Genomics 96 (2010) 17–26

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Prediction of Polycomb target genes in mouse embryonic stem cells

Yingchun Liu [1], Zhen Shao [1], Guo-Cheng Yuan *

Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, and Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA

ARTICLE INFO

Article history:
Received 8 January 2010
Accepted 23 March 2010
Available online 29 March 2010

Keywords:
Polycomb
Epigenetics
Transcription factor sequence motif
ChIP-chip
ChIP-seq
Stem cell
Mouse

ABSTRACT

Polycomb group (PcG) proteins are important epigenetic regulators, yet the underlying targeting mechanism in mammals is still poorly understood. We have developed a computational approach to predict genome-wide PcG target genes in mouse embryonic stem cells. We use TF binding and motif information as predictors and apply the Bayesian Additive Regression Trees (BART) model for classification. Our model has good prediction accuracy. The performance can be mainly explained by five TF features (Zf5, Tcfcp2l1, Ctcf, E2f1, Myc). Our analysis of H3K27me3 and gene expression data suggests that genomic sequence is highly correlated with the overall PcG target plasticity. We have also compared the PcG target sequence signatures between mouse and *Drosophila* and found that they are strikingly different. Our predictions may be useful for *de novo* search for Polycomb response elements (PRE) in mammals.

Introduction

PcG proteins were first discovered in *Drosophila* for their ability to silence Hox genes [1]. Subsequent studies have shown that PcG proteins regulate a large number of genes and play an important role in early development in many species, including human [2–5]. A major function of PcG proteins is to repress the transcriptional activities of their target genes through tri-methylation of the histone H3 on lysine 27 residue (H3K27me3). The mammalian PcG proteins form two major complexes: Polycomb repressive complex 1 (PRC1) and Polycomb repressive complex 2 (PRC2). Both complexes are required for gene silencing [3,5,6], whereas the histone methyltransferase activity is carried out by two PRC2 members: Ezh2 and Ezh1 [3,7,8]. Recent studies have shown that PcG proteins are critical for the maintenance of stem cell identity and cell differentiation [9,10]. In mammalian ESCs, PcG proteins bind to several thousand genes, typically localizing in promoter regions [7,11]. Many of these target genes (hereinafter, target genes mean physical binding targets) are developmental regulators [7,9–15], which are repressed in ESCs but can be activated during cell-differentiation as the PcG complexes disassociate from the promoters.

A fundamental yet unresolved question is how PcG proteins are recruited to specific target genes. The question is relatively well-understood in *Drosophila* but remains unclear in mammals [3,5]. In *Drosophila*, it is known that the DNA sequence plays an important role in PcG targeting. In particular, the PcG proteins are recruited to specific DNA regulatory elements called Polycomb Response Elements (PREs), which can recruit PcG binding independent of the surrounding environment [4]. Bioinformatic studies have shown that the DNA sequences of the PREs can be well characterized by a number of TF motifs including Pho, GAF, and Zeste [16–18]. An important observation from these studies is that proper combinations of multiple of TFs are required for the maintenance of PcG targeting. A high level of accuracy has been achieved by computational methods which have been developed to predict genome-wide PREs by combination of TF [16,19].

In mammals, it remains unclear to what extent the DNA sequence plays a role in PcG targeting. Unsupervised search of mammalian PREs is difficult and has only led to isolated successes [20,21]. Insights can be gained by computational studies aimed at detection of discriminative DNA sequence features. Previous studies have identified a number of TF motifs that are associated with PcG targets [11,22], but it remains unclear to what extent genome-wide target genes can be explained by combinations of different TF motifs. Furthermore, more general DNA sequence features have also been found to be associated with PcG targets, including high CpG density [13,23], high sequence conservation score [23], depletion of DNA transposons [13], and periodic patterns of dinucleotide frequencies [24]. The roles of these more general features are even less understood. More recently, it has been recognized that PcG can physically interact with non-coding RNAs [25–27], providing an indirect mechanism for sequence specific targeting. The genome-wide impact of noncoding RNA mediated targeting is still unclear.

The goal of this paper is to investigate to what extent genome-wide PcG targets in mammals can be explained by combinations of TF binding patterns. We focus on TFs rather than general DNA sequence

* Corresponding author.
E-mail address: gcyuan@jimmy.harvard.edu (G.-C. Yuan).
[1] These authors contributed equally.

0888-7543/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2010.03.012

features to facilitate biological interpretations. To this end, we have developed a computational approach to predict PcG target genes by combining ChIP-chip/seq data (referring to the collection of ChIP-chip and ChIP-seq data, which are chromatin immunoprecipitation followed by microarray or DNA sequencing, respectively) and motif sequence information. Notice that the ChIP-chip/seq data may also contain sequence independent information such as local chromatin configuration. Our model can predict PcG target genes in mouse ESCs with good accuracy. A similar approach based on the sequence information alone has a comparable model performance. The model performance is primarily due to the presence of five TF features (either binding data or motif sequences): Zf5, Tcfcp2l1, Ctcf, E2f1, and Myc.

Although our model is trained based on ESC specific data, the predictions actually reflect information about the overall target plasticity across cell-types. This is supported by several observations including that the genes with high scores tend to be marked by H3K27me3 in multiple lineages, to remain repressed during the early stage of *in vitro* cell differentiation, and to retain H3K27me3 upon Ezh2 deletion. We compared the target sequence properties between *Drosophila* and mouse and identified both similarities and differences. Interestingly, the TF homolog pair Pho/Yy1 seems to play very different roles in PcG targeting between *Drosophila* and mouse.

Results

Enrichment of TF binding and motif sites in PcG target genes

Numerous groups have previously applied ChIP-chip/seq methods to identify genome-wide locations of PcG and H3K27me3 targets in mouse ESCs [7,9–15]. Our analysis here is based on a recent dataset containing genome-wide targets of Ezh1, Ezh2, and loci harboring a H3K27me3 signature, identified by using biotinylation-mediated ChIP (bioChIP) followed by whole-genome tiling array hybridization [7]. The bioChIP approach is advantageous over traditional ChIP because it circumvents the problem associated with antibody availability. The majority of the target sites fall into promoter regions, defined here as the [-8kb, +2kb] region with respect to a transcription start site (TSS). Therefore, we focused on the promoter regions and aimed at detecting discriminative TF features associated with those targeted by the PcG proteins. In total we identified 4010 promoters that overlapped with a PcG target site. We obtained the genome-wide locations of 15 TFs in ESCs previously identified by ChIP-chip/seq experiments [28,29], and then tested whether the TF binding sites were enriched or depleted with PcG target promoters. We identified three enriched TFs: Esrrb, Ctcf, and Tcfcp2l1 (p<1.0E-32; the p-values here and after are obtained from the Fisher exact test with Bonferroni correction if not explicitly stated otherwise) and four depleted ones: Myc, E2f1, Rex1 (Zfp42), and Zfx (p<1.0E-7) (Fig. 1A). Of the above TFs, Myc, Rex1, and Ctcf have been previously linked to PcG mediated gene silencing [28,30]. Interestingly, while Oct4 (Pou5f1), Nanog, and Sox2 share many common targets with PcG [9], their overall enrichment is not significant. This is not surprising because a main function of these pluripotent TFs is to activate ESC-specific genes, which are not targeted by PcG.

