CHROMATIN FLAVORS

Joke G. van Bemmel

Chromatin Flavors:
Chromatin composition and domain organization in *Drosophila melanogaster*

Joke Gerarda van Bemmel

ISBN: 978-94-6182-095-2

Layout and printing: Off Page, www.offpage.nl

Cover design: J. van Bemmel, inspired by a design from Esther Wiegert, www.sterontwerp.nl

Chromatin Flavors:
Chromatin composition and domain organization in *Drosophila melanogaster*

Chromatine smaken:
Chromatine samenstelling en domein organisatie
in *Drosophila melanogaster*

Proefschrift

ter verkrijging van de graad van doctor
aan de Erasmus Universiteit Rotterdam
op gezag van de rector magnificus
Prof.dr. H.G. Schmidt
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
woensdag 9 mei 2012 om 11:30 uur

door

Joke Gerarda van Bemmel
geboren te Heemskerk

| PROMOTIECOMMISSIE

Promotor:	Prof.dr. B. van Steensel
Overige leden:	Prof.dr. C.P. Verrijzer
	Prof.dr. J.N.J. Philipsen
	Dr. M.W.J. Fornerod

| TABLE OF CONTENTS

| LIST OF ABBREVIATIONS

BN	Bayesian Network
BNI	Bayesian Network Inference
CC	Chromatin Component
CdLS	Cornelia de Lange Syndrome
CG	Computed Gene
ChIP	Chromatin Immuno-Precipitation
DamID	DNA Adenine Methyltransferase IDentificication
DBD	DNA Binding Domain
DBF	DNA Binding Factor
EcR	Ecdysone Receptor
FAIRE	Formaldehyde Assisted Identification of Regulatory Elements
FISH	Fluorescence In Situ Hybridization
GO	Gene Ontology
HAT	Histone Methyltransferase
HCNE	Highly Conserved Non-coding Element
HDAC	Histone DeACetylase
HMM	Hidden Markov Model
HP1	Heterochromatin Protein 1
LAD	Lamina Associated Domain
LAM	Lamin
modENCODE	model organism ENCyclopedia Of DNA Elements
NER	Nucleotide Excision Repair
NL	Nuclear Lamina
ORF	Open Reading Frame
PcG	Polycomb Group
RNAi	RNA interference
SU(HW)	Supressor of Hairy Wing
TEV cleavable	Tobacco Etch Virus inducable cleavage
TSS	Transcription Start Site

1

Introduction

Joke G. van Bemmel, Bas van Steensel

History of Chromatin

Chromatin was originally identified by W. Flemming in 1882 as not much more than the stainable substance of the cell nucleus. Flemming named this substance according to the Greek word "chroma", meaning color [1]. In 1911 chromatin was characterized as proteins, named histones, that were attached to nucleic acid (DNA) [2]. In the following years it became clear that chromatin formed the structural basis of genetic information. Not until more than 30 years later the DNA, and not the histone proteins as was widely expected, was identified as the carrier of the genetic information [3]. In 1952 the role of DNA in inheritance was confirmed [4] and not much later Watson and Crick discovered the double-helical structure of DNA [5] based on the DNA crystal analyses of Franklin, Gosling and Wilkins [6-7].

The double helix provided the foundation of the central dogma of molecular biology, which describes the transfer of biological sequence information [8]. Transfer of biological information occurs at three different levels (Figure 1). First, genetic information can be transmitted from one cell to its daughter cells and from parent to progeny, by duplicating the DNA into two identical DNA molecules (replication). Second, genetic information is processed so that genes can exert their function by copying the DNA into mRNA (transcription). Third, this mRNA is used as a template to generate proteins (translation) which in turn exhibit their specific functions. In addition, some transcripts are not being translated into protein. These RNA molecules, called non-coding RNA, are involved in the regulation of translation, transcription and other cellular processes.

This dogma provides the foundation but does not explain the entire complexity of life. Just knowing the information that resides in the DNA is not enough to understand a complex multi-cellular eukaryotic organism, such as a human being. To do so it is important to comprehend how the transcription of DNA into mRNA is regulated. In other words, the function of a gene resides in its DNA sequence, but when and where it is expressed and thus exerts its function is regulated in a time and cell-type dependent manner. If we understand how the transcription of DNA into mRNA is regulated, we can answer questions like: how can the same DNA in every cell give rise to different phenotypes? How can some genes be off while others are on? How can this be different between different cell types?

We now know that histone proteins and other chromatin components are

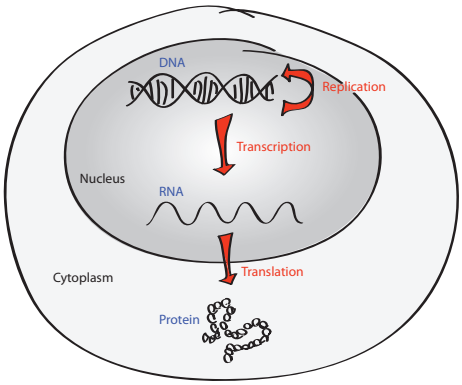


Figure 1 Central Dogma of the Molecular Biology. Replication: Genetic information is transmitted by copying the DNA to another DNA molecule. Transcription: Genetic information is processed by copying the DNA into an mRNA molecule. Translation: mRNA is used as a template to synthesize proteins. (adapted from Hurst's The Heart, V Fuster, RA Walsh, RA Harrington)

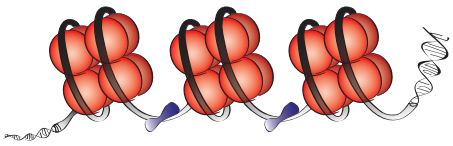


Figure 2 Nucleosome Model. ~146 basepairs of DNA (grey) is wrapped 1.7 times around an octameric core of 4 pairs of histone proteins (red). In between each nucleosome resides a linker region which can be bound by the linker histone H1 (blue).

instrumental in regulating the transcription of DNA into mRNA. However, after the discovery of the double helix, the general interest in chromatin declined. DNA was believed to harbor "all" the information and for a long time chromatin was thought to be "just" a way of packaging the large amounts of DNA (about 2m) into a small nucleus of some µm in diameter. In the 1970s this packaging of DNA by histone proteins was characterized in detail: the acidic DNA was found to be wrapped 1.7 times around an octameric core of alkaline histone proteins, forming the fundamental repeating unit of chromatin, the nucleosome [9-11]. This octameric histone core consists of 4 pairs of the histone proteins H2A, H2B, H3 and H4. In between each nucleosome resides a linker region of around ~20-50 basepairs of DNA which can be bound by linker histones, such as H1 [12-13]. Together the DNA and histone proteins form a so called "beads-on-a-string" structure (Figure 2).

While histone proteins were dismissed as not more than inert packing material, gene transcription was at that time believed to be purely regulated by transcription factors. Transcription factors or sequence-specific DNA-binding factors were found to bind a DNA sequence adjacent to the genes they regulate, in that way directly affecting transcription by either promoting or blocking the recruitment of RNA polymerase (the enzyme that performs the transcription from DNA into mRNA) [reviewed in 14] (see also the following Chromatin Protein section). In the 1980s this view drastically changed. With the discovery that histone proteins were involved in transcriptional repression [15] chromatin came into the picture again. Based on following research as well as earlier findings chromatin appeared to exist in different forms depending on its composition. In turn, the chromatin composition as well as the spatial localization of a gene was found to influence its transcriptional activity. Below these different aspects of chromatin and transcriptional activity will be further discussed.

Chromatin Types

Already in the 1930s chromatin was found to exist in two different forms. Light as well as electron microscopy studies showed differences in chromatin density, which indicated a weakly colored type of chromatin, called euchromatin, and a darkly colored type, called heterochromatin [16] (Figure 3). Because of this observation heterochromatin was believed to be more densely compacted than the lightly stained euchromatin which was thought to have a more open structure. The first link between gene expression and chromatin types came from genomic inversions in *Drosophila* which were found to cause altered gene expression, detected by stochastic changes in eye color [17]. Later, it appeared that the silencing of the eye color gene was caused by inversions which placed the gene next to heterochromatic chromatin [18, summarized in 19].

Figure 3 Euchromatin and Heterochromatin. Electron microscopy image of a plasma cell nucleus from bone marrow showing darkly stained heterochromatin and lightly stained euchromatin. (Courtesy of Prof. DL Schmucker, UCSF School of medicine, San Fransisco)

This indicated a repressive effect of heterochromatin on transcriptional activity.

At the molecular level heterochromatin could be subdivided in two distinct types. One type is marked by histone H3 lysine 9 methylation (which is one of the many histone modifications possible, and will be discussed below) [20] and binding of the non-histone proteins HP1 [21] and Su(var)3-9 [20]. This heterochromatin type is particularly found at (peri)centromeric regions. The other type is marked by histone H3 lysine 27 methylation and binding of Polycomb Group proteins [22] and is primarily found at developmentally regulated genes [for a summarizing table see 23].

For a long time, mainly based on anecdotal evidence, it was thus believed that two types of chromatin existed; active euchromatin and inactive heterochromatin. Heterochromatin in turn could be divided into two types marked by HP1 and PcG proteins. At the beginning of this century, when the genomics era started, evidence accumulated defying this concept. Genome-wide molecular mapping of HP1 showed binding at genes that were actively transcribed and contained histone modifications typical for active chromatin [reviewed in 24, 25-26]. It can thus be concluded that HP1 heterochromatin does not purely act as a uniform repressive chromatin type. Furthermore, only a limited fraction of the genome is bound by either HP1 or PcG proteins, while studies on the expression level of integrated reporter genes have suggested that most of the fly genome is transcriptionally inactive [27-29]. Hence, the concept of euchromatin and two types of heterochromatin is likely to be more complicated and additional types of repressive chromatin are likely to exist.

Histone Modifications

In the previous section we have concluded that there are several types of chromatin which relate to differences in transcriptional activity. We also recognized that different chromatin types correlate with different modifications of the histone proteins. Each of the four histone proteins consists of well structured domains that together make up the centre of the nucleosome. Their N-terminal tails however, are structurally more dynamic and protrude outwards of the nucleosome. Different amino-acid residues at these tails are prone to a diversity of chemical modifications such as acetylation, methylation, phosphorylation, ADP-ribosylation, ubiquitination

and combinations thereof (Figure 4, first panel). That the chemical modification of histones could have a role in transcription was already suggested in 1964 [30], but as mentioned above the role of histones in gene regulation became fully recognized only in the 1980s.

By now a plethora of different modifications at different residues (Table 1) has been identified as well as the enzymes responsible for placing or removing the modifications. These modifications strongly correlated with transcriptional activity and at the beginning of this century Strahl and Allis proposed the so called "histone code" in which specific combinations of histone modifications form a fundamental regulatory mechanism for gene expression by recruiting specific effector proteins or protein complexes with different functional outcomes [31]. Initially a combination of "activating" (euchromatic) or "inactivating" (heterochromatic) modifications were presented, but once genome-wide techniques became widely available and dozens of genome-wide studies were published combinatorial modifications were found specific for example for promoters versus gene bodies [32], for bivalent genes (genes that are

Table 1 The main histone modifications and their transcriptional outcome.

Modification	Histone	Residue	Transcriptional Outcome
Acetylation	H2A	K5	Activation
	H2B	K5, K12, K15, K20	Activation
	H3	K4, K14, K18, K23, K27	Activation
		K9	Histone deposition (activation)
	H4	K5, K12	Histone deposition
		K8, K16	Activation
Methylation	H3	K4	Activation, TSSs & Regulatory Elements
		K79	Activation, Euchromatin
		K9, K27	Silencing
		R17	Activation
		K36	Elongation, Repression cryptic transcription
	H4	R3	Activation
		K20	Silencing
Phosphorylation	H2A	S1, T119	Mitosis
	H2AX	S139	DNA repair
	H3	T3. S10, T11, S28	Mitosis
	H4	S1	Mitosis
Ubiquitination	H2A	K119	Silencing
	H2B	K120	Activation

K= Lysine, R=Arginine, S=Serine, T=Threonine
Information from [37-38]

inactive but are primed to become active) [33], for regulatory elements [34] and other specific regions and functions. Combinatorial histone modifications, although limited in number [reviewed in 35], can thus define multiple types of chromatin with different functions and they allow for better and more precise separation than just eu- and heterochromatin.

The "histone code" has been a popular and useful concept, however one could doubt the implied causality as discussed by Henikoff and Shilatifard in 2011 [36]. They pointed out that the enzymes that place the modifications are for example far from independent, but rely on other factors such as sequence specific transcription factors or small RNAs for their targeting to the genome. This indicates that the (combination of) histone modification(s) can describe but not necessarily determine a chromatin type and its transcriptional outcome. In addition to the causality, one could doubt if the different combinations of histone modifications describe all different chromatin types. Different proteins are known (see next section) that do affect transcriptional outcome independent of a histone modification. For example, histone modifications are only known for two types of repressive chromatin (H3K9me and H3K27me) while, as pointed out above, a third type of repressive chromatin is likely to exist. Rather than looking at histone modifications to define different chromatin types one might therefore consider the different proteins that bind the chromatin.

Chromatin Proteins

So far we have talked about chromatin as DNA wrapped around nucleosomes, consisting of histones that can be modified. We have just pointed out that (modified) histones or DNA itself can be bound by many proteins or protein complexes. Defining chromatin as DNA and all the proteins bound to it would thus be more appropriate. Many of these chromatin proteins have been identified and their function partially elucidated. These proteins can be roughly categorized in different classes, which are not strictly defined and proteins can often belong to multiple classes (Figure 4).

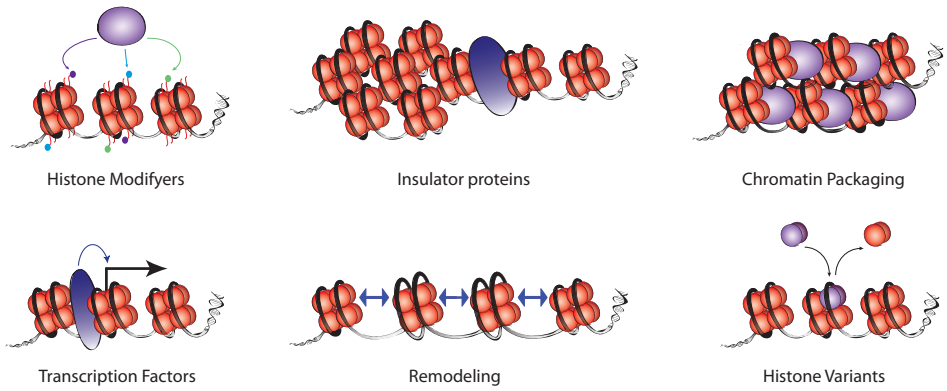


Figure 4 Chromatin proteins. Simplified graphical representations of the different classes of chromatin proteins. Nucleosomes are depicted as in Figure 3 (in the first panel histone tails are added).

Histone Modifying Enzymes and **proteins binding to modified histones** are already mentioned above. A subgroup of histone modifying enzymes are the proteins containing a SET domain, which catalyses methylation of lysine residues on histone tails. Su(var)3-9 and E(z), which both contain a SET domain, can for example specifically methylate respectively H3K9 and H3K27. The HP1 and PcG proteins which are mentioned above are examples of proteins containing a chromodomain. A chromodomain recognizes methylated lysines on a histone tail. The HP1 chromodomain specifically recognizes methylated H3K9 [39-40] while the Pc chromodomain recognizes methylated H3K27 [22].

Chromatin Remodeling Complexes are a class of protein complexes that can move nucleosomes along DNA in an ATP dependent manner. These complexes can be fairly large (up to 12 subunits) and (can) include multiple proteins that interact with either the DNA or with (modified) histones. The ISWI complex for example has been shown to bind to acetylated histones. In addition to sliding nucleosomes along the DNA some remodeling complexes can replace histones by specific histone variants. For a review on remodeling complexes in *Drosophila* see [41].

Histone Variants can specifically replace the canonical histones. The histone variant H2A.Z for example is specifically replacing the canonical H2A at transcriptionally inactive promoters [42], while the histone variant H3.3 is specifically incorporated at active genes and promoters [43] and macro-H2A is specifically present at the female inactive X chromosome [44].

Chromatin Packaging Proteins, include histone H1 but also proteins such as HP1 and Pc which are believed to package the chromatin in a more compacted structure (as discussed above).

Most of the above mentioned proteins or protein complexes are targeted to the chromatin by specific histone modifications. The following class of proteins is rather targeted by specific sequence motifs in the DNA. **Transcription Factors**, also called sequence specific DNA binding factors, are by definition proteins that contain a DNA binding domain through which they bind to a specific DNA sequence within the genome, like enhancer or promoter regions. As mentioned above, transcription factors were the first proteins identified to regulate gene expression. They can regulate the expression of their target genes by either promoting or blocking the recruitment and binding of RNA polymerase to the DNA, by catalyzing histone modifications or by recruiting other chromatin proteins which can regulate transcription. In turn the binding of transcription factors to the DNA is likely to be influenced by histone modifications and chromatin proteins present at their binding sequence.

Insulator Proteins are a specific class of sequence specific DNA binding proteins, which bind to sequence specific insulator elements. Insulator elements are DNA sequence elements that, when bound by insulator proteins, can regulate gene expression by preventing inappropriate interactions between adjacent chromatin domains. They have been shown to block the interaction between a promoter and its enhancer when placed in between [45-46, reviewed in 47]. More interesting with respect to

different chromatin types, some insulators are thought to separate inactive from active chromatin domains by acting as a barrier against spreading of repressive chromatin proteins or histone modifications into neighboring regions [reviewed in 47, 48]. In mammals only two insulator proteins are known, namely CCCTC-binding Factor (CTCF) [49] and USF1 [50], while in *Drosophila* five main insulator proteins have been identified; Suppressor of Hairy-wing (SU(HW)), CTCF, Boundary Element-associated Factor (BEAF-32), Zeste-white 5 (Zw5) -also known as Deformed Wings (DWG)- and GAGA Factor (GAF) [45, 51-55] and [reviewed in 56].

In addition one could consider proteins involved in the key processes of the cell: **Transcription, Repair** and **Replication**. During these processes the cellular machineries performing these tasks need to gain access to the DNA that is packaged into chromatin. Proteins involved in these processes are therefore influencing chromatin composition or are even being part of the chromatin.

Taken together, histone modifications as well as protein composition collectively determine different chromatin types and thereby certain transcriptional outcomes.

Spatial Organization

The classic and electron microscopy used to study chromatin did not only reveal a difference in compaction between euchromatin and heterochromatin, it also showed a striking preference for heterochromatin to localize at the most peripheral regions of the nucleus (Figure 2). This indicated that chromatin is not randomly organized within the nucleus like "spaghetti floating around in a kettle of nucleoplasmic sauce" [57] but rather non-randomly with respect to the periphery. Using fluorescence in situ hybridization (FISH) it was shown that the activity of a gene globally correlated with its radial position. Inactive genes preferentially localize at the periphery whereas active genes often reside in the interior [reviewed in 58, 59]. Genome-wide mapping using the DamID technology (see next section) in *Drosophila* Kc cells showed more than 500 genes to be in molecular contact with lamin, a protein that covers the inside of the nuclear membrane [60]. These genes are strongly repressed and depleted of active histone marks. Application of the same mapping technology at a higher resolution in human lung fibroblasts showed that nuclear-lamina (NL) interactions occur through large continuous genomic domains with sharply defined borders [61] (Figure 5). Furthermore, loss or mutation of lamins in flies and mammals can cause dissociation from the periphery and changes in gene expression [62-63] and tethering of reporter as well as endogenous genes to the lamina can in some cases -but not all- cause gene repression [64-67]. On the other hand, genome-NL interactions were found not to be directly required for gene repression in embryonic stem cells, but other membrane components could still be required [68]. Together, these data indicate a causal but maybe not essential role for this spatial organization in gene repression.

Active genes on the other hand are believed to come together in so-called transcription factories, which are foci with high concentrations of RNA polymerase II [69, reviewed in 70] (Figure 5). Transcription factories are thought to allow

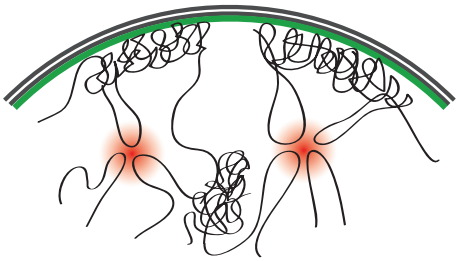


Figure 5 Spatial organization. The genome (black line) is non-randomly organized in the nucleus. Large domains of inactive DNA are localized at the periphery in association with the nuclear lamina (green) while active genes come together in transcription factories (red).

efficient and coordinated transcription, since RNA polymerase II and transcription factors can efficiently be shared by multiple genes. Movement of genes in or out of these factories correlates with respectively activation or repression of transcription [69].

The spatial organization of the genome within the nucleus is thus highly non-random and strongly correlates with gene expression. If we consider the hard coded DNA sequence as a first dimension of information, and the chromatin composition (different chromatin types) as a second, then the spatial organization of the genome is the third dimension adding to the complexity of gene expression.

Taken together, we can conclude that chromatin, defined as the DNA and all associated proteins, plays an essential role in gene regulation. Chromatin protein composition as well as spatial organization can determine the transcriptional activity of a gene. We have discussed how chromatin was traditionally divided into inactive heterochromatin and active euchromatin. However from previous research it became clear that this classification is likely to be more complex. Combinatorial histone modifications allowed a more complex and specific definition of different chromatin types and specific genomic regions. Additionally, many chromatin proteins are known that do affect transcriptional outcome independent of histone modifications. Previous studies mostly aimed to elucidate the function of a single protein or protein complex or they focused on the regulation of a single genomic locus, but a global view and understanding of the chromatin composition and its role in gene regulation was still lacking. This thesis therefore comprises systematic and genome-wide analyses of chromatin protein composition as well as spatial organization in relation to transcriptional activity.

This Thesis

In order to study the composition and function of different chromatin types we make use of a genome-wide mapping technique called DamID [71]. In short, this technique is based on the in vivo expression of a protein of interest fused to E.coli DNA adenine methyltransferase (Dam). Expression of low amounts of this fusion protein leads to preferential DNA methylation of GATCs in the vicinity of the binding site of the protein of interest (Figure 6). To identify methylated sites, these regions are selectively amplified, fluorescently labeled and hybridized to a high-resolution genome-wide tiling array with a probe spacing of ~300bp. In contrast to Chromatin IP (ChIP), the most commonly used mapping technique, DamID does not require antibodies to map the genome-wide binding sites of a protein of interest. This has the major advantage that it allows mapping of pro-

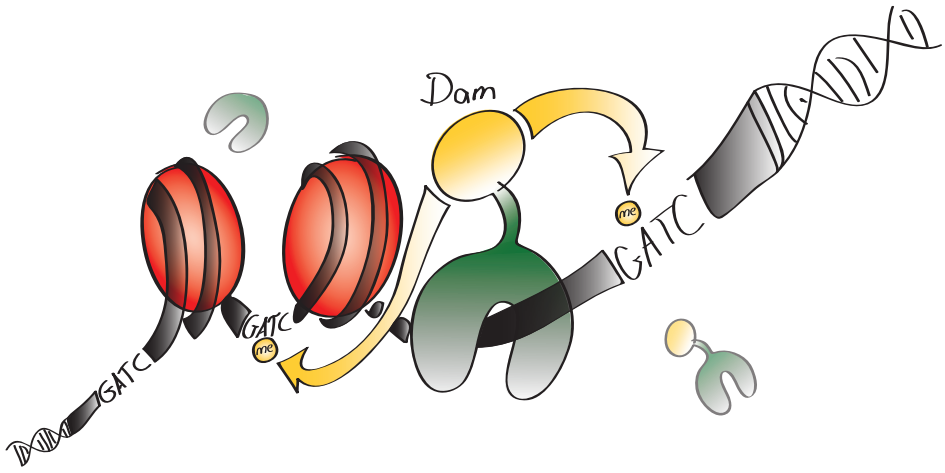


Figure 6 Principle of DamID. A protein of interest (green) fused to Dam (yellow) is expressed in vivo, resulting in adenine methylation (yellow) of GATC sequences in the vicinity of the binding site. (adapted from G. Filion)

teins for which no antibody is available, and it allows mapping in a semi high-throughput manner without the need to optimize and quality assess each antibody.

As a model system, we used *Drosophila melanogaster*. The *Drosophila* genome is significantly smaller than mammalian genomes which provides a better signal to background-noise ratio and allows genome-wide mapping of protein binding sites cost-effectively with a much higher resolution. Furthermore, the complexity of chromatin proteins and complexes in *Drosophila* is lower and has less functional redundancy, which allows mapping of a broad variety of chromatin proteins while covering all functional classes.

In the following chapters we used DamID in *Drosophila* to address some of the issues raised in this introduction.

Chapter 2 is focused on the spatial organization of the genome by identifying which regions of the genome associate with the NL. By doing so, we provide insights into the evolution of spatial organization. Several lines of evidence point towards the existence of a mechanism that regulates genome-NL interactions (discussed in the introduction of Chapter 2). Since insulator proteins are believed to play a key role in chromatin organization we examined the putative roles of five insulator proteins in genome-NL association. This revealed SU(HW) to bind at the border and inside lamina associated domains, thereby fine-tuning the genome-NL interactions.

Chromatin protein composition can affect transcriptional outcome independent of histone modifications. The binding pattern and effect on gene regulation of one such protein complex, Cohesin, is reported in Chapter 3. Cohesin's canonical function is mediating sister chromatid cohesion but it has been suggested to be involved in gene regulation. In a collaboration with Kim Nasmyth's lab we have found Cohesin to bind and regulate

the expression of a distinct set of genes, including those mediating the ecdysone response.

In contrast to Chapter 2 and 3 where we focused on individual chromatin proteins, we aimed in Chapter 4 to generate a global view of the protein composition and function of different chromatin types. We generated a collection of 53 binding maps, covering different classes of chromatin proteins. By integrative analysis of these binding maps we identified 5 distinct chromatin types, which are defined by unique combinations of proteins. Besides the well known Polycomb and HP1 chromatin we found active chromatin (previously called euchromatin) to be divided into two functionally different types. We also found a novel repressive type which covers the majority of all inactive genes.

The molecular architecture of these chromatin types is however still largely unknown and a large fraction of the chromatin proteome is still completely uncharacterized. In Chapter 5 we therefore used a systematic approach to complete our understanding of chromatin composition and organization. We identified 42 novel chromatin proteins and generated a functional network model together with the known proteins from Chapter 4. This network confirms the robustness of the 5 chromatin types, provides insight into the molecular architecture of the chromatin network and offers functional predictions for the novel proteins.

Taken together, the work presented in this thesis provides new insights in the composition as well as the function of chromatin in *Drosophila melanogaster*.

ACKNOWLEDGEMENTS

Many thanks to Wouter Meuleman and Piet Borst for their critical reading and their useful suggestions to improve this chapter.

REFERENCES

1. Flemming, W., Zellsubstanz, Kern und Zelltheilung. F.C.W. Vogel, Leipzig, 1882.
2. Kossel, A., Ueber die chemische beschaffenheit des zellkerns. Munchen Med. Wochenschrift, 1911. 58(65-69).
3. Avery, O.T., C.M. Macleod, and M. McCarty, Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii. J Exp Med, 1944. 79(2): p. 137-58.
4. Hershey, A.D. and M. Chase, Independent functions of viral protein and nucleic acid in growth of bacteriophage. J Gen Physiol, 1952. 36(1): p. 39-56.
5. Watson, J.D. and F.H. Crick, Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature, 1953. 171(4356): p. 737-8.
6. Wilkins, M.H., A.R. Stokes, and H.R. Wilson, Molecular structure of deoxypentose nucleic acids. Nature, 1953. 171(4356): p. 738-40.
7. Franklin, R.E. and R.G. Gosling, Molecular configuration in sodium thymonucleate. Nature, 1953. 171(4356): p. 740-1.
8. Crick, F., Ideas on Protein Synthesis. Symp. Soc. Exp. Biol, 1956.
9. Kornberg, R.D. and J.O. Thomas, Chromatin structure; oligomers of the histones. Science, 1974. 184(139): p. 865-8.
10. Van Holde, K.E., et al., DNA-histone interactions in nucleosomes. Biophys J, 1980. 32(1): p. 271-82.

11. Luger, K., et al., Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature, 1997. 389(6648): p. 251-60.
12. Simpson, R.T., Structure of the chromatosome, a chromatin particle containing 160 base pairs of DNA and all the histones. Biochemistry, 1978. 17(25): p. 5524-31.
13. Travers, A., The location of the linker histone on the nucleosome. Trends Biochem Sci, 1999. 24(1): p. 4-7.
14. Roeder, R.G., The role of general initiation factors in transcription by RNA polymerase II. Trends Biochem Sci, 1996. 21(9): p. 327-35.
15. Kayne, P.S., et al., Extremely conserved histone H4 N terminus is dispensable for growth but essential for repressing the silent mating loci in yeast. Cell, 1988. 55(1): p. 27-39.
16. Heitz, Das Heterochromatin der Moose I. Jb. wiss. Bot., 1928. 69: p. 762-818.
17. Muller, H. and W.S. Stone, Anat Rec, 1930. 47: p. 393-394.
18. Schultz, J., Variegation in Drosophila and the Inert Chromosome Regions. Proc Natl Acad Sci U S A, 1936. 22(1): p. 27-33.
19. Cooper, K.W., Cytogenetic analysis of major heterochromatic elements (especially Xh and Y) in Drosophila melanogaster, and the theory of "heterochromatin". Chromosoma, 1959. 10: p. 535-88.
20. Schotta, G., et al., Central role of Drosophila SU(VAR)3-9 in histone H3-K9 methylation and heterochromatic gene silencing. EMBO J, 2002. 21(5): p. 1121-31.
21. James, T.C. and S.C. Elgin, Identification of a nonhistone chromosomal protein associated with heterochromatin in Drosophila melanogaster and its gene. Mol Cell Biol, 1986. 6(11): p. 3862-72.
22. Fischle, W., et al., Molecular basis for the discrimination of repressive methyl-lysine marks in histone H3 by Polycomb and HP1 chromodomains. Genes Dev, 2003. 17(15): p. 1870-81.
23. Trojer, P. and D. Reinberg, Facultative heterochromatin: is there a distinctive molecular signature? Mol Cell, 2007. 28(1): p. 1-13.
24. Dimitri, P., et al., The paradox of functional heterochromatin. Bioessays, 2005. 27(1): p. 29-41.
25. Hediger, F. and S.M. Gasser, Heterochromatin protein 1: don't judge the book by its cover! Curr Opin Genet Dev, 2006. 16(2): p. 143-50.
26. Fanti, L. and S. Pimpinelli, HP1: a functionally multifaceted protein. Curr Opin Genet Dev, 2008. 18(2): p. 169-74.
27. Handler, A.M. and R.A. Harrell, 2nd, Germline transformation of Drosophila melanogaster with the piggyBac transposon vector. Insect Mol Biol, 1999. 8(4): p. 449-57.
28. Kelley, R.L. and M.I. Kuroda, The Drosophila roX1 RNA gene can overcome silent chromatin by recruiting the male-specific lethal dosage compensation complex. Genetics, 2003. 164(2): p. 565-74.
29. Markstein, M., et al., Exploiting position effects and the gypsy retrovirus insulator to engineer precisely expressed transgenes. Nat Genet, 2008. 40(4): p. 476-83.
30. Allfrey, V.G., R. Faulkner, and A.E. Mirsky, Acetylation and Methylation of Histones and Their Possible Role in the Regulation of Rna Synthesis. Proc Natl Acad Sci U S A, 1964. 51: p. 786-94.
31. Strahl, B.D. and C.D. Allis, The language of covalent histone modifications. Nature, 2000. 403(6765): p. 41-5.
32. Pokholok, D.K., et al., Genome-wide map of nucleosome acetylation and methylation in yeast. Cell, 2005. 122(4): p. 517-27.
33. Bernstein, B.E., et al., A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell, 2006. 125(2): p. 315-26.
34. Heintzman, N.D., et al., Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet, 2007. 39(3): p. 311-8.
35. Rando, O.J., Global patterns of histone modifications. Curr Opin Genet Dev, 2007. 17(2): p. 94-9.
36. Henikoff, S. and A. Shilatifard, Histone modification: cause or cog? Trends Genet, 2011. 27(10): p. 389-96.
37. Suganuma, T. and J.L. Workman, Signals and combinatorial functions of histone

modifications. Annu Rev Biochem, 2011. 80: p. 473-99.
38. Sadri-Vakili, G. and J.H. Cha, Mechanisms of disease: Histone modifications in Huntington's disease. Nat Clin Pract Neurol, 2006. 2(6): p. 330-8.
39. Bannister, A.J., et al., Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. Nature, 2001. 410(6824): p. 120-4.
40. Lachner, M., et al., Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. Nature, 2001. 410(6824): p. 116-20.
41. Bouazoune, K. and A. Brehm, ATP-dependent chromatin remodeling complexes in Drosophila. Chromosome Res, 2006. 14(4): p. 433-49.
42. Guillemette, B., et al., Variant histone H2A.Z is globally localized to the promoters of inactive yeast genes and regulates nucleosome positioning. PLoS Biol, 2005. 3(12): p. e384.
43. Ahmad, K. and S. Henikoff, The histone variant H3.3 marks active chromatin by replication-independent nucleosome assembly. Mol Cell, 2002. 9(6): p. 1191-200.
44. Chadwick, B.P. and H.F. Willard, Histone H2A variants and the inactive X chromosome: identification of a second macroH2A variant. Hum Mol Genet, 2001. 10(10): p. 1101-13.
45. Geyer, P.K. and V.G. Corces, DNA position-specific repression of transcription by a Drosophila zinc finger protein. Genes Dev, 1992. 6(10): p. 1865-73.
46. Kellum, R. and P. Schedl, A group of scs elements function as domain boundaries in an enhancer-blocking assay. Mol Cell Biol, 1992. 12(5): p. 2424-31.
47. Gaszner, M. and G. Felsenfeld, Insulators: exploiting transcriptional and epigenetic mechanisms. Nat Rev Genet, 2006. 7(9): p. 703-13.
48. Sun, F.L. and S.C. Elgin, Putting boundaries on silence. Cell, 1999. 99(5): p. 459-62.
49. Bell, A.C., A.G. West, and G. Felsenfeld, The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. Cell, 1999. 98(3): p. 387-96.
50. Huang, S., et al., USF1 recruits histone modification complexes and is critical for maintenance of a chromatin barrier. Mol Cell Biol, 2007. 27(22): p. 7991-8002.
51. Hagstrom, K., M. Muller, and P. Schedl, Fab-7 functions as a chromatin domain boundary to ensure proper segment specification by the Drosophila bithorax complex. Genes Dev, 1996. 10(24): p. 3202-15.
52. Mihaly, J., et al., In situ dissection of the Fab-7 region of the bithorax complex into a chromatin domain boundary and a Polycomb-response element. Development, 1997. 124(9): p. 1809-20.
53. Zhao, K., C.M. Hart, and U.K. Laemmli, Visualization of chromosomal domains with boundary element-associated factor BEAF-32. Cell, 1995. 81(6): p. 879-89.
54. Gaszner, M., J. Vazquez, and P. Schedl, The Zw5 protein, a component of the scs chromatin domain boundary, is able to block enhancer-promoter interaction. Genes Dev, 1999. 13(16): p. 2098-107.
55. Ohtsuki, S. and M. Levine, GAGA mediates the enhancer blocking activity of the eve promoter in the Drosophila embryo. Genes Dev, 1998. 12(21): p. 3325-30.
56. Bushey, A.M., E.R. Dorman, and V.G. Corces, Chromatin insulators: regulatory mechanisms and epigenetic inheritance. Mol Cell, 2008. 32(1): p. 1-9.
57. Marshall, W.F., Order and disorder in the nucleus. Curr Biol, 2002. 12(5): p. R185-92.
58. Takizawa, T., K.J. Meaburn, and T. Misteli, The meaning of gene positioning. Cell, 2008. 135(1): p. 9-13.
59. Fedorova, E. and D. Zink, Nuclear genome organization: common themes and individual patterns. Curr Opin Genet Dev, 2009. 19(2): p. 166-71.
60. Pickersgill, H., et al., Characterization of the Drosophila melanogaster genome at the nuclear lamina. Nat Genet, 2006. 38(9): p. 1005-14.
61. Guelen, L., et al., Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature, 2008. 453(7197): p. 948-51.
62. Malhas, A., et al., Defects in lamin B1 expression or processing affect interphase chromosome position and gene expression. J Cell Biol, 2007. 176(5): p. 593-603.

63. Shevelyov, Y.Y., et al., The B-type lamin is required for somatic repression of testis-specific gene clusters. Proc Natl Acad Sci U S A, 2009. 106(9): p. 3282-7.
64. Finlan, L.E., et al., Recruitment to the nuclear periphery can alter expression of genes in human cells. PLoS Genet, 2008. 4(3): p. e1000039.
65. Kumaran, R.I. and D.L. Spector, A genetic locus targeted to the nuclear periphery in living cells maintains its transcriptional competence. J Cell Biol, 2008. 180(1): p. 51-65.
66. Reddy, K.L., et al., Transcriptional repression mediated by repositioning of genes to the nuclear lamina. Nature, 2008. 452(7184): p. 243-7.
67. Dialynas, G., et al., The role of Drosophila Lamin C in muscle function and gene expression. Development, 2010. 137(18): p. 3067-77.
68. Kim, Y., et al., Mouse B-type lamins are required for proper organogenesis but not by embryonic stem cells. Science, 2011. 334(6063): p. 1706-10.
69. Osborne, C.S., et al., Active genes dynamically colocalize to shared sites of ongoing transcription. Nat Genet, 2004. 36(10): p. 1065-71.
70. Razin, S.V., et al., Transcription factories in the context of the nuclear and genome organization. Nucleic Acids Res, 2011. 39(21): p. 9085-92.
71. van Steensel, B. and S. Henikoff, Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. Nat Biotechnol, 2000. 18(4): p. 424-8.

The insulator protein SU(HW) fine-tunes nuclear lamina interactions of the *Drosophila* genome

Joke G. van Bemmel[1], Ludo Pagie[1], Ulrich Braunschweig[1], Wim Brugman[3], Wouter Meuleman[1,2,4], Ron M. Kerkhoven[3], Bas van Steensel[1]

PLoS ONE, 5 (11), 2010, e15013

[1]Division of Gene Regulation, [2]Division of Molecular Biology and [3]Central Microarray Facility, Netherlands Cancer Institute, Amsterdam, the Netherlands; [4]Delft Bioinformatics Lab, Delft University of Technology, Delft, the Netherlands

2

| ABSTRACT

Specific interactions of the genome with the nuclear lamina (NL) are thought to assist chromosome folding inside the nucleus and to contribute to the regulation of gene expression. High-resolution mapping has recently identified hundreds of large, sharply defined lamina-associated domains (LADs) in the human genome, and suggested that the insulator protein CTCF may help to demarcate these domains. Here, we report the detailed structure of LADs in *Drosophila* cells, and investigate the putative roles of five insulator proteins in LAD organization. We found that the Drosophila genome is also organized in discrete LADs, which are about five times smaller than human LADs but contain on average a similar number of genes. Systematic comparison to new and published insulator binding maps shows that only SU(HW) binds preferentially at LAD borders and at specific positions inside LADs, while GAF, CTCF, BEAF-32 and DWG are mostly absent from these regions. By knockdown and overexpression studies we demonstrate that SU(HW) weakens genome – NL interactions through a local antagonistic effect, but we did not obtain evidence that it is essential for border formation. Our results provide insights into the evolution of LAD organization and identify SU(HW) as a fine-tuner of genome – NL interactions.

| INTRODUCTION

The nuclear lamina (NL), a dense fibrillar network covering the inside of the nuclear membrane in metazoan cells [reviewed in 1], is thought to represent a major structural element for the nuclear organization of the genome. Close contacts between the NL and chromatin have been observed by electron microscopy [2] and more recently by three-dimensional structured illumination microscopy [3]. Based on FISH studies specific loci are known to preferentially localize at the periphery [reviewed in 4, 5]. Genome-wide mapping using the DamID technology [6] in *Drosophila* Kc cells demonstrated hundreds of genes to be in molecular contact with the NL [7]. These genes are strongly repressed and lack active histone marks. Application of the same mapping technology at a higher resolution in human lung fibroblasts showed that NL interactions occur through large continuous genomic domains with sharply defined borders [8]. In these Lamina Associated Domains (LADs) gene expression is strongly repressed, RNA Polymerase II (RNApolII) and active histone marks are depleted, and repressive histone marks are enriched.

Several observations indicate that LADs are not just passively pushed towards the periphery, but instead are the result of specific NL – genome interactions. Human LAD borders tend to be marked by sequence elements such as outward orientated promoters, CTCF binding sites and CpG islands [8], which indicates that the association with the NL could be controlled by DNA sequence. Furthermore, loss or mutation of lamins in flies and mammals can cause dissociation from the periphery and changes in gene expression, histone modifications and binding of chromatin proteins [9-14]

indicating the functional relevance of genome – lamina associations. In addition, upon differentiation hundreds of genes move from or towards the NL, correlating with their respectively increased and decreased expression levels [7, 15-16]. Taken together this points to the existence of mechanism that regulates genome-NL interactions.

The current knowledge about such a regulatory mechanism is limited. Repressive histone marks are most likely not involved, since loss of histone methyltransferases or DNA methyltransferases do not effect peripheral localization of single loci [15, 17]. Possibly the absence of active histone marks could play a role, since treatment with an HDAC inhibitor has been shown to disrupt molecular interactions with the NL in *Drosophila* Kc cells [7] and to cause dissociation of genes from the nuclear periphery in mammalian cells [18]. Alternatively, DNA-binding proteins that physically interact with the NL could be involved in modulating NL interactions.

Because a subset of human LAD borders is marked by an insulator element, the CTCF binding sequence [8], we reasoned that insulator proteins are likely candidates to be involved in modulating genome-NL interactions. Insulator elements are DNA sequences that, when bound by insulator proteins, are thought to play a key role in chromatin organization by mediating intra- and interchromosomal interactions. They have been shown to block the communication between a promoter and its enhancer when placed in between [19-20, reviewed in 21] and some insulators are thought to separate inactive from active chromatin domains by acting as a barrier against spreading of repressive chromatin proteins or histone modifications into neighbouring regions [reviewed in 21, 22].

In *Drosophila* five main insulator proteins have been identified; Suppressor of Hairy-wing (SU(HW)), CCCTC-binding Factor (CTCF), Boundary Element-associated Factor (BEAF-32), Zeste-white 5 (Zw5) -also known as Deformed Wings (DWG)- and GAGA Factor (GAF) [19, 23-27, reviewed in 28]. Especially SU(HW) is a promising candidate to regulate genome – NL interactions since the SU(HW) complex member TOPORS is shown to interact with lamin proteins [29].

Here we analyze the possible roles of insulator proteins in the regulation of NL- genome interactions in *Drosophila* Kc cells. We report a high resolution map of *Drosophila* genome – NL interactions, showing that *Drosophila* LADs exhibit remarkably similar characteristics as their human counterparts. Comparison to new and published binding maps for all five insulators revealed SU(HW) to be the only insulator protein that preferentially binds at LAD borders and at specific positions inside LADs. By direct functional studies we demonstrate that SU(HW) modulates LAD – NL interactions through a local antagonistic effect. We thus identified SU(HW) as the first protein to fine-tune molecular interactions between the genome and the NL.

| RESULTS

LADs in the *Drosophila* genome

A previous DamID study in *Drosophila* Kc cells identified hundreds of genes that associate with the NL [7]. However, this study lacked the resolution required for a detailed view of genome - NL interaction patterns. We therefore repeated these DamID experiments for LAM (also known als Lamin-Dm0, the only B-type lamin in *Drosophila*), this time using a high-density microarray that queried the entire fly genome with a median probe spacing of ~300 bp. With DamID, DNA adenine methyltransferase (Dam) fused to LAM leaves a stable adenine-methylation 'footprint' *in vivo* at the interaction sites. Previous comparisons to fluoresence in situ hybridization data have indicated that DamID signals obtained with LAM can be interpreted as relative molecular contact frequencies between the NL and the probed genomic locus [7-8]. Note that the Dam-LAM fusion protein is expressed at very low levels, preventing overexpression artifacts.

We averaged the data of two independent DamID experiments, which highly correlated with each other (Pearson correlation of 0.77) and with the previously published low resolution data (Pearson correlation of 0.74). The resulting profile (Figure 1A, Figure S1) shows that the genome in Kc cells is associated with the NL through large continuous domains, alternating with regions of low association. A domain detection algorithm, previously developed for the analysis of human NL interaction data [8] identified a total of 412 *Drosophila* Lamina Associated Domains (LADs) (Table S1, available upon request). These LADs vary in size between 7 and 700 kb, with a median size of ~90kb (red line in Figure 1B). In total they cover 40% of the genome (data not shown).

The *Drosophila* genome is much more compact than the human genome. Although the overall gene counts differ by only 2-fold (respectively 14,449 and ~31,000 genes), the *Drosophila* genome is ~25 times smaller, and *Drosophila* genes are typically shorter and more closely spaced than human genes. These differences in scale raise the question whether the organization of the *Drosophila* genome in LADs also occurs at a smaller scale. Indeed, human LADs are much bigger, ranging from ~0.1Mb to ~10Mb with a median LAD size of 553kb (blue line in Figure 1B) [8]. However, we find that the number of genes per LAD is remarkably similar between the two species, on average respectively 6.8 and 8.5 genes/LAD in human and *Drosophila* cells (Figure 1C). These observations suggest that LAD organization has co-evolved with the linear spacing and size of genes, and that the number of genes, rather than the absolute length of DNA, is an important structural parameter of LADs. Taken together, these results demonstrate that the organization of the genome into LADs is conserved between *Drosophila* and human cells, albeit at different scales.

Drosophila LADs are repressed chromatin domains

Human LADs represent a repressive chromatin environment with a relatively low gene density. To investigate whether *Drosophila* LADs exhibit similar characteristics, we aligned the 412 LADs by their left

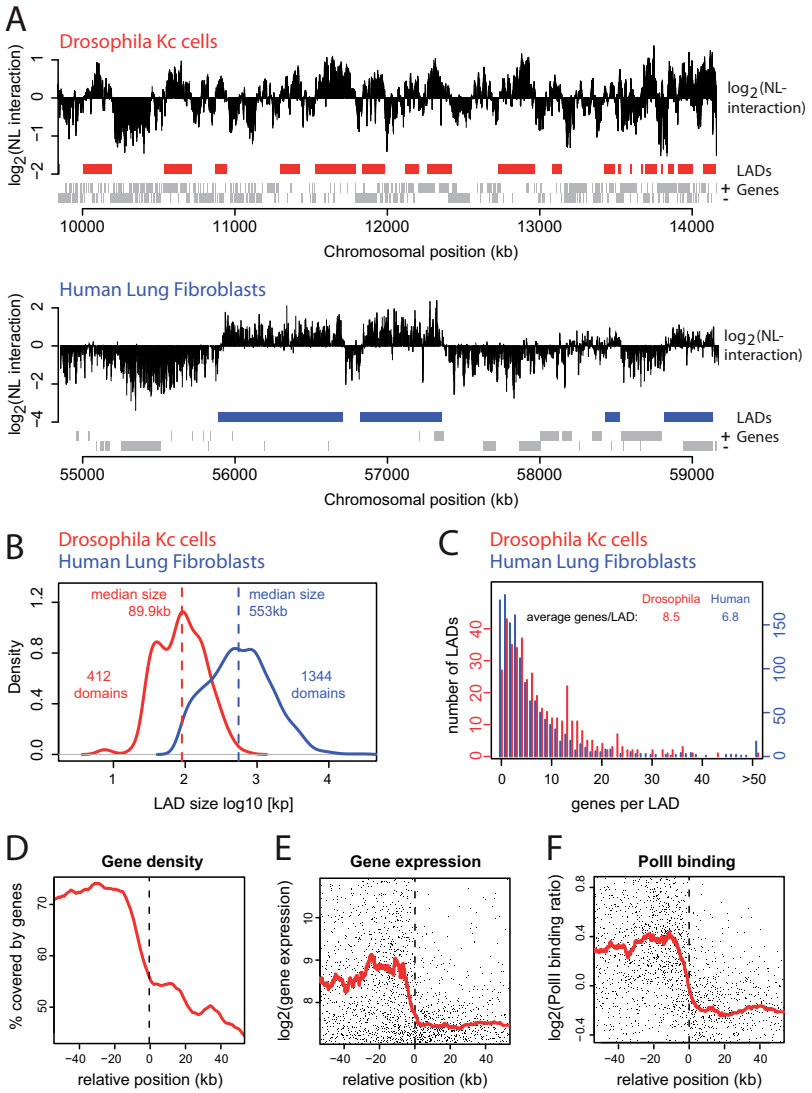


Figure 1 Lamina associated domains in the *Drosophila* genome. (A) Genome – NL interaction maps in *Drosophila* Kc cells and human Lung Fibroblasts along a 4Mb region at respectively chromosome 2L and chromosome 18. Human data are from [8]. Y-axes depict the $\log_2$ transformed Dam-LAM over Dam-only methylation ratio, smoothed by a running median of respectively 15 and 5 probes. Rectangles below each map represent calculated LAD positions for *Drosophila* (red) and human (blue). Grey rectangles at the bottom represent genes at the + and - strand. (B) Distribution of LAD sizes in *Drosophila* (red) and human cells (blue). Dashed lines mark the median LAD sizes. (C) Histogram of the number of genes per LAD. (D-F) Profiles across aligned LAD borders (824 borders, left and mirrored right borders combined). Running window median (red line) and a random subset of 2001 single genes (black dots in E and F). The region around each border from which data was taken ranges from the center of the inter-LAD region to the center of the LAD; this ensures that each data point is used only once. X-axis depicts the position relative to the nearest LAD border; positive coordinates inside, negative coordinates outside LADs. (D) Median gene coverage. (E) mRNA levels in A-values, $(\log_2(Cy5)+\log_2(Cy3))/2$ (F) Median RpII18 occupancy on entire genes as determined by DamID [data from30].

as well as their mirrored right borders and calculated the average profiles for several features across the 824 combined borders. This analysis revealed that 45-55% of the sequence within LADs consists of genes, while outside LADs the gene coverage is ~70% (Figure 1D).

In total, the 412 LADs in Kc cells contain about 30% of all genes. To assess the expression status of these genes we measured the mRNA expression levels of nearly all genes in Kc cells using microarrays and calculated the expression profile across the aligned LADs. As is the case for human LADs, almost all genes inside LADs are expressed at baseline levels, while genes outside LADs display varying and on average higher expression levels (Figure 1E). Consistent with this, the binding of the 18-kDa subunit of RNA polymerase (RpII18) to genes [30] shows a low median level and a low variance inside LADs, compared to inter-LAD regions (Figure 1F). The median mRNA expression levels and RpII18 binding levels both exhibit a sharp transition at LAD borders (red lines in Figure 1E-F), similar to what has been reported for human LADs [8]. Taken together, *Drosophila* LADs exhibit similar characteristics as their human counterparts: they represent a repressive type of chromatin with sharp transitions.

