Biophysical Journal
Article

Biophysical Society

Learning the Formation Mechanism of Domain-Level Chromatin States with Epigenomics Data

Wen Jun Xie[1] and Bin Zhang[1,*]
[1]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts

ABSTRACT Epigenetic modifications can extend over long genomic regions to form domain-level chromatin states that play critical roles in gene regulation. The molecular mechanism for the establishment and maintenance of these states is not fully understood and remains challenging to study with existing experimental techniques. Here, we took a data-driven approach and parameterized an information-theoretic model to infer the formation mechanism of domain-level chromatin states from genome-wide epigenetic modification profiles. This model reproduces statistical correlations among histone modifications and identifies well-known states. Importantly, it predicts drastically different mechanisms and kinetic pathways for the formation of euchromatin and heterochromatin. In particular, long, strong enhancer and promoter states grow gradually from short but stable regulatory elements via a multistep process. On the other hand, the formation of heterochromatin states is highly cooperative, and no intermediate states are found along the transition path. This cooperativity can arise from a chromatin looping-mediated spreading of histone methylation mark and supports collapsed, globular three-dimensional conformations rather than regular fibril structures for heterochromatin. We further validated these predictions using changes of epigenetic profiles along cell differentiation. Our study demonstrates that information-theoretic models can go beyond statistical analysis to derive insightful kinetic information that is otherwise difficult to access.

INTRODUCTION

A remarkable achievement of multicellular organisms is the formation of distinct cell types with the same DNA sequence. The epigenome that consists of covalent modifications to histone proteins and the DNA is expected to play a key role in encoding the cell type diversity (1,2). A prominent feature of the epigenome is that neighboring nucleosomes tend to share similar modifications such that the entire chromatin is partitioned into various continuous regions with well-defined histone modification patterns (3) (i.e., domain-level chromatin states) that include transcriptionally active domains and heterochromatin. Switching the type of these states as cells differentiate can impact the expression of the underlying genes and drive phenotypic changes without altering the DNA sequence.

For epigenetic modifications and domain-level chromatin states to maintain well-defined gene expression profiles and cell phenotypes, there must exist robust mechanisms for their establishment and maintenance (4). Ensuring the sustained stability of these chemical modifications within a cell cycle and through cell division can be a challenging task. Unlike the DNA sequence, histone modifications are intrinsically dynamic and face constant perturbations from enzymes that add or remove these marks (5). Furthermore, roughly half of the histone proteins will be replaced with unmodified ones as the DNA replicates, and these random additions can dramatically alter the epigenome of the daughter cells if left uncorrected (6–8). Therefore, it is not surprising that cells have evolved complex regulatory pathways to ensure the stability of epigenetic modifications. Though significant insights on the robustness of histone modifications can be gained by studying these pathways (9,10), a complete understanding based on them is lacking because many of the molecular players that are crucial for epigenome stability and inheritance remain unknown.

Large-scale sequencing studies have provided valuable data on the genome-wide distribution of epigenetic marks (11). Top-down theoretical models based on these data provide an alternative approach for the mechanistic understanding of domain-level chromatin state formation. They can circumvent the challenges faced by modeling chromatin regulatory pathways, including the incompleteness of these pathways and the lack of kinetic parameters. Existing epigenomics data analysis approaches (12–14) have focused on extracting unique patterns of histone modifications. They have discovered many novel single-nucleosome chromatin states and support the histone code hypothesis (15,16), which states that unique combinations of histone modifications encode

Submitted January 23, 2019, and accepted for publication April 4, 2019.
*Correspondence: binz@mit.edu
Editor: Tamar Schlick.
https://doi.org/10.1016/j.bpj.2019.04.006

distinctive biological outcomes. However, focusing on the correlation between epigenetic marks within the same nucleosome, current approaches have often ignored the extension of these marks across multiple nucleosomes to form domains (17). Consequently, they fail to provide insight into the stability of large domain-level chromatin states and the mechanism leading to their establishment.

We propose an information-theoretic model to identify domain-level chromatin states with distinct histone modification patterns and infer the mechanism for their formation from epigenomics data. This model can be derived and parameterized following the maximal entropy principle. It rigorously accounts for both intra- and internucleosome correlation between epigenetic marks and succeeds in identifying various known states, including super (stretch) enhancers (18,19), broad H3K4me3 domains (20), and heterochromatin (9). Transition path analysis of these domain-level chromatin states revealed that promoters and enhancers could be stabilized at a continuous range of genomic lengths, and more extended states often grow from shorter ones through a sequential maturation process. On the other hand, heterochromatin states exhibit a bistable behavior, and nucleosomes within these states undergo an all-or-none cooperative transition as they become methylated. By correlating the model interaction energies with contact probability between nucleosomes in the three-dimensional (3D) space, we found that our model supports condensed, globular heterochromatin conformations that resemble phase-separated liquid droplets. Our results demonstrate the usefulness of statistical mechanical models and molecular biophysical approaches in interpreting the rich information encoded in epigenomics data.

