Published in final edited form as:

J Proteome Res. 2013 March 1; 12(3): 1467–1477. doi:10.1021/pr3011205.

Recognition of Methylated Peptides by *Drosophila melanogaster* Polycomb Chromodomain

Richard S. L. Stein^{†, #}, Nan Li^{†, #}, Wei He^{†, #}, Elizabeth Komives[†], and Wei Wang^{†, ‡}

Wei Wang: wei-wang@ucsd.edu

[†]Department of Chemistry and Biochemistry, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093-0359, United States

[‡]Department of Cellular and Molecular Medicine, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093-0359, United States

Abstract

Lysine methylation is one of the important post-translational modifications (PTMs) that regulate protein functions. Up to now, proteomic identification of this PTM remains a challenge due to the lack of effective enrichment methods in mass spectrometry experiments. To address this challenge, we present here a systematic approach to predicting peptides in which lysine residues may be methylated to mediate protein–protein interactions. We used the chromodomain of the polycomb protein in *Drosophila melanogaster* as a model system to illustrate the success of this approach. We started with molecular dynamics simulations and free energy analyses on the histone peptides complexed with the polycomb chromodomain to understand how the binding specificity is achieved. We next conducted virtual mutagenesis to quantify each domain and peptide residue's contribution to the domain-peptide recognition, based on which scoring scheme was developed to evaluate the possibility of any lysine-containing peptides to be methylated and recognized by the chromodomain. A peptide microarray experiment on a panel of conserved histone peptides showed a satisfactory prediction accuracy of the scoring scheme. Next, we implemented a bioinformatics pipeline that integrates multiple lines of evidence including conservation, subcellular localization, and mass spectrometry data to scan the fly proteome for a systematic identification of possible methyllysine-containing peptides. These putative chromodomain-binding peptides suggest unknown functions of the important regulator protein polycomb and provide a list of candidate methylation events for follow-up investigations.

© 2013 American Chemical Society

Correspondence to: Wei Wang, wei-wang@ucsd.edu.

[#]**Author Contributions:** These authors made equal contribution.

Notes: The authors declare no competing financial interest.

Keywords

molecular dynamics; MM-GBSA; mutation matrix; peptide microarray; methylated lysine

Introduction

Proteins can exhibit specificity in their binding to methylated lysines — the *Drosophila melanogaster* protein polycomb, a member of the chromodomain group of proteins that regulate chromatin structure and that are involved with genetic repression,¹ binds strongly to trimethylated Lys27 of histone protein H3 (H3K27me3) but only weakly to trimethylated H3K9.^{2,3} Post-translational methylation has also been found on proteins other than histone peptides. The expression of the transient protein p53, for instance, which is involved in tumor suppression, is regulated by lysine methylation effected by the writer protein Set9.⁴ In addition, proteins that interact with histones can themselves be marked by lysine methylation, which can affect their interactions with histone peptides.⁵ While very little is known about the ability of chromodomain proteins to recognize methylated lysines on nonhistone peptides, the polycomb-group protein (PcG) Pc2 has been observed to interact with lysine residues modified by attachment of a small ubiquitin-like modifier (SUMO).^{6,7} This phenomenon suggests that PcGs could have functions of interest that are currently unknown besides genetic regulation by modified histone recognition.

Mass spectrometry can detect many types of post-translation modifications (PTMs) and the proteins that manipulate them, but its use in the specific case of methylation is limited due to the lack of effective enrichment methods. In theory, binding assay experiments can determine the binding affinity of a protein to a wide variety of methylated-lysine peptides, but experimental data has been scarce. By combining a computational study with a scan of the fruit fly's proteome, we hope to shed light on both the total class of proteins that can interact with methylated lysine and the methylated-lysine peptides that can be recognized by

PcG proteins, and acquire a more complete understanding of the functions of this group of proteins.

Methods and Materials

I. Modeling Procedure

The polycomb–H3K27me3 complex structure was obtained from the Protein Data Bank (PDB).⁸ Its PDB entry code is 1PDQ. The other methylation states were formed by manually removing methyl carbons in turn from the structure data in the PDB. The module of LEAP in the AMBER 9.0⁹ software package was used to prepare the models for molecular dynamics simulation. The methylated lysine residues were not recognized by the standard residue library of LEAP and were treated as nonstandard residues. The charges for all nonstandard residues were calculated by Gaussian03,¹⁰ using the HF/6-31G* basis set, and subsequently fitted with RESP.¹¹ The partial charges on all nonstandard residue atoms as calculated by Gaussian03 and RESP are listed in Table S1. The nonstandard residues were all handled by Gaussian03 individually and capped¹² by ACE and NME residues so that they could be treated as complete molecules. The FF03 force field¹³ was used for all atoms in the model, except for the methylated zeta nitrogens of H3K27, for which the parameters of the GAFF force field¹⁴ were used. Each polycomb-histone complex was placed in a rectangular, periodic box of TIP3P water molecules.¹⁵ The dimensions of the box extended 10 Å past the solute system on all sides. When necessary, Cl[−] counterions were added to neutralize the system for each model.

II. Molecular Dynamics Simulation

Each system was minimized in two stages; in the first stage, the water molecules and counterions were relaxed while organic molecules were restrained by a harmonic force of 100 kcal/mol Å². In the second stage, the restraints were released and the entire system was minimized. Each stage consisted of 400 steps of steepest descent followed by 1600 steps of conjugate gradient minimization. After the minimization stages, the systems were heated from 0 to 300 K over 30 ps. The temperature of 300 K and pressure of 1 atm were maintained by a Berendsen thermo- and barostat with a coupling time¹⁶ of 0.2 ps. The SHAKE algorithm was used to restrain all bonds involving hydrogen.¹⁷ Initial temperatures were assigned randomly according to the Maxwell–Boltzmann distribution at the start of the heating run. The systems were then equilibrated for 200 ps followed by production runs lasting 4 ns. During the equilibration and production runs, the heavy atoms in the polycomb protein were restrained by a harmonic force of 10 kcal/mol Å² in order to ensure complex stability. In addition, the charged residues in the PDB crystal structure were checked with PROPKA.^{18–21} All of the Asp and Glu residues in the chromodomain have a pK_a below 5, both near the binding interface and away. Therefore, it is extremely unlikely that these residues will exhibit a protonated state in the complex.

III. Trajectories Are Most Stable at Moderate Run Length

In order to determine the optimal interval within the trajectory for calculating binding free energies, we calculated the root-mean-square deviation (RMSD) for the complex, the H3 peptide, and the polycomb protein over the equilibration and production runs (see Figure

1a). The coordinates of the α -carbon atoms of the residues were the data used to calculate the RMSD values. For the protein, the RMSD remained below 0.20 over the entire 4.2 ns trajectory, as expected (because of the harmonic restraints). Therefore, virtually all fluctuation in the RMSD for the complex is due to the movement of the H3 peptide which was not restrained. Since the RMSD appears to be most stable in the interval between 1.0 and 2.0 ns in the production run (1200–2200 ps in the figure), this interval was chosen for binding free energy calculation analysis. In order to show that the stability of the peptide conformation is not drastically affected by a point mutation in the polycomb protein, we also provide the RMSD of the complex for the protein mutation Y26G in Figure 1b.