Genome-wide identification of *in vivo* binding sites of insulator proteins

Next, we investigated whether specific insulator proteins are involved in regulating LAD formation. Such an activity requires the candidate insulator protein to bind either inside LADs or at LAD borders. We therefore conducted DamID experiments in Kc cells to obtain whole-genome binding maps for the five known *Drosophila* insulator proteins: BEAF-32, CTCF, DWG, GAF and SU(HW). Full-genome binding profiles of DWG and GAF were previously not available for *Drosophila* Kc cells. Although for CTCF, SU(HW) and BEAF-32 such maps have been reported [31], a proper comparison requires that all the insulator profiles are obtained from the same cells and under the same experimental conditions. For each insulator protein we performed two independent DamID experiments, which highly correlated with each other (Pearson correlation coefficients between 0.71 and 0.83) and with previously published low resolution DamID data (Pearson correlation coefficients of 0.64 for GAF [32] and 0.70 for SU(HW) and BEAF32 [33]). We averaged the duplicate datasets to obtain a single full-genome profile for each protein. The resulting binding profiles of all five insulator proteins are generally characterized by sharp peaks of local enrichment (Figure 2A, Figure S2A). A peak detection algorithm (see Methods) identified 2,173 peaks for DWG; 4,027 for BEAF-32; 1,290 for GAF; 2,930 for CTCF; and 2,986 for SU(HW) within the *Drosophila* genome. Each insulator protein has a unique binding pattern, and co-occurrence of different insulators is present but relatively rare (Figure S2B). An exception is formed by DWG and BEAF-32, which exhibit substantial overlap in their binding pattern.

To validate the insulator DamID profiles we first compared four of the profiles with recent chromatin immunoprecipitation (ChIP) data [31, 34]. This shows that most of the insulator binding peaks

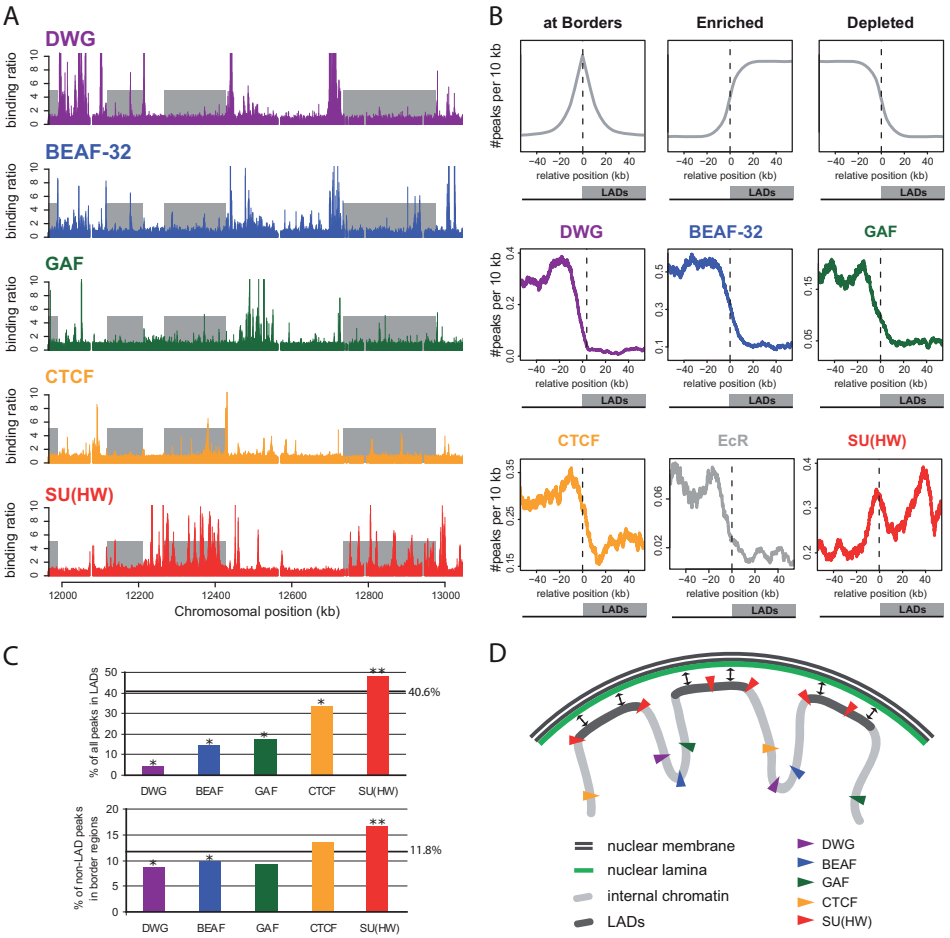


Figure 2 SU(HW) sites are enriched at LAD borders and within LADs. (A) Binding maps of insulator proteins along a 1Mb region on chromosome 2L. Y-axes depict the Dam-insulator over Dam-only methylation ratio (high values are truncated at 10). Grey rectangles represent LADs. (B) Theoretical profiles of features that are respectively enriched at LAD borders ,enriched inside LADs or depleted from LADs (upper panels). Profiles of insulator and EcR binding peaks across aligned LAD borders. 824 borders, left and mirrored right borders combined (lower panels). The region around each border from which data was taken ranges from the center of the inter-LAD region to the center of the LAD; this ensures that each data point is used only once. X-axis depicts the position relative to the nearest LAD border; positive coordinates inside LADs and negative coordinates outside LADs. Y-axes depict the median number of binding peaks within a running window of 10kb. Y-axes are scaled to frequencies within the plotted window, depending on the genome-wide frequency of the feature. (**C**) Percentage of insulator binding peaks within LADs (*top*) and percentage of inter-LAD peaks within a 10kb region just outside LADs (*bottom*). Black horizontal lines represent the percentage expected by chance. **significantly enriched or *depleted compared to random permutation simulations; $p<10^{-3}$. (D) Model of the chromatin organization in LADs, with SU(HW) binding mainly at LAD borders and inside LADs; and DWG, BEAF-32, GAF and CTCF preferentially located in inter-LAD regions.

detected with DamID coincide with corresponding ChIP peaks (Figure S3A). As a second and independent validation method we compared the DamID profiles of SU(HW), CTCF and GAF with the occurrence of their cognate DNA binding motifs. The high co-occurrence of DamID binding sites with the corresponding DNA binding motifs (Figure S3B) further validates the generated profiles. For DWG, ChIP data and a DNA binding motif have not been published before, but the overlap of the DWG and BEAF-32 binding profiles (Figure 2A, Figure S2B) is consistent with the reported direct interaction between DWG and BEAF-32 *in vivo* [35]. We conclude that we have successfully generated high-resolution genome-wide binding profiles for five insulator proteins, together with the NL interaction profile, in one and the same *Drosophila* cell type.

SU(HW) sites are enriched at LAD borders and within LADs

Next, we compared insulator binding profiles with the LAD pattern. Visual inspection led to two observations. First, SU(HW) frequently binds at multiple sites within LADs, while the other four insulator proteins mainly bind outside LADs (Figure 2A and S2A). Second, many LAD borders coincide with a binding peak of one of the insulator proteins.

To address whether these observations generally apply at a genome-wide level, we calculated the density of insulator binding peaks across the 824 aligned LAD borders (Figure 2B). Hypothetically an enrichment at LAD borders or inside LADs would result in respective profiles as depicted in the first two upper panels. The profile of a protein that is depleted from LADs (which would not be expected to be directly involved in LAD formation) is expected to yield a profile as depicted in the third panel. DWG, BEAF-32, CTCF and GAF fall into this latter class: they show similar overall profiles, with the majority of binding sites occurring outside LADs. Their binding frequency gradually decreases across the LAD borders and does not show a peak at the borders themselves, indicating that these four proteins are not specifically enriched at LAD borders. For comparison, the Ecdyson Receptor (EcR) [36], a transcription factor not expected to be linked to NL interactions, yields a similar depletion from LADs. Thus, DWG, BEAF-32, CTCF and GAF are unlikely to play prominent roles in the direct regulation of LAD – NL interactions.

In contrast, SU(HW) exhibits a distinct profile with two prominent regions of enrichment. First, SU(HW) binding peaks are preferentially located in the vicinity of LAD borders, with the highest frequency occurring just outside LADs, at ~4kb from the borders. Second, the profile confirmed the visually observed enrichment of SU(HW) binding peaks inside LADs. Strikingly, within LADs the SU(HW) binding peaks are not equally distributed, but instead are concentrated at a distance of ~40kb from the nearest LAD border (see below). This occurrence of a SU(HW) peak within the 35-45kb region from one or the other border is found in a substantial part (51%) of the LADs of which half the size is at least 45kb (data not shown). Performing the same alignment analysis with ChIP-defined insulator binding peaks [31] results in similar patterns, thereby confirming the

validity of our findings with data from an independent experiment and technique (Figure S3C).

Further statistical analyses confirmed the enrichment of SU(HW) and the depletion of the other four proteins in LADs. For DWG, BEAF-32, CTCF and GAF respectively 4.6%, 14.3%, 17.5% and 33.7% of the binding peaks are located within LADs, while 40.6% is expected by chance. In contrast, 48.1% of the SU(HW) peaks are located within LADs, which is more than expected by chance (top panel Figure 2C). Statistical testing against random permutation simulations showed that this enrichment of SU(HW) and the depletion of the other insulators in LADs is significant ($p < 10^{-3}$). Furthermore, ~17% of the SU(HW) binding peaks outside LADs are located within border regions (defined as the 10kb areas just outside LADs), which is a statistically significant enrichment (Statistical testing against random permutation simulations: $p<10^{-3}$) (bottom panel Figure 2C). None of the four other insulator proteins are significantly enriched in LAD border regions (Statistical testing against random permutation simulations $p>0.01$). DWG and BEAF-32 are even significantly depleted from border regions ($p<10^{-3}$) while GAF and CTCF bind randomly in LAD border regions, as would be expected by chance. In total, 77% of the LADs and 27% of the LAD border regions contain a least one SU(HW) binding peak.

To address whether SU(HW) is specifically enriched inside LADs and not just in repressive chromatin in general, we compared the SU(HW) binding peaks to Polycomb-bound regions (Figure S4), which are large chromatin domains that are mostly transcriptionally inactive [37] and only partly overlap with LADs. The amount of overlap between Polycomb domains and LADs is roughly as may be expected by random chance, namely 40%. Statistical analysis revealed that, in contrast to LADs, Polycomb domains are significantly depleted of SU(HW) binding sites. We find 4.9% of the SU(HW) peaks to be located within Polycomb domains, while 9.3% is expected by chance, showing that the enrichment of SU(HW) is specific for NL-interacting chromatin.

Taken together, these results demonstrate that SU(HW) binding is significantly and specifically enriched just outside LAD borders, as well as at specific locations within LADs (red triangles in Figure 2D). Furthermore, even though DWG, BEAF-32, dCTCF and GAF binding sites occasionally overlap with individual LAD borders, they do not show a statistically significant global preference for LADs or LAD borders (colored triangles in Figure 2D).

Enrichment of SU(HW) is sequence driven

The observed enrichment of SU(HW) and the paucity of the other four insulator proteins in LADs could be dictated by the genomic distribution of the corresponding DNA binding motifs, which is likely since the presence of a SU(HW) DNA binding motif is known to be highly predictive for SU(HW) binding [38]. Alternatively, the patterns could be driven by respectively cooperative or exclusive interactions with other chromatin components in LADs. To discriminate between these two mechanisms we compared the occurrence of protein binding peaks to

the distribution of the corresponding sequence motifs for CTCF, GAF and SU(HW) (respectively grey and colored lines in Figure 3A). This revealed that the distributions of these three insulator proteins and their cognate motifs are highly similar when aligned to LADs and LAD borders. The motifs of CTCF and GAF are depleted in LADs and not specifically enriched in border regions. In contrast, the motif of SU(HW) is enriched within LADs as well as in border regions. Importantly, the SU(HW) motif shows the same prominent enrichment inside LADs at ~40kb from LAD borders as was observed for SU(HW) binding. Thus, the enrichments of SU(HW) at LAD borders and at the +40kb position are to a large extent "hard-coded" in the sequence of the *Drosophila* genome.

The enrichment of SU(HW) at +40kb can not be explained by a subgroup of specifically sized LADs, since alignments of subgroups of differently sized LADs all result in an enrichment at ~40kb (data not shown). The enrichment of SU(HW) at +40kb is also not caused by a genome-wide periodicity of SU(HW) binding

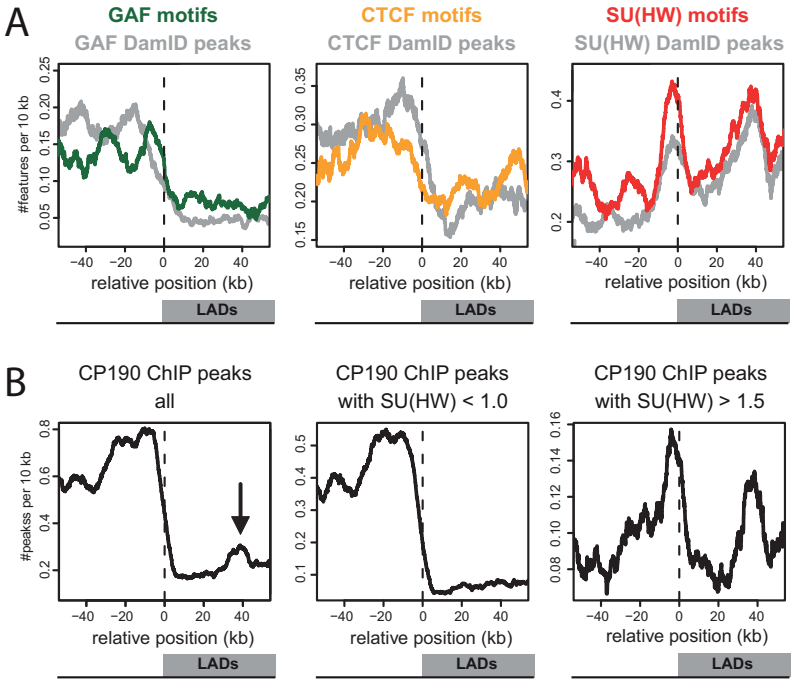


Figure 3 SU(HW) distribution relative to LADs is sequence driven and linked to CP190 binding. (A) Profiles of sequence motifs across aligned LAD borders. X-axis depicts the position relative to the nearest LAD border; positive coordinates inside LADs and negative coordinates outside LADs. Colored lines show the median frequency within a running window of 10kb for sequence motifs, grey lines for DamID identified peaks. (B) Profiles of CP190 peaks [28] across aligned LAD borders; all Cp190 peaks (1st panel), CP190 peaks without SU(HW) binding, defined as peaks with an average $\log_2$(SU(HW)binding ratio) <1.0 (2nd panel), CP190 peaks with SU(HW) binding, defined as CP190 peaks with an average $\log_2$(SU(HW)binding ratio) > 1.5 (3rd panel).

itself: a histogram of the pair-wise distances between all SU(HW) peaks shows no preferential spacing of SU(HW) peaks in the range of 40kb (Figure S5). In addition, we found no significant correlation between the presence of SU(HW) at a LAD border and binding of SU(HW) in the same LAD around +40kb (Fisher's exact test p>0.1, data not shown). Thus, SU(HW) shows preferences for LAD borders as well as for the +40kb position within LADs, but these SU(HW) binding events appear not to be linked. In summary, these results show that the remarkable pattern of SU(HW) relative to LADs is driven by the distribution of its binding motif in the genome.

Enrichment of SU(HW) containing CP190 peaks indicates functionality

To further investigate the remarkable pattern of SU(HW) relative to LADs we analysed the distribution across aligned LAD borders of CP190 binding peaks (defined with ChIP by Bushey et al. [31]). CP190, a subunit of different insulator protein complexes, is thought to be necessary for insulator function since it is essential for both insulator body formation and enhancer blocking activity [31, 39-41]. Figure 3B (1st panel) shows that CP190 is mostly bound outside LADs. Interestingly, the profile also exhibits a modest local peak of enrichment inside LADs, exactly at ~40kb from the LAD borders. CP190 insulator complexes can be divided in at least three different subclasses, containing either SU(HW), CTCF or BEAF-32 [31]. The profile of CP190 binding peaks across LAD borders can therefore be subdivided in peaks that co-localize with SU(HW) and peaks that do not (Figure 3B, 2nd and 3rd panels). Remarkably, CP190 binding peaks that do not contain SU(HW) binding (defined as CP190 peaks with an average $\log_2$(SU(HW)binding ratio) < 1) are depleted from LADs, without any enrichment at +40kb. They probably represent the BEAF-32 and CTCF containing subclasses of CP190 binding peaks. Strikingly, the SU(HW) containing subclass (defined as CP190 peaks with an average $\log_2$(SU(HW)binding ratio) > 1.5) strongly resembles the profile of SU(HW) itself, showing the enrichment just outside LAD borders as well as at +40kb. Taken together, the enrichment of SU(HW) at LAD borders as well as at 40kb inside LADs is confirmed by the binding of CP190, and thus is likely to involve functional insulator protein complexes.

SU(HW) binding antagonizes genome - NL interactions

The surprising pattern of SU(HW) binding relative to LADs suggested two possible roles for SU(HW) in the regulation of genome - NL associations. First, the binding of SU(HW) at LAD borders could help to separate LADs from inter-LADs. Second, the SU(HW) binding inside LADs could modulate NL interactions of LADs. To directly test these hypotheses, we monitored the genome-wide changes in NL interactions after alteration of the expression level of SU(HW). Specifically, we either reduced SU(HW) levels by RNA interference (RNAi), or we increased SU(HW) levels by transfection with a SU(HW) expression vector. We then created new full-genome DamID maps of NL interactions.

For RNAi we used two different, non-overlapping, double-stranded RNA

(dsRNA) fragments to exclude off-target effects. Treatment with a dsRNA fragment derived from the unrelated *white* gene served as a control. Western blot analysis showed that both *su(Hw)* dsRNA fragments caused efficient knockdown of the SU(HW) protein (Figure 4A, 1st panel). Knockdown of SU(HW) had no effect on the doubling time of the cells (data not shown), ruling out secondary effects of an altered cell cycle on the DamID pattern. Elevated levels of SU(HW) were obtained by co-transfection of the DamID plasmids with a vector that drives expression of SU(HW) from an *Act5C* promoter. Western blot analysis showed only a

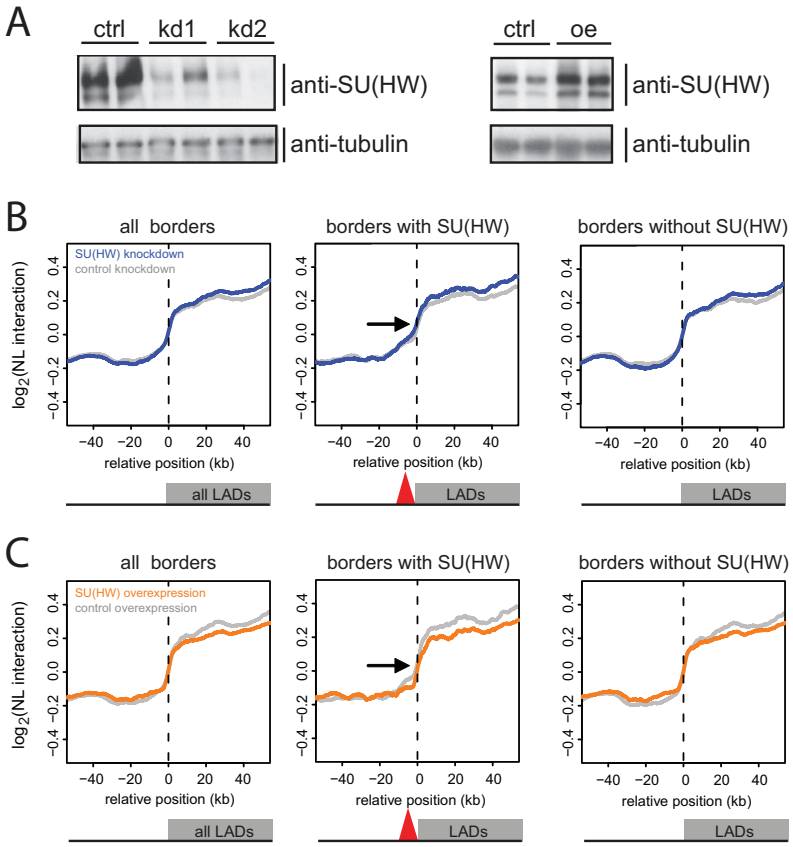


Figure 4 SU(HW) alone is not essential for demarcation of LAD borders. Genome - NL interaction maps after knockdown and overexpression of SU(HW). (A) Western blot analysis of SU(HW) expression levels after knockdown (ctrl: control RNAi; kd1 and kd2: SU(HW) RNAi with two independent dsRNA fragments) and after overexpression (ctrl: control vector oe: overexression by transfection of SU(HW) under an Act5C promoter). 1st lane in each panel: transfected with Dam-LAM, 2nd lane with Dam-only. (B-C) Median NL interaction (log_2 Dam-LAM/Dam ratio) across all aligned LAD borders (824 borders, 1st panel); border regions with SU(HW) present (220 borders, 2nd panel, red triangle represents SU(HW) at the borders), borders without SUH(HW) present (604 borders, 3rd panel). (B) Knockdown of SU(HW) (blue line) and control knockdown (grey line). (C) Overexpression of SU(HW) (orange line), and corresponding control (grey line).

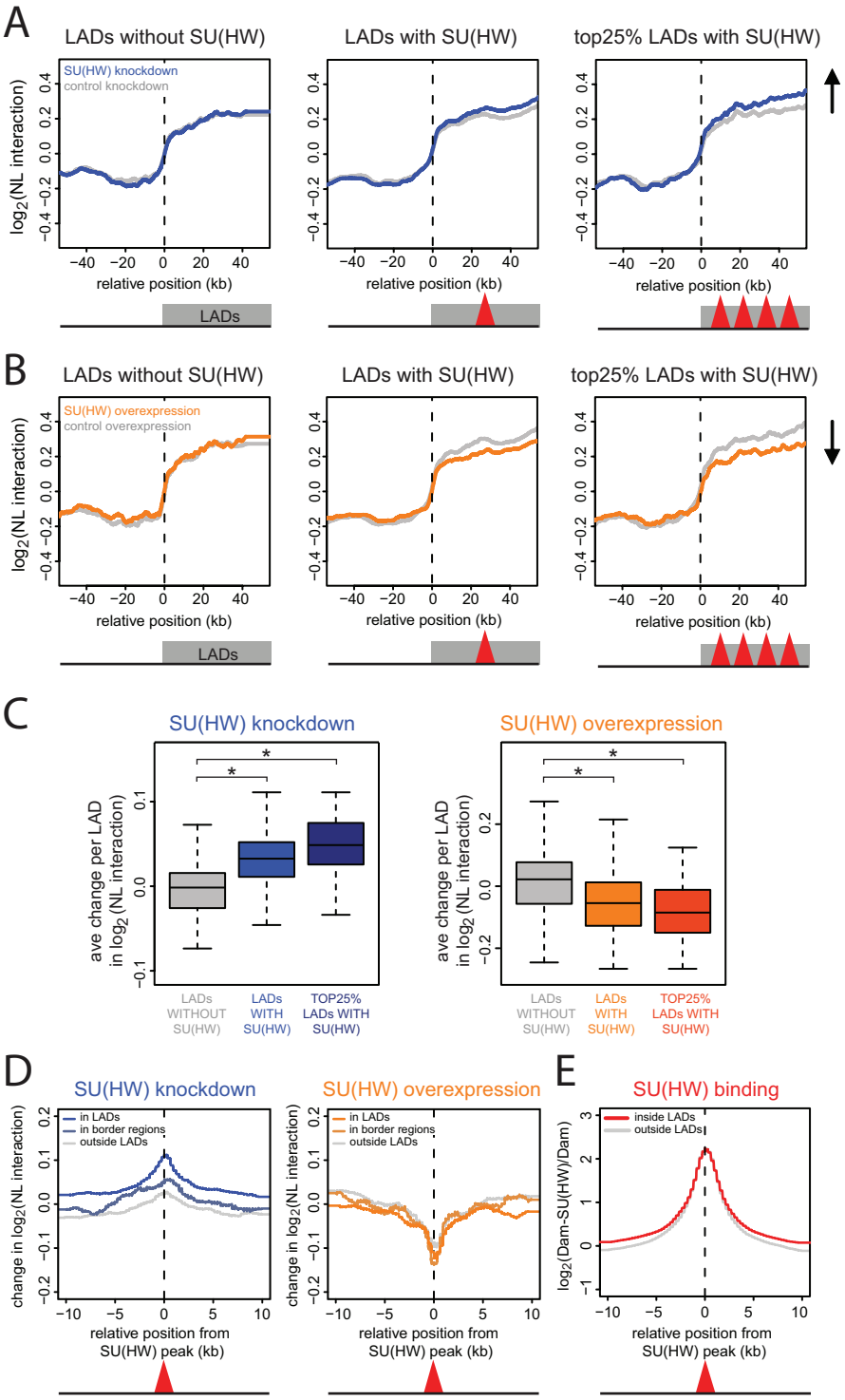
slight increase in expression of SU(HW), presumably because only a minority of cells is transfected (Figure 4A, 2nd panel). However, because the overexpression vector is co-transfected with the DamID vector, overexpression may be expected to be more prominent in cells that express Dam-LAM. We generated DamID maps of NL association for each treatment in two independent experiments.

If SU(HW) is involved in the demarcation of LAD borders, we reasoned that loss of SU(HW) would lead to changes in NL interactions near the borders where SU(HW) is bound, such as expansion or contraction of LADs, or decreased sharpness of borders. However, the median NL interaction profile across LAD borders with SU(HW) bound within 10kb outside the LAD showed no dramatic change at the borders: the median curve did not shift laterally, nor did it change in steepness when comparing SU(HW) knockdown to control cells (blue vs grey in Figure 4B). Overexpression of SU(HW) also had no detectable effect on LAD border sharpness or position (orange vs grey in Figure 4C). We therefore conclude that under the conditions and in the cell type tested, SU(HW) is not essential for the demarcation of LAD borders, despite the clear enrichment of SU(HW) binding sites at LAD borders.

In contrast, we did notice effects of altered SU(HW) levels on genome - NL interactions inside LADs. Knockdown of SU(HW) caused a specific increase in NL interaction levels inside LADs, while overexpression of SU(HW) had the opposite effect inside LADs (color vs grey in the 1st panels of Figure 4B-C). Thus, SU(HW) antagonizes genome – NL interactions inside LADs. To confirm the specificity of this effect, we repeated this analysis after dividing LADs into three classes: LADs without SU(HW) binding peaks inside, LADs with at least one SU(HW) binding peak inside and the 25% of LADs with the highest SU(HW) peak density (Figure 5A-B). This shows that the effects of altered SU(HW) expression levels are restricted to LADs that harbor SU(HW) binding sites, and that the overall effect is proportional to the density of SU(HW) peaks.

To visualize this dependency further we calculated the average change in genome – NL interactions per LAD (Figure 5C), showing that the change in NL interactions is indeed increasing with

Figure 5 SU(HW) is a local antagonist of genome – NL interactions. (A-B) Median NL interaction ($\log_2$ Dam-LAM/Dam ratio) across LADs without SU(HW) peaks (190 borders, 1st panel), LADs with at least one SU(HW) peak (634 borders, 2nd panel, red triangle represents the presence of one or more SU(HW) peaks at any position inside LADs), the 25% of LADs with the highest SU(HW) peak density (206 borders, 3rd panel, red triangles represent high density of SU(HW) peaks). (A) After knockdown of SU(HW) (blue line), after control knockdown (grey line). (B) After overexpression of SU(HW) (orange line), after control overexpression (grey line). (C) Ave changes in NL interaction levels per LAD, for LADs without SU(HW) (grey), LADs with at least one SU(HW) peak (light blue or orange), the 25% of LADs with the highest SU(HW) peak density (dark blue or orange) after knockdown (blue, 1st panel) and overexpression of SU(HW) (orange, 2nd panel). * Wilcoxon test; $p<10^{-3}$ (D) Median changes in NL interaction across aligned SU(HW) peaks (red triangle) inside LADs (bright lines), in border regions (pale lines) and outside LADs (grey lines) after knockdown of SU(HW) (blue, 1st panel) and after overexpression of SU(HW) (orange, 2nd panel). (E) *Cis*-spreading of SU(HW) DamID signals from aligned SU(HW) binding peaks (red triangle), inside LADs (red line) and outside LADs (grey line). ▶


A
LADs without SU(HW)
LADs with SU(HW)
top25% LADs with SU(HW)
SU(HW) knockdown
control knockdown
log2(NL interaction)
relative position (kb)
LADs
B
LADs without SU(HW)
LADs with SU(HW)
top25% LADs with SU(HW)
SU(HW) overexpression
control overexpression
log2(NL interaction)
relative position (kb)
LADs
C
SU(HW) knockdown
SU(HW) overexpression
ave change per LAD in log2(NL interaction)
LADs WITHOUT SU(HW)
LADs WITH SU(HW)
TOP25% LADs WITH SU(HW)
D
SU(HW) knockdown
SU(HW) overexpression
in LADs
in border regions
outside LADs
change in log2(NL interaction)
relative position from SU(HW) peak (kb)
E
SU(HW) binding
inside LADs
outside LADs
log2(Dam-SU(HW)/Dam)
relative position from SU(HW) peak (kb)

higher SU(HW) peak density. A Wilcoxon test between the grouped LADs confirmed NL interactions to significantly increase and decrease after respectively knockdown and overexpression of SU(HW) ($p<10^{-3}$). Repeating this analysis for the individual replicates showed the same trend (Figure S6), thereby confirming the reproducibility of the observed changes in genome – NL interactions.

Together, these results demonstrate that SU(HW) reduces the frequency of NL interactions inside LADs.

The observation that only LADs with SU(HW) binding sites are affected by changes in SU(HW) levels indicates that SU(HW) controls genome – NL interactions locally rather than globally. To investigate the range over which SU(HW) acts *in cis*, we plotted the change in NL interactions as a function of the distance to the nearest SU(HW) binding site, after knockdown and overexpression of SU(HW) (Figure 5D). This reveals that the antagonistic effect of SU(HW) on NL interactions is most pronounced within ~5kb from the SU(HW) binding sites, but extends to >10kb. A weaker effect is also seen at SU(HW) sites in border regions and outside LADs, indicating that in these regions SU(HW) may help to suppress spurious contacts with the NL. Interestingly, the range of the SU(HW) effects on NL interactions corresponds approximately to the width of the SU(HW) binding peaks (Figure 5E). Moreover, an elevated baseline of SU(HW) interactions, extending over >10kb from the binding peak centers, is present specifically inside LADs. Thus, the *cis*-effect of SU(HW) on NL interactions has a similar distribution as the actual contacts of SU(HW) with sequences surrounding the SU(HW) binding sites. Taken together, our data reveal that SU(HW) binding antagonizes genome - NL interactions in a local fashion, over a distance that roughly corresponds to the range over which SU(HW) contacts the genome.

| DISCUSSION

Here, we report the fine structure of LADs in a *Drosophila* cell line, and analyze the genome-wide distribution of five insulator proteins with respect to LADs. The DamID method is particularly suited for the detailed mapping of NL interactions. DamID offers molecular resolution that cannot be obtained by traditional microscopy-based methods such as fluorescent in situ hybridization. This allowed us to compare NL interactions and insulator protein binding sites at a resolution of ~1kb. We discovered that SU(HW) is the only tested insulator protein that preferentially interacts with DNA in LADs and at LAD borders. Furthermore, we demonstrate that SU(HW) modulates genome – nuclear lamina (NL) interactions by local antagonism inside LADs. Because interactions as detected by DamID require at least transient physical contact of the Dam-Lam protein with the chromatin fiber, we interpret changes in DamID signals as local changes in the molecular contact frequency between the NL and the probed genomic locus.

SU(HW) as an antagonist of LAD – NL interactions

The found enrichment of SU(HW) in LADs supports previous ideas about the peripheral localization of SU(HW). SU(HW) is thought to form large aggregates that are often located at the nuclear periphery [42] and a SU(HW) protein complex member, TOPORS, is shown to interact with the NL. The association with the NL seems to be essential since the enhancer blocking activity of the *gypsy* insulator is disturbed by a *lamin* mutation [29]. Our data demonstrate that SU(HW) has an inhibitory effect on genome – NL interactions. This seems at odds with previous observations that the *gypsy* transposable element, which harbors a short array of SU(HW) binding motifs, can target flanking DNA to the nuclear periphery in a SU(HW) – dependent manner [43]. Possibly, the effect of SU(HW) on the nuclear location of *gypsy* is different from that on most LADs, perhaps due to a different sequence context. It is also possible that unknown factors cause SU(HW) to switch from an antagonist to an agonist of LAD – NL interactions depending on the cell type.

At present, it is not known how LADs associate with the NL. Still, we can envision several mechanisms for the antagonistic action of SU(HW). For example, SU(HW) could form bulky DNA-associated complexes that locally disrupt NL interactions, which would be supported by the believe that SU(HW) forms aggregates at the nuclear periphery [42]. In addition distant *gypsy* elements have been suggested to interact with each other [44-45] to form chromatin loops in a SU(HW) dependent manner [46]. Hence, if SU(HW) promotes associations between a LAD and a locus located in the nuclear interior, then naturally the LAD would be less frequently located at the NL.

The presence of a protein inside LADs that antagonizes LAD – NL interactions is somewhat paradoxical. We propose that SU(HW) is important for the fine-tuning of NL interactions. For example, by loosening of LAD – NL associations, SU(HW) may facilitate switches in NL associations of some loci during cellular differentiation. In this respect it is interesting to note that SU(HW) is expressed at particularly high levels in embryos and pupae (http://flybase.org/reports/FBgn0003567.html), when many cells differentiate into new cell types, while the expression is low in larvae and adult flies when relatively few cells change their identity. In addition a fraction of the SU(HW) binding sites has been found to be cell-type specific [31]. Although the antagonistic effect of SU(HW) appears relatively modest at individual loci, the large numbers of SU(HW) sites in many LADs may together modulate chromosome organization to a significant degree.

Evolutionary aspects of LADs and their borders

Our high-resolution map of NL interactions reveals that the fly genome is organized into hundreds of discrete LADs, similar to the human genome. Like in human LADs, most genes in fly LADs exhibit low occupancy by the RNA polymerase II 18kD subunit (RPII18) and they are expressed at low levels. Even though *Drosophila* LADs are about five-fold smaller than human LADs, the number of genes per LAD is remark-

ably similar between the two species. This suggests that LAD organization has co-evolved with the linear spacing and size of genes along the two genomes.

Human and fly LADs are also similar in their demarcation of borders by specific insulator proteins. Surprisingly, the two species employ different insulators for this purpose. In human fibroblasts, a substantial amount of LAD borders is marked by CTCF. In *Drosophila* cells, we observed no such enrichment of CTCF; instead SU(HW) is enriched at LAD borders. SU(HW) is an insect–specific protein [47]. The switch of insulator protein utilization at LAD borders between the two species is remarkable, because this involves not only a functional switch of the two proteins, but simultaneously the co-evolution of the respective binding motifs at many LAD borders.

While the enrichment of SU(HW) near LAD borders is clearly non-random and sequence-based, we were unable to detect significant consequences of the loss or overexpression of SU(HW) on genome – NL interactions around LAD borders. Since SU(HW) alone is not essential for demarcation of LAD borders, we suggest that SU(HW) is redundant with one or more of its partner proteins, such as CP190 or Mod(mdg4) [39, 48]. It is also possible that SU(HW) is only important at LAD borders in specific cell types.

Effects on gene regulation?

Given the strong overall repression of genes in LADs, together with observations that the NL can actively contribute to gene silencing [14, 49-50], it may be expected that the effects of SU(HW) on genome – NL interactions also have impact on gene repression in LADs. Interestingly, gypsy elements, which harbor strong SU(HW) binding sites, can boost transgene expression at many integration sites [51]. Possibly the antagonistic action of SU(HW) on NL interactions contributes to this anti-silencing effect. However, analysis of microarray expression data after SU(HW) knockdown or overexpression in Kc cells (data not shown) did not reveal preferential misregulation of genes in SU(HW)-marked LADs or genes that have respectively increased or decreased levels of NL interactions. Redundant mechanisms may provide robustness to gene repression in LADs and may thus mask a contribution of SU(HW). This notion is supported by the fact that *su(Hw)* mutant flies are viable and exhibit a phenotype that is restricted to impaired oogenesis (Phenotypic Descriptions of Classical Alleles from http://flybase.org/reports/FBgn0003567.html). Finally, it should be considered that the effects of SU(HW) on genome – NL interactions, and thus on the spatial organization of interphase chromosomes, may be important for other nuclear processes, such as the regulation of DNA replication or repair [52].

| MATERIALS AND METHODS

Constructs

Dam-LAM (pDamMyc-Dm$_0$), Dam-Gaf (pNDamMyc-Gaf) and control Dam-only (pNDamMyc) constructs have been described previously [6-7, 34]. To obtain pGWNDamMyc-su(Hw), pGWN-DamMyc-CTCF, and pGWC-Dwg-MycDam, the open reading frames were amplified from cDNA clones (LD15893, GH14774, LD44361, LD45751) and cloned in-frame with Dam by using TOPO cloning and GATEWAY recombination as described [30]. Dam-su(Hw) and Dam-Beaf32 were published before [33].

For overexpression of SU(HW) we constructed vector pA-su(Hw)-iresR, and as a control pAWiresR. To obtain pAWiresRFP, the internal ribosome entry site (IRES) from pUAST-3xEGFP [53] was amplified using primers that incorporate a STOP codon just 5' of the IRES, and SacI sites at both ends of the amplicon. This fragment was cloned into the SacI site of pAWR (The *Drosophila* Gateway[a] Vector Collection, a resource developed by the Murphy lab, Carnegie Institute) to yield pAWiresR. pA-su(Hw)-iresR was obtained by GATEWAY recombination with pAWiresR.

DamID

DamID was performed as described before [54]. In brief, Dam-fusion and Dam-only expression vectors were transfected in parallel into separate dishes of Kc cells by electroporation. Genomic DNA was isolated after 24h and adenine-methylated fragments were amplified from genomic DNA by methylation-specific PCR. 1µg of amplified methylated DNA was labeled with Cy-dye labeled random nonamers (TriLink Biotechnologies, according to NimbleChip Arrays User's Guide: ChIP-chip Analysis v2.0). To correct for nonspecific binding of Dam and local differences in DNA accessibility, methylated fragments of Kc cells transfected with a Dam-only construct were labeled with a different fluorescent dye. 13µg of labeled Dam-fusion and 13µg of Dam-only methylated fragments were pooled and hybridized to microarrays carrying 380,000 60-mer DNA oligonucleotides [55] (Roche-NimbleGen). Median probe spacing is 300bp. For each profile, material from two independent experiments was hybridized in opposite dye orientations over Dam controls. The obtained Dam-fusion/Dam-only ratio reflects the extent of protein binding to each fragment on the microarray, corrected for local differences in chromatin accessibility. Probes are mapped to *Drosophila melanogaster* genome sequence release 4.3.

Knockdown and overexpression of SU(HW)

NL interaction profiles after knockdown of SU(HW) were obtained by using dsRNAs directed against *white* and *su(Hw)*. dsRNAs were *in vitro* transcribed using the RiboMax kit (Promega) from PCR amplicons. PCR amplicons were designed according to the Harvard *Drosophila* RNAi Screening Centre (www.flyrnai.org; *su(Hw)* HFA17074 and MRC020_B05), or as published before for *white* [56]. RNAi treatments were performed as described before [30], with the exception that the treatment was repeated at day 5 and cells were transfected with DamID constructs on day 7.

NL interaction profiles after overexpression of SU(HW) were obtained by co-transfection of DamID vectors and respective overexpression vectors. For overexpression of SU(HW) we used vector pAsu(Hw)iresR. As a control we used the vector pAWiresR. Genomic DNA was isolated after 48h instead of 24h. Expression levels of SU(HW) were monitored with Western blot analysis, presenting the protein expression level within the entire cell population. However, because typically 20-30% of cells are transfected, this yields an underestimate of the degree of overexpression.

SU(HW) antibody

The antibody against the C-terminal peptide of SU(HW) was kindly provided by P. Geyer [57].

Expression analysis

Total RNA was extracted with TRIzol (Invitrogen) and treated with DNaseI. Isolated RNA from three independent cell cultures was labeled with Cy5 and with Cy3 and co-hybridized to INDAC oligo arrays (http://www.indac.net) printed at the NKI Central Microarray Facility, with each oligonucleotide spotted twice. Raw data from three biological replicates were loess normalized per subarray, and averaged. A-values, $(\log_2(Cy5)+\log_2(Cy3))/2$, were used for further analysis.

Data analysis

Microarray data analysis was performed with R [58]. Raw data from two biological replicates were loess normalized, median centered, and dye swap arrays were averaged. For the NL interaction profile after SU(HW) knockdown, the normalized data from the different dsRNA amplicons were averaged as well. To calculate the correlation with previously published low resolution data, the high resolution data were re-sampled to the resolution of the published cDNA microarrays by averaging values for probes from the high resolution array whose center falls within the space of one probe of the cDNA array.

LADs were defined as described in [8]. In short; sharp transitions in the DamID signal were identified using a sliding edge filter (window size 199 probes), and adjacent transitions exceeding a threshold (here 0.3) were combined into domains if at least 70% of the enclosed probes have a positive $\log_2$ ratio. Polycomb domains were taken from [37] and transposed to FlyBase release 4. Insulator peak positions were determined as follows: after applying a running mean of 5 probes, the derivative was calculated over the running-mean with a 7 probe window. In addition FDR-corrected p-values were determined for each probe using linear modeling (LiMMA) [59]. Peaks were assigned at transitions of the derivative from a positive to a negative value (indicating a peak) and where in addition at least three probes were significantly enriched ($p<0.005$). Motif scans were performed using the TFBS Perl module [60] with position weight matrices (PWMs) obtained from literature [38, 61] and the TRANSFAC database [62]. Briefly, PWMs were compared against the genomic sequence and a relative matching score was calculated based on a PWM's information content. A matching score of 85% (CTCF, SU(HW)) and 99% (GAF) was used as it yielded a similar number of matches to the identified in vivo binding sites. Custom R

scripts were used to align data to LAD borders or to SU(HW) peaks; for this purpose, genome-wide positions of all analyzed features were converted to coordinates relative to the nearest border or peak. In case of LADs, data around right-side borders were mirrored and combined with data around left-side borders. The region around borders from which data was taken ranges from the middle of inter-LAD regions to the middle of LADs themselves; this ensures that all datapoints are used only once. Similarly, in case of alignments to Su(HW) peaks, the region ranged halfway to the next peak. Median binding ratios across the aligned borders or peaks were calculated with a running window covering 5% of the data within the aligned region for alignments at LADs (Figure 1D-F, Figure 4B-C, Figure 5A-B), 20% and 10% for changes in NL interaction (Figure 5c) and 10% for aligning at Su(Hw) peaks (Figure 5D-E).

Data availability

DamID and expression data have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession number GSE20313.

| ACKNOWLEDGEMENTS

We thank Marja Nieuwland for mRNA isolation, labeling and microarray hybridization; members of our laboratory, Paul Koppen, Maarten Fornerod and Bernike Kalverda for helpful discussions; and Pam Geyer for providing SU(HW) antibody.

| AUTHOR CONTRIBUTIONS

JvB and BvS conceived and designed the experiments. JvB performed the experiments. UB generated the Lamin DamID profile. WB and RMK performed microarray hybridizations. JvB and LP analyzed the data. WM performed sequence motif searches. JvB and BvS wrote the manuscript.

| REFERENCES

1. Prokocimer, M., et al., *Nuclear lamins: key regulators of nuclear structure and activities.* J Cell Mol Med, 2009. **13**(6): p. 1059-85.
2. Belmont, A.S., Y. Zhai, and A. Thilenius, *Lamin B distribution and association with peripheral chromatin revealed by optical sectioning and electron microscopy tomography.* J Cell Biol, 1993. **123**(6 Pt 2): p. 1671-85.
3. Schermelleh, L., et al., *Subdiffraction multicolor imaging of the nuclear periphery with 3D structured illumination microscopy.* Science, 2008. **320**(5881): p. 1332-6.
4. Takizawa, T., K.J. Meaburn, and T. Misteli, *The meaning of gene positioning.* Cell, 2008. **135**(1): p. 9-13.
5. Fedorova, E. and D. Zink, *Nuclear genome organization: common themes and individual patterns.* Curr Opin Genet Dev, 2009. **19**(2): p. 166-71.
6. van Steensel, B. and S. Henikoff, *Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase.* Nat Biotechnol, 2000. **18**(4): p. 424-8.
7. Pickersgill, H., et al., *Characterization of the Drosophila melanogaster genome at the*

nuclear lamina. Nat Genet, 2006. **38**(9): p. 1005-14.

8. Guelen, L., et al., *Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions.* Nature, 2008. **453**(7197): p. 948-51.
9. Goldman, R.D., et al., *Accumulation of mutant lamin A causes progressive changes in nuclear architecture in Hutchinson-Gilford progeria syndrome.* Proc Natl Acad Sci U S A, 2004. **101**(24): p. 8963-8.
10. Malhas, A., et al., *Defects in lamin B1 expression or processing affect interphase chromosome position and gene expression.* J Cell Biol, 2007. **176**(5): p. 593-603.
11. Scaffidi, P. and T. Misteli, *Reversal of the cellular phenotype in the premature aging disease Hutchinson-Gilford progeria syndrome.* Nat Med, 2005. **11**(4): p. 440-5.
12. Columbaro, M., et al., *Rescue of heterochromatin organization in Hutchinson-Gilford progeria by drug treatment.* Cell Mol Life Sci, 2005. **62**(22): p. 2669-78.
13. Shumaker, D.K., et al., *Mutant nuclear lamin A leads to progressive alterations of epigenetic control in premature aging.* Proc Natl Acad Sci U S A, 2006. **103**(23): p. 8703-8.
14. Shevelyov, Y.Y., et al., *The B-type lamin is required for somatic repression of testis-specific gene clusters.* Proc Natl Acad Sci U S A, 2009. **106**(9): p. 3282-7.
15. Williams, R.R., et al., *Neural induction promotes large-scale chromatin reorganisation of the Mash1 locus.* J Cell Sci, 2006. **119**(Pt 1): p. 132-40.
16. Peric-Hupkes, D., et al., *Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation.* Mol Cell, 2010. **38**(4): p. 603-13.
17. Yokochi, T., et al., *G9a selectively represses a class of late-replicating genes at the nuclear periphery.* Proc Natl Acad Sci U S A, 2009. **106**(46): p. 19363-8.
18. Zink, D., et al., *Transcription-dependent spatial arrangements of CFTR and adjacent genes in human cell nuclei.* J Cell Biol, 2004. **166**(6): p. 815-25.
19. Geyer, P.K. and V.G. Corces, *DNA position-specific repression of transcription by a Drosophila zinc finger protein.* Genes Dev, 1992. **6**(10): p. 1865-73.
20. Kellum, R. and P. Schedl, *A group of scs elements function as domain boundaries in an enhancer-blocking assay.* Mol Cell Biol, 1992. **12**(5): p. 2424-31.
21. Gaszner, M. and G. Felsenfeld, *Insulators: exploiting transcriptional and epigenetic mechanisms.* Nat Rev Genet, 2006. **7**(9): p. 703-13.
22. Sun, F.L. and S.C. Elgin, *Putting boundaries on silence.* Cell, 1999. **99**(5): p. 459-62.
23. Hagstrom, K., M. Muller, and P. Schedl, *Fab-7 functions as a chromatin domain boundary to ensure proper segment specification by the Drosophila bithorax complex.* Genes Dev, 1996. **10**(24): p. 3202-15.
24. Mihaly, J., et al., *In situ dissection of the Fab-7 region of the bithorax complex into a chromatin domain boundary and a Polycomb-response element.* Development, 1997. **124**(9): p. 1809-20.
25. Zhao, K., C.M. Hart, and U.K. Laemmli, *Visualization of chromosomal domains with boundary element-associated factor BEAF-32.* Cell, 1995. **81**(6): p. 879-89.
26. Gaszner, M., J. Vazquez, and P. Schedl, *The Zw5 protein, a component of the scs chromatin domain boundary, is able to block enhancer-promoter interaction.* Genes Dev, 1999. **13**(16): p. 2098-107.
27. Ohtsuki, S. and M. Levine, *GAGA mediates the enhancer blocking activity of the eve promoter in the Drosophila embryo.* Genes Dev, 1998. **12**(21): p. 3325-30.
28. Bushey, A.M., E.R. Dorman, and V.G. Corces, *Chromatin insulators: regulatory mechanisms and epigenetic inheritance.* Mol Cell, 2008. **32**(1): p. 1-9.
29. Capelson, M. and V.G. Corces, *The ubiquitin ligase dTopors directs the nuclear organization of a chromatin insulator.* Mol Cell, 2005. **20**(1): p. 105-16.
30. Braunschweig, U., et al., *Histone H1 binding is inhibited by histone variant H3.3.* Embo J, 2009. **28**(23): p. 3635-45.
31. Bushey, A.M., E. Ramos, and V.G. Corces, *Three subclasses of a Drosophila insulator show distinct and cell type-specific genomic distributions.* Genes Dev, 2009. **23**(11): p. 1338-50.
32. de Wit, E., et al., *Global chromatin domain organization of the Drosophila genome.* PLoS Genet, 2008. **4**(3): p. e1000045.
33. van Steensel, B., et al., *Bayesian network analysis of targeting interactions in chromatin.* Genome Res, 2010. **20**(2): p. 190-200.
34. Moorman, C., et al., *Hotspots of transcription factor colocalization in the genome of Drosophila melanogaster.* Proc Natl Acad Sci U S A, 2006. **103**(32): p. 12027-32.

