Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/4/11/eaau1447/DC1

Supplementary Materials for

Deciphering and engineering chromodomain-methyllysine peptide recognition

Ryan Hard, Nan Li, Wei He, Brian Ross, Gary C. H. Mo, Qin Peng, Richard S. L. Stein, Elizabeth Komives, Yingxiao Wang, Jin Zhang, Wei Wang*

*Corresponding author. Email: wei-wang@ucsd.edu

Published 7 November 2018, *Sci. Adv.* **4**, eaau1447 (2018)
DOI: 10.1126/sciadv.aau1447

The PDF file includes:

Other Supplementary Material for this manuscript includes the following:

(available at advances.sciencemag.org/cgi/content/full/4/11/eaau1447/DC1)

Movie S3 (.avi format). WT CBX1-PAmCherry in HeLa cells.
Movie S4 (.avi format). V22E/K25E/D59S CBX1-PAmCherry in HeLa cells.

Supplementary Information

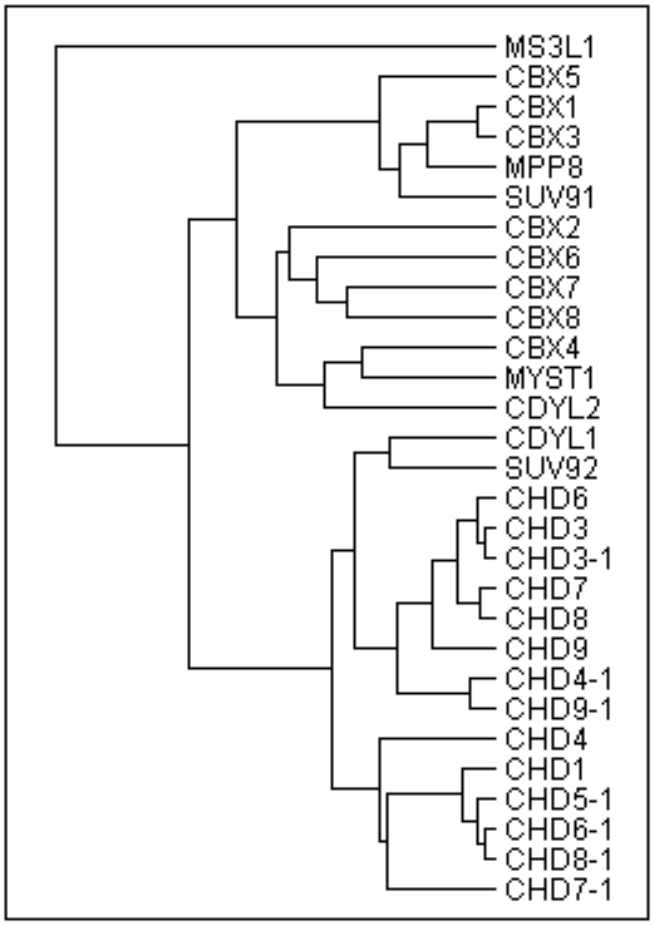

Fig. S1. Hierarchical clustering of chromodomains based upon *z* scores. Hierarchical clustering details of chromodomains based on the *z* scores of their microarray binding intensities towards methylated histone and non-histone peptides.

Fig. S2. Two-dimensional hierarchical clusters of peptides binding to chromodomains based on *z* scores (as separate PDF file). Details of two-dimensional hierarchical clusters of methylated peptides binding to chromodomains using *z* scores of the microarray binding intensities. Peptide sequences and annotations in the six clusters of methylated peptides are listed. Orange lines separate the six clusters.