IV. Determination of Optimal Dielectric Constant ϵ

The calculated binding free energies for the four methylation states of H3K27 complexed with polycomb are given in Table 1. Three different internal dielectric constants (1.0, 2.0 and 4.0) were supplied in the calculations. When the simulation binding free energies are graphed against the logarithms of the experimental dissociation constants ($\ln K_d$), the correlation coefficients for $\epsilon = 1.0, 2.0$, and 4.0 are, respectively, 0.8009, 0.9375, and 0.9851. Therefore, for this system, the binding free energies agree most closely with experiment for a dielectric constant of 4.0, including the reversal phenomenon of H3K27me2 binding more weakly to polycomb than H3K27me1. Surprisingly, the van der Waals component of the interaction is 1.21 kcal/mol weaker for H3K27me2 than H3K27me1. This phenomenon may be due to the dimethylated cation being more energetically stable than the monomethylated one, and thus having less of a tendency to bind to the hydrophobic pocket of the chromodomain.

While the AMBER force field parameters are designed for electrostatic conditions close to vacuum ($\epsilon = 2.0$ or lower), we feel that a value of 4.0 for ϵ is justifiable for the polar interface of this system, and higher values of ϵ can be used for calculations of the binding free energy.²² A complete listing of the interaction components of the binding free energy for all H3K27 methylation states is given in Table S2.

V. Mutation Matrix Procedure

The mutation matrix was formed by identifying the 21 most critical binding residues in polycomb for binding to H3K27me3, according to Table 2. Each of these residues was mutated to all of the other 19 naturally occurring amino acids. This produced $21 \times 19 = 399$ separate models, each a point mutation of the wild-type complex between polycomb and histone H3. In a separate operation, the sequence of the protein was preserved, and each of the H3 residues featured in the model, except for H3K27, was mutated to each of the other 19 amino acids. This produced another $8 \times 19 = 152$ separate models. The methodology is similar to that used in earlier research on the enzyme spermine synthase,²³ with the distinction that our operative binding free energies are relative to the wild-type, as opposed to a “population mean”.

The software program SCAP²⁴ was used to perform the point mutations. The standard procedure of minimization, heating and equilibration (see subsection II) was performed on each of the models, followed by 2 ns of production run. Binding free energies were

calculated from the second half (the 1.0–2.0 ns interval) of the production run for each model.

VI. Trajectory Analysis and Free Energy Calculation

The RMSD for each system during the equilibration and production run phases of the MD was calculated by the PTRAJ module in the AMBER package.⁹ As a representative sample, the RMSD for the polycomb/H3K27me3 wild-type is shown in Figure 7. On the basis of the stabilization of the RMSD after about 1000 ps, the interval for the binding free energy calculation was chosen for the latter half of each production run, between 1.0 and 2.0 ns. MM-GBSA^{25–27} was used to calculate the binding free energies for each polycomb—histone complex using the *mm_gbsa* module in AMBER 9.0, according to the general procedure given in refs 28 and 29. The binding free energies, G_{bind} , were calculated from 100 snapshots taken at equal intervals between the 1.0 and 2.0 ns marks of the production run trajectory. G_{bind} was calculated according to

$$\Delta G_{\text{bind}} = \langle G_{\text{complex}} \rangle - (\langle G_{\text{protein}} \rangle + \langle G_{\text{peptide}} \rangle)$$

Here, $\langle G_{\text{complex}} \rangle$, $\langle G_{\text{protein}} \rangle$, and $\langle G_{\text{peptide}} \rangle$ are the respective individual free energies of the complex, protein and peptide, and each term is calculated by summing the contributions from electrostatic potential E_{elec} , van der Waals potential E_{vdw} , and solvation free energy G_{solv} . In turn, G_{solv} is the sum of the polar contribution G_{polar} and nonpolar contribution G_{nonpolar} to the solvation free energy:

$$G_{\text{complex/protein/peptide}} = E_{\text{elec}} + E_{\text{vdw}} + G_{\text{solv}}$$

$$G_{\text{solv}} = G_{\text{polar}} + G_{\text{nonpolar}}$$

G_{polar} was calculated by setting the value of IGB to 2, activating the generalized Born parameters of Onufriev et al.³⁰ G_{nonpolar} was estimated to be 0.0072 times the solvent-accessible surface area (SASA), as measured by the LCPO method.³¹ In the calculation of the SASA, the probe radius³² was set to 1.4 Å. Three different values (1.0, 2.0, and 4.0) were used for the interior dielectric constant ϵ in separate free energy calculations, while the external dielectric constant was given throughout as 80.0.

To investigate the molecular basis of recognition specificity, we conducted component decomposition of the binding free energy. The contributions to the binding free energy were also broken down for individual residues, and by component (van der Waals, electrostatic, and solvation) for each of the complexes. Decompositions were performed by invoking the DECOMP module in the MM-GBSA program within AMBER. Each H3 or polycomb residue was evaluated according to its overall contribution to the binding interaction.

VII. Proteomics and Mass Spectrometry Analysis

The *D. melanogaster* proteome data were downloaded from Uniprot/Swissprot (2009–06–16).^{43,44} In total, there are 3047 protein sequences, from which we extracted all 9 amino acid long peptides containing a lysine at the eighth position from the N-terminus, in order to obtain direct comparison with H3 as it appears in the PDB crystal structure.

For each candidate peptide, a mutation score was calculated as

$$\text{mutation score} = \sum_{i=1, i \neq 8}^9 M(i, A_i)$$

where i is the amino acid position on the peptide (position 8 of the lysine was omitted), M is the mutation matrix generated from virtual mutagenesis analysis, A_i is the i th amino acid on the candidate peptide.

The conservation score considered the conservation of both the entire peptide and the methyllysine site. We downloaded all the protein sequences in 12 fly proteomes (Table 3) from *flybase*.³³ For each *D. melanogaster* protein that contains a candidate peptide, its homologous proteins in the other 11 fly proteomes were found using NCBI BLAST.³⁴ Multiple sequence alignment was then generated for the 12 homologous proteins using CLUSTALW2,³⁵ and the overall conservation score was calculated as

$$\text{overall conservation score} = \sum_{i=1, i \neq 8}^9 \sum_{m=1}^{11} \text{BLOSUM}(A_i, A_i^m)$$

where m is the index of the 11 *Drosophila* proteomes, i is the amino acid position in the candidate peptide, BLOSUM is the BLOSUM62 mutation matrix,³⁶ A_i is the i th amino acid in the candidate peptide, and A_i^m is the i th amino acid in the homologue of the m th *Drosophila* proteome. The conservation score for the methyllysine site was defined as the number of lysine occurring at position 8 in the multiple sequence alignment. An arbitrary conservation score cutoff of 500 was used in the filtering process. Table S3 lists the investigated peptides in the fruit fly proteome and their conservation scores along with their mutation matrix scores.