Only a few TFs have been directly interrogated by ChIP-chip/seq experiments to date. For other TFs, sequence motifs provide a useful tool for the computational prediction of binding targets. We downloaded all 569 vertebrate TF motifs in the TRANSFAC database (Release 10.1) [31], and scanned genome-wide promoter sequences for matching sites. A match is quantified as a continuous motif matching score. By thresholding the motif matching scores, we also obtained binary calls indicating the presence or absence of a motif site at a given promoter (Materials and Methods). A total of 50 motifs are significantly enriched in the promoter regions of PcG target genes (p<1.0 E-5) although the degree of enrichment is mostly moderate

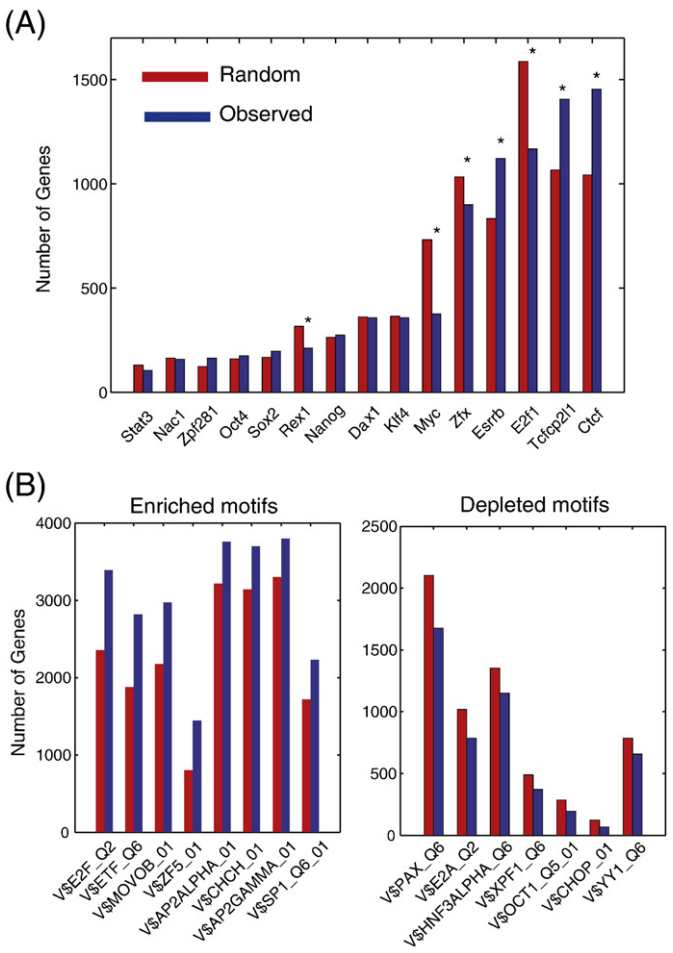


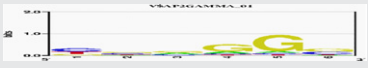
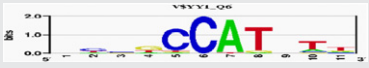
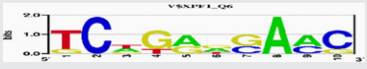
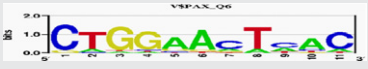
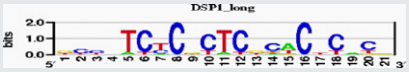
Fig. 1. Enrichment analysis for overlaps between TF and PcG targets. (A) The 15 TFs probed by ChIP-chip/seq experiments in mouse ESCs. The statistically significant one are marked by asterisks (p<1.0E-7 from one-sided Fisher exact test with Bonferroni correction). (B) The most enriched or depleted TF motifs.

(some extreme examples are shown Table 1, and a complete list is shown in Supplemental Table 1). These motifs correspond to 39 distinct TFs due to redundancy. Some of these TFs have been implicated to interact with PcG proteins in the literature, such as E2f family members, Sp1, NRSF/Rest, and Myc [11,32–35]. Interestingly, we also found 7 motifs that are significantly depleted in PcG targets (p<1.0 E-5), of which Pax and Yy1 have been previously linked to PcG binding in the literature [11,36,37]. Our results are consistent with a previous study [11].

Prediction and validation of PcG target genes in mESCs

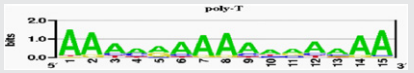
While statistically significant, the relative enrichment of the each associated TF is only moderate. The situation in *Drosophila* is similar and it has been shown that enhanced specificity can be achieved by considering combinatorial binding patterns [16]. This motivated us to build a statistical model to predict genome-wide PcG targets in mouse ESCs from combinations of TF features corresponding to either binding data or motif information. Due to the complex relationship between TF features and PcG binding, we applied a recently developed flexible statistical method called Bayesian Additive Regression Trees (BART) [38], which has been recently shown [39] to be more powerful than a number of traditional methods including Lasso [40] and support vector machine (SVM) [41]. BART is a sum-of-tree model (see Methods and Materials for details). Its superior performance can be attributed to two important properties. First,

Table 1

Motif Name	Bias (mouse)	p-value (mouse)	Bias (Drosophila)	p-value (Drosophila)	Motif Logo
A. Enrichment analysis for mouse TF motifs: (p-values are calculated based on one-sided Fisher exact tests with Bonferroni correction).					
V$E2F_Q2	enriched	0.00E + 00	enriched	5.09E-02	
V$ETF_Q6	enriched	1.26E-240	enriched	8.28E-02	
V$MOVOB_01	enriched	6.41E-177	enriched	6.88E-06	
V$ZF5_01	enriched	4.30E-155	enriched	8.88E-01	
V$AP2ALPHA_01	enriched	5.17E-152	enriched	5.01E-10	
V$CHCH_01	enriched	1.64E-147	enriched	5.01E-10	
V$AP2GAMMA_01	enriched	1.90E-141	enriched	5.01E-10	
V$SP1_Q6_01	enriched	6.33E-71	enriched	9.22E-01	
V$YY1_Q6	depleted	2.58E-06	enriched	3.15E-02	
V$CHOP_01	depleted	1.28E-07	depleted	1.00E + 00	
V$OCT1_Q5_01	depleted	7.19E-08	enriched	1.00E + 00	
V$XPF1_Q6	depleted	1.82E-08	depleted	1.00E + 00	
V$HNF3ALPHA_Q6	depleted	8.86E-12	enriched	7.46E-01	
V$E2A_Q2	depleted	3.50E-19	depleted	1.00E + 00	
V$PAX_Q6	depleted	6.42E-48	enriched	3.39E-01	
B. Enrichment analysis for Drosophila TF motifs: (p-values are calculated based on one-sided Fisher exact tests with Bonferroni correction).					
DSP1_long	enriched	1.00E + 00	enriched	7.98E-23	
GAF_short	enriched	1.52E-18	enriched	8.18E-21	
DSP1_short	enriched	3.54E-03	enriched	4.02E-18	
GAF_long	enriched	7.56E-29	enriched	4.55E-06	
PHO_short	depleted	2.64E-12	enriched	3.46E-04	
GT_repeat	enriched	3.67E-18	enriched	4.18E-04	
PHOL_short	depleted	1.82E-02	depleted	1.19E-03	
PHO_long	depleted	7.77E-07	enriched	3.38E-02	
ZESTE	enriched	1.00E + 00	depleted	9.56E-02	

(continued on next page)

Table 1 (*continued*)

Motif Name	Bias (mouse)	p-value (mouse)	Bias (Drosophila)	p-value (Drosophila)	Motif Logo
B. Enrichment analysis for Drosophila TF motifs: (p-values are calculated based on one-sided Fisher exact tests with Bonferroni correction).					
PHOL_long	depleted	2.08E-02	depleted	1.00E + 00	
polyA	depleted	6.15E-09	depleted	1.00E + 00	
TGC_triplet	enriched	1.34E-16	enriched	1.00E + 00	

BART is able to capture nonlinear relationships among the sequence features, which is important since it has been shown in *Drosophila* that combinations of TFs are important for PcG targeting [16]. Second, BART integrates information from a large number of classifiers thereby enhancing the robustness against heterogeneity among different PcG targets.

For each gene, we combined the matching scores for the 569 downloaded motifs with the ChIP-chip/seq data for the 7 differentially associated TFs mentioned above (also see Fig. 1A). The resulting 576 dimensional TF features were used as predictors for PcG targeting. The detailed implementation is described in the Materials and Methods section. The model performance was evaluated using a variation of the three-fold cross-validation (Materials and Methods). The average receiver operating characteristic (ROC) curve and the corresponding area under the curve (AUC) score are shown in Fig. 2A. A random classification model corresponds to an AUC score of 0.5, and a perfect classification model has an AUC score of 1.0. Our model has good prediction accuracy. It has an AUC of 0.8266 when all 576 TF features are included as predicting variables (full-version) and of 0.7951 if only information from the 569 motifs (motif-only version) is included, suggesting that prediction accuracy is primarily contributed by the DNA sequence. We compared the model performance of SVM and Lasso and found that BART indeed provided higher prediction accuracy (Supplemental Fig. 1).

