Published online 28 January 2012 Nucleic Acids Research, 2012, Vol. 40, No. 10 4385–4395
doi:10.1093/nar/gks045

Chromatin domain boundary element search tool for *Drosophila*

Arumugam Srinivasan and Rakesh K. Mishra*

Centre for Cellular and Molecular Biology, Council for Scientific and Industrial Research (CSIR),
Uppal Road, Hyderabad, 500007, India

Received December 15, 2010; Revised January 6, 2012; Accepted January 10, 2012

ABSTRACT

Chromatin domain boundary elements prevent inappropriate interaction between distant or closely spaced regulatory elements and restrict enhancers and silencers to correct target promoters. In spite of having such a general role and expected frequent occurrence genome wide, there is no DNA sequence analysis based tool to identify boundary elements. Here, we report chromatin domain Boundary Element Search Tool (cdBEST), to identify boundary elements. cdBEST uses known recognition sequences of boundary interacting proteins and looks for 'motif clusters'. Using cdBEST, we identified boundary sequences across 12 *Drosophila* species. Of the 4576 boundary sequences identified in *Drosophila melanogaster* genome, >170 sequences are repetitive in nature and have sequence homology to transposable elements. Analysis of such sequences across 12 *Drosophila* genomes showed that the occurrence of repetitive sequences in the context of boundaries is a common feature of drosophilids. We use a variety of genome organization criteria and also experimental test on a subset of the cdBEST boundaries in an enhancer-blocking assay and show that 80% of them indeed function as boundaries *in vivo*. These observations highlight the role of cdBEST in better understanding of chromatin domain boundaries in *Drosophila* and setting the stage for comparative analysis of boundaries across closely related species.

INTRODUCTION

Eukaryotic genomes contain a large number and variety of regulatory elements that control the cell type and context dependent expression patterns of genes. Much of the genome that does not code for proteins, contains these regulatory elements often in the close proximity of the target gene but frequently also at far away locations. Specific and appropriate interaction of regulatory elements governs the complex and regulated expression of genes. However, much of the mechanism involved in this process remains to be understood and, in particular, it is far from clear how the specificity of interactions among regulatory elements is achieved. It is known that given access by bringing together in transgenic context or by chromosomal rearrangements, enhancers as well as silencers can act on almost any promoter. It is also known that expression of a transgene construct in different independent transgenic lines often depends on the site of insertion in the genome, a phenomenon referred to as position effect, due to the influence of local regulatory elements on the transgene. These observations suggest that in order to restrict infidel and distantly located elements to appropriate target promoters the genome is subdivided and highly organized by means of functionally independent 'chromatin domains' (1).

Critical to this chromatin domain model are chromatin domain boundary elements, the DNA sequences that define the borders of chromatin domains and capable of blocking enhancer-promoter interactions. Chromatin domain boundary elements were first identified in *Drosophila melanogaster*, as specialized chromatin structures that border the two heat shock genes in 87A7 heat shock locus. The DNA sequences that are responsible for the bordering effect were named as scs and scs′ (specialized chromatin structures), and they became the first molecularly defined boundary elements (2). Thereafter, several of boundary elements have been identified in *D. melanogaster* at various genomic locations, the notable ones are *gypsy*, *Mcp*, *Fab-7* and *Fab-8* (3–8). The *gypsy* boundary is part of naturally occurring *gypsy* retrotransposon and the *Mcp*, *Fab-7* and *Fab-8* boundary elements are present within the bithorax complex [BX-C] of *Drosophila* and are required for domain specific expression of *Abd-B*

*To whom correspondence should be addressed. Tel: +91 40 27192658; Fax: +91 40 27160591; Email: mishra@ccmb.res.in gene. Apart from *Drosophila,* the boundary elements were also identified in variety of organisms ranging from yeast to human (9–11).

Experimental identification of boundary elements in *Drosophila* involves the two main functional assays. First, the enhancer blocker assay where the reporter gene is driven by a minimal promoter and a strong enhancer and when a boundary element is placed in-between the enhancer and promoter, the expression level of reporter gene is eliminated or reduced (12). The second assay is known as insulation from position effect where a reporter gene is flanked by boundary elements that to insulate from chromosomal position effects leading to uniform levels of expression in independent transgenic lines (13). Though, these assays have been successfully used for various boundary elements, they have an inherent disadvantage of having to produce transgenic animal lines and as a result, these assays cannot be applied for genome-wide screening. As an alternative to testing in transgenic flies, recently boundary elements have been assayed in *Drosophila* cells (14,15). More recently, a cell culture based barrier assay has been used where ability of a test DNA to prevent spread of repressive chromatin has been assayed (16). These testing methods also involve making of constructs and establishing cell lines, and, therefore, limit their applicability at genome level analysis. Based on the functional assays used, boundary elements are also referred as enhancer-blocking insulator or barrier element.

Though it is accepted that boundary elements divide the genome into domains, the mechanism by which they establish independent functional domains remains unclear and various models that have been proposed so far, suggest that they may use more than one mechanism (10,17). All the models agree that boundary elements exert their function through interaction with various DNA-binding proteins and associated factors. To date, in *Drosophila* six DNA-binding proteins are shown to directly interact with boundary elements. These include BEAF, Zw5, GAGA Factor (GAF), Su(Hw), dCTCF and recently identified Elba factor (18–27). A boundary element may require one or more of these DNA-binding proteins to function as a boundary *in vivo* (21,26,27).

Recent studies have reported the genome-wide binding profiles of various boundary binding proteins using ChIP-on-chip approach (28–32). Although these binding profiles can precisely map the *in vivo* binding sites of individual proteins, it is not clear how far these individual protein binding profiles can define functional boundaries in the genome since they define only the binding sites of proteins rather than defining complete boundary sequence. The experimentally identified boundary size varies from 431 bp to ~2.5 kb (Supplementary Table S1). Moreover, these proteins are also involved in other nuclear functions such as transcriptional activation or silencing apart from their boundary function (28,33). *In vivo* binding site analysis of a boundary interacting factor does not necessarily indicate an associated boundary function.

Here, we report a bioinformatics tool, chromatin domain Boundary Element Search Tool (cdBEST) for identification of potential boundary sequences in *Drosophila*. Using cdBEST, we identified 93109 boundaries across 12 *Drosophila* species. Our approach identified 4576 boundaries for *D. melanogaster,* including several repetitive boundaries. We also analysed these boundaries for their context in terms of flanking genes and experimentally tested 19 cdBEST boundaries in *Drosophila* S2 cells for enhancer-blocking activity and found that great majority of these cdBEST boundaries indeed function as boundaries *in vivo*.

MATERIALS AND METHODS

Sequence data

The known boundary sequences were retrieved from National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/) by following the information provided in published literatures (Supplementary Table S1). We used *D. melanogaster* genome release version 5.2 of NCBI for boundary prediction. The following are the accession numbers for various chromosomal arms of *D. melanogaster* that were used in this study: NT_033777.2 (3R), NT_037436.3 (3L), NT_033778.3 (2R), NT_033779.4 (2L), NC_004354.3 (X) and NC_004353.3 (4).

cdBEST

Calculation of motif frequencies and fold enrichment values. To calculate genome level motif frequencies (f_g), we searched for the occurrences of individual boundary motifs (m_i–m_n) in the *Drosophila* genome and divided the total number of occurrences with genome size. Similarly a boundary level frequency (f_b) was also calculated by considering boundary sequence alone for each boundary motif (see Supplementary Table S2 for details). The ratio between genome level frequency and boundary level frequency for a given motif m_i is defined as fold enrichment value (FEV) for that particular motif.

$$\text{FEV of motif } m_i \ (FEVm_i) = f_b(m_i)/f_g(m_i)$$

Boundary score. We calculated the boundary score by taking the overall summation of motif occurrences (Om_{i-n}) and their multiplication with FEV ($FEVm_{i-n}$).

$$\text{Boundary score} = \sum_{i=1}^{n} (\mathrm{Om_i} \times \mathrm{FEVm_i})$$

Implementation and availability of cdBEST. The cdBEST algorithm is implemented in Perl as two variants, the cdBEST basic and advanced. The basic version is a command line script requires Perl alone, can be used in Linux and Mac computers directly without any additional requirements. cdBEST_basic uses FASTA formatted sequence files for boundary search and produces text output files. The advanced version has a simple GUI (graphical user interface), where user can choose various parameters for searching boundaries. In addition to text files, this version produces an image output showing boundaries and gene annotations together with the scale for easy correlation. cdBEST_advanced requires Perl-Tk module for generating GUI and BioPerl modules (34) to draw image output files. In addition cdBEST search results can be uploaded to various genome browsers such as FlyBase GBrowse and UCSC genome Browser as custom tracks to view along with any other feature. The cdBEST source file for different OS platforms with a readme file is available for download at http://www.ccmb.res.in/rakeshmishra/cdBEST.html.