We collected 817 332 *D. melanogaster* mass spectra from the PRIDE database.³⁷ All the *D. melanogaster* proteins in Uniprot/Swissprot were used as the reference database when scanning possible methyllysines by any of the three methods, X! TANDEM,³⁸ OMSSA,³⁹ and INSPECT.⁴⁰ For the scanning parameter, the parent mass error was set to 1 Da, fragment mass error was set to 0.2 Da, and modifications under consideration included mono-, di-, and trimethylation. If a lysine was identified as methylated by any of the three programs, we considered the peptide as potentially methylated to keep it in the analysis. In total, we found 10 838 possible methyllysines in the *D. melanogaster* proteome.

VIII. Peptide Microarray Measurement and Analysis

The chromodomain (a.a. 1–90) of polycomb protein was expressed as GST-tagged fusion protein and purified as described in ref 41. The purity of the protein was examined by SDS-PAGE electrophoresis followed by both Coomassie staining and Western blot using an Anti-GST-HRP conjugate (Santa Cruz Biotechnology). The concentration of purified protein was determined by BCA assay (Amresco). A total of 46 trimethylated histone peptides were synthesized by Sigma Aldrich (desalted, mass spectrometry checked). The peptides were then printed as triplets onto glass slides (ArrayIt), together with a Cy3 marker and an anti-GST (mouse monoclonal) antibody (Thermo) as references.

The peptide microarray was rinsed with TBST buffer (25 mM Tris, 125 mM NaCl, 0.05% Tween-20, pH 8) followed by blocking with 5% nonfat milk in TBST (room temperature 1 h or 4 °C overnight). The slide was then incubated at 4 °C with polycomb-GST fusion protein at a final concentration of 5 mM in 5% nonfat milk/TBST for 6 h. After being washed three times for 10 min each with TBST, the slide was incubated with an anti-GST mouse monoclonal IgG antibody (Thermo) at a final concentration of 1 $\mu\text{g/mL}$ in 5% nonfat milk/TBST and shaken gently for 1 h at room temperature. A secondary antimouse IgG Dylight-488 conjugated antibody (Thermo) was added to a final concentration of 0.1 $\mu\text{g/mL}$ after three more cycles of 10 min washing with TBST. The slide was shaken for 1 h at room temperature and washed three times with TBST.

The dried microarray slides were scanned using a Hamamatsu NanoZoomer 2.0HT slide scanning system (Neuroscience Light Microscopy Facility, UCSD). Data quantification were processed using the microarray processing software ImageJ,⁴² where the fluorescent intensity of a microarray spot was defined as its own intensity minus the background intensity around it on the scanned image.

Results

I. Overall Procedure

Molecular dynamics (MD) simulations were performed on *D. melanogaster* polycomb protein complexed with a peptide from the histone protein H3. First, the binding affinities of wild-type polycomb to unmodified H3 and H3K27me1, H3K27me2, and H3K27me3 were determined by simulations and free-energy calculation. The models were derived from crystal structures; the structures of the polycomb–H3 complexes are shown in Figure 2. The contributions of each individual residue in both polycomb and H3 to the binding interaction were derived by energy decomposition analysis (see Materials and Methods), leading to the discovery of the most critical polycomb residues. Next, mutation matrices were formed both for polycomb and for H3 to estimate the effect of point mutation in each residue critical for binding. Subsequently, peptide microarray binding data were used to verify our prediction of binding to polycomb on fly histone peptides, which demonstrated the satisfactory prediction ability for the H3 mutation matrix. Finally, a scan of the fruit fly proteome was conducted to find additional peptides with methylated lysine and estimate their binding affinity relative to the wild-type H3 peptide by using the mutation matrix. A further refinement was conducted through a combined analysis of data from conservation, subcellular location, and mass

spectrometry evidence. The final list of possible polycomb binders provided insight into the quantity and character of proteins that can interact with polycomb.

II. Simulation Results

A. Residues in Polycomb Protein Critical for Binding to H3—To shed light on how the methylated peptides are recognized by the polycomb chromodomain, we first identified the critical residues for binding. Having determined that $\epsilon = 4.0$ is the optimal dielectric constant for deriving accurate binding energies with this model (see Materials and Methods, subsection IV), we performed residue decomposition²⁸ on the polycomb/H3K27me3 complex for this value. The results of this calculation are given in Table 2. The two most important residues in polycomb for the binding interaction are Tyr26 (−4.82 kcal/mol) and Trp47 (−5.08 kcal/mol), which are two of the three bulky hydrophobic residues that form a hydrophobic cage for van der Waals attraction of H3K27me3. The third residue of the cage, Trp50, is also high on the critical residue list (−2.36 kcal/mol). Other key hydrophobic residues in polycomb for the interaction are Leu24 (−2.57 kcal/mol), Val25 (−3.29 kcal/mol), Leu64 (−2.84 kcal/mol), and Leu68 (−2.26 kcal/mol), while the most important charged polycomb residues are Asp23 (−1.82 kcal/mol), Glu29 (−1.22 kcal/mol), Glu58 (−1.18 kcal/mol), Asp65 (−2.59 kcal/mol), and Arg67 (−1.95 kcal/mol). A crucial neutral polar residue is Asn62 (−3.15 kcal/mol). Apart from H3K27me3 (−13.69 kcal/mol), the most important residues in H3 for the binding interaction are H3R26 (−8.41 kcal/mol), H3K23 (−6.66 kcal/mol), H3A25 (−5.41 kcal/mol), H3A24 (−4.27 kcal/mol), H3S28 (−2.09 kcal/mol), and H3T22 (−1.78 kcal/mol). The most important electrostatic interactions in the polycomb/H3 complex, from visualization of the trajectory, are Asp23 with H3R26; Glu29 with H3K23; Glu58 with H3S28; and Asp65 with H3K23.

B. Polycomb Protein Mutation Matrix—The mutation matrix for the polycomb protein in complex with H3K27me3 is given in Figure 3. The following observations can be made about trends in the mutation matrix. First, mutations of the eight or nine residues in polycomb (Tyr26, Trp47, Val25, Asn62, Leu64, Asp65, Leu24, Trp50, and possibly Arg67) that are most important to the binding interaction generally weaken the binding affinity, especially in the case of Trp47. Second, mutations in the lower half of the matrix (residues less critical for binding) tend to make the interaction with H3K27me3 more favorable, with the exception of Glu58. Third, mutations to positively charged amino acids (lysine and arginine) as well as to glycine, proline, and to a lesser extent, alanine, tend to make the binding interaction more unfavorable. Finally, mutations to negatively charged residues (aspartic and glutamic acid) tend to make the binding interaction more favorable.

