Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome

Cell

Three-Dimensional Folding and Functional Organization Principles of the *Drosophila* Genome

Tom Sexton,[1,4] Eitan Yaffe,[2,4] Ephraim Kenigsberg,[2] Frédéric Bantignies,[1] Benjamin Leblanc,[1] Michael Hoichman,[2] Hugues Parrinello,[3] Amos Tanay,[2,*] and Giacomo Cavalli[1,*]

[1]Institut de Génétique Humaine, UPR 1142, CNRS, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France

[2]Department of Computer Science and Applied Mathematics and Department of Biological Regulation, Weizmann Institute of Science, Rehovot 76100, Israel

[3]Montpellier GenomiX IBiSA, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France

[4]These authors contributed equally to this work

*Correspondence: amos.tanay@weizmann.ac.il (A.T.), giacomo.cavalli@igh.cnrs.fr (G.C.)

DOI 10.1016/j.cell.2012.01.010

SUMMARY

Chromosomes are the physical realization of genetic information and thus form the basis for its readout and propagation. Here we present a high-resolution chromosomal contact map derived from a modified genome-wide chromosome conformation capture approach applied to *Drosophila* embryonic nuclei. The data show that the entire genome is linearly partitioned into well-demarcated physical domains that overlap extensively with active and repressive epigenetic marks. Chromosomal contacts are hierarchically organized between domains. Global modeling of contact density and clustering of domains show that inactive domains are condensed and confined

Understanding chromosome structure fully is therefore a fundamental task in genomic and epigenetic research, and different hypotheses on the causative or consequential nature of chromosomal folding patterns have major implications on our understanding of how genetic information is encoded and interpreted. Various mathematical models have been proposed to explain the effects of different physical factors on chromosome fiber folding (reviewed in Heermann, 2011; Lieberman-Aiden et al., 2009; Mateos-Langerak et al., 2009; Münkel et al., 1999; Sachs et al., 1995), but high-resolution, genome-wide measurements of DNA fragment interdistances or interaction frequency are required for their rigorous assessment.

The development of the chromosome conformation capture (3C) technique, which allows detection of genomic regions that are in close proximity in vivo (Dekker et al., 2002), and its integration with genomic methods have allowed chromatin topology

in metazoans. The recent explosion in sequencing throughput, the advent of new computational techniques to analyze 3C-based datasets (de Wit et al., 2008; Yaffe and Tanay, 2011), and extensive data on linear epigenomic marks (Ernst et al., 2011; Filion et al., 2010; Kharchenko et al., 2011) put us in an unprecedented position to obtain and understand high-resolution chromatin interactions in the context of genome function.

Here we present the application of a modification of the Hi-C approach to derive a high-resolution contact map of fly embryonic chromosomes. Most importantly, chromosomes are shown to be organized hierarchically in a highly functional manner. First, multiple genes build chromosomal domains (10–500 kb) delimited by sharp boundaries defined by insulator binding, peaks of DNase hypersensitivity, and/or the active histone mark H3K4me3. Second, we show that domains are folded distinctively and interact hierarchically as units. Inactive chromosomal domains cluster together and are strongly associated with the chromosome territory. On the other hand, active domains are less compact and more likely to form interchromosomal contacts with other active, but not inactive, domains. We develop a probabilistic model to predict the contact map structure from simple principles, including the domain structure and the clustering of active and repressed domains. This leads to the systematic identification of contacts that cannot be explained by these observed principles of global chromosomal organization, including specific contacts between PcG-regulated domains. The data therefore demonstrate the multiple levels at which chromosomal architecture interacts with genome function.

RESULTS

A Comprehensive Map of Chromosomal Contacts in Fly Embryonic Nuclei

To develop a robust and simplified version of the Hi-C technique (Figure 1A), we collected nuclei from fixed embryos and performed 3C using the frequently cutting restriction enzyme DpnII. Ligation and DNA purification as for conventional 3C were followed by sonication and size selection of long (~800 bp) products, resulting in strong enrichment for sequences including at least two DpnII-ligated fragments (see Extended Experimental Procedures and Figure S1A, available online, for details and experimental validation). We sequenced over 362 million paired-end tags and processed them extensively to generate genome-wide quantification of contact intensities (extending Yaffe and Tanay, 2011) (Figures S1B–S1E; see Extended Experimental Procedures). As shown in Figure 1B, the genome-wide, technically corrected contact map that we derived from over 118 million stringently filtered sequences effectively covers interactions inside the *Drosophila* chromosome arms, between arms, and between chromosomes. As predicted and observed previously (Dekker et al., 2002; Lieberman-Aiden et al., 2009), interaction frequencies are reduced with increasing distance in base pairs along the linear chromosome. The global decay in contact frequency as a function of genomic distance is similar for all chromosome arms (Figure 1C) and can be approximated by a power law with a scaling exponent of −0.85 (with positive deviations from the regime in the 10–100 kb interval and the 1–10 Mb interval). Consistent with the idea of constrained chromosomal territories (Cremer and Cremer, 2010; Lanctôt et al., 2007), the frequency of interarm contacts (2L-2R and 3L-3R) is on average similar to the intensity of the most distal intraarm contacts, and interchromosomal contacts are on average 4-fold less frequent. The map confirms other known cytological features of fly embryonic nuclei by revealing preferential contacts between all telomeres (Figure 1D) and demonstrating coclustering of centromeres and the heterochromatin-rich chromosome 4 (Lowenstein et al., 2004). To validate and demonstrate the resolution of the map, we used it to reproduce local contact profiles at loci that were studied previously using 4C (Bantignies et al., 2011; Tolhuis et al., 2011). As exemplified in Figure 2A (an analogous map generated from a 36 million read pilot experiment is shown in Figure S1F), a contact matrix for a 14 Mb region in chromosome 3R can be dissected into informative local profiles (termed virtual 4C, Figure 2B), which clearly recapitulate long-range contacts between two PcG-regulated Hox gene clusters, ANT-C (Antennapedia complex) and BX-C (bithorax complex). This coassociation was previously observed by DNA FISH and 4C studies (Bantignies et al., 2011; Grimaud et al., 2006; Tolhuis et al., 2011), validating our Hi-C technique. However, the Hi-C map goes further than these initial observations, quantitatively demonstrating a 10-fold enrichment in contact frequency between the two Hox clusters over flanking genomic regions and flexibly providing data on additional virtual 4C viewpoints (Figures S2A and S2B).