Drosophila S2 cell based enhancer blocker assay

To make NPG vector (neomycin PE enhancer GFP) construct, we used pGreen H-Pelican vector as starting material (35). We assembled the final NPG construct (Figure 4A) in three steps. (i) Vector backbone: by DraIII digestion, we removed white gene and religated vector. (ii) Neomycin-hsp70 promoter: we assembled the neomycin gene (amplified from pEGFP) and hsp70 promoter in pBluescript vector. Using ApaI/KpnI double digestion, this cassette was removed and cloned in vector backbone. (iii) PE enhancer: we amplified PE enhancer (*twist* gene's Proximal Element enhancer) from the *Drosophila* genomic DNA using specific primers carrying KpnI sites, and cloned in vector backbone (36). In the final NPG construct the test fragments can be cloned in NotI and BamHI sites. We PCR amplified various test fragments (known boundaries and predicted boundary elements) and cloned in pGEM-T Easy vector. Using NotI sites that are present in pGEM-T Easy vector, we removed the inserts and cloned to NPG vector.

We cultured *Drosophila* S2 cells in Schneider's Insect Medium (supplemented with 10% serum). We used one microgram of Qiagen column purified DNA to transfect 1 ml of cells (1×10^6 cells/ml). Effectene transfection reagent kit (Qiagen) was used for transfections. After 36 h of transfection we added G418 to cells with the final concentration 1 mg/ml. Once in every 4 days, we replaced the old media with fresh media (containing G418). After 6 weeks of culturing, we assayed the cells for enhancer-blocking activity. Flow cytometry analyses were done on MoFlo cell sorter using 0.5 ml of cells and 20 000 events were scored.

Polycomb and H3K27me3 data analysis

We downloaded the Polycomb and H3 me3K27 ChIP-chip data from ArrayExpress database (accession code E-MEXP-535) (37). Using Perl scripts and FlyBase coordinate converter, we converted release four coordinates to release five coordinates and extracted the data for regions of our interest and written the data in GFF format. By using BioPerl modules we generated image files in which gene information and predicted boundaries are plotted along with ChIP-chip data and manually counted the boundaries that lie very close to borders of strong PcG sites.

FlyAtlas gene expression profile comparison

We downloaded FlyAtlas data set containing the expression data for 17 adult tissues, eight larval tissues and S2 cell line (38). This data set has the expression profiles for 18 880 probes which represents *Drosophila* 18 500 transcripts covering around 13 500 genes. As data is referenced based on Probe Set IDs, we added gene names, start, stop positions and FlyBase IDs to each probe by following an annotation file (*Drosophila*_2.na28.annot.csv). Using a Perl script and chromosome specific gene lists (prepared from release 5.2 GenBank files) we separated out the data for each chromosome and wrote it as separate files. Then we mixed the predicted boundaries to this chromosome specific data and sorted the data using start positions. Using another Perl script, we extracted the gene pairs that flank predicted boundaries and compared their expression profiles. Here we called a gene pair as 'differentially expressed' if their expression change direction is in negative correlation (i.e. the change direction is 'Up' verses 'Down' or 'Down' verses 'Up') for one or more tissues. In order to make sure that the 'Up' change direction of a gene reflects its presence, we used Affymetrix call Presence/Absence values. We considered Affymetrix Presence/Absence call value 0 or 1 as a mark for absence, value 3 or 4 as mark presence of transcripts.

RESULTS

cdBEST

The experimentally identified boundary elements in *Drosophila* share common functional properties; however they do not posses any significant sequence similarity. In general, *cis*-regulatory elements are often enriched by small sequence motifs. Boundaries elements are no exception to this phenomenon and they do have several such motifs that serve as interacting sites for boundary interacting proteins (Table 1). With the objective to identify new boundary elements, we analysed the distribution of these motifs in the euchromatic portion of the *D. melanogaster* genome. The DNA motifs we analysed include BEAF (19), Zw5 (20), GAF (21), Su(Hw) (23,24), Elba (27), CTCF (31,32,39) and *Fab-7* Motif (F7M) (Mishra,R.K., unpublished data). The distribution pattern shows that the smaller motifs, BEAF, Zw5 and GAF are more randomly distributed in the genome, but occur as clusters in boundary regions (Table 1). The larger motifs like Su(Hw), CTCF are non-randomly distributed in the genome and highly enriched in the boundary regions.

Based on motif clustering and enrichment, we divided the boundaries into five types: *Fab-7* type, *Fab-8* type, *SCS* type, *SCS′-BE28* type and *gypsy* type (Table 2). For each boundary type, we evaluated for criteria like, the total number of motifs, predominant motif(s) and the average gap between motifs. Taking inputs from these analyses we built a boundary element search tool, cdBEST that looks for boundary type specific motif clusters under a defined set of constraints (Table 2). We included a scoring method in this tool to eliminate false predictions. The boundary score for a given sequence is calculated by the overall summation of motif occurrences multiplied by their respective FEV. In general FEV of motifs are calculated by comparing their occurrence frequencies in positive verses background data sets (40–42). We used known boundary regions as positive data set and *Drosophila* genome sequence as a background data. The calculated FEV (Table 1) are incorporated in the tool to derive the boundary score. We set boundary type specific minimum required scores in cdBEST (Table 2).

To test the tool for its efficiency, we used set of known boundary sequences (Supplementary Table S1), cdBEST picked up 10 of the 11 boundary sequences as hits with varying scores (Supplementary Table S3). *SF1* was the only boundary that failed to a give hit because of its poor motif content. The *gypsy* boundary achieved a highest score of 1722.6, while the SCS boundary received the lowest score 43.97. cdBEST did not yield even a single hit when regulatory element sequences such as enhancers and polycomb response elements (PREs) were used as input (Supplementary Table S4) indicating the accuracy of the tool. A test run on a sequence region that covers *Drosophila* Bithorax complex (BX-C) identifies 12 boundaries including previously known *Mcp*, *Fab-6*, *Fab-7* and *Fab-8* boundaries (Figure 1 and Supplementary File S1).

Whole genome analysis for boundaries in *D. melanogaster*

Drosophila melanogaster genome release 5.2 was used as input for boundary search. Each chromosome was separately analysed using 750 bp as set window size and 10 bp as window slide. Under these conditions, we retrieved 4576 boundaries in the whole genome (Table 3 and Supplementary File S2). cdBEST correctly identified the reported boundaries, even-skipped, TER94, Abd-Bm and myoglianin-eyeless (ME), which were not included in our positive data set (43–46). The average domain size deciphered by predicted boundaries varies from 19 to 31 kb for various chromosome arms. Density of predicted boundary was greater on the X chromosome despite having a low gene density and moderate size.

Boundaries with repetitive occurrence are associated with transposable elements

To find multicopy or repetitive boundaries in the *D. melanogaster* genome, we carried out BLAST sequence alignments among the predicted boundary sequences. We used an identity of >90% over a stretch of 100-bp sequence to call repetitive boundaries in the genome. Among the 4576 predicted boundaries, we retrieved 55 groups of repetitive boundaries containing 239 individual elements. The number of boundaries within a group ranges from 2 to 39 (Table 4 and Supplementary Table S5). We also found the known gypsy boundary in the list with two copies. This led to the assumption that many of these multicopy boundaries may be associated with transposable elements. To test this, we compared these repetitive boundary sequences with known transposable element sequences from databases, FlyBase and Repbase (47,48). Of the 239 multicopy boundaries that are found in the *D. melanogaster* genome, 173 showed significant sequence similarity (>90% identity over 100-bp sequence) with transposable elements. *Drosophila melanogaster* has 96 known families of transposable elements that covers the ~5% of the euchromatic part of the genome (49,50). Out of 173 boundaries of repetitive nature that are identified by cdBEST 110 boundaries maps to nine of these families (Table 4 and Supplementary Table S5), indicating that only a small subset of transposable elements have boundary activity.