C. Histone H3 Peptide Mutation Matrix—Three aspects of this problem are particularly worthy of investigation; first, to find polypeptides other than histones that contain methylated lysine; second, to understand the mechanisms by which a chromodomain protein recognizes a peptide containing a methylated lysine residue, and finally, to obtain as much functional information as possible about the polypeptides that bind to chromodomain protein. Therefore, in another set of simulations, point mutations were performed on residues of the H3 peptide in turn (positions 20 to 28, except for H3K27me3 which was kept conserved), while the polycomb protein was held to the wild-type. The graphical mutation

matrix is given in Figure 4 (see Table S4 for the numerical mutation matrix). In general, mutations of peptide residues tend to have a more favorable effect on the binding interaction than mutations of polycomb residues, especially those which are foremost in importance for binding to H3. Several trends are evident from the mutation matrix data: first, mutations of H3R26 tend to have a weakening effect on the binding affinity; second, mutations of H3A25 tend to make the binding interaction more favorable; third, mutations of H3 residues to Gly, Asp, Asn, and Pro make the binding energy more unfavorable, especially in the near vicinity of H3K27me3; and fourth, mutations of H3 residues to tyrosine and tryptophan have a strengthening effect on the binding interaction.

III. Experimental Validation of Theoretical Predictions

A. Experimental Identification of Binders—Our goal is to find peptides in the fly proteome that contain a methylated lysine residue that can be recognized by the polycomb chromodomain (polycomb-binding peptides, or PBPs). The template peptide H3K27me3 is nine amino acids long and has the sequence of LATKAARKS, the second lysine of which is H3K27. We thus searched for peptides of nine-amino-acid length with a lysine on the eighth residue that show similar or better binding than that of H3 (H3K27me3). For this purpose, we implemented a three-stage strategy to narrow down the searching space in a stepwise manner. In addition to removing peptides with a mutation score worse than the H3K27me3 peptide, we implemented filters of sequence conservation, subcellular location, and mass-spectrometry evidence to further reduce false positives (see below for details).

There are a total of 46 histone peptide sequences (nine a.a. long) with a single lysine-trimethylation site which are conserved between the human and fruit fly H2, H3, and H4 proteins (Table S5). These were screened for polycomb protein chromodomain binding by peptide microarray. Each peptide sequence was screened as triplets, and the fluorescence intensity was quantified with background subtraction individually for each spot from the microarray images by ImageJ,⁴² and the whole data set was fit to a mixed Gaussian distribution model (Figure 5, Table 4; see Table S6 for the intensity data for individual polycomb-peptide complexes). A sequence was identified as a positive binder when at least two out of the three microarray spots exhibited a fluorescence intensity that made a *p*-value cutoff of 0.001 from the background Gaussian distribution in the mixed model. Twelve out of the 46 sequences were identified as positive binders under these criteria. Ten of these microarray-identified binders also yielded a high mutation matrix score in an independent prediction (Table 5).

B. Prediction Performance of Polycomb Mutation Matrix on Fly Histone Peptides—For a thorough and critical assessment of the prediction performance of the mutation matrix, we used microarray to test the binding between polycomb and the 28 histone peptides that it was predicted to bind to, as well as 18 peptides that were not predicted to bind. By doing this, both true positive and true negative can be identified and a comprehensive analysis of prediction performance is feasible.

The microarray result identified 12 peptides as array-positive and 34 as array-negative with a *p*-value cutoff of 0.001. A peptide with a mutation matrix score equal to or more favorable

than the wild-type H3 peptide was predicted to be polycomb-binding. Among 12 true binders, 10 were predicted as binders by the mutation matrix score; among 34 true nonbinders, the mutation matrix predicted 16 as nonbinders, resulting in a sensitivity and accuracy of $10/12 = 0.833$ and $(10 + 16)/46 = 0.565$ respectively and an area under the receiver operating curve (ROC) of 0.66. The result is satisfactory considering that we did not construct any complicated feature structure and used one single score to discriminate binders against non-binders (Figure 6). To further justify the prediction ability of the mutation score, we also calculated its average for both binders and nonbinders. For 12 binders, the average score is -5.48 with a standard deviation of 8.31; for 34 nonbinders, the average score is 0.16 with a standard deviation of 10.72. The p -value of the t test between the two distributions is 0.037 using the R software package, which demonstrates a strong trend of discrimination between binder and nonbinder by the mutation matrix score.

IV. Proteomics and Gene Ontology Analysis

On the basis of the encouraging agreement between the predictions of our mutation matrix and the experimental data, our simulations can be used as a tool in scanning the proteome for polycomb-binding peptides. Since there is no systematic way to identify methylated lysines in a high-throughput manner, our computational predictions provide a candidate list of putative methylations to serve as a starting point for experimental study.

A. Proteome Scan, Filters, and General Analysis Results—We searched for candidate PBPs in the *D. melanogaster* proteome. The overall scheme is shown in Figure 7. We first identified 103 524 candidate peptides, each of which is a sequence of nine amino acids for which the eighth residue is lysine. Mutation scores were calculated for all the peptides, and 58 131 peptides with favorable binding energy (mutation score no higher than zero) were kept.

If a peptide binds to the polycomb chromodomain, it is expected to be conserved through evolution since mutations may disrupt the functional interaction. It is thus reasonable to remove peptides that are not conserved at all the nine amino acids or at the methylated lysine (the eighth) position. By considering the conservation score for the entire peptide and the lysine site, we kept 9919 peptides for further analysis.

Since the polycomb protein is localized in the nucleus and functions as a reader of chromatin modifications, we focus on identifying its interacting partners that are also compartmented to the nucleus. For the proteins without annotation of their subcellular location, we chose to keep them as their predicted interactions with polycomb may shed light on their functional roles. As a result, 5792 peptides from proteins with a subcellular location annotated to be either “nucleus” or “unknown” in Uniprot's general annotation^{43,44} passed this filter.

Because recognition by polycomb protein requires the lysine residue to be methylated, the mass-spectrometry evidence of lysine-methylated peptides can provide supporting information for PBP identification. We collected all the publicly available mass spectra and performed a scan of the *D. melanogaster* proteome to identify lysine methylation. Even though the mass spectrometry experiments were not designed to detect methyllysine, such additional evidence can help to identify the most possible binding peptides of the polycomb

chromodomain. We found that 539 among the 5792 peptides showed possible addition of the methyl group to lysine as suggested by the mass spectrometric signals, and as such were the most likely in the set to be PBPs.

