Developmental Cell, Vol. 5, 759–771, November, 2003, Copyright ©2003 by Cell Press

# Genome-Wide Prediction of Polycomb/Trithorax Response Elements in *Drosophila melanogaster*

**Leonie Ringrose,[1,4] Marc Rehmsmeier,[2,4] Jean-Maurice Dura,[3] and Renato Paro[1,*]**

[1]ZMBH
Universität Heidelberg
Im Neuenheimer Feld 282
69120 Heidelberg
Germany
[2]Universität Bielefeld
International NRW Graduate School
in Bioinformatics and Genome Research
Postfach 100131
33501 Bielefeld
Germany
[3]Institut de Génétique Humaine
CNRS UPR1142
141, Rue de la Cardonille
34396 Montpellier
France

## Summary

**Polycomb/Trithorax response elements (PRE/TREs) maintain transcriptional decisions to ensure correct cell identity during development and differentiation. There are thought to be over 100 PRE/TREs in the *Drosophila* genome, but only very few have been identified due to the lack of a defining consensus sequence. Here we report the definition of sequence criteria that distinguish PRE/TREs from non-PRE/TREs. Using this approach for genome-wide PRE/TRE prediction, we identify 167 candidate PRE/TREs, which map to genes involved in development and cell proliferation. We show that candidate PRE/TREs are bound and regulated by Polycomb proteins in vivo, thus demonstrating the validity of PRE/TRE prediction. Using the larger data set thus generated, we identify three sequence motifs that are conserved in PRE/TRE sequences.**

## Introduction

Polycomb/Trithorax response elements (PRE/TREs) are epigenetic switchable elements. They maintain the previously determined transcriptional state of their associated genes over many cell generations, thus ensuring a memory of transcriptional history. The Polycomb group (PcG) proteins mediate transcriptional repression while the trithorax group (trxG) proteins act antagonistically, maintaining activation (Orlando, 2003).

To date, identification of *Drosophila* PRE/TREs has relied upon functional assays such as transgene analysis or chromatin immunoprecipitation. These approaches have identified PRE/TREs and the genes they regulate at five loci: the homeotic genes of the bithorax (BX-C) and Antennapedia complexes (ANT-C), and the *engrailed*, *polyhomeotic*, and *hedgehog* genes (Mihaly et al., 1998 and references therein; Bloyer et al., 2003; Maurange and Paro, 2002). The emerging picture is that PRE/TREs may play a global role in maintaining correct cell identity (Orlando, 2003; Zuckerkandl, 1999); thus, it is of fundamental importance to identify them and their associated genes. Many PRE/TREs are thought to exist in the *Drosophila* genome, because Polycomb and trithorax group proteins bind to over 100 sites on polytene chromosomes from larval salivary glands (Zuckerkandl, 1999 and references therein). However, PRE/TREs are typically a few kilobases long, whereas the resolution of polytene mapping is in the range of several hundreds of kilobases, and thus it does not enable identification of individual PRE/TREs or the genes they regulate.

The PRE/TREs thus far identified all show similar properties when taken out of their endogenous context and inserted elsewhere in the genome. These properties include pairing-sensitive repression of adjacent reporter genes in a manner that is genetically dependent on the PcG and trxG, and recruitment of PcG/trxG proteins to the site of transgene insertion. These functional similarities indicate that PRE/TREs must share common DNA sequence features. However, alignment of known PRE/TRE sequences reveals little similarity between them, and thus it has so far not been possible to define a consensus sequence helpful for the identification of the many other PRE/TREs in the *Drosophila* genome.

Nevertheless, several short motifs that are required for PRE/TRE function have been identified. These include binding sites for three sequence-specific DNA binding proteins: the Pleiohomeotic protein (PHO), a PcG member (Brown et al., 1998; Mihaly et al., 1998), and the GAGA factor (GAF; Strutt et al., 1997) and the zeste protein (Z; Saurin et al., 2001; Hur et al., 2002), both of which are trxG members. Each of these motifs occurs at least once in all known PRE/TREs, and thus one might expect that other PRE/TREs could be identified by searching for these motifs. However, such an approach is limited by the shortness of the GAF binding site and the degeneracy of the PHO and Z consensus sites (Table 1), and so all of them will occur with a certain frequency at random in any DNA sequence. Furthermore, GAF and Z regulate many genes independently of the PcG/trxG system, and thus functional sites occur in many regulatory regions that are not PRE/TREs. The same may also be true for PHO. Thus, for these motifs to contribute to true PRE/TRE function, additional features such as their spacing relative to one another, or other motifs, must put them in their correct context.

The greatest obstacle to the identification of PRE/TREs and the genes they regulate has been the lack of a sequence-based search tool that can accurately distinguish between PRE/TREs and non-PRE/TREs. Here we report the development of such a tool and its application to the prediction of PRE/TREs in the *Drosophila* genome, identifying over 100 candidate PRE/TREs and their associated genes. Furthermore, we show that predicted PRE/TREs are bound and regulated by

*Correspondence: paro@sun0.urz.uni-heidelberg.de
[4]These authors contributed equally to this work.

Table 1. Sequence Motifs

| Name | Sequence |
|---|---|
| G | GAGAG |
| G10 | GAGAGAGAGA |
| PS | GCCAT |
| PM | CNGCCATNDNND |
| PF | GCCATHWY |
| EN 1 | GSNMACGCCCC |
| Z | YGAGYG |

Motifs were defined for the purposes of computer searching as shown. G, GAGA factor (GAF) binding site (Strutt et al., 1997). G10, extended GAGA site; up to one mismatch was allowed. PS, core site bound by the Pleiohomeotic protein (PHO). PM, PHO consensus according to Mihaly et al. (1998). PF, PHO consensus according to Fritsch et al. (1999). EN 1 (Kassis et al., 1989). Up to one mismatch was allowed. Z, zeste binding site (Hur et al., 2002). Nucleotides are named according to the UIPAC code.

Table 2. Training Data Set: PRE/TREs and Non-PRE/TREs

| | Single Score | Pair Score |
|---|---|---|
| PRE/TREs | | |
| *bxd* | 13.27 | 441.60 |
| *iab-2* | 8.66 | 211.84 |
| *Fab-7* | 8.57 | 189.39 |
| *en Dm* | 6.00 | 111.42 |
| *Scr10X.2* | 7.54 | 91.77 |
| *ph p* | 4.13 | 88.19 |
| *en Dv* | 8.09 | 85.65 |
| *ph d* | 2.86 | 71.01 |
| *Mcp* | 8.31 | 49.90 |
| *Scr10X.1* | 4.56 | 38.23 |
| *Scr8.2Xba* | 4.73 | 27.15 |
| *iab-8* | 4.89 | 25.23 |
| Non-PRE/TREs | | |
| *white* | 6.03 | 45.47 |
| *hsp23* | 5.57 | 31.37 |
| *hsp27* | 4.05 | 27.50 |
| *hsc70-3* | 2.14 | 22.38 |
| *yellow* | 5.46 | 21.57 |
| *linotte* | 4.29 | 17.55 |
| *rosy* | 3.24 | 7.99 |
| *hsp67B* | 1.97 | 7.93 |
| *hsp68* | 2.92 | 7.71 |
| *hsp26* | 1.39 | 7.33 |
| *Polycomb* | 1.31 | 5.16 |
| *hsp83* | 0.37 | 1.64 |
| *hsc70-1* | 1.09 | 1.22 |
| *hsc70-4* | 1.15 | 0.71 |
| *hsc70-2* | 0.37 | 0.00 |
| *hsp22* | −0.18 | −58.81 |

The PRE/TREs and non-PRE/TREs used as training data are shown. PRE/TREs were selected according to published coordinates (see Experimental Procedures). Non-PRE/TRE sequences contain the upstream regulatory region of each gene up to the transcription start site, obtained from Flybase (http://flybase.bio.indiana.edu/). Column 2 shows the highest score of each sequence based on single motifs, and column 3 shows the highest scores based on paired motifs. For computational reasons, the 10 kb *Scr10Xba* PRE/TRE was separated into two subfragments. The paired motif scores for the two subfragments suggest that the main *Scr10Xba* PRE/TRE is in the *Scr10X.2* subfragment.

PcG proteins in vivo. This analysis not only expands the current repertoire of PRE/TRE sequences and associated genes, but also provides several unexpected insights into PRE/TRE function.

## Results

### Clusters of Single Motifs Do Not Define PRE/TREs

Unlike coding DNA, the sequence of a regulatory element is not read as a linear code in vivo, but rather in the context of chromatin, which may juxtapose distant sequences, and preferentially expose some sequences while obscuring others. Detection of similarities between regulatory sequences requires an approach that takes account of the three-dimensional space in which they operate. To evaluate sequence similarities between PRE/TREs, we thus developed a strategy that detects motif clustering without imposing constraints on motif order.

GAGA, Z, and PHO binding sites (Table 1) are required for the function of several PRE/TREs. In addition, deletion of a short motif (EN1; Table 1) from the *engrailed* PRE/TRE abrogates silencing function (Kassis et al., 1989). To ask whether we can distinguish PRE/TREs from non-PRE/TREs on the basis of these motifs, we compared sequences from 11 PRE/TREs with 16 non-PRE/TRE regulatory sequences (Table 2). This non-PRE/TRE training set includes promoters of genes that are regulated by GAF and Z (e.g., *hsp22* and *white*). By including these sequences, we aimed to define criteria whereby the occurrence of these sites in PRE/TREs can be distinguished from their occurrence at promoters regulated by GAF or Z.