In additional to TRANSFAC, there are two additional large-scale motif databases available: JASPAR [42] and UniPROBE [43]. We recognized that each database has its own advantages and drawbacks therefore tested the model performance by using these other databases instead. For each of these additional TFs, we obtained the gene-specific motif matching scores and used these matching scores as predictors as for TRANSFAC. However, these other two databases did not perform as well as TRANSFAC (Supplemental Fig. 2). This is probably due to the fact that TRANSFAC is more comprehensive than the other two databases, which consist of higher quality motifs.

Another concern is that either the model performance or the detected TF features may be sensitive to the experimental platform where the binding data were obtained. To address this issue, we noticed that the TFs Oct4, Nanog, Sox2, Klf4, and Myc, were queried in both Chen *et al.* [29] and Kim *et al.* [28] studies. In the above analysis, the binding information for these common TFs were obtained from ref. [28]. We repeated the analysis by using the data from ref. [29] instead. The model performance is similar (Supplemental Fig. 3).

To gain biological insights into which TFs are most informative, we ranked the TF features according to the BART counts (Supplemental Table 2). A higher ranked TF feature is considered more important since its distribution is more informative of PcG targeting status. We repeated the above procedure and built reduced versions of the model by keeping only the top ranked TF features as the predicting variables. With 18 features only, the prediction accuracy is nearly the same as the full-version (AUC = 0.8270). Strikingly, the model still performs reasonably well (AUC = 0.79) even with only 5 features, corresponding to E2f1, Zf5, Tcfcp2l1, Myc, and Ctcf, suggesting that the model performance is mainly due to these five features. Of these 5 features, 3 are PcG-enriched (Zf5, Tcfcp2l1, Ctcf) and 2 are PcG-depleted (E2f1, Myc), suggesting that both types of features are important for global PcG targeting.

We then asked which genes were most likely to be PcG targets based on the full-version model. Since our model was built upon three different training sets and each gave a slightly different result, we averaged the predicted propensity scores associated with each gene. To avoid using observed PcG binding information for its own

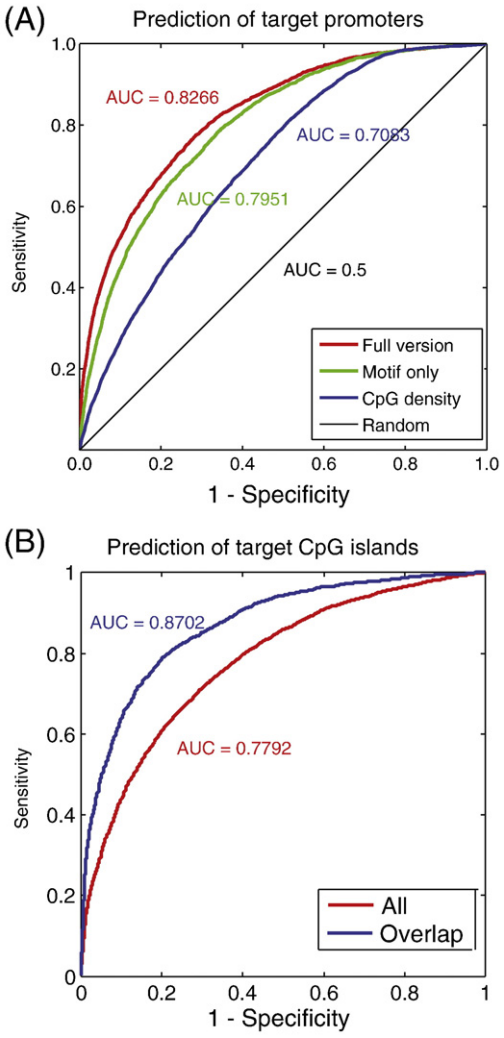


Fig. 2. Prediction accuracy for the BART model. (A) ROC curves and corresponding AUC scores for genome-wide PcG target promoter predictions. We present results for two versions of the BART model: the full-version (red) and the motif-only version (green). Also shown are results based on the CpG density (blue) and random guess (black). (B) ROC curves and corresponding AUC scores for PcG positive CpG island predictions. The two curves correspond to all CpG islands (red) and the overlapping subset between our study and Ku *et al.* 2008 (blue), respectively.

"prediction", we only considered the testing set for averaging. The results for all mouse genes are shown in Supplemental Table 3. Of the 400 genes with the highest predicted propensity scores, 368 (92%) correspond to experimentally identified targets, suggesting these genes are strongly sequence dependent. Gene Ontology (GO) enrichment analysis suggests these genes are highly enriched with regulators both for multicellular organismal development (FDR = 1.94E-20, where FDR stands for false discovery rate) and for cellular metabolic process (FDR = 9.13E-18). The prediction specificity gradually decreases with less stringent criteria, but still 49% (compared to 19.2% expected at random) of the top 5000 predicted targets correspond to experimentally identified targets, with a sensitivity of 61% (compared to 23.9% expected at random).

TF motifs provide additional target information than CpG density

Previous studies have implicated a role of the CpG density and highly conserved non-coding elements (HCNE) in PcG recruitment [4,13,23]. However, it remains unclear to what extent these sequence features are useful for genome-wide target predictions and whether TF motifs provide additional information. For comparison, we ranked the genes based on the CpG density and conservation score, respectively, and selected the top ranked genes as predictors for PcG targets. The CpG density is indeed a good predictor (AUC = 0.7083), as expected, although less so than our TF-based model (Fig. 2A). On the other hand, the sequence conservation score is not a good predictor (AUC = 0.5806). This is not inconsistent with previous results [23]. While the majority of PcG targets fall into highly conserved regions, there are also many non-target promoters that are also highly conserved.

A different approach has been used in a recent study to predict genome-wide PcG target CpG islands in mouse ESCs based on motif analysis [11]. This approach first identifies discriminative TF motifs, then predicts target CpG islands by counting the number of PcG-enriched vs depleted motifs. About two thirds of the targets are correctly predicted by this method. To compare the model performance, we applied the motif-only version of our model to predict PcG target CpG islands. Based on the ChIP-chip data in ref. [7], we identified 6602 target CpG islands and 9346 non-targets. The prediction of our model for target CpG islands is also good (AUC = 0.7792, Fig. 2B). For comparison, we selected the overlapping subset of CpG islands for which our annotations agreed with that in ref. [11], containing 2595 target and 7573 non-target CpG islands. For this subset, our model has an 88.7% specificity at a cutoff value corresponding to a 66.6% sensitivity. Therefore, the two approaches perform similarly well in prediction of the total targets, although the performance of the method in ref. [11] might be slightly overestimated since the PcG target status of the CpG islands were also used for model training. The main difference is that the prediction outcomes of our model are more quantitative compared to those in ref. [11]. These quantitative predictions can be used to detect subtle differences among the PcG targets, thereby providing new biological insights as discussed in the next section.

DNA sequence is important for PcG target plasticity

While our prediction model was based on the DNA sequence information along with the binding profiles of seven TFs in ESCs, we asked to what degree the predictions were valid in other lineages. To this end, we analyzed a ChIP-seq dataset containing genome-wide target information for H3K27me3, which is the histone modification mark catalyzed by PcG proteins, in three cell lines: ESC, mouse embryonic fibroblast (MEF), and neural precursor cells (NPC) [14]. The H3K27me3 targets change significantly across cell lines. Out of 3426 genes that are marked by H3K27me3 in at least one lineage, only 516 are marked in all three lineages. The genes predicted with highest propensity scores tend to be marked by H3K27me3 in all three lineages. We divided all genes into four groups based on the frequency they are marked by H3K27me3 in the three lineages. We compared the distribution of propensity scores for different groups and observed a consistent trend: higher propensity scores are associated with more stably marked genes (Fig. 3A, Wilcoxon rank sum test $p<6.0E-21$). This trend suggests that the propensity score of a gene may reflect the intrinsic PcG binding stability which is independent of lineages, and the target plasticity is highest for genes with intermediate propensity scores.