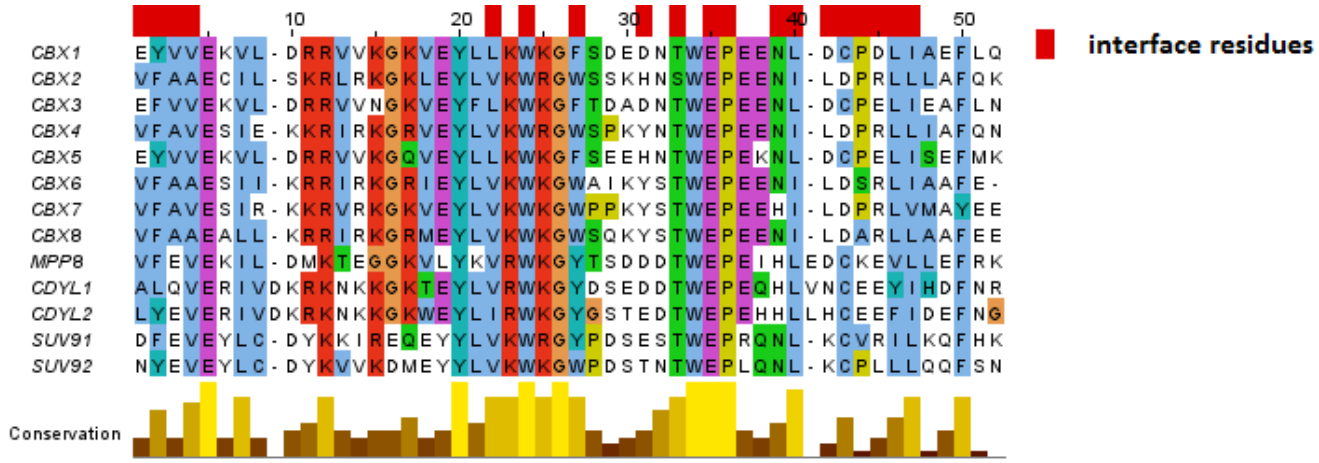

Fig. S3. Multiple sequence alignment of the chromodomains. Protein sequences of the 13 chromodomains that bind to peptides with single chromodomains were extracted from Uniprot/Swissprot (https://web.expasy.org/docs/swiss-prot_guideline.html) and ClustalW2.0 (https://www.ebi.ac.uk/Tools/msa/clustalw2/) was used to generate multiple sequence alignment of these 13 chromodomains, which were used in building the molecular interaction energy components-support vector machine (MIEC-SVM) model (*32*).