The Genome Is Partitioned into Well-Demarcated Physical Domains

To systematically explore the *Drosophila* chromosomal contact map, we developed a quantitative probabilistic approach to model as much of the structure as possible using simple folding principles and inferring local and global chromosomal properties while refining the model progressively. As a first step toward this goal, we modeled variation in local Hi-C connectivity as illustrated in Figure 2C. According to the global genomic trend (Figure 1C), two regions X and Y can be predicted to contact with a probability $P(\Delta L_{basic})$ that depends only on their genomic distance ΔL_{basic}. By fitting a *distance-scaling factor* to each restriction fragment, our model rescales genomic distances such that the prediction of contact probabilities based on modified distances ($P(\Delta_{scaled})$) is improved. As illustrated in Figure 2C, frequently contacting elements may be modeled more accurately if the scaling factors of fragments in between them are low. On the other hand, fragments with high scaling factors allow modeling of physical chromosomal borders, as contacts crossing them appear less likely than otherwise expected. Analysis of the distance-scaling factors that were inferred from the data using an iterative maximum likelihood algorithm, and the comparison of these factors to contact intensities around the Hi-C matrix diagonal (Figures 2D and 2E), revealed striking contact-enriched submatrices that were flanked sharply by chromosomal elements with high scaling factors. This led to the systematic identification of *physical domains* as contiguous chromosomal regions that are flanked by distance-scaling peaks (Figure 2E). The 1,169 physical domains that were defined this way (Table S1; for distribution of sizes, see

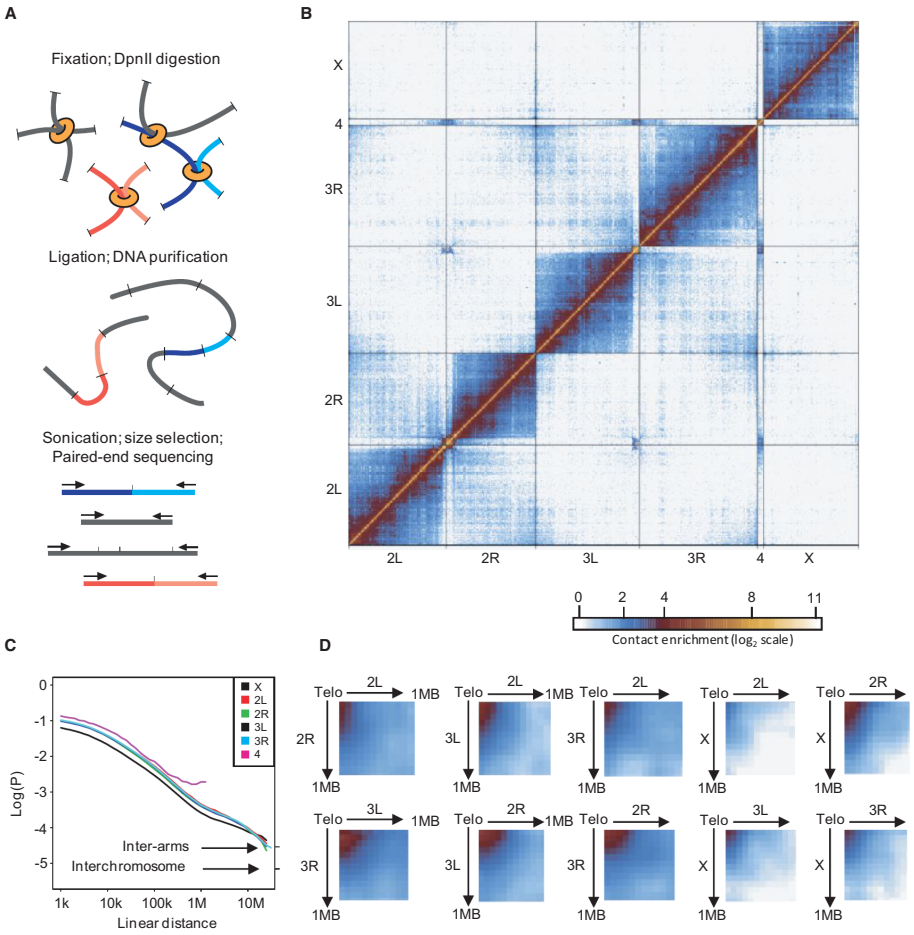

Figure 1. A High-Resolution Normalized Contact Map for the Fly Genome

(A) The simplified Hi-C technique. Nuclei are fixed and digested with the frequently cutting enzyme DpnII and ligated to capture interactions between fragments that were covalently linked during fixation (shown for four chromatin interactions). Purified DNA is sonicated and products ~800 bp (a size larger than two DpnII fragments on average) are size selected and sequenced.