A closer examination of the data suggests that the propensity score is dependent not only on the overall target frequency but also on the specific lineage in which a gene is targeted. Since our model is trained based on ESC data, the target status in ESC has a greater impact on the propensity score compared with the other cell lines. We refined each group by specifying the exact target pattern, resulting in 8 subgroups. We found that the average propensity score for (ESC+; MEF-; NPC-) was indeed higher than that for (ESC-; MEF+; NPC+) (Supplemental Table 3). However, after controlling for the H3K27me3 pattern in other cell lines, the target status in a specific cell line is always positively correlated with propensity scores. Thus the propensity scores indeed reflect the target plasticity.

Since dynamic changes of PcG binding affinity occur during *in vitro* cell differentiation [14,15], we were interested to test an association between the PcG residence time and the propensity scores. However, such a test cannot be done directly due to the lack of time-course PcG binding data. To circumvent this difficulty, we recognized that the PcG target status is strongly associated with gene expression and analyzed a time-course gene expression dataset, where gene expression levels were measured at Day 0, Day 2, Day 6 and Day 8 after induced differentiation by LIF removal [7]. We reasoned that, if the propensity scores were indeed associated with the intrinsic PcG binding stability, then the highly scoring genes should remain repressed during differentiation while other genes tend to be more readily activated. Therefore, we divided the PcG target genes detected in ESCs into groups by propensity scores and compared the expression levels for each gene group. The gene groups with higher propensity scores indeed remain repressed for a longer time (Fig. 3B). Interestingly, the expression levels of different PcG target gene groups are already markedly different in ESCs, probably reflecting variable PcG binding stability among the target genes in ESCs.

We next tested whether the propensity scores may also be predictive of PcG target stability under genetic perturbations. Recent ChIP-chip experiments have shown that depletion of either Ezh2 or Jumonji (Jmj, Jarid2) severely damages but does not completely abolish either PcG or H3K27me3 targeting [7,44]. Of the 4010 PcG target genes in the wild-type mouse ESC, 1161 (29%) retain the H3K27me3 mark in the Ezh2 mutant ESC. This group of targets is denoted as $Ezh2^{-/-,\ H3K27me3+}$ to be distinguished from the rest, denoted as $Ezh2^{-/-,\ H3K27me3-}$. We found that the targets with higher propensity scores are indeed more likely to retain the H3K27me3 mark in the Ezh2 mutant (Supplemental Table 4). For example, of the 400 PcG targets in wild-type with the highest propensity scores, 289 (72%) belong to the $Ezh2^{-/-,\ H3K27me3+}$ category, substantially higher than the overall frequency (29%). The overall distribution of the propensity scores is shifted to the right for the $Ezh2^{-/-,\ H3K27me3+}$ genes compared to the $Ezh2^{-/-,\ H3K27me3-}$ genes (Fig. 3C, t-test $p<1E-100$). Consistent with these results, the enrichment bias for the TF features is also stronger for the $Ezh2^{-/-,\ H3K27me3+}$ genes (Fig. 3D). The prediction for the $Ezh2^{-/-,\ H3K27me3+}$ genes (AUC = 0.9003) is also more accurate than for the $Ezh2^{-/-,\ H3K27me3-}$ genes (AUC = 0.7329).

Taken together, the above results strongly suggest that the propensity score, which is mainly determined by the genomic sequence, is highly correlated with intrinsic PcG binding stability and that genes with intermediate propensity scores tend to be associated with significant target plasticity.

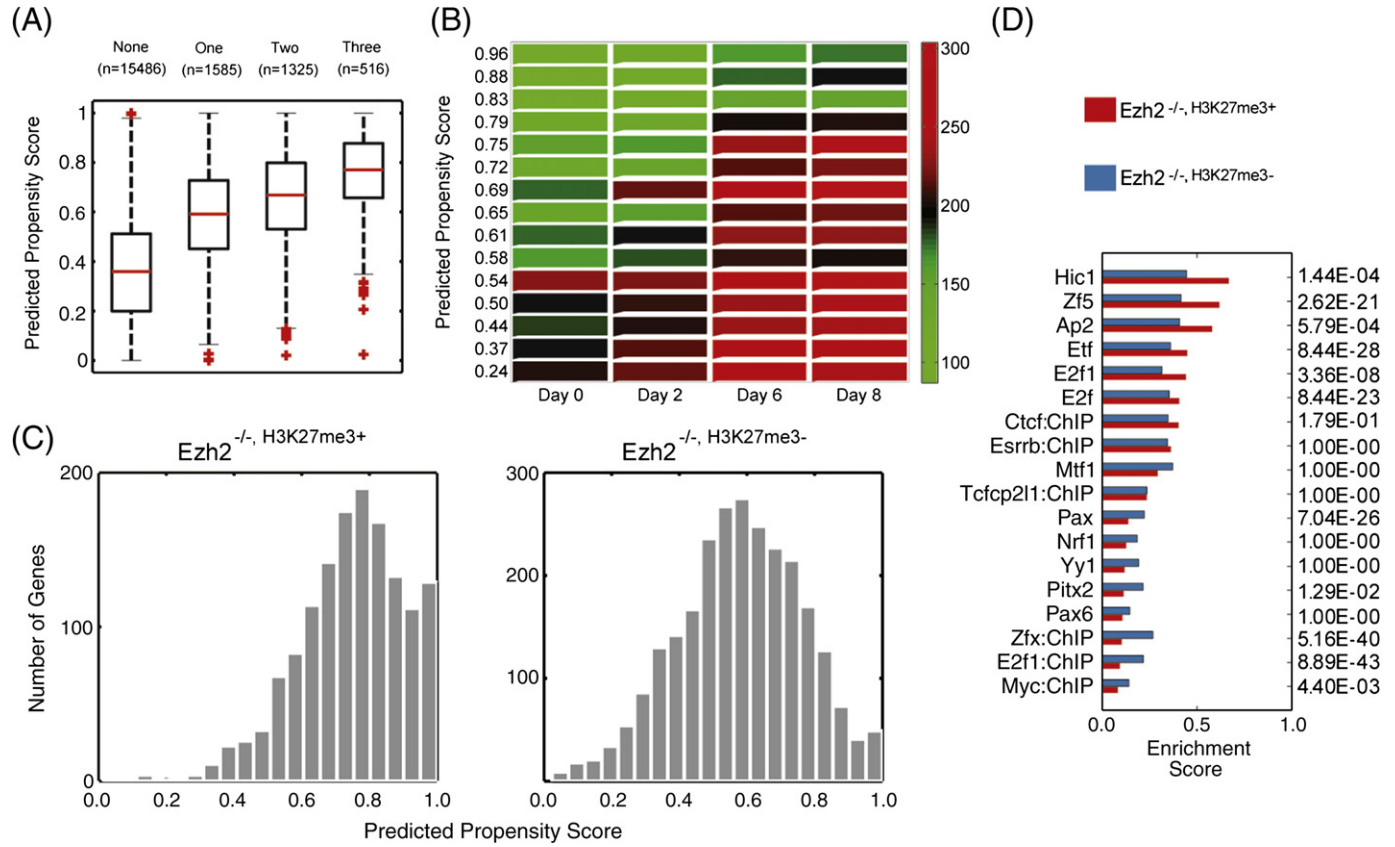


Fig. 3. Predicted propensity scores reflect the overall PcG target plasticity. (A) Comparison of the propensity score distribution among different gene groups with similar H3K27me3 profiles. The number of lineages in which the genes are marked by H3K27me3 is shown above the figure. The number of genes in each group is also shown (in parentheses). (B) Time course gene expression level analysis. The PcG target genes in ESCs are divided into 15 roughly equal-sized groups associated with similar propensity scores (mean values shown on the left). The heat map indicates the mean mRNA expression level within each group at different time points after LIF removal. (C) Comparison of the propensity score distributions for the Ezh2$^{-/-,\ H3K27me3+}$ and Ezh2$^{-/-,\ H3K27me3-}$ genes, which correspond to the subset of PcG targets that either retain or lose the H3K27me3 mark in the Ezh2$^{-/-}$ mutant ESCs. (D) Enrichment score for overlap between the top 18 TF features and Ezh2$^{-/-,\ H3K27me3+}$ or Ezh2$^{-/-,\ H3K27me3-}$ targets. The label ":ChIP" after certain TFs is used to indicate that target information is based on ChIP-chip/seq data. The enrichment score is defined as the ratio of the observed frequency of a TF feature among PcG targets over the frequency expected by chance.