Application of cdBEST in other *Drosophila* species

Encouraged by the performance of cdBEST in *D. melanogaster* genome, we wanted to extent cdBEST prediction to 11 sequenced non-melanogaster species (50). Considering the evolutionary closeness of these species we expected that the boundary elements to be conserved and cdBEST might be able to pickup boundaries across these species. First, we asked whether cdBEST can predict prominent boundaries, such as Fab-7 and Fab-8 in these species. For this, we used region(s) that covers Bithorax complex to predict boundaries and found hits that are orthologous to these two boundaries in all *Drosophila* species except *grimshawi*, where Fab-8 alone was predicted (Supplementary Figure S1). As cdBEST correctly recognizes these two test boundaries in 10 out of the 11 species, we applied cdBEST for genome-wide boundary prediction. We downloaded the assembled genome sequence of these 11 species from FlyBase (FB2011_05 Release) and screened for contigs that are >200 kb (47). Each genomic chromosome/contig/scaffold was subjected to boundary search and the total number of boundaries was counted using an automated script. cdBEST identified 88533 boundaries for these 11 non-melanogaster *Drosophila* species. The entire prediction data can be downloaded from our website (http://www.ccmb.res.in/rakeshmishra/cdBEST.html). Some species show very high number of boundaries when compared to other species (Figure 2 and Supplementary Table S6). *Drosophila mojavensis* has the highest number of predicted boundaries (15781) among all 12 species in-spite of not having the largest genome size. We also searched for the occurrence of repetitive boundaries in each of these species and extended the transposable elements verses repetitive boundaries comparison. In the end, we found large number of boundaries that are associated with transposable elements across these species (Supplementary Table S7). As shown in Figure 2, the percentage of repetitive boundaries closely follows the repeat contents of these 12 *Drosophila* genomes (50). *D. ananassae* and *D. virilis* are the only two species that are having higher repeat boundary percentage than their overall repeat content. This may be because of species specific repeat sequences with boundary potential are present in these species. The highest copy-number boundary (2702 copies), is indeed a species specific repeat sequence of *D. ananassae*.

Epigenomic context of boundaries identified by cdBEST

Boundary elements that mark the borders of repressive domains. During the early embryonic development, the active and inactive chromatin regions are marked by specific post-translational histone modifications in a

Table 1. Motif frequency and fold enrichment in boundary compared to whole genome

S. No.	Motif[a]	Motif sequence[b]	Boundary level frequency		Whole genome frequency[c,d]	Fold enrichment in boundaries
			Boundary	Occurrence[c]		
1	BEAF	CGATA	SCS′ and BE28	9.15	1.292	7.09
2	Zw5	GCTGMG	SCS	5.03	0.963	5.23
3	GAF	GAGAG	Fab-7	4.89	1.344	3.64
4	Su(Hw)-M1	YRYTGCATAYYY	–	–	0.022	156.55[e]
	Su(Hw)-M2	YWGCMTACTTHY	(2L-203)[f]	3.47	0.022	156.55
5	Elba	MCAATAAG	Fab-7 and Fab-8	0.99	0.069	14.21
6	CTCF-M1	MHRGRKGKCGCY	Fab-8	2.49	0.016	150.91
	CTCF-M2	YAGRKGKCGC	Fab-8	1.25	0.020	61.58
	CTCF-M3	RRCGCCMYCYRKY	Fab-8	1.25	0.008	165.67
7	*Fab-7motif*	CCAATTGG	Fab-7	1.63	0.022	73.02

[a] Motif names are defined based on the binding protein for the purpose of computer searching. M1, M2 and M3 are alternative or additional binding motifs of the protein. [b] IUPAC code. [c] Occurrence per kb. [d] Whole genome used here includes only the Euchromatic regions (X, 2L, 2R, 3L, 3R and 4) of release 4.1. [e] Assigned based on the value obtained for Su(Hw)-M2 Motif, as both have similar genomic frequencies. [f] 2L-203, 3.09, 3.28, 2L-203, X-103 and y-45.

Table 2. Five boundary types and prediction criteria for new boundaries

Boundary type	Motif cluster	Boundary mapping criteria		
		Specific feature	Motif/gap[a]	Score
1. *Fab-7*	**GAF**[6], **Elba**[1], **F7M**[2], Zw5[2]	≥2 kinds of motifs	8/90	60
2. *Fab-8*	**CTCF-M1**[2], **M2**[1], **M3**[1], GAF[2], Elba[1], BEAF[1]	≥1 CTCF motif(s)	2/90	75
3. *SCS*	**Zw5**[9], BEAF[3]	Two Zw5 motifs[b]	8/90	37
4. *SCS′*	**BEAF**[8], Zw5[2]	Six BEAF motifs[c]	6/90	43
BE28	**BEAF**[7], Zw5[2]			
5. *Gypsy*	**Su(Hw)-M1**[11], **M2**[1]	≥2 Su(HW) motifs	2/125	313
2L-203	**Su(Hw)-M1**[2], **M2**[3], CTCF[1], Zw5[1], BEAF [1]			
X-103	**Su(Hw)-M1**[3], **M2**[3]			

Motifs in bold are the predominant/experimentally tested motifs in a particular boundary type, numbers in bracket indicate their occurrences. [a] Motifs/gap combination shows the number of total motifs required and with allowed average gap. [b] Here two high affinity motifs (Zw5 motif flanked by next Zw5 motif with 13 bases as maximum allowed gap) are required. [c] Here two high affinity motifs (BEAF motif flanked by next BEAF motif with 16 bases as maximum allowed gap) are required.

tissue or cell type manner (51). The marked histone status is further inherited or maintained in subsequent generations with the help of Polycomb and trithorax group of proteins (51–53). The regions that are covered by Polycomb proteins and marked by H3K27me3 modifications are shown to form repressive domains in the genome (53,54). As Polycomb proteins can spread on chromatin over a considerable distance and repress the gene activity, boundary elements are thought to be involved in limiting the spread of repressive chromatin and define the borders repressive of domains (54–57). In this background, we analysed an available Polycomb binding data (37) derived from *Drosophila* S2 cells and asked whether cdBEST can locate boundaries that can limit the Polycomb spreading or define the borders of repressive domains. We used a data set that consist 95 strong PcG sites as repressive domains. The ninety five repressive domains yielded 190 border regions for analysis. Overall, of the 190 border regions we analysed, 108 regions have well positioned boundaries (data not shown).

Here, we show two specific regions as examples and discuss them in detail. The first region, NK homeodomain region, has a repressive domain marked by H3K27me3, containing CG31179 and C15 genes that are transcriptionally silent in S2 cells. However, the repressive domain is immediately flanked by hypomethylated regions that contain transcriptionally active genes CG7922 and CG7956. cdBEST predicts boundary 3R_591 and 3R_592 and mark the limits of the repressive domain (Figure 3A). The second region is *dco-Sox100B* region where a presumptive PRE lies between the two divergently transcribed genes (Figure 3B). PRE can spread silencing effect to long distance and one of the possible ways by which nearby genes are protected from the silencing effects could be the intervention of a boundary between the active and silent regions. cdBEST finds boundary 3R_913 near *dco* promoter. It is interesting to note that boundaries closely associated with promoters have been reported earlier (43).