(B) Genome-wide normalized contact map for fly embryonic nuclei. Each element in the matrix represents the ratio between the number of observed contacts in a two-dimensional genomic bin and the number expected from a probabilistic model correcting for systematic biases.

(C) Graph of contact probability as a function of distance in linear genomic space, plotted separately for each chromosome. The X chromosome trend is corrected by a factor of 4/3 to account for the mixed male:female population. The contact probabilities for pairs of restriction fragments on different arms of the same chromosome, or on different chromosomes, are indicated on the right.

(D) Parts of the genome-wide contact map are shown to highlight stronger than expected contacts between each of the five mappable telomeric regions of the genome.

See also Figure S1.

Figure S2C) reorganize the 130 Mb nonrepetitive fly genome in a remarkably concise and well-demarcated fashion. The clear partitioning of the contact map into domains was also observed in an independent, lower-resolution Hi-C experiment (Figure S2D) and suggests that quantitative understanding of chromosomal contacts in *Drosophila* can be facilitated by in-depth study of the structure, epigenetic definition, and higher-order interactions of physical domains.

Physical Domains Reflect Epigenetic Domains

Previous extensive profiling of histone modifications and binding of chromatin proteins has promoted the view of the fly genome as a set of epigenetic domains, correlated with the regulation of the underlying loci (Filion et al., 2010; Kharchenko et al., 2011). We therefore studied the link between the contact map's physical domains and a comprehensive reference collection of linear epigenetic profiles available for *Drosophila*. Systematic screening of 403 linear epigenomic profiles indicated a strong statistical association of the physical domains with 315 of the available marks ($p < 0.001$; FDR-adjusted resampled chi-square test; Table S2, see also Figure S2E), highlighting the physical and linear epigenomic domains as strongly linked chromosomal properties. This association was observed even though the profiles were generated by different techniques at varying developmental stages, suggesting that some of the domain structure we characterized may be present at multiple stages and conditions. To annotate physical domains given these extensive data we used unsupervised (Figure S3) and supervised (Figure 3) clustering of physical domains given average epigenetic enrichments. In both cases, we identified four major domain classes characterized by a clear biological function, good overlap with previously characterized epigenomic domains (Filion et al., 2010) (Figure S4A), and a remarkable demarcation of their characteristic epigenetic marks over the domain borders (Figure 3D).

Out of the four classes, "Null" domains were not enriched for any available mark (except for a weak enrichment in binding of the insulator protein Su(Hw)) and spanned a large proportion (59%; 492 domains) of the genome. As previously described (Filion et al., 2010; Kharchenko et al., 2011; Schwartz et al., 2010), these null or void chromatin domains are on average larger than those from the other classes (Figures 3A and 3B) and have a very low transcriptional output (Figure 3C), despite having comparable gene densities to the other classes. Transcriptionally active ("Active") domains, associated with histone marks such as H3K4me3, H3K36me3, and hyperacetylation (Figure 3E), comprise 42% of the domains and 22% of the genome. The other two classes of physical domains entail well-described forms of repressive chromatin: one bound by PcG proteins and associated with the mark H3K27me3 ("PcG"), and one bound by the heterochromatin proteins HP1 and Su(var)3-9 and associated with H3K9me2 ("HP1/Centromere"). The latter class is associated with classical heterochromatin and therefore is predominantly centromeric and has a low coverage within the nonrepetitive, mappable part of the genome analyzed in this study. As expected, these two classes tend to form larger domains than Active domains and have a low transcriptional output. The distribution of scaling factors inside the three classes of repressive domain was similar, showing lower values than observed in Active domains (Figure S4B). In summary, our contact data and clustering analysis show that physical three-dimensional chromatin domains are characterized by highly specific epigenetic properties (Figures 3F, S4C, and S4D).

Domain Borders Are Defined by Insulator Binding Sites

A systematic screen of chromatin profiles showed that several factors are associated with the borders of physical domains (based on Kolmogorov-Smirnov analysis; see Table S2). The list includes the insulator proteins CP190, CTCF, and Beaf-32, the mitotic spindle protein Chromator (Rath et al., 2004), DNase hypersensitivity, and the active histone mark H3K4me3, all of which show a striking enrichment at domain borders compared to background regions (Figure 4A) or compared to control promoter regions (Figure S5A). Conversely, analysis of the distribution of the distance-scaling factors at peaks of CP190 and Chromator (Figure 4B) showed that physical boundaries are globally observed at sites bound by these proteins. Boundary behavior was significant but somewhat weaker at peaks of Beaf-32, H3K4me3, and CTCF and was minimal at peaks of Su(Hw), suggesting that domain demarcation is observed in a smaller fraction of these binding sites. Hierarchical clustering of domain borders according to their epigenetic makeup showed that multiple recurrent combinations of insulators and active marks are observed at domain borders (Figures S5B–S5D). These results are compatible with the idea that domain demarcation may be actively defined by insulator binding sites. Alternatively, a proportion of the boundaries may be indirectly determined by transcriptional activity next to a silent domain. To distinguish between these two hypotheses, we plotted the distribution of each of the marks on borders demarcating specific combinations of epigenetic domains (Figure 4C). Remarkably, we discovered that combinations of CP190, Beaf-32, CTCF, and/or Chromator are distributed symmetrically around the inferred physical boundary point with possible preferences of Beaf-32 for borders of Active domains (appearing 10 times more often than expected) and CTCF for borders of PcG domains (enriched 11-fold) (Figure 4C). Strong enrichment of H3K4me3 marks is observed in borders involving Active domains, but in these cases the H3K4me3 peak is located toward the Active domain, on average peaking 500 bp from the nearby CP190 site. Moreover, demarcation of Null and PcG domains showed strong insulator presence but weak or no H3K4me3 enrichment. Therefore, we conclude that physical boundaries of chromatin domains are determined by insulator proteins that may or may not be flanked by transcriptionally active sites. The data suggest that although a rich repertoire of sequence-specific insulator binding sites (CTCF, Su(Hw)) are observed consistently in heterogeneous nuclei populations and across developmental stages, only a fraction of them, and specifically the fraction that is cobound by CP190 and/or Chromator, serve as de facto domain boundaries at the majority of the nuclei in the developmental stage assayed here. It may be hypothesized that insulator sites not demarcating physical domains may serve as scaffolds for the formation of borders in other conditions or smaller fractions of the population.