35. Blanton, J., M. Gaszner, and P. Schedl, *Protein:protein interactions and the pairing of boundary elements in vivo.* Genes Dev, 2003. **17**(5): p. 664-75.
36. Gauhar, Z., et al., *Genomic mapping of binding regions for the Ecdysone receptor protein complex.* Genome Res, 2009. **19**(6): p. 1006-13.
37. Tolhuis, B., et al., *Genome-wide profiling of PRC1 and PRC2 Polycomb chromatin binding in Drosophila melanogaster.* Nat Genet, 2006. **38**(6): p. 694-9.
38. Adryan, B., et al., *Genomic mapping of Suppressor of Hairy-wing binding sites in Drosophila.* Genome Biol, 2007. **8**(8): p. R167.
39. Pai, C.Y., et al., *The centrosomal protein CP190 is a component of the gypsy chromatin insulator.* Mol Cell, 2004. **16**(5): p. 737-48.
40. Gerasimova, T.I., et al., *Coordinated control of dCTCF and gypsy chromatin insulators in Drosophila.* Mol Cell, 2007. **28**(5): p. 761-72.
41. Mohan, M., et al., *The Drosophila insulator proteins CTCF and CP190 link enhancer blocking to body patterning.* Embo J, 2007. **26**(19): p. 4203-14.
42. Gerasimova, T.I. and V.G. Corces, *Polycomb and trithorax group proteins mediate the function of a chromatin insulator.* Cell, 1998. **92**(4): p. 511-21.
43. Gerasimova, T.I., K. Byrd, and V.G. Corces, *A chromatin insulator determines the nuclear localization of DNA.* Mol Cell, 2000. **6**(5): p. 1025-35.
44. Comet, I., et al., *PRE-mediated bypass of two Su(Hw) insulators targets PcG proteins to a downstream promoter.* Dev Cell, 2006. **11**(1): p. 117-24.
45. Kyrchanova, O., et al., *Orientation-dependent interaction between Drosophila insulators is a property of this class of regulatory elements.* Nucleic Acids Res, 2008. **36**(22): p. 7019-28.
46. Byrd, K. and V.G. Corces, *Visualization of chromatin domains created by the gypsy insulator of Drosophila.* J Cell Biol, 2003. **162**(4): p. 565-74.
47. Schoborg, T.A. and M. Labrador, *The Phylogenetic Distribution of Non-CTCF Insulator Proteins Is Limited to Insects and Reveals that BEAF-32 Is Drosophila Lineage Specific.* J Mol Evol, 2009.
48. Gerasimova, T.I., et al., *A Drosophila protein that imparts directionality on a chromatin insulator is an enhancer of position-effect variegation.* Cell, 1995. **82**(4): p. 587-97.
49. Finlan, L.E., et al., *Recruitment to the nuclear periphery can alter expression of genes in human cells.* PLoS Genet, 2008. **4**(3): p. e1000039.
50. Reddy, K.L., et al., *Transcriptional repression mediated by repositioning of genes to the nuclear lamina.* Nature, 2008. **452**(7184): p. 243-7.
51. Markstein, M., et al., *Exploiting position effects and the gypsy retrovirus insulator to engineer precisely expressed transgenes.* Nat Genet, 2008. **40**(4): p. 476-83.
52. Lankenau, D.H., M.V. Peluso, and S. Lankenau, *The Su(Hw) chromatin insulator protein alters double-strand break repair frequencies in the Drosophila germ line.* Chromosoma, 2000. **109**(1-2): p. 148-60.
53. Wang, Z., et al., *Taste representations in the Drosophila brain.* Cell, 2004. **117**(7): p. 981-91.
54. Greil, F., C. Moorman, and B. van Steensel, *DamID: mapping of in vivo protein-genome interactions using tethered DNA adenine methyltransferase.* Methods Enzymol, 2006. **410**: p. 342-59.
55. Choksi, S.P., et al., *Prospero acts as a binary switch between self-renewal and differentiation in Drosophila neural stem cells.* Dev Cell, 2006. **11**(6): p. 775-89.
56. Greil, F., et al., *Distinct HP1 and Su(var)3-9 complexes bind to sets of developmentally coexpressed genes depending on chromosomal location.* Genes Dev, 2003. **17**(22): p. 2825-38.
57. Parnell, T.J., et al., *An endogenous suppressor of hairy-wing insulator separates regulatory domains in Drosophila.* Proc Natl Acad Sci U S A, 2003. **100**(23): p. 13436-41.
58. R-Development-Core-Team, *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2010.
59. Smyth, G.K. and T. Speed, *Normalization of cDNA microarray data.* Methods, 2003. **31**(4): p. 265-73.
60. Lenhard, B. and W.W. Wasserman, *TFBS: Computational framework for transcription factor binding site analysis.* Bioinformatics, 2002. **18**(8): p. 1135-6.
61. Holohan, E.E., et al., *CTCF genomic binding sites in Drosophila and the organisation of the bithorax complex.* PLoS Genet, 2007. **3**(7): p. e112.
62. Matys, V., et al., *TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.* Nucleic Acids Res, 2006. **34**(Database issue): p. D108-10.

| SUPPORTING INFORMATION

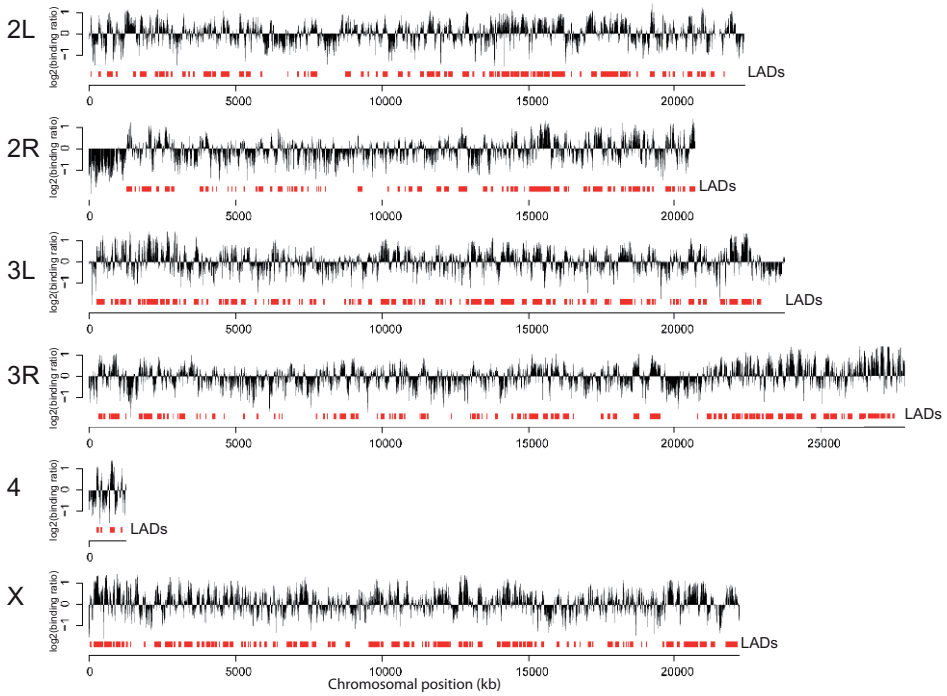


Figure S1 Genome – NL interaction map in *Drosophila* Kc cells on all chromosomes. Y-axes depict the $\log_2$ transformed Dam-LAM over Dam-only methylation ratio with a running median of 15 probes. Red rectangles represent LADs.

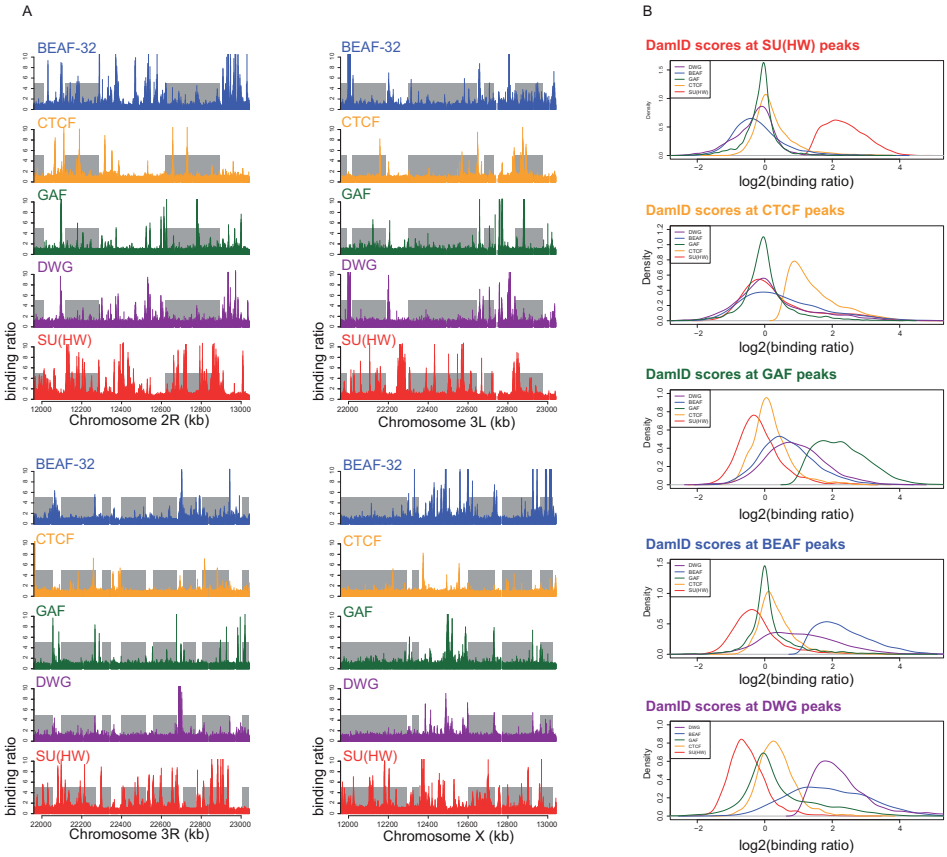


Figure S2 Insulator protein binding map in *Drosophila* Kc cells. (A) Insulator protein binding maps at four arbitrarily chromosomal regions. Y-axes depict the Dam-insulator over Dam-only methylation ratio. Grey rectangles represent LADs. (B) Co-occurrence of insulator proteins indicated by a density plot of the log2 transformed binding ratio of each insulator protein (colored lines) at the binding peaks of each insulator protein (different panels).

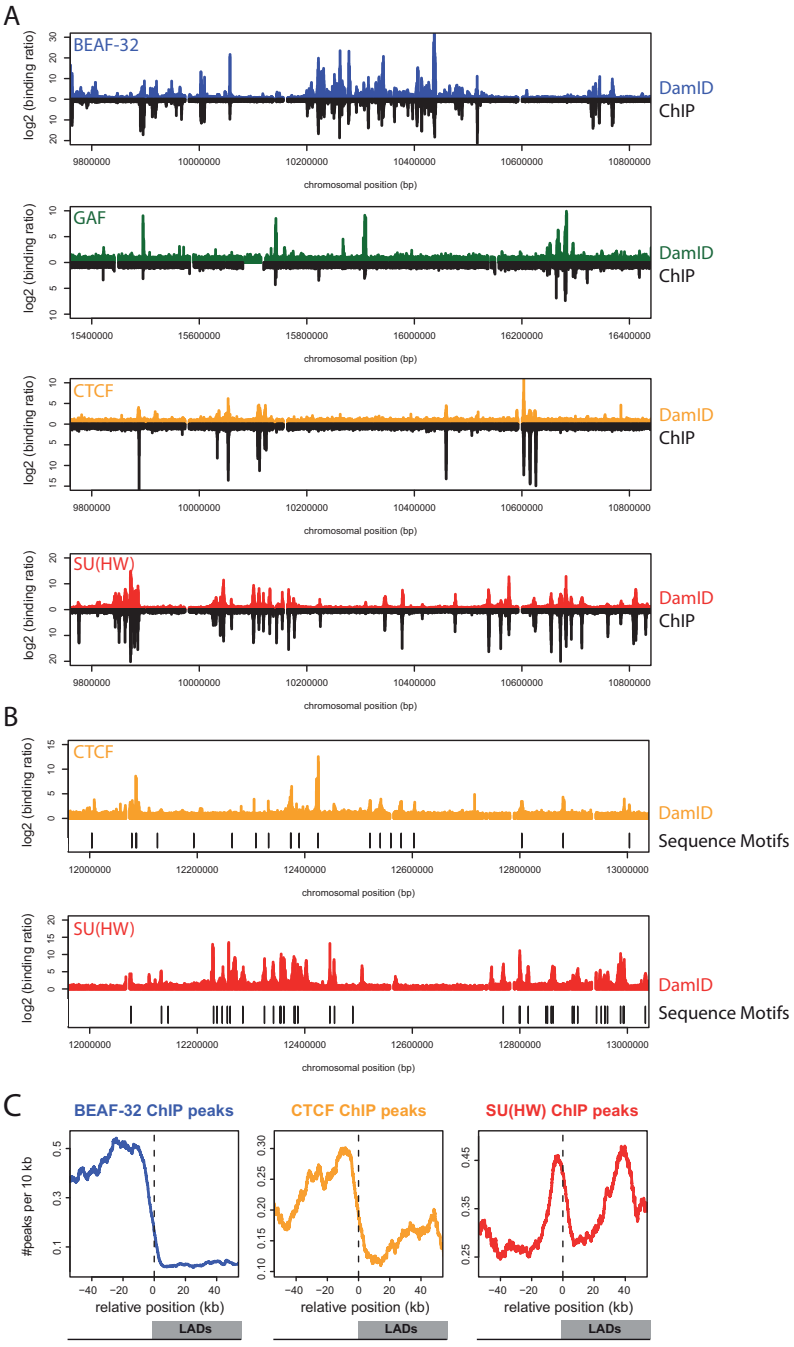


Figure S3 DamID maps are consistent with ChIP and sequence motif distributions. (A) Binding maps of BEAF-32, GAF, CTCF and SU(HW) at random regions of chromosome 2L for Dam-insulator over Dam-only methylation ratios (colored lines) versus ChIP scores (black). (B) DamID binding maps of CTCF and SU(HW) (colored lines) at chromosome 2L versus the location of corresponding sequence motifs (black).

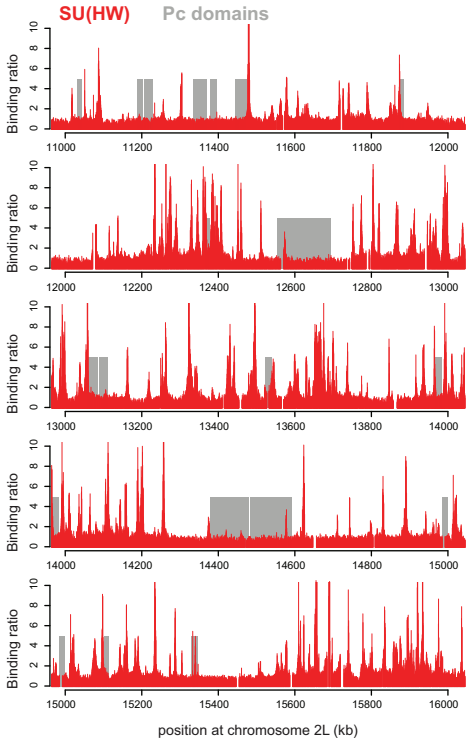


Figure S4 No SU(HW) enrichment in Polycomb domains. (A) Binding maps of insulator proteins along a five sequential 1Mb regions at chromosome 2L. Y-axes depict the linear Dam-SU(HW) over Dam-only methylation ratio. Grey rectangles represent the Polycomb domains.

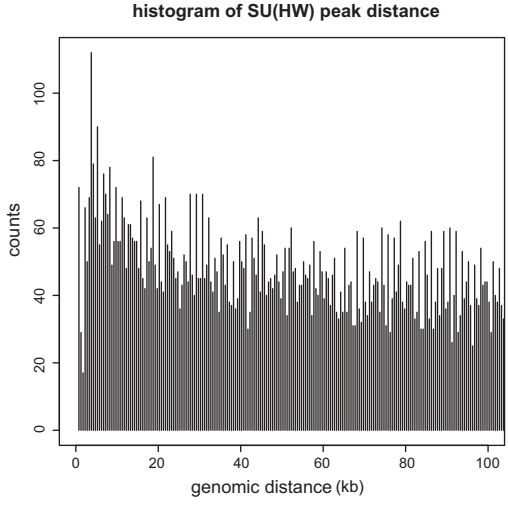


Figure S5 No preferential spacing of SU(HW) peaks in a range of 40kb. Histogram of the pair-wise distances between all SU(HW) peaks. X-axis depicts genomic distance between the peaks.

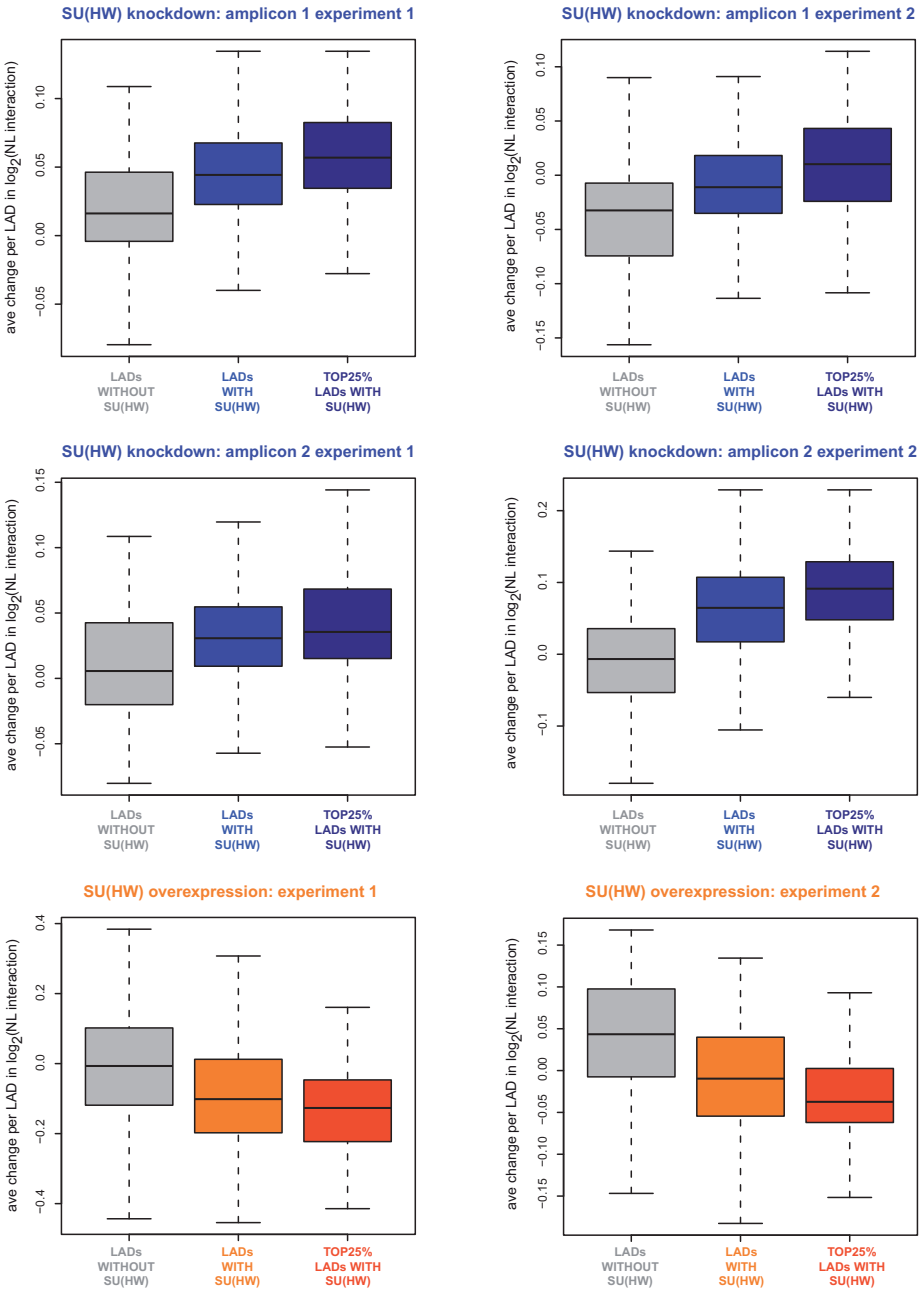


Figure S6 Changes in NL interaction after altering SU(HW) expression levels are reproducible. Ave changes in NL interaction levels per LAD, for LADs without SU(HW) (grey), LADs with at least one SU(HW) peak (light blue or orange), the 25% of LADs with the highest SU(HW) peak density (dark blue or orange) after knockdown with amplicon 1 (blue, upper panles), knockdown with amplicon 2 (blue, middle panels) and overexpression of SU(HW) (orange, lower panels). First experiment (left panels), second experiment (right panels).

A direct role for cohesin in gene regulation and ecdysone response in *Drosophila* salivary glands

Andrea Pauli[1,5], Joke G. van Bemmel[2], Raquel A. Oliveira[1], Takehiko Itoh[3], Katsuhiko Shirahige[4], Bas van Steensel[2], Kim Nasmyth[1]

Curr Biol, 20 (20), 2010, 1787-98

[1]Department of Biochemistry, University of Oxford, Oxford, UK; [2]Division of Gene Regulation, Netherlands Cancer Institute, Amsterdam, The Netherlands; [3]Laboratory of In Silico Functional Genomics, Graduate School of Bioscience, Tokyo Institute of Technology, Japan; [4]Laboratory of Genome Structure and Function, Research Center for Epigenetic Disease, Institute of Molecular and Cellular Biosciences, The University of Tokyo, Japan; [5]Current address: Department of Cellular and Molecular Biology, Harvard University, Cambridge, USA

3

| ABSTRACT

Developmental abnormalities observed in Cornelia de Lange Syndrome (CdLS) have been genetically linked to mutations in the cohesin machinery. These and other recent experimental findings have led to the suggestion that cohesin, in addition to its canonical function of mediating sister chromatid cohesion, might also be involved in regulating gene expression. We report that cleavage of cohesin's kleisin subunit in post-mitotic *Drosophila* salivary glands induces major changes in the transcript levels of many genes. Kinetic analyses of changes in transcript levels upon cohesin cleavage reveal that a subset of genes responds to cohesin cleavage within a few hours. In addition, cohesin binds to most of these loci, suggesting that cohesin is directly regulating their expression. Amongst these genes are several that are regulated by the steroid hormone Ecdysone. Cytological visualization of transcription at selected Ecdysone-responsive genes reveals that puffing at *Eip74EF* ceases within an hour or two of cohesin cleavage, long before any decline in Ecdysone Receptor associated with this locus. We conclude that cohesin regulates expression of a distinct set of genes, including those mediating the Ecdysone response.

| INTRODUCTION

The regulation of gene expression essential for normal animal development is largely mediated by sequence specific transcription factors. One of the more mysterious aspects of developmentally regulated transcription concerns how transcription factors bound to remote regulatory sequences modulate transcription of genes many kilobases away while having no effect on neighboring genes. These distant factors must either slide long distances along chromatin fibres or else interact directly with those factors bound close to the start of transcription, with intervening chromatin forming a loop. Due to their proposed roles in chromatin looping, it is suspected that factors that regulate chromatin topology might have key roles in modulating transcription. One such factor is cohesin, a multi-subunit complex essential for sister chromatid cohesion necessary for mitotic chromosome segregation [1]. Cohesin's Smc1, Smc3, and Rad21/Scc1 subunits form a three-membered ring within which sister chromatin fibres are entrapped in a process that requires a separate "cohesin loading factor" composed of the Scc2 and Scc4 proteins. By entrapping unreplicated DNAs, cohesin could in principle hold distant sequences of the same chromatid together (in *cis*), using the same topological principle by which sister DNAs are held together in trans.

Cohesin clearly functions in processes besides sister chromatid cohesion because it is associated with chromatin in most if not all quiescent cells [2] and is essential for pruning of postmitotic neurons, at least partly by regulating levels of Ecdysone Receptor [3-4]. Whether or not cohesin regulates transcription has hitherto been investigated mainly by analyzing the effects of its depletion using RNA interference. Depletion of its Rad21/Scc1 subunit causes twofold changes in expression of the *H19* and *IGF2* genes in HeLa cells [2] and little or no effect on induc-

ibility of the gene encoding Interferon-γ in T cells despite destroying a putative loop between its enhancer and promoter sequences [5]. In *Drosophila* BG3 tissue culture cells, up to 10- to 50-fold changes in the level of transcripts from the *enhancer of split* and *invected-engrailed* loci were detected 6 days after RNAi treatment [6]. Intriguingly, substantial changes in mRNA levels for these transcripts were only observed 3 days following RNAi treatment. Though insightful, these experiments have a number of limitations. The effects on transcription are either modest or they are only seen long after cohesin depletion and might therefore be secondary effects due to chromosome missegregation, defective DNA repair, or some other hitherto uncharacterized state of stress induced by a loss cohesin activity.

Another line of evidence hinting at a role for cohesin in transcriptional control is the finding that inactivation of one allele of *Nipped-B*, the *Drosophila* ortholog of *Scc2*, alters long-range enhancer-promoter interactions at the homeotic loci *cut* and *Ultrabithorax* (*Ubx*) at least when compromised by a gypsy retrotransposon [7-9]. Moreover, mutating *Rad21* in zebrafish reduces expression of the hematopoietic transcription factors *RUNX1* and *RUNX3* during development [10], while mutations in *mau-2*, the *C. elegans Scc4* ortholog, cause defects in axon guidance [11-12]. Particularly striking is the finding that Cornelia de Lange syndrome (CdLS), a multi-system developmental disorder, is caused (in more than 50% of cases) by haplo-deficiency of *NIPBL/Delangin*, the human *Scc2/Nipped-B* ortholog [13-15]. Because tissue culture cells derived from CdLS patients have apparently normal sister chromatid cohesion, dysregulated gene expression during embryonic development has been suggested as potential cause. There are indeed minor changes in the expression of certain genes in *NIPBL+/-* mice (up to 2.5-fold) [16] and CdLS-patient-derived cell lines (up to 4-fold) [17], but these so far do little to explain the developmental defects associated with CdLS, which could in principle be due to defective DNA repair at crucial stages of development.

Ideally, an investigation of cohesin's role in transcription should aim to observe the immediate consequences of the complex's inactivation in cells that are neither undergoing mitosis nor replicating their DNA. Sister chromatid cohesion is normally destroyed at the onset of anaphase by separase-mediated cleavage of cohesin's Rad21/Scc1 α-kleisin subunit, which destroys its topological entrapment of chromatin fibres by opening the cohesin ring [18-19]. This process can be reproduced in an inducible manner using TEV (Tobacco Etch Virus) protease (TEV) in strains of *Drosophila melanogaster* whose α-kleisin Rad21 contains TEV cleavage sites [3, 20]. We describe here the effect on gene expression of TEV-induced Rad21 cleavage in a non-proliferating tissue, which constitutes conclusive evidence that cohesin has a direct role in regulating transcription.

| RESULTS

Transcriptional changes within salivary glands due to cohesin cleavage

To analyse cohesin's role in gene regulation, we used a heat-inducible transgene (*hs-TEV*) to induce TEV in terminally differentiated third instar *Drosophila* salivary glands expressing either wild type or TEV-cleavable myc_{10}-tagged Rad21 protein ($Rad21^{TEV}$) (see outline in Figure 1A). This tissue undergoes multiple rounds of endoreplication (repeated cycles of S- and G-phases without intervening mitoses or cell division), giving rise to transcriptionally active giant polytene chromosomes containing ~1000 closely aligned sister DNAs. We have shown previously that heat shock induction of TEV in late third instar larvae (at a time when there is no further replication in salivary glands) removes TEV-cleavable Rad21 protein from chromosomes within four hours, without any obvious change in their morphology [3]. Late third instar salivary glands are therefore an ideal tissue to study the putative role for cohesin in gene expression since possible changes in transcript levels cannot be attributed to changes in chromosome morphology, to defective DNA repair during DNA replication, or to side-effects caused by chromosome missegregation.

RNAs were isolated from salivary glands expressing wild type (+ cohesin) or TEV-cleavable Rad21 (- cohesin) 10-12 hours after heat shock induction of TEV. This time-point was chosen since the heat shock transcriptional response will have abated but most Rad21 containing TEV sites remains cleaved, and newly synthesized $Rad21^{TEV}$ does not re-accumulate until about 16 hours after the heat shock (Figure S1A and [3]). Both RNA samples were converted to cDNA, labeled with Cy3 and Cy5 respectively, and hybridized to INDAC FL003 arrays containing 18,240 transcript-specific 70mer oligonucleotides. Analysis of seven arrays, each hybridized to an independently generated sample-pair (Figure S1B and C) revealed major differences in the levels of certain transcripts. Cohesin cleavage caused 78 transcripts to increase and 55 to decrease at least four-fold. Moreover, 419 genes (262 up and 157 down) genes changed at least 1.5-fold (Figure 1B and Table S1, available upon request), which suggests that cohesin may function both as an activator and repressor of transcription. Apart from the highly downregulated divergently transcribed *Sgs1* and *hoe2* gene pair, differentially expressed genes are not clustered in the genome and are implicated in a variety of biological processes (Gene Ontology (GO) Enrichment analysis, Table S2, available upon request).

qRT-PCR analysis of selected candidates confirmed differential expression of 16 out of 19 differentially expressed genes tested (Figure 1B and data not shown), revealing up to 100-fold changes. In many cases, changes in transcript levels were accompanied by corresponding changes in association of RNA Polymerase II (Pol-II) with transcription units (Figure 1C, see Figure S2A/B for additional examples), as measured by ChIP-CHIP analysis. Thus, TEV cleavage of cohesin depleted Pol-II from the *EcR* locus, while it increased Pol-II's association with the *ush* locus. Consistent with the relatively small set of differentially expressed genes after cleavage of cohesin, major changes

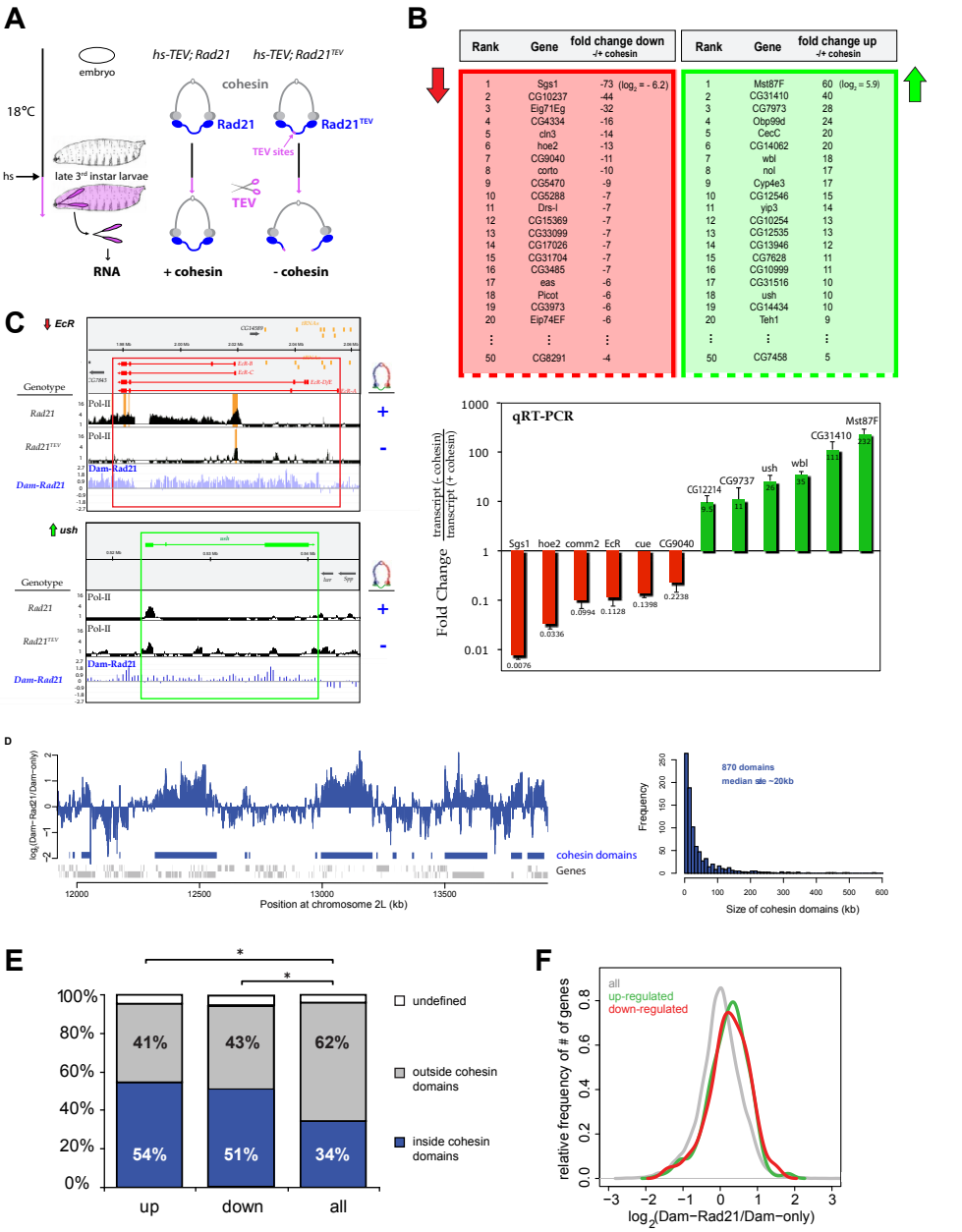


Figure 1 Cleavage of cohesin causes major transcriptional changes in salivary glands. Transcriptional changes in salivary glands in the absence versus presence of cohesin were assessed after heat shock-induced TEV cleavage of cohesin. Green indicates upregulation, red downregulation in the absence of cohesin. (A) Outline of the hs-TEV system used for differential gene expression profiling in salivary glands. Larvae carrying the *hs-TEV* construct and containing cohesin complexes with wildtype or TEV-cleavable (purple arrow) Rad21 were raised at 18°C before TEV protease was induced ubiquitously in late third instar larvae by heat shock (hs, 45 min, 37°C). Salivary glands were dissected 10-12 hours after hs, followed by RNA isolation. Rad21 ▶

in the binding of Pol-II were confined to rare loci.

The transcriptional program within salivary glands changes both during the third instar larval stage and at the transition to the pupal stage [21]. It was therefore possible that the observed differences in gene expression were related to differences in the developmental stage, especially when comparing fly stocks carrying different transgenes (*Rad21^{TEV}* animals develop slightly slower at 18°C than Rad21 animals). To exclude that the changes in gene expression after TEV cleavage of cohesin were due to any minor developmental difference caused by the presence of TEV sites within Rad21, we analysed Pol-II profiles from salivary glands of *Rad21^{TEV}* larvae in the absence of TEV protease induction (+ cohesin) and performed qRT-PCR analysis to compare transcript levels of selected differentially expressed candidates in larvae of different ages with those in the absence versus presence of cohesin. The results of these experiments implied that the majority of changes in gene expression after TEV cleavage of cohesin were indeed due to loss of cohesin (Figure S3A and B).

Differentially expressed genes are preferentially bound by cohesin

To analyse cohesin's distribution in salivary glands we used DamID (DNA adenine methylase identification) [22-23], which involves detecting sites of adenine methylation in transgenic strains expressing bacterial DNA adenine methyltransferase (Dam) fused to a protein of interest, in this case Rad21. DamID was chosen since we were not successful to obtain sufficient high quality Rad21myc-ChIPed starting material, most likely due to poor efficiency of myc-antibodies for ChIP. Because Dam fusion proteins must be expressed at very low levels to ensure

▶ is shown in blue, TEV in purple (see also Figure S1A). (B) (top panel) List of the 20 most downregulated and 20 most upregulated genes upon cohesin cleavage in salivary glands identified by microarray analysis (see also Figure S1B/C and Table S1 and S2, available upon request). Genes are sorted in descending order based on their average fold change in transcript levels in the absence versus presence of cohesin across seven independent microarrays. A minus indicates fold downregulation. Genes at rank 50 are also given. Differential expression of selected candidates was confirmed by quantitative real-time PCR (qRT-PCR) (bottom panel). Each bar represents the average fold change in the absence versus presence of cohesin of at least 3 independent experiments (error bars are standard deviations of the mean). Gene names are given above each bar, fold changes are given below/inside each bar. (**C**) ChIP-CHIP analysis of the distribution of Pol-II (black plots) in *Rad21* (+ cohesin) and *Rad21^{TEV}* (- cohesin) salivary glands 10-12 hours after heat shock induction of TEV protease. Cohesin binding (blue plots) in salivary glands was assessed by DamID (Dam-Rad21). ChIP-CHIP data is represented as fold enrichment of IP over Input (MAT scores; log-scale; highly enriched regions ($p < 0.0001$) are coloured in orange). DamID data is represented as the relative enrichment of methyl-adenine marked DNA from Dam-Rad21 glands over Dam-only glands ($\log_2$ scale). *EcR* (downregulated, red box) and *ush* (upregulated, green box) loci are shown as representative examples (see also Figure S2). (D) Rad21-bound domains ("cohesin domains") across a randomly chosen 2 Mb chromosomal region of chromosome 2L. Shown is the relative enrichment of DamRad21 versus Dam-only signal ($\log_2$ scale). Rad21 domains are highlighted as blue bars; genes are indicated as grey bars. The size distribution of the total number of 870 Rad21 domains in salivary glands is shown at the right. (E) Differentially expressed genes are enriched in Rad21 binding. Shown are percentages of transcriptional start sites (TSSs) of cohesin-dependent genes (up- or downregulated) and of all genes that localized inside (blue), outside (grey) or at the border (white) of Rad21 bound regions. The asterisk (*) indicates Fisher's exact test: $p < 0.01$. (F) Average Rad21 binding at the TSSs of upregulated genes (green), downregulated genes (red) and all genes (grey).

specificity of methylation, it is not possible to assess their functionality directly. We therefore measured the ability of mRNAs, encoding Rad21 or Dam-Rad21, to rescue mitotic defects associated with TEV-induced cleavage of cohesin during syncytial divisions in *Rad21^{TEV}* embryos [20]. mRNAs encoding Dam-Rad21 were, similarly to those encoding wild type Rad21, able to rescue precocious sister chromatid separation caused by TEV injection (Supplemental Movies 1, 2 and 3, available upon request), suggesting that the Dam-Rad21 fusion protein is functional, at least in conferring sister chromatid cohesion. The Dam-Rad21 binding profile, namely the relative enrichment of methyl-adenine marked DNA fragments from *Dam-Rad21* third instar salivary glands over a *Dam-only* control, shows Rad21 enrichment at large regions of the genome containing one or more transcription units (Figure 1D, see Figure S2A/B for additional examples). A domain detection algorithm (see Experimental Procedures for details) identified a total of 870 Rad21 bound regions (so-called cohesin domains) varying in size from ~2 to ~650 kb, with a median size of ~20 kb (Figure 1D). Cohesin domains cover 33% of the genome and contain 34% of all the transcription start sites (TSSs) (defined as the 1 kb region downstream of the transcriptional start). These cohesin domains identified in salivary glands substantially overlap with the Smc1-bound domains identified by ChIP-CHIP in cultured cells [24]. The TSSs of genes defined as cohesin-bound by Misulovin and colleagues [24] are significantly enriched inside our cohesin domains ($p<10^{-5}$, data not shown), confirming the validity of our approach.

Notably, cohesin domains are significantly enriched in genes that are differentially expressed upon loss of cohesin. 54% of TSSs of upregulated genes and 51% of the TSSs of downregulated genes localize within a cohesin domain, which is a significantly (Fisher's exact test, up: $p=2.36e^{-11}$; down: $p=1.12e^{-05}$) larger number compared to only 34% of all TSSs (Figure 1E). Calculation of the average Rad21 binding at the TSS of each gene confirmed that TSSs of cohesin-dependent genes are significantly enriched for Rad21 binding (Wilcoxon test, up: $p=4.3e^{-11}$; down: $p=3.19e^{-08}$) (Figure 1F). Together, these results are consistent with previous reports [6, 17] and suggest that cohesin may indeed be the primary cause of the transcriptional changes observed for more than half of the genes whose expression changes after cohesin cleavage.

Cohesin is an essential regulator of the transcriptional response to Ecdysone

We noticed that several of the differentially expressed genes had previously been implicated in the 20-hydroxyecdysone (Ecdysone) response, including the *Ecdysone Receptor* (*EcR*) itself, whose protein level is reduced by cohesin cleavage in postmitotic neurons [3-4]. Encouraged by these findings, we addressed whether cohesin may have a general role in the transcriptional regulation of Ecdysone-responsive genes. Comparison of the published list of 555 genes whose expression levels changes upon Ecdysone treatment in cultured larval organs [25] to our list of differentially expressed genes after cohesin cleavage revealed that out of the 424 Ecdysone-responsive genes of which we had expression data, 33 (7.7%) were differentially expressed (18

up and 15 down) after TEV cleavage of cohesin, a significantly larger number than expected by chance (2.7%, Fisher's exact test: $p=1.19^{-07}$) (Figure 2A). Plotting the changes in gene expression after cohesin cleavage for Ecdysone-responsive genes versus all genes confirmed that Ecdysone-responsive genes are preferentially up- or downregulated after cohesin removal (Siegel-Tukey test: $p=1.04e^{-28}$) (Figure 2B), suggesting that cohesin plays a so far unrecognized role as mediator of the transcriptional response to Ecdysone in larval salivary glands.

In support of a direct (versus indirect) regulatory role for cohesin in the Ecdysone response, our statistical analysis revealed that the TSSs of 23 out of the 33 Ecdysone-responsive genes whose expression changed following cohesin cleavage localized within a cohesin domain (see Figures 2C and S2A/B for examples). This is a significantly (Fisher's exact test: $p=2.03e^{-03}$) larger number than expected by chance, namely 69.7% versus 46.6% of all Ecdysone-responsive genes or versus 34.5% of all genes of which differential expression was measured (Figure 2D). In addition, calculation of the average Rad21 binding at TSSs confirmed that the TSSs of Ecdysone-responsive genes - and especially of the subset of genes that are differentially expressed after cohesin cleavage - are preferentially bound by Rad21 (Figure 2E).

Timed cohesin cleavage specifically in salivary glands

Changes in transcript levels ten hours after cohesin cleavage do not necessarily imply that cohesin directly regulates transcription even if cohesin is present at the locus in question. For example, members of the Ecdysone signaling gene family are induced sequentially by a pulse of the steroid hormone Ecdysone that initiates the larval-to-pupal transition. Because transcription of *EcR* is also reduced upon cohesin cleavage, it is possible that the effect of cohesin cleavage on the aforementioned genes is merely due to reduced levels of EcR protein. To distinguish primary from secondary effects, it was therefore essential to evaluate the kinetics of changes in transcript levels occur upon cohesin cleavage.

The heat shock system used to induce TEV in our original screen has a number of limitations for this purpose. First, early effects could be missed because strong heat shocks have a drastic effect on transcription. Second, cohesin re-appears on chromosomes between 15 and 20 hours after *hs-TEV* induction due to re-synthesis of $Rad21^{TEV}$ and degradation of TEV protease (Figure S1A, Figure 3B and data not shown), which precludes evaluation of long term effects of cohesin cleavage. Third, the heat shock promoter is transcribed in all larval tissues and effects on transcription in one tissue (in this case salivary glands) might in principle be caused by changes that had occurred in another. For example, cohesin cleavage would have drastic and pleiotropic effects on proliferating neuroblasts, muscle cell precursors, and imaginal disc cells.

To control both the timing and tissue specificity of cohesin cleavage, TEV protease with a nuclear localization signal was expressed from a transgene (*UAST-NLS-TEV*) whose promoter contained multiple Gal4 binding sites [3]. Tissue-specificity was conferred by a second transgene *F4-Gal4* [26] that produces the Gal4 transcriptional activator protein from a salivary gland-specific "driver". For

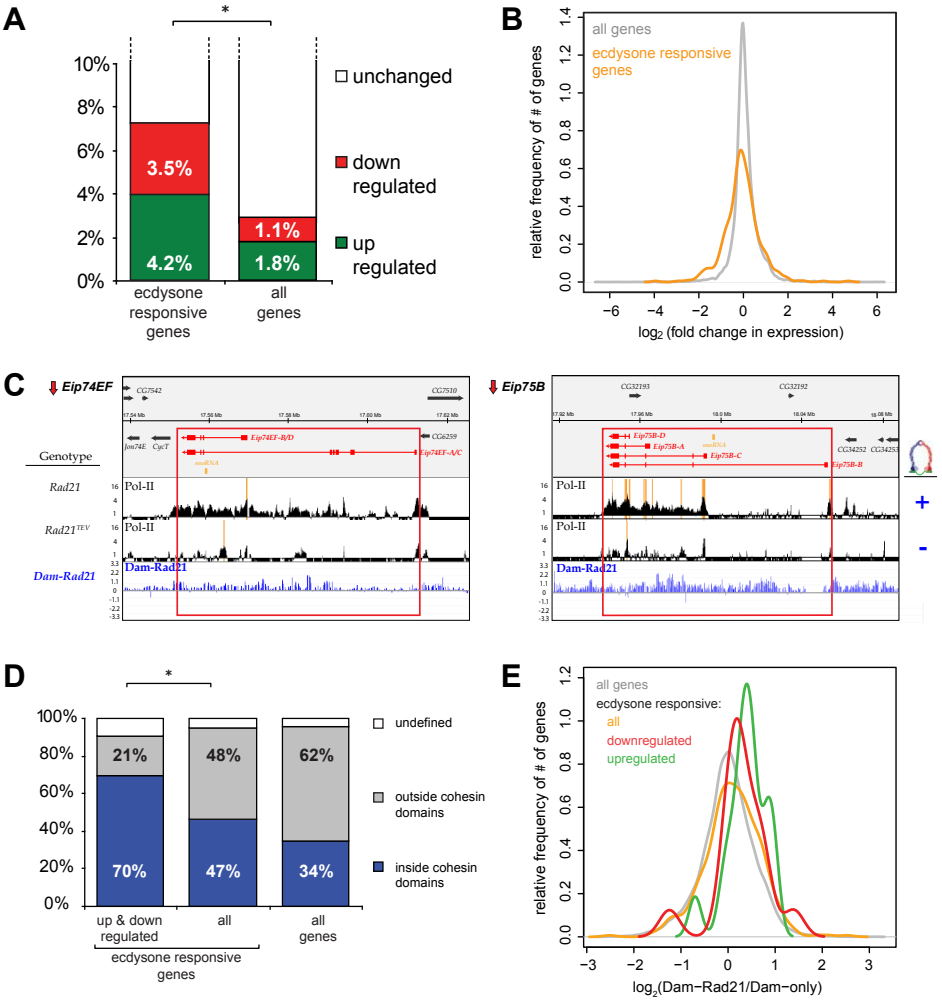


Figure 2 Cohesin regulates the expression of Ecdysone-responsive genes. Differential gene expression in salivary glands after cohesin cleavage was assessed at Ecdysone-responsive genes [25]. (A) Percentages of Ecdysone-responsive and of all genes that are upregulated (green), downregulated (red) and unchanged (white) after TEV cleavage of cohesin. The asterisk (*) indicates Fisher's exact test: $p < 0.01$. (B) $\log_2$-fold changes in gene expression after TEV cleavage of cohesin for Ecdysone- responsive genes (yellow) versus all genes (grey). (C) Pol-II and Rad21-binding profiles at *Eip74EF* and *Eip75B* are shown as representative examples for Ecdysone-responsive loci (see Figure legend 1(D) for further details; see also Figure S2). (D) Ecdysone-responsive genes that are differentially expressed upon cohesin cleavage are enriched in Rad21 binding. Shown are percentages of transcriptional start sites (TSSs) of Ecdysone-responsive cohesin-dependent genes, of all Ecdysone-responsive genes and of all genes that localized inside (blue), outside (grey) or at the border (white) of Rad21 bound regions. The asterisk (*) indicates Fisher's exact test: $p < 0.01$. (E) Average Rad21 binding at the TSSs of four different categories of genes: Ecdysone-responsive genes that are upregulated (green) or downregulated (red) after cohesin cleavage, all Ecdysone-reponsive genes and all genes (grey).

clarity, we will refer to the combination of *UAST-NLS-TEV* and *F4-Gal* as *SG-TEV*. Lastly, temporal control of transcription was conferred by a third transgene *tubGal80^{ts}* that expresses (ubiquitously) the temperature sensitive Gal80 protein, which binds to and inhibits Gal4 at 18°C (the permissive temperature) but not at 30°C (the restrictive temperature) [27]. Though complex, this system ensures that TEV is only expressed in salivary glands upon transfer of larvae to the restrictive temperature (see Figure 3A). Importantly, normal development and the transcriptional programs that underlie it continue after larvae are shifted permanently to 30°C and it should therefore be possible to evaluate both short and long term effects of tissue specific cohesin cleavage (see Figure 3B).

Salivary gland development occurs normally at 18°C in both *Rad21^{TEV}* and control (*Rad21*) larvae containing *SG-TEV/tubGal80^{ts}*, demonstrating efficacy of *tubGal80^{ts}* in repressing *SG-TEV* at this temperature. In contrast, transfer of animals to 30°C during early larval stages blocks salivary gland growth in *Rad21^{TEV}* but not in *Rad21* larvae (data not shown), confirming that cohesin has an essential role during salivary gland endocycles [3] despite the lack of chromosome segregation. To address cohesin's role in salivary glands that have completed their endocycles, late third instar larvae were shifted from 18°C to 30°C. Western blots of salivary gland extracts showed that TEV is undetectable prior to the temperature shift, accumulates within four hours of the shift, and remains at high levels thereafter (Figure 3C). Accumulation of TEV was accompanied by a permanent decline in TEV-cleavable Rad21 but not wild type Rad21. Importantly, neither TEV protease nor Rad21^{TEV} cleavage fragments could be detected in *Rad21^{TEV}* larvae from which salivary glands had been removed (data not shown), confirming the tissue-specificity of *F4-Gal4*.

Chromosome spreads showed that Rad21^{TEV} but not wild type Rad21 disappeared from polytene chromosomes within 2-4 hours of the temperature shift (Figs. 3D and 5A). The sustained removal of cohesin, only possible using the *SG-TEV/tubGal80^{ts}* system, revealed a number of post-transcriptional, most likely stress-induced alterations at late (> 24 hours) time points, e.g. increase in actin protein levels, clipping of histone H3 [28], and the dramatic appearance of what appears to be a post-translationally modified version of the large subunit of Pol-II (Figure 3C).