To discover unknown functions of the polycomb protein, we conducted GO clustering analysis using DAVID⁴⁵ on the 357 proteins containing the 539 filtered peptides (Table S7). We found that GO terms of “regulation of transcription” and “histone modification” are the most enriched, accounting for five out of six top enriched GO clusters. Moreover, several important proteins were included in the list, such as mediator subunits 23 and 31, which involves recruitment of RNA polymerase II to promoter region that leads to the start of transcription, and sex combs extra (Sce), subunit of PRC1 complex regulating gene repression in the context of methylated H3K9 and H3K27. Also, several histone-related methyltransferase, demethyltransferase, and acetyltransferase can be found in the list, such as Art4, Jhd1, Kdm4B, and Tip60.

B. Chromatin Modification—Polycomb's known function is related to H3K9me3/H3K27me3 mediated gene repression by PRC1. From our analysis, we found that polycomb protein could also involve in gene regulation through chromatin modification with several other proteins, including histone methyltransferase, acetyltransferase, demethyltransferase, and several nucleosome remodeling proteins.

ASH1, CARM1, MOF, and KDM4B are related to gene regulation by either adding histone marks or remove marks. ASH1 and CARM1 are two methyltransferases. ASH1 specifically trimethylates H3K4, an activation mark, and prevents inappropriate polycomb silencing of important genes during development.^{46,47} CARM1 methylates H3R17 and activates transcription through chromatin remodeling.⁴⁸ MOF is a probable histone acetyltransferase to H4K16 on male X chromosome to balance the gene expression between male and female.⁴⁹ KDM4B is a probable demethyltransferase that demethylates methylated H3K9 and H3K36. All these proteins are related to either adding activation marks, H3K4me3 (ASH1), H3R17me2 (CARM1), and H4K16ac (MOF), removing repression mark, H3K9me3 (KDM4B), or removing activation mark, H3K36me3 (KDM4B). All our finding indicates that polycomb may be involved in a much more complicated gene regulation network related to chromatin modification than what we have known so far.

RING1 and CAF1, two other putative binding targets of polycomb, are components of PRC1⁵⁰ and PRC2,⁵¹ respectively. Intriguingly, polycomb itself is also a component of PRC1. We speculated that the binding to RING1 is related to recruit polycomb to form PRC1 complex and the binding to CAF1 is related to recruit PRC1 to PRC2.

CAF1 is also a component of chromatin assembly factor 1 (CAF-1), the nucleosome remodeling and deacetylase complex (NuRD), and the nucleosome remodeling factor (NuRF).^{52–56} IPYR, another component of NuRF,^{54,57} is also included in the putative binder list. These interactions inferred that polycomb may have further function in chromatin modification related gene regulation.

C. Transcription Regulation—We were surprised to find that our potential polycomb binders include several transcription regulation proteins. Most of these proteins are involved in the *Drosophila* development or gene transcription. Putative polycomb binders that are functional in development include hemeotic gene (TSH),⁵⁸ eye development (SRC42 and BARH1),^{59,60} neural development (SOS and PNT1),^{61,62} and other development-related genes (VAV, DORS, EAF, CIC, and APTE).^{63–67} Putative polycomb binders involving transcription include TAF,⁶⁸ the largest component of the TFIID basal transcription factor complex; MED23 and MED31, components of Mediator complex and coactivators of nearly all Pol II-dependent genes; RPA2, the second largest component of RNA polymerase I; RPB2 and RPB9, components of RNA polymerase II; and RPC6, component of RNA polymerase III. Until now, little is known about the function of the interaction of polycomb in transcription regulation. Although we can speculate that polycomb may regulate the gene transcription process by binding to these important protein factors, further experiments are still needed to explore the unrevealed realm for polycomb and its regulatory functions.

Discussion

We investigated the polycomb–histone H3 complex with the use of MD. By specifying the dielectric coefficient in the calculation of the binding free energy, we were able to replicate the relative binding affinities of polycomb to H3K27me3, H3K27me2, H3K27me1, and unmodified H3, including the counterintuitive phenomenon of the chromodomain binding more strongly to H3K27 when monomethylated than dimethylated. We performed residue decomposition to identify the residues in polycomb most critical for binding to H3K27me3, as well as the relative contributions of the H3 residues to the binding interaction. These residues formed the basis of a mutation matrix, in which we analyzed the effect of mutating any single one of these residues to other amino acids.

With an estimate of the effect of mutating a given residue in the polycomb–H3K27me3 complex, we scanned the *D. melanogaster* proteome to find peptides containing methylated lysine that could potentially be recognized by polycomb chromodomain, apart from histones. Coupled with additional criteria including conservation, subcellular localization and mass spectrometry data, we have identified a set of nonhistone peptides that likely interacts with the polycomb protein mediated by chromodomain-methyllysine recognition. Considering the lack of proteomic approaches to determine methylated lysine, our approach not only provides a promising alternative to suggest candidates for follow-up investigation of the biological significance of methylation in the polycomb partner proteins but also reveals unknown functions of the important transcriptional regulator.

In our previous work, we have developed computational methods to identify unmodified peptides recognized by proteins such as those including the SH3 domain^{69,70} or regulatory subunit of protein kinase A (PKA).⁷¹ The present study represents a significant expansion in searching for methyllysine mediated protein–protein interactions in the proteome. More importantly, our approach can be readily applied to identify any PTMs and suggest their possible interacting partners, which will open a new avenue of illustrating the mechanisms of how a specific PTM regulates cellular functions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was partially supported by NIH (R01GM085188 to W.W.).