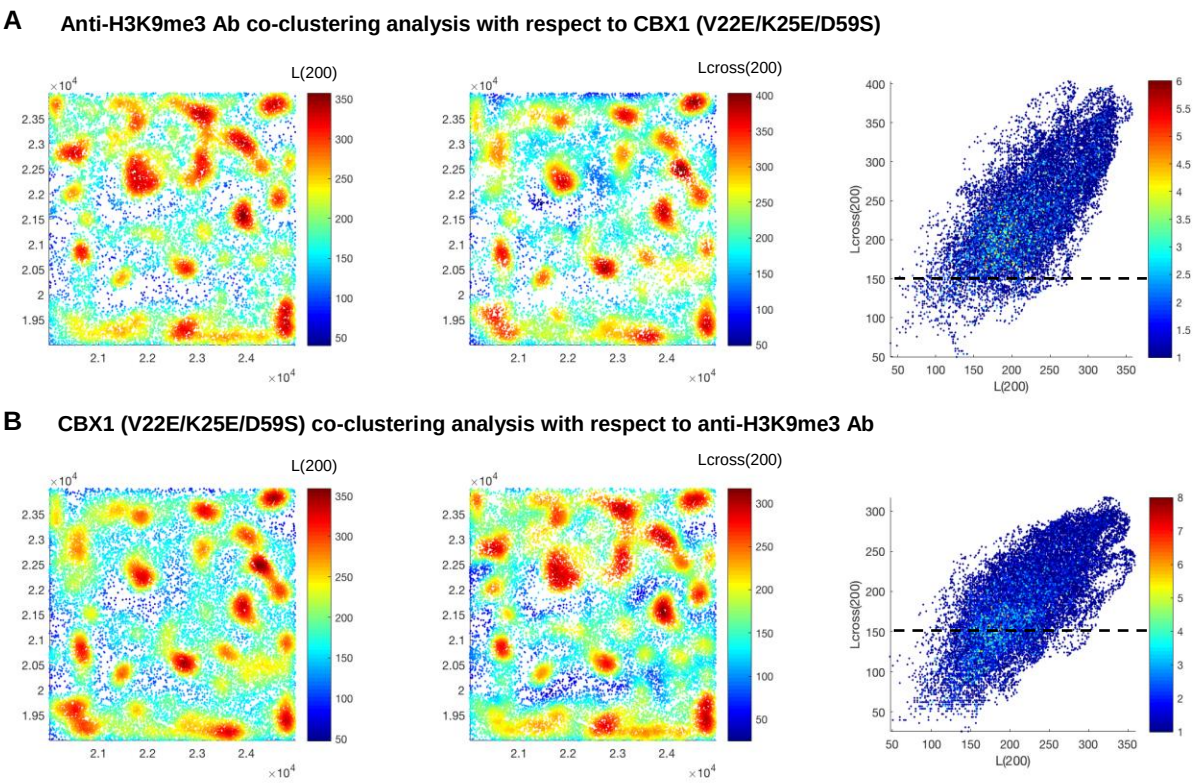

Fig. S4. Getis-Franklin single-molecule coclustering analysis (*35*) of H3K9me3 and the CBX1 (V22E/K25E/D59S) chromodomain. An anti-H3K9me3 antibody, labeled with Alexa647 2° antibody, and FLAG-CBX1 chromodomain (V22E/K25E/D59S), labeled with Alexa568 2° antibody, were co-imaged in fixed HeLa cells using STORM microscopy. (**A**). Representative scatter plot showing molecular localizations for the anti-H3K9me3 antibody, color-coded with the values for L(200) (left) and Lcross(200) (center), which reflect the number of localizations in its own species or of the other species, respectively, within a 200 nm radius of each localization. A scatter plot of Lcross(200) and L(200) is shown (right). Localizations with a Lcross(200) score above 150 were considered co-clustered. (**B**). Representative scatter plots showing molecular localizations of CBX1 chromodomain (V22E/K25E/D59S) in the same region, color-coded with L(200) (left) or Lcross(200) (center). A scatter plot of Lcross(200) vs. L(200) is shown (right).

Table S1. List of chromodomains screened by peptide microarray.

Protein	Chromodomain	Cloned Sequence
CBX1	21-79	20-73
CBX2	12-70	9-66
CBX3	30-88	29-81
CBX4	11-69	8-65
CBX5	20-78	18-75
CBX6	11-69	8-65
CBX7	11-69	8-62
CBX8	11-69	8-61
CHD1	272-364, 389-452	262-448, 379-448
CHD3	494-594, 631-673	529-579, 529-688
CHD4	494-594, 622-697	521-582, 521-679
CHD5	497-554, 592-653	489-543
CHD6	292-343, 375-439	282-359, 282-434
CHD7	800-867, 882-947	790-867, 790-940
CHD8	642-709, 724-790	632-709, 632-783
CHD9	690-761, 773-839	680-757, 680-832
MPP8	59-118	55-121
CDY1/CDY2	6-66	1-71
CDYL1	61-121	61-122
CDYL2	7-67	1-70
SUV91	43-101	40-106
SUV92	47-105	42-107
MYST1	69-123	67-124
MS3L1	32-90	29-94

Table S2. Four hundred sixty-seven peptide sequences printed on microarray (as separate Excel file).

Table S3. Sequence and averaged signal intensities of identified binders from the peptide microarray for all 29 chromodomains (as separate Excel file).

Table S4. Receptor-ligand residue pairs after LASSO feature selection (as separate Excel file)

Table S5. Nested cross-validation performed to evaluate overfitting in the training process (as separate Excel file).

Table S6. List of ranked sites for the CBX1 chromodomain that were considered for randomization (as separate Excel file). The lower the number, the higher the value was for that position.

Movies S1 to S4. Movies from raw PALM images of WT CBX1-PAmCherry and V22E/K25E/D59S CBX1-PAmCherry expressed transiently in either HeLa or MEF cells and imaged using near total internal reflection fluorescence (TIRF) microscopy (excitation with 561 nm laser, photoactivation with 405 nm laser). Video 1: WT CBX1 in MEF cells, Video 2: Mutant CBX1 in MEF cells, Video 3: WT CBX1 in HeLa cells, Video 4: Mutant CBX1 in HeLa cells.