Controlled salivary gland-specific cohesin cleavage reveals both rapid and slow changes in gene expression

To identify genes whose expression is altered before such pleiotropic changes in cell physiology take place and which are therefore good candidates for being directly regulated by cohesin, we used the *SG-TEV/tubGal80^{ts}* system combined with qRT-PCR analysis to measure the transcript levels of a subset of cohesin-dependent genes in salivary glands of late third instar larvae (staged upon collection) over a 48-hour period following induction of cohesin cleavage. All mRNA levels were normalized using tubulin mRNA, which did not alter upon cohesin cleavage using the heat shock system. Figure 4 plots the ratios of *Rad21^{TEV}* and *Rad21* mRNA

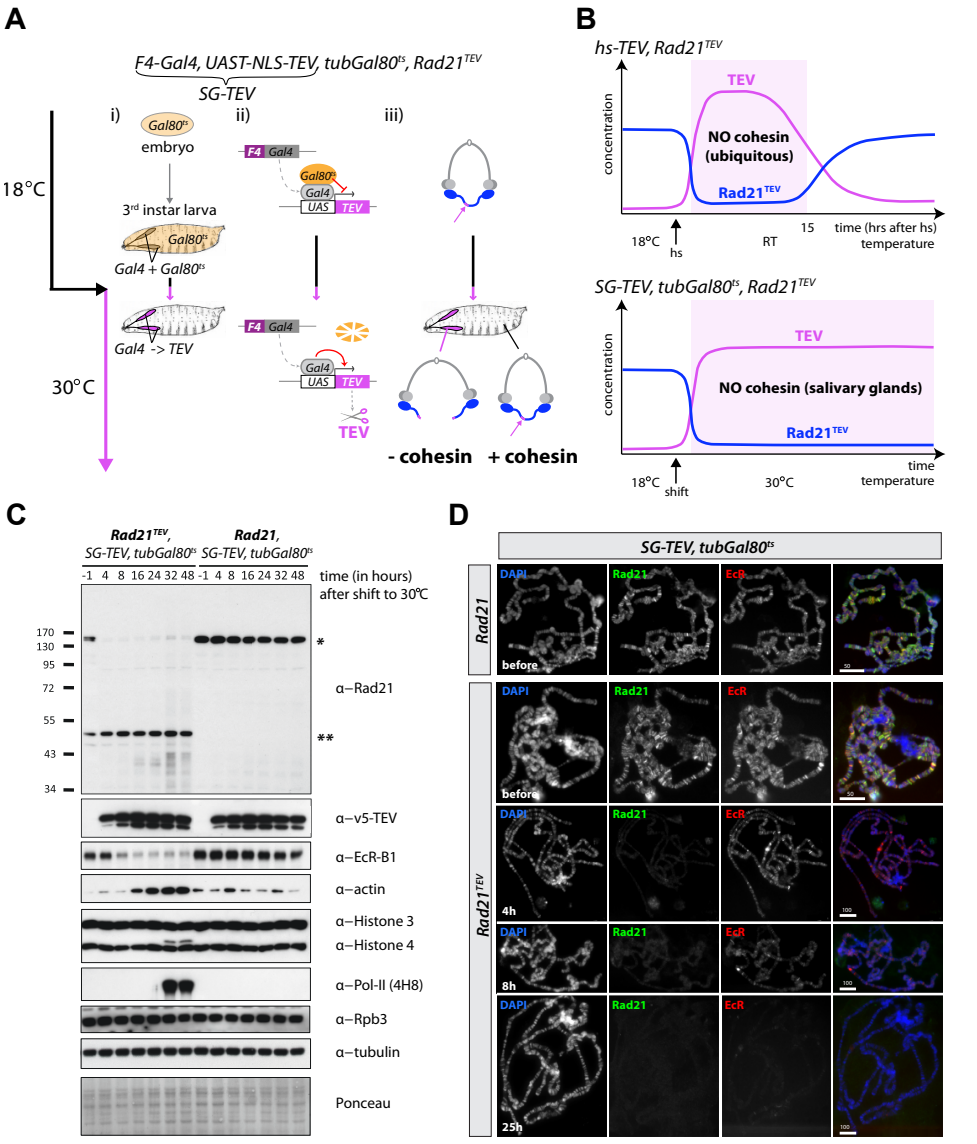


Figure 3 Timed salivary gland-specific cleavage of cohesion. (A) Outline of the tubGal80^{ts}/SG-TEV system that enables control of timing and salivary-gland specificity of cohesin cleavage. Larvae surviving on *Rad21^{TEV}* and encoding *tubGal80^{ts}* and *SG-TEV* (*F4-Gal4, UAST-NLS-TEV*) were raised at 18°C, the permissive temperature for the ubiquitously expressed Gal4-inhibitor Gal80^{ts}. By shifting late third instar larvae to the restrictive temperature (30°C), Gal80^{ts} is degraded, which enables the salivary gland-specific Gal4-driver F4-Gal4 to induce TEV protease expression specifically in salivary glands (i). (ii) and (iii) summarize the states of transcription and cohesin, respectively, in salivary glands at 18°C and 30°C. Rad21 is shown in blue, TEV in purple, Gal80^{ts} in orange. (B) Schematic comparison of the effects of the hs-TEV versus SG-TEV/tubGal80^{ts} systems on the level of functional cohesin complexes (presence of uncleaved Rad21^{TEV}) after induction of TEV protease. Concentrations of TEV (purple) and Rad21^{TEV} (blue) are plotted against time. While Rad21^{TEV} starts to re-accumulate in the hs-TEV system about 15 hours after TEV induction due to degradation of TEV and ▶

levels following TEV induction. Importantly, the ratio for a cohesin-independent gene *Rpll215*, which encodes the large subunit of Pol-II, remained close to one at all time points (Figure 4). Of six transcripts down-regulated by cohesin cleavage, three (*EcR-B1*, *Eip74EF*, *comm2*) declined to minimal levels within four hours of the temperature shift while three (*Sgs1, Eip75B, CG31698*) declined more gradually reaching minimal levels only after 16 hours (Figure 4, black lines). Up-regulated genes could also be divided into early and late response categories. Transcript levels of *wbl* and *CG12214* showed negligible changes during the first four hours and increased to maximal levels only after 16 hours while *ush* (and to a lesser extent *Mst87F*) increased within four hours. Time-courses of mRNA levels following heat shock induced cohesin cleavage confirmed the pattern of early and late responses (Figure 4, grey lines), at least over the first 16 hours (using this system analysis beyond 16 hours was not possible due to recovery of full-length Rad21^{TEV} protein - Figure S1A). Notably, the TSSs of all five early response genes (3 down and 2 up) that were identified by our time-course analysis are located inside a Rad21-bound domain (data not shown), which strongly suggests that they are indeed directly regulated by cohesin.

Rapid reduction of *Eip74EF* mRNA is not due to loss of Ecdysone Receptor

The decline of Ecdysone-regulated genes could be partly or wholly an indirect effect caused by the rapid decline of *EcR-B1* mRNA levels following cohesin cleavage. To address this, we used Western blots to measure the level Edysone Receptor protein and polytene chromosome spreads to measure its association with specific loci after shifting *SG-TEV/tubGal80^{ts}* larvae to 30°C. This revealed little or no change in protein levels or chromosomal association of EcR-B1 during the first four hours (Figure 3C and D). The ten-fold decrease in *Eip74EF* transcript levels within this period (Figure 4) cannot therefore be attributed to a lack of the hormone receptor. A lack of receptor could be responsible for the steep decline between 8 and 16 hours of mRNAs from the *Sgs1* locus, which encodes an Ecdysone-inducible salivary gland-specific glue protein and at which we detect only background levels of cohesin (Figure 4 and Figure S3A). A decline in EcR might also contribute to the gradual decrease of *Eip75B* between 4 and 16 hours (Figure 4).

re-synthesis of Rad21^{TEV} (top), continuous expression of TEV in salivary glands at 30°C prevents re-appearance of full-length Rad21^{TEV} in this tissue in the SG-TEV/tubGal80^{ts} system (bottom). (C) Western blot analysis of salivary gland extracts from *SG-TEV, tubGal80^{ts}* larvae surviving either on Rad21^{TEV} or Rad21. Extracts were prepared before (t = -1; TEV off) and at different time points after shifting third instar larvae to 30°C (TEV on). Blots were probed with the indicated antibodies. Full-length Rad21 (*) and the C-terminal TEV cleavage fragment (**) as well as full-length Histone H3 (<) and N-terminally clipped Histone H3 (<<), are indicated. Tubulin and Ponceau stainings were used as loading controls. (D) Representative polytene chromosome spreads from *SG-TEV, tubGal80^{ts}* crawling third instar larvae surviving either on Rad21 or Rad21^{TEV} were prepared before (TEV off) and at various times after shifting third instar larvae to 30°C (TEV on). Polytene chromosomes were co-immunostained with antibodies against Rad21 and EcR-B1. The chromosome morphology was visualized by DAPI staining. In the overlay (right panels), DAPI is shown in blue, Rad21 in green and EcR-B1 in red. Scale bars are 50 µm (top two rows) and 100 µm (bottom three rows).

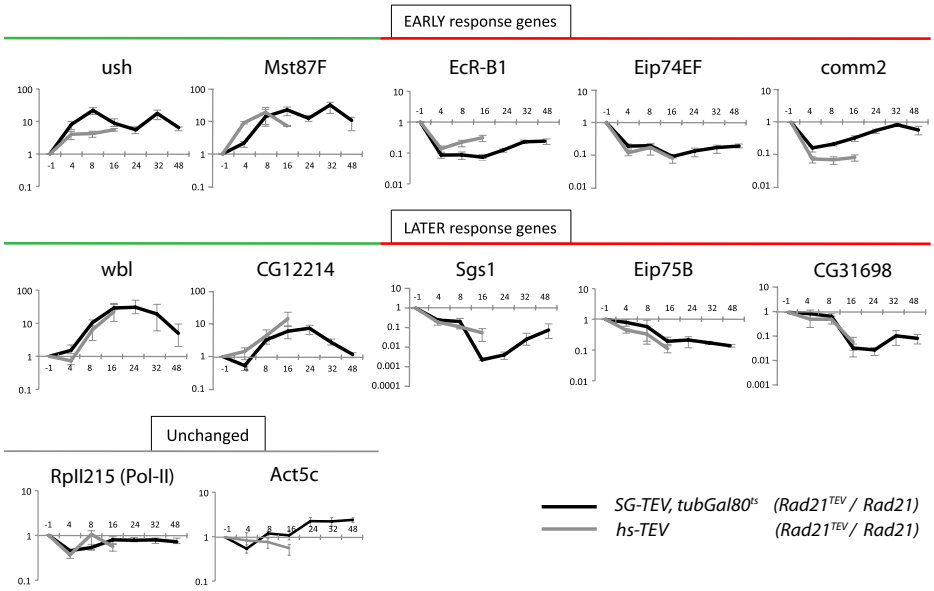


Figure 4 Loss of cohesin in salivary glands causes both rapid and slow changes in transcript levels. The kinetics of transcriptional changes upon cleavage of cohesin in salivary glands were assessed by qRT-PCR analysis, using the *SG-TEV/tubGal80^{ts}* (black lines) and *hs-TEV* (grey lines) systems. Plots show tubulin-normalized, averaged fold differences in transcript levels in the absence (Rad21^{TEV}) versus presence (Rad21) of cohesin over time (in hours, t = -1 indicates time point before TEV induction), obtained from three independent time-courses per TEV-expression system (error bars show standard deviations). Genes were classified as "early" and "late" response genes based on rapid or gradual changes in transcript levels within the first 4 or 16 hours, respectively. Green indicates upregulation, red downregulation, grey no change in the absence of cohesin.

Cohesin is required for puffing at Eip74EF

Eip74EF and *Eip75B* belong to a group of genes whose activity can be visualized cytologically on polytene chromosomes in late third instar larvae [29]. High rates of transcription induced by the late third instar peak of Ecdysone cause a characteristic decondensation or "puffing" of these loci, which happen to be located at adjacent bands, namely 74 and 75. Because puffing at band 75 occurs within 5-10 minutes of puffing at band 74, creating highly characteristic "twin" puffs at this stage of development [21], the 74/75 region is particularly easy to locate on polytene spreads (Figure 5A; see also model in Figure 5D) and enabled us to monitor reliably puffing at both loci within the same spread.

Polytene chromosome spreads were prepared from *Rad21* (control, + cohesin) and *Rad21^{TEV}* (- cohesin) late third instar larvae before and after salivary gland-specific TEV cleavage of cohesin. To monitor association of cohesin and Ecdysone Receptor with particular loci along chromosomes, spreads were co-immunostained with antibodies against Rad21 and EcR-B1. Plotting of normalized puff sizes at bands 74 and 75 (see methods) along y and x axes, respectively, confirmed that

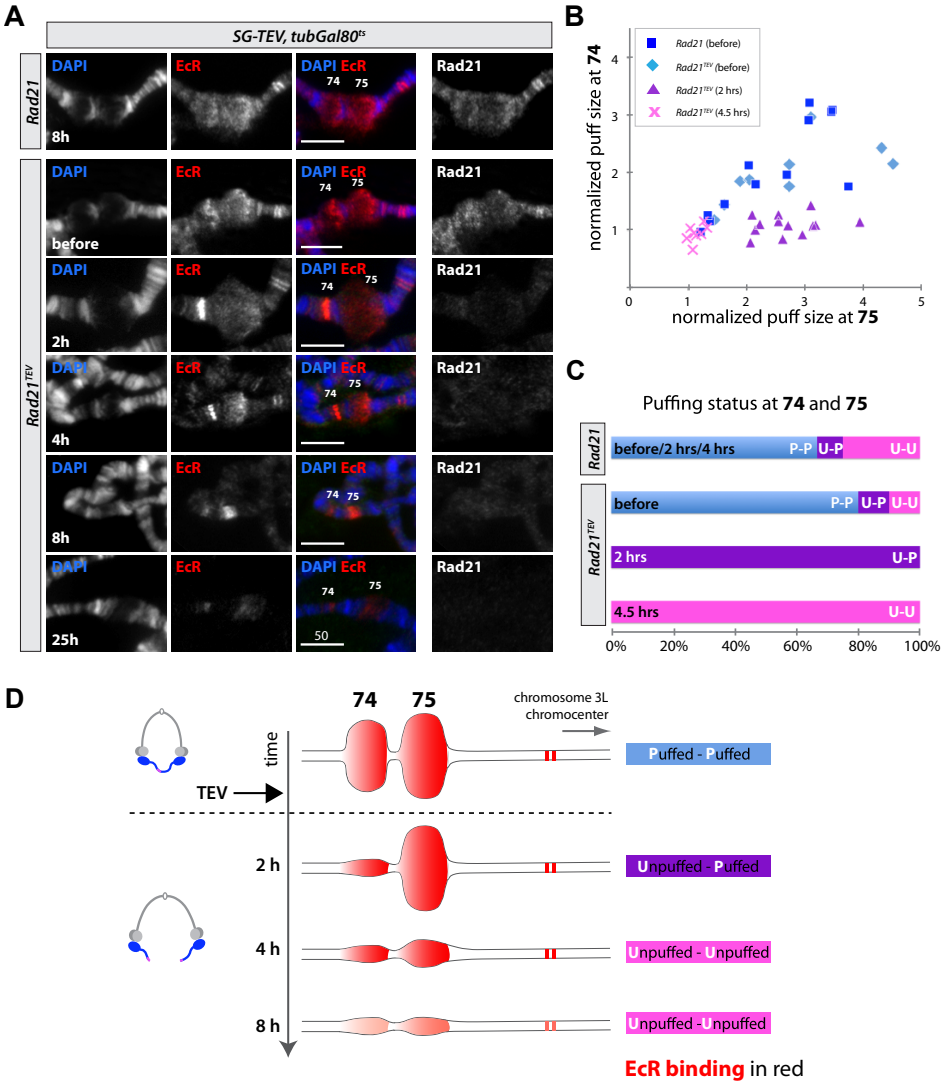


Figure 5 Cohesin is required for puffing at *Eip74EF* and *Eip75B*. The Ecdysone-induced late third instar puffing response at the neighbouring chromosomal bands 74 and 75 (harbouring *Eip74EF* and *Eip75B*, respectively) was analysed by polytene chromosome spreads before and after salivary gland-specific TEV protease induction in *SG-TEV/tubGal80^{ts}* late third instar larvae surviving either on Rad21^{TEV} or Rad21. (A) Representative polytene chromosome spreads from salivary glands in the presence (*Rad21*$_{8h}$ and *Rad21^{TEV}*$_{before}$) and in the absence of cohesin (*Rad21^{TEV}*- later time points) were co-immunostained with antibodies against Rad21 and EcR-B1. The chromosome morphology was visualized by DAPI staining. Positions of bands 74 and 75 are indicated in overlays (DAPI in blue, EcR-B1 in red). The two adjacent sharp EcR-stained bands at the right side of band 75 (towards the chromocenter) were used to identify the 74/75 locus. Scale bars are 50 µm. (B) Quantification of normalized puff sizes (see Methods) at bands 74 and 75 in the presence of cohesin (blue squares and light blue diamonds) and at 2 hrs (purple triangles) and 4.5 hrs (pink crosses) after cohesin cleavage. Each data point represents the normalized puff size at band 74 plotted against the normalized puff size at band 75 from the same polytene chromosome spread. (C) Changes in puffing status at bands 74 and 75 upon cleavage of cohesin. Based on the ratio and absolute sizes of the 74/75 puffs, each twin locus was classified as either puffed-puffed (P-P, blue) with both 74 and 75 decondensed, unpuffed-puffed (U-P, purple) with 74 condensed and 75 decondensed, or unpuffed-unpuffed (U-U, pink) (puffed loci: normalized width >1.5). The graph plots the percentage of spreads belonging to each category for each experimental condition. (D) The scheme illustrates the distinct puffing stages at 74/75 observed before and after TEV cleavage of cohesin. EcR binding to 74 and 75 as well as to the two adjacent bands towards the chromocenter is highlighted in red.

there was no significant difference in the puffing pattern of *Rad21* and *Rad21*^{TEV} salivary glands before TEV protease induction (Figure 5B; blue symbols). Even though there was considerable variation in the size of puffs between spreads (due presumably to small differences in the developmental state between larvae), puff size at 74 correlated with puff size at 75 within individual chromosomes. The constant ratio of normalized puff sizes at 74 and 75 is consistent with simultaneous induction and regression of these twin puffs [21]. Immunostaining showed that both puffs were enriched for EcR-B1 (Figure 5A), as has been previously described [30]. Both puffs were also enriched for Rad21, which is consistent with our DamID binding profile for cohesin at theses loci (Figs. 2C and S2B).

Neither puffing nor EcR-B1- and Rad21-binding patterns changed upon shifting *Rad21* (control) animals to 30°C (data not shown). In contrast, TEV protease induction caused rapid and major alterations to the 74/75 chromosomal region in *Rad21*^{TEV} salivary glands. Thus, Rad21 staining at both loci (as well as at all chromosomal loci, data not shown) declined to background levels within two hours, implying complete cleavage and dissociation of cohesin within this short time period (Figure 5A). Strikingly, cohesin cleavage was accompanied by a dramatic reduction of puffing at band 74. Because high levels of EcR-B1 persisted at this locus during this period, a sharp intense band of EcR-B1 staining was created at 74 (Figure 5A; 2h panels). Puffs at band 75, on the other hand, were little affected two hours after the temperature shift, with the result that there was a dramatic reduction in the ratio of puff sizes at 74 and 75 within chromosomes sampled at the two hour time point (Figure 5A and 5B - purple triangles; see also Figure 5C for averaged quantification of the puffing status at 74 and 75). The contraction of band 74 but not band 75 so soon after cohesin's removal without any detectable loss of EcR suggests that cohesin plays a direct role in transcribing the former. Though band 75 remained in a puffed state in the absence of cohesin for a longer period than band 74, it too declined by four hours, notably before any noticeable decline in EcR-B1 associated with the locus (Figure 5A and 5B – pink crosses; see also quantification in Figure 5C). Transcription from band 75 might therefore require cohesin activity at this locus as well as EcR protein.

| DISCUSSION

Development of a method to cleave Rad21 with TEV protease in a time and tissue specific manner has enabled us to assess the immediate and long-term consequences of cohesin inactivation on transcription in third instar salivary glands from *D. melanogaster*. We chose this post-mitotic tissue to ensure that any effects of cohesin inactivation on the transcriptional apparatus could not be attributed to indirect or "knock on" effects of chromosome missegregation or defective DNA repair due to the absence of cohesin's canonical function, namely sister chromatid cohesion. Despite this precaution, cohesin cleavage causes, from 24 hours onwards,

major changes in cellular physiology, some of which most likely reflect a general stress-related response (see Figure 3C). We cannot at this stage ascertain whether these highly pleiotropic events are triggered by changes in gene expression that precede them or by the loss of a novel cohesin function that we are currently unaware of. In either case, our observations demonstrate that it is very difficult to attribute functions to cohesin in regulating gene expression merely by observing the long-term consequences of its inactivation. Changes in gene expression that only occur 24 or more hours after cohesin's removal from chromosomes could be secondary or tertiary events triggered by fundamental changes in cell physiology.

Our observations reveal an additional complication in interpreting gene expression changes. Several of the genes whose expression is affected by cohesin cleavage are genes regulated by the Ecdysone Receptor, whose abundance declines after 8 hours due presumably to an almost immediate, cohesin cleavage-dependent decline in its mRNA. Thus, the precipitous decline in Sgs1 mRNAs that takes place between 8 and 16 hours could be caused by the lack of Ecdysone Receptor and not by the lack of cohesin per se. Such phenomena could explain many late responses to cohesin inactivation.

Given these considerations, it is clear that in order to attribute a role for cohesin in regulating a gene on the basis of changes in its expression upon cohesin inactivation, it is necessary to demonstrate a change in transcription as soon as cohesin has been removed from chromosomes and, crucially, long before any major change in cell physiology or in the concentration of other transcription regulators. Two genes stand out in this regard, namely *EcR* encoding the Ecdysone Receptor and *Eip74EF* encoding an Ecdysone-dependent transcription factor. *Eip74EF* is a particularly good candidate as heavy transcription of this gene in third instar larvae gives rise to a cytologically visible puff. Cohesin is associated with this puff and its removal by Rad21 cleavage is accompanied by an immediate cessation of puffing. Crucially, contraction of band 74 caused by Rad21's removal takes place several hours before any decline in Ecdysone Receptor associated with it. We suggest therefore that cohesin present at *Eip74EF* has a direct role in maintaining transcription of the gene. We have no reason to believe that the same is not also true for *EcR*, though we have not observed it at a cytological level. Because transcription of most genes is unaffected by cohesin cleavage, it is striking that transcription of *EcR* as well as a direct target gene, *Eip74EF*, appear both to be directly regulated by cohesin. Our finding that Ecdysone-responsive genes in general are enriched in cohesin domains and preferentially mis-regulated following cohesin cleavage suggests a common aspect of the transcription process at these loci that render them particularly dependent on cohesin. It is conceivable that the interplay between the core set of gene regulatory mechanisms (transcription factors, enhancers, promoters etc) was insufficient to achieve the precise control that was required to orchestrate the dramatic Ecdysone-induced changes that occur during the larval-to-pupal metamorphosis. It is also conceivable that cohesin, due to its ability to encircle chromatin strands,

was particularly suited to fulfill this role, either by facilitating interactions between distant DNA-elements in cis or by its ability to slide along DNA (see below).

While *Eip74EF* may be the best example of a gene directly regulated by cohesin, it is by no means the only candidate. Reduced puffing at its twin, the adjacent *Eip75B*, also occurs before any obvious decline in Ecdysone Receptor at this locus. While the drop in *Eip75B* mRNAs that occurs 8 hours after induction of Rad21 cleavage may be due to a decline in Ecdysone Receptor, the more modest decrease that occurs earlier may be due to a direct effect of cohesin's dissociation from the locus. There are other genes, for example *comm2*, whose mRNAs decline rapidly upon cohesin cleavage and these also may be directly regulated by cohesin. Interestingly, transcripts from at least two genes, namely *ush* and *Mst87F*, rise rapidly after cohesin cleavage, suggesting that while cohesin promotes transcription at certain genes it exerts repression at others.

Cohesin's canonical function is to mediate sister chromatid cohesion. It is currently thought to perform this by entrapping sister DNAs inside a tripartite ring formed by its Smc1, Smc3, and Rad21/Scc1 subunits [1]. This raises the important question whether cohesin regulates gene expression using a similar topological principle. With this in mind it has been repeatedly proposed that cohesin might regulate gene expression by facilitating the formation or maintenance of loops between remote regulatory elements and promoter regions. Such loops have not been visualized directly but have instead been inferred from co-precipitation of remote DNA sequences following formaldehyde fixation. According to this somewhat indirect assay, long term cohesin depletion reduces interaction between an enhancer at the 3' end of the *H19* gene with a remote CTCF binding site that controls imprinting of the *IGF2-H19* locus [31]. Loss of the putative loop between the CTCF binding site and the *H19* enhancer is thought to enable the enhancer now to activate the neighboring *IGF2* gene. Cohesin depletion also disrupts a similar type of long-range interaction between distant (cohesin-associated) CTCF sites at the *INFG* locus, though in this case, cohesin depletion has little effect on inducibility of the locus by cytokine [5]. The observation that cohesin in *Drosophila* is – unlike to its enrichment at CTCF binding sites in human cells [2, 32], associated with large domains raises the possibility that it can also regulate transcription by means other than formation of loops between remote regulatory elements. By entrapping DNAs inside rings capable of sliding along chromatin, cohesin complexes may provide a potentially mobile platform for the stable association of other factors necessary for regulating (positively or negatively) the movement of polymerases through transcription units. Cohesin's intriguing potential to modulate chromatin, together with its binding to regions covering several transcription units are seemingly at odds with our finding that differentially expressed genes are not clustered in the genome. Whatever the activity is that cohesin brings along, our data suggests that its absence affects only a subset of genes that are normally exposed to it. Our identification of Ecdysone-responsive

genes as a class of cohesin-dependent genes highlights that there might exist still unknown common determinants or gene-specific regulators that render a gene susceptible to changes in cohesin binding.

| EXPERIMENTAL PROCEDURES

Fly strains

Flies surviving on myc_{10}-tagged wildtype Rad21 (*Rad21*: w^{1118}; +/+; $Rad21^{ex3}$, $P[w^+$, *tubpr<Rad21-*myc_{10}*<SV40])* or myc_{10}-tagged TEV-cleavable Rad21 (*Rad21*^{TEV}: w^{1118}; +/+; $Rad21^{ex15}$, $P[w^+$, *tubpr<Rad21(550-3TEV)-*myc_{10}*<SV40])* and heat-inducible TEV-protease expressing flies (*hs-TEV*: w^{1118}; *hs-NLS-v5-TEV-*NLS_2; $Rad21^{ex3}$*/TM6B Tb ubiquitin-GFP*) have been described previously [3]. For salivary gland specific TEV-cleavage, the $tubGal80^{ts}$ transgene [33] was recombined with the *Rad21*-null mutant $Rad21ex^{15}$ and the nuclear *UAST-NLS-TEV* transgene [3] and crossed to the F4-Gal4 driver line [26] to generate w^{1118}; *F4-Gal4*; $tubGal80^{ts}$ $Rad21ex^{15}$ *UAST-NLS-TEV/Tm6B Tb ubiquitin-GFP* flies. For DamID, flies carrying *5xUAST-Dam-myc-Rad21* (*Dam-Rad21*) were generated by standard P-element-mediated transgenesis (BestGene). Flies carrying *5xUAST-Dam* (*Dam-only*) have been published before [34]. For details on the cloning see Supplemental Experimental Procedures (available upon request). To test for the functionality of *Dam-myc-Rad21* constructs, rescue experiments were performed after TEV-cleavage of Rad21^{TEV} in syncytial embryos. Embryo preparation, synthesis of mRNA coding for wildtype and Dam-myc-tagged Rad21, and mRNA/TEV protease injections were performed as previously described [20]. A complete list of genotypes of all fly strains used in this study can be found in (available upon request) Supplemental Experimental Procedures.

Gene expression profiling

Virgin female flies expressing Rad21 with (*Rad21*^{TEV}) or without (transgenic or endogenous *Rad21*) TEV-cleavage sites as their sole source of Rad21 were crossed to male flies carrying heat shock inducible TEV-protease in a *Rad21*-null background (*hs-TEV* flies). Crosses were kept at 18°C under non-crowded conditions. TEV protease expression was induced in late third instar Tb^- GFP^- larvae by heat-shock (45 minutes in a 36.5°C water bath, followed by 10-12 hours incubation at RT). For each of the 7 microarray experiments, 10-20 salivary gland pairs of *Rad21*^{TEV} (-cohesin) and *Rad21* (+ cohesin) crawling third instar larvae (staged upon collection) were dissected and total RNA was isolated using Trizol® Reagent (Invitrogen) according to manufacturer's instructions. cDNA preparation, Cy3- and Cy5-labeling of the sample pairs, hybridization to *Drosophila* long oligonucleotide cDNA arrays (FL002), array scanning, normalization and basic statistical analysis (Bioconductor package, Limma) were performed at the FlyChip facility in Cambridge, UK. Data is presented as vsn-normalized $\log_2$ ratio ($\log_2$(-cohesin)/(+cohesin)).

qRT-PCR analysis of selected candidate genes was performed according

to standard procedures. For details see Supplemental Experimental Procedures (available upon request).

DamID analysis of cohesin binding in salivary glands

Genomic DNA was isolated from salivary glands of homozygous *Dam-Rad21* or *Dam-only* crawling third instar larvae. *In vivo* methylated DNA was amplified as described before [22]. Differentially labeled fragments of both samples were pooled and hybridized to microarrays carrying 380,000 60-mer DNA oligonucleotides [35] (Roche-NimbleGen) with a median probe spacing of 300 bp. Probes were mapped to *D.mel* Release 5 genome. Microarray data analysis was performed with R (http://www.r-project.org). Raw data was loess normalized, median centered and dye swap arrays were averaged. Rad21 domains were defined using a two-state Hidden Markov Model. Further details are available in Supplemental Experimental Procedures (available upon request). All downstream analyses were performed using custom made R-scripts, which are available upon request.

ChIP-CHIP analysis

Pol-II Chromatin Immunoprecipitations (Chromatin-IPs) of third instar larval salivary glands, using the CTD4H8 mouse anti-Pol-II antibody (Upstate), were performed according to [36] and [37] with minor modifications (see Supplemental Experimental Procedures for details, available upon request).

Immunolabeling and analysis of polytene chromosome squashes

Polytene chromosome spreads were prepared according to standard procedures and stained o.n. at 4°C with primary antibodies (gp-anti-Rad21 (1:500), mouse-anti-EcR-B1 (1:200)). Immuno-complexes were detected with Alexa-conjugated secondary antibodies and mounted using Vectashield mounting medium containing DAPI (Vector Laboratories). Flourescent images were acquired with an AXIO Imager.Z1 microscope (Zeiss) equipped with 40x and 63x EC Plan-Neofluar oil objectives and a CoolSNAP HQ CCD camera (Photometrics) using MetaMorph software (Universal Imaging).

To analyze puffing of bands 74 and 75, only chromosome spreads in which those chromosome loci could be identified unambiguously (based on the characteristic twin puffed morphology and neighbouring EcR double bands, see Figure 5D) were taken into consideration. The width of each puff was measured and normalized to the average width of 3 neighboring bands.

Western Blot analysis

Western Blot analysis was performed from dissected third instar larval salivary glands and whole larvae according to standard protocols. All antibodies used are listed in Supplemental Experimental Procedures (available upon request).

Data availability

Gene expression, DamID, and Pol-II ChIP-CHIP data have been deposited to NCBI's Gene Expression Omnibus and are accessible through GEO Series accession numbers GSE21844 (gene expression and Pol-II) and GSE21874 (DamID).

| ACKNOWLEDGEMENTS

We thank B. Edgar, J. Mellor and J. Mummery-Widmer (J. Knoblich lab) for reagents; Guillaume Filion for helping with the cohesin domain definition; B. Fischer (Fly-CHIP Cambridge) and Y. Katuo for help with microarray analysis and CHIP-CHIP experiments, respectively; the central microarray facility from the NKI for DamID-array hybridization; J. Metson, P. Guna and Wendy Talhout for technical assistance; and all members of the K.N. and B.v.S. laboratories for discussions and comments on the manuscript. A.P. currently holds an EMBO Long-Term Fellowship. R.A.O. holds a post-doctoral fellowship from the Fundação para a Ciência e a Tecnologia of Portugal. T.I. is supported by a Grant-in-Aid for Young Scientists (A) from the JSPS. Work in the laboratory of K.S. is supported by a grant of the Cell Innovation Program from the MEXT and Grant-in-Aid for Scientific Research (S) from the JSPS. Work in the laboratory of B.v.S. is supported by a Netherlands Genomics Initiative grant. Work in the laboratory of K.N. is supported by grants from Medical Research Council (MRC) and Wellcome Trust.

| REFERENCES

1. Nasmyth, K. and C. Haering, Cohesin: Its Roles and Mechanisms. Annu Rev Genet, 2009. 43: p. 525-558.
2. Wendt, K.S., et al., Cohesin mediates transcriptional insulation by CCCTC-binding factor. Nature, 2008. 451(7180): p. 796-801.
3. Pauli, A., et al., Cell-type-specific TEV protease cleavage reveals cohesin functions in Drosophila neurons. Developmental Cell, 2008. 14(2): p. 239-51.
4. Schuldiner, O., et al., piggyBac-based mosaic screen identifies a postmitotic function for cohesin in regulating developmental axon pruning. Developmental Cell, 2008. 14(2): p. 227-38.
5. Hadjur, S., et al., Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. Nature, 2009.
6. Schaaf, C.A., et al., Regulation of the Drosophila Enhancer of split and invected-engrailed gene complexes by sister chromatid cohesion proteins. PLoS ONE, 2009. 4(7): p. e6202.
7. Rollins, R.A., P. Morcillo, and D. Dorsett, Nipped-B, a Drosophila homologue of chromosomal adherins, participates in activation by remote enhancers in the cut and Ultrabithorax genes. Genetics, 1999. 152(2): p. 577-93.
8. Rollins, R., et al., Drosophila nipped-B protein supports sister chromatid cohesion and opposes the stromalin/Scc3 cohesion factor to facilitate long-range activation of the cut gene. Mol Cell Biol, 2004. 24(8): p. 3100-11.
9. Dorsett, D., et al., Effects of sister chromatid cohesion proteins on cut gene expression during wing development in Drosophila. Development, 2005. 132(21): p. 4743-53.
10. Horsfield, J., et al., Cohesin-dependent regulation of Runx genes. Development, 2007. 134(14): p. 2639-49.
11. Benard, C.Y., et al., mau-2 acts cell-autonomously to guide axonal migrations in Caenorhabditis elegans. Development, 2004. 131(23): p. 5947-58.
12. Seitan, V., et al., Metazoan Scc4 homologs link sister chromatid cohesion to cell and axon migration guidance. PLoS Biol, 2006. 4(8): p. e242.
13. Krantz, I.D., et al., Cornelia de Lange syndrome is caused by mutations in NIPBL, the human homolog of Drosophila melanogaster Nipped-B. Nat Genet, 2004. 36(6): p. 631-5.
14. Tonkin, E.T., et al., NIPBL, encoding a homolog of fungal Scc2-type sister chromatid cohesion proteins and fly Nipped-B, is mutated in Cornelia de Lange syndrome. Nat Genet, 2004. 36(6): p. 636-41.

15. Musio, A., et al., X-linked Cornelia de Lange syndrome owing to SMC1L1 mutations. Nat Genet, 2006. 38(5): p. 528-30.
16. Kawauchi, S., et al., Multiple organ system defects and transcriptional dysregulation in the Nipbl(+/-) mouse, a model of Cornelia de Lange Syndrome. PLoS Genetics, 2009. 5(9): p. e1000650.
17. Liu, J., et al., Transcriptional dysregulation in NIPBL and cohesin mutant human cells. PLoS Biol, 2009. 7(5): p. e1000119.
18. Uhlmann, F., et al., Cleavage of cohesin by the CD clan protease separin triggers anaphase in yeast. Cell, 2000. 103(3): p. 375-86.
19. Gruber, S., C. Haering, and K. Nasmyth, Chromosomal cohesin forms a ring. Cell, 2003. 112(6): p. 765-77.
20. Oliveira, R.A., et al., Cohesin cleavage and Cdk inhibition trigger formation of daughter nuclei. Nat Cell Biol, 2010. 12(2): p. 185-92.
21. Ashburner, M., Patterns of puffing activity in the salivary gland chromosomes of Drosophila. VI. Induction by ecdysone in salivary glands of D. melanogaster cultured in vitro. Chromosoma, 1972. 38(3): p. 255-81.
22. Greil, F., C. Moorman, and B. van Steensel, DamID: mapping of in vivo protein-genome interactions using tethered DNA adenine methyltransferase. Methods Enzymol, 2006. 410: p. 342-59.
23. van Steensel, B. and S. Henikoff, Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. Nat Biotechnol, 2000. 18(4): p. 424-8.
24. Misulovin, Z., et al., Association of cohesin and Nipped-B with transcriptionally active regions of the Drosophila melanogaster genome. Chromosoma, 2008: p. in press.
25. Beckstead, R.B., G. Lam, and C.S. Thummel, The genomic response to 20-hydroxyecdysone at the onset of Drosophila metamorphosis. Genome Biol, 2005. 6(12): p. R99.
26. Weiss, A., et al., Continuous Cyclin E expression inhibits progression through endoreduplication cycles in Drosophila. Curr Biol, 1998. 8(4): p. 239-42.
27. McGuire, S.E., et al., Spatiotemporal rescue of memory dysfunction in Drosophila. Science, 2003. 302(5651): p. 1765-8.
28. Santos-Rosa, H., et al., Histone H3 tail clipping regulates gene expression. Nat Struct Mol Biol, 2009. 16(1): p. 17-22.
29. Zhimulev, I.F., et al., Polytene chromosomes: 70 years of genetic research. Int Rev Cytol, 2004. 241: p. 203-75.
30. Yao, T.P., et al., Functional ecdysone receptor is the product of EcR and Ultraspiracle genes. Nature, 1993. 366(6454): p. 476-9.
31. Nativio, R., et al., Cohesin is required for higher-order chromatin conformation at the imprinted IGF2-H19 locus. PLoS Genet, 2009. 5(11): p. e1000739.
32. Parelho, V., et al., Cohesins functionally associate with CTCF on mammalian chromosome arms. Cell, 2008. 132(3): p. 422-33.
33. Buttitta, L.A., et al., A double-assurance mechanism controls cell cycle exit upon terminal differentiation in Drosophila. Dev Cell, 2007. 12(4): p. 631-43.
34. Vogel, M.J., et al., High-resolution mapping of heterochromatin redistribution in a Drosophila position-effect variegation model. Epigenetics Chromatin, 2009. 2(1): p. 1.
35. Choksi, S.P., et al., Prospero acts as a binary switch between self-renewal and differentiation in Drosophila neural stem cells. Dev Cell, 2006. 11(6): p. 775-89.
36. Adelman, K., et al., Efficient release from promoter-proximal stall sites requires transcript cleavage factor TFIIS. Mol Cell, 2005. 17(1): p. 103-12.
37. Yao, J., et al., Intranuclear distribution and local dynamics of RNA polymerase II during transcription activation. Mol Cell, 2007. 28(6): p. 978-90.

| SUPPLEMENTARY FIGURES

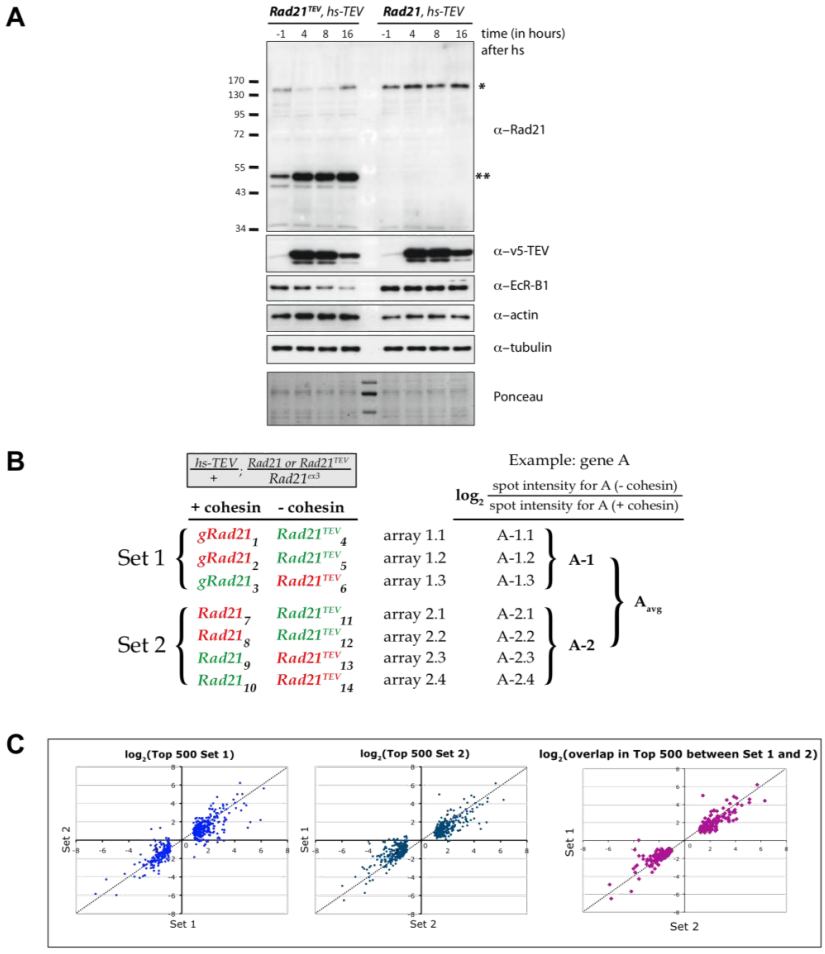


Figure S1 Differential gene expression profiling by microarrays after TEV cleavage of cohesion. (A) Cleavage of cohesin in salivary glands using the hs-TEV system. Western blot analysis of salivary gland extracts from *hs-TEV* crawling third instar larvae surviving either on Rad21TEV or Rad21. Extracts were prepared before (t = -1; TEV off) and at different time points after a 45 minute heat shock at 37°C (the end of heat shock was set as t = 0). Blots were probed with the indicated antibodies. Full-length Rad21 (*) and the C-terminal TEV cleavage fragment (**) are marked. Tubulin and Ponceau stainings were used as loading controls. (B) Overview of the seven individual microarray experiments and subsequent data analysis used for differential gene expression profiling of third instar larval salivary glands after heat shock-induced TEV cleavage of cohesin. Two sets of microarray experiments were performed, with three arrays in Set 1 (sample pairs *gRad21* and *Rad21TEV*) and four arrays in Set 2 (sample pairs *Rad21* and *Rad21TEV*). Cy3- (red) and Cy5 (green) labelled independent biological samples (numbered from 1 to 14) with (*gRad21* or *Rad21*) and without cohesin (*Rad21TEV*) were hybridized in pairs to seven arrays. Spot intensities for each locus were measured for each channel separately and used to calculate log2-ratios (spot intensity in the presence versus absence of cohesin). Log2-ratios were averaged across each set (e.g. values A-1 and A-2 for gene A) and across both sets (Aavg). (C) Scatter plots comparing log2-ratios of the Top 500 genes of the two sets of experiments. Classification of genes as Top 500 was based on Limma statistical analysis. In the top two graphs, the log2-ratios of the Top 500 genes of Set 1 (light blue, left) were plotted against the corresponding log2-ratios of Set 2 and vice versa (dark blue, right). In the bottom graph (purple), log2-ratios of the 227 genes common between the Top 500 of Set 1 and Set 2 were plotted against each other. The diagonal indicates a hypothetical perfect correlation between both sets of data.

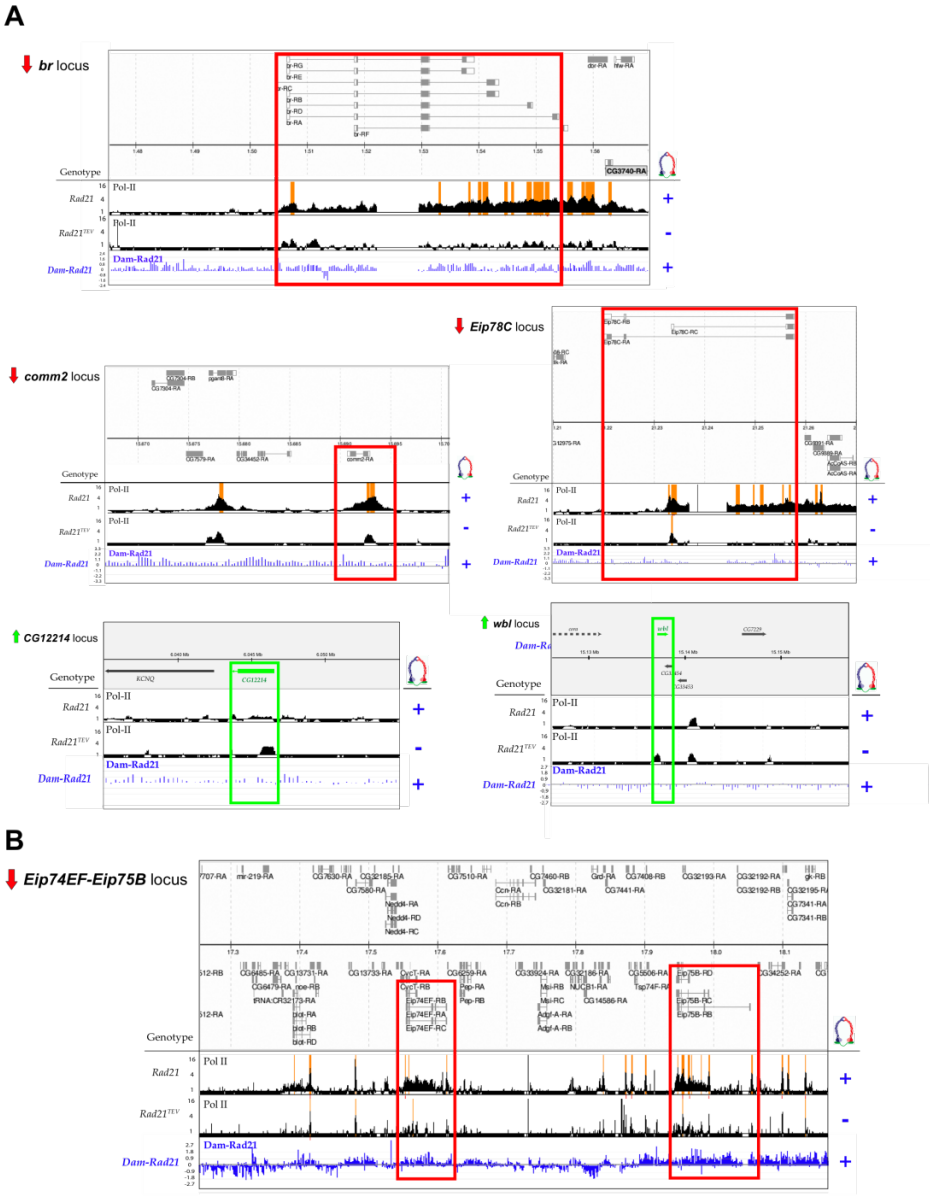


Figure S2 RNA Polymerase II (Pol-II) and cohesin binding profiles at differentially expressed loci. (A) Pol-II and cohesin binding at representative genomic loci whose expression changes upon cleavage of cohesin. ChIP-CHIP analysis was used to assess the distribution of Pol-II in *Rad21* (+ cohesin) and *Rad21TEV* (- cohesin) salivary glands 10-12 hours after heat shock induction of TEV protease (black plots). ChIP-CHIP data is represented as fold enrichment of IP over Input (MAT scores; logscale; highly enriched regions ($p < 0.0001$) are coloured in orange). Cohesin binding in salivary glands was assessed by DamID (Dam-Rad21; blue plots) and is represented as the relative enrichment of methyladenine marked DNA from Dam-Rad21 glands over Dam-only glands (log2 scale). (B) Pol-II and cohesin binding at the Ecdysone-regulated *Eip74EF* and *Eip75B* loci. For further details see (A).

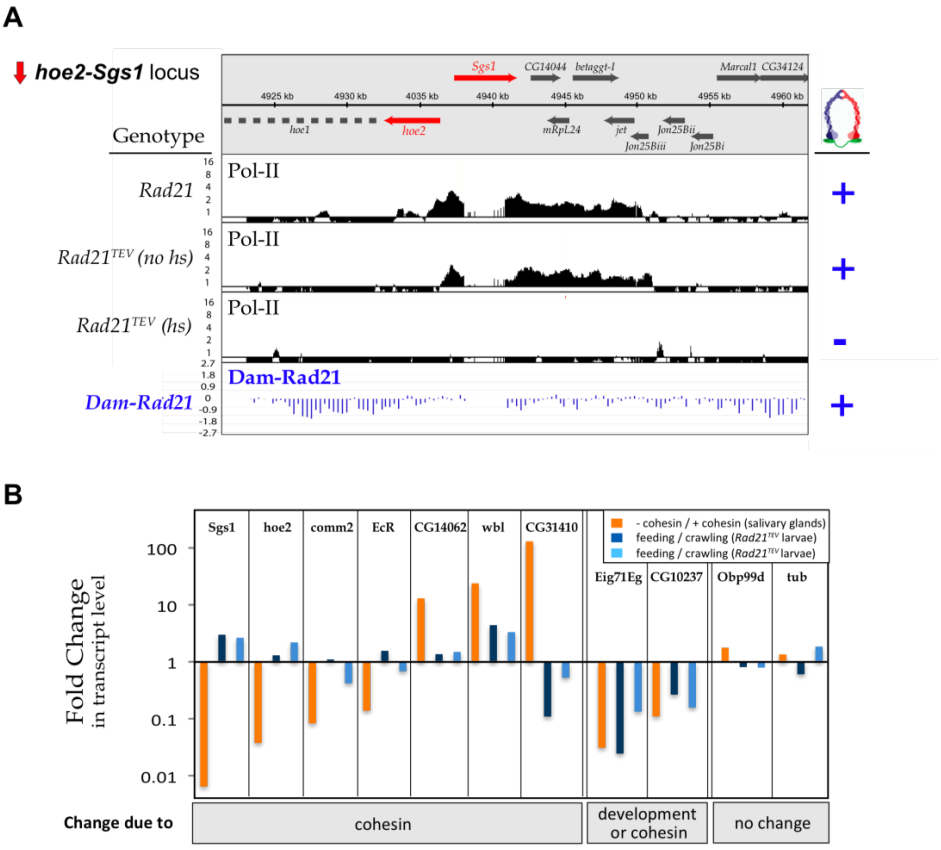


Figure S3 The majority of differences in gene expression is caused by loss of Cohesion. (A) RNA Polymerase II (Pol-II) and cohesin binding profiles at the downregulated *Sgs1-hoe2* locus. Experimental details are as described in Figure legend S2, except that the Pol-II distribution is also shown for *Rad21TEV* salivary glands carrying the uninduced *hs-TEV* transgene (no hs, + cohesin). Note that cohesin binds only at background levels to this locus (blue plot, DamID data). (B) Comparison of gene expression differences observed after cohesin cleavage in salivary glands with those in younger versus older larvae. Transcript levels of 10 candidate genes (tubulin served as non-differentially expressed control) were measured by qRT-PCR. For each locus, the fold-change in transcript level in the absence versus presence of cohesin (orange bars) was compared to the fold-change in transcript level in feeding (younger) versus crawling (older) third instar *Rad21TEV* (dark blue) or *Rad21* (light blue) larvae. Each value represents the average of at least two independent experiments.

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

Guillaume J. Filion*,[1], Joke G. van Bemmel*,[1], Ulrich Braunschweig*,[1], Wendy Talhout[1], Jop Kind[1], Lucas D. Ward[2,3,4], Wim Brugman[5], Ines de Castro Genebra de Jesus[1,6], Ron M. Kerkhoven[5], Harmen J. Bussemaker[2,3], Bas van Steensel[1]

Cell, 143 (2), 2010, 212-24

*These authors contributed equally

[1]Division of Gene Regulation, Netherlands Cancer Institute, Amsterdam, the Netherlands; [2]Department of Biological Sciences, Columbia University, New York, USA; [3]Center for Computational Biology and Bioinformatics, Columbia University, New York, USA; [4]Current address: Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, USA; [5]Central Microarray Facility, Netherlands Cancer Institute, Amsterdam, the Netherlands; [6]Current address: Genome Function Group, MRC Clinical Sciences Centre, Imperial College School of Medicine, London, United Kingdom

4

| ABSTRACT

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 repressive chromatin type that covers about half of the genome and lacks classic heterochromatin markers. Furthermore, transcriptionally active euchromatin consists of two types that differ in molecular organization and H3K36 methylation, and regulate distinct classes of genes. Finally, we provide evidence that the different chromatin types help to target DNA-binding factors to specific genomic regions. 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 [1-2]. 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 was divided into heterochromatin and euchromatin. There is now ample evidence that a finer classification is required. 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; it is a repressive type of chromatin that primarily regulates genes with developmental functions [3]. 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 [4]. Direct comparison of genome-wide binding maps indicates that PcG and HP1 heterochromatin are non-overlapping [5].