Evolutionary conservation of PcG target sequence specificity

Although the genomic sequences between *Drosophila* and mammals are very different, the core PcG complex members are conserved between *Drosophila* and mammals and the target genes tend to have similar functions [9,18,45]. Therefore it is interesting to investigate to what extent the PcG targeting mechanism is evolutionarily conserved. Genome-wide PcG targets in *Drosophila* have been identified by using ChIP-chip experiments [18]. Based on this dataset, we analyzed the enrichment of different TF motifs with respect to the PcG target genes. Because our goal was to compare the target sequence properties between the two species, we included both the 569 vertebrate TF motifs and a set of 12 *Drosophila*-specific TF motifs in our analysis (Methods and Materials). The *Drosophila*-specific motifs were selected from the literature based on their implication with PcG targeting [16,18]. All these motifs are distinct from mouse motifs except for Pho, which is similar to the mouse Yy1 motif. In fact, the TFs Yy1 and Pho are homologous to each other. We found several similar sequence features: 5 *Drosophila*–specific motifs and 4 vertebrate-specific motifs are significantly enriched or depleted with the same bias in mouse and *Drosophila* PcG targets (p<1E-5) (Table 1).

Unexpectedly, our analysis also identified a striking difference between the two species. For *Drosophila* the TF Pho is the single most important factor for PcG recruitment [16–19]. In mouse, the homolog protein Yy1 is structurally similar to Pho and recognizes similar DNA sequences (Fig. 4A). Transfected Yy1 can interact with PcG proteins and partially rescue the Pho mutation phenotype in *Drosophila* [37]. Therefore we expected that Yy1 binding should be positively correlated with PcG targets in mouse. However, our analysis shows that the Pho/Yy1 motif sites were PcG-enriched in *Drosophila* but depleted in mouse. By searching the literature, we found that similar results have been previously found in the literature using different approaches [11,12]. However, the role of Yy1 in PcG targeting in mammals remains unclear.

We adapted our BART model to predict genome-wide PcG target genes in *Drosophila,* replacing the predictors by the motif matching scores corresponding to the 12 *Drosophila*-specific motifs. We used the ChIP-chip data in ref. [18] for model training. The model performance was evaluated as discussed above. We found that the overall prediction accuracy is slightly poorer compared to mouse (AUC = 0.7877, Fig. 4B. The predicted propensity scores for all *Drosophila* genes are shown in Supplemental Table 5.) We applied the *Drosophila*-derived model to predict PcG targets in mouse and found the performance is rather poor (AUC = 0.5450). Removing the 4 Pho related motifs improved the prediction accuracy but only slightly (AUC = 0.5696). Next we applied the mouse-derived model to predict PcG targets in *Drosophila*. The model performance is also poor (AUC = 0.6070) and only slightly improved by removing the 4 Yy1-related motifs from model construction (AUC = 0.6671) (Fig. 4B). The above results suggest significant differences between the target sequence signatures between *Drosophila* and mammals.

Despite the large overall differences between genome sequences in mouse and *Drosophila*, the Hox gene clusters are highly conserved and are targeted by PcG in both species. Because our above analysis

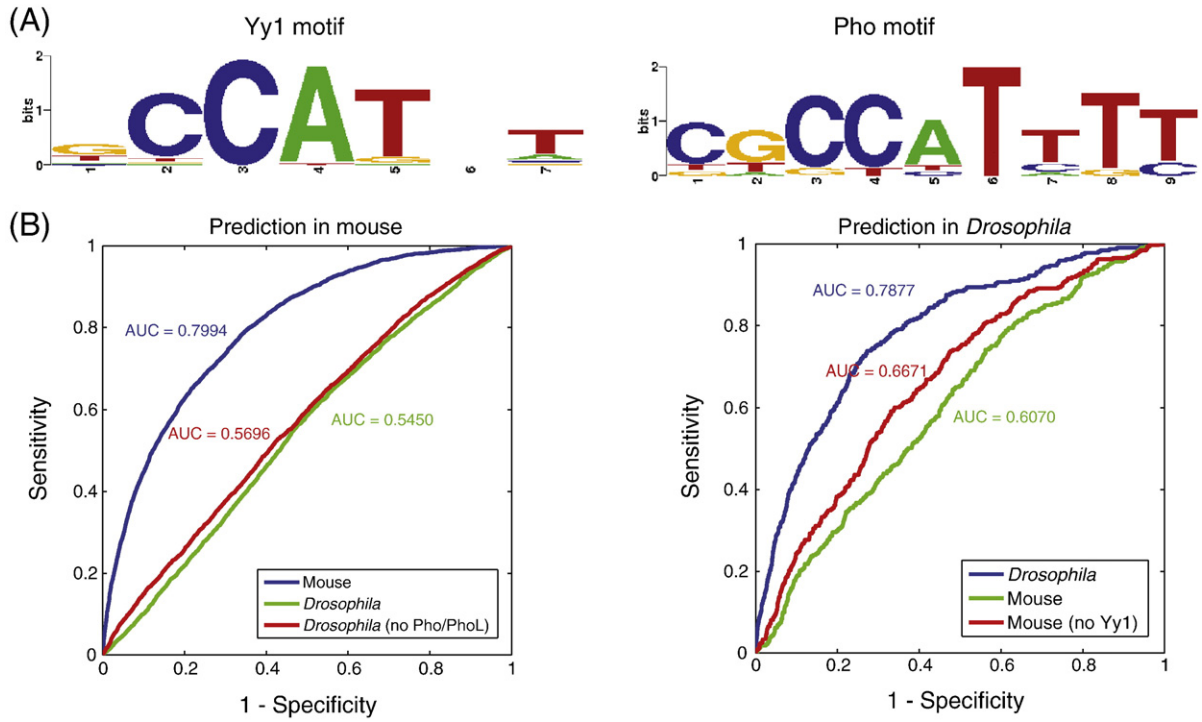


Fig. 4. The mouse and *Drosophila* PcG targets are associated with different sequence signatures. (A) The Pho and Yy1 motifs are similar. (B) ROC curves and corresponding AUC scores for cross-specifies prediction.

suggests that the PcG target sequences are very different between the species, we were interested to test whether the conservation of PcG targeting can be explained by divergence of promoter sequences. *Drosophila* contains eight Hox genes spread over two clusters: ANT-C and BX-C. The five Hox genes in ANT-C: lab, pb, Dfd, Scr, and Antp, are strongly homologous to the mammalian Hoxb genes: Hoxb1, Hoxb2, Hoxb4, Hoxb5, and Hoxb6. Therefore, we compared the PcG and TF binding patterns between the ANT-C and Hoxb clusters. All five Hox genes at the ANT-C region in *Drosophila* are PcG targets (Fig. 5A), whereas nine out of the ten Hoxb genes in mouse are also targeted by PcG proteins. Our model correctly identified almost all the PcG targets in these regions (Fig. 5). Strikingly, all the PcG targets at the ANT-C locus in *Drosophila* are bound by Pho, whereas none of the Hoxb cluster genes in mouse contain a Yy1 motif site. In summary, we found that the PcG target sequences are different even at conserved targets, suggesting strongly divergent evolution of the PcG targeting mechanism.

Discussion

We have developed a computational approach to predict PcG target genes based on combinations of TF features. Our model predicts genome-wide PcG target genes in mouse ESCs with good accuracy, especially for the top predicted candidates. The prediction power is mainly contributed by five TF features (Zf5, Tcfcp2l1, Ctcf, E2f1, Myc). By comparison with a previous approach [11], we found that the two approaches performed similarly well in predicting the total number of PcG targets. However, the outcomes of our model are more quantitative and can be used to predict PcG targets with various stringencies. The top predicted candidates from our model are much more accurate than the overall average and may potentially serve as a guide for *de novo* search for PREs. On the other hand, we also recognize that a mere association does not necessarily imply active recruitment, and that it is possible that the active recruitment sites are located not in promoters but elsewhere.