To determine whether the single motifs of Table 1 can indeed be used to define PRE/TREs, we assigned a weight to each motif (Figure 1A). The weight reflects the extent to which a given motif occurs more often per kilobase in the PRE/TRE training set than in non-PRE/TREs. A weight of zero indicates equal abundance in both training sets. For example, GAF binding sites obtained weights close to zero. ZESTE binding sites also have a low weight. All three variants of the PHO site have a higher weight, indicating that they are more abundant in PRE/TREs.

To assess whether clustering of these motifs might characterize PRE/TREs, we used the weights from Figure 1A to calculate scores for windows of 500 bp across each sequence. For a given window, each motif was counted, and this number was multiplied by the weight of the motif itself. The values thus generated for each motif were added together to give a score for that window. In this way, favored motifs are "amplified" by their weights and will lead to high scores. This procedure was applied to both training sets, calculating scores for a 500 bp window moved in steps of 100 bp across each sequence. The highest score calculated for each sequence is shown in Table 2, column 2. Surprisingly, this approach gave only a poor separation of PRE/TREs from non-PRE/TREs. Smaller or larger window sizes did not improve the separation. Only six of the PRE/TRE sequences scored higher than the highest scoring non-PRE/TRE. Moreover, the difference in score between the highest non-PRE/TRE (*white*; 6.03) and the highest scoring PRE/TRE (*bxd*; 13.27) is only 2-fold.

We conclude that, although some motifs occur more often in PRE/TREs, clustered single motifs are not sufficient to distinguish between PRE/TREs and non-PRE/TREs in our training data set.

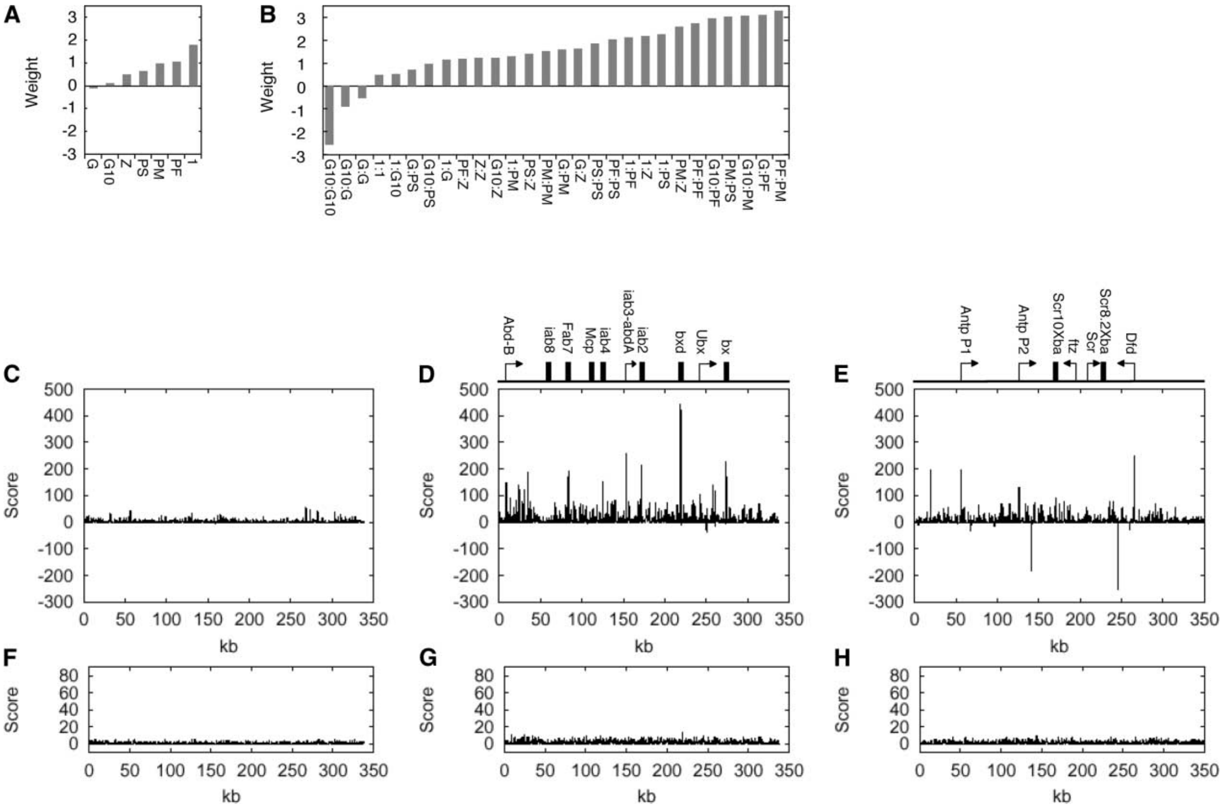

Figure 1. Using Single and Paired Motifs to Find PRE/TREs in the BX-C and ANT-C

(A) Weights for single motifs. Motif 1 = EN 1 (Table 1).

(B) As for (A), but calculated for pairs of motifs occurring within a distance of 220 bp.

(C–H) Score plots for test sequences.

(C–E) Scores calculated using weights for paired motifs as in (B).

(F–H) Scores from weights for single motifs as in (A).

(C and F) 350 kb random sequence.

(D and E) Diagrams show positions of PRE/TREs (black bars). Arrows indicate start site and direction of transcription of the homeotic genes and *ftz*. Eight peaks correspond to characterized PcG/trxG binding fragments not in training data

(D) *Abd-B* promoter (8,422–12,618); *Abd-B* introns (34,378–35,078); *iab-4* (123,772–127,219); *abd-A* promoter and *iab-3* (152,879–153,578); *Ubx* promoter (241,078–243,227); *bx* PRE/TRE (273,301–274,960).

(E) *Antp*P1 (55,707–59,706); *Antp*P2 (118,258–126,591).

(D and G) Bithorax complex (BX-C).

(E and H) Antennapedia complex (ANT-C).

## Clustered Motif Pairs Distinguish PRE/TREs from Non-PRE/TREs

The failure of clustered single motifs to distinguish the PRE/TREs from the non-PRE/TREs reflects the fact that these two data sets are very similar to each other in simple terms of motif composition and clustering. This prompted us to ask whether in PRE/TREs, particular combinations of motifs might work in concert, imposing a stringent constraint on the distance between pairs of similar or different binding sites.

To examine this idea, we analyzed the occurrence of all 28 possible pairwise combinations of the seven motifs. A pair was defined as two motifs occurring in any orientation on either strand within a distance of 220 bp or less. This is the approximate distance between adjacent nucleosomal linkers, and is the optimal distance for short-range looping in chromatin (Ringrose et al., 1999). As described above for the single motifs, we compared the PRE/TRE training set to the non-PRE/TREs, and calculated a weight for each motif pair (Figure 1B). These weights cover a wider range of positive and negative values than those for single motifs, and strongly disfavor GAF sites paired with themselves. Interestingly, Z:Z or GAF:Z pairs have a 3- to 4-fold higher weight than GAF or Z motifs alone (Figure 1A). Combinations of the various PHO sites with themselves, or with either GAF or Z, are strongly favored. We next asked whether these pairs are sufficient to distinguish between the two training sets. We calculated scores for each sequence as described above. Table 2, column 3 shows the highest score for each sequence. The paired motifs performed far better in this test than the single sites: nine of the eleven PRE/TREs achieved a higher score than the highest scoring non-PRE/TRE. The separation was not improved by other pair distances and window sizes (not shown). Strikingly, the difference in score between the highest scoring non-PRE/TRE (*white*; 45.47) and the highest scoring PRE/TRE (*bxd*; 441.60) is about 10-fold, indicating that motif pair scoring can generate robust separation between PRE/TREs and non-PRE/TREs.

In summary, this analysis shows that scoring based on clustered motif pairs distinguishes almost all PRE/TREs from non-PRE/TREs in our training sets, and suggests that this approach may enable detection of PRE/TREs in more complex data sets.

### Clustered Motif Pairs Correctly Identify PRE/TREs in the BX-C and ANT-C

The PRE/TRE training data set contains some but not all documented PRE/TREs. Eight further sites of PcG and trxG binding have been identified experimentally in the bithorax (BX-C) and Antennapedia (ANT-C) complexes (Zink et al., 1991; Simon et al., 1993; Strutt et al., 1997; Orlando et al., 1998). To ask whether motif-based scoring can correctly identify these PRE/TREs, we calculated scores in the BX-C and ANT-C. These complexes are each about 350 kb long. As a negative control, we used a randomly generated sequence of 350 kb. We calculated scores for each of these sequences for both single motifs and motif pairs using a 500 bp window as described above (Figures 1C–1H). The mean scores from the analysis of the random sequence were used to set equivalent vertical scales for single and paired score plots. In the BX-C and ANT-C, scoring for single motifs (Figures 1G and 1H) showed that there are very few individual peaks that score higher than the random sequence (Figure 1D). This confirms the poor performance of single motifs in separating PRE/TREs from non-PRE/TREs.

In contrast, scoring for motif pairs on the BX-C and ANT-C generated plots with many individual peaks that are clearly discernable above the background level (Figures 1D and 1E). Inspection of the precise coordinates of each of these peaks showed that four of them correspond to PRE/TREs in our training set (*Fab-7*, *iab-2*, *bxd*, and *Scr10X.2*). The training data set contained three further PRE/TREs (*iab-8*, *Mcp*, and *Scr 8.2 Xba*), which are not evident as strong peaks in the score plots. These may thus represent a subclass of PRE/TREs that do not conform to the criteria we have determined.