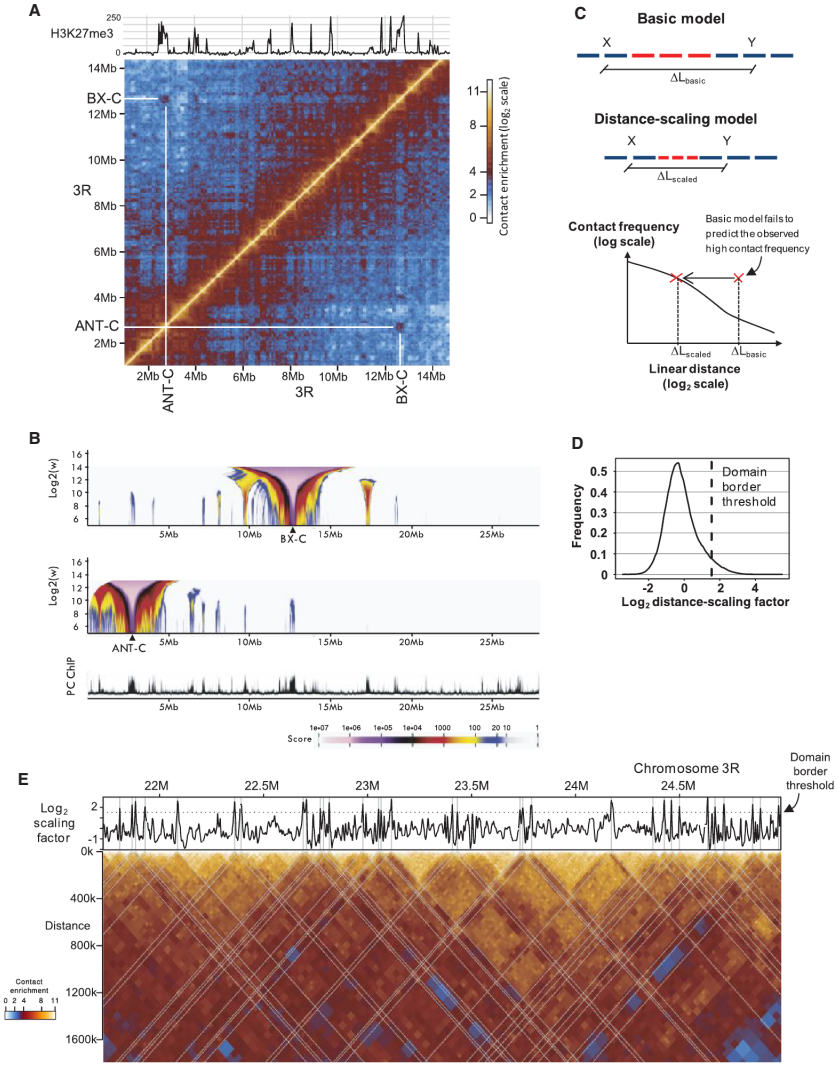

Figure 2. Virtual 4C Profiles Validate the Hi-C Genome-wide Map

(A) A normalized contact map for a 14 Mb region in chromosome 3R spanning the ANT-C and BX-C loci. A linear H3K27me3 ChIP-seq profile (Kharchenko et al., 2011) spanning the same region is shown above. The 10 Mb contact between the ANT-C and BX-C Hox loci is indicated.

(B) Virtual 4C domainograms generated from the Hi-C dataset, using the BX-C (top panel) or ANT-C (middle panel) locus as bait, and assessing interactions with the remainder of chromosome 3R. The x axis denotes the genomic position of the assessed interacting region; the y axis is the size of the window used to assess

Repressed Physical Domains Are Folded Distinctively and Interact Hierarchically

In addition to the identification of physical domains and boundaries, the rescaling of genomic distances according to inferred distance scaling (Figure 2C) gave rise to a model with improved fit between the observed and expected contact intensities (Figures 5A and 5B). Analysis of residual contacts showed that, despite this improvement, contacts between elements within many domains remained stronger than expected by the model (Figure 5C). Systematic estimation of the intradomain contact intensity as a function of the genomic distance (Figure 5D) showed that domains within the three repressive epigenetic classes (Null, PcG, and HP1/Centromere) show a distinctive decay exponent (−0.7), which is significantly different from the exponent observed for Active domains (−0.85). The different exponents suggest that repressive physical domains are governed by an intrinsic folding regime that is fundamentally different from the regime controlling active chromatin regions. This regime may be characterized by different chromatin compositions, such as varying prevalence of linker histone H1, which is relatively depleted from Active chromatin (Figure 3E), and can mediate different means of nucleosome packaging (reviewed in Robinson and Rhodes, 2006; Weintraub, 1984). Importantly, the distinct folding patterns of repressive domains form hierarchical building blocks that define higher-order chromosomal structure. As exemplified in Figure 5E and summarized statistically in Figures 5F and 5G (see Figure S5E for the control), we observed relatively uniform contact intensities between pairs of elements within a fixed pair of repressive domains even when the genomic distance between these chromosomal elements varies by more than 100 kb. According to these observations, repressive domains promote a hierarchical chromosomal organization by folding genomic regions into physical modules, and by facilitating longer-range contact between whole domains rather than between individual elements within the domains.