Supplementary Methods and Materials

Non-histone peptides selected from the human proteome

In our previous work *(19, 20)*, we developed a bioinformatics pipeline integrating multiple filters to select non-histone peptides that are possibly methylated and bound by chromodomains (Figure 1A and see details in *(19, 20)*). Briefly, we first identified all the 9-amino-acid-long peptides that contain lysine at the 8th position in the human proteome. If the peptides are involved in protein interaction with chromodomains, they are likely conserved across species (Sequence Conservation). Because chromodomains recognize methyllysine, we searched 30 million mass spectra from human tissues for evidence of methylation at the 8th Lys in the candidate peptides (Mass Spectrometry). Peptides that passed the first two filters were subject to examination of their structural features characterizing peptides bound by the chromodomains: low propensity to form α helix (Secondary Structure) and accessible for binding (Solvent Accessibility). Because we were interested in identifying non-histone proteins that may contribute to regulating chromatin remodeling, we required that the proteins containing the peptide exist in the nucleus (Cellular Compartment). Furthermore, we used the CBX6 chromodomain as the template and applied additional criteria to narrow down the candidate peptides. Because interacting proteins are often co-expressed, we required proteins containing the candidate peptides share similar expression profiles with CBX6 across diverse cell types or conditions (Gene Co-Expression). To prioritize the peptides to be printed on the microarray, we ranked them based on the estimated binding affinity (Estimated Binding Score).

Peptide Microarray experiments

A total of 29 chromodomains were expressed as GST fusion proteins using pGEX-KG vector in *E. coli* strain BL21. Protein expression was induced at O.D.(600 nm) = 0.7 using 0.4 mM IPTG at 20°C overnight and purified on GST-Bind™ resin (Novagen) based on a previously described protocol *(20)*. 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 467 unmodified and modified 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% non-fat milk in TBST (room temperature 1 hour or 4°C overnight). The slide was then incubated at 4°C with chromodomain-GST fusion protein at a final concentration of 5 μM in 5% non-fat milk/TBST for 12 hours. After washing three times for 10 minutes each with TBST, the slide was incubated with an anti-GST mouse

monoclonal IgG antibody (Thermo) at a final concentration of 1 µg/mL in 5% non-fat milk/TBST and shaken gently for 1 hour at room temperature. A secondary anti-mouse IgG Dylight-488 conjugated antibody (Thermo) was added to a final concentration of 0.1 µg/mL after three more cycles of 10 minute washes with TBST. The slide was shaken for 1 hour at room temperature and washed three times with TBST.

Data Acquisition and Analysis

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, where the fluorescent intensity of a microarray spot was defined as the signal mean intensity minus the mean background intensity around it on the scanned image. For each peptide, the fluorescent intensity of the printed triplet was individually measured and all the intensities from one array were fit to a mixed Gaussian distribution with two components (non-binding background vs binder) *(19, 20)*. A statistical cutoff of $p < 0.05$ (from the background Gaussian distribution) was selected to quantitatively distinguish the binders from the non-binders.

Cluster Generation Process

The clusters were generated according to the dendrogram (see figure S2). First, we selected clusters that at least contain more than 30 peptides and got 12 clusters. Then these clusters were merged into 6 clusters with unique patterns. We made cluster 3 an independent cluster regardless of its size owing to its unique pattern. There are a few small clusters that contained 2, 2, 16 and 17 peptides respectively. We merged them into clusters 1 and 2 based on the pattern similarity. Also, small clusters that contains 11, 1, and 5 peptides were merged into clusters 5 and 6 respectively.

Fluorescence Polarization

Fluorescein-labeled methylated histone peptides (Karebay) were titrated with chromodomains in Tris buffer (25 mM Tris-HCl, pH 8, 125 mM NaCl) at room temperature for K_D determination. Final peptide concentration was 1 nM and incubation time was 30 minutes for each titration before reading. Data were acquired on a DTX 880 Multimode Detector Beckman Coulter plate reader with excitation filter at 485 nm and two emission filters at 535 nm equipped with polarizers. The dissociation constant (K_D) values were obtained by fitting data to a nonlinear regression equation with GraphPad Prism 4 software.

Sequence pattern analysis

For each chromodomain, the amino acid propensity of the 9 sites was calculated from the sequences of all single tri-methylated binders and illustrated by web logo (http://weblogo.berkeley.edu/logo.cgi). No obvious sequence pattern was detected other than the 8th position, which represents the tri-methylated lysine site.