Remarkably, eight of the other high-scoring peaks in the BX-C and ANT-C correspond exactly to the eight documented fragments for which PcG/trxG binding has been demonstrated (see legend to Figure 1 for details). In addition to the 12 peaks for documented PRE/TREs, we further observe four strong peaks. The first BX-C peak (Figure 1D), at 40 kb, may be the *iab-9* PRE/TRE. The second peak, at 260 kb, lies between the *Ubx* promoter and the *bx* PRE/TRE, and may contribute to *bx* PRE/TRE function. In the ANT-C (Figure 1E), the two additional strong peaks are at 15 kb (upstream of the *Antp* promoter PRE/TREs) and at 265 kb. This latter peak corresponds exactly to the promoter of the *Deformed* gene, which is the third homeotic gene of the ANT-C, but for which no PRE/TRE has previously been defined.

Taken together, these results show that, in the more challenging context of the BX-C and the ANT-C, scoring for clustered motif pairs not only finds PRE/TREs that were in the training set, but also correctly identifies all of the eight other documented sites of PcG/trxG binding, as well as revealing several additional candidate PRE/TREs.

### Genome-Wide PRE/TRE Prediction Identifies 167 Candidate PRE/TREs

The success of scoring for clustered motif pairs in the context of the BX-C and ANT-C prompted us to use it to search for PRE/TREs in the *Drosophila* genome. We calculated scores for a 500 bp window moved in 100 bp steps across the entire sequenced *Drosophila* genome (117 Mb). To determine the significance of these scores, we performed the same operation on a random sequence that was 100 times longer, and used the empirical score distribution to express each score in terms of an E value. For a given score, the E value is the number of times one expects to find that score (or higher) in the *Drosophila* genome. Accordingly, a score with an E value of 1.0 would be expected to occur only once by chance in the genome. Analysis of the random sequence showed that an E value of 1.0 corresponds to a score of 157. We chose this E value as the cutoff for the prediction of PRE/TREs in the *Drosophila* genome, and thus sequences that score below 157 will not be detected by this analysis. There may be many true PRE/TREs in the genome that have a score below 157 (Table 2), but we expect only one non-PRE/TRE to score so well. Because our aim was not to find all PRE/TREs but to find real new PRE/TREs, we reasoned that selectivity should take priority over sensitivity.

Calculation of scores in the *Drosophila* genome identified 167 hits for which the E value is 1.0 or less (Figure 2; the sequence and genomic position of each hit is available at http://www.techfak.uni-bielefeld.de/marc/pre/). Comparison of the cytological positions of the predicted PRE/TREs shows an excellent agreement with immunocytologically mapped PcG and trxG binding sites (Figure 2; Zuckerkandl, 1999 and references therein). The 167 hits group into 115 cytological bands. Ninety-one of these bands correspond to binding sites for PcG or trxG proteins. The 24 predicted bands that are not thus accounted for may be bound in polytene chromosomes at levels that are undetectable by immunological staining, or they may be targets for the PcG and trxG in other tissues. Our predicted hits cover about 50% of all the immunologically detected PcG and trxG binding sites. The other binding sites may contain PRE/TREs that fall below our cutoff score of 157. From this we estimate that our prediction covers about half of the PRE/TREs in the genome. Because we predict 167 individual PRE/TRE sequences with an E value of 1.0 or less, we expect the genome to contain about 334 PRE/TREs in total.

### Predicted PRE/TREs Map to Genes Involved in Development and Cell Proliferation

To identify genes that may be regulated by the predicted PRE/TREs, we used the Flybase genome annotation (release 3.1) to find the gene closest to each hit (Table 3; Supplemental Table S1 at http://www.developmentalcell.com/cgi/content/full/5/5/759/DC1). Although we cannot be sure that the closest gene is the one that is regulated by the predicted PRE/TRE, we found that 118 (70%) of the hits are overlapping or very close to the nearest gene (less than 5 kb away), and thus these genes are good candidates for regulation. The other 30% of PREs that are further away from the closest gene may in fact

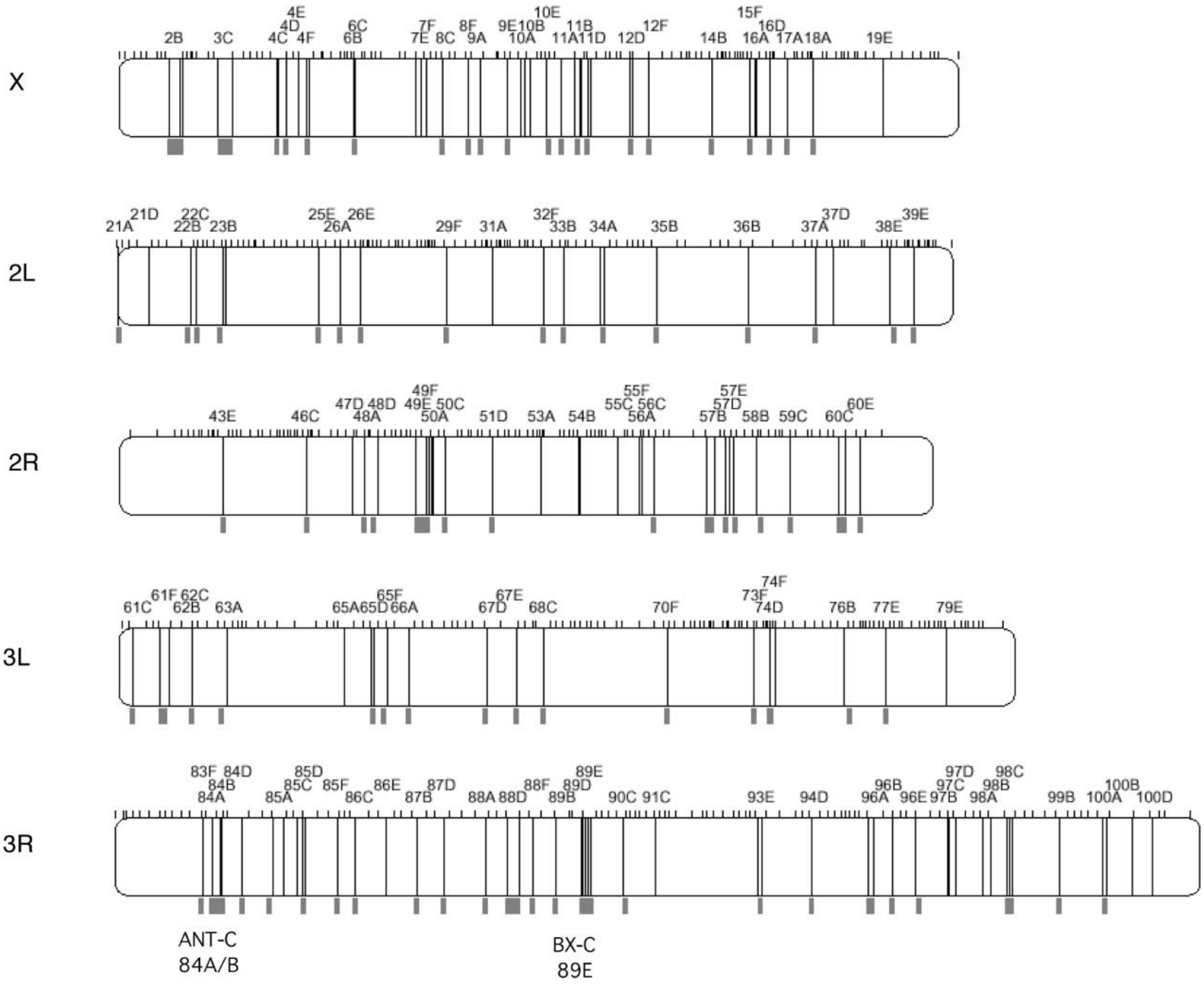

Figure 2. Genome-Wide PRE/TRE Prediction

Hits for which E = 1.0 or less are shown as black lines on each *Drosophila* chromosome arm. Each hit is 500–900 bp long. The cytological position of each hit, according to the Gadfly genome annotation at http://flybase.bio.indiana.edu/, is shown above the chromosome. Below each arm, the positions of PcG and trxG binding sites on *Drosophila* polytene chromosomes that correspond to predicted hits are shown (gray bars).

regulate other genes. The exact distance of each PRE hit from its closest gene, as well as Fbgn numbers and links to Flybase entries, are available at http://www.techfak.uni-bielefeld.de/∼marc/pre/). The 102 genes for which functional information is available comprise genes that are known to be regulated by the PcG, genes involved in determining cell identity, and several genes with unexpected functions.

We predict PRE/TREs in a number of genes whose regulation is known to depend on the PcG proteins. As expected, the high-scoring PRE/TREs of the BX-C and ANT-C, including that of the *Deformed* gene, were also significant hits in the genome search (Table 3; Figures 1D and 1E). Interestingly, we identify a candidate PRE/TRE in the *engrailed* gene, immediately adjacent to the published PRE/TRE that was used in our training set (Kassis, 1994) and closer to the transcription start. The published PRE/TRE has a score of 111.42, and thus falls below our cutoff score of 157. The predicted PRE/TRE has a score of 189.53. This strongly suggests that the full *engrailed* PRE/TRE is longer than anticipated. In addition, we predict PRE/TREs in three other genes that are regulated by the PcG: *caudal* (Moreno and Morata, 1999; Beuchle et al., 2001; Figures 3C–3G), *even skipped* (Smouse et al., 1988; Figure 3), and *knirps* (McKeon et al., 1994). The identification of candidate PRE/TREs in known PcG target genes demonstrates the reliability of the PRE/TRE prediction approach, and is a strong indication that the other genes at which we predict PRE/TREs are likely to be true targets of PcG/trxG regulation.