Two of the most informative TF features, corresponding to Myc and E2f1, belong to families of TFs that recognize similar DNA sequences but have different biological functions. While Myc is a potent activator of cell-cycle related genes, another family member, Mga, is a repressor [46]. Similarly, E2f1 and E2f6 are from the same family but the former is an activator whereas the latter is a repressor [47]. Indeed, previous studies have found that E2f6 interacts with Mga and a number of PcG proteins [32,48]. Interestingly, the binding sites of Myc and E2f1 are both depleted in PcG targets, but their corresponding motif sites are enriched. We recognize that a TF motif may correspond to multiple TFs, since multiple proteins may recognize the same DNA sequence. Therefore the motif analysis suggests that different members within a TF family may play antagonistic roles in PcG targeting.

Although our model training involves only one specific cell line (ESC), the predicted propensity scores are strongly associated with the overall PcG target plasticity across different cell lines. Highly scored genes are likely to be marked by H3K27me3 in multiple lineages, to remain repressed during induced differentiation, and to retain H3K27me3 in Ezh2 mutants. On the other hand, it is important to recognize that the lineage-specificity of PcG targeting is ultimately controlled by the concerted action of many factors, including not only TFs but also chromatin modifiers [49,50] and noncoding RNAs [25,26,51]. A comprehensive understanding of the PcG targeting mechanism will undoubtedly require the integration of multiple types of information together.

We found significant differences between the PcG target sequence properties in mouse and in *Drosophila*. Surprisingly, while nearly all *Drosophila* PcG targets contain a Pho binding site, the Pho/Yy1 motif is actually depleted in mouse. This difference cannot simply be explained by the overall genomic sequence differences between the two species, since the target sequence difference is also present at conserved genes such as the Hox gene clusters. These results suggest that Pho/Yy1 may play very different roles in PcG targeting between the two species. We recognize that it is also possible that the Yy1

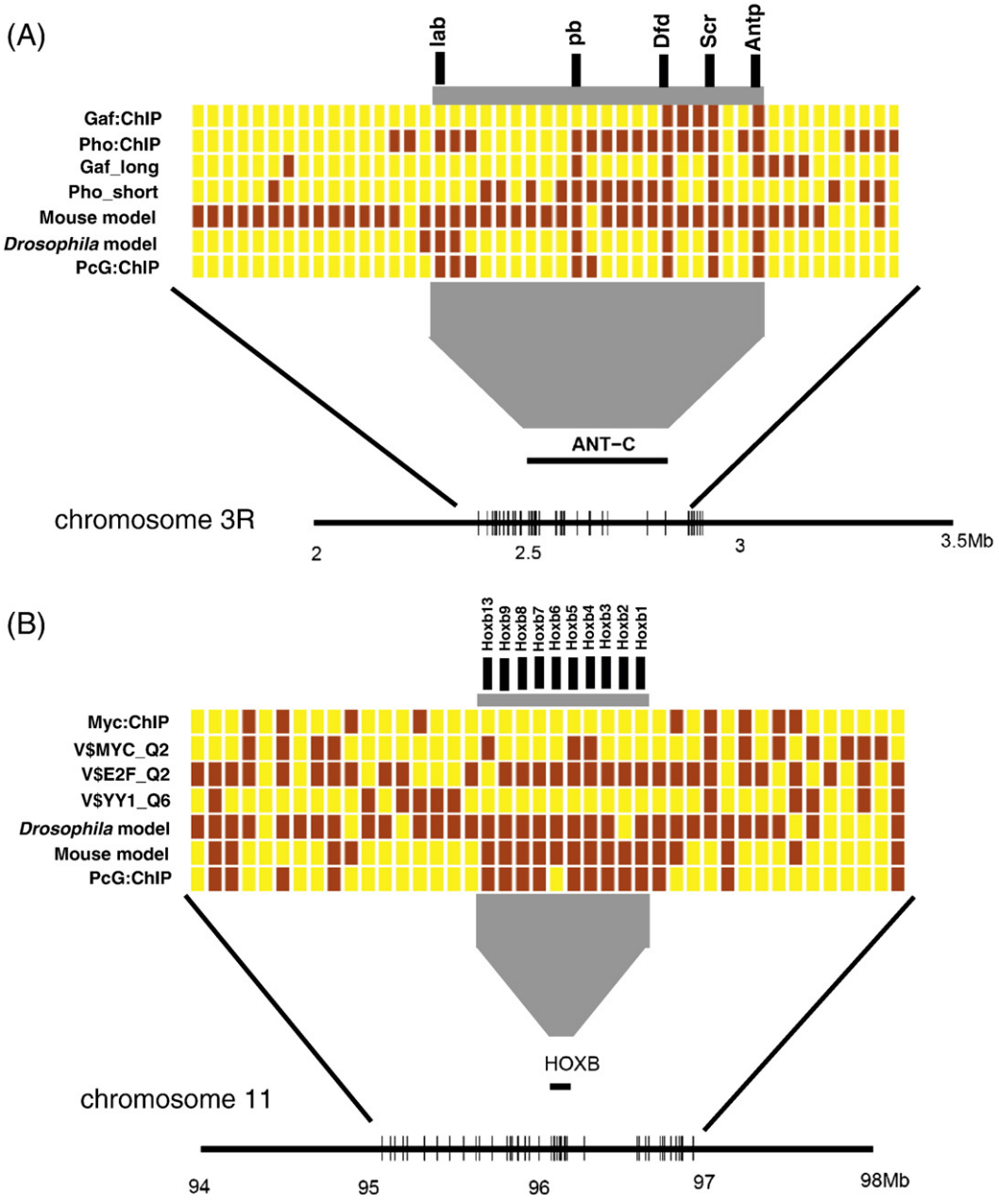


Fig. 5. The Hox clusters are targeted by PcG in both mouse and *Drosophila* but their promoter sequences show different properties. (A) The *Drosophila* ANT-C region. (B) The mouse Hoxb cluster. Each colored box represents a protein-coding gene, in the order of their chromosomal locations. The TSS coordinates of the genes are shown as vertical lines in the bottom of the figure. The color indicates either presence (red) or absence (blue) of a specific feature labeled on the left. The locations of the Hox genes are marked in the above. The label ":ChIP" after certain TFs is used to indicate that target information is based on ChIP-chip data.

binding sites located elsewhere may play a positive role in PcG recruitment through chromatin looping. It will be very interesting to closely examine the role of Yy1 in PcG targeting by combined experimental and bioinformatic methods in the future.

Materials and methods

Mouse genome sequences and annotations

Mouse genome sequences, gene annotations, and conservation scores are downloaded from UCSC Genome Browser (*http://genome.ucsc.edu/*). The genome sequence is based on the February 2006 (mm8) genome assembly. Gene annotations are based on Refseq mRNA, including a total of 21,115 Refseq entries, which map to 18,912 unique gene symbols. The promoter of a gene is defined as the region between -8kb and +2kb with respect to the TSS as in previous studies [7,28]. Conservation scores are calculated from PhastCons [52]. The conservation score of a gene is set to be the highest conservation score of the genomic location mapped to its promoter region. The CpG density of a gene is defined as the frequency of the dinucleotide CpG in the promoter region. Gene Ontology analysis is done by using the DAVID Gene Functional Classification Tool (*http://david.abcc.ncifcrf.gov/*).

Target genes of PcG and TFs identified from ChIP-chip and ChIP-seq experiments

The Ezh1, Ezh2, and H3K27me3 targets are previously identified by ChIP-chip experiments using whole-genome tiling arrays [7]. A promoter is called a target if it overlaps with either an Ezh1 or an Ezh2 peak (i.e., by at least one base pair) in the wild-type mouse. The

genome-wide targets for 9 TFs, Oct4, Sox2, Klf4, Myc, Nanog, Dax1 (Nr0b1), Rex1, Zpf281, and Nac1 (Nacc1), are identified by ChIP-chip using promoter tiling arrays [28]. The targets for an additional 6 TFs, Stat3, Zfx, Esrrb, E2f1, Tcftp2l1, and Ctcf, are identified by ChIP-seq experiments [29] and mapped to overlapping promoters. For the TFs investigated by both studies, the target locations are based on ref. [28].