Multiple sequence alignment of CBX chromodomains

Protein sequences of the 13 chromodomains that bind to peptides with single chromodomains were extracted from Uniprot/Swissprot (https://web.expasy.org/docs/swiss-prot_guideline.html) and ClustalW2.0 (https://www.ebi.ac.uk/Tools/msa/clustalw2/) was used to generate multiple

sequence alignment of these 13 CBX chromodomains with some local adjustment using structural information, which were used in building the MIEC-SVM model (*32*) (figure S3).

Building the MIEC-SVM model

Template complex structures of chromodomain-peptide interaction

The chromo-peptide complex structures were obtained from either PDB or structural modeling. The peptide in each chromo-peptide structure must be 9 residues (truncated if more than 9) with tri-methylated lysine on the 8^{th} position.

Among the 13 chromodomains, 7 domains, including CBX1 (1GUW), CBX2 (3H91), CBX3 (2L11), CBX5 (3FDT), CBX6 (3I90), CBX7 (2L12), and MPP8 (3QO2), have available chromo-peptide complex structures with at least 9 residues of the peptide in PDB. The complex structures were used as templates.

Another 4 domains, including CBX4 (2K28), CBX8 (3I91), CDYL1 (2DNT), and SUV91 (3MTS), have either chromodomain-only structures or chromo-peptide structure with peptides shorter than 9 residues. A structural alignment based modeling was used to construct respective chromo-peptide structures. First, the chromodomains in the chromo-only structure and each of the four crystal chromo-peptide templates (3H91, 3FDT, 3I90, and 3QO2) were aligned by the program LSQKAB (*36*) in the CCP4 software package (http://www.ccp4.ac.uk/). Then, based on the structural alignment, the peptide conformations from the four crystal templates were taken to the chromo-only structure to form four candidate complex templates. Finally, these candidate templates were optimized by molecular dynamics (MD). The one with the best RMSD and no steric clash (heavy atom distance < 3 Å) was selected as the template for further modeling.

The remaining two domains have no available structure (CDYL2 and SUV92). Their chromodomain structures were modeled by homology modeling using MODELLER (https://salilab.org/modeller/) and the chromo-peptide complex structure was constructed by the three-step procedure described above.

Force field parameters in structural modeling

The topology and coordinate files were prepared for the 13 chromo-peptide systems by tleap in AMBER11 *(21)*. AMBER ff03 force field *(22)* was used for all the standard amino acids and AMBER gaff force field *(23)* for modified amino acids. Electrostatic potential of modified residues was calculated by Gaussian09 *(37)* using Hartree-Fock HF/6-31G* basis set and their atomic charges were obtained using the RESP method *(38)* implemented in the program antechamber *(39)* in the AMBER package. TIP3P water boxes *(40)* were added around the protein molecule to 12 Å. The charge neutrality for each system was ensured through adding counter-ions Na^+ or Cl^-.

Conformational sampling

For each chromo-peptide template, molecular dynamics (MD) simulation was performed for conformational optimization and sampling. The system was relaxed by 10,000 steps of energy minimization with the first 3,000 steps of steepest descent followed by 7,000 steps of conjugate gradient minimization. After relaxation, the system was heated from 0 K to 300 K in 60 ps under NVT ensemble. Then, 5 ns of equilibrium and production run were performed under NPT

ensemble. SHAKE *(41)* was used to constrain all bonds involving hydrogen atoms. Langevin dynamics and isotropic position scaling were used for temperature and pressure control. Time step was set to 1 fs. Binding interface residue backbone RMSD was evaluated for the 13 chromodomains to verify the equilibrium. After the production run, 8 snapshots were evenly selected from the trajectory between 3 to 5 ns as chromo-peptide binding complex templates for each system. The chromo-peptide binding complex templates were mutated *in silico* to each of the 457 peptides by SCWRL4 (*42*). Restrained by the computational cost, we performed 5,000 steps of energy minimization instead of MD simulation to optimize all the complex structures obtained through mutation.