References

1. Paro R, Hogness DS. The Polycomb protein shares a homologous domain with a heterochromatin-associated protein of *Drosophila*. *Proc Nat Acad Sci U S A*. 1991; 88:263–267.
2. Fischle W, et al. Molecular basis for the discrimination of repressive methyl-lysine marks in histone H3 by Polycomb and HP1 chromodomains. *Genes Dev*. 2003; 17:1870–1881. [PubMed: 12897054]
3. Min J, Zhang Y, Xu RM. Structural basis for specific binding of Polycomb chromodomain to histone H3 methylated at Lys 27. *Genes Dev*. 2003; 17:1823–1828. [PubMed: 12897052]
4. Chuikov S, Kurash JK, et al. Regulation of p53 activity through lysine methylation. *Nature*. 2004; 432:353–360. [PubMed: 15525938]
5. Rathert P, Dhayalan A, et al. Protein lysine methyltransferase G9a acts on non-histone targets. *Nat Chem Biol*. 2008; 4:344–346. [PubMed: 18438403]
6. Colby T, Matthai A, Boeckelmann A, Stuible HP. SUMO-Conjugating and SUMO-Deconjugating Enzymes from Arabidopsis. *Plant Physiol*. 2006; 142:318–332. [PubMed: 16920872]
7. Miura K, Jin JB, Hasegawa PM. Sumoylation, a post-translational regulatory process in plants. *Curr Opin Plant Biol*. 2007; 10:495–502. [PubMed: 17720613]
8. Berman HM, Westbrook JD, Feng Z, Gilliland GL, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Res*. 2000; 28:235–242. [PubMed: 10592235]
9. Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, Merz KM, Onufriev A, Simmerling C, Wang B, Woods RJ. The Amber biomolecular simulation programs. *J Comput Chem*. 2005; 26:1668–1688. [PubMed: 16200636]
10. Frisch, MJ.; Pople, JA., et al. Gaussian 03. Gaussian Inc.; Wallingford, CT: 2004. Revision C. 02<http://www.gaussian.com/>
11. Bayly CI, Cieplak P, Cornell WD, Kollman PA. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges—the Resp model. *J Phys Chem*. 1993; 97:10269–10280.
12. Cieplak P, Cornell WD, Bayly C, Kollman PA. Application of the Multimolecule and Multiconformational RESP Methodology to Biopolymers: Charge Derivation for DNA, RNA, and Proteins. *J Comput Chem*. 1995; 16:1357–1377.
13. Duan Y, Wu C, Chowdhury S, Lee MC, Xiong G, et al. A point-charge force field for molecular mechanics simulations of proteins. *J Comput Chem*. 2003; 24:1999–2012. [PubMed: 14531054]
14. Wang JM, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and testing of a general amber force field. *J Comput Chem*. 2004; 25:1157–1174. [PubMed: 15116359]
15. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. *J Chem Phys*. 1983; 79:926–935.
16. Berendsen HJC, Postma JPM, Vangunsteren WF, Dinola A, Haak JR. Molecular-dynamics with coupling to an external bath. *J Chem Phys*. 1984; 81:3684–3690.
17. Ryckaert J, Ciccotti G, Berendsen HJC. Numerical-integration of cartesian equations of motion of a system with constraints: Molecular-dynamics of n-alkanes. *J Comput Phys*. 1977; 23:327–341.
18. Li H, Robertson AD, Jensen JH. Very fast empirical prediction and interpretation of protein pKa values. *Proteins*. 2005; 61:704–721. [PubMed: 16231289]
19. Bas DC, Rogers DM, Jensen JH. Very fast prediction and rationalization of pKa values for protein-ligand complexes. *Proteins*. 2008; 73:765–783. [PubMed: 18498103]

20. Olsson MHM, Søndergaard CR, Rostkowski M, Jensen JH. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa predictions. *J Chem Theory Comp.* 2011; 7(2): 525–537.
21. Søndergaard CR, Olsson MHM, Rostkowski M, Jensen JH. Improved treatment of ligands and coupling effects in empirical calculation and rationalization of pKa values. *J Chem Theory Comp.* 2011; 7(7):2284–2295.
22. Wang W, Kollman PA. Free energy calculations on dimer stability of the HIV protease using molecular dynamics and a continuum solvent model. *J Mol Biol.* 2000; 303:567–582. [PubMed: 11054292]
23. Zhang Z, Norris J, Schwartz C, Alexov E. In silico and in vitro investigations of the mutability of disease-causing missense mutation sites in spermine synthase. *PLoS One.* 2011; 6(5):e20373. [PubMed: 21647366]
24. Xiang Z, Honig B. Extending the accuracy limits of prediction for side-chain conformations. *J Mol Biol.* 2001; 311:421–430. [PubMed: 11478870]
25. Qiu D, Shenkin PS, Hollinger FP, Still WC. The GB/SA continuum model for solvation, a fast analytical method for the calculation of approximate Born radii. *J Phys Chem A.* 1997; 101:3005–3014.
26. Kollman PA, Massova I, Reyes CM, Kuhn B, Huo S, et al. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc Chem Res.* 2000; 33:889–897. [PubMed: 11123888]
27. Wang W, Donini O, Reyes CM, Kollman PA. Biomolecular simulations: recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions. *Annu Rev Biophys Biomol Struct.* 2001; 30:211–243. [PubMed: 11340059]
28. Hou T, Chen K, McLaughlin WA, Lu B, Wang W. Computational analysis and prediction of the binding motif and protein interacting partners of the Abl SH3 domain. *PLoS Comput Biol.* 2006; 2:0046–0055.
29. Hou T, McLaughlin WA, Wang W. Evaluating the potency of HIV-1 protease drugs to combat resistance. *Proteins.* 2008; 71:1163–1174. [PubMed: 18004760]
30. Onufriev A, Bashford D, Case DA. Modification of the generalized Born model suitable for macromolecules. *J Phys Chem B.* 2000; 104:3712–3720.
31. Weiser J, Shenkin PS, Still WC. Approximate solvent-accessible surface areas from tetrahedrally directed neighbor densities. *Biopolymers.* 1999; 50:373–380. [PubMed: 10423546]
32. Pellegrini E, Field MJ. A Generalized-Born Solvation Model for Macromolecular Hybrid-Potential Calculations. *J Phys Chem A.* 2002; 106:1316–1326.
33. Tweedie S, Ashburner M, Falls K, Leyland P, McQuilton P, Marygold S, Millburn G, Osumi-Sutherland D, Schroeder A, Seal R, Zhang H. FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res.* 2009; 37:D555–559. [PubMed: 18948289]
34. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990; 215:403–410. [PubMed: 2231712]
35. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* 2003; 31:3497–3500. [PubMed: 12824352]
36. Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A.* 1992; 89:10915–10919. [PubMed: 1438297]
37. Vizcaino JA, Cote R, Reisinger F, Foster JM, Mueller M, Rameseder J, Hermjakob H, Martens L. A guide to the Proteomics Identifications Database proteomics data repository. *Proteomics.* 2009; 9:4276–4283. [PubMed: 19662629]
38. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics.* 2004; 20:1466–1467. [PubMed: 14976030]
39. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH. Open mass spectrometry search algorithm. *J Proteome Res.* 2004; 3:958–964. [PubMed: 15473683]

