2010  
e) Greil et al., 2003  
f) de Wit et al., 2008  
g) Orian et al., 2003  
h) Tolhuis et al., 2006  
i) Pindyurin et al., 2007  
j) Braunschweig et al., 2009
Supplementary Material

A

B

C

D

Robustness of the classification

% Faithful segmentations

% Overlap with current segmentation

Subsample size
Supplementary Figure 1 (associated with Fig. 1)

Validation of DamID and robustness analysis of the 5-type segmentation. (A) Comparisons of DamID (Braunschweig et al. 2009) and ChIP for histone H1 (this study), SU(HW), and CTCF (Bushey et al. 2009). All datasets were subjected to running mean smoothing with a window of three data points. Genes on the top and bottom strand are depicted as lines with blocks indicating exons. (B) The five-state classification is robust to the quantification method. DamID profiles for each protein were binarized before projection onto the first three principal components. The shape of the cloud is different from the one shown in main Figure 1B, but the five types still form separated clouds in the first three principal components. (C) Sensitivity analysis of the segmentation. The Principal Component Analysis and HMM procedure was applied as in main Figure 1 to datasets where one or more proteins were left out. It is expected that smaller sets will not always exhibit all five states; for example, a subset that lacks the four PcG proteins will not identify the BLUE state. In that case, the loci formerly assigned to BLUE will be assigned to another color by the HMM. We call such a segmentation “unfaithful”. On the contrary, “faithful” segmentations show no replacement of a color by another. The percentage of overlap between the current segmentation (based on the complete dataset) and unfaithful segmentations is irrelevant: it mostly reflects the size of the type that has been replaced. For example if BLUE has been replaced by another color at least 20% of the calls will differ (i.e. all the former BLUE calls). For subsets of 35-51 proteins, 50 samples were drawn at random without replacement. For subsets of 52 proteins, all possible 53 subsets were tested. For each subsample size, the percentage of faithful segmentations (thin lines) and their mean percentage of overlap with the current segmentation (thick lines) were determined. Vertical bars represent ± standard deviations. The plot shows that larger sets of proteins give an increasingly reliable discovery of the five states. Faithful segmentations show substantial agreement with the current definition, even for smaller sample sizes. Thus, identification of the states is sensitive to protein choice, but the classification itself is robust. (D) Minimal set of proteins defining the five types. A carefully selected set of 5 proteins (histone H1, PC, HP1, MRG15 and BRM) summarizes the segmentation in 5 types. The data points were projected on their first three principal components, showing that the types are clearly visible with this minimal set of proteins. The coloring was obtained by applying the HMM to the first three principal components, exactly as was done for the 53 proteins. The agreement with the 53 profile segmentation is 85.5%.
Supplementary Figure 2 (associated with Fig. 3)
Localization of GREEN and BLUE chromatins supports their identity with known heterochromatin types. (A) Chromosomal maps of the pericentric region of chromosome 2 and of the entire chromosome 4. GREEN chromatin domains are shown, and other types are collectively represented in grey. Constrictions symbolize centromeres. (B) Close-up view of the HOX gene clusters. The HOX genes are known to be Pc targets in Kc167 cells. Genes on the top and bottom strands are shown as boxes.
Chromatin types differ widely in their total protein occupancy. (A) Sample plot of the total occupancy (out of 53 mapped proteins) on a 1Mb segment of chromosome 2L. The height of each vertical line indicates the number of proteins bound to a locus. The color of the line indicates the local chromatin type. (B) Histograms of total occupancy distributions per chromatin type. Grey vertical dashed lines indicate median values.