Boundaries that separate domains of differentially expressed genes

As boundary elements can organize neighbouring genes into independent chromatin domains, one would expect that the 5′- and 3′-flanking genes of a boundary element may have independent expression profiles. To test this,

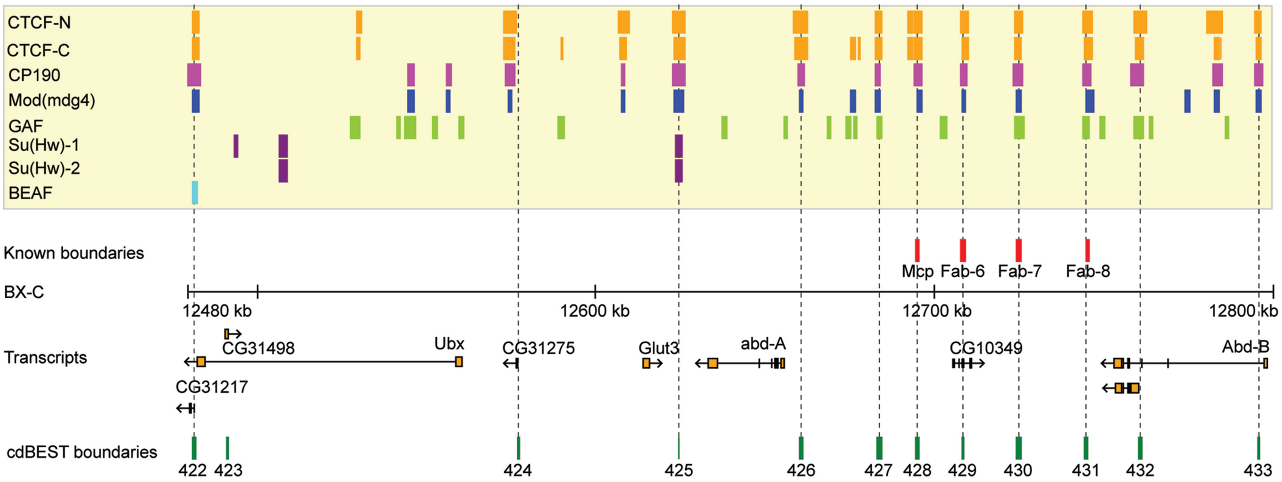

Figure 1. cdBEST analysis for the boundaries in the *Drosophila* Bithorax Complex. A 320 kb region of chromosome 3R, which consist of the BX-Complex is drawn according to scale. The upper yellow panel shows the *in vivo* binding profiles of various boundary proteins [plotted using a data from the recent study (64)]. The known boundaries were mapped and shown as red boxes. The lower panel shows annotated genes and cdBEST predicted boundaries with boundary numbers (corresponding to chr3R prediction). Dashed vertical lines show alignment of the cdBEST predictions against *in vivo* binding profiles of boundary interacting proteins.

Table 3. Whole genome analysis for boundary elements using cdBEST

Chromosome arm	Size (bp)	No. of boundaries	Boundary frequency [per 100 kb]	No. of genes	Gene density [genes/100 kb]	Average domain size[a] (kb)	Average genes per domain
2L	23011 544	784	3.41	2766	12.0	29.4	3.5
2R	21 146 708	830	3.92	3088	14.6	25.5	3.7
3L	24 543 557	793	3.23	2848	11.6	31.0	3.6
3R	27 905 053	953	3.42	3547	12.7	29.3	3.7
4	1 351 857	52	3.85	90	6.7	26.0	1.7
X	22 422 827	1164	5.19	2314	10.3	19.3	2.0
Whole genome	120 381 546	4576	3.80	14 653	12.2	26.3	3.2

[a]Domain size was calculated by dividing the chromosome size with number of boundaries.

Table 4. Transposon associated multicopy boundary elements in *D. melanogsater*

S. No.	Predicted boundary	Number of copies	Associated transposon	Predominant motif(s)
1	X_52	39	Doc	GAF Elba
2	2L_14	23	blood	BEAF
3	2R_83	8	Rt1a	CTCF GAF
4	4_2/4_3	12	GATE	BEAF CTCF
5	X_1143	7	G-element	BEAF
6	2L_768	6	Rt1b	CTCF
7	4_48	5	TART-A	BEAF GAF
8	2L_86	5	mdg3	BEAF
9	X_921	5	297	GAF F7M

we compared the expression profiles of 5′- and 3′-flanking genes of predicted boundary elements using the publicly available FlyAtlas data set (38). On the basis of intergenic and gene flank criteria, we have shortlisted 2559 boundaries for this comparison. For each boundary we extracted 5′- and 3′-flanking genes using a gene table and compared their expression profiles using a Perl script. The expression profile includes the data for 17 adult tissues, eight larval tissues and a cell line (S2 cells). We termed two genes as 'differentially expressed', if their expression profiles are in negative correlation for one or more tissues (See 'Materials and Methods' section for details). Further 497 boundaries were removed from the list as their flanking gene's expression profile is not available in FlyAtlas data set. In the end, we have isolated 1545 boundaries that are having differentially expressed gene pairs as flanking genes (Supplementary File S3). These 1545 boundaries constitute 75% of total intergenic boundaries that we considered in this analysis. The boundaries 3R_592 and 3R_913 are also appeared in this analysis as they separate the differentially expressed genes i.e. C15-CG7956 and dco-Sox100B respectively.

Enhancer-blocking activity of the boundaries identified by cdBEST analysis

To assay predicted boundary elements, we designed a construct that can be used in *Drosophila* S2 cells. We call this construct as NPG (Neomycin-PE enhancer-GFP) and it

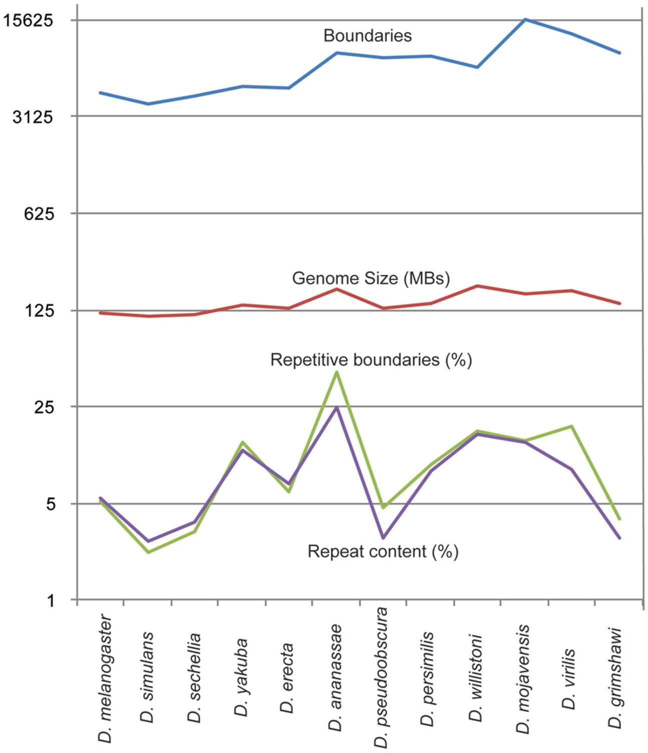

Figure 2. Boundaries and their repetitive nature in 12 *Drosophila* species. Four different data series, boundaries, repetitive boundaries, genome sizes and their repeat contents are plotted in logarithmic scale covering all 12 *Drosophila* species. Repetitive boundaries curve closely follows the repeat contents of the genomes indicating a strong positive correlation between them (i.e. genomes with higher repeat content are more likely to have higher number of repetitive boundaries).

has two reporters, neomycin gene for selection of the plasmid and GFP to assay the enhancer-blocking activity (Figure 4A). We cloned two known boundaries (*Fab-7 and Fab-8*) and 19 predicted boundaries of *D. melanogaster* (see Supplementary Table S8 for list of primes) in this construct and transfected them in S2 cells. We also included NG construct, a minimal version of NPG that lacks PE enhancer. We selected for stable integrants by growing the transfected S2 cells in G418 containing culture medium for 6 weeks and assayed for enhancer blocker activity using Flow cytometry analysis. The result show that *Fab-7* and *Fab-8* function as strong enhancer blockers in this assay, as their percentage of cells that express GFP is comparable to NG (Figure 4B). Of the nineteen cdBEST boundaries that were assayed fifteen showed enhancer-blocking activity (11 strong & 4 moderate). In the remaining four, two elements that are close to *Abd-B* promoter had higher GFP fluorescence than NPG vector (Figure 4B). It is interesting to note that one of the tested boundary, 4_29, is repetitive in nature and it also showed strong enhancer-blocking activity.