Clustering of the Domain Contact Map Reveals Active and Inactive Genomic Fractions

The rescaling of genomic distances and the identification of physical boundaries and domains gave rise to a model explaining the distribution of contact intensities around the Hi-C matrix diagonal. However, these essentially short-range effects cannot account for the numerous cases of longer-range interactions between domains that we observed in the Hi-C data. A simple model that could theoretically explain multiple long-range contacts is based on global physical clustering of active and inactive chromosomal elements (Lieberman-Aiden et al., 2009; Simonis et al., 2006). Such clustering may result in preferential pairings between active or between inactive domains. To test whether this type of organization is indeed present in the *Drosophila* Hi-C data, we generated a coarse-grained, domain-level contact matrix for each chromosome arm and clustered it into two. We note that this model technically corrects for experimental biases, which are variable in different physical domains, and takes into account scaled genomic distances and the physical domain structure (Figures 6A, 6B, and S6A–S6C). The resulting clusters were remarkably well defined at the epigenetic level, where in all chromosome arms, one cluster included the vast majority (93%–98%) of the Active class domains (Figures 6C and S6D). This result indicated that within each chromosome arm, domains are indeed organized into an active and an inactive higher-order cluster. Contact frequencies decayed similarly as a function of genomic separation between domains within the active or inactive clusters or between domains from the two clusters (Figure 6D). However, interchromosomal contacts were stronger for associations between domains in the active clusters of two different chromosomes (Figure 6E). In contrast, weaker interchromosomal intensities of interactions were observed between pairs of inactive cluster domains or mixed pairs of active and inactive cluster domains. This was also supported by analysis of interchromosomal contact enrichment between the predefined epigenetic classes (Figure S6E). Taken together, contacts between inactive cluster domains are generally confined to their chromosomal territory and show lower probability for interchromosomal contacts. On the other hand, active cluster domains can reach out of their chromosomal territories to form interchromosomal contacts with other active cluster domains at low specificity.

Functionally Specific Long-Range Contacts Extend beyond Global Folding Principles

We summarized all observations described above in the hierarchical domain model for chromosomal contacts (Figure 7A). The model uses scaled genomic distances, demarcation of genomic regions into physical domains, and global partitioning of chromosomes into active and inactive clusters to quantitatively predict contact probabilities on a genomic scale. We then tested the extent by which this model captures the large-scale structure of the Hi-C matrix, finding that, although model predictions indeed recapitulate much of the plaid-like structure of the matrix, many residual hotspots of contacts are still left unaccounted for (Figure 7B). Overall, we identified 268 putative remote contacts with at least 2.8-fold ($\log_2 > 1.5$) enrichment in contact intensity over the rich background model (Table S3). Classification of contacts according to epigenetic class (Figures S7A and S7B) identified, for example, 30 pairs of long-range contacts between Polycomb domains (compared to 6 expected assuming random pairing of genomic elements; see examples

interaction with the bait (logarithmic scale). The color denotes the score for the significance of the interaction. Bottom panel is a Polycomb ChIP profile (Schuettengruber et al., 2009), demonstrating that the majority of long-range interactions with the two Hox clusters are for PcG target genes.
(C) Schematic illustration of genomic distance scaling.
(D) Distribution of inferred distance-scaling factors. The threshold used for demarcating physical domains is indicated.
(E) Normalized map of contact frequencies in an ~3 Mb region of chromosome 3R, represented as for Figure 2A, but rotated by 45° so that the y axis denotes genomic separation of the interacting elements. The distance-scaling factors for the region are depicted on top, and physical domain boundaries are indicated by a white grid.
See also Figure S2.