MATERIALS AND METHODS

Calculating epigenetic mark correlation

For a comprehensive characterization of the chromatin landscape, we analyzed the collective behavior of 12 epigenetic marks. These marks include histone H3 lysine 4 monomethylation/dimethylation/trimethylation (H3K4me1/me2/me3), which are important for identifying enhancer and promoter regions; H3 lysine 9 acetylation (H3K9ac) and H3 lysine 27 acetylation (H3K27ac), which are associated with increased activation of gene promoters and enhancers, respectively; H3 lysine 9 trimethylation (H3K9me3) and H3 lysine 27 trimethylation (H3K27me3), which are signatures of constitutive heterochromatin and facultative heterochromatin; H3 lysine 36 trimethylation (H3K36me3), H3 lysine 79 dimethylation (H3K79me2), and H4 lysine 20 monomethylation (H4K20me1), which are indicative of transcribed gene regions; DNase I hypersensitive sites that probe exposed DNA and open chromatin regions; and an important histone variant H2A.Z.

We obtained genome-wide profiles of these marks from the ROADMAP epigenomics project (11). Data from the IMR90 cell line were collected because of their high quality. We then binarized the epigenetic profiles with a Poisson background model at a resolution of 200 bp that corresponds to the nucleosome repeat length (21). From the binarized data, the mean occupancy for epigenetic mark i $\left(\sum_{n=1}^{N} s_i^n/N\right)$, the intranucleosome mark correlations $\left(\sum_{n=1}^{N} s_i^n s_j^n/N\right)$, and the correlation coefficient between marks i and j that are separated by l nucleosomes $\left(\sum_{n=1}^{N-l} s_i^n s_j^{n+l}/(N-l)\right)$ were calculated. In these notations, the binary variable s_i^n indicates whether the i-th mark is present ($=1$) at the n-th nucleosome or not ($=0$), and N is the genome length. We note that the experimental data used here is obtained from a bulk average, but epigenetic marks are under constant remodeling in individual cells. The correlation coefficients between histone marks determined here, therefore, only represent a mean-field approximation to the actual values. Single cell epigenomics data are becoming available (22) and will provide more accurate estimations for the correlation between histone marks that are present on the same DNA molecule at the same time.

Parameterizing the information-theoretic model

Using the mean occupancy and mark correlations as experimental constraints, parameters in the energy function $E(s, L)$ defined in Eq. 8 can be determined with a Boltzmann learning algorithm (23). This algorithm minimizes the cross entropy

$$S^*(\boldsymbol{\theta}) = \sum_p F_p^{\exp}\theta_p - \ln Z(\boldsymbol{\theta}), \tag{1}$$

where $\boldsymbol{\theta} = \{h_i, J_{ij}, K_i^t\}$ is the collection of model parameters, Z is the partition function of the model, and $F^{\exp}$ is the collection of experimental constraints. It is straightforward to demonstrate that at the stationary point of $S^*(\theta)$, the simulated ensemble averages match with experimental constraints.

The minimum of $S^*(\theta)$ can be found with the steepest gradient descent method that iteratively updates the parameters with the following expression

$$\theta_p^{t+1} = \theta_p^t - \alpha\left(\frac{\partial S^*}{\partial \theta_p}\right)_{\theta^t}, \tag{2}$$

where the gradient of S^* is defined as:

$$\frac{\partial S^*}{\partial \theta_p} = F_p^{\exp} - \frac{\partial \ln Z(\theta)}{\partial \theta_p} = F_p^{\exp} - F_p^{\text{model}}. \tag{3}$$

F^{model} is the collection of simulated ensemble averages for mean mark occupancy and correlation coefficients between marks.

To accelerate the steepest gradient descent in the relevant direction and dampen oscillations, a momentum step was included when updating the parameters:

$$\begin{cases} \theta_p^{t+1} = \phi_p^t - \alpha\left(\dfrac{\partial S^*}{\partial \theta_p}\right)_{\theta^t} \\ \phi_p^{t+1} = \theta_p^{t+1} + \beta\left(\theta_p^{t+1} - \theta_p^t\right) \end{cases}. \tag{4}$$

We used a learning rate of $\alpha = 0.002$ and $\beta = 0.9$. At each iteration t, we used the Metropolis Monte Carlo algorithm to sample the Boltzmann distribution and estimate the ensemble averages F_p^{model}. We further applied the parallel tempering technique to enhance the sampling efficiency. A total of 13 replicas was used with temperatures evenly distributed between 0.8 and 2.0, with each replica lasting for 10^6 Monte Carlo steps per iteration. The parameters were set as zero at the beginning and were updated until the relative error defined as $\varepsilon = \sum_i |f_i^{\text{model}} - f_i^{\exp}| / \sum_i f_i^{\exp}$ is less than 5%.