DISCUSSION

Chromatin domain boundaries are the key regulatory elements that help in packaging the genome and regulating gene expression as they are known to subdivide the genome into independent functional domains (9,58–61). Mapping the boundary elements at the genome level, therefore, gives a global view of the structural and functional organization of the genome. Unlike coding or other

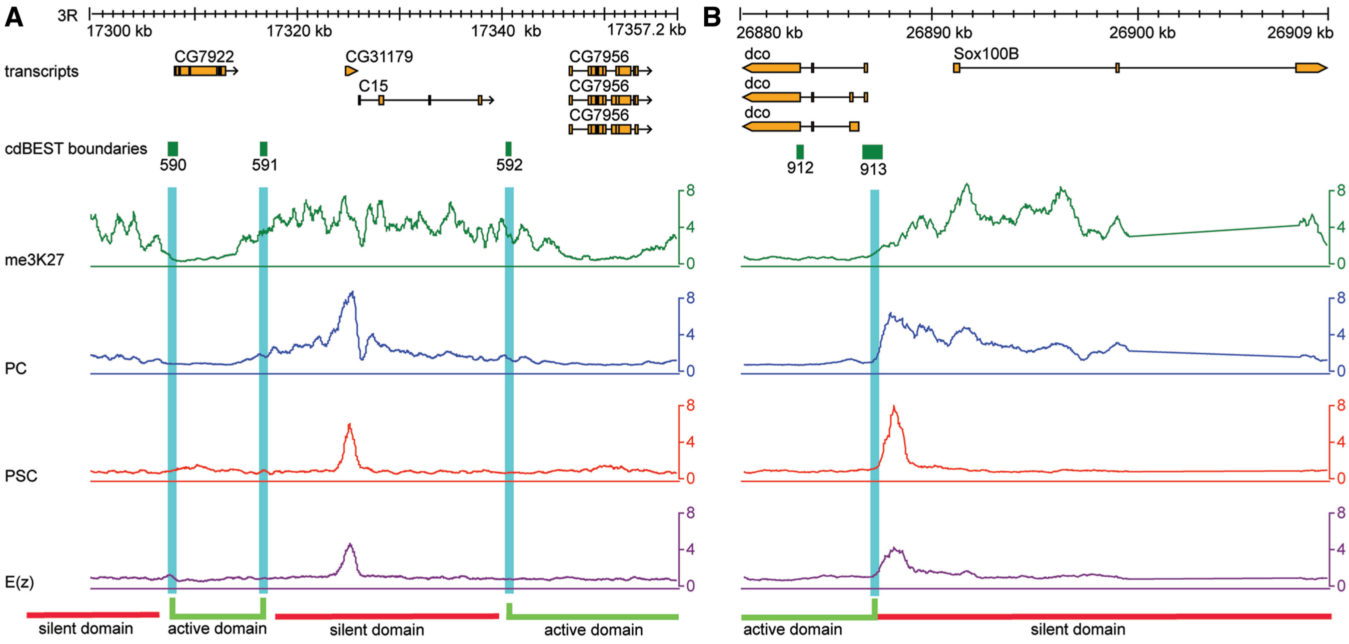

Figure 3. Predicted boundary elements mark the borders of Polycomb mediated repressed domains**.** The A and B parts are two representative regions of chromosome 3R of *Drosophila* genome. Upper panels show the predicted boundaries and annotated gene transcripts with scale. Lower panels show the binding profiles (ChIP/input ratio) for H3K27me3, PC, PSC and E(Z) proteins obtained from previously published ChIP-chip study (37).

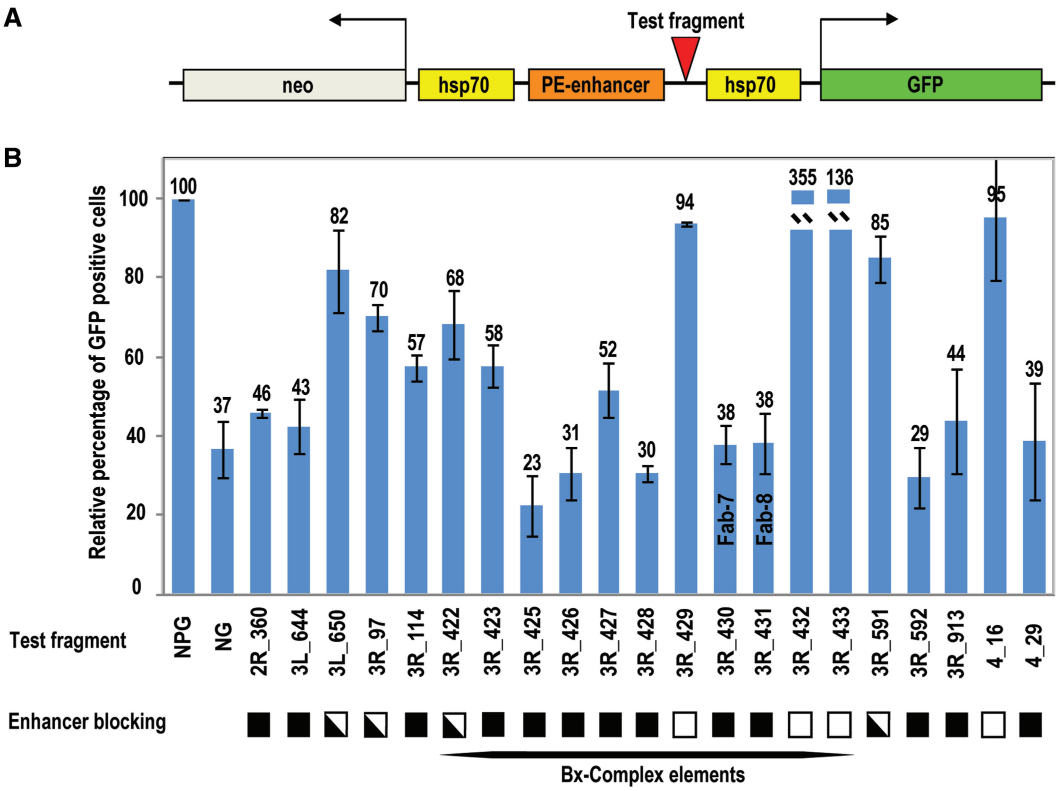

Figure 4. Predicted boundary elements function as enhancer blockers in *Drosophila* S2 cells. (**A**) The enhancer-blocking assay vector, NPG, showing the *neo* resistance gene, the PE enhancer, the GFP reporter gene and the test DNA insertion site. If the test DNA blocks enhancer-promoter communication, the stably transfected cells would have a lesser number of GFP positive cells. (**B**) Flow cytometry analysis was used to determine number of GFP positive cells. For each test DNA, percentage of GFP positive cells was calculated and plotted relative to NPG vector transfection. Filled black boxes indicate strong enhancer-blocking activity and half-filled ones indicate moderate activity and empty boxes show weak or no blocking activity.

regulatory elements, boundaries do not have prominent sequence features that are common to all boundaries. Although a boundary can replace another boundary in the endogenous locus and rescue the function, they lack any apparent primary sequence similarity (62). This, therefore, rule out simple sequence comparison based search for such elements in the genome. Biochemical analysis of several boundaries elements in *Drosophila* earlier have led to the identification of small sequence motifs that are recognition sites of the DNA-binding proteins involved in the boundary function. We noticed that all the boundaries contain a cluster of such motifs and several of them are common in subset of boundaries. Based on this motif clustering, here we describe a bioinformatics approach, cdBEST, to identify boundaries in *D. melanogaster* genome. We were also able to assign common sequence features and derive boundary type specific criteria needed to predict various subclasses of boundaries that exist in *Drosophila* genome (30,63).