Calculation of MIECs

Residue pair-wise energy decomposition on minimized structures was performed by mm_pbsa.pl in the AMBER package (*21*). For each residue pair between the chromodomain and the peptide, the interaction energy was decomposed into four terms: van der Waals energy ΔE_{vdw}, electrostatic energy ΔE_{ele}, polar contribution to the desolvation energy ΔE_{gb}, and non-polar contribution to the desolvation energy ΔE_{sa}. Dielectric constant of 1 was used to calculate ΔE_{ele}. ΔE_{gb} was calculated using the generalized Born (GB) model with parameters developed by Onufriev et al. *(43)*. The interior and the exterior dielectric constants in the GB calculation were set to 1 and 80, respectively. ΔE_{sa} was estimated according to the solvent accessible surface area (SASA) as $\Delta E_{sa} = 0.0072 \times \text{SASA}$.

MIEC profile was generated based on the energy decomposition result. It consists of chromo-peptide MIEC profile and peptide internal MIEC profile. For chromo-peptide MIEC profile, all residue pairs less than 10 Å were included to reflect binding characteristics of the chromo-peptide interactions. For peptide internal MIEC profile, MIECs of the adjacent peptide residue pairs were calculated to represent the conformational preference of the peptide. MIEC profile for each chromo-peptide interaction contains 158 chromo-peptide residue pairs and 8 peptide-peptide pairs (664 energy components in total).

Feature Selection for the MIEC-SVM model

LASSO (Least Absolute Shrinkage and Selected Operator) logistic regression method was applied to the MIEC profile to select informative features to construct the MIEC-SVM model. Package "glmnet" (*44*) in R was used to train and test LASSO logistic regression models. The 222 MIEC components with non-zero coefficients (table S4) were kept as informative features for discriminating the chromo-peptide interactions.

A nested cross validation was performed to evaluate whether the over-fitting problem exists in the current training process. All data was randomly divided into two parts: the training set (90% of the data) and the test set (10% of the data). LASSO logistic regression-based feature selection was applied to the whole training set to select informative features. Then, 3-fold cross validation was performed on the training set with the informative features to select an optimal combination of SVM kernel and hyper-parameter (such as C, gamma, and kernel parameters). Lastly, the SVM model was trained on the training set using the selected features and SVM parameters. This model was used to make predictions on the test set. The comparison of the prediction performance between CV and the test set was used to evaluate the over-fitting issue on the training process. Such a design is used to avoid the use of any information from the test set in the

feature, kernel, and hyper-parameter selection processes. To avoid sampling bias, the random split of training and test set is repeated 10 times.

The nested CV result supports the lack of an over-fitting issue in the current training process since the prediction performance on the training set (CV result) and the test set is comparable (table S5). Moreover, the polynomial kernel and related kernel parameters (Poly1 in table S5) are selected from the 3-fold cross validation results because the parameter combination is selected as the best combination most frequently observed among the 10 repeats of the cross validation.

Training and testing of the MIEC-SVM model

All SVM training and tests were conducted using the LIBSVM package (*25*). The polynomial kernel function was used. Both 3-fold cross validation and leave-one-domain-out (LODO) test were performed to evaluate the prediction accuracy of the MIEC-SVM model. For 3-fold cross validation tests, the peptides were randomly divided into three groups. The SVM models were trained on any two groups and tested on the third group. Cross validations were repeated for 500 times to avoid overfitting and the average area under the curve (AUC) was used to evaluate the prediction performance. For leave-one-domain-out (LODO) test, one domain and all its associated interacting data were left out for testing while the model was trained on the remaining data. The LODO test was conducted for each of the 13 chromodomains and the average AUC was reported.

Background distribution for Jensen-Shannon Divergence

Jensen-Shannon divergence (JS divergence) is a symmetric metric to evaluate the distance between two distributions. Each JS divergence in Figure 3 (JS Divergence of SVM decision values) was calculated from the two foreground datasets of SVM decision values. To generate the background distribution for JS divergence, JS divergences were calculated between the larger foreground dataset and random background datasets which had the same number to the smaller foreground dataset and were randomly selected from all 5,941 SVM decision values. The random selection was repeated 1 million times. The mean and standard deviation were calculated from the random distribution. *P*-values were computed by assuming the background distribution to be a Gaussian distribution.