Identifying domain-level chromatin states using the information-theoretic model

To identify robust combinatorial patterns of epigenetic marks and domain-level chromatin states, we searched for basins of attractions supported

by the information-theoretic model using steepest gradient descent optimization. Like other Hopfield-like systems (14,24), these basins are local minima that represent patterns with high probabilities of appearance. Local minima are defined as states whose energies are lower than all the neighboring configurations. The neighborhood of a state includes all configurations that differ from the state with only one epigenetic mark. Following this definition, the closest local minimum for any configuration can be found by the steepest gradient descent algorithm. In this algorithm, we iteratively select out the configuration that has the lowest energy in the neighborhood of the lowest energy state from the previous iteration until convergence.

To identify the set of local minima that most resemble the configurations observed in genome-wide profiles of epigenetic marks, we first partitioned the genome into unmodified and modified regions. The modified regions were identified as stretches of genomic segments that do not contain any gaps longer than 25 nucleosomes. Gap regions do not exhibit any epigenetic marks. The local minimum for unmodified regions will always be the ground state. For modified regions, we further divided them into configurations with 25 nucleosomes that correspond to the system size of our model. For each 25-nucleosome long region, we performed the steepest gradient descent optimization to find the corresponding local minimum (i.e., domain-level chromatin states).

Calculating transition rates between domain-level chromatin states

We used Monte Carlo simulations to probe connections between domain-level chromatin states and to calculate transition rates. Specifically, for a pair of states A and B, we first selected a large set of paths that originated from A and ended up in B from a trajectory that has sufficiently traversed the entire energy landscape. From the path ensemble ρ_{AB}, the configuration of the transition state connecting A and B is identified with the energy

$$E_{AB} = \min_{p\in\mathbb{P}_{AB}} \max_{s\in p} E(\boldsymbol{s}), \tag{5}$$

where $E(\boldsymbol{s})$ is the energy of configuration $\boldsymbol{s}$ along the transition path p, and its detailed expression is provided in Eq. 8. The energy barrier is then determined as $\Delta E_{AB} = E_{AB} - E_A$ for the transition from A to B and $\Delta E_{BA} = E_{AB} - E_B$ for the reverse transition. From these barriers, we estimate the transition rates using the Arrhenius equation $r_{AB} = \frac{1}{\tau} e^{-\Delta E_{AB}/k_B T}$, where τ is the timescale for observing a fluctuation in histone modification. Note that it's nontrivial to efficiently sample the huge phase space. We therefore employed the generalized Wang-Landau algorithm (25) to bias the simulations for enhanced transitions between states (see Supporting Materials and Methods for more details).

Calculating the most probable transition path between domain-level chromatin states

To determine the most probable pathway connecting two domain-level chromatin states A and B, we applied the transition path theory to the kinetic network. Transition path theory is a probabilistic framework to analyze the mechanism of reaction or transition and provides the reactive rate of state transformation (26). The core concept in transition path theory is the committor probability. The committor probability (q_i^+) at state i is defined as the probability that the system reaches state B first before visiting state A. Thus, the committor probability can be solved by the simple relation with the transition probability matrix

$$T_{iB} = q_i^+ - \sum_{k\neq A} T_{ik} q_k^+. \tag{6}$$

The transition probability matrix T was constructed from the rate matrix r using the expression $T = e^{r\Delta t}$. A lag time of $\Delta t = 1$ was used.

Among all the transitions between the two domain-level chromatin states, only a fraction is reactive. The effective flux of transition between intermediate states $i \rightarrow j$ contributing to the path $A \rightarrow B$ is

$$f_{ij} = \pi_i \left(1 - q_i^+\right) T_{ij} q_j^+, \tag{7}$$

where π_i is the equilibrium distribution of domain-level chromatin state i that can be determined from the relation $\pi T = \pi$. Then, we picked up the most probable transition pathway connecting A and B as the pathway with the greatest flux (f_{ij}) along it.