We used several approaches to validate the cdBEST predicted boundaries. Using an available genome-wide epigenetic profiling data of *Polycomb* group of proteins, we looked if a predicted boundary was seen between the repressive and active regions of the genome based on epigenetic mark. Several predicted boundaries, indeed, were found in such locations supporting that boundaries subdivide genome into functional domains (Figure 3). In an independent approach, we also observed that around 75% of the genes that flank predicted boundary are differentially expressed in one or more tissues/cell lines. This supports the anticipated role of the predicted boundaries in their genomic context. The strongest support for the relevance of the cdBEST predicted boundary elements comes from the direct demonstration of their enhancer-blocking activity. We tested 19 predicted boundaries and found that 15 of them function as enhancer blockers in *Drosophila* S2 cells (Figure 4). These results allow us to conclude that the predicted boundaries are indeed functional elements in the *Drosophila* genome.

Several approaches have been used recently to address chromatin domains and boundaries in *Drosophila* and human. One of these approaches is computational search of motifs across the genome that are the sites of interaction for individual boundary interacting protein (24,40). This approach leaves other factors that may co-occupy the boundary region while cdBEST uses ‘motif cluster’ approach, where cluster of boundary motifs is preferred over single motif. In addition, cdBEST includes all the known boundary motifs and covers various boundary types that are present in *Drosophila*, which makes the search more comprehensive.

cdBEST also has the flexibility of changing parameters and constraints and can be set to individual motif search too.

Another approach to identify boundaries at genome scale is by ChIP based *in vivo* occupancy of individual boundary proteins (64–66) which is experimental equivalent of the above discussed computation approach. A major difficulty in this approach is that majority of the boundary interacting proteins in *Drosophila* are also involved in other nuclear functions such as transcriptional repressor or activator and, therefore, each site detected for interaction may not necessarily reflect boundary (65). Furthermore, ChIP experiments are often performed using a single cell line, or mixed tissue such as embryo which may not reflect the complexities involved in each and every tissue and cell types (30). To investigate this, we compared our cdBEST boundaries with a published ChIP data (64). As indicated in Figure 1, 11 out of the 12 predicted boundaries that are present in BX-C had clear overlapping ChIP signals. We find that ~55% of the cdBEST boundaries have an overlapping ChIP signal for one or more boundary proteins. While this is a reasonable agreement, it is possible that at least some of the remaining 45% cdBEST predicted boundaries may be tissue specific and may not be bound by proteins in cells where they are not functional. Also, since cdBEST used sequence motifs of additional boundary proteins, for example, Zw5, Elba and F7M, and genome scale ChIP data is not available for these proteins, cdBEST can still predict boundaries dependent on above mentioned factors. We also noticed several instances where a site identified as binding region for a boundary protein *in vivo* (for example, CP190) does not have the consensus DNA sequence motif on boundaries (64,67). Considering that boundaries can cluster together, some may not have direct binding sites and may be recruited through protein–protein interactions (67). Such boundaries can show up in ChIP based analyses but will be missed in recognition motif based predictions. Since cdBEST uses experimentally tested sites in the context of their boundary function and the scoring system has been optimized keeping 'true boundary motifs' context in the consideration, it has stringent predictive value. In addition, cdBEST has a clear advantage over ChIP approach as it can be applied any other closely related genomes.

The third genome scale boundary search approach has been to look at the transition regions in profiling of histone modifications or chromatin proteins that define chromatin domains (68,69). This approach is novel and most recent with the limitation that it is human specific and does not offer any tool which can be applied to other genomes (69). Although we have cdBEST for *D. melanogaster*, since it uses motif based approach it offers a tool which can be optimized in many other closely related genomes as seen from our boundary search results in non-melanogaster drosophilids. A related study that used genome scale profiling of more than 50 chromatin proteins shows five principal chromatin types that are present in *Drosophila* Kc-167 cells (68). In this study 8428 chromatin domains have been identified with the median size of 6.5 kb. Their data provides a fair idea of chromatin domains that are present in *Drosophila* cells. We explored how frequently the transition regions of these chromatin domains coincided with the boundaries defined by cdBEST. Of the 4576 cdBEST boundaries, 21% (977) overlap within 2 kb sequence that was used as the transition region, while the rest of the boundaries are found be located inside these domains. We took a close look at the BX-Complex region that has a series of well identified and studied boundaries separating independent regulatory domains (11). While all the BX-C boundaries were mapped by cdBEST and validated by enhancer-blocking assays (Figure 4B), in the 'five principal type chromatin' study the entire BX-Complex is marked as a single BLUE chromatin that corresponds to PcG chromatin. While this is in agreement with the chromatin state of Kc167 cell line, where BX-C genes are repressed by PcG proteins, it does not reflect the dynamic and cell type specific redistribution of chromatin types. Since cdBEST uses the primary sequence alone as the input, it extracts all possible boundaries that depend on the motifs used even if they may not functionally exist in a particular cell type or state. Such an inclusive approach gives the global picture the genome organization.

Any whole genome analysis is not complete, specially, in higher eukaryotes, unless it takes into account the repetitive elements. Several lines of studies indicate role of repetitive DNA in boundary function (16,70–72). In cdBEST based analysis, we also find several boundaries that occur multiple times in the genome and majority of them turned out to be associated with transposable elements (Supplementary Table S5). Prior to this analysis, *gypsy* and *Idefix* were the only transposable elements in *Drosophila* known to have boundary function and further experiments may identify many such elements and link these transposable elements to regulatory function. Boundary activity associated with repetitive sequences is of special significance as it provides means to regulate number of loci with fewer protein factors in a coordinated manner (72).

In conclusion, cdBEST is a reliable tool to detect boundaries at whole genome scale in *D. melanogaster* and many other drosophilids. With the help of cdBEST, we can annotate a significant portion of the genome (~3%) as boundary elements. As majority of the boundary interacting proteins are conserved among insects (73,74), cdBEST can be easily adapted to other insect genomes to search boundary element sequences and annotate their genomes for boundaries. With the increasing number of species whose genome sequences are being made available, for example, i5k project of 5000 insects and other arthropods (75), tools like cdBEST will be helpful to analyse and understand features of genome organization and function.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–8, Supplementary Figure 1 and Supplementary Files 1–3.

ACKNOWLEDGEMENTS

The authors thank Jyotsna Singh for her help in NPG assay vector preparation and G. Srinivas for help in Flow Cytometry data analysis.

FUNDING

RKM laboratory is supported by Council of Scientific and Industrial Research (CSIR), India and research grants from Department of Biotechnology (DBT) and Department of Science and Technology (DST) of the Government of India. Funding for open access charge: Centre for Cellular and Molecular Biology, Hyderabad, India.

Conflict of interest statement. None declared.