Motif analysis

Position weight matrices (PWM) are downloaded from the three databases: 569 vertebrate motifs from TRANSFAC (Release 10.1) [31], 99 vertebrate motifs from JASPAR [42], and 387 mouse motifs from UniPROBE [43] For each motif M, the raw motif matching score at a promoter region P is calculated as

$$\max_{S\in P}\left[\log\frac{P(S|M)}{P(S|B)}\right],$$

where the background frequency (B) of different nucleotides is estimated using promoter sequences only. The raw motif matching scores are normalized by dividing the maximum possible score. The statistical significance of a motif score is quantified by a p-value which is estimated based on the distribution of motif scores for 10,000 DNA sequences of 10 kb length randomly selected from the mouse genome. Binary calls for motif sites are determined by a cutoff score value corresponding to $p=0.005$. Motifs obtained from each databases are analyzed separately.

The enrichment score for a TF feature in PcG target genes is defined as the ratio of the observed frequency of the feature among PcG targets over the frequency expected by chance. For analysis of $Ezh2^{-/-,\ H3K27me3+}$ (and $Ezh2^{-/-,\ H3K27me3-}$, respectively) genes, the numerator is defined similarly, but the denominator is modified by excluding $Ezh2^{-/-,\ H3K27me3-}$ (and $Ezh2^{-/-,\ H3K27me3+}$, respectively) genes. The corresponding p-values are calculated based on one-sided Fisher exact tests with Bonferroni correction.

For *Drosophila*, we consider 12 motifs that have been previously implicated in PcG recruitment [16,18], including nine motifs corresponding to five TFs (Pho, PhoL, Dsp1, Gaf and Zeste) and three motifs (GT repeat, TCG triplet, and PolyA) obtained from *de novo* motif search. The enrichment analysis is done as described above.

BART model

BART is a Bayesian sum-of-trees model. Briefly, the BART model can be represented as $Y=\left(\sum_{j=1}^{J} g(x;T_j,M_j)\right)+\varepsilon$, where $g(x;T_j,M_j)$ is a regression tree and can be viewed as an independent classifier. The output of each "tree" is obtained by following the "branches" which bifurcate at a series of "nodes". A node represents a binary rule like "does a TF X bind to a gene G"? The end of each branch is called a terminal node, where a constant outcome is assigned to all the genes that satisfy the same set of binary rules. The final outcome of BART is obtained by summing up the outputs from individual "trees".

Mathematically speaking, T_j and M_j represent the topological structure and the terminal node values associated with the j-th tree, respectively. For each run, 200 trees are sampled and iteratively refined by the Gibbs sampler, which updates each tree (T_j,M_j) by calculating the posterior distribution $(T_j,M_j)|T_{(j)},M_{(j)},Y$, based on the current setting $(T_{(j)},M_{(j)})$, i.e., all other trees excluding (T_j,M_j). Four types of updating changes are allowed at each time with pre-determined probabilities: growing a terminal node, pruning a pair of terminal nodes, changing a non-terminal rule, and swapping a rule between parent and child. The proposed change is either accepted or rejected based on the Metropolis-Hastings algorithm [53]. The iteration is repeated until convergence. The computations are done by using the *BayesTree* package in *R* with default parameters [38].

The main outcome of the BART model is a propensity score measuring the likelihood of a given gene as a PcG target. In addition, the relative importance of each predicting variable is evaluated by the BART counts, that is, counting the number of times the variable appears as non-terminal nodes. It is possible that a TF corresponds to multiple motif variants. In this case, only the maximum count is considered.

Model validation

We use a three-fold cross-validation to validate our prediction model. The total PcG targets are divided into three equal-sized subsets, each containing 1323 genes. Each target subset is matched with a randomly selected set of 1323 non-target genes. At each time, one of the matched sets is used as the training set, whereas all remaining genes formed the corresponding testing set. This procedure is repeated three times so that every balanced gene set is selected for training exactly once.

The prediction accuracy is evaluated by using the receiver operating characteristic (ROC) curve, which plots the sensitivity, defined as $TP/(TP+FN)$, against the specificity, defined as $TN/(TN+FP)$. Notice that the x-axis actually measures 1 – specificity. The area under the ROC curve (AUC) is used as a summary score to quantify the overall accuracy.

Model versions

Several versions of the BART model are obtained differing in the predicting variable selections. Each promoter is assigned with a 576 dimensional covariate vector, containing the binary outcome (presence or absence) of ChIP-chip/seq data for the 7 differentially distributed TFs and the normalized matching scores for the 569 motifs downloaded from TRANSFAC. For the full-version model, all 576 TF features are used as predicting variables. For the motif-only version, the ChIP-chip/seq data for 7 TFs are excluded. In addition, reduced versions are also built by using smaller subsets of predicting variables. Here the TF features are ranked based on their BART count in the full-version model. Only the top TF features are selected as predicting variables. All versions of the BART model are trained in the same procedure as described above.

Alternative classifiers

For comparison, we also build prediction models by using two traditional classifiers: Lasso [40] and SVM [41]. Lasso is a shrinkage and selection method for linear regression. It was originally developed for regular linear regression. In our analysis, we use the R package *grplasso* which is a modified version for logistic regression [54]. For SVM, we used the package gist (*http://www.bioinformatics.ubc.ca/gist/*) to do the analysis.

CpG island analysis

15,948 **C**pG islands are downloaded from the UCSC Genome Browser. A 10kb window centered at the mid-position of each CpG island is used for annotation and prediction. A CpG island is called PcG-positive if the corresponding 10 kb window overlaps with an Ezh1 or an Ezh2 peak in the wild-type. The motif matching scores associated with each CpG island are calculated as above with the exception that the motif scanning is done by using the DNA sequence in the 10kb window. The propensity score for each CpG island is calculated by applying the motif-only version model. The PcG-positive CpG islands in ref. [11] are obtained by communication with Dr. Bradley Bernstein and referred to as Ezh2 positive in their paper.

Comparative analysis

Drosophila genome sequences, gene annotations, and conservation scores are downloaded from the Flybase (*http://flybase.bio.indiana.*

edu/). The genome sequence is based on the *Drosophila melanogaster* genome assembly release 4.3. A total of 14,449 genes are annotated. The promoter of a gene is again defined as the region between -8kb and +2kb with respect to the annotated transcription start sites.

Twelve PcG-related motifs are obtained from the literature. In particular, Zeste, polyA, TCG-triplet, and GC-repeat are obtained from ref. [16]. DSP1_long, DSP1_short, GAF_long, GAF_short, PHO_long, PHO_short, PHOL_long, PHOL_short are obtained from ref. [18]. Comparison between mouse and Drosophila motifs is done by using STAMP [54]. The motif matching scores are calculated as above. For each motif, the 4000 genes with highest matching scores are annotated as motif sites. This is roughly equal to the number of Pho ChIP peaks. The BART model is built as above while using the 12 *Drosophila* motifs as predicting variables and the ChIP-chip data in ref. [18] for training. This *Drosophila*-trained model is then applied to mouse as described above.

Acknowledgments

We thank Drs. Xiaohua Shen, Jonghwan Kim, Kimberly Glass, Catherine Yan, and Stuart Orkin for helpful discussions. We also thank Drs. Bradley Bernstein, Mythily Ganapathi, Leonie Ringrose, and Marc Rehmsmeier for assistance with data related to the cited publications. This research was supported by a Claudia Adams Barr Award.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2010.03.012.