In further support of this argument is the identity of the genes themselves (Table 3; Supplemental Table S1). They include 26 transcription factors, of which 10 contain a homeobox (e.g., *homothorax* and *unc- 4*.) Apart from the known role of the PcG and trxG in maintaining cell identities established in the embryo, the PcG/trxG proteins also play distinct roles in maintaining gene expression patterns during oogenesis (Paro and Zink, 1993), in later larval development (Maurange and Paro, 2002), and in specifying neuronal identity (Smouse et

Table 3. Genes with Predicted PRE/TREs

| Score | Gene Name | Cyt | Score | Gene Name | Cyt | Score | Gene Name | Cyt |
|---|---|---|---|---|---|---|---|---|
| 447.845 | Ubx (bxd) | 89E | 202.514 | arm | 2B | 174.304 | pum | 85C |
| 387.172 | Dr | 99B | 202.245 | CG17062 | 50A | 172.208 | cato | 53A |
| 374.963 | CG4774 | 96E | 202.042 | Antp (upstream) | 84B | 171.904 | ben | 12D |
| 368.457 | H15 | 25E | 201.802 | CG3483 | 60C | 171.601 | vg | 49E |
| 341.501 | CG12540 | 12F | 201.724 | kni | 77E | 171.528 | EG:100G7.6 | 3C |
| 333.710 | slou | 93E | 201.513 | CG2111 | 9E | 171.160 | scrib | 97C |
| 330.146 | noc | 35B | 201.269 | CG12877 | 98B | 170.170 | BEST:GH14656 | 54B |
| 327.640 | CG32336 | 61F | 201.192 | dx | 6C | 169.456 | CG8861 | 85D |
| 316.161 | eve | 46C | 200.446 | CG32183 | 74F | 169.347 | CG12092 | 19E |
| 312.769 | salm | 32F | 198.537 | CG12524 | 67D | 169.099 | CG3650 | 60E |
| 308.982 | B-H2 | 16A | 196.928 | CG13438 | 57B | 169.069 | CG4712 | 49F |
| 304.813 | prod | 56A | 196.880 | svp | 87B | 168.668 | CG31714 | 31A |
| 291.670 | unc-4 | 16D | 196.791 | CG32169 | 73F | 168.643 | lilli | 23B |
| 284.565 | Idgf4 | 9A | 195.093 | Pdp1 | 66A | 168.300 | Tl | 97D |
| 277.114 | CG12454 | 12D | 194.295 | CG3843 | 88D | 168.061 | Cha | 91C |
| 276.013 | scrib | 97B | 193.938 | dhd | 4F | 167.943 | Hex-A | 8F |
| 272.709 | CG5070 | 15F | 193.791 | Nedd4 | 74D | 167.460 | CG15543 | 100A |
| 272.428 | CG15198 | 10B | 192.877 | CG3754 | 11D | 167.322 | Cp36 | 7F |
| 271.709 | Ef1alpha48D | 48D | 192.768 | B-H1 | 16A | 167.172 | CG15381 | 22C |
| 271.477 | CG7417 | 56C | 191.560 | mm | 54B | 167.038 | ninA | 67E |
| 267.298 | cad | 38E | 189.612 | CG7552 | 88D | 166.843 | Shab | 63A |
| 267.265 | CtBP | 87D | 189.525 | en | 48A | 166.560 | CG31613 | 39E |
| 263.459 | CG5988 | 17A | 189.388 | Abd-B (Fab-7) | 89E | 166.498 | Abd-B (promoter) | 89E |
| 256.118 | CG1961 | 10A | 188.056 | CG12516 | 98C | 166.334 | CG15880 | 21D |
| 254.976 | CG12661 | 8C | 187.712 | Abd-B (iab-9?) | 89E | 165.918 | CG2543 | 11B |
| 254.417 | CG9299 | 76B | 187.612 | pum | 85D | 165.541 | trh | 61C |
| 254.250 | fas | 50C | 187.494 | CG32169 | 73F | 165.381 | Sdc | 57E |
| 251.909 | Ets65A | 65A | 187.016 | CG17048 | 50A | 165.116 | fd96Ca | 96B |
| 251.383 | dnc | 3C | 186.544 | CG32465 | 84D | 164.549 | CG1139 | 62B |
| 249.920 | CG7710 | 91C | 186.359 | disco | 14B | 164.544 | Act57B | 57B |
| 249.714 | Ten-m | 79E | 185.931 | bi | 4C | 164.437 | l(2)gl | 21A |
| 245.480 | CG14503 | 55C | 185.589 | CG14355 | 88A | 164.394 | klg | 94D |
| 242.271 | cv-2 | 57D | 185.169 | CG5075 | 34A | 163.770 | CG3394 | 60C |
| 238.700 | CG3918 | 6B | 184.958 | CG7710 | 91C | 163.323 | tll | 100A |
| 237.947 | chic | 26A | 184.888 | CG13500 | 58B | 163.310 | Rh5 | 33B |
| 234.300 | bi | 4C | 183.500 | CG9469 | 96A | 162.558 | CG9896 | 59C |
| 232.155 | abd-A promoter | 89E | 183.319 | seq | 49F | 162.345 | CG15183 | 83F |
| 231.411 | CG13972 | 98A | 183.042 | pnr | 89B | 162.099 | CG8112 | 85A |
| 228.012 | beat-IIIb | 36B | 181.894 | CG7378 | 18A | 161.987 | CG2560 | 11A |
| 227.629 | CG15344 | 7E | 181.731 | CG18375 | 57D | 161.960 | CG1841 | 10E |
| 224.569 | Dfd | 84A | 179.974 | rg | 4F | 161.149 | CG5142 | 34A |
| 223.042 | aPKC | 51D | 179.941 | CG12657 | 7E | 161.111 | BcDNA:GH24095 | 86E |
| 222.346 | abd-A (iab-2) | 89E | 178.470 | CG12626 | 10A | 160.746 | EG:BACH61I5.1 | 2B |
| 219.772 | Ubx (bx) | 89D | 178.337 | CG31797 | 37D | 160.507 | CG7024 | 4D |
| 216.544 | Antp (P1) | 84B | 178.337 | sano | 55F | 160.064 | CG12425 | 98C |
| 215.350 | Obp85a | 85A | 178.276 | CG2750 | 11A | 159.609 | Ptp61F | 61F |
| 213.479 | tomosyn | 11B | 177.909 | CG1340 | 100A | 159.500 | CG12852 | 98C |
| 213.447 | eya | 26E | 177.029 | nAcRalpha-96Aa | 96A | 159.194 | CG14830 | 65F |
| 211.722 | nAcRalpha-96Aa | 96A | 176.542 | CG32090 | 68C | 158.854 | CG12814 | 85F |
| 211.571 | CG4645 | 11D | 176.427 | Map205 | 100D | 158.417 | CG1499 | 100B |
| 208.221 | Prat2 | 65D | 175.799 | CG32986 | 29F | 158.326 | shn | 47D |
| 208.049 | CG15467 | 4E | 175.346 | CG18812 | 43E | 157.940 | tinc | 90C |
| 204.324 | CG10570 | 37A | 175.304 | CG10908 | 22B | 157.716 | CG6125 | 88F |
| 204.093 | CG9425 | 70F | 175.177 | lbe | 93E | 157.607 | CG3123 | 23B |
| 203.981 | hth | 86C | 175.118 | Rx | 57B | 157.072 | CG3729 | 2B |
| 203.424 | Map205 | 100D | 174.634 | vvl | 65D | 174.304 | pum | 85C |

The gene closest to each of the 167 top scoring hits, according to BDGP release 3.1, is shown. More information, such as links to the sequence of each hit and the Flybase entry for each gene, is available at http://www.techfak.uni-bielefeld.de/~marc/pre/. The cytological position of each gene is also given. The PRE/TRE prediction algorithm is available at http://bibiserv.techfak.uni-bielefeld.de/predictor/.

al., 1988). Concordant with these expectations, we predict 17 PcG/trxG target genes with a role in embryonic patterning, 10 genes with a role in oogenesis, and 27 that are involved in cell fate specification at later larval stages. Remarkably, this latter group includes 13 genes that are involved in nervous system development, of which 10 have a role in eye development (e.g., *seven up* and *eyes absent*). Thus, it appears that the PcG and trxG directly regulate genes involved in a surprisingly wide spectrum of developmental pathways.

Another interesting class is those genes that are involved in regulating cell proliferation, and for which mutations generate tumors (Table 3). These include the tumor suppressors *lethal(2) giant larvae* and *proliferation disrupter*. We also identify two p53-like transcription factors (*bifid* and *H15*). These findings are striking given