Selection of Candidate Sites to Randomize on the CBX1 Chromodomain

The rationale of the site selection was to find the sites that contribute most to binding and also particularly to recognition of H3K9me3. For this purpose, we should identify the residues where the binding energy profile of CBX1-H3K9me3 differs significantly from that of the CBX1-nonbinders. To achieve the goal, we used the MIEC to characterize the binding energy between CBX1 and the peptide.

First, we selected sites that are important for H3K9me3 binding. The total number of interacting residue pairs is 50*9=450 and each interacting residue pair is represented by two interaction types: polar (electrostatic + generalized Born) and non-polar (van der Waals + surface area). Each of the 900 MIEC components from the CBX1-H3K9me3 binding profile is compared to the respective components from the 389 CBX1-nonbinder MIECs. For each comparison, a two sided p-value is calculated to quantify the deviation of CBX1-H3K9me3's MIEC component to the

non-binder distribution. Then, two filters are applied to remove trivial components that do not have a major contribution to the binding energy: 1) contribution filter: for one component, the mean value between 68 binders and 389 nonbinders must be larger than 0.2 kcal/mol; 2) conservation filter: the site should not be conserved (identity≤9/13) in the multiple sequence alignment of 13 human chromo domains (figure S3). Third, for each CBX1 site, the mean logarithm p-value of all corresponding components is calculated. We then ranked all sites using the mean log-p value.

Second, we selected sites that are generally important for binding. For this purpose, we compared all CBX1 binders to a non-binder. Because H3K27me3 peptide is similar to H3K9me3 (4 amino acids in common) and the WT CBX1 chromodomain does not bind to H3K27me3, we used H3K27me3 to represent non-binders. We used a similar procedure as described above except that each component of CBX1-H3K27me3 was compared to that of 68 CBX1-binder MIECs.

Then, we combined the CBX1-H3K9me3 and CBX1-generally important residue lists into one list, which resulted in selection of sites 59, 60, 62, 25, and 22 (see table S6). Sites 21 and 56, which also ranked high in the prediction, were not chosen because they were part of the aromatic cage critical for methyl-lysine recognition (site 21) or because they pointed to the conserved C-terminus in the truncated peptide structure (site 56, which points to serine for both H3K9 and H3K27).

Two-Color Stochastic Optical Reconstruction Microscopy (STORM) Imaging Co-Localization Analysis

For two-color STORM imaging, acquisitions for the two channels were interleaved. Image analysis and reconstruction were performed using the N-STORM software package. Before the co-clustering analysis was performed, over-counting of localizations due to persistence of the fluorescence of an emitter for multiple frames and blinking was corrected by grouping localizations within 100nm separated by less than a particular dark time (t_d) and treating them as coming from a single emitter. The appropriate dark time for each fluorophore was determined using a published method *(45)*.The number of localizations after such grouping was determined for varying values of t_d, and the data were fit to the following equation

$$N(t_d) = N(1 + n_{blink,1}e^{\frac{(1-t_d)}{t_{off,1}}} + n_{blink,2}e^{\frac{(1-t_d)}{t_{off,2}}})$$

Here, $N(t_d)$ is the number of localizations after grouping with a particular dark time, N is the real number of fluorescent molecules, $n_{blink,1}$ and $n_{blink,2}$ are the number of dark state conversions occurring with average dark state lifetimes $t_{off,1}$ and $t_{off,2}$, respectively. For Alexa568, the fit showed that 0.8 blinks per molecule occurred for dark states with a lifetime of 11.4 frames, and 1.4 blinks per molecule occurred for dark states with a lifetime of 90 frames. For Alexa647, the fit showed that 0.3 blinks per molecule occurred for dark states with a lifetime of 9.2 frames, and 0.7 blinks per molecule occurred for dark states with a timescale of 171 frames. Values of the

dark time were chosen to be approximately double the lifetime of the longer blinking timescale, or 180 frames for Alexa568 and 340 frames for Alexa647.

Co-clustering analysis was performed using a published method *(35)*, based on Getis and Franklin's local point pattern analysis using MATLAB software. For each localization in one channel, the number of localizations in the other channel within 200nm was counted, with that statistic being called L(200)cross. A threshold of L(200)cross of 150 was set for a particular localization to be considered co-localized with a cluster of the other channel. We then determined the percentages of localizations in each channel that were localized within clusters of the other channel for each cell.