RESULTS

Information-theoretic model predicts long-range internucleosome correlations

Chromatin immunoprecipitation followed by deep sequencing is a powerful method for characterizing the epigenome at high resolution and has helped to determine genome-wide profiles of numerous epigenetic marks across hundreds of cell types (11,27). Here, we introduce an information-theoretic approach to analyze these data and investigate the interrelationship of epigenetic marks.

The information-theoretic model describes a chain of N-interacting nucleosomes, each one of which is characterized by a total of 12 epigenetic marks (Fig. 1). These marks are selected for a comprehensive characterization of the chromatin landscape, and their biological importance are explained in the Materials and Methods: Calculating epigenetic mark correlation. The model's potential energy adopts the following form

$$E(\mathbf{s}, L) = \sum_{n,i} \left[h_i s_i^n + \sum_{j>i} J_{ij} s_i^n s_j^n + \sum_{l=1}^{L} K_i^l s_i^n s_i^{n+l} \right]. \tag{8}$$

The binary variable s_i^n indicates whether the i-th mark is present ($= 1$) at the n-th nucleosome or not ($= 0$). The parameters h_i, J_{ij}, and K_i^l measure the overall propensity for the appearance of the i-th mark, the coupling strength between marks i and j on the same nucleosome, and the coupling strength between the same mark i separated by l nucleosomes, respectively. We include internucleosome interactions up to L nucleosomes, and the total number of parameters equals $78 + 12 \times L$. As shown in the Supporting Materials and Methods, $E(s, L)$ is the most probable model that maximizes the information entropy while reproducing the experimental mean and pair-wise correlation of epigenetic marks. Maximal entropy models have been successfully applied to study a wide variety of problems, including protein structure prediction and genome folding (28–35). A related model with only intranucleosome correlations has been proposed to study single-nucleosome chromatin states (14,36). The added internucleosome interactions here are crucial for studying the spreading of epigenetic marks and the formation of long-range domain-level states.

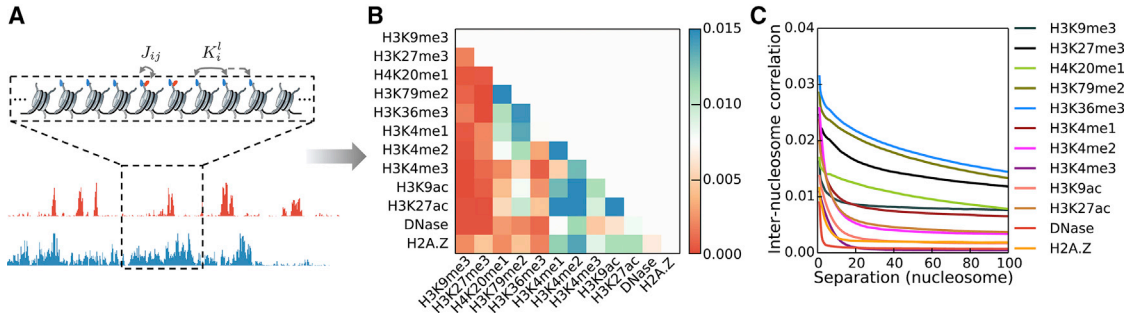

FIGURE 1 Epigenetic marks exhibit strong intra- and internucleosome correlations. (*A*) Shown is an illustration of the information-theoretic model for studying domain-level chromatin state formation (*top panel*). The model explicitly considers intra- (J_{ij}) and internucleosome (K_i^l) coupling between epigenetic marks, and the values of the coupling strengths are derived from epigenomics data (*bottom panel*). The fundamental unit of this model is a 200-bp-long genomic segment that includes both the core nucleosome and the linker DNA. A total of 12 epigenetic marks is used to describe the state of each segment. For simplicity, only two marks are shown. (*B*) Pair-wise correlation between epigenetic marks on the same nucleosome is shown. (*C*) Self-correlation for epigenetic marks on different nucleosomes as a function of the genomic separation is shown. The characteristic length scales of these correlation coefficients measure the tendency for epigenetic marks to spread across multiple nucleosomes. To see this figure in color, go online.

As detailed in the Materials and Methods, parameters in $E(s, L)$ can be derived with a Boltzmann learning algorithm (Parameterizing the information-theoretic model) using experimental data collected from IMR90 cells (Calculating epigenetic mark correlation). For efficient computational sampling, we limited the system size N to 25 nucleosomes and adopted the periodic boundary condition. As shown in Table S1, the correlation length for most epigenetic marks is significantly less than 25. In the meantime, the periodic boundary condition ensures that the long-range correlation between some epigenetic marks that give rise to large, periodic domain-level chromatin states can be modeled accurately in a finite system.