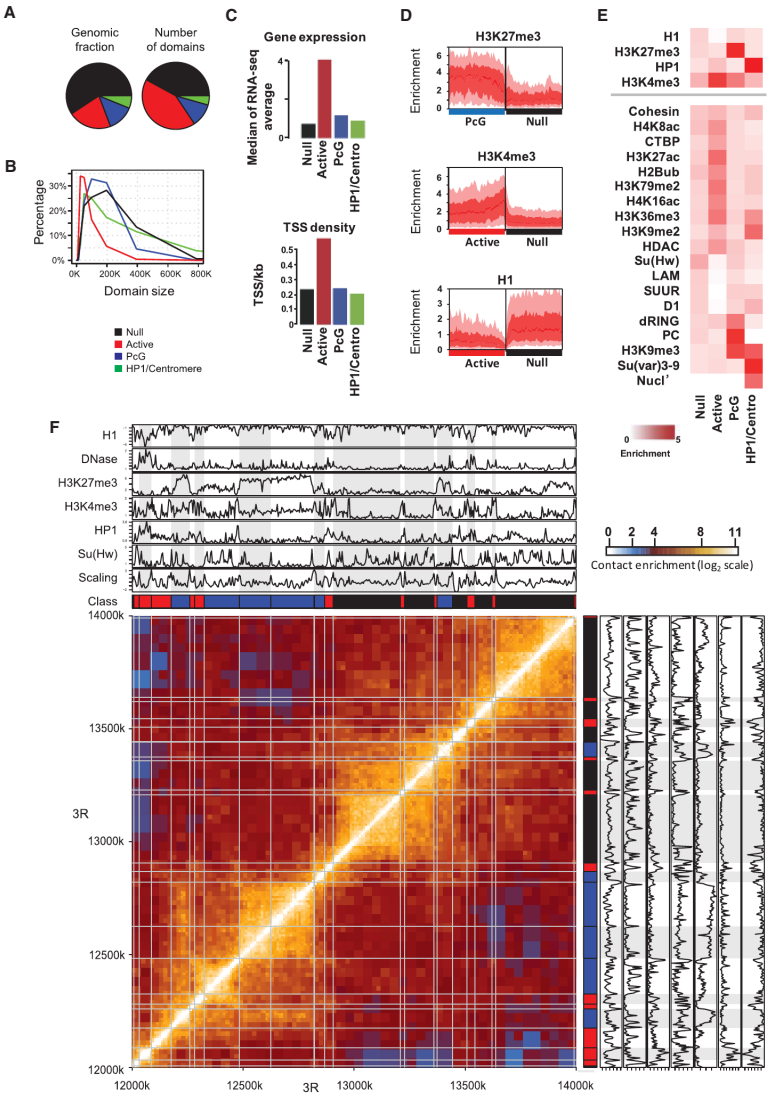

Figure 3. Physical Domains Reflect Epigenetic Domains

(A) Pie charts depicting the numbers of physical domains assigned to each epigenomic class (right) and their respective coverage of the nonrepetitive *Drosophila* genome (left).

in Figures S7C–S7E). Other striking examples of specific long-range contacts are shown in Figures 7C and S7F–S7H. The *hbs* (*hibris*) and *sns (sticks and stones)* genes code for two paralogous immunoglobulin superfamily members from genomic regions that lack traces of segmental duplications and are therefore robustly characterized by Hi-C. Remarkably, we observed very high colocalization (32-fold enrichment in Hi-C; 48% coassociation by DNA FISH, compared with 7% for the control coassociation between *hbs* and *synj*; $p = 1 \times 10^{-21}$, two-tailed Fisher's exact test; Figure 7C) of the two physical domains (measuring 100 kb and 150 kb) containing the two genes and spanning a distance of 6 Mb. This high level of colocalization was also observed in an independent, lower-resolution experiment (Figure S7H). *sns* and *hbs* are both highly expressed in fusion-competent myoblasts but are largely repressed in most other embryonic nuclei, including in the embryonic head where FISH experiments were performed. Genetic and functional studies suggested that the two proteins colocalize at the plasma membrane (Artero et al., 2001) and function redundantly in myoblasts (Shelton et al., 2009), suggesting that nuclear colocalization may be involved in coregulation of the two genes. In summary, these results show that global *Drosophila* chromosomal architecture can be predicted from a few simple principles including local distance scaling, partitioning of the chromosome into physical domains, and global clustering of domains into active and inactive fractions. Beyond these generic genome-wide effects, specific contacts are also formed, highlighting such contacts as likely candidates to play a direct regulatory role.

DISCUSSION

A Genome-wide Chromosome Contact Map for Fly Embryonic Nuclei

We have developed a simplified Hi-C procedure for minimally biased profiling of chromosomal contacts on a genomic scale. Using this technique, we comprehensively and accurately characterized chromosomal architecture in *Drosophila melanogaster* embryonic nuclei. The chromosomal contact map we derived relaxes the classical trade-off between coverage and resolution in the study of chromosome structure. The data provide us with sufficient resolution to observe local contact profiles derived from 4C (Figure 2) and consistently deliver such resolution for essentially any genomic locus. The effective resolution limitations of the map depend on the features being studied. Demarcation of physical domains can be achieved within a precision of one or a few DpnII fragments (i.e., of ~1 kb), as many fragments with high expected contact probability contribute to their identification. On the other hand, detection of long-range contacts with statistical confidence greatly depends on their absolute intensity compared to the background, which decays significantly with genomic separation. For example, based on the current sequencing depth, the decay in background contact probability with genomic distance (Figure 1C), and the average DpnII restriction site density, we estimate that a contact with 4-fold enrichment over the background could be confidently detected at a resolution of ~10 kb for genomic separations of 100 kb, a resolution of ~30 kb for genomic separations of 1 Mb, and ~125 kb for interchromosomal coassociations. Regardless of these considerations, and despite the fact that the experiment assayed a large and heterogeneous set of nuclei, the derived Hi-C map reveals a clear structure and allows for multiple chromosome folding principles to be explored systematically. The implications of the *Drosophila* map are therefore far reaching, and the analysis presented here can be viewed as a baseline on which further efforts directed to understand genomic and epigenomic patterns at particular cell states or genetic backgrounds can be developed.