References

[1] E.B. Lewis, A gene complex controlling segmentation in Drosophila, Nature 276 (5688) (1978) 565–570.
[2] A. Sparmann, M. van Lohuizen, Polycomb silencers control cell fate, development and cancer, Nat. Rev. Cancer 6 (11) (2006) 846–856.
[3] B. Schuettengruber, et al., Genome regulation by polycomb and trithorax proteins, Cell 128 (4) (2007) 735–745.
[4] L. Ringrose, R. Paro, Polycomb/Trithorax response elements and epigenetic memory of cell identity, Development 134 (2) (2007) 223–232.
[5] J.A. Simon, R.E. Kingston, Mechanisms of polycomb gene silencing: knowns and unknowns, Nat. Rev. Mol. Cell Biol. 10 (10) (2009) 697–708.
[6] Y.B. Schwartz, V. Pirrotta, Polycomb silencing mechanisms and the management of genomic programmes, Nat. Rev. Genet. 8 (1) (2007) 9–22.
[7] X. Shen, et al., EZH1 mediates methylation on histone H3 lysine 27 and complements EZH2 in maintaining stem cell identity and executing pluripotency, Mol. Cell 32 (4) (2008) 491–502.
[8] R. Margueron, et al., Ezh1 and Ezh2 maintain repressive chromatin through different mechanisms, Mol. Cell 32 (4) (2008) 503–518.
[9] L.A. Boyer, et al., Polycomb complexes repress developmental regulators in murine embryonic stem cells, Nature 441 (7091) (2006) 349–353.
[10] T.I. Lee, et al., Control of developmental regulators by Polycomb in human embryonic stem cells, Cell 125 (2) (2006) 301–313.
[11] M. Ku, et al., Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains, PLoS Genet. 4 (10) (2008) e1000242.
[12] S.L. Squazzo, et al., Suz12 binds to silenced regions of the genome in a cell-type-specific manner, Genome Res. 16 (7) (2006) 890–900.
[13] B.E. Bernstein, et al., A bivalent chromatin structure marks key developmental genes in embryonic stem cells, Cell 125 (2) (2006) 315–326.
[14] T.S. Mikkelsen, et al., Genome-wide maps of chromatin state in pluripotent and lineage-committed cells, Nature 448 (7153) (2007) 553–560.
[15] F. Mohn, et al., Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors, Mol. Cell 30 (6) (2008) 755–766.
[16] L. Ringrose, et al., Genome-wide prediction of Polycomb/Trithorax response elements in Drosophila melanogaster, Dev. Cell 5 (5) (2003) 759–771.
[17] T. Fiedler, M. Rehmsmeier, jPREdictor: a versatile tool for the prediction of cis-regulatory elements, Nucleic Acids Res. 34(Web Server issue) (2006) W546–W550.
[18] B. Schuettengruber, et al., Functional anatomy of polycomb and trithorax chromatin landscapes in Drosophila embryos, PLoS Biol. 7 (1) (2009) e13.
[19] A. Hauenschild, et al., Evolutionary plasticity of polycomb/trithorax response elements in Drosophila species, PLoS Biol. 6 (10) (2008) e261.
[20] A. Sing, et al., A vertebrate Polycomb response element governs segmentation of the posterior hindbrain, Cell 138 (5) (2009) 885–897.
[21] C.J. Woo, et al., A region of the human HOXD cluster that confers polycomb-group responsiveness, Cell 140 (1) (2010) 99–110.
[22] Q. Zhou, et al., A gene regulatory network in mouse embryonic stem cells, Proc. Natl. Acad. Sci. U. S. A. 104 (42) (2007) 16438–16443.
[23] A. Tanay, et al., Hyperconserved CpG domains underlie Polycomb-binding sites, Proc. Natl. Acad. Sci. U. S. A. 104 (13) (2007) 5521–5526.
[24] G.C. Yuan, Targeted recruitment of histone modifications in humans predicted by genomic sequences, J. Comput. Biol. 16 (2) (2009) 341–355.
[25] J. Zhao, et al., Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome, Science 322 (5902) (2008) 750–756.
[26] J.L. Rinn, et al., Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs, Cell 129 (7) (2007) 1311–1323.
[27] A.M. Khalil, et al., Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression, Proc. Natl. Acad. Sci. U. S. A. 106 (28) (2009) 11667–11672.
[28] J. Kim, et al., An extended transcriptional network for pluripotency of embryonic stem cells, Cell 132 (6) (2008) 1049–1061.
[29] X. Chen, et al., Integration of external signaling pathways with the core transcriptional network in embryonic stem cells, Cell 133 (6) (2008) 1106–1117.
[30] T. Li, et al., CTCF regulates allelic expression of Igf2 by orchestrating a promoter-polycomb repressive complex 2 intrachromosomal loo, Mol. Cell. Biol. 28 (20) (2008) 6473–6482.
[31] V. Matys, et al., TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes, Nucleic Acids Res. 34(Database issue) (2006) D108–D110.
[32] H. Ogawa, et al., A complex with chromatin modifiers that occupies E2F- and Myc-responsive genes in G0 cells, Science 296 (5570) (2002) 1132–1136.
[33] C. Attwooll, et al., A novel repressive E2F6 complex containing the polycomb group protein, EPC1, that interacts with EZH2 in a proliferation-specific manner, J. Biol. Chem. 280 (2) (2005) 1199–1208.
[34] J.L. Brown, et al., An Sp1/KLF binding site is important for the activity of a Polycomb group response element from the Drosophila engrailed gene, Nucleic Acids Res. 33 (16) (2005) 5181–5189.
[35] J.M. Trimarchi, et al., The E2F6 transcription factor is a component of the mammalian Bmi1-containing polycomb complex, Proc. Natl. Acad. Sci. U. S. A. 98 (4) (2001) 1519–1524.
[36] D.P. Satijn, et al., The polycomb group protein EED interacts with YY1, and both proteins induce neural tissue in Xenopus embryos, Mol. Cell. Biol. 21 (4) (2001) 1360–1369.
[37] L. Atchison, et al., Transcription factor YY1 functions as a PcG protein in vivo, Embo J. 22 (6) (2003) 1347–1358.
[38] H. Chipman, E. George, R. McCulloch, Bayesian ensemble learning, Neural information processing systems, 2006.
[39] Q. Zhou, J.S. Liu, Extracting sequence features to predict protein-DNA interactions: a comparative study, Nucleic Acids Res. 36 (12) (2008) 4137–4148.
[40] R. Tibshirani, Regression Shrinkage and Selection via the Lasso, J. R. Stat. Soc.: Ser. B: Methodol. 58 (1) (1996) 267–288.
[41] V.N. Vapnik, *Statistical learning theory*. Adaptive and learning systems for signal processing, communications, and control, Wiley, New York, 1998, p. 736, xxiv.
[42] A. Sandelin, et al., JASPAR: an open-access database for eukaryotic transcription factor binding profiles, Nucleic Acids Res. 32 (Database issue) (2004) D91–D94.
[43] D.E. Newburger, M.L. Bulyk, UniPROBE: an online database of protein binding microarray data on protein-DNA interactions, Nucleic Acids Res. 37(Database issue) (2009) D77–D82.
[44] X. Shen, W. Kim, Y. Fujiwara, Y. Liu, M.R. Mysliwiec, G.C. Yuan, Y. Lee, S.H. Orkin, Jumonji modulates Polycomb activity and self-renewal versus differentiation of stem cells, Cell 139 (7) (2009) 1303–1314.
[45] Y.B. Schwartz, et al., Genome-wide analysis of Polycomb targets in Drosophila melanogaster, Nat. Genet. 38 (6) (2006) 700–705.
[46] P.J. Hurlin, et al., Mga, a dual-specificity transcription factor that interacts with Max and contains a T-domain DNA-binding motif, EMBO J. 18 (24) (1999) 7019–7028.
[47] S. Gaubatz, J.G. Wood, D.M. Livingston, Unusual proliferation arrest and transcriptional control properties of a newly discovered E2F family member, E2F-6, Proc. Natl. Acad. Sci. U. S. A. 95 (16) (1998) 9190–9195.
[48] N.B. La Thangue, Transcription. Chromatin control–a place for E2F and Myc to meet, Science 296 (5570) (2002) 1034–1035.
[49] F. Lan, et al., A histone H3 lysine 27 demethylase regulates animal posterior development, Nature 449 (7163) (2007) 689–694.
[50] M.G. Lee, et al., Demethylation of H3K27 regulates polycomb recruitment and H2A ubiquitination, Science 318 (5849) (2007) 447–450.
[51] M. Guttman, et al., Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals, Nature 458 (7235) (2009) 223–227.
[52] A. Siepel, et al., Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes, Genome Res. 15 (8) (2005) 1034–1050.
[53] W. Hastings, Monte Carlo samplling methods using Markov chains and their applications, Biometrika 87 (1) (1970) 97–109.
[54] S. Mahony, P.V. Benos, STAMP: a web tool for exploring DNA-binding motif similarities, Nucleic Acids Res. 35(Web Server issue) (2007) W253–W258.