40. Tanner S, Shu H, Frank A, Wang LC, Zandi E, Mumby M, Pevzner PA, Bafna V. InsPecT: identification of post-translationally modified peptides from tandem mass spectra. *Anal Chem.* 2005; 77:4626–4639. [PubMed: 16013882]
41. Hou T, Xu Z, Zhang W, McLaughlin WA, Case DA, Xu Y, Wang W. Characterization of domain-peptide interaction interface: a generic structure-based model to decipher the binding specificity of SH3 domains. *Mol Cell Proteomics.* 2009; 8:639–649. [PubMed: 19023120]
42. Rasband, WS. ImageJ. U.S. National Institutes of Health; Bethesda, Maryland, USA: 1997–2005. <http://rsbweb.nih.gov/ij/>
43. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* 2010; 38:D142–148. [PubMed: 19843607]
44. Jain E, Bairoch A, Duvaud S, Phan I, Redaschi N, Suzek BE, Martin MJ, McGarvey P, Gasteiger E. Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics.* 2009; 10:136. [PubMed: 19426475]
45. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nat Protocols.* 2009; 4(1):44–57. <http://david.abcc.ncifcrf.gov/home.jsp>.
46. Beisel C, et al. Histone methylation by the *Drosophila* epigenetic transcriptional regulator Ash1. *Nature.* 2002; 419:857–862. [PubMed: 12397363]
47. Klymenko T, et al. The histone methyltransferases Trithorax and Ash1 prevent transcriptional silencing by Polycomb group proteins. *EMBO Rep.* 2004; 5:373–377. [PubMed: 15031712]
48. Cakouros D, et al. An arginine-histone methyltransferase, CARMER, coordinates ecdysone-mediated apoptosis in *Drosophila* cells. *J Biol Chem.* 2004; 279:18467–18471. [PubMed: 14976192]
49. Hilfiker A, et al. Mof, a putative acetyl transferase gene related to the Tip60 and MOZ human genes and to the SAS genes of yeast, is required for dosage compensation in *Drosophila*. *EMBO J.* 1997; 16:2054–2060. [PubMed: 9155031]
50. Francis NJ, et al. Chromatin compaction by a Polycomb group protein complex. *Science.* 2004; 306:1574–1577. [PubMed: 15567868]
51. Tie F, et al. The *Drosophila* Polycomb group proteins ESC and E(Z) are present in a complex containing the histone-binding protein p55 and the histone deacetylase RPD3. *Development.* 2001; 128:275–286. [PubMed: 11124122]
52. Tyler JK, et al. The p55 subunit of *Drosophila* chromatin assembly factor 1 is homologous to a histone deacetylase-associated protein. *Mol Cell Biol.* 1996; 16:6149–6159. [PubMed: 8887645]
53. Martinez-Balbas MA, et al. *Drosophila* NURF-55, a WD repeat protein involved in histone metabolism. *Proc Natl Acad Sci U S A.* 1998; 95:132–137. [PubMed: 9419341]
54. Gdula DA, et al. Inorganic pyrophosphatase is a component of the *Drosophila* nucleosome remodeling factor complex. *Genes Dev.* 1998; 12:3206–3216. [PubMed: 9784495]
55. Beall EL, et al. Role for a *Drosophila* Myb-containing protein complex in site-specific DNA replication. *Nature.* 2002; 420:833–837. [PubMed: 12490953]
56. Taylor-Harding B, et al. p55, the *Drosophila* ortholog of RbAp46/RbAp48, is required for the repression of dE2F2/RBF-regulated genes. *Mol Cell Biol.* 2004; 24:9124–9136. [PubMed: 15456884]
57. Tsukiyama T, et al. Purification and properties of an ATP-dependent nucleosome remodeling factor. *Cell.* 1995; 83:1011–1020. [PubMed: 8521501]
58. Fasano L, et al. The gene teashirt is required for the development of *Drosophila* embryonic trunk segments and encodes a protein with widely spaced zinc finger motifs. *Cell.* 1991; 64:63–79. [PubMed: 1846092]
59. Takahashi F, et al. Regulation of cell-cell contacts in developing *Drosophila* eyes by Dsrc41, a new, close relative of vertebrate c-src. *Genes Dev.* 1996; 10:1645–1656. [PubMed: 8682295]
60. Saigo K, et al. Identification of a different-type homeobox gene, BarH1, possibly causing Bar (B) and Om(1D) mutations in *Drosophila*. *Proc Natl Acad Sci U S A.* 1991; 88:4343–4347. [PubMed: 1674606]
61. Bonfini L, et al. The Son of sevenless gene product: a putative activator of Ras. *Science.* 1992; 255:603–606. [PubMed: 1736363]

62. Klaembt C, et al. The *Drosophila* gene pointed encodes two ETS-like proteins which are involved in the development of the midline glial cells. *Development*. 1993; 117:163–176. [PubMed: 8223245]
63. Dekel I, et al. Identification of the *Drosophila melanogaster* homologue of the mammalian signal transducer protein, Vav. *FEBS Lett*. 2000; 472:99–104. [PubMed: 10781813]
64. Gross I, et al. Dorsal-B, a splice variant of the *Drosophila* factor Dorsal, is a novel Rel/NF-kappaB transcriptional activator. *Gene*. 1999; 228:233–242. [PubMed: 10072776]
65. Smith ER, et al. Regulation of the transcriptional activity of poised RNA polymerase II by the elongation factor ELL. *Proc Natl Acad Sci U S A*. 2008; 105:8575–8579. [PubMed: 18562276]
66. Goff DJ, et al. Establishment of dorsal-ventral polarity of the *Drosophila* egg requires capicua action in ovarian follicle cells. *Development*. 2001; 128:4553–4562. [PubMed: 11714680]
67. Cohen B, et al. Apterous, a gene required for imaginal disc development in *Drosophila* encodes a member of the LIM family of developmental regulatory proteins. *Genes Dev*. 1992; 6:715–729. [PubMed: 1349545]
68. Kokubo T, et al. *Drosophila* 230-kD TFIID subunit, a functional homolog of the human cell cycle gene product, negatively regulates DNA binding of the TATA box-binding subunit of TFIID. *Genes Dev*. 1993; 7:1033–1046. [PubMed: 8504928]
69. Xu Z, Hou T, Li N, Xu Y, Wang W. Proteome-wide detection of Abl1 SH3-binding peptides by integrating computational prediction and peptide microarray. *Mol Cell Proteomics*. 2012; 11(1):O111 010389. [PubMed: 22023807]
70. Hou T, Li N, Li Y, Wang W. Characterization of Domain-Peptide Interaction Interface: Prediction of SH3 Domain-Mediated Protein-Protein Interaction Network in Yeast by Generic Structure-Based Models. *J Proteome Res*. 2012; 11(5):2982–2995. [PubMed: 22468754]
71. Hou T, Li Y, Wang W. Prediction of peptides binding to the PKA RII α subunit using a hierarchical strategy. *Bioinformatics*. 2011; 27(13):1814–1821. [PubMed: 21586518]
72. Theoretical and Computation Biophysics Group. University of Illinois; Urbana-Champaign: <http://www.ks.uiuc.edu/Research/vmd/>
73. R: A language and environment for statistical computing. <http://www.r-project.org/>

Figure 1. RMSD for the equilibration and production runs of polycomb protein complexed with H3K27me3: (a) wild-type; (b) mutation Y26G. The RMSD for the H3 peptide is shown in green, for the entire complex in red, and for the polycomb protein in blue.