REFERENCES

1. Labrador,M. and Corces,V.G. (2002) Setting the boundaries of chromatin domains and nuclear organization. *Cell*, **111**, 151–154.
2. Udvardy,A., Maine,E. and Schedl,P. (1985) The 87A7 chromomere. Identification of novel chromatin structures flanking the heat shock locus that may define the boundaries of higher order domains. *J. Mol. Biol.*, **185**, 341–358.
3. Gdula,D.A., Gerasimova,T.I. and Corces,V.G. (1996) Genetic and molecular analysis of the gypsy chromatin insulator of Drosophila. *Proc. Natl Acad. Sci. USA*, **93**, 9378–9383.
4. Karch,F., Galloni,M., Sipos,L., Gausz,J., Gyurkovics,H. and Schedl,P. (1994) Mcp and Fab-7: molecular analysis of putative boundaries of cis-regulatory domains in the bithorax complex of Drosophila melanogaster. *Nucleic Acids Res.*, **22**, 3138–3146.
5. Hagstrom,K., Muller,M. and Schedl,P. (1996) Fab-7 functions as a chromatin domain boundary to ensure proper segment specification by the Drosophila bithorax complex. *Genes Dev.*, **10**, 3202–3215.
6. Zhou,J., Barolo,S., Szymanski,P. and Levine,M. (1996) The Fab-7 element of the bithorax complex attenuates enhancer-promoter interactions in the Drosophila embryo. *Genes Dev.*, **10**, 3195–3201.
7. Mihaly,J., Hogga,I., Gausz,J., Gyurkovics,H. and Karch,F. (1997) In situ dissection of the Fab-7 region of the bithorax complex into a chromatin domain boundary and a Polycomb-response element. *Development*, **124**, 1809–1820.
8. Barges,S., Mihaly,J., Galloni,M., Hagstrom,K., Muller,M., Shanower,G., Schedl,P., Gyurkovics,H. and Karch,F. (2000) The Fab-8 boundary defines the distal limit of the bithorax complex iab-7 domain and insulates iab-7 from initiation elements and a PRE in the adjacent iab-8 domain. *Development*, **127**, 779–790.
9. Bell,A.C., West,A.G. and Felsenfeld,G. (2001) Insulators and boundaries: versatile regulatory elements in the eukaryotic genome. *Science*, **291**, 447–450.
10. West,A.G., Gaszner,M. and Felsenfeld,G. (2002) Insulators: many functions, many mechanisms. *Genes Dev.*, **16**, 271–288.
11. Mishra,R.K. and Karch,F. (1999) Boundaries that demarcate structural and functional domains of chromatin *J. Biosci.*, **24**, 377–399.
12. Kellum,R. and Schedl,P. (1992) A group of scs elements function as domain boundaries in an enhancer-blocking assay. *Mol. Cell. Biol.*, **12**, 2424–2431.
13. Kellum,R. and Schedl,P. (1991) A position-effect assay for boundaries of higher order chromosomal domains. *Cell*, **64**, 941–950.
14. Ciavatta,D., Rogers,S. and Magnuson,T. (2007) Drosophila CTCF is required for Fab-8 enhancer blocking activity in S2 cells. *J. Mol. Biol.*, **373**, 233–239.
15. Li,M., Belozerov,V.E. and Cai,H.N. (2008) Analysis of chromatin boundary activity in Drosophila cells. *BMC Mol. Biol.*, **9**, 109.
16. Kim,J.H., Ebersole,T., Kouprina,N., Noskov,V.N., Ohzeki,J., Masumoto,H., Mravinac,B., Sullivan,B.A., Pavlicek,A., Dovat,S. *et al.* (2009) Human gamma-satellite DNA maintains open chromatin structure and protects a transgene from epigenetic silencing. *Genome Res.*, **19**, 533–544.
17. Valenzuela,L. and Kamakaka,R.T. (2006) Chromatin insulators. *Annu. Rev. Genet.*, **40**, 107–138.
18. Zhao,K., Hart,C.M. and Laemmli,U.K. (1995) Visualization of chromosomal domains with boundary element-associated factor BEAF-32. *Cell*, **81**, 879–889.
19. Cuvier,O., Hart,C.M. and Laemmli,U.K. (1998) Identification of a class of chromatin boundary elements. *Mol. Cell. Biol.*, **18**, 7478–7486.
20. Gaszner,M., Vazquez,J. and Schedl,P. (1999) The Zw5 protein, a component of the scs chromatin domain boundary, is able to block enhancer-promoter interaction. *Genes Dev.*, **13**, 2098–2107.
21. Schweinsberg,S., Hagstrom,K., Gohl,D., Schedl,P., Kumar,R.P., Mishra,R. and Karch,F. (2004) The enhancer-blocking activity of the Fab-7 boundary from the Drosophila bithorax complex requires GAGA-factor-binding sites. *Genetics*, **168**, 1371–1384.
22. Spana,C., Harrison,D.A. and Corces,V.G. (1988) The Drosophila melanogaster suppressor of Hairy-wing protein binds to specific sequences of the gypsy retrotransposon. *Genes Dev.*, **2**, 1414–1423.
23. Parnell,T.J., Viering,M.M., Skjesol,A., Helou,C., Kuhn,E.J. and Geyer,P.K. (2003) An endogenous suppressor of hairy-wing insulator separates regulatory domains in Drosophila. *Proc. Natl Acad. Sci. USA*, **100**, 13436–13441.
24. Ramos,E., Ghosh,D., Baxter,E. and Corces,V.G. (2006) Genomic organization of gypsy chromatin insulators in Drosophila melanogaster. *Genetics*, **172**, 2337–2349.
25. Parnell,T.J., Kuhn,E.J., Gilmore,B.L., Helou,C., Wold,M.S. and Geyer,P.K. (2006) Identification of genomic sites that bind the Drosophila suppressor of Hairy-wing insulator protein. *Mol. Cell. Biol.*, **26**, 5983–5993.
26. Moon,H., Filippova,G., Loukinov,D., Pugacheva,E., Chen,Q., Smith,S.T., Munhall,A., Grewe,B., Bartkuhn,M., Arnold,R. *et al.* (2005) CTCF is conserved from Drosophila to humans and confers enhancer blocking of the Fab-8 insulator. *EMBO Rep.*, **6**, 165–170.
27. Aoki,T., Schweinsberg,S., Manasson,J. and Schedl,P. (2008) A stage-specific factor confers Fab-7 boundary activity during early embryogenesis in Drosophila. *Mol. Cell. Biol.*, **28**, 1047–1060.
28. Jiang,N., Emberly,E., Cuvier,O. and Hart,C.M. (2009) Genome-wide mapping of boundary element-associated factor (BEAF) binding sites in Drosophila melanogaster links BEAF to transcription. *Mol. Cell. Biol.*, **29**, 3556–3568.
29. Emberly,E., Blattes,R., Schuettengruber,B., Hennion,M., Jiang,N., Hart,C.M., Kas,E. and Cuvier,O. (2008) BEAF regulates cell-cycle genes through the controlled deposition of H3K9 methylation marks into its conserved dual-core binding sites. *PLoS Biol.*, **6**, 2896–2910.
30. Bushey,A.M., Ramos,E. and Corces,V.G. (2009) Three subclasses of a Drosophila insulator show distinct and cell type-specific genomic distributions. *Genes Dev.*, **23**, 1338–1350.
31. Holohan,E.E., Kwong,C., Adryan,B., Bartkuhn,M., Herold,M., Renkawitz,R., Russell,S. and White,R. (2007) CTCF genomic binding sites in Drosophila and the organisation of the bithorax complex. *PLoS Genet.*, **3**, e112.
32. Smith,S.T., Wickramasinghe,P., Olson,A., Loukinov,D., Lin,L., Deng,J., Xiong,Y., Rux,J., Sachidanandam,R., Sun,H. *et al.* (2009) Genome wide ChIP-chip analyses reveal important roles for CTCF in Drosophila genome organization. *Dev. Biol.*, **328**, 518–528.
33. Chopra,V.S., Srinivasan,A., Kumar,R.P., Mishra,K., Basquin,D., Docquier,M., Seum,C., Pauli,D. and Mishra,R.K. (2008) Transcriptional activation by GAGA factor is through its direct interaction with dmTAF3. *Dev. Biol.*, **317**, 660–670.
34. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I., Lapp,H. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
35. Barolo,S., Carver,L.A. and Posakony,J.W. (2000) GFP and beta-galactosidase transformation vectors for promoter/enhancer analysis in Drosophila. *Biotechniques*, **29**, 726, 728, 730, 732.