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 additional types of repressive chromatin that have remained unnoticed?

In order to address these questions we generated genome-wide location maps of 53 broadly selected chromatin proteins and four key histone modifications in *Drosophila* cells, providing a rich description of chromatin composition along the genome. By integrative computational analysis 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 type of repressive 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 embryonic *Drosophila melanogaster* cell line Kc167 (Figure 1A and Figure S1A). In order to obtain 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) (Table S1). 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 [6], 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 [7] and can be unreliable for proteins with short residence times [8-9]. 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, so that even transient interactions may be detected [10]. 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 (Figure S1B), which was also reported in previous comparative studies [11-12]. 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.

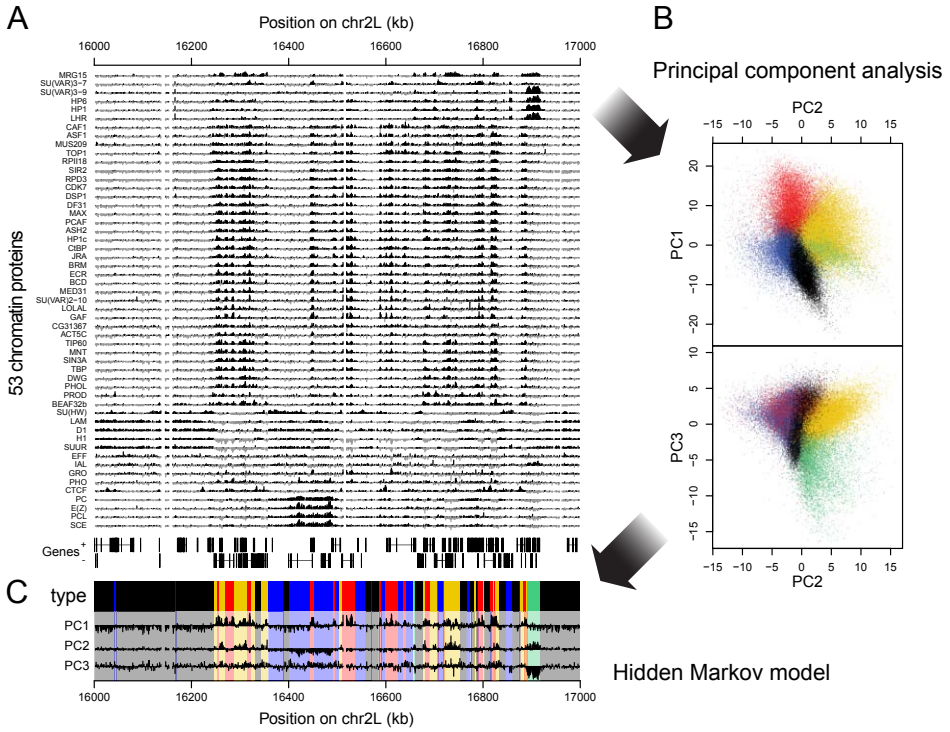


Figure 1 Overview of protein binding profiles and derivation of the 5-type chromatin segmentation. (A) Sample plot of all 53 DamID profiles (log_2 enrichment over Dam-only control). Positive values are plotted in black, negative values in grey for contrast. Below the profiles, genes on both strands are 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). See also Figure S1 and Table S1.

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 (Figure 1A). Nevertheless, several sets of proteins exhibit profiles that are similar. Some similarities were anticipated, such as for PC, PCL, SCE and E(Z), which are all PcG proteins [3]; and for HP1, SU(VAR)3-9, LHR and HP6, which are part of classic HP1-type heterochromatin [13]. 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 [14]. There is a prominent 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 Methods, available upon request). 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 are 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.

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 (Figure 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 (Figure S1C), showing that this result is robust to different quantification methods.

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, available upon request). To avoid semantic confusion, and in line with the Greek word *chroma* (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 (Figure 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 Figure S1D.

Domain organization of chromatin types

The five types of chromatin differ substantially in their genome coverage, numbers of domains, and numbers of genes (Figure 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 (Figure 2B). 441 domains are larger

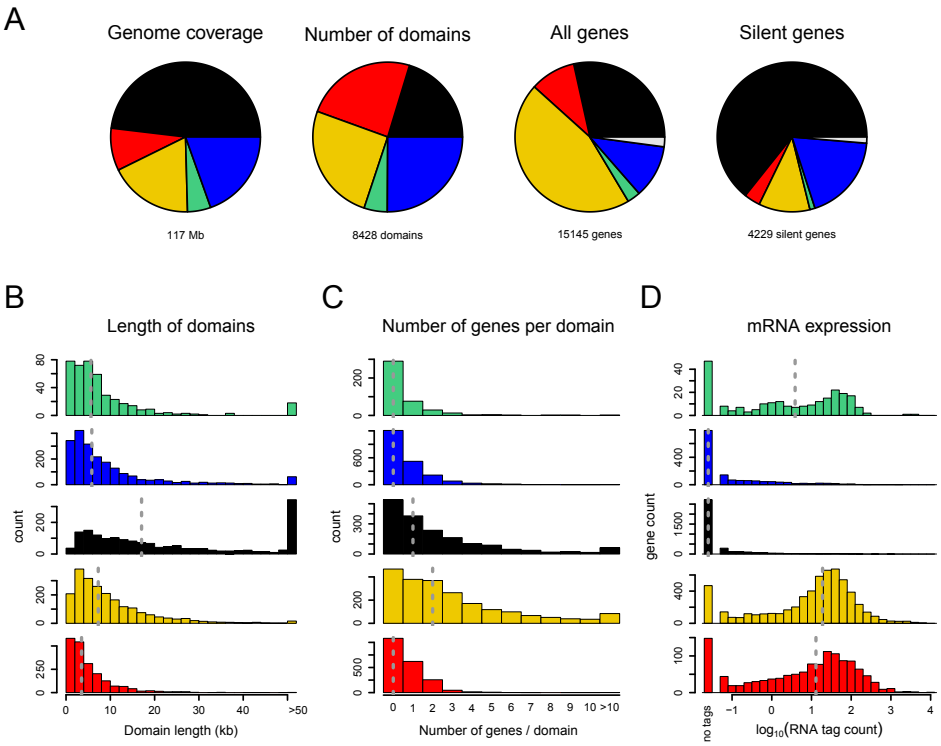


Figure 2 Characteristics of the five chromatin types. (A) Coverage and gene content of chromatin domains of each type. The chromatin type of a gene is defined as the chromatin type at its transcription start site (TSS). Grey sectors correspond to genes whose TSS maps at the transition between two chromatin 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.

than 50 kb, and 155 are larger than 100 kb, with the largest domain being 737 kb. Many individual domains include multiple neighboring genes (Figure 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 (Figure 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 cor-

responds to classic heterochromatin that is marked by SU(VAR)3-9, HP1, and the HP1-interacting proteins LHR and HP6. As described previously [13, 15], this type of chromatin is prominent in pericentric regions and on chromosome 4 (Figure S2A). 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 [4] . Indeed, H3K9me2 is highly and specifically enriched in GREEN chromatin (Figure 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 (Figure S2B). Furthermore, genome-wide ChIP of H3K27me3, the histone mark that is generated by E(Z) and recognized by PC [3] is highly enriched in BLUE chromatin (Figure 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 [16-17]. The presence of EFF in BLUE chromatin is consistent with a reported role of this protein in PcG-mediated silencing [18].

BLACK chromatin is the prevalent type of repressive chromatin

BLACK chromatin covers 48% of the probed genome and is thus by far the most abundant type (Figure 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 (Figure 2B). BLACK chromatin is overall relatively gene-poor (Figure 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 (Figure 2D). This is in agreement with the low coverage of BLACK chromatin by RPII18, a subunit shared by all three RNA polymerases (Figure 3A) and a lack of the active histone marks H3K4me2 and H3K79me3 as detected by ChIP (Figure 3B). We note that the majority of silent genes in the genome are located in BLACK chromatin (Figure 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 (Figure 3A). 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 (Figure 4A).

Given that genes in BLACK chromatin are expressed at very low levels, we asked whether BLACK chromatin actively represses transcription, or merely

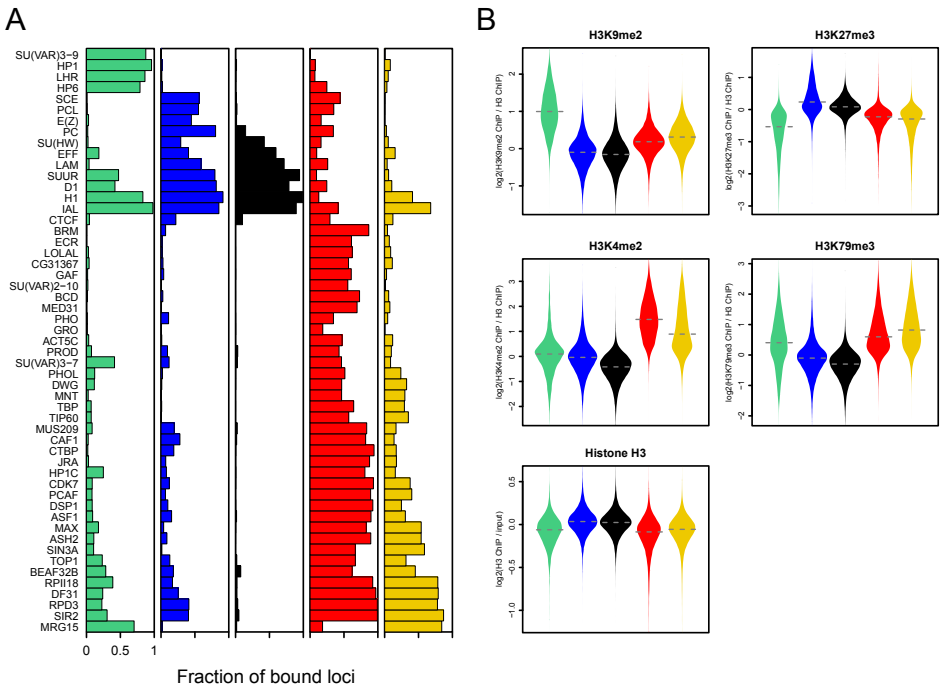


Figure 3 Chromatin types are characterized by distinctive protein combinations and histone modifications. (A) Fraction of all probed genomic loci within each chromatin type that is bound by each protein. Bound loci were determined separately for each protein as described in the text. (B) Levels of histone H3 and four histone modifications as determined by genome-wide ChIP. The distribution of values is shown as "violin plots", which are symmetrized density plots of binding values per chromatin type: the wider the violin, the more data points are associated to that value. Dashed horizontal lines indicate the median binding value for each chromatin type. Histone modification ChIP data were normalized to H3 occupancy. See also Figure S2.

forms secondary to a lack of transcription. In the former model, transgenes inserted into BLACK chromatin may exhibit reduced transcription, while in the latter model transgenes should be unaffected. To test this, we examined a dataset of 2,852 random P-element insertions that carry a mini-*white* eye color reporter gene. For each of these insertions the expression level was previously scored and the integration site mapped [19]. Strikingly, of 307 insertions located in BLACK regions 36% exhibited various degrees of *w* silencing, compared to 13% genome-wide (Figure 4B). Moreover, repression of transgene insertions in BLACK chromatin is more pronounced than in BLUE and GREEN chromatin. This result strongly indicates that BLACK chromatin has an active role in transcriptional silencing.

Developmental regulation of genes in BLACK chromatin.

Not all genes in BLACK regions are expected to remain silenced in various tissues. Indeed, a survey of tissue expression profiling data [20] indicates that genes in BLACK chromatin can become active, although their expression tends to be restricted to a few tissues only (Figure

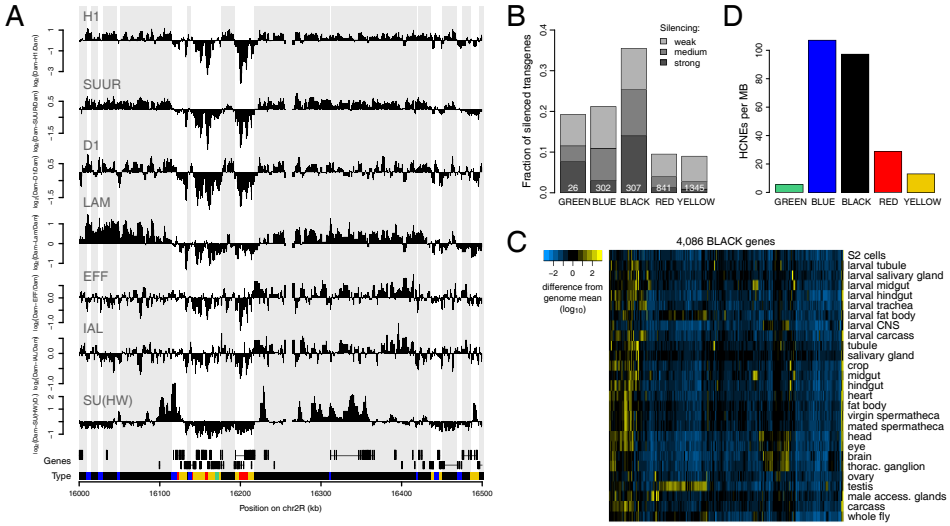


Figure 4 Properties of BLACK chromatin. (A) Sample plots of binding profiles of the six proteins that are the most prevalent in BLACK chromatin. Genes on both strands as well as chromatin types are depicted below the profiles. Grey blocks in the background correspond to BLACK chromatin domains. (B) Silencing of a *white* reporter gene in 2,852 P-element insertions in adult eyes [19] separated by chromatin type in Kc cells. The fraction of silenced insertions is higher among those overlapping with BLACK regions than in the rest of the genome ($p<2.2*10^{-16}$, Chi-squared test). (C) Relative expression levels (log_{10} scale, normalized to genome-wide average) of BLACK genes in various tissues [20]. (D) Density of highly conserved non-coding elements (HCNEs) per chromatin type.

4C). This suggests that BLACK chromatin domains as defined in Kc167 cells can be remodeled into a different chromatin type in some cell types. Consistent with this dynamic regulation, BLACK chromatin is particularly rich in highly conserved non-coding elements (HCNEs) (Figure 4D), which are thought to mediate gene regulation [21]. The density of HCNEs in BLACK chromatin is comparable to that in BLUE chromatin, which harbors many developmentally regulated genes [22], and is much higher than in the other three chromatin types. Together, these data suggest that BLACK chromatin is at least in part under 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 (Figure 2D), and levels of RNA polymerase (Figure 3A), H3K4me2 and H3K79me3 are typically high, whereas levels of H3K9me2 and H3K27me3 are low (Figure 3B).

RED and YELLOW chromatin share various chromatin proteins (Figure 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 [7].

Other proteins that are highly abundant in both RED and YELLOW chromatin include DF31, a little-studied protein that drives chromatin decondensation *in vitro* [23]; ASH2, a homolog of a subunit of a H3K4 methyltransferase complex in yeast and vertebrate cells [24]; and MAX, a DBF that is part of the MYC network of regulators of growth and proliferation [25].

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 (Figure 3A). Among these are the nucleosome remodeling ATPase Brahma (BRM); the regulator of chromosome structure SU(VAR)2-10; the Mediator subunit MED31; the 55 kDa subunit of CAF1, present in various histone-modifying complexes [26-27]; and several DBFs including the ecdysone receptor (ECR), GAGA factor (GAF), and Jun-related antigen (JRA).

These differences in protein composition prompted us to investigate the timing of DNA replication during S-phase, which is known to differ in relation with chromatin marks [28]. Analysis of a genome-wide replication timing map from Kc167 cells [29] 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 (Figure 5A). This coincides with a strong enrichment of origin recognition complex (ORC) binding in RED chromatin as mapped by ChIP [30] (Figure 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. Because human MRG15 has previously been reported to bind H3K36me3 [31], we compared the fine distribution of MRG15 and H3K36me3 along genes within the two chromatin types [32]. Indeed, both are highly enriched along genes in YELLOW chromatin, but nearly absent from RED chromatin (Figure 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 [2, 33]. 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 (Figure 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 [34] are

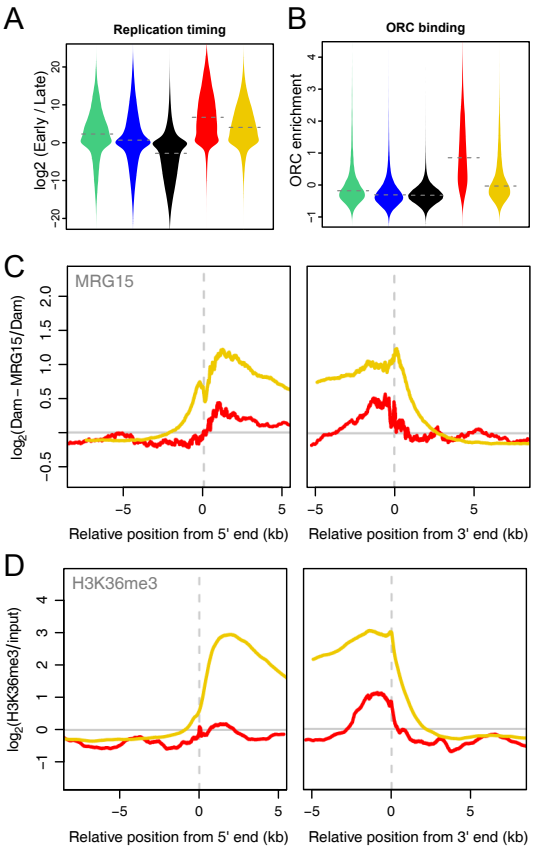


Figure 5 RED and YELLOW are two distinct types of euchromatin. (A) Violin plots of replication timing [29] per chromatin type. (B) Violin plots of origin of replication complex 2 (ORC2) binding [30] per chromatin type. (C) Average binding of MRG15 around 5′ and 3′ ends of genes in RED and YELLOW chromatin. Left panel, alignment to transcript 5′ ends; right panel, alignment to 3′ ends. Only genes that are entirely within one chromatin type are depicted. (D) Average enrichment of H3K36me3 [32], plotted as in (C).

highly enriched in YELLOW chromatin, while genes with more restricted expression patterns are depleted (Figure 6A). Consistent with this, Gene Ontology (GO) analysis revealed that universal cellular functions such as "ribosome", "DNA repair" and "nucleic acid metabolic process" are almost exclusively found in YELLOW chromatin (Figure 6B), while genes in RED chromatin are linked to more specific processes such as "receptor binding", "defense response", "transcription factor activity" and "signal transduction" (Figure 6C). Such specific functions and expression patterns require complex mechanisms of gene regulation. Indeed, intergenic regions in RED domains contain about twofold more HCNEs than YELLOW chromatin (Figure 4D), although not as much as BLACK and BLUE chromatin. Furthermore, genome wide formaldehyde-assisted identification of regulatory elements (FAIRE) [35-36] points to a high density of regulatory chromatin complexes in RED chromatin (Figure 6D).

Motif binding by DBFs is guided by chromatin types

Chromatin can affect the ability of DBFs to bind to their cognate binding sequences, which is thought to explain

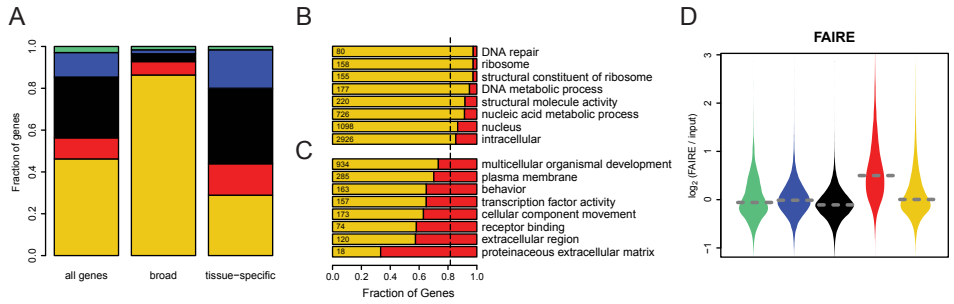


Figure 6 Genes in RED and YELLOW differ in regulation and function. (A) Distribution of genes having "broad" and "tissue-specific" expression patterns (defined in [34]) over the five chromatin types. Left bar shows distribution of all genes for comparison. (B)-(C) GO slim categories that are significantly enriched (B) or depleted (C) in RED compared to YELLOW genes. Bars indicate the fraction of RED and YELLOW genes for the given category (BLACK, GREEN and BLUE are not considered here). Vertical dotted line represents the distribution expected by random chance. The total number of RED and YELLOW genes within each category are indicated on the left. (D) Violin plots of the $\log_2$ FAIRE signal per chromatin type [36]. See also Figure S3.

why *in vivo* most DBFs bind to only a small subset of their recognition motifs in the genome [37]. We investigated how the five chromatin types might modulate DBF-DNA interactions. We focused on five DBFs in our dataset (JRA, MNT, GAF, CTCF and SU(HW)) for which the sequence-specificity is well-characterized. We first calculated the expected genomic binding pattern of each DBF, based on the occurrence of sequence motifs that match the known DBF recognition motif. The exactness of these matches is taken into account, yielding for each DamID-probed locus a predicted relative affinity for the DBF [38]. Genome-wide comparison of this sequence-based predicted affinity and actual protein occupancy indicated only weak to moderate correlations (Spearman's rho ranging from 0.04 to 0.35; dashed grey curves in Figure 7A; Figure S4). This suggests that chromatin indeed has substantial modulating effects on DBF-motif interactions.

We then repeated this correlation analysis by chromatin type. Surprisingly, this revealed that each DBF has its own dependence on chromatin context (Figure 7A and Figure S4): GAF and JRA both bind to their respective motif variants over a range of affinities in RED chromatin, but not in the other chromatin types; MNT binds to its motifs only in RED and YELLOW; CTCF preferentially binds its motifs in RED and BLUE chromatin; SU(HW) recognizes its motifs most efficiently in BLACK, BLUE and RED chromatin. Thus, each of the five chromatin types is conducive to DNA binding by specific subsets of DBFs. Some chromatin types may also weakly bind certain DBFs independently of DNA interactions, as suggested by the varying DamID baseline levels in loci that lack high-affinity motifs (e.g. for SU(HW) and CTCF; Figure 7A).

Four out of five DBFs exhibit a preference for their motif in RED chromatin. We wondered whether RED chromatin might have an intrinsic property such as 'openness' or nucleosome remodeling activity that would generally facilitate DBF access. To test this, we generated a DamID profile

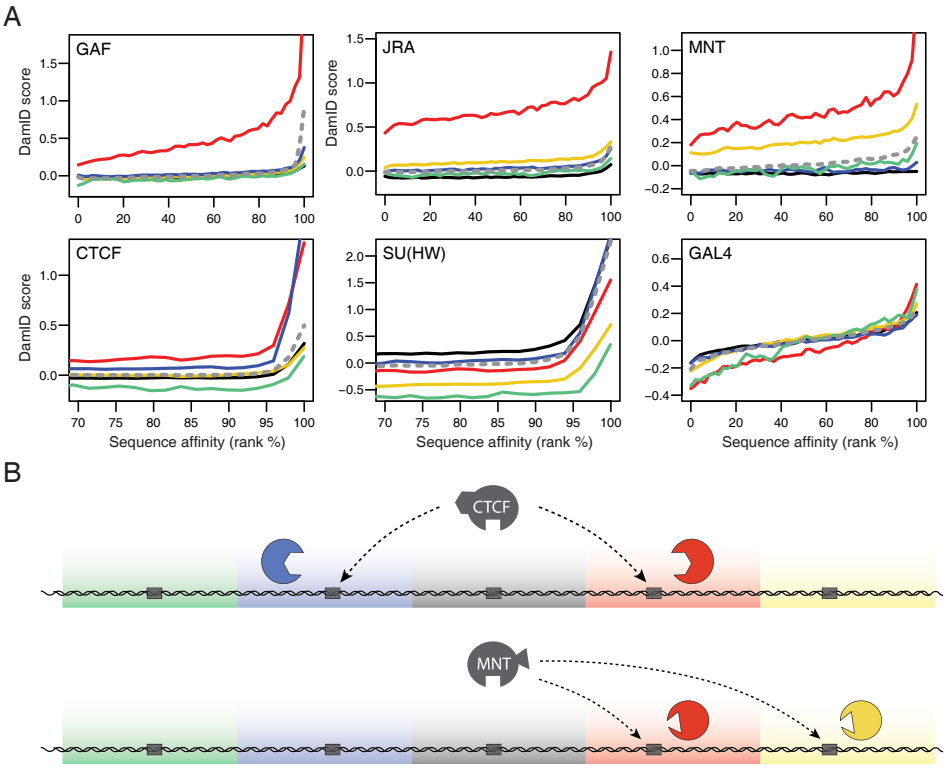


Figure 7 Binding of DBFs to their cognate motifs is differentially guided by chromatin types. (A) Correlations between predicted DNA affinity and actual binding detected by DamID, genome-wide (grey dashed lines) or for each chromatin type (solid lines), for six DBFs as indicated. Curves are loess-fitted lines; raw data is shown in Figure S4. (B) Cartoon model depicting the specific guidance of DBFs to their cognate motifs in only certain chromatin types, illustrated for CTCF and MNT. DBF binding to its cognate motif (grey box) is guided by protein-protein interactions. The presence of specific interactors (colored shapes) only in some chromatin types may account for targeting. See also Figure S4.

for the DNA-binding domain (DBD) of yeast Gal4. This foreign DBD is not expected to have specific protein-protein interactions with *Drosophila* chromatin, and its recognition motif occurs randomly throughout the fly genome. We observed similar interactions of Gal4-DBD with its cognate motifs in all five chromatin types (Figure 7A, bottom right panel). This indicates that RED chromatin does not have a general positive effect on protein-DNA interactions, and that high DBF occupancy in this chromatin type is more likely due to specific targeting mechanisms for each DBF. In summary, these results indicate that the five chromatin types together act as guides that help to target DBFs to specific regions of the genome, even though the cognate binding motifs are broadly distributed (Figure 7B).

| DISCUSSION

By systematic integration of 53 protein location maps we found 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, DBF targeting, 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

Identifying five chromatin states out of the binding profiles of 53 proteins comes out as a surprisingly low number (one can form approximately 10^{16} subsets of 53 elements). 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. We cannot exclude that the accumulation of binding profiles of additional proteins would reveal other novel chromatin types. We also anticipate that the pattern of chromatin types along the genome will vary between cell types. For example, many genes that are embedded in BLACK chromatin (defined in Kc167 cells) are activated in some other cell types (Figure 4C). Thus, the chromatin of these genes is likely to switch to an active type.

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 (Figure S1E). 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.

BLACK chromatin: a distinct type of repressive chromatin

Previous work on the expression of integrated reporter genes [39-41] had suggested that most of the fly genome is transcriptionally repressed, contrasting with the low coverage of PcG and HP1-marked chromatin. BLACK chromatin, which consists of a previously unknown combination of proteins and covers about half of the genome, may account for these observations. Essentially all genes in BLACK chromatin exhibit extremely low expression levels, and transgenes inserted in BLACK chromatin are frequently silenced, indicating that BLACK chromatin constitutes a strongly repressive environment. Importantly, BLACK chro-

matin is depleted of PcG proteins, HP1, SU(VAR)3-9 and associated proteins, and is also the latest to replicate, underscoring that it is different from previously characterized types of heterochromatin (here identified as BLUE and GREEN chromatin).

The proteins that mark BLACK domains provide important clues to the molecular biology of this type of chromatin. Loss of LAM, EFF or histone H1 causes lethality during *Drosophila* development [42-44]. Extensive *in vitro* and *in vivo* evidence has suggested a role for H1 in gene repression, most likely through stabilization of nucleosome positions [45-47]. The enrichment of LAM points to a role of the nuclear lamina in gene regulation in BLACK chromatin [48], consistent with the long-standing notion that peripheral chromatin is silent [49]. Depletion of LAM causes derepression of several LAM-associated genes [50], while artificial targeting of genes to the nuclear lamina can reduce their expression [51-52], suggesting a direct repressive contribution of the nuclear lamina in BLACK chromatin. D1 is a little-studied protein with 11 AT-hook domains. Overexpression of D1 causes ectopic pairing of intercalary heterochromatin [53], suggesting a role in the regulation of higher-order chromatin structure. SUUR specifically regulates late replication on polytene chromosomes [54], which is of interest because BLACK chromatin is particularly late-replicating. EFF is highly similar to the yeast and mammalian ubiquitin ligase Ubc4 that mediates ubiquitination of histone H3 [55-56], raising the possibility that nucleosomes in BLACK chromatin may carry specific ubiquitin marks. These insights suggest that BLACK chromatin is important for chromosome architecture as well as gene repression and provide important leads for further study of this previously unknown yet prevalent type of 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 [57]. Why active genes in RED chromatin lack H3K36me3 remains to be elucidated.

The remarkably high protein occupancy in RED chromatin suggests that RED domains are "hubs" of regulatory activity. This may be related to the predominantly tissue-specific expression of genes in RED chromatin, which presumably requires many regulatory proteins. We note that our DamID assay integrates protein binding events over nearly 24 hours, so it is likely that not all proteins bind simultaneously; some proteins may bind only during a specific stage of the cell cycle. It is highly unlikely that the high protein occupancy in RED chromatin originates 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, ORC also shows a specific enrichment in RED chromatin, even though it was mapped by ChIP, by another laboratory and on another detection platform [30]. Fourth, DamID of Gal4-DBD does not show any enrichment in RED chromatin.

RED chromatin resembles DBF binding hotspots that were previously discovered in a smaller-scale study in *Drosophila* cells [11]. Discrete genomic regions targeted by many DBFs have recently also been found in mouse ES cells [58], hence it is tempting to speculate that an equivalent of RED chromatin may also exist in mammalian cells. Housekeeping and dynamically regulated genes in budding yeast also exhibit a dichotomy in chromatin organization [59] which may be related to our distinction between YELLOW and RED chromatin. The observations that RED chromatin is generally the earliest to replicate and strongly enriched in ORC binding, suggest that this chromatin type may be not only involved in transcriptional regulation but also in the control of DNA replication.

Chromatin types as guides for DBF targeting

Our analysis of DBF binding indicates that the five chromatin types together act as a guidance system to target DBFs to specific genomic regions. This system directs DBFs to certain genomic domains even though the DBF recognition motifs are more widely distributed. We propose that targeting specificity is at least in part achieved through interactions of DBFs with particular partner proteins that are present in some of the five chromatin types but not in others (Figure 7B). The observation that yeast Gal4-DBD binds its motifs with nearly equal efficiency in all five chromatin types suggests that differences in compaction among the chromatin types represent overall a minor factor in the targeting of DBFs. Although additional studies will be needed to further investigate the molecular mechanisms of DBF guidance, the identification of five principal types of chromatin provides a firm basis for future dissection of the roles of chromatin organization in global gene regulation.

| MATERIAL & METHODS

Constructs

DamID constructs used for this study are listed in Table S1. New constructs were cloned by TOPO cloning and GATEWAY recombination as described [36] or by Cre-mediated recombination. For the latter we generated an acceptor vector containing the Hsp70 promoter upstream of myc-epitope tagged Dam, using the Creator Acceptor Vector Construction Kit (Clontech, 631618). Chromatin protein open reading frames from pDNR-Dual donor vectors (Drosophila Genomics Resource Center, Bloomington) were cloned into the acceptor vector using the Creator™ DNA Cloning Kit (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 [13]. Only MNT, GRO and IAL gave weak nuclear signals but were not discarded because MNT

and GRO were successfully mapped by DamID in previous studies [25, 60] and IAL binds metaphase chromosomes [61].

DamID, ChIP and microarrays

DamID assays were carried out under standardized conditions as described previously [11] with a minor modification: proteins were grouped in sets sharing the same Dam-only 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. ChIP and subsequent linear amplification reactions were done as described [62] using anti-H3K27me3 (07-449) and anti-H3K4me2 (07-030) from Upstate Biotechnology; anti-H3K9me2 (1220), and anti-H3 (1791) from Abcam; affinity-purified anti-H1 serum [36]; and anti-H3K79me3 [63] kindly provided by Fred van Leeuwen. Fluorescent labeling of DamID and ChIP samples and two-color hybridizations on custom-designed 385k NimbleGen arrays [36] were performed according to NimbleGen's array users guide, version 4.0. Arrays 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. This open reading frame was masked before further data analysis.

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) RNA tag sequencing was carried out on an Illumina Solexa GAII using the tag profiling kit with DpnII. Two RNA samples yielded 7.4 and 9.0 million reads. Tags were mapped by BLAST, requiring at most 2 mismatches and 11 consecutively matching bases. Only the tags mapping to the last GATC of a transcript (FlyBase release 5.8) 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.

Data availability and analysis

DamID, ChIP and expression data, binarized DamID data and a list of the coordinates of all identified chromatin domains are available from NCBI's Gene Expression Omnibus, accession number GSE22069. Computational methods are described in the Supplementary Methods (available upon request).

| 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; Reuven Agami, Fred van Leeuwen, Wouter Meuleman, Ludo Pagie and Aleksey Pindyurin for helpful suggestions. Supported by an EMBO Long-term Fellowship to J.K.; National Institutes of Health grants T32GM008798, R01HG003008, and U54CA121852 to L.D.W. and H.J.B.; and grants from the Netherlands Genomics Initiative, NWO-ALW VICI and an EURYI Award to B.v.S.

| REFERENCES

1. Berger, S.L., The complex language of chromatin regulation during transcription. Nature, 2007. 447(7143): p. 407-12.
2. Rando, O.J. and H.Y. Chang, Genome-wide views of chromatin structure. Annu Rev Biochem, 2009. 78: p. 245-71.
3. Sparmann, A. and M. van Lohuizen, Polycomb silencers control cell fate, development and cancer. Nat Rev Cancer, 2006. 6(11): p. 846-56.
4. Hediger, F. and S.M. Gasser, Heterochromatin protein 1: don't judge the book by its cover! Curr Opin Genet Dev, 2006. 16(2): p. 143-50.
5. de Wit, E., F. Greil, and B. van Steensel, High-resolution mapping reveals links of HP1 with active and inactive chromatin components. PLoS Genet, 2007. 3(3): p. e38.
6. Collas, P., The state-of-the-art of chromatin immunoprecipitation. Methods Mol Biol, 2009. 567: p. 1-25.
7. Wang, Z., et al., Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. Cell, 2009. 138(5): p. 1019-31.
8. Gelbart, M.E., et al., Genome-wide identification of Isw2 chromatin-remodeling targets by localization of a catalytically inactive mutant. Genes Dev, 2005. 19(8): p. 942-54.
9. Schmiedeberg, L., et al., A temporal threshold for formaldehyde crosslinking and fixation. PLoS One, 2009. 4(2): p. e4636.
10. van Steensel, B., J. Delrow, and S. Henikoff, Chromatin profiling using targeted DNA adenine methyltransferase. Nat Genet, 2001. 27(3): p. 304-8.
11. Moorman, C., et al., Hotspots of transcription factor colocalization in the genome of Drosophila melanogaster. Proc Natl Acad Sci U S A, 2006. 103(32): p. 12027-32.
12. Negre, N., et al., Chromosomal distribution of PcG proteins during Drosophila development. PLoS Biol, 2006. 4(6): p. e170.
13. Greil, F., et al., HP1 controls genomic targeting of four novel heterochromatin proteins in Drosophila. Embo J, 2007. 26(3): p. 741-51.
14. Pindyurin, A.V., et al., SUUR joins separate subsets of PcG, HP1 and B-type lamin targets in Drosophila. J Cell Sci, 2007. 120(Pt 14): p. 2344-51.
15. Ebert, A., et al., Histone modification and the control of heterochromatic gene silencing in Drosophila. Chromosome Res, 2006. 14(4): p. 377-92.
16. Czermin, B., et al., Physical and functional association of SU(VAR)3-9 and HDAC1 in Drosophila. EMBO Rep, 2001. 2(10): p. 915-9.
17. Tie, F., et al., A 1-megadalton ESC/E(Z) complex from Drosophila that contains polycomblike and RPD3. Mol Cell Biol, 2003. 23(9): p. 3352-62.
18. Fauvarque, M.O., et al., Dominant modifiers of the polyhomeotic extra-sex-combs phenotype induced by marked P element insertional mutagenesis in Drosophila. Genet Res, 2001. 78(2): p. 137-48.
19. Babenko, V.N., et al., Paucity and preferential suppression of transgenes in late replication domains of the D. melanogaster genome. BMC Genomics, 2010. 11: p. 318.

20. Chintapalli, V.R., J. Wang, and J.A. Dow, Using FlyAtlas to identify better Drosophila melanogaster models of human disease. Nat Genet, 2007. 39(6): p. 715-20.
21. Engstrom, P.G., et al., Genomic regulatory blocks underlie extensive microsynteny conservation in insects. Genome Res, 2007. 17(12): p. 1898-908.
22. Tolhuis, B., et al., Genome-wide profiling of PRC1 and PRC2 Polycomb chromatin binding in Drosophila melanogaster. Nat Genet, 2006. 38(6): p. 694-9.
23. Crevel, G., H. Huikeshoven, and S. Cotterill, Df31 is a novel nuclear protein involved in chromatin structure in Drosophila melanogaster. J Cell Sci, 2001. 114(Pt 1): p. 37-47.
24. Nagy, P.L., et al., A trithorax-group complex purified from Saccharomyces cerevisiae is required for methylation of histone H3. Proc Natl Acad Sci U S A, 2002. 99(1): p. 90-4.
25. Orian, A., et al., Genomic binding by the Drosophila Myc, Max, Mad/Mnt transcription factor network. Genes Dev, 2003. 17(9): p. 1101-14.
26. Martinez-Balbas, M.A., et al., Drosophila NURF-55, a WD repeat protein involved in histone metabolism. Proc Natl Acad Sci U S A, 1998. 95(1): p. 132-7.
27. Tie, F., et al., The Drosophila Polycomb Group proteins ESC and E(Z) are present in a complex containing the histone-binding protein p55 and the histone deacetylase RPD3. Development, 2001. 128(2): p. 275-86.
28. Gilbert, D.M., Replication timing and transcriptional control: beyond cause and effect. Curr Opin Cell Biol, 2002. 14(3): p. 377-83.
29. Schwaiger, M., et al., Chromatin state marks cell-type- and gender-specific replication of the Drosophila genome. Genes Dev, 2009. 23(5): p. 589-601.
30. MacAlpine, H.K., et al., Drosophila ORC localizes to open chromatin and marks sites of cohesin complex loading. Genome Res, 2010. 20(2): p. 201-11.
31. Zhang, P., et al., Structure of human MRG15 chromo domain and its binding to Lys36-methylated histone H3. Nucleic Acids Res, 2006. 34(22): p. 6621-8.
32. Bell, O., et al., Accessibility of the Drosophila genome discriminates PcG repression, H4K16 acetylation and replication timing. Nat Struct Mol Biol, 2010. 17(7): p. 894-900.
33. Lee, J.S. and A. Shilatifard, A site to remember: H3K36 methylation a mark for histone deacetylation. Mutat Res, 2007. 618(1-2): p. 130-4.
34. Tomancak, P., et al., Global analysis of patterns of gene expression during Drosophila embryogenesis. Genome Biol, 2007. 8(7): p. R145.
35. Giresi, P.G., et al., FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res, 2007. 17(6): p. 877-85.
36. Braunschweig, U., et al., Histone H1 binding is inhibited by histone variant H3.3. Embo J, 2009. 28(23): p. 3635-45.
37. Beato, M. and K. Eisfeld, Transcription factor access to chromatin. Nucleic Acids Res, 1997. 25(18): p. 3559-63.
38. Foat, B.C., A.V. Morozov, and H.J. Bussemaker, Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. Bioinformatics, 2006. 22(14): p. e141-9.
39. Handler, A.M. and R.A. Harrell, 2nd, Germline transformation of Drosophila melanogaster with the piggyBac transposon vector. Insect Mol Biol, 1999. 8(4): p. 449-57.
40. Kelley, R.L. and M.I. Kuroda, The Drosophila roX1 RNA gene can overcome silent chromatin by recruiting the male-specific lethal dosage compensation complex. Genetics, 2003. 164(2): p. 565-74.
41. Markstein, M., et al., Exploiting position effects and the gypsy retrovirus insulator to engineer precisely expressed transgenes. Nat Genet, 2008. 40(4): p. 476-83.
42. Lenz-Bohme, B., et al., Insertional mutation of the Drosophila nuclear lamin Dm0 gene results in defective nuclear envelopes, clustering of nuclear pore complexes, and accumulation of annulate lamellae. J Cell Biol, 1997. 137(5): p. 1001-16.
43. Cenci, G., et al., UbcD1, a Drosophila ubiquitin-conjugating enzyme required for proper telomere behavior. Genes Dev, 1997. 11(7): p. 863-75.
44. Lu, X., et al., Linker histone H1 is essential for Drosophila development,

the establishment of pericentric heterochromatin, and a normal polytene chromosome structure. Genes Dev, 2009. 23(4): p. 452-65.
45. Laybourn, P.J. and J.T. Kadonaga, Role of nucleosomal cores and histone H1 in regulation of transcription by RNA polymerase II. Science, 1991. 254(5029): p. 238-45.
46. Wolffe, A.P. and J.J. Hayes, Chromatin disruption and modification. Nucleic Acids Res, 1999. 27(3): p. 711-20.
47. Woodcock, C.L., A.I. Skoultchi, and Y. Fan, Role of linker histone in chromatin structure and function: H1 stoichiometry and nucleosome repeat length. Chromosome Res, 2006. 14(1): p. 17-25.
48. Pickersgill, H., et al., Characterization of the Drosophila melanogaster genome at the nuclear lamina. Nat Genet, 2006. 38(9): p. 1005-14.
49. Towbin, B.D., P. Meister, and S.M. Gasser, The nuclear envelope--a scaffold for silencing? Curr Opin Genet Dev, 2009. 19(2): p. 180-6.
50. Shevelyov, Y.Y., et al., The B-type lamin is required for somatic repression of testis-specific gene clusters. Proc Natl Acad Sci U S A, 2009. 106(9): p. 3282-7.
51. Reddy, K.L., et al., Transcriptional repression mediated by repositioning of genes to the nuclear lamina. Nature, 2008. 452(7184): p. 243-7.
52. Finlan, L.E., et al., Recruitment to the nuclear periphery can alter expression of genes in human cells. PLoS Genet, 2008. 4(3): p. e1000039.
53. Smith, M.B. and K.S. Weiler, Drosophila D1 overexpression induces ectopic pairing of polytene chromosomes and is deleterious to development. Chromosoma, 2010. 119(3): p. 287-309.
54. Zhimulev, I.F., et al., Influence of the SuUR gene on intercalary heterochromatin in Drosophila melanogaster polytene chromosomes. Chromosoma, 2003. 111(6): p. 377-98.
55. Singh, R.K., et al., Histone levels are regulated by phosphorylation and ubiquitylation-dependent proteolysis. Nat Cell Biol, 2009. 11(8): p. 925-33.
56. Liu, Z., R. Oughtred, and S.S. Wing, Characterization of E3Histone, a novel testis ubiquitin protein ligase which ubiquitinates histones. Mol Cell Biol, 2005. 25(7): p. 2819-31.
57. Li, B., et al., Infrequently transcribed long genes depend on the Set2/Rpd3S pathway for accurate transcription. Genes Dev, 2007. 21(11): p. 1422-30.
58. Chen, X., et al., Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell, 2008. 133(6): p. 1106-17.
59. Tirosh, I. and N. Barkai, Two strategies for gene regulation by promoter nucleosomes. Genome Res, 2008. 18(7): p. 1084-91.
60. Bianchi-Frias, D., et al., Hairy transcriptional repression targets and cofactor recruitment in Drosophila. PLoS Biol, 2004. 2(7): p. E178.
61. Giet, R. and D.M. Glover, Drosophila aurora B kinase is required for histone H3 phosphorylation and condensin recruitment during chromosome condensation and to organize the central spindle during cytokinesis. J Cell Biol, 2001. 152(4): p. 669-82.
62. Kind, J., et al., Genome-wide analysis reveals MOF as a key regulator of dosage compensation and gene expression in Drosophila. Cell, 2008. 133(5): p. 813-28.
63. Schubeler, D., et al., The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. Genes Dev, 2004. 18(11): p. 1263-71.

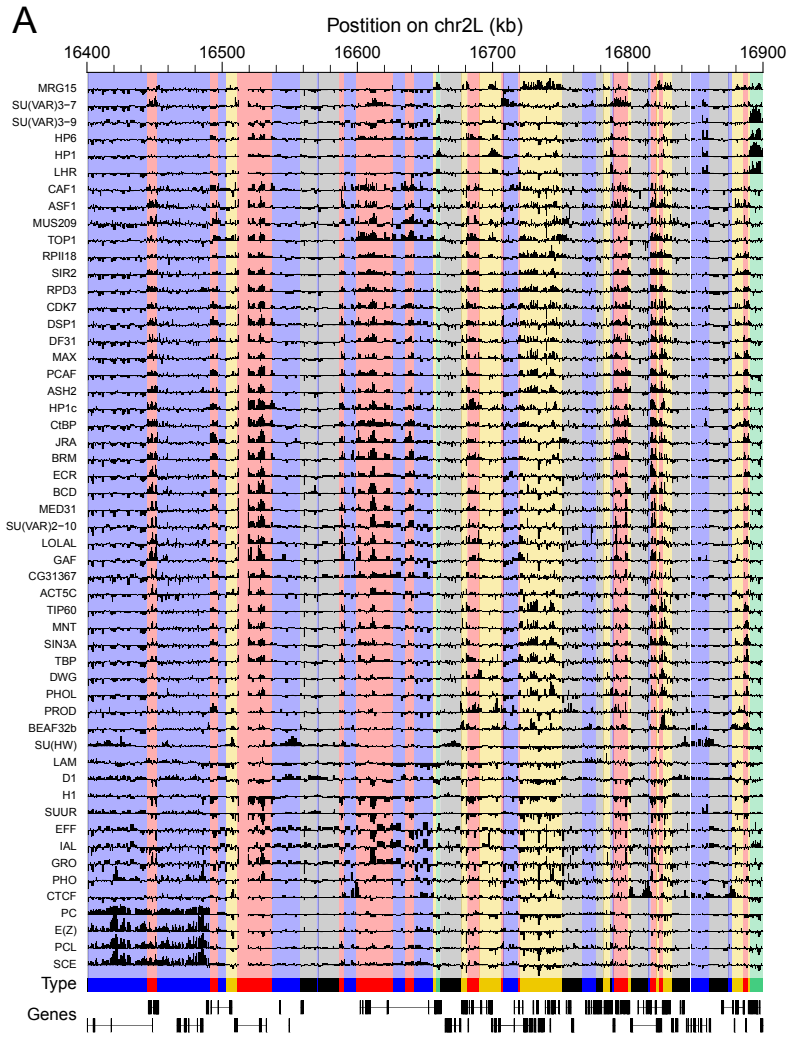
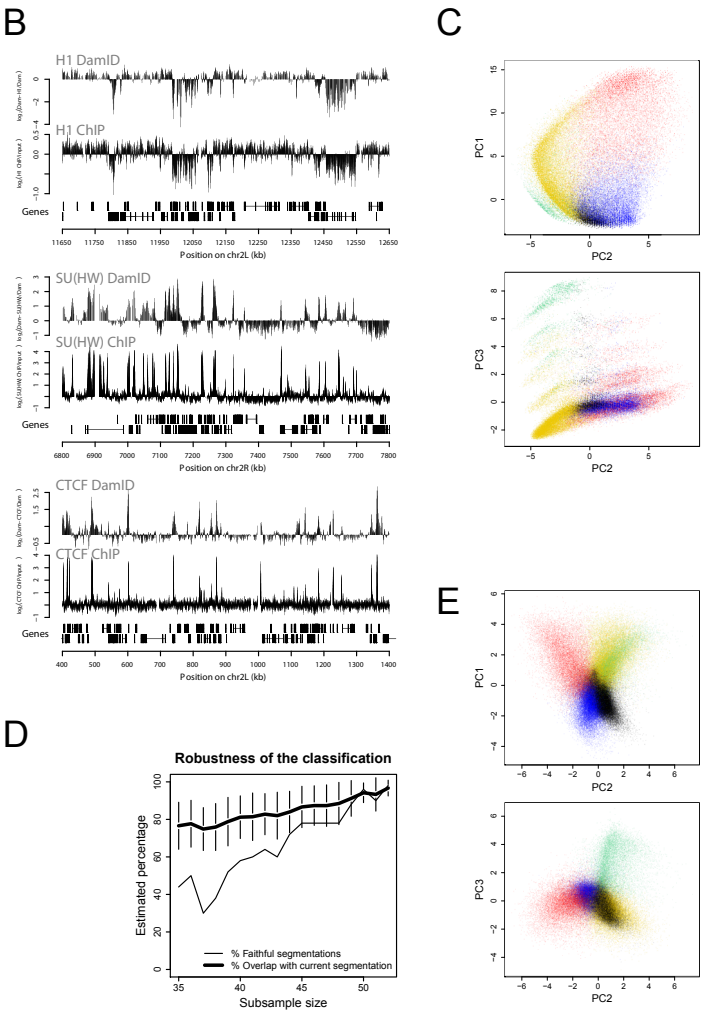


Figure S1 Validation of DamID and robustness analysis of the 5-type segmentation. (A) Magnification of the 53 DamID profiles as depicted in Figure 1. Traces show $\log_2$ DamID ratios, and width of black bars represents the length of DpnI restriction fragments. (B) Comparisons of DamID (Braunschweig 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. (C) 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. (D) 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. (E) 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%.

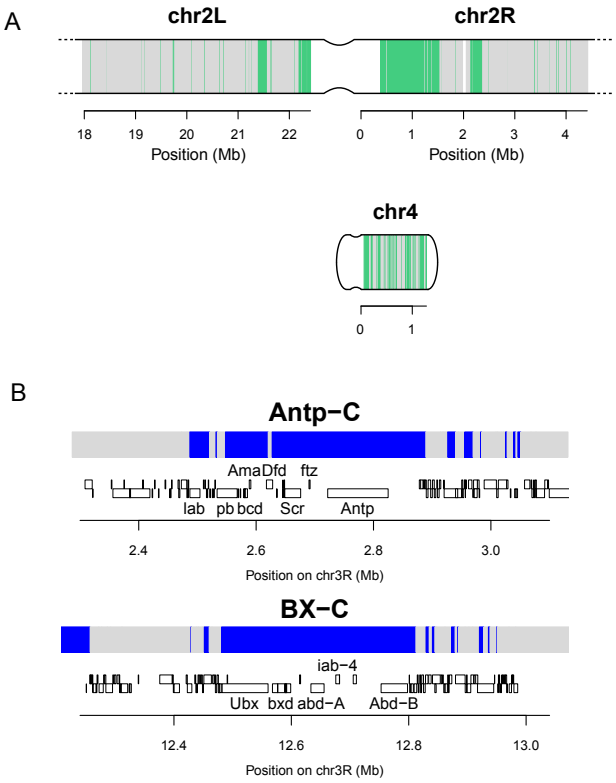


Figure S2 Localization of GREEN and BLUE chromatin supports their identity with known heterochromatin types. (A) Chromosomal maps of the pericentric region of chromosome 3 and of the entire chromosome 4. GREEN chromatin domains are shown, and other types are collectively represented in grey. Constrictions 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.