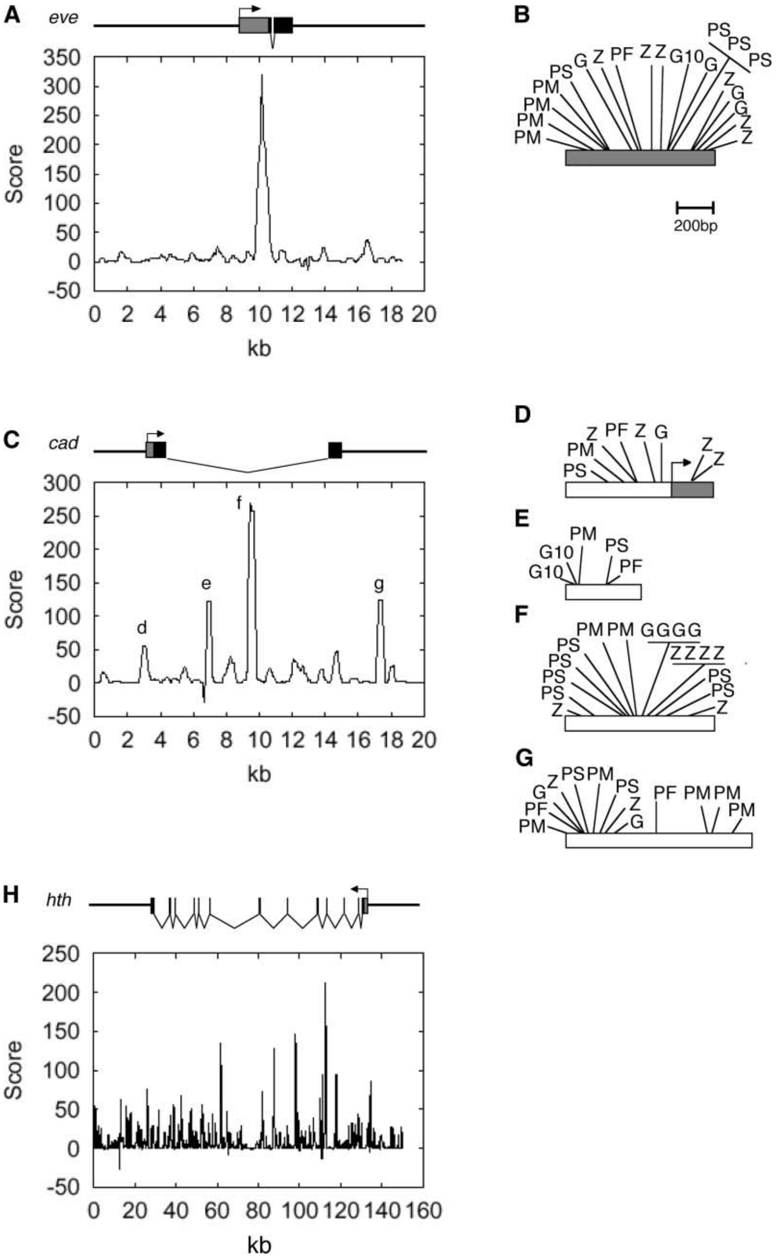

Figure 3. Predicted PRE/TREs in *even skipped, caudal*, and *homothorax*

(A) *even skipped* (46C).

(B) Predicted *even skipped* PRE/TRE (chromosome 2R; 5,039,100–5,039,899). Underlined motifs indicate consecutive repeats spaced at intervals of 4 bp or less.

(C) *caudal* (38E).

(D–G) Predicted PRE/TREs in the *caudal* gene.

(D) Chromosome 2L; 20,739,900–20,740,699.

(E) 20,744,100–20,744,499.

(F) 20,746,400–20,747,199.

(G) 20,754,500–20,755,500.

(H) *homothorax* (86C). Transcription start sites (arrows), noncoding (stippled boxes), and coding exons (black boxes) are indicated.

the fact that the PcG and trxG play a role in the control of cellular proliferation and tumorigenesis in vertebrates (Jacobs and van Lohuizen, 2002). So far, a connection between the PcG/trxG and tumorigenesis in *Drosophila* has not been reported. The genes we have identified may provide this link.

In summary, genome-wide PRE/TRE prediction identifies candidate PRE/TREs at high resolution, providing not only the exact location of PRE/TREs in known PcG targets but also identifying additional putative target genes whose functions give a clue to the diverse roles of the PcG and trxG in epigenetic regulation.

### Cooperation between PRE/TREs at the Promoter and at a Distance

Known PRE/TREs usually occur in pairs or groups. To determine whether predicted PRE/TREs fall into a similar pattern, we examined the regions around 70 of the high-scoring hits to map additional peaks that score below 157 but which may nevertheless have PRE/TRE function (Figure 3). We found that only 10% of predicted PRE/TREs occur as single peaks, like that of the *even skipped* gene (Figure 3A). These peaks are, without exception, positioned at or near (within 800 bp) of the transcription start site, and all except one are associated with short genes (<3.5 kb). The other 90% of hits we examined are accompanied by one or more additional peaks, like those of the *caudal* gene (Figure 3C). In all cases, we found a peak scoring 50 or more positioned at or near the transcription start. The most extreme example of PRE/TRE grouping we found was the *homothorax* gene (Figure 3H).

Closer examination of these promoter peaks revealed that they are not simply composed of Z and GAF binding sites, but all contain PHO motifs as well (Figure 3D and see below). For example, in the *caudal* gene (Figures

3C–3G), the promoter peak is small, and is not significant in the context of the genome. However, in the context of the gene, the precise placing of this peak at the promoter, its motif composition (Figure 3D), and the presence of other high-scoring peaks nearby strongly suggest that this sequence may function to bring PcG and trxG proteins bound at the other stronger PRE/TREs (Figures 3E–3G) into the vicinity of the promoter.

In conclusion, we observe a PRE/TRE peak at the promoters of all genes we examined, and we propose that the PcG and trxG are brought to the promoter by direct binding to these PRE/TREs, which is stabilized in most cases by other PRE/TREs nearby.

### Candidate PRE/TREs Are Bound and Regulated by the PcG In Vivo

To determine whether the PRE/TREs we have predicted are indeed targets for PcG regulation in vivo, we used chromatin immunoprecipitation (ChIP) to detect enrichment for PC binding in *Drosophila* Schneider cells. We tested 43 candidate PRE/TREs from genes with known function, chosen to evenly represent the full range of scores covered by our 167 hits (Figure 4A). As controls we included six known PRE/TREs for which PC enrichments are published (Strutt and Paro, 1997; Strutt et al., 1997). Five of these (*bxd, bx, iab-2, abd-A* promoter, and *engrailed*) show enrichment for PC binding in SL-2 cells. For the sixth, *Fab-7*, it has been shown that although this element is a bona fide PRE/TRE in transgenic assays, it is not enriched in SL2 cells, presumably because these cells represent a single tissue type in which *Fab-7* is not bound by PC protein (Strutt et al., 1997). Thus, this fragment serves as a negative control. As additional negative controls we tested 43 fragments that do not contain PRE/TRE sequences.

Twenty-nine of the candidate PRE/TREs showed higher than 2-fold enrichment for PC binding (Figure 4), whereas none of the negative control fragments were enriched above 1.4-fold. Fourteen candidate PRE/TRE fragments were enriched less than 2-fold, a similar level to that observed for *Fab-7*. Thus, these fragments might not be PRE/TREs. Alternatively, these 14 fragments may be, like *Fab-7*, true PRE/TREs that are not enriched for PC binding in these cells. Eighteen fragments were enriched at similar levels to the *abd-A* promoter and *engrailed* (2- to 4-fold), while 11 fragments fall into the range of *bxd*, *bx*, and *iab-2* (4- to 12-fold). The fact that the majority of predicted PRE/TREs are enriched for PC binding at levels comparable to known PRE/TREs strongly suggests that they are targets for PcG regulation in vivo.

To further test predicted PRE/TREs by independent means, we selected four of them covering a range of enrichments in the ChIP experiment for transgenic analysis: *proliferation disrupter* (*prod*; 1.6-fold enrichment), *caudal* (*cad*; 2-fold), *atypical Protein Kinase C* (*aPKC*; 2-fold), and *eyes absent* (*eya*; 4.5-fold). Three of the predicted PRE/TREs are in introns of their associated genes. The exception is *proliferation disrupter*, for which the PRE/TRE we tested is 2 kb downstream of the annotated gene end (there is a second peak near the promoter). For each sequence, a 3 kb fragment containing the predicted PRE/TRE was amplified by PCR from genomic DNA and cloned into the pUZ P element vector (Cavalli and Paro, 1998) upstream of the *miniwhite* reporter gene, which gives a red eye color. These constructs were used to generate transgenic flies, and the effects of the candidate PRE/TRE on *miniwhite* were analyzed.

Known PRE/TREs induce pairing-sensitive repression and variegation of *miniwhite* in a manner that is genetically dependent on the PcG and trxG. Repression is typically stronger in flies raised at 25°C compared to 18°C (Fauvarque and Dura, 1993; Kassis, 1994). Consistent with these expectations, all the constructs analyzed showed strong repression of *miniwhite* activity in the heterozygote state (compare Figure 4B with Figures 4D, 4F, 4H, and 4J), and several lines for each construct showed variegation (not shown). For all lines, silencing was stronger in flies raised at 25°C compared to 18°C (not shown). Furthermore, for *aPKC, prod*, and *cad*, in several lines the eye color of homozygotes was similar to, or lighter than, that of heterozygotes, indicating pairing-sensitive repression (compare top [heterozygote] and bottom [homozygote] panels in Figures 4F–4K). Finally, several lines for each construct showed loss of *miniwhite* repression in a *ph*$^{410}$ heterozygous mutant background (Figures 4L, 4N, 4P, and 4R), compared to a genetic background in which *ph* was wild-type (Figures 4M, 4O, 4Q, and 4S). Similar results were obtained with the *Pc*$^{XL5}$ mutant allele (not shown).

In conclusion, this analysis indicates that all four of the PRE/TREs we tested are regulated by the PcG in vivo. This, together with the ChIP analysis described above, confirms the validity of the PRE/TRE prediction approach.

### Three PRE/TRE Sequence Motifs

The detection of PRE/TREs by prediction generates a large data set that can be used to search for further common sequence features. To this end, we scanned the 30 highest scoring PRE/TRE hits for motifs that occur significantly more often in PRE/TREs than in randomly generated sequence (Bailey and Elkan, 1994). We found five significant motifs, shown as sequence logos in Figure 5 (Schneider and Stephens, 1990). Not surprisingly, but reassuringly, we found two known motifs, the GAF and PHO binding sites (Figures 5A and 5B). We did not find the Z binding motif by this analysis, although it occurs as frequently as GAF in the 30 sequences we analyzed (Figure 5F; http://www.techfak.uni-bielefeld.de/∼marc/pre/). This is probably due to the shortness and degeneracy of the Z motif (Table 1), and suggests that other such short motifs will also be missed by this approach.