36. Jiang,J., Kosman,D., Ip,Y.T. and Levine,M. (1991) The dorsal morphogen gradient regulates the mesoderm determinant twist in early Drosophila embryos. *Genes Dev*., **5**, 1881–1891.
37. Schwartz,Y.B., Kahn,T.G., Nix,D.A., Li,X.Y., Bourgon,R., Biggin,M. and Pirrotta,V. (2006) Genome-wide analysis of Polycomb targets in Drosophila melanogaster. *Nat. Genet*., **38**, 700–705.
38. Chintapalli,V.R., Wang,J. and Dow,J.A. (2007) Using FlyAtlas to identify better Drosophila melanogaster models of human disease. *Nat. Genet*., **39**, 715–720.
39. Bartkuhn,M., Straub,T., Herold,M., Herrmann,M., Rathke,C., Saumweber,H., Gilfillan,G.D., Becker,P.B. and Renkawitz,R. (2009) Active promoters and insulators are marked by the centrosomal protein 190. *EMBO J*., **28**, 877–888.
40. Xie,X., Mikkelsen,T.S., Gnirke,A., Lindblad-Toh,K., Kellis,M. and Lander,E.S. (2007) Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc. Natl Acad. Sci. USA*, **104**, 7145–7150.
41. Eden,E., Lipson,D., Yogev,S. and Yakhini,Z. (2007) Discovering motifs in ranked lists of DNA sequences. *PLoS Comput. Biol*., **3**, e39.
42. Fiedler,T. and Rehmsmeier,M. (2006) jPREdictor: a versatile tool for the prediction of cis-regulatory elements. *Nucleic Acids Res*., **34**, W546–W550.
43. Ohtsuki,S. and Levine,M. (1998) GAGA mediates the enhancer blocking activity of the eve promoter in the Drosophila embryo. *Genes Dev*., **12**, 3325–3330.
44. Fujioka,M., Wu,X. and Jaynes,J.B. (2009) A chromatin insulator mediates transgene homing and very long-range enhancer-promoter communication. *Development*, **136**, 3077–3087.
45. Chopra,V.S., Cande,J., Hong,J.W. and Levine,M. (2009) Stalled Hox promoters as chromosomal boundaries. *Genes Dev*., **23**, 1505–1509.
46. Sultana,H., Verma,S. and Mishra,R.K. A BEAF dependent chromatin domain boundary separates myoglianin and eyeless genes of Drosophila melanogaster. *Nucleic Acids Res*., **39**, 3543–3557.
47. McQuilton,P., St Pierre,S.E. and Thurmond,J. FlyBase 101 - the basics of navigating FlyBase. *Nucleic Acids Res*, **40**, D706–D714.
48. Jurka,J., Kapitonov,V.V., Pavlicek,A., Klonowski,P., Kohany,O. and Walichiewicz,J. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res*., **110**, 462–467.
49. Kaminker,J.S., Bergman,C.M., Kronmiller,B., Carlson,J., Svirskas,R., Patel,S., Frise,E., Wheeler,D.A., Lewis,S.E., Rubin,G.M. *et al*. (2002) The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. *Genome Biol*., **3**, RESEARCH0084.
50. Clark,A.G., Eisen,M.B., Smith,D.R., Bergman,C.M., Oliver,B., Markow,T.A., Kaufman,T.C., Kellis,M., Gelbart,W., Iyer,V.N. *et al*. (2007) Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, **450**, 203–218.
51. Ringrose,L. and Paro,R. (2004) Epigenetic regulation of cellular memory by the Polycomb and Trithorax group proteins. *Annu. Rev. Genet*., **38**, 413–443.
52. Schwartz,Y.B. and Pirrotta,V. (2008) Polycomb complexes and epigenetic states. *Curr. Opin. Cell Biol*., **20**, 266–273.
53. Papp,B. and Muller,J. (2006) Histone trimethylation and the maintenance of transcriptional ON and OFF states by trxG and PcG proteins. *Genes Dev*., **20**, 2041–2054.
54. Kahn,T.G., Schwartz,Y.B., Dellino,G.I. and Pirrotta,V. (2006) Polycomb complexes and the propagation of the methylation mark at the Drosophila ubx gene. *J. Biol. Chem*., **281**, 29064–29075.
55. Sigrist,C.J. and Pirrotta,V. (1997) Chromatin insulator elements block the silencing of a target gene by the Drosophila polycomb response element (PRE) but allow trans interactions between PREs on different chromosomes. *Genetics*, **147**, 209–221.
56. van der Vlag,J., den Blaauwen,J.L., Sewalt,R.G., van Driel,R. and Otte,A.P. (2000) Transcriptional repression mediated by polycomb group proteins and other chromatin-associated repressors is selectively blocked by insulators. *J. Biol. Chem*., **275**, 697–704.
57. Mendenhall,E.M. and Bernstein,B.E. (2008) Chromatin state maps: new technologies, new insights. *Curr. Opin. Genet. Dev*, **18**, 109–115.
58. Bushey,A.M., Dorman,E.R. and Corces,V.G. (2008) Chromatin insulators: regulatory mechanisms and epigenetic inheritance. *Mol. Cell*, **32**, 1–9.
59. Wallace,J.A. and Felsenfeld,G. (2007) We gather together: insulators and genome organization. *Curr. Opin. Genet. Dev*., **17**, 400–407.
60. Gaszner,M. and Felsenfeld,G. (2006) Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat. Rev. Genet*., **7**, 703–713.
61. Brasset,E. and Vaury,C. (2005) Insulators are fundamental components of the eukaryotic genomes. *Heredity*, **94**, 571–576.
62. Iampietro,C., Cleard,F., Gyurkovics,H., Maeda,R.K. and Karch,F. (2008) Boundary swapping in the Drosophila Bithorax complex. *Development*, **135**, 3983–3987.
63. Gurudatta,B.V. and Corces,V.G. (2009) Chromatin insulators: lessons from the fly. *Brief. Funct. Genomic. Proteomic*., **8**, 276–282.
64. Negre,N., Brown,C.D., Shah,P.K., Kheradpour,P., Morrison,C.A., Henikoff,J.G., Feng,X., Ahmad,K., Russell,S., White,R.A. *et al*. A comprehensive map of insulator elements for the Drosophila genome. *PLoS Genet*., **6**, e1000814.
65. Negre,N., Brown,C.D., Ma,L., Bristow,C.A., Miller,S.W., Wagner,U., Kheradpour,P., Eaton,M.L., Loriaux,P., Sealfon,R. *et al*. A cis-regulatory map of the Drosophila genome. *Nature*, **471**, 527–531.
66. Martin,D., Pantoja,C., Fernandez Minan,A., Valdes-Quezada,C., Molto,E., Matesanz,F., Bogdanovic,O., de la Calle-Mustienes,E., Dominguez,O., Taher,L. *et al*. Genome-wide CTCF distribution in vertebrates defines equivalent sites that aid the identification of disease-associated genes. *Nat. Struct. Mol. Biol*., **18**, 708–714.
67. Pai,C.Y., Lei,E.P., Ghosh,D. and Corces,V.G. (2004) The centrosomal protein CP190 is a component of the gypsy chromatin insulator. *Mol. Cell*, **16**, 737–748.
68. Filion,G.J., van Bemmel,J.G., Braunschweig,U., Talhout,W., Kind,J., Ward,L.D., Brugman,W., de Castro,I.J., Kerkhoven,R.M., Bussemaker,H.J. *et al*. Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell*, **143**, 212–224.
69. Wang,J., Lunyak,V.V. and Jordan,I.K. Genome-wide prediction and analysis of human chromatin boundary elements. *Nucleic Acids Res*, **40**, 511–529.
70. Lunyak,V.V., Prefontaine,G.G., Nunez,E., Cramer,T., Ju,B.G., Ohgi,K.A., Hutt,K., Roy,R., Garcia-Diaz,A., Zhu,X. *et al*. (2007) Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. *Science*, **317**, 248–251.
71. Brasset,E., Bantignies,F., Court,F., Cheresiz,S., Conte,C. and Vaury,C. (2007) Idefix insulator activity can be modulated by nearby regulatory elements. *Nucleic Acids Res*., **35**, 2661–2670.
72. Kumar,R.P., Senthilkumar,R., Singh,V. and Mishra,R.K. (2010) Repeat performance: how do genome packaging and regulation depend on simple sequence repeats? *Bioessays*, **32**, 165–174.
73. Gray,C.E. and Coates,C.J. (2005) Cloning and characterization of cDNAs encoding putative CTCFs in the mosquitoes, Aedes aegypti and Anopheles gambiae. *BMC Mol. Biol*., **6**, 16.
74. Schoborg,T.A. and Labrador,M. (2009) The phylogenetic distribution of non-CTCF insulator proteins is limited to insects and reveals that BEAF-32 is Drosophila lineage specific. *J. Mol. Evol*, **70**, 74–84.
75. Robinson,G.E., Hackett,K.J., Purcell-Miramontes,M., Brown,S.J., Evans,J.D., Goldsmith,M.R., Lawson,D., Okamuro,J., Robertson,H.M. and Schneider,D.J. Creating a buzz about insect genomes. *Science*, **331**, 1386.