Domain-Based Hierarchical Chromosomal Architecture

The Hi-C map is rich in local and global structure, describing contact frequencies that vary within five orders of magnitude. We wished to explain the distributions of contact frequencies in the map using quantitative models based on the simplest principles and justified any progressive increases in model complexity by proven discrepancies between the data and a simpler version of the model. One of the most remarkable patterns we observed in the map was the partitioning of chromosomes into physical domains, which showed up in the matrix as diagonal submatrices with high contact intensities (e.g., Figures 2E and 3F). We used a quantitative probabilistic model to show that contacts inside these domains are governed by a distinct regime that cannot be attributed to denser contacts or more compact chromosomal structure alone (Figures 5A–5C). Further analysis showed that physical domains form the backbones of a hierarchical chromosome structure (Figures 5E–5G), as the contact intensities between genomic elements are mostly determined by the identities of the domains containing them, rather than the element's location within the domain. Previous lower-resolution exploration of human chromosome architecture identified a global power law decay of contact frequency with genomic separation and used this to propose a fractal globule model of chromosome folding (Lieberman-Aiden et al., 2009).

(B) Frequency plot showing the distribution of the physical domain sizes for each epigenomic class.
(C) Transcriptional activity (top) and gene density (bottom) distributions for physical domains assigned to each epigenomic class.
(D) Distribution of H3K27me3, H3K4me3, and H1 on borders delimiting domains of specific classes (color-coded): light pink denotes 10^{th}–25^{th} and 75^{th}–90^{th} percentiles, red indicates 25^{th}–75^{th} percentiles, dark red line indicates the median.
(E) Heat map showing enrichment of specific epigenetic marks within each of the four identified classes of physical domains. The epigenetic marks used for supervised clustering are depicted at the top.
(F) Hi-C normalized contact map for an ~2 Mb region of chromosome 3R, shown alongside the profiles for selected epigenetic marks and a color-coded indication of physical domains and their epigenetic classes. A white grid on the contact map denotes domain boundaries. Color code of classes: black—Null; red—Active; blue—PcG; green—HP1/Centromere.
See also Figures S3 and S4 and Tables S1 and S2.

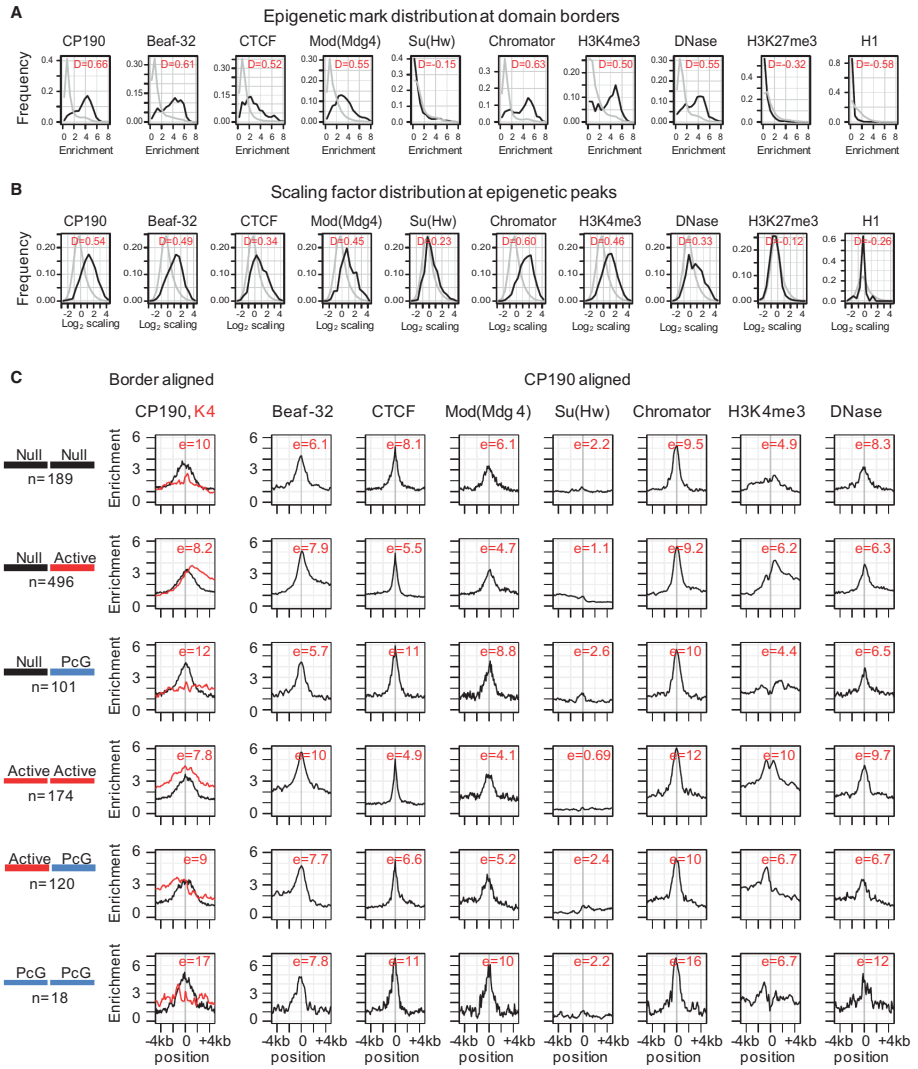

Figure 4. Domain Demarcation at Insulator Binding Sites

(A) Enrichment of epigenetic marks at domain borders. Frequency plots comparing distributions of ChIP-seq or ChIP-chip enrichment values of various proteins and histone modifications at physical domain borders (black curves) and in the whole genome (gray curves). Marks with a right shift of the black curve compared to the gray thus have enriched binding at physical domain boundaries. Kolmogorov-Smirnov D-statistics are indicated on each panel, reflecting in all cases highly significant behaviors ($p << 10^{-10}$). Negative controls (H1, H3K27me3) are shown to the right.