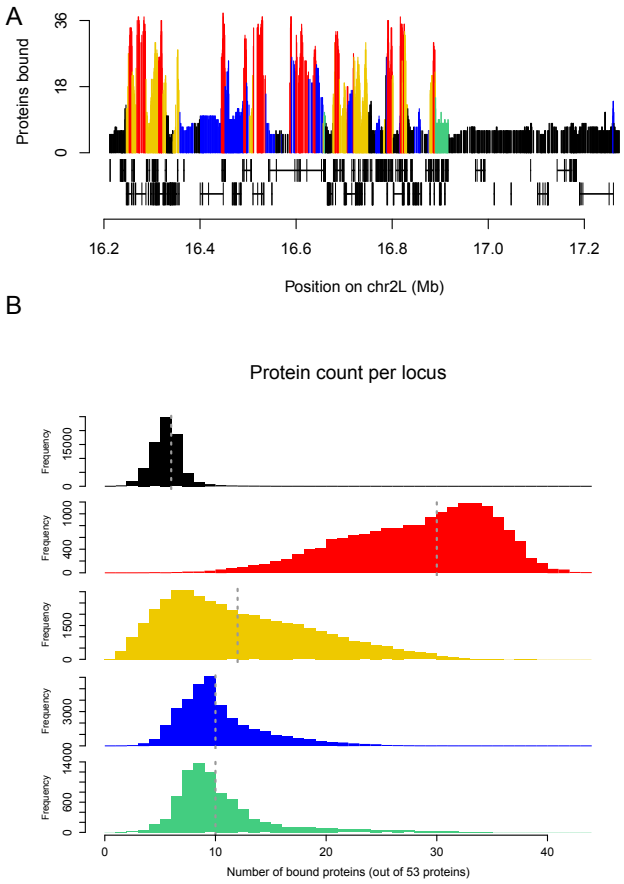


Figure S3 Chromatin types differ widely in their total protein occupancy. (A) Sample plot of the total occupancy (out of 53 mapped proteins) on a 1Mb segment of chromosome 2L. The height of each vertical line indicates the number of proteins bound to a locus. The color of the line indicates the local chromatin type. (B) Histograms of total occupancy distributions per chromatin type. Grey vertical dashed lines indicate median values.

712aa57b6f364d4c9b55374161d46eb6

Figure S4 Chromatin types guide DBF binding to their motifs. (A) Optimized position-specific affinity matrices for 5 *Drosophila* DBFs and Gal4-DBD (see Supplementary Methods for details, available upon request). (B) Distributions of relative affinities of GATC fragments mapping within each of the five types. For each chromatin type, DamID scores of GATC fragments ranked by the predicted affinity of a 2 kb window around its center. Grey lines represent the loess fit. Last row: superimposed loess fits for the five chromatin types. Colored numbers indicated the Spearman's rank correlation coefficients of DamID value with predicted affinities.

4

Table S1 List and accession numbers of proteins mapped with DamID. Membership to chromatin protein class as well as vectors used for cloning are indicated.

	Name	FlyBase Protein Accession	Plasmid Backbone	Transcription Factor	Transcription Machinery	Pericentric Heterochromatin	Polycomb Group	Nuclear Envelope	
1	ACT5C	FBpp0070787	pGWMycDam						
2	ASF1	FBpp0074715	pCreDamMyc						
3	ASH2	FBpp0084040	pCreDamMyc						
4	BCD	FBpp0081165	pDamMyc	X					
5	BEAF32b	FBpp0086572	pGWMycDam						
6	BRM	FBpp0075279	pDamMyc						
7	CAF1	FBpp0082511	pGWDamMyc						
8	CDK7	FBpp0070723	pCreDamMyc		X				
9	CG31367	FBpp0292134	pCreDamMyc						
12	CtBP	not specified	pDamMyc						
10	CTCF	FBpp0076588	pGWDamMyc						
11	D1	FBpp0081468	pGWDamMyc						
13	DF31	FBpp0085273	pCreDamMyc						
14	DSP1	FBpp0088319	pGWDamMyc	X			X		
15	DWG	FBpp0070465	pGWMycDam						
16	E(Z)	FBpp0076008	pCreDamMyc				X		
17	ECR	not specified	pDamMyc	X					
18	EFF	FBpp0082477	pCreDamMyc						
19	GAF	FBpp0089419	pDamMyc	X					
20	GRO	FBpp0084337	pMycDam						
21	H1	FBpp0085248	pGWDamMyc						
22	HP1	FBpp0079251	pDamMyc			X			
23	HP1c	FBpp0083702	pMycDam						
25	HP6	FBpp0077134	pDamMyc			X			
26	IAL	FBpp0079769	pCreDamMyc			X			
27	JRA	FBpp0087499	pDamMyc	X					
28	LAM	FBpp0078733	pDamMyc					X	
24	LHR	FBpp0086073	pDamMyc			X			
29	LOLAL	FBpp0085956	pCreDamMyc	X			X		

	Histone Modifying Enzyme	Binds Histone Modification	Cofactor	Chromatin Assembly/ Remodeling	Trithorax Group	Structural Protein	Insulator	Replication	Plasmid Published	Profile published before
				X					This paper	
				X					This paper	
					X				This paper	
									a)	
							X		b)	d)
				X	X				a)	
				X				X	This paper	
									This paper	
									This paper	
			X						c)	
							X		d)	d)
						X			e)	
						X			This paper	
									e)	
							X		d)	d)
	X								This paper	
									a)	
	X								This paper	
					X		X		d)	d)
			X						a)	
						X			f)	f)
		X				X			f)	
			X						f)	
									f)	
									This paper	
									a)	
									d)	d)
									f)	
									This paper	

Table S1 ***Continued***.

	Name	FlyBase Protein Accession	Plasmid Backbone	Transcription Factor	Transcription Machinery	Pericentric Heterochromatin	Polycomb Group	Nuclear Envelope	
30	MAX	FBpp0074785	pMycDam	X					
31	MED31	FBpp0078496	pCreDamMyc		X				
32	MNT	FBpp0070554	pMycDam	X					
33	MRG15	FBpp0082579	pCreDamMyc						
34	MUS209	FBpp0078496	pCreDamMyc						
35	PC	FBpp0078059	pDamMyc				X		
36	PCAF	FBpp0075701	pGWDamMyc						
37	PCL	FBpp0085914	pCreDamMyc				X		
38	PHO	FBpp0088268	pMycDam	X			X		
39	PHOL	FBpp0088268	pCreDamMyc	X			X		
40	PROD	FBpp0085785	pCreDamMyc						
41	RPD3	FBpp0073173	pMycDam						
42	RPII18	FBpp0078433	pDamMyc		X				
43	SCE	FBpp0084614	pMycDam				X		
44	SIN3A	FBpp0087003	pDamMyc						
45	SIR2	FBpp0080015	pMycDam						
46	SU(HW)	FBpp0082404	pGWDamMyc						
47	SU(VAR)2-10	FBpp0087657	pCreDamMyc						
48	SU(VAR)3-7	FBpp0082204	pDamMyc			X			
49	SU(VAR)3-9	FBpp0082583	pMycDam			X			
50	SUUR	FBpp0075933	pDamMyc			X			
51	TBP	FBpp0071596	pCreDamMyc		X				
52	TIP60	FBpp0070636	pGWMycDam						
53	TOP1	FBpp0073822	pGWMycDam						

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

	Histone Modifying Enzyme	Binds Histone Modification	Cofactor	Chromatin Assembly/ Remodeling	Trithorax Group	Structural Protein	Insulator	Replication	Plasmid Published	Profile published before
									a)	
			X						This paper	
									h)	
		X							This paper	
								X	This paper	
		X							i)	
	X								This paper	
									This paper	
									a)	
									This paper	
						X			This paper	
	X								a)	
									f)	f)
									i)	
	X								h)	
	X								c)	
							X		b)	
						X			This paper	
									e)	
	X								f)	
								X	j)	
									This paper	
	X								This paper	
						X			b)	

A network model of the molecular organization of chromatin in *Drosophila*

Joke G. van Bemmel[1], Guillaume J. Filion[1,2], Wendy Talhout[1,3], Arantxa Rosado[1], Bas van Steensel[1]

This manuscript is submitted for publication

[1]Division of Gene Regulation, Netherlands Cancer Institute, Amsterdam, the Netherlands; [2]Current address: Centro Regulación Genomica (CRG), Barcelona, Spain; [3]Current address: VSBN, Beverwijk, The Netherlands

5

| ABSTRACT

Chromatin governs gene regulation and genome maintenance, yet a substantial fraction of the chromatin proteome is still unexplored. Moreover, a global model of the chromatin protein network is lacking. By screening >100 candidates we identify 42 novel *Drosophila* chromatin proteins with specific genomic binding patterns. Bayesian network modeling of the binding profiles of these and 70 known chromatin components yields a detailed blueprint of the *in vivo* chromatin protein network. We demonstrate functional compartmentalization of this network, and predict functions for most of the novel chromatin proteins, including roles in DNA replication and repair, and gene activation and repression.

| INTRODUCTION

Chromatin is the ensemble of nuclear DNA and all associated proteins and RNA molecules. The chromatin proteome is highly complex: it consists of histones that carry a wide range of post-translational modifications, and hundreds of other proteins that can interact with the DNA, with (modified) histones, and with each other. The precise protein composition of the chromatin along the genome determines patterns of gene expression, DNA replication and other functions.

Several studies have recently employed large-scale genome-wide profiling of chromatin components to gain insight into the global protein composition and organization of chromatin. Chromatin immunoprecipitation (ChIP) analyses of large series of histone marks in *Arabidopsis thaliana*, *Caenorhabditis elegans, Drosophila melanogaster* and mouse and human cells identified 4-9 prevalent major combinatorial patterns (chromatin "states" or "types") which are related to genome function[1-6]. An integrative analysis of DamID maps of 53 chromatin proteins in *Drosophila* Kc167 cells uncovered five principal chromatin types, defined as five distinct combinations of proteins that are recurrent throughout the genome. These five chromatin types include the well known HP1-heterochromatin and Polycomb chromatin, two functionally different active chromatin types and a poorly characterized repressive type that covers the majority of all inactive genes.[7]

Despite these advances, our understanding of the overall molecular architecture of chromatin is still far from complete. Here, we address two of the major hurdles that preclude the construction of a 'big picture' view of chromatin. First, a large fraction of the chromatin proteome is still completely uncharacterized. This raises the question whether chromatin types or protein complexes exist that so far have eluded identification and characterization. It is thus important to systematically survey the so far unexplored components of the chromatin proteome. A second major challenge is to chart the intricate network of protein interactions that underlies the formation of the principal chromatin types. Many individual protein complexes have been analyzed in detail by biochemical approaches, yet it has remained difficult to construct comprehensive network models of chromatin

beyond a handful of proteins. As a consequence, an overview of the interaction network of chromatin proteins is still lacking.

Here, we report an integrated approach that begins to tackle these two challenges, using *D. melanogaster* as a model system. First, we have conducted a broad survey to identify and characterize a large number of hitherto unknown chromatin proteins. Specifically, we have used the DamID technology to generate genome-wide binding maps of >100 candidate novel chromatin proteins. For 42 of these proteins this yielded a specific genomic binding pattern, which identifies these proteins as novel chromatin components. Second, we have employed a Bayesian network approach to integrate genome-wide binding data of the 42 novel proteins together with 70 known chromatin components, resulting into a graphical model of the network of interactions among all of these 112 proteins. This model shows that the novel proteins are distributed throughout the network and thus contribute to most functional categories of chromatin proteins. Combination with orthogonal datasets reveals functional compartmentalization of the protein network and predicts functions for most of the novel proteins, some of which we validated by additional experiments. Taken together, our systematic approach provides a framework for a global understanding of chromatin, and identifies and annotates dozens of novel chromatin proteins.

| RESULTS

A DamID mapping survey identifies 42 novel chromatin proteins.

In order to identify new chromatin components, we used the DamID technology to systematically test 112 candidate *Drosophila* proteins for specific interactions with the genome[7-9]. We reasoned that any protein that yields a clear and non-random genome-wide DamID profile must closely interact with specific genomic regions, thereby qualifying as a chromatin protein. We subjected the candidate chromatin proteins to this test in *Drosophila* Kc167 cells.

The 112 candidate proteins were selected as follows. First, in order to qualify as 'novel', we required that there was no prior direct experimental evidence that the proteins are components of chromatin, based on all publications listed in FlyBase (http://flybase.org). Second, we chose candidates with a high probability of being a chromatin protein, either because they contain a domain that is often found in chromatin proteins, such as a DNA binding domain, chromodomain, BESS domain, or HMG-box (52 proteins total), or an RNA-binding domain (21 proteins); or because they were found to interact with one or more known chromatin proteins in a high-throughput yeast two-hybrid (Y2H) screen[10] (39 proteins) (Table S1). Third, we required candidate proteins to be endogenously expressed in Kc167 cells, as judged from RNA tag sequencing data[7]. Fourth, for efficient construction of Dam-fusion proteins, we focused on proteins for which the open reading frame (ORF) was available in a recombination vector (*Drosophila* Genomics Resource

Center, Bloomington). In the resulting list of candidate proteins, the vast majority has not been characterized at all, which is illustrated by the fact that in FlyBase they were only known by a systematic identifier ("CG" for Computed Gene, followed by a number). A small fraction of the proteins has been linked previously to a function unrelated to chromatin.

We subjected all 112 proteins to DamID experiments, each as two independent biological replicates. In an initial agarose gel based assay (see Methods) 33 Dam-fusion proteins did not yield detectable adenine methylation of the genome; these proteins were not analyzed further. For the remaining 79, adenine-methylated DNA was amplified and hybridized to genomic tiling arrays with a median probe spacing of ~300 bp. This yielded reproducible, non-random binding profiles for 42 proteins (Figure 1), while the remaining 37 proteins produced profiles that were poorly reproducible or consisted mostly of random noise. Thus, out of 112 candidates, we identified 42 novel chromatin proteins, each with a non-random genome-wide binding profile. Thirty-five of these were only known by a CG-identifier. We will refer to these proteins as CC1 (Chromatin Component) through CC35 (Table S2).

A reference set of 70 known chromatin components

In order to provide a frame of reference for the interpretation of the binding profiles of the novel proteins, we also compiled a broad collection of genome-wide maps of known chromatin components, all generated in the same Kc167 cell line. For this we used our previously published set of DamID maps of 53 chromatin proteins and ChIP-on-chip maps for 5 histone marks[7], supplemented with new DamID profiles for 10 additional proteins that are known to be components of chromatin (Figure S1). The latter were chosen to further broaden the spectrum of known chromatin proteins and include among others the chromatin remodeling ATPase ISWI[11], the cohesin subunit RAD21[12], and the SAGA complex subunit ADA2B[13]. We also included two previously published profiles of nuclear pore components, which were obtained by DamID under identical experimental conditions[14]. These 12 profiles extend the collection of known chromatin components mapped in Kc167 cells to a total of 65 proteins and 5 histone marks, which serve as a reference set for the analysis of the 42 novel proteins, as described below.

Baysian Network model of interactions among chromatin components

Next, we used the collection of binding maps to construct a network model of the interplay among all 112 chromatin components. Such a model provides insight into the overall molecular organization of chromatin, but also places each of the novel proteins into a functional context. For this purpose we applied Bayesian Network Inference (BNI)[15-18], a statistical method that previously has been used successfully to model the interactions among smaller sets of known chromatin components[19-22]. We used BNI in combination with a bootstrapping approach to assign a confidence score (ranging from 0 to 100%) to each potential interaction[18,20,23] (Table S3, available upon request). We chose 70% as a cutoff because it yields a network that combines acceptable accuracy with sufficient coverage, as will be

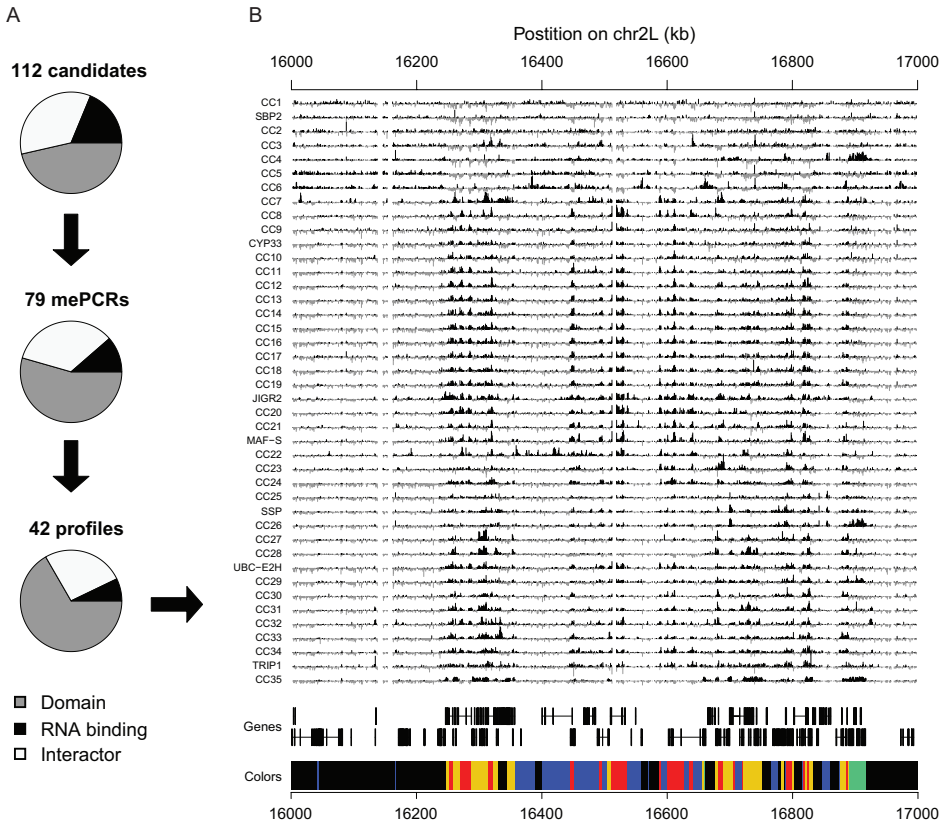


Figure 1 Identification of 42 novel chromatin proteins by generation of DamID binding profiles. (A) Overview of the testing of 112 candidate proteins by DamID. Pie-charts show the representation of the different selection criteria for the initial selected candidate proteins, for proteins which exhibit detectable DNA methylation (mePCR) when fused to Dam and for proteins which yielded a specific and reproducible binding profile. (B) DamID binding profiles of the 42 novel chromatin proteins along a 2Mb region on chromosome 2L. Y-axis depicts $\log_2$ enrichment of Dam-fusion over Dam-only control, positive values are plotted in black and negative values in gray for contrast. Rows were arranged by genome-wide hierarchical clustering. Below the profiles, genes on both strands are depicted as lines with blocks indicating exons. Colored domains represent the previously identified chromatin types[7].

discussed below. This network consists of 98 proteins connected by 141 edges (Figure 2). Fourteen proteins are not connected in this network because their highest confidence scores are below the cut-off, ranging from 22.3% to maximally 64.7%. We tentatively connected each of these proteins to the network by its highest confidence score (dotted lines in Figure 2), resulting into a network of 155 interactions among 112 proteins. We will refer to this network as BN_{70}.

Several biochemical interpretations are possible for the predicted 'interactions' in this network. First, two interacting proteins may be in direct physical contact; second, the proteins may both be part of the same protein complex; third, one

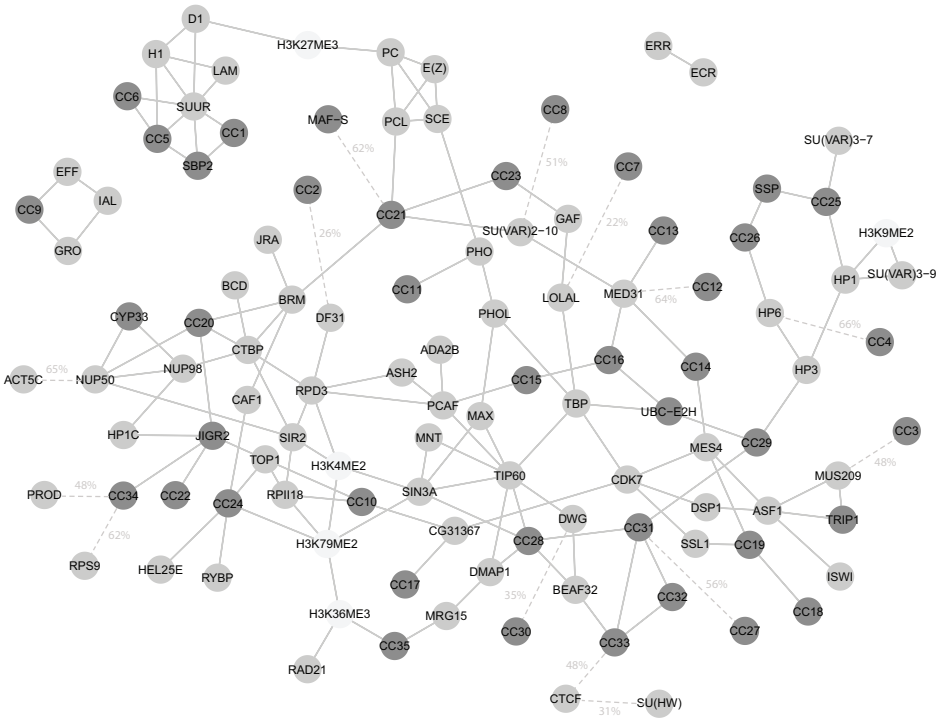


Figure 2 Bayesian Network model of the interactions among 112 novel and known chromatin components. Nodes represent chromatin components. Dark grey, newly identified chromatin proteins; grey, known chromatin proteins; light grey, histone marks. Edges represent predicted interactions with a confidence score of at least 70%. Proteins for which all confidence scores are below 70% are tentatively connected to the network via the edge that has the highest score (dotted lines).

protein may locally assist in the genomic targeting of its partner without being part of the same complex. An example of the latter is a chromatin remodeling protein that facilitates the DNA binding of a transcription factor by removing a nucleosome. While these different types of interactions cannot be discriminated without additional experimental data, all are important for the molecular architecture of chromatin.

BN_{70} recovers many known protein complexes and direct molecular interactions. For example, it connects the four Polycomb Group (PcG) proteins PC, E(Z), ESC and PCL, and recovers the known biochemical interaction of PC with the histone mark H3K27me3. BN_{70} also recapitulates the known direct binding of HP1 to H3K9me2 and SU(VAR)3-9[24-26] and of HP3 to both HP1 and HP6[10,27]. Furthermore, it connects SSL1 and CDK7, which are both subunits of TFIIH[28]; Nup98 and Nup50, which are both nuclear pore proteins[29]; MRG15, DMAP1 and TIP60, which co-purify in one complex[30]; and GAF and LOLAL which interact directly[10]. Because the random chance of two proteins being connected in this network is ~0.025,

these anecdotal observations are unlikely to occur by chance. Rather, they indicate that BN_{70} captures many *bona fide* associations of diverse nature.

Nevertheless, BN_{70} predicts some interactions that may be incorrect. For example, it links MNT via SIN3A to MAX, while it is thought that MNT and MAX form a heterodimer and SIN3A interacts only with MNT[31]. Thus, BNI may have incorrectly modeled the precise local connectivity among the three proteins, even though it correctly clustered them within the entire network. Thus, individual predicted interactions must be treated with caution.

In order to assess the overall reliability of BN_{70}, we systematically investigated the level of agreement with published literature. The curated literature database of FlyBase[32] lists more than 8,500 papers that report on one or more of the 65 known *Drosophila* proteins in our dataset (publications on histone marks are not tracked by FlyBase). A computational survey of this database reveals that at a 70% cutoff pairs of known proteins directly linked are about 10 times more frequently co-cited than those which are not linked (Figure S2A). As an indication of the recovery of known interactions, 58% of the linked pairs among the 65 known proteins are co-cited at least twice in the FlyBase literature list (Figure S2B). This provides support that these protein pairs are indeed biochemically or functionally related.

We also compared the predicted interactions in BN_{70} to the DroID database of protein-protein interactions[33]. DroID lists 33 high-confidence physical interactions among the known proteins that we mapped. Of these interactions, 11 are also present in BN_{70}, which is ~11 times more than may be expected by chance (Figure S2C). Similarly, BN_{70} is about 9.5-fold enriched for *genetic* interactions listed in the BioGRID database[34] (Figure S2D), indicating that the network not only reflects biochemical interactions, but also functionally relevant interactions. We note that full recovery of previously reported interactions should not be expected, because many of these may be cell type- or condition-specific, and an unknown fraction may be false-positives. Taken together, the strong congruence with published data points to a high level of accuracy of the predicted interactions in BN_{70}.

Bayesian network model predicts function of novel proteins

The novel chromatin proteins in BN_{70} do not cluster together but are interspersed among the known chromatin proteins (Figure 2). Because directly linked pairs of known proteins tend to be functionally related (see above), we reasoned that novel chromatin proteins are also likely to share functions with neighboring known chromatin proteins in BN_{70}. This 'guilt-by-association' logic yielded several functional predictions, some of which are highlighted here.

DNA replication: the *Drosophila* Proliferating Cell Nuclear Antigen (PCNA) ortholog MUS209[35] and the histone chaperone ASF1[36] are key parts of the DNA replication machinery. Both proteins are linked in BN_{70} to the novel chromatin component TRIP1, which was so far only known as a regulator of protein translation[37]. In addition, CC3 is connected to MUS209, albeit with low confidence

value (48%). We therefore hypothesized that TRIP1 and possibly CC3 are involved in the control of DNA replication. To test this, we studied the effects of knockdown of these proteins by RNA interference (RNAi) (Figure 3A-B). Analysis of cellular DNA content showed that cells lacking TRIP1 accumulate in G1 (Figure 3C) and fail to incorporate 5-ethynyl-2′-deoxyuridine (EdU), indicating that DNA replication is blocked (Figure 3D). No such effect could be found for CC3. We conclude that TRIP1 is essential for S-phase entry, a function that is likely to be related to its close association with both PCNA and ASF1 in chromatin.

DNA repair: Our dataset contains two known chromatin components which are involved in DNA nucleotide excision repair (NER): MES4, a DNA polymerase

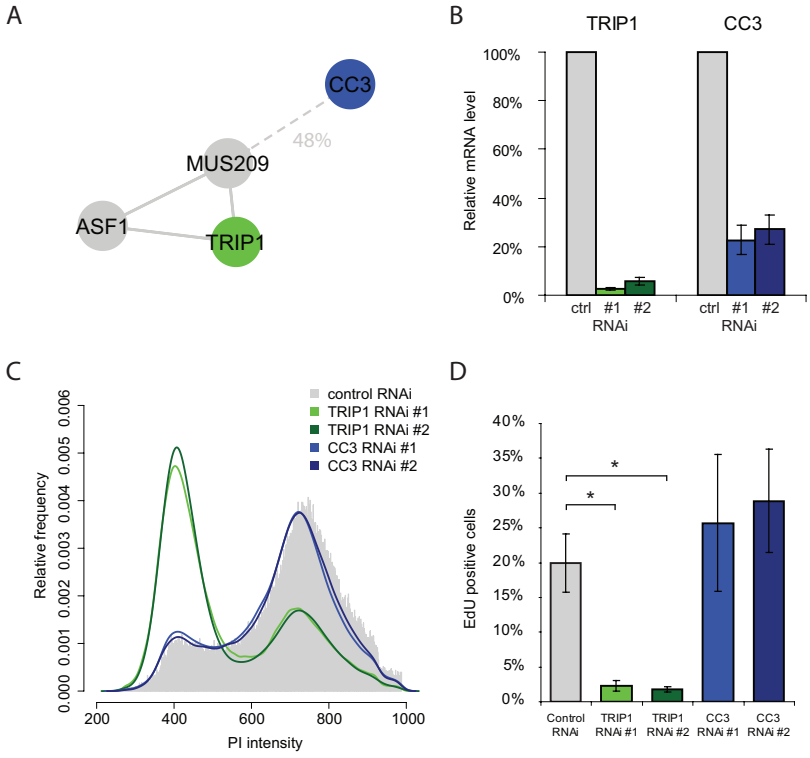


Figure 3 TRIP1 is essential for S-phase entry. (A) BN_{70} connectivity of replication machinery components MUS209 and ASF1, and the novel chromatin components TRIP1 and CC3. (B) Knockdown efficiency determined by quantitative RT-PCR analysis of TRIP1 and CC3 mRNA expression levels, after treatment with dsRNA fragments against the control gene *white* (grey), *Trip1* (green) and *CC3* (blue) with two independent dsRNA fragments each (#1 and #2). Y-axis depicts average mRNA level of three replicates relative to the control knockdown. (C) Cell cycle profile after control knockdown (grey histogram), TRIP1 knockdown (green lines) and CC3 knockdown (blue lines). For each sample ~50.000 cells were counted. X-axis depicts DNA content determined by propidium iodide (PI) labeling and the Y-axis the averaged relative frequency of two replicate experiments. (D) EdU incorporation after TRIP1 knockdown (Fischer's Exact test: * indicates $p<0.05$). Colored bars represent knockdowns as in A. Y-axis depicts average percentage of nucleotide incorporating cells of two replicates. For each treatment 5 fields (about 200-1000 cells) were scored.

epsilon subunit[38] and SSL1, a component of the TFIIH complex which has a dual role in initiation of transcription and NER[39]. Both proteins are connected to CC19, which in turn is linked to CC18. These novel proteins may therefore be involved in NER.

Nuclear pore components: In BN_{70} the novel chromatin proteins CC20 and CYP33 are connected to one or both of the nuclear pore proteins NUP50 and NUP98. The association of CC20 with nuclear pore components is strongly supported by a reported Y2H interaction with another nuclear pore protein, Nup54[33]. Interestingly, CC20 is in turn linked to the transcriptional regulators BRM, CTBP and JRA and may thus help to connect nuclear pore proteins to chromatin.

Histone modifications: Six of the proteins that are known to be involved in histone acetylation and deacetylation are connected to each other in the network, namely PCAF (Histone Acetyltransferase (HAT)), ADA2B (part of HAT complex), TIP60 (HAT), SIR2 (Histone Deacetylase (HDAC)), RPD3 (HDAC) and SIN3A (part of HDAC complex)[40]. This cluster is connected to two novel chromatin proteins, CC28 and CC15. These novel chromatin proteins are thus likely to be involved in histone (de-)acetylation. Furthermore, we find CC35 connected to H3K36me3. Its human ortholog PSIP1, which assists genomic integration of Human Immunodeficiency Virus[41], was recently found to bind directly to H3K36me3 (W. Bickmore, personal communication). It is thus highly probable that CC35 also binds to this histone mark.

Novel proteins in the principal chromatin types

We previously reported that the genome of *Drosophila* Kc167 cells is segmented into five principal chromatin types, which are defined by distinct combinations of proteins. Each of these five types was named after a color: GREEN (HP1-heterochromatin); BLUE (Polycomb chromatin); BLACK (a novel repressive chromatin type that covers nearly 50% of the genome), and YELLOW and RED (transcriptionally active chromatin primarily associated with housekeeping and tissue-specific genes respectively)[7]. These chromatin types were identified by principal component analysis of genome-wide DamID profiles of 53 known chromatin proteins. The same analysis of the 54 newly generated DamID profiles confirms this result (Figure 4A). Only BLUE chromatin is not as prominent, which we attribute to the absence of novel BLUE proteins in the new dataset (see below). We conclude that a completely independent set of proteins largely confirms the classification into five principal chromatin types, indicating that this classification is robust. An implication of this finding is that it is highly unlikely that additional principal chromatin types exist in Kc167 cells.

Because each chromatin type has unique functional properties[7], we studied the contributions of the novel proteins to these chromatin types. Figure 4B-F visualizes the individual protein occupancies in each chromatin type as five color shades projected onto BN_{70}. The five colors each occupy mostly coherent territories in the network, indicating that the overall organization of BN_{70} reflects the partitioning into 5 chromatin types. Only

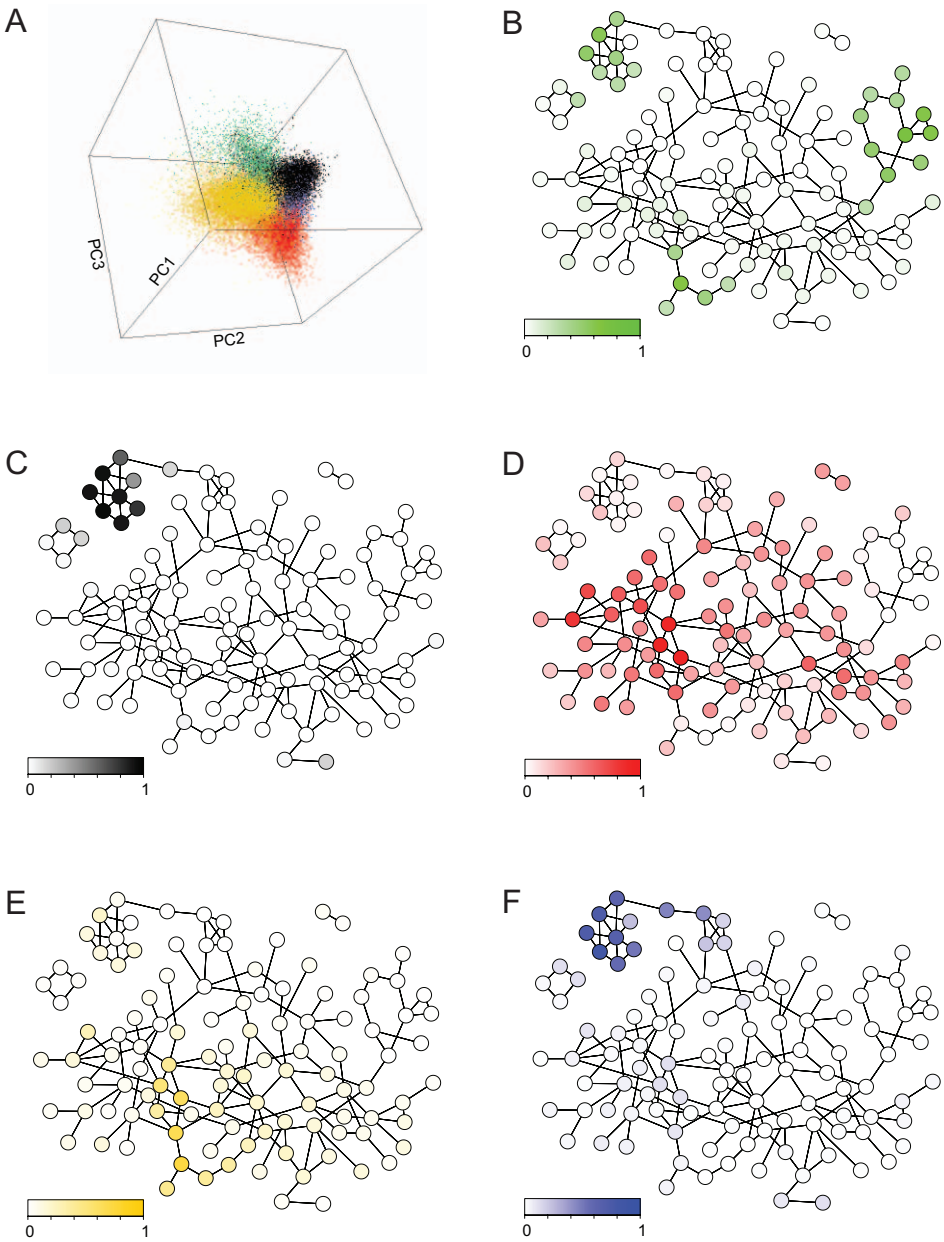


Figure 4 Novel proteins in the principal chromatin types. (A) Principal component analysis of the newly generated binding data of the 42 novel and the 12 known chromatin proteins. Figure represents three dimensional projection of the first three principal components. Each dot represents a probed locus. Loci are colored according to the previously identified chromatin types[7]. Separation of the colors confirms the previous classification into five known chromatin types. (B-F) Occupancy of all mapped chromatin components within each of the five principal chromatin types, projected onto BN_{70} as color scales. A value of 1 on the color scale means that 100% of the genomic loci in this chromatin type are bound by the protein represented by the node.

GREEN chromatin appears to be divided over three separate sub-regions, which may warrant subdivision of GREEN chromatin into distinct subtypes. However, three-dimensional visualization of the entire BN shows that the GREEN subregions cluster at one side of the network (Supplementary Movie 1, available upon request). Note that several proteins are shared among multiple chromatin types, illustrating the combinatorial nature of the chromatin types. For example, histone H1 is abundantly present in BLACK, BLUE, as well as GREEN chromatin.

This color visualization links many of the novel proteins to specific chromatin types. For example, it identifies five novel chromatin proteins that are abundant in GREEN chromatin (CC4, CC26, SSP, CC25, CC29, clustered in the upper-right corner of the network) (Figure 4B). GREEN chromatin corresponds to classic heterochromatin, which is enriched in pericentric regions. Remarkably, CC4 contains a chromodomain, a class of protein domains that is also present in the GREEN marker proteins HP1, HP6 and Su(var)3-9; while CC25 harbors a BESS domain, which is a homodimerization domain that is also present in SU(VAR)3-7, to which it is directly connected in BN70. Moreover, a yeast-2-hybrid screen found CC26 to exhibit a moderately strong interaction with HP1[10].

To further investigate these five proteins, we conducted microscopy localization studies with GFP (Green Fluorescent Protein) fusions. This showed that CC26 and CC4 localize preferentially in the chromocenter (Figure S3), which consists of HP1-rich pericentric heterochromatin. SSP is concentrated in smaller speckles which also generally overlap with the chromocenter. In the course of our studies, SSP was found to localize in subnuclear aggregates reminiscent of chromocenters, dependent on Wingless signaling[42]. We found CC25 in the chromocenter as well as in euchromatin, and CC29 primarily in euchromatin. We attribute the predominant euchromatic localization of CC29 to the connection of this protein to components of RED chromatin, which is a type of euchromatin.

Because proteins that associate with pericentric chromatin often depend on HP1 for their localization[27,43], we tested whether the chromocenter association of GFP tagged CC26 and CC4 was disrupted upon depletion of HP1 by RNAi. Interestingly, CC26 was lost from the chromocenter after knockdown of HP1, while CC4 exhibited no significant change in its nuclear distribution (Figure S5). From this we conclude that HP1 is required for recruitment of CC26 but not CC4 to pericentric heterochromatin.

BLACK chromatin is a largely unexplored repressive type of chromatin that covers nearly half of the non-repetitive genome in *Drosophila* Kc167 cells. Little is known about the molecular composition of BLACK chromatin[7]. Four of the new chromatin proteins are enriched in BLACK chromatin. Together with the previously identified BLACK proteins D1, histone H1, SUUR and LAM, they form a separate sub-network with a high internal connectivity (Figure S4C), suggesting that these eight members of BLACK chromatin may form one large complex. Interestingly, two novel BLACK proteins, SBP2 and CC5, are evolutionarily conserved RNA-binding proteins, suggesting

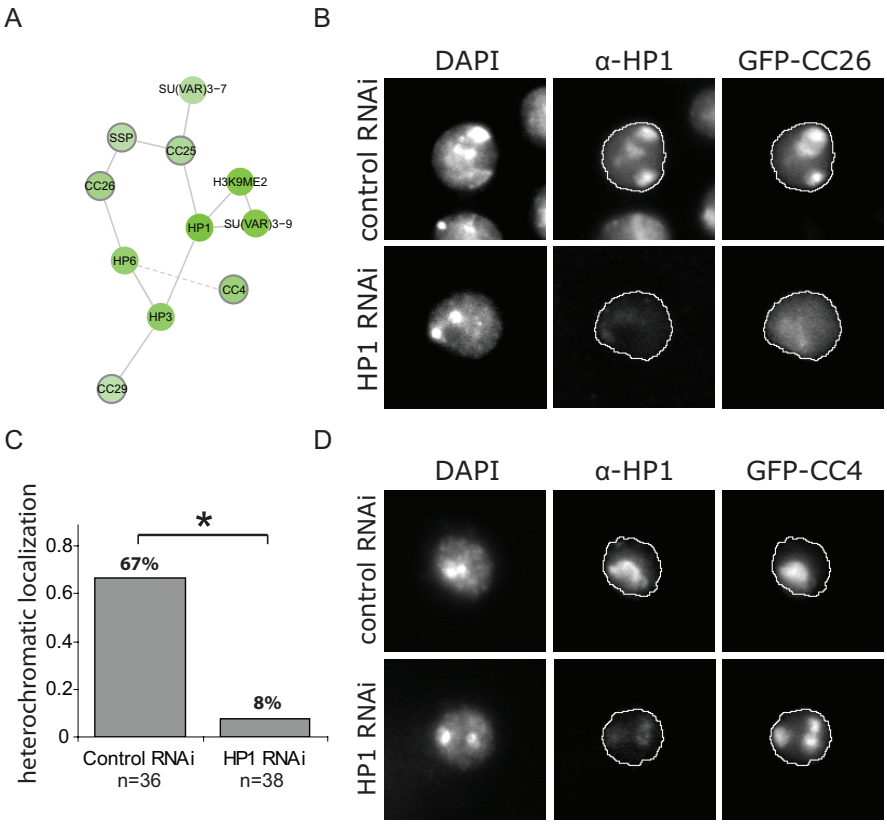


Figure 5 HP1 is required for recruitment of CC26 to pericentric heterochromatin. (A) Cluster of GREEN chromatin including five novel proteins (grey border). Green color scale represents the occupancy of each protein within GREEN chromatin as in Figure 4B. (B) Microscopy images of Kc167 cells transfected with GFP tagged CC26 and stained with anti-HP1 antibody and DAPI, after knockdown of *white* (negative control) (top panel) or of HP1 (bottom panel). White lines indicate the nuclear rim (based on DAPI signal). (C) Percentage of cells with chromocenter enrichment of GFP-CC26, scored in a blind experimental setup; * indicates $p<0.05$ according to Fischer's Exact test; n, numbers of cells scored. Qualitative analysis of an independent replicate experiment showed the same altered localization pattern of CC26 after HP1 knockdown (data not shown). (D) Chromocenter localization of GFP-CC4. HP1 knockdown does not result in a detectable loss of this pattern.

a role for RNA in BLACK chromatin. CC5 is an ortholog of budding yeast Nip7, a conserved nucleolar protein involved in ribosome biogenesis[44]. The third protein, CC6, contains a C2H2-type zinc-finger, which could mediate direct binding to DNA. Finally, CC1 is an ortholog of human DCAF12, which interacts with cullin/RING E3 ligases that are part of the COP9 signalosome (CSN), which in turn has been implicated in chromatin-mediated gene repression[45]. Taken together, these four proteins provide several new clues to the nature and molecular composition of BLACK chromatin.

RED and YELLOW chromatin primarily cover transcriptionally active parts of the genome. They have partially over-

lapping protein compositions (Figure 4D-E). Previously, only one protein was found (MRG15) that specifically binds in YELLOW and not in RED chromatin[7]. The new dataset reveals two proteins with the same characteristic; one novel chromatin protein, CC35, and the cohesin subunit RAD21. Both proteins are linked to H3K36me3, the YELLOW-specific histone mark, in BN_{70}. About half of all novel proteins show a strong enrichment in RED chromatin. This strengthens the notion that RED chromatin has an extremely high compositional complexity[7].

None of the novel proteins are specifically enriched in BLUE chromatin (Figure 4F). This chromatin type has been extensively analyzed (reviewed in [46,47-48]), perhaps leaving very few BLUE proteins to be discovered. In summary, this analysis provides new insights into the molecular architecture of four out of five principal chromatin types.

Predicting gene regulation function

Next, we investigated possible roles of the novel proteins in gene regulation. We first analyzed the expression levels of the target genes of each protein in Kc167 cells (digital gene expression from [7]). Low activity of its direct target genes may indicate a repressive function of a chromatin protein; conversely, high activity of the target genes may point to an activating role. Visualization of the average target gene expression levels (Figure 6A) shows that the clusters of BLACK and BLUE chromatin proteins have the lowest target gene expression levels, consistent with a repressive function[7].

A more accurate prediction of the regulatory role of a protein may be derived from the developmental expression dynamics of each protein in relation to its target genes. For a *bona fide* activator, it is expected that its own expression throughout development is positively correlated with the expression levels of its target genes, and conversely for a repressor. Using this logic to identify candidate activators and repressors, we tested each protein for significant positive or negative correlation with its target genes across 17 tissues[49] (Figure 6B). For example, PC and JRA, which are well-known repressive and activating proteins, show a negative and positive correlation respectively (red curves in Figure 6B). Other predicted repressors are, as expected, mainly components of BLACK and BLUE chromatin (such as CC5) while activators are mainly part of RED and YELLOW chromatin (such as CC35) (Figure 6B-C). Among the novel chromatin proteins this strategy identified 16 candidate activators and 5 candidate repressors.

Network compartmentalization of biological functions

Finally, we reasoned that functions of a chromatin protein may be predicted by Gene Ontology (GO) analysis of the collection of genes bound by this protein. We systematically conducted GO enrichment analyses, focusing on the main GO branch "biological process", because this tends to be the most informative in terms of cellular and organismal physiology. Of the 112 mapped proteins and histone marks, 108 yielded at least one significantly enriched GO category among their respective target gene sets. Remarkably, visualization of specific GO category enrichments in BN_{70} typically highlights clusters of connected

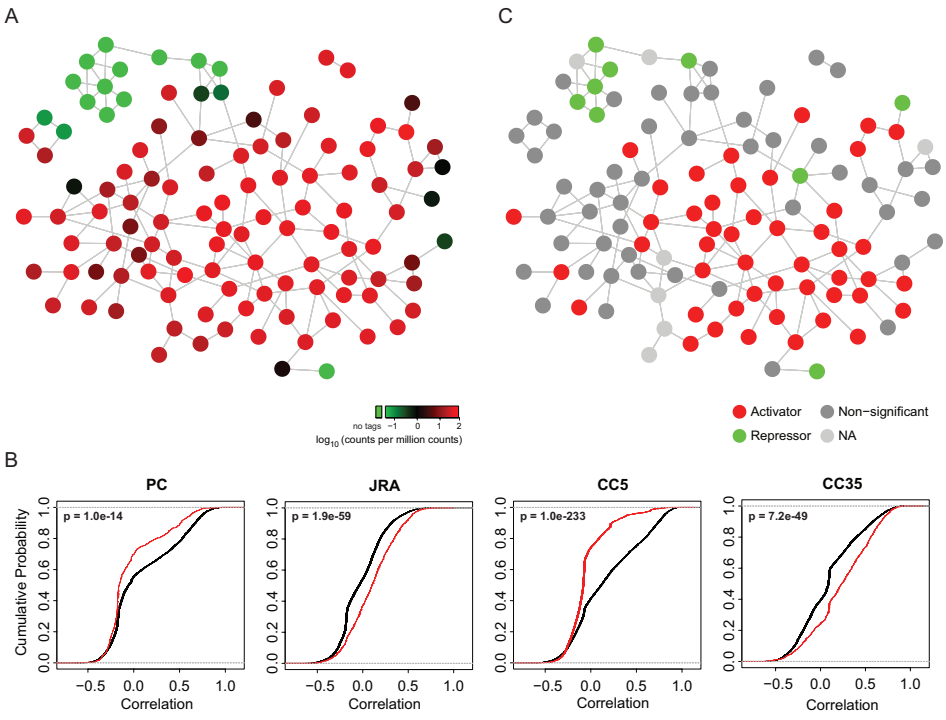


Figure 6 Predicting regulatory function. (A) Average mRNA expression levels of the target genes of each protein in BN_{70} plotted as a color scale. (B) Examples of cumulative probability plots of the correlations between the expression level of a chromatin protein and each of its target genes across 17 tissues and developmental stages[49] (red curves). The same is done for the correlation with each non-target gene (black curves). A significant shift of the red curve to the left or right relative to the black curve predicts repressive (PC and CC5) or activating (JRA and CC35) functions, respectively. (C) BN_{70} highlighting predicted activators (red), putative repressors (green), non-significant proteins (dark grey), and histone marks (light (grey) ($p < 0.0001$).

proteins within the network (Figure 7 and Figure S4, available upon request), indicating that sets of interconnected proteins within the chromatin network are dedicated to the regulation of specific biological processes. These clusters vary strongly in size between the GO categories. A striking example is "neurological system process", which largely coincides with BLACK chromatin (Figure 7A). A similarly confined cluster of several PcG proteins together with CC11 and CC21 is enriched for genes involved in "digestive tract development" (Figure 7B). In contrast, the GO category "embryo development" is covered by a much larger group of chromatin proteins (Figure 7C), as may be expected for this much broader function. The function "tRNA metabolic process" coincides largely with the core of YELLOW chromatin (Figure 7D), consistent with the previously noted role of YELLOW chromatin in housekeeping functions[7]. Interestingly, "stem cell differentiation" and "cell proliferation" exhibit strongly overlapping enrichments (Figure

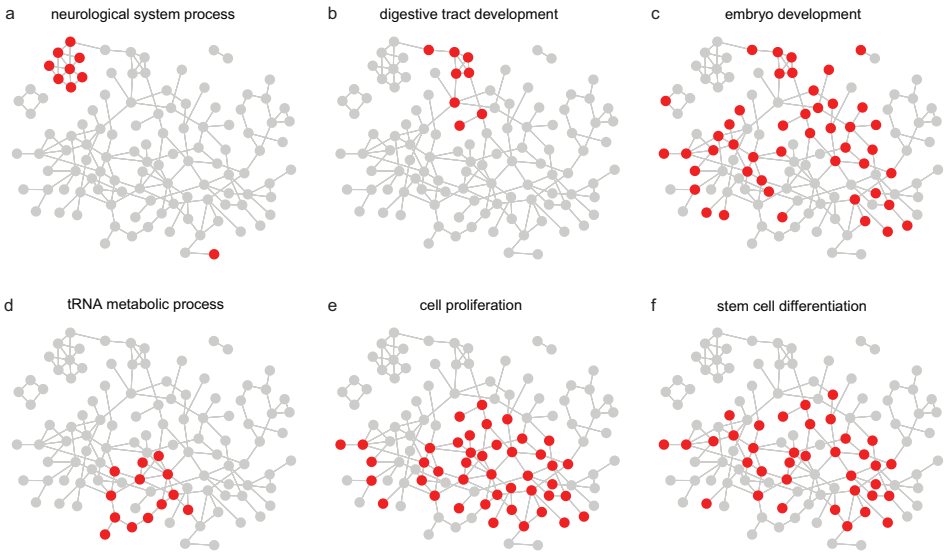


Figure 7 Functional compartmentalization of BN_{70}. (A-F) Six examples illustrating the enrichment of GO categories among the target genes of clusters of proteins in BN_{70}. Proteins with their target genes significantly enriched for the indicated GO category are depicted in red. A complete overview of all significant GO categories is shown in Supplementary Figure 4.