Nevertheless, we found three additional motifs. The first, which we call GTGT (Figure 5C), is found several times in 14 of the sequences. A search for factors that bind such a site using the TRANSFAC database (http://transfac.gbf.de/TRANSFAC/) was unsuccessful. The second motif, poly T (Figure 5C), is found several times in almost all 30 PRE/TRE sequences analyzed. Some variants of this site match the binding consensus for the hunchback protein, which has been shown to be an early regulator at some PRE/TREs (Qian et al., 1991). The third motif, TGC triplets, occurs several times in 13 of the PRE/TRE sequences. No binding factor for this sequence has yet been identified.

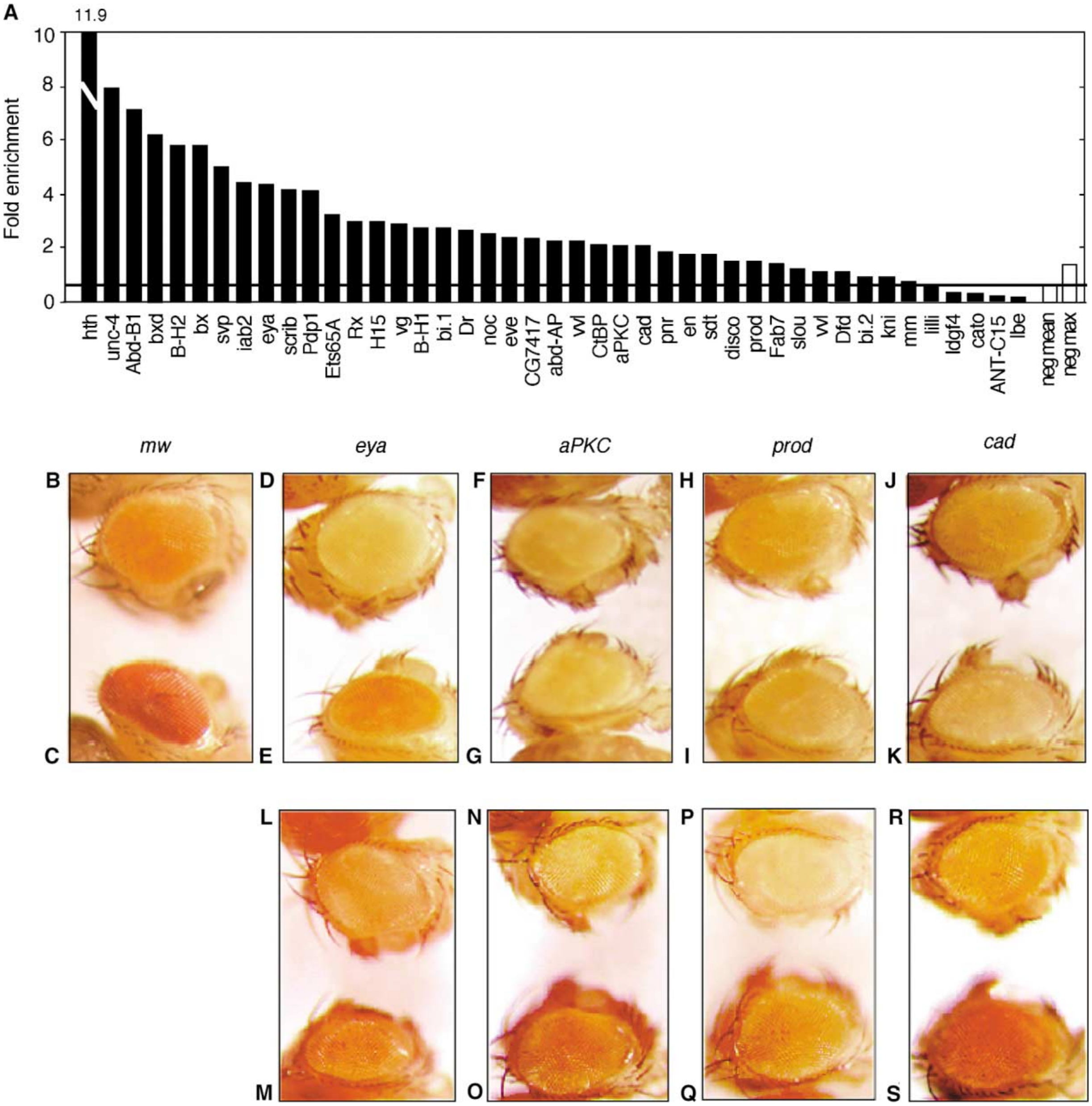

**Figure 4. Polycomb Group Proteins Bind and Regulate Predicted PRE/TREs In Vivo**

(A) Chromatin immunoprecipitation was performed to assess enrichment of PC on candidate PRE/TREs in *Drosophila* Schneider cells. Abd-B1: peak in BX-C at 40 kb (Figure 1D); abd-AP: *abd-A* promoter. ANT-C15, peak in ANT-C at 15 kb (Figure 1E). Fold enrichment of PC is shown for each candidate PRE/TRE (black bars). The mean and maximum enrichments for 43 control non-PRE/TRE fragments are shown (white bars). Similar results were obtained in three independent experiments; the figure shows one representative experiment.

(B–S) Effects of four candidate PRE/TREs on *miniwhite* expression in transgenic flies. Flies are heterozygous for the transgene.

(B and C) *miniwhite* transgene without flanking PRE/TRE. Intensity of eye color indicates *miniwhite* expression level.

(B, D, F, H, and J) Heterozygotes.

(C, E, G, I, and K) Homozygotes. Flies were raised at 25°C.

(L–S) Effects of $ph^{410}$ mutation on *miniwhite* at 18°C.

(L, N, P, and R) Heterozygote transgene in wild-type *ph* background.

(M, O, Q, and S) Heterozygote transgene in $ph^{410}$ heterozygote mutant background.

To further examine these three motifs, we evaluated motif occurrence in all 167 predicted PRE/TREs and in the promoter peaks described above. Figures 5F and 5G show the percentage of sequences that have a given motif at least once. In contrast to the known GAF, Z, and PHO motifs, our three motifs each occur in only a subset of predicted and known PRE/TREs, and do not occur significantly together. These motifs may thus each define a subclass of PRE/TREs. Consistent with this idea, we found that some of the lowest scoring known PRE/TRE sequences from Table 2 indeed contain one or more of our motifs (see http://www.techfak.uni-bielefeld.de/marc/pre/ for full listings).

Although we found no correlation between particular

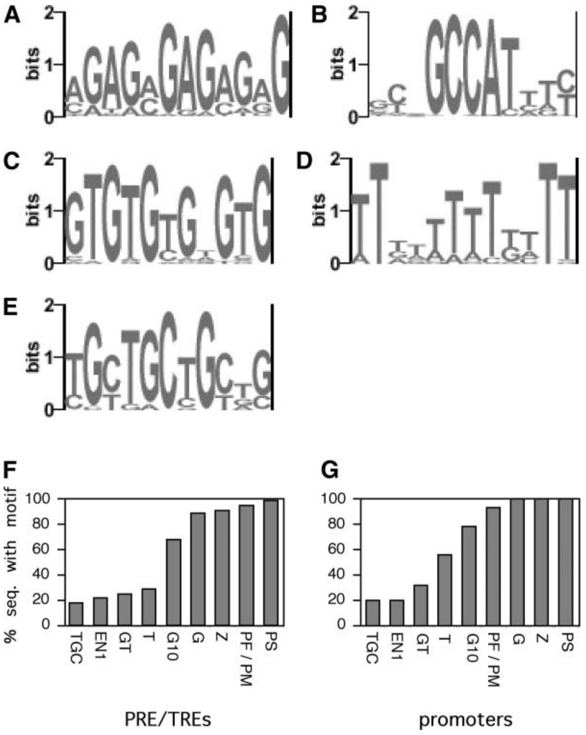

Figure 5. PRE/TRE Motifs and Their Occurrence in PRE/TREs and Promoters

Motifs occurring significantly in the top 30 scoring PRE/TREs are represented as sequence logos. For a given position in the logo, the degree of conservation is reflected by the total height of the drawing (i.e., the total information, measured in bits; Schneider and Stephens, 1990). The frequency with which each nucleotide occurs at that position is represented by the proportion of the total height that is occupied by each letter.

(A) GAGA (E = $3.0 \times 10^{-116}$).

(B) PHO (E = $2.8 \times 10^{-42}$).

(C) GTGT (E = $5.2 \times 10^{-23}$).

(D) Poly T (E = $2.4 \times 10^{-14}$).

(E) TGC triplets (E = $3.7 \times 10^{-6}$).

(F and G) Single motif occurrences in predicted PRE/TREs and at promoter peaks. The percentage of sequences in which a motif occurs at least once is shown.

(F) Known and predicted PRE/TREs. This analysis comprises the 167 predicted PRE/TRE hits and the 8 PREs of the training set that score below 157.

(G) Promoters. This analysis comprises 16 of the 167 hits that lie at the promoter of the nearest gene, and 25 additional promoter peaks scoring 50 or above.

sites and high scores, we did find a negative correlation between numbers of GAF/Z and PHO sites (a correlation coefficient of −0.78, indicating that when many GAF/Z sites are present, there are few PHO sites, and vice versa). This suggests that each PRE/TRE may have a preferred ground state, in which it is either predisposed to silencing (many PHO sites) or to activation (many GAF/Z sites).

In summary, this analysis identifies three motifs that occur significantly in association with known PRE/TRE motifs. Further functional characterization of these motifs and the proteins that bind them may contribute to a more complete definition of the sequence requirement for PRE/TRE function, and of subclasses of PRE/TREs.