Although we observe a roughly similar global decay curve for *Drosophila* chromatin (Figure 1C), higher-resolution analysis of contact decays within the context of physical domains challenges this model and suggests that, in scales of 10–100 kb, the predominant factor affecting chromosome folding is the modular organization. This promotes hierarchical chromosomal organization as an attractive paradigm to facilitate functional epigenetic organization but leaves open questions about the scales at which it may be observed in different genomes that vary significantly in size and gene density.

Physical Domains Are Epigenomic Domains

Remarkably, the physical domains we inferred from the Hi-C contact map were compatible with numerous linear epigenetic profiles describing enrichment for histone modification or DNA-binding factors (Figure 3D). Thus the physical domains, which are key fundamental units of chromosome folding, are reflected and possibly caused by their underlying epigenetic marks. Large silent chromosomal regions that are either enriched with repressive histone marks (H3K27me3 or HP1/H3K9me2) or void of any detectable epigenetic enrichment were shown to form modular chromosomal entities, which are interspersed with small domains associated with active chromosomal marks. By analyzing the epigenomic marks at the borders of physical domains (Figure 4), we observed that a transition in transcriptional activity (as indicated by peaks of H3K4me3) is sometimes sufficient to disturb the compaction of flanking repressive chromatin domains. This may result in the formation of ''punctuated'' repressed domains, with active genes forming ''passive'' physical boundaries. However, in most cases, we find that insulator proteins, particularly CP190 and Chromator, sharply demarcate the borders of physical domains. This is especially apparent at borders marked by both CP190 and H3K4me3, as CP190 binds precisely at the physical domain boundary, with the H3K4me3 peak shifted ~500 bp toward the Active domain (Figure 4C). Interestingly, a recent study suggested that binding of the ''accessory'' insulator protein CP190 is required for a functional insulator (Wood et al., 2011). In agreement with this, we find that CP190 correlates most strongly with physical boundary domains, whereas many regions bound by the DNA sequence-specific binding insulator proteins CTCF and Su(Hw) are not linked to physical domain boundaries (Figure 4B). Chromator emerged from our analyses as another major factor organizing physical domains. Although little is known about the function of the mitotic spindle protein during interphase, Chromator has been shown to be necessary for the maintenance of polytene chromosome structure (Rath et al., 2006). Our findings appear to extend the structural function of Chromator to diploid embryonic nuclei. By providing an architectural context to epigenomic chromatin domains, the Hi-C map thus provides a reference epigenomic model, directing future efforts for analyses of the correlations between hundreds of measured linear epigenomic profiles (Ernst et al., 2011; Filion et al., 2010; Kharchenko et al., 2011).

Global Chromosomal Architecture Sets the Stage for Specific Contacts

Chromosomes clearly fold in a complicated, heterogeneous regime (e.g., Figure 2A). In order to make any reasoned claims about the significance of previously reported individual cases of long-range chromatin interactions, it is important to first understand the basic principles of what defines ''standard'' folding of a chromosome fiber. This Hi-C dataset puts us in an unprecedented position to formulate and test hypotheses on chromatin folding with progressively more complex quantitative models (Figure 7A). First, we accounted for heterogeneity in contact density (Figure 2C), facilitating identification of physical chromatin modules and their hierarchical pattern of folding. Next, we were able to group physical domains into two clusters (annotated postfactum as active or inactive) based on their intrachromosomal contacts and to generally describe interdomain contacts as those within or between clusters (Figure 6A). This supported and extended previous findings on the relationship between transcriptional activity and position within chromosome territories (Chambeyron and Bickmore, 2004; Würtele and Chartrand, 2006). Although the combined model explains much of the chromosome folding behavior, specific long-range chromatin interactions were still apparent. One group of functional long-range interactions that has already been investigated and is clearly visible in the Hi-C map (Figures S7C–S7E) associates PcG-regulated genes that co-occupy Polycomb bodies (Bantignies et al., 2011; Grimaud et al., 2006; Tolhuis et al., 2011).

In summary, this Hi-C study has provided a fundamental chromatin interaction map framework, providing the basis for mathematical models to assess the link between chromosome structure and function. The characterization of hierarchically folded discrete physical modules, which may be epigenetically defined, forms a hitherto unappreciated base from which more complicated chromosome topologies can arise. We posit that this and future Hi-C datasets, combined with specific perturbation experiments, will inform more sophisticated mathematical models of chromosome folding, forming a foundation for new important insights into what shapes nuclear structure and how this in turn affects genome function.

(B) Distributions of distance-scaling factors at peaks of the same epigenetic marks. The genome-wide distribution of distance-scaling factors is denoted by gray curves. Thus factors with a right shift of the black curve compared to the gray have more binding sites that are likely to correspond to physical domain boundaries. Kolmogorov-Smirnov D-statistics are indicated on each panel, again reflecting highly significant behavior ($p < < 10^{-10}$).

(C) Shown are spatial distributions of medians of CP190 (black) and H3K4me3 (red) levels aligned to physical domain borders (left) and medians of several additional marks (black) aligned to CP190 peaks within borders (right). Analysis was performed separately for different combinations of adjacent physical domains from specific epigenetic classes (colored rectangles). The fold enrichment over the background binding level (e – fraction of borders with a binding site divided by the background frequency of the binding site) is indicated for CP190 (left panel) and each of the denoted factors (right panel) at each border type.

See also Figure S5.