Figure 2.

The polycomb/H3K27me3 complex. (a, frontal view; b, back view) Only residues in polycomb critical to binding to H3 are shown. Hydrophobic residues are shown in brown; acidic residues, in orange; basic residues, in turquoise, and polar residues, in violet. The backbone of the remainder of polycomb is shown by a green ribbon, and H3 by a licorice model. H3R26 is shown in red; H3K27me3, in orange; H3S28, in yellow; H3A24 and H3A25, in magenta; and H3L20 through H3K23, in light blue. (c) Orthographic view of the polycomb/H3K27me3 complex. The polycomb protein is shown by molecular surface rendering, and all residues are green except for Tyr26, Trp47, and Trp50, which are lime green, sea green, and olive green, respectively. H3 is shown by licorice model; H3R26 is orange while H3K27me3 is yellow, and the rest of H3 is gray. Images generated by VMD.⁷²

Figure 3. Mutation matrix for the 21 residues in the polycomb protein deemed most important for binding to H3K27me3. Relative favorability of binding interaction is indicated by cell color: Greener cells indicate stronger binding relative to wild-type, while redder cells indicate weaker binding and yellow cells are near neutral. The striped yellow and black cells are equivalent to the wild-type protein. Residue rows are ordered by descending importance for H3K27me3 binding.

Figure 4.

Mutation matrix for the eight residues in H3 apart from K27 that occur in the crystal structure used for the simulation. Relative favorability of binding interaction is indicated by cell color: greener cells indicate stronger binding relative to wild-type, while redder cells indicate weaker binding and yellow cells are near neutral. The striped yellow and black cells are equivalent to the wild-type peptide.

Figure 5.

The microarray data of polycomb protein chromodomain binding to histone H3/H4 peptides were fitted to a mixed Gaussian model with two components using the mixtool package in R⁷³ with maximum likelihood estimation (black: overall; blue: nonbinding background distribution; red: binding distribution).

Figure 6.

ROC curve for polycomb-binding and non-polycomb-binding histone peptides. TP = true positive; TN = true negative; FP = false positive; FN = false negative. Sensitivity = $TP/(TP + FN)$; false positive rate = $FP/(FP + TN)$; AUC = area under curve.

Figure 7. Proteome scan flowchart, showing the filtering of candidate peptides to identify polycomb binders.

Table 1
Polycomb–H3 Complex Binding Free Energies

Complex	binding free energy (kcal/ mol) for $\epsilon = 1.0$	binding free energy (kcal/ mol) for $\epsilon = 2.0$	binding free energy (kcal/ mol) for $\epsilon = 4.0$	experimental dissociation constant (μM) ²
unmodified H3	-55.27 ± 3.35	-67.96 ± 2.79	-74.31 ± 3.09	>1000
H3K27me1	-59.93 ± 3.46	-71.97 ± 2.99	-78.00 ± 3.33	20 ± 3
H3K27me2	-62.71 ± 3.79	-72.56 ± 3.28	-77.48 ± 3.64	28 ± 4
H3K27me3	-69.78 ± 2.91	-79.84 ± 2.68	-84.87 ± 2.98	5 ± 1

Table 2
Polycomb–H3 Binding Free Energy Residue Decomposition^a

residue	contrib in polycomb/H3K27me3	residue	contrib in polycomb/H3K27me3
H3L20	-0.08 ± 0.54	Glu29	-1.22 ± 0.34
H3A21	-0.50 ± 0.20	Trp47	-5.08 ± 0.42
H3T22	-1.78 ± 0.42	Lys48	-0.13 ± 0.26
H3K23	-6.66 ± 0.56	Trp50	-2.36 ± 0.20
H3A24	-4.27 ± 0.33	Tyr54	-0.68 ± 0.12
H3A25	-5.41 ± 0.53	Thr56	-1.32 ± 0.12
H3R26	-8.41 ± 0.64	Glu58	-1.18 ± 0.72
H3K27me3	-13.69 ± 0.67	Pro59	-0.64 ± 0.06
H3S28	-2.09 ± 0.67	Asn62	-3.15 ± 0.55
Asp23	-1.82 ± 0.56	Ile63	-1.14 ± 0.13
Leu24	-2.57 ± 0.36	Leu64	-2.84 ± 0.24
Val25	-3.29 ± 0.26	Asp65	-2.59 ± 0.19
Tyr26	-4.82 ± 0.42	Arg67	-1.95 ± 0.53
Ala27	-1.69 ± 0.22	Leu68	-2.26 ± 0.25
Ala28	-1.89 ± 0.26	Glu73	-0.61 ± 0.02

^aThe residue Lys48 is included since, although its net contribution is only -0.13 kcal/mol, the van der Waals component to that contribution is relatively high at -0.71 kcal/mol.

Table 3
List of *Drosophila* Species Used in Proteome Scan

<i>Drosophila ananassae</i>	<i>Drosophila erecta</i>
<i>Drosophila grimshawi</i>	<i>Drosophila melanogaster</i>
<i>Drosophila mojavensis</i>	<i>Drosophila persimilis</i>
<i>Drosophila pseudoobscura</i>	<i>Drosophila sechellia</i>
<i>Drosophila simulans</i>	<i>Drosophila virilis</i>
<i>Drosophila willistoni</i>	<i>Drosophila yakuba</i>

Table 4
Mixed Gaussian Fitting Parameters of Microarray Data of Polycomb Protein Chromodomain Binding to H3 and H4 Peptides

polycomb-H3/H4	Gaussian distribution G1 (non-binding background)	Gaussian distribution G1 (binding)
proportion (%)	0.5722818	0.4277182
mean (μ)	0.5849581	1.7184761
standard deviation (σ)	0.2535892	1.1745391

Table 5
Identified Histone Binders for Polycomb Protein Chromodomain

sequence	histone (on human histone)	lysine site
GRIHRHLKS	H2AZ	37
GGAKRHRKV	H4	20
IQAVLLPKK	H2A1, H2AX, H2A3, H2A1D, H2A2C, H2A2A, H2A1B, H2A1C, H2A1J, H2A2B, H2A1A, H2A1H, H2AJ	117
RTKQTARKS	H3.3	9
RIHRLLRKG	H2A1A	35
EGTKAVTKY	H2B1A, H2B1B, H2B1C, H2B1D, H2B1H, H2B1J, H2B1K, H2B1M, H2B1L, H2B1N, H2B1O, H2B2E, H2B2F, H2B3B, H2BFS	117
LATKAARKS	H3.3	27
MDVVYALKR	H4	91
PRKQLATKA	H3.3	23
DSGKAKAKA	H2AV	15
ASKDLKVKR	H2AZ, H2AV	79
REIRRYQKS	H3.3	56