## Discussion

We have determined criteria for PRE/TRE prediction and used them to search the *Drosophila* genome. Several lines of evidence indicate that these predictions are meaningful. The performance of the prediction method on the BX-C and ANT-C, where it correctly identifies eight PRE/TREs that were not in the training set, is compelling evidence that PRE/TREs can indeed be identified by their sequence. Furthermore, in the genome search, our predictions show an excellent correspondence with previously mapped cytological locations of PcG/trxG binding. This evidence, coupled with the fact that PC binds many candidate PRE/TREs in vivo, and tested fragments behave as PRE/TREs in transgenic assays, argues strongly in favor of a real biological role for the predicted PRE/TREs as epigenetic regulators.

As candidates for regulation, we identified the closest annotated gene to each predicted PRE/TRE. Although it is possible that more distant genes may in fact be regulated by the sequences we have identified, these genes are good candidates for the following reasons: first, we predict PRE/TREs in five genes that are known to be regulated by the PcG and trxG. Second, in 10% of 70 candidates we examined in detail, a single PRE/TRE is found at the promoter. It is difficult to imagine how such a PRE/TRE could regulate another gene. Finally, in the other 90%, the main hit is accompanied by a second peak at or near the promoter. This configuration is seen in all known PRE/TRE-regulated genes, without exception, and therefore we conclude that the possession of a PRE/TRE peak at the promoter is a hallmark of regulation (the complete PRE/TRE configuration of any sequence of interest can be examined by using our prediction program at http://bibiserv.techfak.uni-bielefeld.de/predictor).

Our study offers four main contributions to the understanding of PRE/TRE function. First, we define a larger set of sequences, which will facilitate the more complete definition of PRE/TRE sequence requirements. We identify three motifs that may contribute to this goal. We expect that the definition of the minimal requirement for PRE/TRE function will not be a trivial task. Analysis of motif composition and order in the 167 predicted PRE/TREs revealed that there is a great diversity of patterns, with no preferred linear order. It is possible that each different pattern of motifs reflects a subtly different function. On the other hand, the concept of a linear order of motifs may well be irrelevant, because these elements operate in the three-dimensional context of chromatin. The fact that such a diversity of PRE/TRE designs exist indicates that the vast majority of them would defy detection by conventional pattern-finding algorithms, and underlines the advantages of the approach we describe here.

Although we found no linear constraints on motif order, the fact that only motif pairs, and not single motifs, were able to identify PRE/TREs strongly suggests that this close spacing of sites has functional significance. Multiple sites may work in concert, to promote cooperative binding of similar proteins (e.g., repeated PHO sites) or to provoke competition between dissimilar proteins (e.g., closely spaced GAF and PHO sites). In addition, in chromatin, only a subset of sites will be exposed and optimally available for binding at any one time, while others will be occluded by nucleosomes. The trxG includes nucleosome remodeling machines, raising the intriguing possibility that remodeling of PRE/TREs in chromatin may contribute to epigenetic switching by exposing different sets of protein binding sites.

Second, we observe a PRE/TRE peak at the promoter of all the genes we examined. This strongly suggests that promoter binding is a general principle of PRE/TRE function. It has been reported that PcG proteins can interact with general transcription factors (Saurin et al., 2001; Breiling et al., 2001). It has hitherto been unclear whether the observed PcG/trxG binding at promoters of the genes they regulate (Strutt et al., 1997; Orlando et al., 1998) is mediated indirectly via such an interaction, or whether the PcG and trxG bind directly to PRE/TREs at the promoters. The high scores we observe at promoters favor the latter interpretation.

Third, we show that in most cases, PRE/TREs do not occur in isolation, but are accompanied by one or more other peaks nearby (Figure 3). These grouped PRE/TREs may create multiple attachment sites for PcG and trxG proteins, which come together to build a fully operational complex at the promoter. Alternatively, grouped PRE/TREs may be individually regulated by tissue-specific enhancers as in the BX-C. Thus, each of the many PRE/TREs of the *homothorax* gene (Figure 3H) may interact with the promoter PRE/TRE in different tissues. This idea is consistent with the fact that *homothorax* has specific roles in diverse developmental processes (Starling Emerald and Cohen, 2001).

Finally, we expand the current list of about ten PcG/trxG target genes to over 150 genes, identifying candidates for epigenetic regulation. The genes thus identified encompass every stage of development, suggesting that the PcG/trxG are global regulators of cellular memory. Experiments to further investigate and compare this regulation for individual genes are currently underway.

## Experimental Procedures

### Accession Numbers and Coordinates of PRE/TRE and Non-PRE/TRE Sequences

PRE/TREs in the Bithorax complex: U31961: *iab-8* (Barges et al., 2000), 59,200–64,582. *Fab-7* (Mihaly et al., 1997), 82,553–86,163. *Mcp* (Busturia et al., 1997), 109,688–114,288. *iab-2* (Shimell et al., 2000), 170,000–173,000. *bxd*, 218,249–219,807 (Chan et al., 1994). PRE/TREs in the Antennapedia complex (Gindhart and Kaufman, 1995); AE001573: *Scr10Xba.1*, 161,142–163,700. *Scr10Xba.2*, 169,500–170,718. *Scr8.2Xba*, 220,703–226,000. *engrailed* PRE/TREs (Kassis, 1994; Kassis et al., 1989) *D. melanogaster*, M29285, 459–2,003. *D. virilis*, M29286, 487–2,327. *polyhomeotic* PRE/TREs (Fauvarque and Dura, 1993; Bloyer et al., 2003), Z98269, proximal (*ph p*), 14,651–16,619; distal (*ph d*) 2,069–4,446. Non-PRE/TREs: these sequences contain the upstream regulatory region of each gene up to the transcription start site, obtained from Flybase (http://flybase.bio.indiana.edu/): *hsc70-1* AE003536, 224,561–225,431. *hsc70-2* AE003698, 123,759–124,188. *hsc70-3* AE003487, 88,268–93,230. *hsc70-4* AE003708, 44,714–45,244. Heat shock genes in AE003552: *hsp22*, 184,572–184,940. *hsp23*, 190,362–192,821. *hsp26*, 188,111–189,055. *hsp27*, 193,690–195,000. *hsp67B*, 183,240–183,854. *hsp68* AE003746, 25,889–27,105. *hsp83* AE003477, 128,230–128,928. *linotte* AE003662, 14,009–15,617. *Polycomb* AE003594, 44,604–45,155. *rosy* AE003698, 110,679–111,780. *white* AE003425, 151,170–154,200. *yellow* AE003417, 105,668–112,676.

### Computational Methods

Motifs and their reverse complements are translated into regular expressions (patterns) by replacing degenerate nucleotides with corresponding character sets. For motif G10, all possible one-mismatch patterns are added. This procedure results in 54 different patterns. The genome (or any other input sequence) is searched for single pattern occurrences. When in motif pair mode, single occurrences are combined to pairs if their distance does not exceed 220 bp. Single occurrences or pair occurrences are weighted and summed up in a 500 bp window that is moved over the input sequence in steps of 100 bp. Motif weights are defined as log-odds scores: the score S(p) for a motif or motif pair p is defined as $S(p) = \ln(f(p|PRE)/f(p|non\text{-}PRE))$ where f(p|) is the number of occurrences of p per 1 kb in the PRE/TRE or non-PRE/TRE training sequences, respectively. To get a significance estimation of scores, we searched a random sequence of size 100 times the *Drosophila melanogaster* genome (11.7 Gb) and used the empirical score distribution to define score thresholds for E values of 1.0, 0.1, and 0.05 in the actual genome search. The random sequence was generated i.i.d. according to the nucleotide distribution of the *D. melanogaster* genome. The time and space consumption of the algorithm is linear with respect to the length of the input sequence and (if in motif pair mode) quadratic with respect to the number of single motifs. The algorithm is implemented in C++ and uses the LEDA library and the C regexp utility. Searching the *D. melanogaster* genome in pair mode takes 6 min on an UltraSparc III 900MHz. The program is available on the Bielefeld Bioinformatics Server (http://bibiserv.techfak.uni-bielefeld.de/predictor) for the generation of score plots for uploaded input sequences.

### Immunoprecipitation of Crosslinked Chromatin from SL-2 Cells

Crosslinking, immunoprecipitation with antibodies against PC, and preparation of PCR material for use as hybridization probe were as described previously (Strutt et al., 1997). Radioactive probes were prepared from PC IP, mock IP (without antibody), and genomic DNA cut with HaeIII restriction enzyme using the Ready Prime kit (Amersham). Probes were hybridized to S&S Nytran Supercharge nylon filters (Scleicher & Schuell). Filters were spotted using an S&S Minifold I dot blotting apparatus (Schleicher & Schuell), with PCR fragments approximately 1 kb in length corresponding to the 43 predicted PRE/TRE fragments or negative control fragments, nine of which correspond to low-scoring regions of genes for which we predict PRE/TREs, and 32 of which were identified from the BX-C map as sequences that do not contain PRE/TREs (Strutt et al., 1997). These PCR products were amplified from *Drosophila* genomic DNA using primers designed according to the genomic sequence. Hybridization signals were quantified by phosphorimager analysis. Variation in fragment concentration at each spot was corrected using the signal from genomic DNA probe, and the enrichment of PC IP with respect to mock IP was calculated.