7E-F), suggesting tightly coordinated regulation of these two processes by a large cluster of chromatin proteins that includes several novel proteins. In total, 40 of the novel proteins are enriched for specific GO categories, providing predictions of their regulatory functions (Table S4, available upon request). Taken together, this analysis reveals extensive functional compartmentalization of the chromatin protein network.

| DISCUSSION

The high complexity of the chromatin proteome poses a major obstacle towards the understanding of the overall molecular architecture of chromatin. On one hand, it is crucial that the full complement of chromatin proteins in a cell is systematically identified and characterized. On the other hand, data-driven computational methods are required for the construction of global models of the chromatin interactome. The strategy described here begins to address both needs.

First, by taking advantage of the DamID method, we have compiled one of the broadest compendia of genome-wide maps of chromatin protein binding in a metazoan cell. We nearly doubled the available collection of binding maps in Kc167 cells[7,14] and included maps for dozens of novel proteins. Second, we have used BNI to place each known and novel protein into its molecular context of putative partner proteins, from which functional predictions can be derived. Through

this approach we identified TRIP1 to be involved in DNA replication and predicted candidate proteins to be involved in DNA damage repair, gene regulation by nucleoporins, binding to modified histones, and other functions. We emphasize that a link between two proteins in BN_{70} does not necessarily imply a direct physical interaction; in reality a 'hidden' (not yet mapped) protein or protein complex may bridge the two proteins or two proteins can bind to the same loci but not at the same point in time. With further expansion of the collection of mapped proteins (both known and novel), the network model is expected to become more fine-grained.

The network model not only predicts functions of individual proteins, it also provides insight into the overall molecular organization of chromatin. At a global level, BN_{70} recapitulates the compartmentalization into the five principal chromatin types that we reported previously[7]. It also reveals how these chromatin types are interconnected and partially overlap. Furthermore, it provides a first draft of the topology of the protein connections within each chromatin type. For example, it indicates a very high connectivity among the proteins that define BLACK chromatin, suggesting that these proteins form one complex. In contrast, RED chromatin consists of a vast network of proteins, containing specific sub-networks (Figure 4) which coincide with the compartmentalization of the GO enrichments. It is striking that most GREEN chromatin components are not associated with specific GO categories. This suggests that the regulatory role of GREEN chromatin in specific biological processes is minor relative to that of the other four chromatin types. In contrast, BLACK chromatin is linked to several specific GO categories, and four of the BLACK proteins (including the novel proteins CC5 and SBP2) are predicted to have a repressive function (Figure 6). These results underscore the notion that BLACK chromatin is a repressive chromatin type dedicated to restricting the expression of tissue-specific gene sets.

In summary, this study substantially extends our understanding of the molecular network underlying *Drosophila* chromatin, and provides a general framework for similar systematic analyses of chromatin in other species.

| MATERIALS AND METHODS

Constructs

DamID constructs used in this study are described in Table S2. The full length open reading frame (ORF) of ISWI was cloned into pDamMyc[50]. pDamMycRad21 was described previously[12]. For the other chromatin proteins, full length ORFs (*Drosophila* Genomics Resource Centre) were cloned into an acceptor vector using Cre-mediated recombination as described[7]. To obtain GFP fusion proteins of the putative GREEN chromatin proteins the ORFs were amplified and cloned in-frame with GFP in a vector containing an Act5C promoter by using TOPO cloning and GATEWAY recombination as described[51].

DamID and Microarrays

DamID assays were carried out in Kc167 cells under standardized conditions as described[7]. DNA adenine methylation levels by the fusion protein were qualitatively estimated by analysing the products of the PCR amplification of the methylated fragments on an agarose gel[9]. We only continued with microarray hybridization if Dam-fusion proteins yielded a smear of methylated fragments stronger than the background smear obtained after transfection of cells with a plasmid lacking the Dam ORF. The identity of the hybridized material was tracked and verified as described[7].

Knockdown by RNAi

RNAi experiments were performed using dsRNAs directed against the gene products of interest. dsRNAs were in vitro transcribed using the RiboMax kit (Promega) from PCR amplicons. PCR amplicons used were published before (white and HP1 [52]) or designed by the Harvard Drosophila RNAi Screening Centre (www.flyrnai.org, BKN46519 and VDRC14057 against TRIP1 and 62037 and VDRC11589 against CC3). Knockdown of HP1 was performed as described in [52]. After 3 days cells were transfected with GFP-fusion constructs. HP1 knockdown was monitored by fluorescence microscopy using the C1A9 anti-HP1 antibody (Developmental Studies Hybridoma Bank, University of Iowa). Randomized images were blindly scored for localisation patterns of the GFP fusion proteins.

Knockdown of TRIP1 and CC3 was done in S2-DRSC cells (an isolate of S2 specialized for RNAi, Drosophila Genome Resource Center). 2 million cells in 1ml serum free medium (BPYE+ 1:100 PenStrep + 2mM L-Glutamine) were treated with ~50ug dsRNA and incubated for one hour. After 1 hour 1ml of medium with 10% serum was added. After 5 days cells were harvested for subsequent analysis. To monitor knockdown mRNA was isolated with TRIzol (Invitrogen), treated with DNAseI and reverse transcribed to allow analysis by qPCR using CyberGreen. For normalization amplicons for Twinstar, Fmo-2 and GAPDH2 were used. For cell-cycle profiling cells were fixed o/n in 70% EtOH, RNase treated and stained with propidium iodide. DNA content was determined with a FACS Calibur and results were analysed using FCS express. To monitor nucleotide incorporation RNAi treated cells were incubated after 5 days with 0.5uM EdU (Click-It system, Invitrogen) for 1.5hrs. Cells further treated as described in the Click-iT imaging manual using Alexa 647.

Data analysis

Microarray data. Normalization and analyses were performed with R (R Development Core Team, 2011). Raw data from two biological replicates were loess-normalized, median-centered, and dye swap arrays were averaged. The data was further processed only if it passed the following quality control criteria, based on experience with ~ 250 hybridizations on this microarray platform. 1) Microarray hybridizations were spiked with oligonucleotides giving a fingerprint that had to be matched on the MA plot. 2) The correlation between replicates had to be higher than 0.3. 3) The autocorrelation between neighboring probes at lag 2 had to be higher than 0.3. 4) Both replicates must cluster together in a hierarchical cluster-

ing involving all ~ 250 experiments. For the latter analysis we filtered the DamID scores with a running median (window width 3) and used the 1-r dissimilarity measure for clustering, where r is the correlation coefficient between experiments.

Target identification. We discretized the profiles into *target* and *non target* loci using a 3-state Hidden Markov Model (HMM) modified from [7]. In the course of this study, it became apparent that many protein profiles feature three types of domains: enriched, intermediate and depleted. The 2-state HMM randomly identified intermediate domains as *targets* or *non targets*. In contrast, the 3-state HMM never failed to identify the intermediate state and returned sensible calls. When run on profiles previously discretized by a 2-state HMM (such as PC or H3K27me3), the 3-state HMM gave good concordance on high confidence targets, showing that this discretization scheme is conservative. Even though we do not fully grasp the biological significance of the intermediate states, the 3-state HMM was adopted for being both general and stringent.

BNI. BNI was performed as described[20] using Banjo version 2.0.0 (http://www.cs.duke.edu/~amink/software/banjo/), with the following modifications. Input data consisted of binary scores generated for each individual protein by a 3-state HMM as above, scoring the 'high' state as 1 and the 'medium' and 'low' states as 0. For each BNI bootstrap iteration, 15,000 genomic positions were randomly sampled from a total of 180,575, without replacement; thus, the average genomic sampling interval is 7896bp, which ensures approximate independence between neighboring data points because the resolution of DamID is ~1-2kb. A file listing the Banjo parameter settings is provided as Table S5 (available upon request). Bootstrap scores for each protein pair are provided in Table S3 (available upon request). The graph representation of the network was generated using Cytoscape[53]. A Force-Directed Layout was applied on all protein pairs with a bootstrap score above 70%. Bootstrap values were used to determine the weight for the repelling and attraction force between the nodes. Force-Directed Layout settings were as follows: default spring coefficient 1e-4, default spring length 50, default node mass 3.0, number of iterations 100. To prevent overlapping nodes and edges, node positions were subsequently manually edited, taking into account the weight of the edges. Proteins without any bootstrap value above 70% were manually connected at the outside of the network according to their highest bootstrap value. The layout file containing node coordinates was exported and plotted in R using the network package.

For 3D visualization of the network, we used BioLayout Express version 2.1 for Mac with default settings to calculate the optimal edge-score weighted 3D topology. As input the entire matrix of BNI bootstrap scores was used (including edges below the 70% threshold), except that scores <4% were set to 0 to prevent excessive influence of background noise. The layout file containing node coordinates was exported and used to generate a 3D movie of the network using the rgl() R package.

Analysis of overlap with public databases. We downloaded from FlyBase the list (version 17 Oct 2011) of all papers (FBrf identifiers) reporting on one or

more of the 107 proteins (or their corresponding genes) in our dataset. Histone marks were not considered because they are not explicitly tracked by FlyBase. We removed publications describing more than 8 proteins, because these papers typically report very long lists of proteins or genes that are not specific enough for our study. For each possible pair of known proteins we then counted the number of unique FBrf entries that these two proteins share. For all protein pairs we calculated a co-citation score, which is defined as the number of publications that report on both proteins, divided by the product of the total numbers of publications for each of the two proteins. This normalization adjusts for the fact that frequently studied proteins are more likely to be co-cited than rarely studied proteins. For comparison of BN_{70} to a database of physical protein-protein interactions, we downloaded protein interaction confidence scores from http://www.droidb.org[33] (version 2011_08) and only considered interactions with confidence scores >0.41, a previously estimated optimal cutoff[54]. For a list of previously reported genetic interactions we used BioGRID database[34] version 3.1.81.

Predicting gene regulatory function. We determined the correlation between the expression level of a protein and the expression level of each target gene over 17 tissues[49]. The same is done for the correlation with each non-target gene. In order to test if the expression level of a protein has a positive or negative correlation with the expression of its target genes we have statistically tested (Wilcoxon-test) the shift between the curves and corrected for multiple testing. In addition we only considered a correlation to be significant in case it also shifted compared to 0, i.e. the correlation with the target genes changed instead of the correlation with the non-target genes.

Gene ontology analysis. Gene mapping and GO-gene association information was obtained from FlyBase release 5.41 (ftp://flybase.org). The Gene Ontology OBO specifications were downloaded from http://archive.geneontology.org/latest-termdb/go_daily-termdb.obo-xml.gz on October 20, 2011. Genes were considered 'target' of a chromatin protein if their transcription start site was located < 1 kb from a target locus (see target identification). To focus on the general biological processes, only the terms annotated 'generic slim', or with a direct ancestor annotated 'generic slim', were considered (307 terms in total). Genes without GO annotations or not mapped on the microarray platform were removed from the background set. Enrichment was considered significant if the Benjamini-Hochberg-corrected p-value of the hypergeometric test was lower than 0.05. Input files, scripts and reproducibility walkthrough are available for download from http://research.nki.nl/vansteensellab/.

Data availability. DamID data have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession number GSE36175.

| ACKNOWLEDGEMENTS

We thank Celine Moorman for cloning of pDamMycISWI, Andrea Pauli for providing pDamMycRad21, Wim Brugman and Ron Kerkhoven for microarray hybridizations and members of our laboratory for helpful discussions. This research was supported by the Netherlands Genomics Initiative and NWO-ALW VICI.

| AUTHOR CONTRIBUTIONS

JGvB and BvS prepared the manuscript. JGvB, GJF and BvS designed the experiments. JGvB, GJF, WT and AR performed the experiments. GJF designed and performed the bioinformatics analyses with the help of JGvB and BvS.

| REFERENCES

1. Kharchenko, P.V. et al. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. Nature 471, 480-5 (2011).
2. Roy, S. et al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science 330, 1787-97 (2010).
3. Roudier, F. et al. Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. EMBO J 30, 1928-38 (2011).
4. Liu, T. et al. Broad chromosomal domains of histone modification patterns in C. elegans. Genome Res 21, 227-36 (2011).
5. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol 28, 817-25 (2010).
6. Ernst, J. et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 473, 43-9 (2011).
7. Filion, G.J. et al. Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. Cell 143, 212-24 (2010).
8. van Steensel, B., Delrow, J. & Henikoff, S. Chromatin profiling using targeted DNA adenine methyltransferase. Nat Genet 27, 304-8 (2001).
9. Greil, F., Moorman, C. & van Steensel, B. DamID: mapping of in vivo protein-genome interactions using tethered DNA adenine methyltransferase. Methods Enzymol 410, 342-59 (2006).
10. Giot, L. et al. A protein interaction map of Drosophila melanogaster. Science 302, 1727-36 (2003).
11. Sala, A. et al. Genome-wide characterization of chromatin binding and nucleosome spacing activity of the nucleosome remodelling ATPase ISWI. EMBO J 30, 1766-77 (2011).
12. Pauli, A. et al. A direct role for cohesin in gene regulation and ecdysone response in Drosophila salivary glands. Curr Biol 20, 1787-98 (2010).
13. Muratoglu, S. et al. Two different Drosophila ADA2 homologues are present in distinct GCN5 histone acetyltransferase-containing complexes. Mol Cell Biol 23, 306-21 (2003).
14. Kalverda, B., Pickersgill, H., Shloma, V.V. & Fornerod, M. Nucleoporins directly stimulate expression of developmental and cell-cycle genes inside the nucleoplasm. Cell 140, 360-71 (2010).
15. Pearl, J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference (Morgan Kaufmann, San Francisco, 1988).
16. Cooper, G.F. & Herskovits, E. Machine Learning: A Bayesian method for the induction of probabilistic networks from data 309-347 (1992).
17. Friedman, N. Inferring cellular networks using probabilistic graphical models. Science 303, 799-805 (2004).

18. Pe'er, D. Bayesian network analysis of signaling networks: a primer. Sci STKE 2005, pl4 (2005).
19. Yu, H., Zhu, S., Zhou, B., Xue, H. & Han, J.D. Inferring causal relationships among different histone modifications and gene expression. Genome Res 18, 1314-24 (2008).
20. van Steensel, B. et al. Bayesian network analysis of targeting interactions in chromatin. Genome Res 20, 190-200 (2010).
21. Lv, J. et al. Discovering cooperative relationships of chromatin modifications in human T cells based on a proposed closeness measure. PLoS One 5, e14219 (2010).
22. Cui, X.J., Li, H. & Liu, G.Q. Combinatorial patterns of histone modifications in Saccharomyces cerevisiae. Yeast 28, 683-91 (2011).
23. Friedman, N., Linial, M., Nachman, I. & Pe'er, D. Using Bayesian networks to analyze expression data. J Comput Biol 7, 601-20 (2000).
24. Lachner, M., O'Carroll, D., Rea, S., Mechtler, K. & Jenuwein, T. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. Nature 410, 116-20 (2001).
25. Bannister, A.J. et al. Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. Nature 410, 120-4 (2001).
26. Schotta, G. et al. Central role of Drosophila SU(VAR)3-9 in histone H3-K9 methylation and heterochromatic gene silencing. EMBO J 21, 1121-31 (2002).
27. Greil, F., de Wit, E., Bussemaker, H.J. & van Steensel, B. HP1 controls genomic targeting of four novel heterochromatin proteins in Drosophila. EMBO J 26, 741-51 (2007).
28. Aoyagi, N. & Wassarman, D.A. Genes encoding Drosophila melanogaster RNA polymerase II general transcription factors: diversity in TFIIA and TFIID components contributes to gene-specific transcriptional regulation. J Cell Biol 150, F45-50 (2000).
29. Suntharalingam, M. & Wente, S.R. Peering through the pore: nuclear pore complex structure, assembly, and function. Dev Cell 4, 775-89 (2003).
30. Kusch, T. et al. Acetylation by Tip60 is required for selective histone variant exchange at DNA lesions. Science 306, 2084-7 (2004).
31. Loo, L.W. et al. The transcriptional repressor dMnt is a regulator of growth in Drosophila melanogaster. Mol Cell Biol 25, 7078-91 (2005).
32. Drysdale, R. FlyBase : a database for the Drosophila research community. Methods Mol Biol 420, 45-59 (2008).
33. Murali, T. et al. DroID 2011: a comprehensive, integrated resource for protein, transcription factor, RNA and gene interactions for Drosophila. Nucleic Acids Res 39, D736-43 (2011).
34. Stark, C. et al. The BioGRID Interaction Database: 2011 update. Nucleic Acids Res 39, D698-704 (2011).
35. Henderson, D.S., Banga, S.S., Grigliatti, T.A. & Boyd, J.B. Mutagen sensitivity and suppression of position-effect variegation result from mutations in mus209, the Drosophila gene encoding PCNA. EMBO J 13, 1450-9 (1994).
36. Schulz, L.L. & Tyler, J.K. The histone chaperone ASF1 localizes to active DNA replication forks to mediate efficient DNA replication. FASEB J 20, 488-90 (2006).
37. Lasko, P. The drosophila melanogaster genome: translation factors and RNA binding proteins. J Cell Biol 150, F51-6 (2000).
38. Ravi, D. et al. A network of conserved damage survival pathways revealed by a genomic RNAi screen. PLoS Genet 5, e1000527 (2009).
39. Wang, Z. et al. The yeast TFB1 and SSL1 genes, which encode subunits of transcription factor IIH, are required for nucleotide excision repair and RNA polymerase II transcription. Mol Cell Biol 15, 2288-93 (1995).
40. Kuo, M.H. & Allis, C.D. Roles of histone acetyltransferases and deacetylases in gene regulation. Bioessays 20, 615-26 (1998).
41. Llano, M. et al. An essential role for LEDGF/p75 in HIV integration. Science 314, 461-4 (2006).
42. Taniue, K. et al. Sunspot, a link between Wingless signaling and endoreplication in Drosophila. Development 137, 1755-64 (2010).
43. Delattre, M., Spierer, A., Tonka, C.H. & Spierer, P. The genomic silencing of position-effect variegation in Drosophila melanogaster: interaction between the

heterochromatin-associated proteins Su(var)3-7 and HP1. J Cell Sci 113 Pt 23, 4253-61 (2000).
44. Zanchin, N.I., Roberts, P., DeSilva, A., Sherman, F. & Goldfarb, D.S. Saccharomyces cerevisiae Nip7p is required for efficient 60S ribosome subunit biogenesis. Mol Cell Biol 17, 5001-15 (1997).
45. Chamovitz, D.A. Revisiting the COP9 signalosome as a transcriptional regulator. EMBO Rep 10, 352-8 (2009).
46. Sparmann, A. & van Lohuizen, M. Polycomb silencers control cell fate, development and cancer. Nat Rev Cancer 6, 846-56 (2006).
47. Morey, L. & Helin, K. Polycomb group protein-mediated repression of transcription. Trends Biochem Sci 35, 323-32 (2010).
48. Sawarkar, R. & Paro, R. Interpretation of developmental signaling at chromatin: the Polycomb perspective. Dev Cell 19, 651-61 (2010).
49. Chintapalli, V.R., Wang, J. & Dow, J.A. Using FlyAtlas to identify better Drosophila melanogaster models of human disease. Nat Genet 39, 715-20 (2007).
50. van Steensel, B. & Henikoff, S. Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. Nat Biotechnol 18, 424-8 (2000).
51. Braunschweig, U., Hogan, G.J., Pagie, L. & van Steensel, B. Histone H1 binding is inhibited by histone variant H3.3. EMBO J 28, 3635-45 (2009).
52. Greil, F. et al. Distinct HP1 and Su(var)3-9 complexes bind to sets of developmentally coexpressed genes depending on chromosomal location. Genes Dev 17, 2825-38 (2003).
53. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13, 2498-504 (2003).
54. Yu, J. & Finley, R.L., Jr. Combining multiple positive training sets to generate confidence scores for protein-protein interactions. Bioinformatics 25, 105-11 (2009).

SUPPLEMENTARY DATA

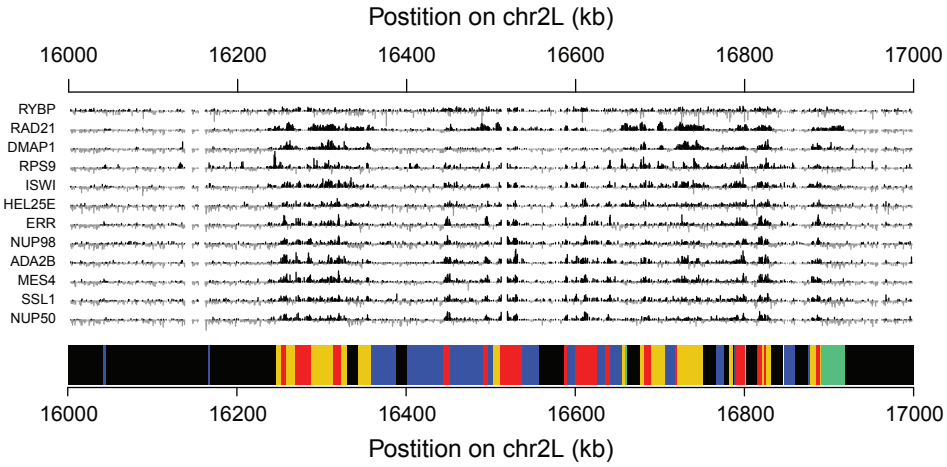


Figure S1 DamID binding profiles of 12 known chromatin components along a 2Mb region on chromosome 2L. Y-axis depicts $\log_2$ enrichment over Dam-only control, positive values are plotted in black and negative values in gray for contrast. Colored domains epresent the previously identified chromatin types (*Filion et al., Cell 2010*). Nup98 and Nup50 data are from (*Kalverda et al., Cell 2010*)

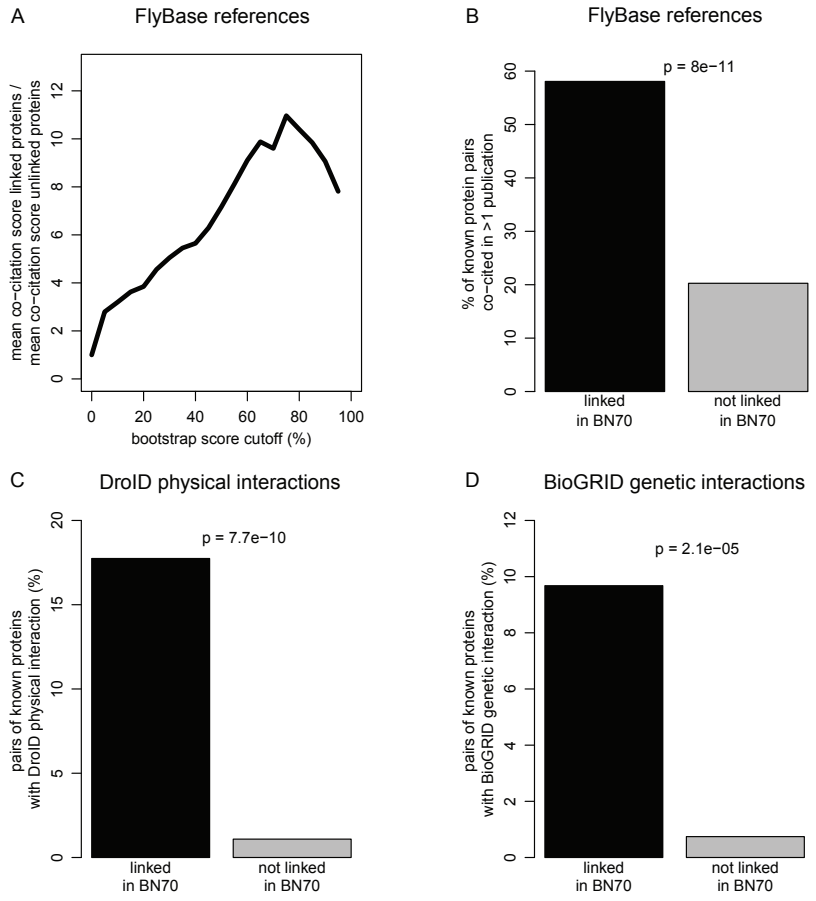


Figure S2 Agreement of the Bayesian network model with published data. (A) Enrichment of literature co-citations reported in FlyBase among 65 known chromatin proteins that are linked in the Bayesian Network, for various bootstrap score cutoff levels. The co-citation score of a pair of known proteins is defined as the number of publications that report on both proteins, normalized by the product of the total numbers of publications for each of the two proteins. This normalization adjusts for the fact that frequently studied proteins are more likely to be co-cited than rarely studied proteins. The vertical axis indicates the enrichment of the co-citation scores among protein pairs that are linked in the Bayesian Network, relative to co-citation scores among unlinked proteins. (B-C) Proportion of known protein pairs linked (black) or not linked (grey) in BN70 for (B) protein pairs co-cited at least twice in the FlyBase literature list; (c) protein pairs with physical interactions according to the DroID database, and (D) protein pairs with genetic interactions of the corresponding genes according to BioGRID. P-values according to Fisher's exact test.

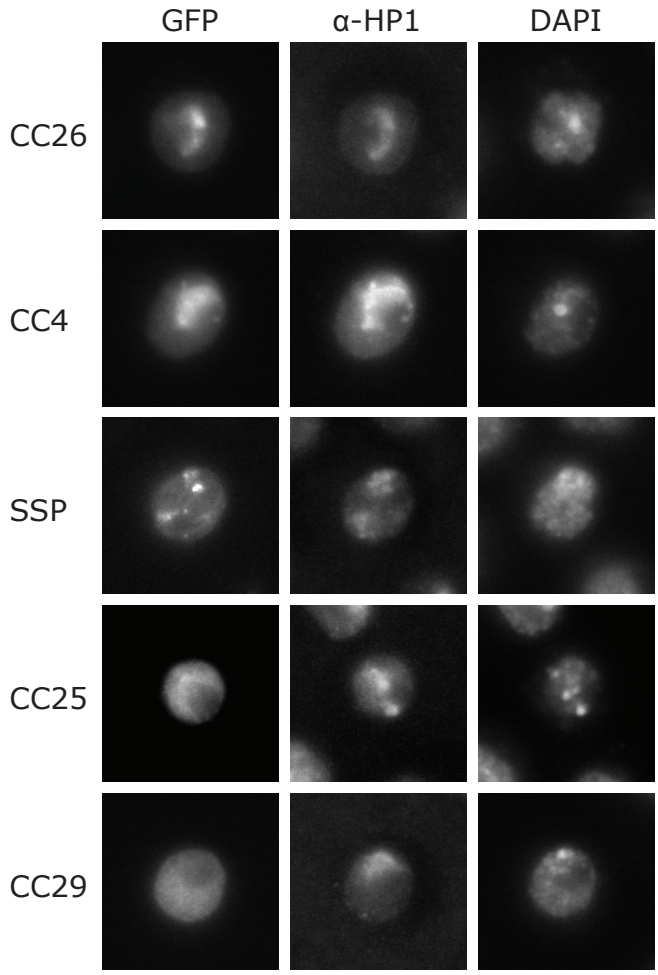


Figure S3 Intra-nuclear localization of novel GREEN components. Microscopy images of Kc cells transfected with GFP-tagged proteins and stained with anti-HP1 antibody and DAPI.

Table S1 FlyBase accession numbers, gene IDs, selection criteria, and DamID result for all tested candidate proteins.

FlyBase gene	Name	FlyBase Gene Accession	DamID result	Selection criterium
CG2621	CG2621	FBgn0003371	cloned	I
CG4612	CG4612	FBgn0035016	cloned	R
CG2183	CG2183	FBgn0033273	cloned	R
CG11791	CG11791	FBgn0039266	cloned	I
CG8187	CG8187	FBgn0034027	cloned	I
CG11982	CG11982	FBgn0037653	cloned	D
CG4424	CG4424	FBgn0038765	cloned	D
CG3919	CG3919	FBgn0036423	cloned	D
CG6570	CG6570	FBgn0008651	cloned	I
CG5920	sop	FBgn0004867	cloned	R
CG12129	CG12129	FBgn0033475	cloned	R
CG10754	CG10754	FBgn0036314	cloned	R
CG5190	CG5190	FBgn0034351	cloned	R
CG5315	CG5315	FBgn0038984	cloned	I
CG3613	qkr58E-1	FBgn0022986	cloned	R
CG14718	CG14718	FBgn0037939	cloned	R
CG8314	CG8314	FBgn0034057	cloned	D
CG1407	CG1407	FBgn0033474	cloned	D
CG15437	morgue	FBgn0027609	cloned	I
CG17838	CG17838	FBgn0038826	cloned	R
CG10327	TBPH	FBgn0025790	cloned	R
CG17257	CG17257	FBgn0031495	cloned	D
CG8597	lark	FBgn0011640	cloned	R
CG6049	CG6049	FBgn0037081	cloned	R
CG17198	CG17198	FBgn0039366	cloned	D
CG6627	Dnz1	FBgn0027453	cloned	D
CG13628	Rpb10	FBgn0039218	cloned	D
CG4510	Surf6	FBgn0038746	cloned	I
CG5144	CG5144	FBgn0035957	cloned	I
CG6654	CG6654	FBgn0038301	cloned	I
CG10157	CG10157	FBgn0039099	cloned	I
CG11722	CG11722	FBgn0037777	cloned	I
CG7109	mts	FBgn0004177	cloned	I
CG4609	CG4609	FBgn0014163	mePCR	I

Table S1 Continued.

FlyBase gene	Name	FlyBase Gene Accession	DamID result	Selection criterium
CG6961	CG6961	FBgn0030959	mePCR	R
CG6712	CG6712	FBgn0032408	mePCR	R
CG14816	CG14816	FBgn0023517	mePCR	I
CG16788	RnpS1	FBgn0037707	mePCR	R
CG6693	CG6693	FBgn0037878	mePCR	I
CG7154	CG7154	FBgn0031947	mePCR	D
CG6474	e(y)1	FBgn0000617	mePCR	D
CG13123	CG13123	FBgn0032150	mePCR	D
CG4258	dbe	FBgn0020305	mePCR	R
CG1681	CG1681	FBgn0030484	mePCR	R
CG1681	CG1681	FBgn0030484	mePCR	I
CG9705	CG9705	FBgn0036661	mePCR	D
CG6272	CG6272	FBgn0036126	mePCR	D
CG8385	Arf79F	FBgn0010348	mePCR	I
CG31324	CG31324	FBgn0051324	mePCR	I
CG18543	mtrm	FBgn0010431	mePCR	R
CG6842	CG6842	FBgn0027605	mePCR	I
CG6664	CG6664	FBgn0036685	mePCR	I
CG18319	ben	FBgn0000173	mePCR	I
CG5641	CG5641	FBgn0038046	mePCR	D
CG14283	mRpL55	FBgn0038678	mePCR	D
CG5927	Her	FBgn0030899	mePCR	D
CG4217	TFAM	FBgn0038805	mePCR	D
CG9205	CG9205	FBgn0035181	mePCR	I
CG10147	CG10147	FBgn0035702	mePCR	D
CG3576	Lag1	FBgn0040918	mePCR	D
CG12175	tth	FBgn0030502	mePCR	D
CG10318	NC2alpha	FBgn0034650	mePCR	D
CG11596	CG11596	FBgn0023522	mePCR	I
CG12235	Arp11	FBgn0031050	mePCR	I
CG3350	bigmax	FBgn0039509	mePCR	D
CG5684	Pop2	FBgn0036239	mePCR	D
CG7911	CG7911	FBgn0039735	mePCR	I
CG18591	CG18591	FBgn0031962	mePCR	I
CG5623	CG5623	FBgn0038357	mePCR	I

Table S1 Continued.

FlyBase gene	Name	FlyBase Gene Accession	DamID result	Selection criterium
CG11555	CG11555	FBgn0086856	mePCR	I
CG8882	Trip1	FBgn0015834	profile	I
CG3771	CG3771	FBgn0023130	profile	I
CG1847	CG1847	FBgn0030345	profile	I
CG5168	CG5168	FBgn0032246	profile	I
CG4886	CG4886	FBgn0028382	profile	R
CG11738	l(1)G0004	FBgn0027334	profile	R
CG5846	CG5846	FBgn0032171	profile	D
CG2129	CG2129	FBgn0030008	profile	D
CG3313	CG3313	FBgn0037980	profile	D
CG7928	CG7928	FBgn0039740	profile	D
CG4617	CG4617	FBgn0029936	profile	D
CG7745	CG7745	FBgn0033616	profile	D
CG15514	CG15514	FBgn0039712	profile	D
CG17385	CG17385	FBgn0033934	profile	D
CG11723	CG11723	FBgn0031391	profile	D
CG5181	CG5181	FBgn0031909	profile	D
CG3909	CG3909	FBgn0027524	profile	D
CG7006	CG7006	FBgn0039233	profile	R
CG16817	CG16817	FBgn0037728	profile	I
CG7818	CG7818	FBgn0032016	profile	D
CG14962	CG14962	FBgn0035407	profile	D
CG9797	CG9797	FBgn0037621	profile	D
CG7946	CG7946	FBgn0039743	profile	I
CG3838	CG3838	FBgn0032130	profile	D
CG4936	CG4936	FBgn0038768	profile	D
CG10949	CG10949	FBgn0032858	profile	D
CG10267	CG10267	FBgn0037446	profile	D
CG5245	CG5245	FBgn0038047	profile	D
CG11802	CG11802	FBgn0030346	profile	I
CG17153	CG17153	FBgn0036248	profile	D
CG12744	CG12744	FBgn0033459	profile	D
CG7357	CG7357	FBgn0038551	profile	D
CG8289	CG8289	FBgn0030854	profile	D
CG8924	CG8924	FBgn0030710	profile	D

Table S1 Continued.

FlyBase gene	Name	FlyBase Gene Accession	DamID result	Selection criterium
CG7066	Sbp2	FBgn0087039	profile	I
CG2257	Ubc-E2H	FBgn0029996	profile	I
CG33213	CG33213	FBgn0053213	profile	D
CG17383	jigr1	FBgn0039350	profile	D
CG9954	maf-S	FBgn0034534	profile	D
CG15436	CG15436	FBgn0031610	profile	D
CG2540	CG2540	FBgn0030411	profile	I
CG5274	CG5274	FBgn0036987	profile	I

cloned	no detectable adenine methylation of the genome
mePCR	detectable adenine methylation of the genome, but no specific and reproducible binding profile
profile	specific and reproducible binding profile
D	contains a protein **D**omain, which is often found in chromatin proteins
I	**I**nteracts with a known chromatin protein in Y2H screen
R	**R**NA binding domain

Table S2 FlyBase accession numbers, gene IDs, names and DamID plasmid backbone of the 42 novel as well as the 10 additional known chromatin proteins.

Name	FlyBase Gene Accession	FlyBase gene	Transcript	cDNA clone ID	Plasmid Backbone
CC1	FBgn0037980	CG3313	CG3313-RA	LD21841	pCreDamMyc
CC2	FBgn0027334	CG11738	l(1)G0004-RA	LD21667	pCreDamMyc
CC3	FBgn0038047	CG5245	CG5245-RA	IP01468	pCreDamMyc
CC4	FBgn0030854	CG8289	CG8289-RA	LD36501	pCreDamMyc
CC5	FBgn0039233	CG7006	CG7006-RA	GM12126	pCreDamMyc
CC6	FBgn0031610	CG15436	CG15436-RA	LD32747	pCreDamMyc
CC7	FBgn0033934	CG17385	CG17385-RA	IP01247	pCreDamMyc
CC8	FBgn0033616	CG7745	CG7745-RA	LD32860	pCreDamMyc
CC9	FBgn0030710	CG8924	CG8924-RB	LD19131	pCreDamMyc
CC10	FBgn0027524	CG3909	CG3909-RA	LD21537	pCreDamMyc
CC11	FBgn0032016	CG7818	CG7818-RA	LD06016	pCreDamMyc
CC12	FBgn0032171	CG5846	CG5846-RA	LP07441	pCreDamMyc
CC13	FBgn0036987	CG5274	CG5274-RA	SD10847	pCreDamMyc
CC14	FBgn0030345	CG1847	CG1847-RA	HL02936	pCreDamMyc
CC15	FBgn0035407	CG14962	CG14962-RA	AT25633	pCreDamMyc
CC16	FBgn0030346	CG11802	CG11802-RA	LD30509	pCreDamMyc
CC17	FBgn0032246	CG5168	CG5168-RA	LD41958	pCreDamMyc
CC18	FBgn0037728	CG16817	CG16817-RA	LD23532	pCreDamMyc
CC19	FBgn0030411	CG2540	CG2540-RA	GH24869	pCreDamMyc
CC20	FBgn0032130	CG3838	CG3838-RB	LD04047	pCreDamMyc
CC21	FBgn0033459	CG12744	CG12744-RA	GH03826	pCreDamMyc
CC22	FBgn0030008	CG2129	CG2129-RA	LD35215	pCreDamMyc
CC23	FBgn0023130	CG3771	CG3771-RA	LD13641	pCreDamMyc
CC24	FBgn0031909	CG5181	CG5181-RA	LD31278	pCreDamMyc
CC25	FBgn0031391	CG11723	CG11723-RA	LD29420	pCreDamMyc
CC26	FBgn0038551	CG7357	CG7357-RA	LD33778 [a]	pCreDamMyc
CC27	FBgn0039740	CG7928	CG7928-RA	LD15405	pCreDamMyc
CC28	FBgn0029936	CG4617	CG4617-RA	LD46272	pCreDamMyc
CC29	FBgn0053213	CG33213	CG33213-RA	LD18667	pCreDamMyc
CC30	FBgn0039712	CG15514	CG15514-RA	LD24318	pCreDamMyc
CC31	FBgn0037621	CG9797	CG9797-RA	LD30467	pCreDamMyc
CC32	FBgn0037446	CG10267	CG10267-RA	LD25151	pCreDamMyc
CC33	FBgn0038768	CG4936	CG4936-RA	LD08906	pCreDamMyc
CC34	FBgn0032858	CG10949	CG10949-RA	GH22016	pCreDamMyc

Table S2 Continued.

Name	FlyBase Gene Accession	FlyBase gene	Transcript	cDNA clone ID	Plasmid Backbone
CC35	FBgn0039743	CG7946	CG7946-RA	LD23804	pCreDamMyc
CYP33	FBgn0028382	CG4886	CG4886-RA	LD35248	pCreDamMyc
JIGR2	FBgn0039350	CG17383	CG17383-RA	LD33329	pCreDamMyc
MAF-S	FBgn0034534	CG9954	CG9954-RA	GH02096	pCreDamMyc
SBP2	FBgn0087039	CG7066	Sbp2-RA	GH01354	pCreDamMyc
SSP	FBgn0036248	CG17153	CG17153-RA	LD31163	pCreDamMyc
TRIP1	FBgn0015834	CG8882	Trip1-RA	LD24026	pCreDamMyc
UBC-E2H	FBgn0029996	CG2257	Ubc-E2H-RA	LD13772	pCreDamMyc

Name	FlyBase Gene Accession	FlyBase gene	Transcript	cDNA clone ID	Plasmid Backbone
ADA2B	FBgn0037555	CG9638	CG9638-RA	LD24257	pCreDamMyc
DMAP1	FBgn0034537	CG11132	DMAP1-RA	LD35228	pCreDamMyc
ERR	FBgn0035849	CG7404	CG7404-RA	GH28308	pCreDamMyc
HEL25E	FBgn0014189	CG7269	Hel25E-RA	LD23644	pCreDamMyc
ISWI	FBgn0011604	CG8625	RA, RB and RC only differ in untranslated region	RH13158	pDamMyc
MES4	FBgn0034726	CG11301	CG11301-RA	IP14609	pCreDamMyc
RAD21	FBgn0026057	CG17436	vtd-RA	published [b]	pDamMyc
RPS9	FBgn0010408	CG3395	RpS9-RB	LD32106	pCreDamMyc
RYBP	FBgn0034763	CG12190	CG12190-RA	LD18758	pCreDamMyc
SSL1	FBgn0037202	CG11115	CG11115-RA	GH08526	pCreDamMyc

[a] contains a silent mutation in the ORF
[b] plasmid is published before (Pauli et al 2010)

6

Discussion

Joke G. van Bemmel, Bas van Steensel

In the previous Chapters I have presented 4 different studies that examine chromatin composition and its role in gene regulation. Each of these studies focused on a different aspect of chromatin organization, as introduced in the first Chapter of this thesis. Together these studies provided new insights in the composition as well as the function of chromatin in *Drosophila melanogaster*. In this Chapter I will take these new insights together and put them in the perspective of the 5 defined chromatin types.

Defining chromatin types

We have found the Drosophila genome to be organized in 5 distinct chromatin types. However this does not imply that there are only five combinations of proteins possible along the genome. As already pointed out in Chapter 4, each type could be further divided in subtypes. For example, within RED and YELLOW chromatin, promoters have a distinctly different protein composition than gene bodies. The differences between the 5 chromatin types are however much more distinct than the differences within each type. It should be noted that large amounts of high quality high-resolution data sets will most likely always yield an extremely large number of possible proteins combinations (chromatin types) that are significantly different form each other. However, this does not necessarily imply that all of these chromatin types are biologically different and relevant. Several other groups also employed large-scale genome-wide profiling of chromatin components to gain insight into chromatin organization [1-7, reviewed in 8]. Indeed different studies in different species defined different numbers of chromatin types, ranging from 4 up to even 51 different types. Notably, the studies finding large numbers of chromatin types in addition applied other algorithms or added a clustering in order to group their different states into 4 to 9 main chromatin types.

In addition, one can not completely exclude that the number of defined states could depend on the choice of mapped chromatin components. At the time that we defined the 5 principal chromatin types, the existence of an additional chromatin type could not be ruled out completely, since we could not exclude that marker proteins for such an additional type were missing in our dataset. Interestingly, no additional types could be identified after adding 42 novel so far unknown chromatin components (Chapter 5, Figure 4A). This indicates that in Kc cells additional chromatin types are very unlikely to exist. However this does not exclude that different combinations of chromatin proteins, and thus different chromatin types, could occur in other cell types or tissues.

Chromatin types compared

While we have identified chromatin types based on DamID profiles of 53 chromatin proteins in Kc cells, the modENCODE consortium has published chromatin types based on ChIP profiles of 18 histone modifications in a very similar cell type, S2 cells [5]. Although based on completely different sources the chromatin types show a striking overlap, as well as some essential differences (Figure 1).

We showed the majority of the inactive genes to have a BLACK chromatin type, which is the most predominant chromatin type covering almost half of the genome. BLACK chromatin is almost completely covered by the "Transcription-

ally silent" state from the modENCODE project (Figure 1, top left panel). Like BLACK chromatin, this state also covers the largest part of the genome (Figure 1, bottom right panel) emphasizing the similarity between the two.

We have found two types of active genes; RED genes which are tightly regulated and YELLOW genes which are broadly expressed and enriched for H3K36me3 (Chapter 4, Figure 6). Our RED chromatin largely overlaps mainly with the "Regulatory Regions (Enhancers)" state from the modENCODE consortium (Figure 1, top right panel). The regulatory role of this state is indicated by the enrichment of enhancer binding proteins as well as known enhancer sequences, which fits with our finding that RED genes are active and tightly regulated. YELLOW chromatin on the other hand mainly overlaps with the "Transcription Elongation" and "Promoters and Transcription Start Sites (TSSs)" states (Figure 1, mid left panel). The "Transcription Elongation" state is defined as such because of the enrichment of the elongation marker H3K36me3, which is also the histone modification that distinguishes this state. The "Promoters and Transcription Start Sites (TSSs)" state is classified as such because it is found at active promoters and TSSs. This is in agreement with our description of YELLOW genes as being active and enriched for the elongation marker H3K36me3. In concordance with our distinction between RED and YELLOW genes the modENCODE consortium found that the presence or absence of their "Regulatory Regions (Enhancers)" state is the most common difference in the chromatin composition of expressed genes (that are 1 kb or longer).

BLUE chromatin however, does show less similarity with the modENCODE states. Besides overlapping with the "Polycomb-mediated repression" state it also overlaps largely with the "Transcriptionally silent" and "Active Introns" states (Figure 1, mid right panel). This partial overlap could be explained by the different approaches used to generate the data on which the chromatin types are based. Our approach differs from the modENCODE approach in the technique used (DamID vs ChIP), the cell type (Kc cells vs S2 cells) and in the mapped chromatin components (chromatin proteins versus histone marks). In contrast to RED, YELLOW or BLACK genes, Polycomb repressed genes could for example be differently regulated between Kc and S2 cells. Alternatively, Polycomb binding might be not strictly correlated to the presence of the histone modification H3K27me3. In addition it should be noted that the "Active Introns" state, which overlaps with ~30% of BLUE, is differently annotated in the different modENCODE publications [4,5], indicating that this state might be less robust or less well defined.

GREEN chromatin overlaps with two of the modENCODE states. One part of GREEN chromatin overlaps with the "Pericentromeric heterochromatin" state, which is the classical description of HP1 chromatin. Interestingly the other part of GREEN chromatin overlaps with "Transcription elongation". As discussed in the introduction, HP1 is known to not only bind at pericentromeric heterochromatin but also at the gene body of actively transcribed genes [reviewed in 9, 10-11]. In our chromatin type definition both kinds of HP1 binding are defined as GREEN chromatin while the two modENCODE

states nicely distinguish this duality in HP1 bound chromatin and its correlation with transcriptional outcome.

Overall, the modENCODE consortium aimed to provide a detailed annotation of all genomic sequences, while we rather aimed to describe the chromatin organization. The differences in chromatin type definitions are clearly reflecting these different aims.

Lamina associated chromatin

In Chapter 2 we found 40% of the genome to be in molecular contact with the nuclear lamina (NL). This NL association appeared to be one of the core characteristics of BLACK chromatin. About 75% of the BLACK chromatin is in contact with the NL (Chapter 4, Figure 3) and in the network model Lamin clearly clusters together with other BLACK chromatin components (Chapter 5, Figure 4). Interestingly, this finding provides new insights into the molecular composition of the chromatin that resides at the NL. Other core components of BLACK chromatin are the known proteins SU(HW), EFF, SUUR, D1, H1 and IAL. In Chapter 2 we already found one of these components, SU(HW), to weaken genome-NL interactions through a local antagonistic effect. Although being the first protein identified to modulate genome-NL interactions in *Drosophila*, it became clear that SU(HW) is not the driving force behind LAD formation. Assuming that LAD formation is not a random process, other BLACK chromatin components are thus likely to be involved in LAD formation.

In the network from Chapter 5 Lamin is closely connected to H1, SUUR and D1. Interestingly, preliminary experiments showed elevated levels of H1 but not SUUR and D1 to specifically increase genome - NL association (unpublished work from R Lim and JG van Bemmel). In addition, the network reveals 4 novel BLACK components to be indirectly linked to Lamin and thus to be potential candidates to influence genome – NL association.

Our identification of proteins modulating genome - NL interactions confirms the existence of a regulating mechanism, excluding the possibility that LADs are merely the consequence of active chromatin being at the interior and the inactive chromatin thus being passively pushed towards the periphery.

Insulator proteins in the chromatin types

Insulator proteins are believed to play an essential role in genome organization by facilitating or preventing enhancer-promoter interactions and by preventing spreading of histone modifications and or chromatin proteins to adjacent sequences to separate different chromatin types [reviewed in 12]. In Chapter 2 we generated genome-wide binding profiles for all known *Drosophila* insulator proteins, which have been integrated in the chromatin type definition from Chapter 4 and the network analysis from Chapter 5. Except for SU(HW) the insulator proteins almost exclusively bind in the active chromatin types (RED and YELLOW), suggesting insulator proteins to be essential for gene activity. This is emphasized by the fact that in Chapter 5 we predicted BEAF and DWG to have an activating effect on gene expression.

Interestingly their abundance is higher in RED than in YELLOW chromatin. RED chromatin has a higher density of regulatory elements than YELLOW chromatin

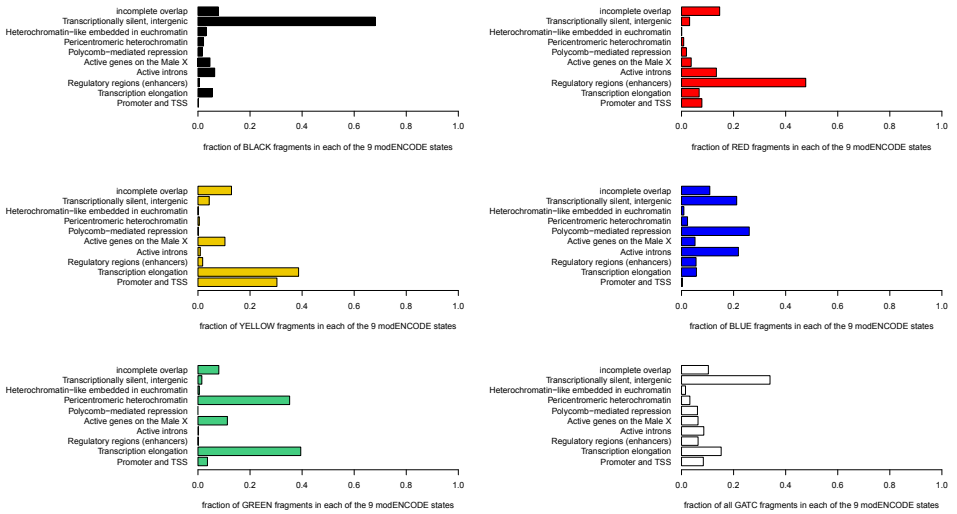


Figure 1 Comparison with the chromatin types defined by the modENCODE consortium. Each plot depicts which fraction (x-axis) of our chromatin type (indicated by the color) overlaps with each of the chromatin states defined by the modENCODE consortium (bars). This fraction is defined based on complete overlap of GATC fragments. Fragments which are only partly overlapping are not included in one of the 9 modENCODE states, but depicted separately in the top bar. The bottom right plot (white) shows the fraction of all GATC fragments within each modENCODE state.

(Chapter 4, Figure 4 and Figure 6). It is therefore tempting to speculate that insulator proteins preferentially bind in RED chromatin in order to regulate the contacts between regulatory elements -such as enhancers- and promoters. BEAF32 and DWG could for example be involved in facilitating the enhancer-promoter interactions since they are found to be enriched exactly at the transcription start sites of active genes (data not shown). CTCF on the other hand is not enriched at transcription start sites (data not shown), suggesting that it might rather be involved in the separation of enhancers from neighboring genes. In contrast, SU(HW) is the only insulator protein that preferentially resides in the inactive BLACK chromatin. Although BLACK chromatin is transcriptionally silent it is extremely rich in putative regulatory elements (Chapter 4, Figure 4D). SU(HW) could thus, besides modulating genome - NL lamina interactions, be involved in separating these regulatory elements from the promoters to ensure silencing. Since BLACK genes are known to become active in specific tissues (Chapter 4, Figure 4C), it would therefore be interesting to analyze how the binding of SU(HW) changes upon differentiation.