### DNA Constructs

For the predicted PRE/TREs of *caudal*, *eyes absent*, *atypical protein kinase C*, and *proliferation disrupter*, a 3 kb fragment containing the highest PRE/TRE score was amplified from *Drosophila* genomic DNA by PCR using the Expand High Fidelity PCR kit (Roche). The genomic coordinates of the fragments thus amplified are as follows: *caudal*: 2L, 20,745,045–20,748,082; *eyes absent*: 2L, 6,536,321–6,539,275; *atypical protein kinase C*: 2R, 10,021,402–10,024,422; *proliferation disrupter*: 2R, 14,029,608–14,032,676. Inclusion of NotI and SpeI restriction sites in the PCR primers enabled directional cloning of each PRE/TRE into the pUZ vector (Cavalli and Paro, 1998) upstream of the *LacZ* and *miniwhite* reporter genes. The PCR fragments are oriented such that the direction of transcription of the endogenous gene is the same as that of *LacZ* and *miniwhite*.

### Generation and Analysis of Transgenic Flies

Injections were performed by standard procedures (Voie and Cohen, 1998) into a homozygous $w^{1118}$; +/+; +/+ strain. Between five and eight independent transgenic lines were analyzed for each construct. In all experiments, the eye color of flies of the same age and sex were compared. To test for pairing-sensitive repression of *miniwhite* expression, stocks carrying the transgene on the second or third chromosome over a balancer (CyO or TM3,Sb) were self-crossed and homozygous progeny were compared with their heterozygote siblings. To avoid the potential effects of balancer chromosomes on eye color, balanced and unbalanced heterozygotes were compared for each line. To test the effects of the $ph^{410}$ mutation on *miniwhite* expression, male flies from homozygote stocks of each transgenic line (second or third chromosome) were crossed to virgins from a homozygous $w,ph^{410}$ strain, and to $w^{1118}$ virgins. Male progeny from each cross were again crossed to $w,ph^{410}$ and $w^{1118}$ virgins. Female progeny from the $w,ph^{410}$ and $w^{1118}$ crosses were compared with each other. To test the effect of the $Pc^{XL5}$ mutation, flies from each transgenic line were crossed to $Pc^{XL5}$/TM3Sb stocks.

## Acknowledgments

We thank Heidi Ehret for technical assistance and Marc Hild for help with chromatin immunoprecipitations. We also extend special thanks to Christian Popp for analysis of transgenic lines. The work of L.R. was supported by an RTN grant from the European Community. R.P. received support from the European Community, the Deutsche Forschungsgemeinschaft, and the Fonds der Chemischen Industrie. J.-M.D. received support from the Association pour la Recherche sur le Cancer (ARC n° 5658).

Received: April 16, 2003
Revised: July 17, 2003
Accepted: October 8, 2003
Published: November 3, 2003

## References

Bailey, T.L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, AAAI Press, Menlo Park, CA.

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.

Beuchle, D., Struhl, G., and Müller, J. (2001). Polycomb group proteins and heritable silencing of *Drosophila Hox* genes. Development *128*, 993–1004.

Bloyer, S., Cavalli, G., Brock, H.W., and Dura, J.M. (2003). Identification and characterization of polyhomeotic PREs and TREs. Dev. Biol. *261*, 426–442.

Breiling, A., Turner, B., Bianchi, M.E., and Orlando, V. (2001). General transcription factors bind promoters repressed by Polycomb group proteins. Nature *412*, 651–655.

Brown, J.L., Mucci, D., Whiteley, M., Dirksen, M.L., and Kassis, J.A. (1998). The *Drosophila* Polycomb group gene pleiohomeotic encodes a DNA binding protein with homology to the transcription factor YY1. Mol. Cell *1*, 1057–1064.

Busturia, A., Wightman, C.D., and Sakonju, S. (1997). A silencer is required for maintenance of transcriptional repression throughout *Drosophila* development. Development *124*, 4343–4350.

Cavalli, G., and Paro, R. (1998). The *Drosophila Fab-7* chromosomal element conveys epigenetic inheritance during mitosis and meiosis. Cell *93*, 505–518.

Chan, C.S., Rastelli, L., and Pirrotta, V. (1994). A Polycomb response element in the *Ubx* gene that determines an epigenetically inherited state of repression. EMBO J. *13*, 2553–2564.

Fauvarque, M.-O., and Dura, J.-M. (1993). *polyhomeotic* regulatory sequences induce developmental regulator-dependent variegation and targeted P-element insertions in *Drosophila*. Genes Dev. *7*, 1508–1520.

Fritsch, C., Brown, J.L., Kassis, J.A., and Müller, J. (1999). The DNA-binding polycomb group protein pleiohomeotic mediates silencing of a *Drosophila* homeotic gene. Development *126*, 3905–3913.

Gindhart, J.G., and Kaufman, T.C. (1995). Identification of Polycomb and trithorax group responsive elements in the regulatory region of the *Drosophila* homeotic gene *Sex combs reduced*. Genetics *139*, 797–814.

Hur, M.W., Laney, J., Jeon, S.H., Ali, J., and Biggin, M.D. (2002). Zeste maintains repression of *Ubx* transgenes: support for a new model of Polycomb repression. Development *129*, 1339–1343.

Jacobs, J.J., and van. Lohuizen, M. (2002). Polycomb repression: from cellular memory to cellular proliferation and cancer. Biochim. Biophys. Acta *1602*, 151–161.

Kassis, J.A. (1994). Unusual properties of regulatory DNA from the *Drosophila engrailed* gene: three "pairing-sensitive" sites within a 1.6-kb region. Genetics *136*, 1025–1038.

Kassis, J.A., Desplan, C., Wright, D.K., and O'Farrell, P.H. (1989). Evolutionary conservation of homeodomain-binding sites and other sequences upstream and within the major transcription unit of the *Drosophila* segmentation gene *engrailed*. Mol. Cell. Biol. *9*, 4304–4311.

Maurange, C., and Paro, R. (2002). A cellular memory module conveys epigenetic inheritance of *hedgehog* expression during *Drosophila* wing imaginal disc development. Genes Dev. *16*, 2672–2683.

McKeon, J., Slade, E., Sinclair, D.A., Cheng, N., Couling, M., and Brock, H.W. (1994). Mutations in some Polycomb group genes of *Drosophila* interfere with regulation of segmentation genes. Mol. Gen. Genet. *244*, 474–483.

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.

Mihaly, J., Mishra, R.K., and Karch, F. (1998). A conserved sequence motif in Polycomb-response elements. Mol. Cell *1*, 1065–1066.

Moreno, E., and Morata, G. (1999). *Caudal* is the Hox gene that specifies the most posterior *Drosophila* segment. Nature *400*, 873–877.

Orlando, V. (2003). Polycomb, epigenomes, and control of cell identity. Cell *112*, 599–606.

Orlando, V., Jane, E.P., Chinwalla, V., Harte, P.J., and Paro, R. (1998). Binding of trithorax and Polycomb proteins to the bithorax complex: dynamic changes during early *Drosophila* embryogenesis. EMBO J. *17*, 5141–5150.

Paro, R., and Zink, B. (1993). The Polycomb gene is differentially regulated during oogenesis and embryogenesis of *Drosophila melanogaster*. Mech. Dev. *40*, 37–46.

Qian, S., Capovilla, M., and Pirrotta, V. (1991). The *bx* region enhancer, a distant cis-control element of the *Drosophila Ubx* gene and its regulation by *hunchback* and other segmentation genes. EMBO J. *10*, 1415–1425.

Ringrose, L., Chabanis, S., Angrand, P.O., Woodroofe, C., and Stewart, A.F. (1999). Quantitative comparison of DNA looping *in vitro* and *in vivo*: chromatin increases effective DNA flexibility at short distances. EMBO J. *18*, 6630–6641.

Saurin, A.J., Shao, Z., Erdjument-Bromage, H., Tempst, P., and Kingston, R.E. (2001). A *Drosophila* Polycomb group complex includes Zeste and dTAFII proteins. Nature *412*, 655–660.

Schneider, T.D., and Stephens, R.M. (1990). Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. *18*, 6097–6100.

Shimell, M.J., Peterson, A.J., Burr, J., Simon, J.A., and O'Connor, M.B. (2000). Functional analysis of repressor binding sites in the *iab-2* regulatory region of the *abdominal-A* homeotic gene. Dev. Biol. *218*, 38–52.

Simon, J., Chiang, A., Bender, W., Shimell, M.J., and O'Connor, M. (1993). Elements of the *Drosophila* bithorax complex that mediate repression by Polycomb group products. Dev. Biol. *158*, 131–144.

Smouse, D., Goodman, C., Mahowald, A., and Perrimon, N. (1988). *polyhomeotic*: a gene required for the embryonic development of axon pathways in the central nervous system of *Drosophila*. Genes Dev. *2*, 830–842.

Starling Emerald, B., and Cohen, S.M. (2001). Limb development: getting down to the ground state. Curr. Biol. *11*, R1025–R1027.

Strutt, H., and Paro, R. (1997). The polycomb group protein complex of *Drosophila melanogaster* has different compositions at different target genes. Mol. Cell. Biol. *17*, 6773–6783.

Strutt, H., Cavalli, G., and Paro, R. (1997). Co-localization of Polycomb protein and GAGA factor on regulatory elements responsible for the maintenance of homeotic gene expression. EMBO J. *16*, 3621–3632.

Voie, A.-M., and Cohen, S. (1998). Germline transformation of *Drosophila melanogaster*. In Cell Biology: A Laboratory Handbook, Second Edition, Volume 3 (San Diego, CA: Academic Press), pp. 510–517.

Zink, B., Engstrom, Y., Gehring, W.J., and Paro, R. (1991). Direct interaction of the Polycomb protein with *Antennapedia* regulatory sequences in polytene chromosomes of *Drosophila melanogaster*. EMBO J. *10*, 153–162.

Zuckerkandl, E. (1999). Sectorial gene repression in the control of development. Gene *238*, 263–276.