These speculations however assume that these proteins purely act as insulating proteins. One should keep in mind that this characteristic function is mainly established under non-endogenous conditions, like reporter essays and in transgenic flies. Alternatively, insulator proteins could therefore very likely rather act directly at their genomic binding site, independent from any insulating action,

similar to traditional sequence specific DNA binding factors. Binding of BEAF32 and DWG at active promoters could for example simply be essential to aid recruitment of the transcription machinery.

As pointed out above, insulator proteins are believed to not only separate or facilitate enhancer - promoter interactions but also to separate specific chromatin types by preventing spreading. This notion makes it tempting to speculate about a role for insulator proteins in the border formation of the different chromatin types. Although, insulator proteins clearly bind inside the chromatin domains and not exclusively at the borders (as discussed above and visible in Chapter 4, Figure 1) we did find a preference for some of the insulator proteins to bind at the border of specific domains (data not shown). CTCF, for example, is known to prevent spreading of H3K27 methylation [13], and we indeed found CTCF to be enriched at the borders of BLUE chromatin domains. SU(HW), which we found to be enriched at the borders of LADs (Chapter 2), also appeared to be enriched, but only slightly, at borders of BLUE chromatin domains. Although, these and some other specific enrichments could be found it became apparent that not all chromatin types are marked by an insulator protein and not all insulator proteins have a preference to bind at the border of specific domains.

RAD21 as a chromatin component

Cohesin, of which the canonical function is to mediate sister chromatid cohesion, is suggested to be involved in gene regulation. In Chapter 3 and Chapter 5 we indeed found RAD21, a subunit of cohesion, to bind specific genomic regions in inter-phase cells. Interestingly, RAD21 appeared to be one of the few proteins that preferentially bind more in YELLOW than in RED chromatin. Similarly to other YELLOW specific proteins it also binds in GREEN chromatin. Together, this indicates RAD21 to be a novel core component of YELLOW chromatin.

In the network, generated in Chapter 5, RAD21 is connected to H3K36me3, which is believed to mark transcription elongation [14]. Additionally, we found RAD21 to bind in large domains containing multiple genes (Chapter 5, Figure S1 and Chapter 3, Figure 1). These two observations together with the ring shaped structure of cohesion, suggest a role as a sort of loading platform for other factors that regulate RNA polymerase II movement during elongation. Already in 2004, cohesin in budding yeast has been suggested to be pushed towards the end of transcription units by RNA polymerase II [15]. Knowing now that cohesin directly regulates its target genes (Chapter 3) it is more likely that cohesion either moves along with RNA polymerase II instead of being pushed by it or binds along transcription units thereby aiding recruitment of the transcription machinery.

In mammals however, cohesin has been suggested to regulate gene expression by mediating interactions between CTCF sites [16-18]. Visual inspection of the RAD21 binding profile from Chapter 5 with the insulator binding profiles from Chapter 2 shows RAD21 not to co-localize with CTCF but with BEAF32, although only partly (data not shown). A gene regulatory role for cohesion by mediating insulator actions could thus potentially be conserved via BEAF32. However, we found RAD21 in fly cells

to bind in large domains containing multiple genes (Chapter 5, Figure S1 and Chapter 3, Figure 1), while insulator proteins and mammalian cohesion exhibit a distinct, more focal binding pattern (Chapter 2, Figure 2 and [16]). Although a conserved function via BEAF32 would fit with the notion that during evolution mammalian CTCF has possibly acquired all the functions of the different subclasses of *Drosophila* insulator proteins [19], it is thus more likely that cohesion regulates gene expression differently in mammals and flies.

Repair and replication proteins in a chromatin context

As pointed out in Chapter 1, proteins involved in the key processes of the cell, like transcription, repair and replication, are influencing chromatin composition and/or are being part of chromatin themselves. We found RNA polymerase II, the core component of the transcription machinery, as expected to be enriched at active genes. For proteins involved in repair and replication it is however less obvious to predict their potential binding profile. Since repair as well as replication proteins both act on the entire genome we anticipated their binding profiles to be uniform. Interestingly, in Chapter 4 and 5 we did find specific binding sites for proteins involved in respectively replication and nucleotide excision repair (NER). This could first of all indicate that these proteins are specific components of chromatin in which they might carry out a chromatin-related function different from their canonical function. Second, repair or replication complexes might assemble at specific genomic locations where they stay until they are required and relocated to sites of action. In case of the replication proteins these sites could potentially reflect replication origins. Comparing the binding profiles of the replication proteins (Chapter 5) with origin recognition complex (ORC) binding sites [20] does reveal a clear overlap. This however only explains the enrichment inside RED chromatin but not the replication protein specific binding, given that other RED chromatin components such as ASF1 and ISWI are also or maybe even more enriched at ORC binding sites (data not shown).

More likely, the specific binding profile of replication proteins reflects stalling of the replication machinery at for example damaged sites while the binding profile of NER proteins reflects preferential repair at specific sites. Certain sites in the genome could for example be more prone to UV damage than others. Since UV irradiation causes pyrimidine dimers, regions with stretches of neighboring pyrimidines could potentially be more susceptible to this type of damage. Alternatively, specific damaged sites could be recognized more efficiently than others depending on their chromatin composition. Accordingly, transcription activators, ATP-driven remodeling complexes and histone acetyltransferases are found to cause a local and transcription-independent increase in nucleotide excision repair [21]. Such proteins are enriched in RED chromatin, thereby providing an alternative or additional explanation for the enrichment of repair proteins in RED chromatin. Alternatively, the NER proteins might recognize distortions in the DNA helix caused by other factors than UV damage [22], such as sequence bias.

Chromatin types during differentiation

We have identified the 5 chromatin types in cultured Drosophila cells and they are very likely to be different in different cell types and tissues. For example, BLACK genes are known to become active in specific tissues, so their chromatin composition will have to turn into YELLOW, RED or GREEN in these tissues. It is thus clear that chromatin types will change into each other, but will all of them change? YELLOW chromatin is for example unlikely to change very much between different tissues. It contains mainly housekeeping genes which are uniformly expressed over different tissues and they will therefore keep a YELLOW chromatin composition in different tissues. Since RAD21 preferentially binds YELLOW chromatin we can use its binding data to obtain a preliminary idea about YELLOW chromatin during differentiation. Comparing the RAD21 binding data in Salivary glands (Chapter 3) with the RAD21 binding data in Kc cells (Chapter 4) indeed shows a great degree of similarity (data not shown). One would expect the other chromatin types to be more dynamic, but it is not clear which type will be able to change into which other type.

Based on their protein composition (Chapter 4, Figure 3) and their clustering in the network (Chapter 5, Figure 4) some chromatin types seem to be more comparable than others. BLACK for example is defined by a combination of proteins that are as well binding in BLUE chromatin and GREEN chromatin is in addition to the classical heterochromatin components also bound by the YELLOW specific components. Conversely, BLUE and YELLOW share almost no components at all. This raises the question of chromatin types with overlapping protein compositions will change into each other more frequently than types with completely distinct protein compositions. Analyzing the chromatin types in fly tissues will be instrumental to answer these questions. As mentioned in Chapter 4 the classification into 5 types can be largely reconstituted with a subset of only 5 marker proteins, which increases the feasibility of a tissue specific study of chromatin types enormously. However, the possible existence of additional chromatin types in different tissues could be a potential caveat of such an approach.

Spatial organization of the Chromatin types

As pointed out above, we found large parts of BLACK (~75%) and also BLUE (~60%) chromatin to be in molecular contact with the nuclear lamina. BLUE and BLACK chromatin thus preferentially localize at the periphery, but what about the other chromatin types? RED chromatin is extremely protein dense; specifically the occupancy of transcription factors is very high. This high protein occupancy could reflect interactions between distant loci/genes allowing proteins to be in contact with multiple genomic regions at the same time. In other words, this high protein occupancy suggests that genes in RED chromatin might come together in transcription factories which contain high concentrations of RNA polymerase II as well as transcription factors [reviewed in 23]. Knowing this, one could also speculate that the above discussed enrichment of BEAF32 and DWG at transcription start sites facilitates these interactions. BLUE chromatin is known to come together in nuclear space, since it has

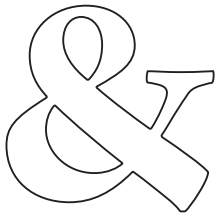
already been shown that Polycomb (which is the core component of BLUE chromatin) domains interact with each other [24-25]. GREEN chromatin domains are also likely to interact with other GREEN domains since heterochromatic regions are known to aggregate together in the chromocenters [26] and ectopic HP1 is found to promote interactions with other HP1 bound regions [27].

An elegant study from the Dekker lab reporting the three-dimensional architecture of the whole human genome revealed the human genome to be spatially segregated in two compartments [28]. They found active regions to preferentially interact with other active regions while inactive regions mainly interacted with other inactive regions. It would thus be interesting to see how the 5 different chromatin types behave in respect to chromosomal interactions. Since we know the Drosophila genome to be organized in 5 chromatin types, a spatial organization into 5 distinct compartments in which specific chromatin types preferentially interact with each other would be very likely.

REFERENCES

1. Roudier, F., et al., Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. EMBO J, 2011. 30(10): p. 1928-38.
2. Liu, T., et al., Broad chromosomal domains of histone modification patterns in C. elegans. Genome Res, 2011. 21(2): p. 227-36.
3. Gerstein, M.B., et al., Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science, 2010. 330(6012): p. 1775-87.
4. Roy, S., et al., Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science, 2010. 330(6012): p. 1787-97.
5. Kharchenko, P.V., et al., Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. Nature, 2011. 471(7339): p. 480-5.
6. Ernst, J. and M. Kellis, Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol, 2010. 28(8): p. 817-25.
7. Ernst, J., et al., Mapping and analysis of chromatin state dynamics in nine human cell types. Nature, 2011. 473(7345): p. 43-9.
8. Baker, M., Making sense of chromatin states. Nat Methods, 2011. 8(9): p. 717-22.
9. Dimitri, P., et al., The paradox of functional heterochromatin. Bioessays, 2005. 27(1): p. 29-41.
10. Hediger, F. and S.M. Gasser, Heterochromatin protein 1: don't judge the book by its cover! Curr Opin Genet Dev, 2006. 16(2): p. 143-50.
11. Fanti, L. and S. Pimpinelli, HP1: a functionally multifaceted protein. Curr Opin Genet Dev, 2008. 18(2): p. 169-74.
12. Valenzuela, L. and R.T. Kamakaka, Chromatin insulators. Annu Rev Genet, 2006. 40: p. 107-38.
13. Bartkuhn, M., et al., Active promoters and insulators are marked by the centrosomal protein 190. EMBO J, 2009. 28(7): p. 877-88.
14. Lee, J.S. and A. Shilatifard, A site to remember: H3K36 methylation a mark for histone deacetylation. Mutat Res, 2007. 618(1-2): p. 130-4.
15. Lengronne, A., et al., Cohesin relocation from sites of chromosomal loading to places of convergent transcription. Nature, 2004. 430(6999): p. 573-8.
16. Wendt, K.S., et al., Cohesin mediates transcriptional insulation by CCCTC-binding factor. Nature, 2008. 451(7180): p. 796-801.
17. Rubio, E.D., et al., CTCF physically links cohesin to chromatin. Proc Natl Acad Sci U S A, 2008. 105(24): p. 8309-14.
18. Parelho, V., et al., Cohesins functionally associate with CTCF on mammalian chromosome arms. Cell, 2008. 132(3): p. 422-33.

19. Bushey, A.M., E. Ramos, and V.G. Corces, Three subclasses of a Drosophila insulator show distinct and cell type-specific genomic distributions. Genes Dev, 2009. 23(11): p. 1338-50.
20. MacAlpine, H.K., et al., Drosophila ORC localizes to open chromatin and marks sites of cohesin complex loading. Genome Res, 2010. 20(2): p. 201-11.
21. Frit, P., et al., Transcriptional activators stimulate DNA repair. Mol Cell, 2002. 10(6): p. 1391-401.
22. Missura, M., et al., Double-check probing of DNA bending and unwinding by XPA-RPA: an architectural function in DNA repair. EMBO J, 2001. 20(13): p. 3554-64.
23. Razin, S.V., et al., Transcription factories in the context of the nuclear and genome organization. Nucleic Acids Res, 2011. 39(21): p. 9085-92.
24. Tolhuis, B., et al., Interactions among Polycomb domains are guided by chromosome architecture. PLoS Genet, 2011. 7(3): p. e1001343.
25. Bantignies, F., et al., Polycomb-dependent regulatory contacts between distant Hox loci in Drosophila. Cell, 2011. 144(2): p. 214-26.
26. Heitz, E., Heterochromatin, Chromocentren, Chromomeren (Vorläufige Mitteilung). Ber. Dtsch. Bot. Ges. , 1929. 47: p. 274-284.
27. Seum, C., et al., Ectopic HP1 promotes chromosome loops and variegated silencing in Drosophila. EMBO J, 2001. 20(4): p. 812-8.
28. Lieberman-Aiden, E., et al., Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science, 2009. 326(5950): p. 289-93.



| SAMENVATTING

Chromatine is het complex van DNA, RNA en alle bijbehorende eiwitten. De exacte eiwit samenstelling van het chromatine beinvloedt genexpressie, DNA replicatie en andere functies. In dit proefschrift hebben we gebruik gemaakt van DamID om de chromatine samenstelling en organisatie in *Drosophila melanogaster* te bestuderen.

In Hoofdstuk 2 hebben we ons gericht op de ruimtelijke organisatie van het genoom in de kern, door genoom - nucleaire lamina (NL) interacties te bestuderen. Van het menselijk genoom was al bekend dat het georganiseerd is in honderden grote domeinen die met de lamina in contact zijn. Deze domeinen noemen we Lamina-Associated-Domains (LADs). Het was echter niet duidelijk of deze organisatie vergelijkbaar is in Drosophila, en er waren tot dusver geen eiwitten bekend die de interacties reguleren. Wij vonden dat vliegen LADs, net als de menselijke LADs, een sterk repressieve chromatine-omgeving vertegenwoordigen en dat ze ongeveer hetzelfde aantal genen bevatten. Een vergelijkende analyse tussen de geïdentificeerde LADs en bindingsprofielen van alle vijf bekende isolator eiwitten liet zien dat SU(HW) het enige isolator eiwit is dat bij voorkeur in en op de randen van LADs bindt. Door depletie en over-expressie van SU(HW) lieten we zien dat SU(HW) lokaal de genoom - NL interacties verzwakt. Deze resultaten geven inzicht in de evolutie van de ruimtelijke organisatie van het genoom, en identificeren SU(HW) als het eerste eiwit dat genoom - NL interacties moduleert.

In Hoofdstuk 3 bestudeerden we een eiwit, RAD21, wat deel uitmaakt van het cohesine complex dat verantwoordelijk is voor de cohesie tussen zuster chromatiden tijdens de celdeling. Het is gesuggereerd dat cohesine ook betrokken is bij het reguleren van genexpressie in niet-delende cellen. Met behulp van induceerbare splitsing van RAD21 in *Drosophila* speekselklieren heeft Kim Nasmyth's lab laten zien dat cohesine de expressie van een specifieke groep genen reguleert. Door middel van een DamID bindingsprofiel van RAD21 in speekselklieren hebben we aangetoond dat deze genen gebonden zijn door RAD21. Bij elkaar tonen deze resultaten aan dat cohesine een direct effect heeft op de expressie van genen.

In Hoofdstuk 2 en 3 bestudeerden we individuele chromatine-eiwitten, en tot zover zijn de meeste publicaties ook gericht op individuele gen-clusters of eiwitten. Een globaal overzicht en begrip van chromatine organisatie ontbrak echter nog. Het doel van Hoofdstuk 4 was daarom om een overzicht te genereren van de eiwit samenstelling en functie van verschillende soorten chromatine. We produceerden een database van 53 bindings profielen en identificeerden 5 verschillende soorten chromatine, die gedefinieerd worden door een unieke combinatie van eiwitten. We hebben deze vijf typen genoemd naar de kleuren GROEN, BLAUW, ZWART, GEEL en ROOD. Wij vonden dat GROEN en BLAUW chromatine overeen komen met de twee bekende Polycomb en HP1 chromatine typen. De meeste van de inactieve genen worden echter gebonden door het nieuwe ZWARTE chromatine type, wat ongeveer 50% van het genoom omvat. Wij vonden dat het

transcriptioneel actieve euchromatin, wat tot nu toe werd beschouwd als één universeel chromatine type, in feite bestaat uit twee zeer verschillende typen, ROOD en GEEL. Genen in GELE regio's zijn over het algemeen betrokken bij de huishoudelijke functies en hebben een algemeen expressie patroon. Genen in RODE gebieden hebben daarentegen veel specifiekere functies en zijn strikter gereguleerd. Verder is gebleken dat deze verschillende soorten chromatine invloed hebben op de targeting van DNA-bindende factoren naar specifieke plekken in het genoom. Deze bevindingen geven een globaal overzicht van chromatine diversiteit en domein organisatie.

De moleculaire structuur van deze chromatine typen is echter nog onbekend en een groot deel van het chromatine proteoom is nog onontgonnen. In Hoofdstuk 5 hebben we eerst een systematische aanpak gevolgd om de ongekarakteriseerde fractie van het chromatine proteoom in kaart te brengen, wat resulteerde in de identificatie van 42 nieuwe chromatine-eiwitten. Ten tweede hebben we een data-gedreven model van het chromatine eiwit-netwerk gegenereerd. We toonden aan dat dit netwerk functioneel gecompartimentaliseerd is en we voorspelden de functies van het grootste deel van de nieuwe chromatine eiwitten, waaronder rollen in DNA replicatie en reparatie, en gen-activering en repressie. Bovendien bevestigt dit netwerk dat de definitie van de 5 chromatine typen robust is. Deze resultaten dragen bij aan een beter begrip van chromatine organisatie.

Al met al toont het werk in dit proefschrift aan dat het *Drosophila* genoom georganiseerd is in domeinen van verschillende chromatine typen, met elk een specifieke rol in genregulatie en ruimtelijke organisatie. Bovendien verbreedt dit werk de kennis van het chromatine proteoom en verschaft het nieuwe inzichten in de moleculaire samenstelling en functie van chromatine in *Drosophila*.

| SUMMARY

Chromatin is the ensemble of nuclear DNA, RNA and all associated proteins. The precise protein composition of the chromatin along the genome determines patterns of gene expression, DNA replication and other functions. In this thesis we have used DamID to study chromatin composition and organisation in *Drosophila melanogaster*.

In Chapter 2 we focused on the spatial organization of the genome in the nucleus, by analyzing genome - nuclear lamina (NL) interactions. The human genome was already known to be organized in hundreds large lamina-associated domains (LADs). It was however not clear if this organization is conserved in *Drosophila* and no proteins were known to regulate this association. We found that fly LADs, like human LADs, represent a strong repressive chromatin environment and contain about the same number of genes. Comparing the identified LADs to binding maps of all five known insulator proteins revealed SU(HW) to be the only insulator protein that preferentially binds inside LADs and at LAD borders. By knockdown and over-expression of SU(HW) we showed that SU(HW) locally weakens genome – NL interactions. These results provide insights into the evolution of spatial organization and identify SU(HW) as the first protein to fine-tune genome - NL interactions.

In Chapter 3 we studied a protein, RAD21, which is part of the cohesin complex that mediates sister chromatid cohesion during cell division. It has been suggested that cohesin might also be involved in regulating gene expression in interphase cells. Using inducible cleavage of RAD21 in *Drosophila* salivary glands Kim Nasmyth's lab found that cohesin regulates the expression of a distinct set of genes. By generating a DamID binding profile of RAD21 in salivary glands we have shown that differentially expressed genes after cohesin cleavage are bound by RAD21. Combined, these results demonstrate that cohesin acts directly on genes to regulate their expression.

In Chapter 2 and 3 we studied individual chromatin proteins, and so far most publications also focused on individual gene clusters or proteins. A global view and understanding of chromatin organization was however still lacking. In Chapter 4 we therefore aimed to generate a global view of the protein composition and function of different chromatin types. We generated a collection of 53 binding maps and identified 5 distinct chromatin types, which are defined by unique combinations of proteins. We named these five types by the colors GREEN, BLUE, BLACK, YELLOW and RED. We found that GREEN and BLUE chromatin correspond to the two well-known Polycomb and HP1 chromatin types. The majority of the inactive genes are bound by the novel BLACk chromatin type, which covers about 50% of the genome. We found that transcriptionally active euchromatin, which was untill now considered as one universal chromatin type, in fact consists of two very distinct types, RED and YELLOW. Genes in YELLOW regions tend to be involved in housekeeping functions and are broadly expressed, whereas genes in RED regions have more specific functions and are highly regulated. Furthermore, we found that these

different chromatin types help to target DNA-binding factors to specific genomic regions. These findings provide a global view of chromatin diversity and domain organization.

The molecular architecture of these chromatin types is however still largely unknown and a large fraction of the chromatin proteome is still completely unexplored. In Chapter 5 we first used a systematic approach to survey the uncharacterized fraction of the chromatin proteome and identified 42 novel chromatin proteins. Second, we generated a data-driven model of the chromatin protein network. We demonstrated functional compartmentalization of this network, and predicted functions for most of the novel chromatin proteins, including roles in DNA replication and repair, and gene activation and repression. Furthermore, this network confirmed the robustness of the definition of the 5 chromatin types. These results contribute to a better understanding of the organization of chromatin.

Taken together, the work presented in this thesis shows that the *Drosophila* genome is organized in domains of distinct chromatin types, each having specific roles in gene regulation and spatial organization. Furtermore this work broadens our knowledge of the chromatin proteome and provides novel insights in the molecular composition as well as the function of chromatin in *Drosophila*.

| SAMENVATTING VOOR IEDEREEN

Zoals de titel van dit proefschrift al aangeeft gaat dit proefschrift over chromatine–smaken, of in andere woorden over de samenstelling en organisatie van chromatine. Maar wat is chromatine eigenlijk? En waarom zijn we daar in geïnteresseerd? Hieronder zal ik proberen uit te leggen wat chromatine is en samenvatten welke resultaten er in dit proefschrift beschreven staan.

Van DNA naar eiwit

DNA is, zoals u waarschijnlijk wel weet, de drager van erfelijke informatie. DNA bestaat uit twee in elkaar gedraaide strengen: de bekende dubbele helix. Deze strengen zijn opgebouwd uit nucleotiden. De volgorde van deze nucleotiden bepaalt de erfelijke informatie. Een stukje DNA waarin één specifieke erfelijke eigenschap -bijvoorbeeld de kleur van je ogen- ligt vastgelegd wordt een gen genoemd. Op één DNA molecuul liggen duizenden genen. Al ons DNA kunt u eigenlijk beschouwen als een soort kookboek waarin de genen de recepten zijn en de nucleotiden de letters.

Elk gen codeert een bepaald eiwit. Eiwitten zijn de belangrijkste moleculen in je lichaam, bijna alles in je cellen bestaat uit of wordt gemaakt door eiwitten. De genen waarin de kleur van je ogen ligt vastgelegd coderen bijvoorbeeld voor gepigmenteerde eiwitten terwijl andere genen bijvoorbeeld coderen voor eiwitten die essentieel zijn om een spiercel te bouwen. Om een eiwit te maken dat gecodeerd wordt door een gen, wordt het DNA eerst gekopieerd naar mRNA. Dit mRNA wordt vervolgens gebruikt om een eiwit te bouwen uit aminozuren (zie ook Figuur 1 in Hoofdstuk 1). Aminozuren zijn de bouwstenen van eiwitten. Als we dit weer vergelijken met het kookboek, dan is het mRNA een gekopieerd recept aan de hand waarvan het gerecht (het eiwit) wordt gemaakt.

Gen regulatie

De mens, en alle andere organismen, is opgebouwd uit cellen. Cellen kunnen uiteenlopende kenmerken hebben: een spiercel is bijvoorbeeld duidelijk verschillend van een cel in de gekleurde iris in je oog. Alhoewel cellen verschillende eigenschappen kunnen hebben, is het DNA in elke cel toch identiek. Als het DNA identiek is, dan bevatten alle cellen dus precies dezelfde genen, wat betekent dat ze exact dezelfde recepten bevatten om eiwitten te maken. Toch bevatten niet alle cellen dezelfde eiwitten. Daaruit volgt dat niet alle genen gebruikt worden in iedere cel. Genen kunnen "aan" of "uit" staan in verschillende celtypen. In ons voorbeeld is het duidelijk dat in een spiercel het gen (ofwel het recept) voor oogkleur niet wordt gebruikt en dat dit gen dus staat uitgeschakeld.

Het aan- en uitschakelen van genen wordt voor een groot deel bepaald door specifieke eiwitten die aan het DNA binden. Sommige eiwitten maken het bijvoorbeeld mogelijk dat een gen gekopiëerd wordt naar mRNA, dat wil zeggen dat ze het kopiëren van een recept uit het kookboek mogelijk maken, terwijl andere eiwitten dit voorkomen. Genen die aan staan zullen dus gebonden zijn door andere eiwitten dan genen die uit staan. De combinatie van DNA en alle eiwitten die aan het DNA binden noemen we *chromatine*. Verschillende genen hebben

dus verschillende chromatine samenstellingen. In dit proefschrift hebben we chromatinesamenstelling bestudeerd door te analyseren welke eiwitten waar aan het DNA binden en hoe dit zich verhoudt tot het aan- en uitschakelen van genen.

DamID

Om te onderzoeken welk eiwit waar aan het DNA bindt hebben we gebruik gemaakt van een techniek genaamd DamID. Met DamID hang je een soort stempeltje aan het eiwit dat je wilt bestuderen. Zodra het eiwit aan het DNA bindt, laat de stempel een markering achter op het DNA (zie ook Figuur 6 in Hoofdstuk 1). Het DNA met de markering kan vervolgens uit de cellen worden geïsoleerd, en door de markering "af te lezen" kunnen we bepalen waar het eiwit op het DNA zat. Wij hebben deze techniek toegepast in gekweekte cellen van de fruitvlieg. De fruitvlieg, *Drosophila melanogaster*, is één van de meest bestudeerde modelorganismen. De principes van chromatinesamenstelling en genexpressie (= het kopiëren van een gen naar een mRNA-molecuul) zijn in fruitvliegjes hetzelfde als in mensen. Het fruitvliegenoom is echter veel kleiner, en het bestuderen daarvan is daarom veel gemakkelijker en betaalbaarder.

Chromatine smaken

Met behulp van de DamID-techniek hebben we voor 53 eiwitten geanalyseerd waar ze aan het DNA binden (Hoofdstuk 4). We vonden dat ze aan het DNA binden in 5 verschillende combinaties. Laten we om deze vinding te verduidelijken een stuk DNA vergelijken met een boterham. Stel dat u 53 ingrediënten in uw koelkast heeft, dan kunt u deze ingrediënten combineren en een eindeloze (53x52x51 … … x3x2x1= 4,3e69, dit is een getal van 70 cijfers!) hoeveelheid van boterhammen met verschillende smaken maken. Bepaalde ingrediënten combineren echter beter samen dan andere. Kaas en tomaat gaan goed samen op een broodje, terwijl hagelslag en tomaten niet zo goed combineren. Dit beperkt het aantal mogelijke combinaties. Bij DNA bleken chromatine-eiwitten aan het DNA te binden in 5 verschillende combinaties, in ons voorbeeld komt dat overeen met 5 boterhammen met 5 verschillende combinaties van ingredienten. In lijn met de betekenis van het Griekse woord "chroma" (kleur), hebben we deze vijf chromatine-smaken genoemd naar de kleuren GROEN, BLAUW, ZWART, GEEL en ROOD.

GROEN en BLAUW chromatine bleek overeen te komen met combinaties van eiwitten die al langer bekend en uitgebreid bestudeerd zijn. Lange tijd werd aangenomen dat uitgeschakelde genen één van deze twee chromatine-smaken hadden. Wij vonden juist dat twee derde van de genen die zijn uitgeschakeld de nieuw ontdekte ZWARTE chromatine-smaak hebben. Slechts een klein deel van de inactieve genen hebben een BLAUWE chromatine-smaak, en een groot deel van de genen met een GROENE chromatine-smaak blijken niet eens echt uitgeschakeld te zijn. De ontdekking van ZWART chromatine helpt ons dus om te begrijpen hoe een groot deel van de genen uitgeschakeld wordt.

Van genen die ingeschakeld staan werd altijd gedacht dat deze allemaal werden gebonden door dezelfde combinatie van chromatine eiwitten. Wij ontdekten dat deze genen eigenlijk twee verschillende smaken chromatine kunnen hebben, namelijk ROOD of GEEL. Genen die altijd ingeschakeld staan, ook

in verschillende soorten cellen en weefsels, bleken een GELE chromatine-smaak te hebben, terwijl genen die alleen in specifieke celtypen of weefsels aan staan een RODE chromatine-smaak hebben. In ons eerdere voorbeeld zou dat betekenen dat een gen dat zowel aan staat in de ogen als in de spiercellen waarschijnlijk een GELE chromatine-smaak heeft, terwijl een gen dat aan staat in de oog cellen en uit in de spiercellen, waarschijnlijk een RODE chromatine-smaak zal hebben als het is ingeschakeld en waarschijnlijk een BLAUWE of ZWARTE als het is uitgeschakeld. Hier wordt momenteel verder onderzoek naar gedaan in verschillende weefsels van het fruitvliegje.

RAD21 in GEEL chromatine

Genen die aan staan kunnen dus twee smaken hebben, ROOD of GEEL. In Hoofdstuk 4 hebben we aangetoond dat de eiwitsamenstelling van deze twee soorten chromatine bijna hetzelfde is. Slechts één eiwit bleek specifiek aanwezig in GEEL chromatine en niet in ROOD chromatine. In Hoofdstuk 3 onderzochten we het eiwit RAD21, waarvan bekend is dat het betrokken is bij de scheiding van DNA tijdens de celdeling. Met DamID hebben we aangetoond dat RAD21 ook in cellen die niet aan het delen zijn aan het DNA bindt. We vonden dat het aan specifieke genen bindt, en een rol speelt bij het aan- en uitschakelen van deze genen. In Hoofdstuk 5 vonden we vervolgens dat dit eiwit veel meer in GEEL dan in ROOD chromatine bindt, en dus waarschijnlijk belangrijk is voor GEEL chromatine.

De ruimtelijke organisatie van chromatine

Het aan en uit zetten van genen wordt niet alleen gereguleerd door de eiwitten die aan het DNA binden maar ook door de lokalisatie in de celkern. De celkern is een compartiment in de cel die al het DNA bevat, en gescheiden is van de rest van de cel door een membraan. Om te analyseren welke delen van het genoom zich binnen in de celkern bevinden en welke delen zich aan de rand bevinden hebben we de eerder beschreven DamID stempel gefuseerd aan het eiwit Lamin (Hoofdstuk 2). Lamins vormen de lamina die zich aan de binnenzijde van het kernmembraan bevindt. Al het DNA dat contact maakt met de lamina, en dus met de stempel, wordt gemarkeerd. Dit kunt u vergelijken met een bal (de celkern) die vol zit met wol (het DNA). Als u de binnenkant van de bal rood verft, dan zal de wol die de wand van de bal raakt rood worden. Door de wol uit de bal te halen kunt u nu zien welke delen zijn rood geschilderd zijn en dus tegen de rand zaten, en welke delen niet gekleurd zijn en in het midden zaten. Hiermee hebben we gevonden dat 40% van het genoom van de fruitvlieg in contact is met de lamina, en zich dus aan de rand van de celkern bevindt. Interessant is dat het DNA dat aan de rand zit voornamelijk inactieve genen bevat, terwijl de genen die aan staan voornamelijk in het midden zitten.

We hebben nu dus twee aspecten die samenhangen met het aan en uitzetten van genen, de combinatie van eiwitten (de smaak van het chromatine) en de lokalisatie in de kern. Maar hoe hangen die twee dan met elkaar samen? Uit de vinding dat inactieve genen ZWART en som BLAUW

zijn, volgt dat BLAUW en ZWART chromatine waarschijnlijk aan de rand van de kern zullen zitten. Dat is inderdaad wat we vonden toen we chromatine-smaken vergeleken met de stukken DNA die in contact zijn met de lamina.

Tot nu toe was er geen enkel eiwit bekend dat de interacties tussen het genoom en de lamina kan regelen. Specifieke eiwitten, die men isolator-eiwitten noemt, worden verondersteld een belangrijke rol te spelen in de chromatine-organisatie. Wij hebben daarom vijf van deze isolator-eiwitten bestudeerd en vonden dat één van deze eiwitten, SU(HW), de interacties tussen het genoom en de lamina plaatselijk kan verminderen.

Nieuwe chromatine-eiwitten

Tot dusver hebben we alleen bekende chromatine-eiwitten onderzocht. Maar het is zeer waarschijnlijk dat er nog vele andere on-ontdekte chromatine-eiwitten zijn. Om nieuwe chromatine-eiwitten te vinden hebben we de DamID stempel gehangen aan 112 eiwitten waarvan we voorspeld hebben dat ze onderdeel zijn van chromatine, maar waarvan niemand dat tot nu toe had aangetoond (Hoofdstuk 5). We hebben gekeken of deze 112 eiwitten een markering achterlieten op het DNA, en vonden dat 42 eiwitten dat inderdaad doen. Een eiwit met een stempel laat alleen een markering achter als het in contact komt met het DNA. We hebben dus 42 nieuwe eiwitten ontdekt die contact kunnen maken met DNA en deel uit maken van chromatine.

Vervolgens vroegen we ons af of dit zou betekenen dat we nu meer chromatine smaken zouden vinden. Tot nu toe keken we naar 53 eiwitten en vonden 5 combinaties. Nu hebben we namelijk nog veel meer eiwitten (of in het geval van het voorbeeld, meer ingrediënten in onze koelkast die we kunnen gebruiken om verschillende smaken broodjes te maken). Het blijkt dat zelfs met 42 extra eiwitten er nog steeds 5 combinaties optreden. De nieuw ontdekte eiwitten maken allemaal deel uit van één van de vijf eerder beschreven chromatine smaken.

In Hoofdstuk 5 hebben we vervolgens een model-netwerk gemaakt dat de interacties tussen alle chromatine-eiwitten laat zien. In dit netwerk vinden we clusters van eiwitten met een vergelijkbare biologische functie of chromatine smaak. Dit netwerk levert ons nieuwe inzichten in de samenstelling en de functie van chromatine in de fruitvlieg.

Samengevat, hebben we laten zien dat het DNA is georganiseerd in 5 chromatine-smaken en dat elke smaak een specifieke rol heeft in het aan en uit zetten van genen. Bovendien bleken inactieve genen voornamelijk aan de rand van de celkern te ziten en vonden we een eiwit dat bij deze lokalisatie betrokken is. We vonden dat het eiwit dat betrokken is bij de scheiding van DNA tijdens de celdeling ook aan genen bindt en deze reguleert. En tot slot ontdekten we 42 nieuwe chromatine-eiwitten en maakten een model van de moleculaire organisatie van het chromatine.

| CURRICULUM VITAE

Joke van Bemmel werd geboren op 17 april 1982 te Heemskerk. In 2000 behaalde zij haar VWO diploma aan het Bertrand Russell College te Krommenie. Datzelfde jaar begon zij de studie Bio-medische wetenschappen aan de Universiteit van Amsterdam, waar zij haar Bachelor diploma behaalde in 2003. Na een minor Wetenschaps Dynamica begon zij in 2004 aan de Master opleiding Bio-medische wetenschappen. Tijdens deze master heeft ze eerst de aanwezigheid van histone varianten in verschillende weefsels bestudeerd onder begeleiding van Maike Stam in de onderzoeksgroep van Roel van Driel aan de Universiteit van Amsterdam. Vervolgens heeft ze in Berlijn, Duitsland, in de groep van Cristina Cardoso onder begeleiding van Sabine Görisch de replicatie timing van specifieke chromatine typen bestudeerd. In 2006 behaalde zij haar Master diploma Cum Laude, waarna zij in 2007 begon als onderzoeker in opleiding in de groep van Bas van Steensel aan het Nederlands Kanker Instituut te Amsterdam. De resultaten van dat onderzoek staan beschreven in dit proefschrift. Na haar promotie zal Joke van Bemmel als post-doc de rol van chromatine in X-inactivatie gaan bestuderen in de groep van Joost Gribnau aan het Erasmus MC te Rotterdam.

| LIST OF PUBLICATIONS

Casas-Delucchi CS, **van Bemmel JG**, Haase S, Herce HD, Nowak D, Meilinger D, Stear JH, Leonhardt H, Cardoso MC. Histone acetylation controls replication timing of constitutive heterochromatin. Nucleic Acids Res, 40 (1), 2012, 159-69

van Bemmel JG, Pagie L, Braunschweig U, Brugman W, Meuleman W, van Steensel B. The Insulator Protein SU(HW) Fine-Tunes Nuclear Lamina Interactions of the *Drosophila* Genome. PLoS ONE, 5 (11), 2010, e15013

Pauli A, **van Bemmel JG**, Oliveira RA, Itoh T, Shirahige K, van Steensel B, Nasmyth K. A direct role for cohesin in gene regulation and ecdysone response in Drosophila salivary glands. Curr Biol, 20 (20), 2010, 1787-98

Filion GJ*, **van Bemmel JG***, Braunschweig U*, Talhout W, Kind J, Ward LD, Brugman W, de Castro IJ, Kerkhoven RM, Bussemaker HJ, van Steensel B. Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. Cell, 143 (2), 2010, 212-24
***these authors contributed equally**

van Steensel B, Braunschweig U, Filion GJ, Chen M, **van Bemmel JG**, Ideker T. Bayesian network analysis of targeting interactions in chromatin. Genome Res, 20 (2), 2010, 190-200

| ACKNOWLEDGEMENTS

En dan ligt er na ruim 5 jaar ineens een boekje. Veel mensen hebben daar aan bijgedragen, zowel aan het wetenschappelijke deel als aan mijn 'geestelijk welzijn'. Ik heb een fantastische tijd gehad op het NKI, en niet in de laatste plaats dankzij alle fijne collega's. Ik wil dan ook graag van de gelegenheid gebruik maken om iedereen te bedanken.

Allereerst **Bas**, als wetlab bioloog was het best spannend om in een genomics lab aan de slag te gaan. Het was dan ook erg fijn dat je me altijd de ruimte hebt gegeven om mezelf te ontwikkelen, maar daarnaast ook altijd klaar stond voor de nodige begeleiding. Onze samenwerking heb ik als inspirerend en leerzaam ervaren en ik waardeer je creativiteit en immer positieve kijk. Bedankt voor een wat mij betreft in alle aspecten ontzettend geslaagde promotie periode, ik had me geen betere begeleider kunnen wensen. **Hein te Riele, Jos Jonkers, Maarten Fornerod, Reuven Agami, Fred van Leeuwen** als wisselende leden van mijn begeleidingscommissie heb ik door de jaren heen veel advies van jullie mogen ontvangen. Bedankt voor jullie input en kritische kijk. **Lodewijk Wessels**, dan wel geen lid van mijn begeleidingscommissie maar ondanks dat stond ook jij altijd klaar voor een adviserend woord, bedankt daarvoor. Chapter 3 of this thesis would not have been there without **Andrea Pauli**. Andi, I really enjoyed our collaboration and your visit in Amsterdam. It was very stimulating to work together. I hope we keep in touch.

De laatste twee hoofdstukken van dit proefschrift leken twee megalomane projecten die onmogelijk door één promovendi alleen uitgevoerd hadden kunnen worden. Ze vereisten een groepsaanpak, waarin nauw werd samengewerkt en de verschillende vaardigheden van collega onderzoekers samen kwamen. Het was soms een uitdaging, maar bovenal heb ik het als een voorrecht beschouwd om al die jaren in een team te mogen samenwerken waarin we elkaar versterkten en stimuleerden. **Uli**, I would have been nowhere in the lab without you. Thank you for teaching me the secrets of DamID, RNAi and R (I am still learning from your scripts). I miss your knowledge, your opinions and most of all your humor. All the best in Toronto together with Sasha. Sasha, you know how to host a party and cook good food, thanks for the good times! **Wendy**, zonder jouw was dit boekje er niet geweest. Ik bewonder je nauwkeurigheid en doorzettingsvermogen bij het cloneren en mappen van alle eiwitten. En daarnaast had je ook altijd nog tijd voor de bestellingen, een gezellig gesprek, het ontdooien van een vriezer, het luisteren naar mijn frustraties of een discussie met inkoop. Veel succes aan het brandwonden front. **Guillaume,** you are a tremendous bio-informatician. I really enjoyed working together to combine the wet and the dry lab. Thanks for not getting tired with telling me which statistical test to use when. You do not only have a talent for data analysis but also for 'people-skills', both from which I have learned a lot. I am sure you are (going to be) a great group leader. I am looking forward to read your first senior-author paper. **Arantxa**, the latest addition to our team. I love your hands on mentality and appreciate your positive character. It is great to see how you "just get things done". Thanks for joining the profiling of the endless amount of proteins. Good luck with the transposon project!

&

Vanaf dag één heb ik me thuis gevoeld in het lab, mede dankzij de jongens in de "bio-informatica"-kamer. Dank dat ik als beginnende wetlab oio bij jullie in de kamer 'mocht' zitten. Ik heb veel van jullie geleerd maar vooral ook veel gelachen. **Wouter**, dank voor alle gein & ongein, de zinnige & onzinnige gesprekken, de gevraagde & ongevraagde R adviezen en natuurlijk het gefeest in Berlijn, op Lowlands en waar niet. Jouw aanwezigheid staat garant voor zowel kritische en waardevolle wetenschappelijke input als voor een goed feestje (en helaas ook voor de bijbehorende kater). Ben benieuwd of je natte lab-skills net zo goed zijn als je droge. **Ludo**, R-class is the best! Ik en hoofdstuk 2 zouden nergens zijn geweest zonder CG-tools. Keep up the good work. **Lars**, je had altijd wel een leuk nieuw bandje, een mooi fiets avontuur of een verhaal uit het oosten van het land. Als ik ooit nog eens "op de fiets stap" ben je de eerste die het hoort. Veel geluk samen met je dames. **Elzo**, bedankt voor de hulp bij mijn eerste stapjes in de wondere wereld van R. Met je sarcasme en humor zorgde je altijd voor een vrolijke noot. Veel succes in de wetenschap en geluk met Iris. Leuk om jullie nog regelmatig tegen te komen. **Daan**, de nuchterheid zelve. Zowel jouw muziek als jouw literatuur kennis zijn onevenaarbaar. Waarschuw je even voordat je world domination in gang zet? Of toch maar eerst een post-doc? Succes!

Deze kamer indeling bestaat allang niet meer, collega's kwamen en gingen, we zijn verhuisd maar altijd was er de basis van een fijne groep vanSteenseltjes. **Maartje**, een verademing dat wetenschappers ook gouden laarzen kunnen dragen in plaats van 'geitenwollen sokken'. Ik bewonder je professionele instelling en ik heb je gezelligheid en hulpvaardigheid erg gewaardeerd. **Jop**, jouw creativiteit is bewonderenswaardig, jij komt er wel. Dank voor de gezelligheid, de wetenschappelijke input, de goede gesprekken en de tango lessen. Sleep well! (Al moet ik toegeven dat ik je semi-irritante slaaploosheids buien stiekem best gezellig vond). **Dominika**, you are truly a wonderful person. Thanks for appreciating and pointing out the beautiful things in life. You were a great partner-in-crime in Paris, and a good friend in Amsterdam. I wish you all the best in Bristol and beyond. **Alex**, I am still thinking of a molecular-biology question you are not able to answer. I am impressed by your knowledge and accurate labwork. Looking forward to see where your -undoubtedly successful- scientific career will lead you. **Katy**, always the first to initiate a creative birthday present, to organize a lab activity or to bring home-made cookies. I enjoyed having you as a colleague. **Carolyn**, I am impressed how you combine wet and dry lab and got your project going quickly and efficiently. Meanwhile you don't forget to enjoy life in Amsterdam. Good going! **Joris,** al voor ik je kende was ik onder de indruk van je beurs aanvraag. In combinatie met jouw humor en inzet zit het wel snor met jouw in het lab. Take good care of the color-flies. **Mario**, thanks for continuously providing us with Italian goodies (loved the cheese!). I hope the colors bring you more luck than the RNA project.

Buiten de vanSteensel groep wil ik graag iedereen van de **CMF** bedanken, en in het bijzonder **Ron** en **Wim**. Alhoewel micro-arrays inmiddels al ouderwets en bijna vergeten zijn, zou dit proefschrift er niet zijn geweest zonder. Dank voor het opzetten van de in-house

hybridisatie. En -ik heb ze geteld- mega dank voor het hybridiseren van het indrukwekkende aantal van 332 arrays die nodig waren voor dit proefschrift. **Tom** en **Marcel** dank voor het reilen en zeilen op P2. Op B4: **Mariette** en **Suzanne**, wat zouden we moeten zonder jullie? **Iris** (CG5181 gaat het worden hoor, ooit...) en de overige gist-mensen, het was fijn een lab met jullie te delen. **Bernike**, bedankt voor je input in de verschillende projecten en voor het delen van praktisch tips en reagentia. Drinken we een kopje koffie in Rotterdam? **Roel,** dank voor je inbreng aan goede muziek in de cel kweek. **Jarno** (dank voor de promotie-regel-tips), **Marieke** en de overige Agami's, met jullie was het alles behalve saai op B4.

Then there is this group of people who were there, following more or less the same path, going to Texel, sharing meetings and courses, having fun at borrels and party's, and sharing frustrations and PhD progress. **Eva** (Dank voor de taart en de kopjes thee in gezellige en mindere gezellige tijden. Succes met de laatste loodjes), **Chris** en **Sietske** (All the best in the US!), **Marieke V**, **Bastiaan**, **Johan**, **Jeroen**, **Izhar**, **Andrej**, **Kitty**, **Fra**, **Joep**, **Yme**, **Julian**, **Jorma** and everyone else, thanks for the good times.

Ook al zijn de volgende mensen niet direct betrokken bij mijn promotie onderzoek, ze stonden wel aan het begin en creëerden de mogelijkheden voor deze promotie. **Roel van Driel**, bij jou ligt de kiem voor mijn fascinatie met chromatine. Jij was de eerste die ons studenten echt uitdaagde en ons liet kennis maken met echte wetenschap in plaats van boeken. Vervolgens gaf je me de mogelijkheid een stage te doen in je lab, ook al was ik eigenlijk ergens anders ingedeeld, dank daarvoor. **Maike**, mijn eeuwige dank dat je me zo precies en exact proeven hebt leren doen. Ik heb bewondering voor je precisie, en heb er nog altijd profeit van. **Marieke**, time flies en ik ben blij dat onze vriendschap, die in het lab begon, verder reikt dan dat. Ik wens jou en JB alle geluk van de wereld, want dat verdien je. **Cristina Cardoso** thanks for the great and instructive time in your lab. I am still impressed by your incredible energy, motivation and speed in thinking as well as talking. **Sabine**, I am still grateful for teaching me a bit of your organizational skills. I would have been lost without them in this project with hundreds of different proteins.

Ondanks dat ik me had voorgenomen een kort dankwoord te schrijven (wat jammerlijk mislukt is) wil ik hier ook graag de mensen buiten de wetenschap bedanken voor hun steun en gezelligheid. Ik had namelijk niet zo'n leuke tijd in de wetenschap kunnen hebben zonder een leuke tijd buiten het lab. Na een dag in het lab is sporten altijd mijn uitlaatklep geweest. Mijn klimmaatjes **Sanne, Joost, Sander, José** en **Melanie**, het zijn altijd fijne avonden vol activiteit en lekkere biertjes. Bedankt voor de gezelligheid en ook het geduld als ik weer eens te laat was of eerst nog een lab frustratie moest ventileren voor we dan eindelijk konden gaan klimmen. Beter een goede buur dan een verre vriend: **Peter en Esther**, zo fijn om naast jullie te wonen, dat er nog maar veel wijntjes in de tuin mogen volgen. Esther, ik sta bij je in het krijt voor de inDesign les. **Anika** en **Gianna**, jullie zorgen ervoor dat ik niet in een complete wetenschaps-nerd verander (ik zal nooit meer blauwe vloerbedekking kopen, beloofd). Dank voor de fijne lunch-dates, high-tea's,

shopsessies, etentjes en andere gezelligheid. **Maartje**, ik ken je al zo lang en het is fijn om te merken hoe onze vriendschap zich blijft ontwikkelingen en me iedere keer weer positief verrast. Ik ben blij dat jij en **Bram** weer in het land zijn, en ik hoop nog veel goede gesprekken en fijne buiten activiteiten met jullie te delen.

Wout, you've made me a different person, daar kan ik je niet genoeg voor bedanken. Je bent een bijzonder mens, en ik koester de herinneringen aan een fijne tijd. Ik hoop dat we ooit weer eens als vrienden samen kunnen genieten van een lekker biertje of een mooie berg.

Dewi, Paul, Mark en **Roos** jullie zijn een stel super de puper vrienden, stelling 11 is voor jullie! Hoeveel ik jullie waardeer en waarom vertel ik graag nog eens in levende lijve. Dewi en Mark, ik ben zo blij dat we ooit hebben besloten samen het bestuur aan te gaan. **Paul**, ik ben jaloers op je kritische blik en bijna kinderlijke nieuwsgierigheid. Ik waardeer onze wetenschappelijke discussies en vind het een eer dat je naast me staat als paranimf. Nu op naar je eigen promotie, die zonder twijfel succesvol gaat zijn.

Carmen, tjeez ik weet niet eens waar ik moet beginnen. De peuterspeelzaal, de vele punk-bandjes en de Weiver avonden, onze studie tijd, Indonesië, klimmen, borrels, lief en leed, gezelligheid, wederzijdse adviezen of troostende woorden.... onze vriendschap is een contstante en super waardevolle factor in m'n leven. En –hoe clicheé – ik heb er geen woorden voor om dat hier te benoemen. Dank dat je mijn parnimf wilt zijn. **Arjen**, ik had me geen liever, fijner en leuker vriendje kunnen wensen voor Carmen dan jij.

Tot slot mijn ouders en broertje. **Martijn**, pannenkoek, ondanks -of misschien wel dankzij- dat we het nooit met elkaar eens zijn (stelling 9 is voor jou) ben je een super toffe broer. Ik ben blij dat we elkaar vaak spreken en vind het bijzonder om te merken dat je er altijd voor me bent. **Pap** en **mam**, jullie opvoeding staat aan het begin van dit boekje. Jullie hebben me altijd gestimuleerd, mijn nieuwsgierigheid geprikkeld en me kritisch laten nadenken. Al vroeg wisten jullie de wetenschapper in mij los te maken en ging ik met veel plezier naar NEMO, zat ik op een basisschool waar ik mijn eigen excursies mocht regelen of deed ik mee aan "Professor-Post". Dank daarvoor. Daarnaast, en dat is me eigenlijk veel meer waard, staan jullie altijd voor ons klaar en zijn jullie mijn rots in de branding.

Dit boekje is voor jullie.

Joke