To probe the effect of internucleosome interactions and identify the minimalist model with the least parameters that succeeds in capturing the correlation between epigenetic marks, we studied a series of systems with increasing L. Fig. 2, *A* and *B* present the results obtained from a model without any internucleosome interactions ($L = 0$). As indicated by the blue dots, this model succeeds in reproducing intranucleosome correlations that were provided as experimental constraints. The relative error, which is defined as the total absolute difference between experimental constraints and modeled values normalized by the sum of experimental constraints, is less than 5%. Furthermore, higher-order intranucleosome correlations, which are measured by the population of all the possible combinatorial patterns formed by the 12 epigenetic marks and are not included as experimental constraints, are accurately predicted as well (Fig. 2 *A*, *green*). On the other hand, the simulated internucleosome correlations between all pairs of epigenetic marks separated by less than 13 nucleosomes differ significantly from the experimental values (Fig. 2 *B*).

The performance of models that explicitly consider internucleosome coupling are shown in Fig. 2, *C* and *D*. We measured the model performance using the quality of the linear regression between simulated and experimental internucleosome correlations and the relative error between the two. As shown in Fig. 2 *C*, the slope and R-squared of the linear fit improve systematically, and the difference between simulation and experiment continuously decreases up to $L = 5$. Explicit results for the $L = 5$ model are provided in Fig. 2 *D*. Further increasing L (see Fig. S1) or introducing internucleosome cross-mark coupling to the energy function $E(s, L)$ (see Fig. S2) does not improve the model's performance. We therefore conducted all following analyses using the model with $L = 5$.

Information-theoretic energy landscape supports domain-level chromatin state formation

A central goal for the computational analysis of epigenomics data is to systematically characterize combinatorial patterns of histone modifications that are of functional importance and appear more frequently than random. Previous studies have focused on histone modifications within the same nucleosome to identify single-nucleosome chromatin states (13,36). Here, we investigate whether distinct patterns also emerge across multiple nucleosomes to form domain-level chromatin states. Toward that end, we searched for basins of attractions supported by the energy landscape of the parameterized information-theoretic model. These energy basins offer representative arrangements of epigenetic marks that are of high population and are ideal candidates for domain-level chromatin states. In the following, we will focus our discussion on the top 100 most populated states identified with the steepest descent algorithm (see Materials and Methods) because they cover over 94% of the whole genome.

Fig. 3 *A* illustrates the average epigenetic mark profiles for the top 100 domain-level chromatin states. The most populated state, which we term as the ground state, exhibits no histone modifications. Constitutive and facultative heterochromatin states are, respectively, identified with

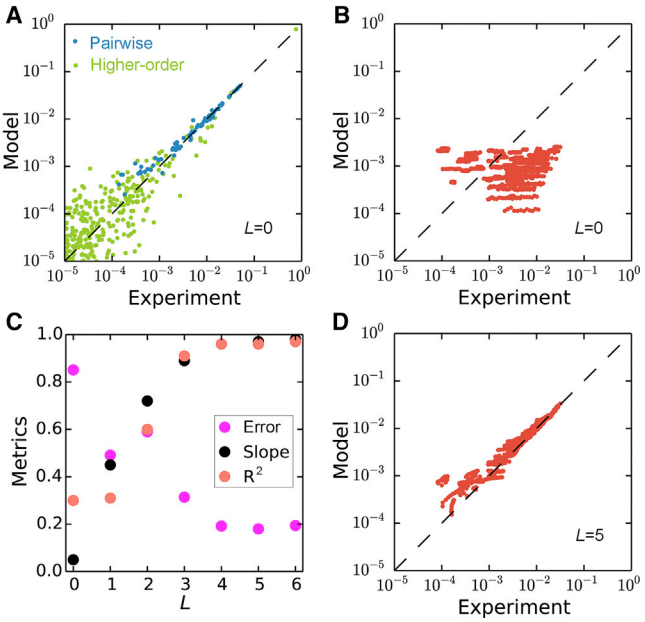

FIGURE 2 Parameterization and validation of information-theoretic models with different internucleosome interaction cutoff length L. (*A*) Shown is a comparison between experimental and simulated ($L = 0$) pair-wise (*blue*) and higher-order intranucleosome correlations (*green*). (*B*) Shown is a comparison between experimental and simulated ($L = 0$) internucleosome correlations for all pairs of epigenetic marks separated by less than 13 nucleosomes. (*C*) Quantitative measurements of the performance for models with different L are shown. (*D*) Shown is a comparison between experimental and simulated ($L = 5$) internucleosome correlations. The same experimental data as in part (*B*) are used for plots in (*C* and *D*). To see this figure in color, go online.

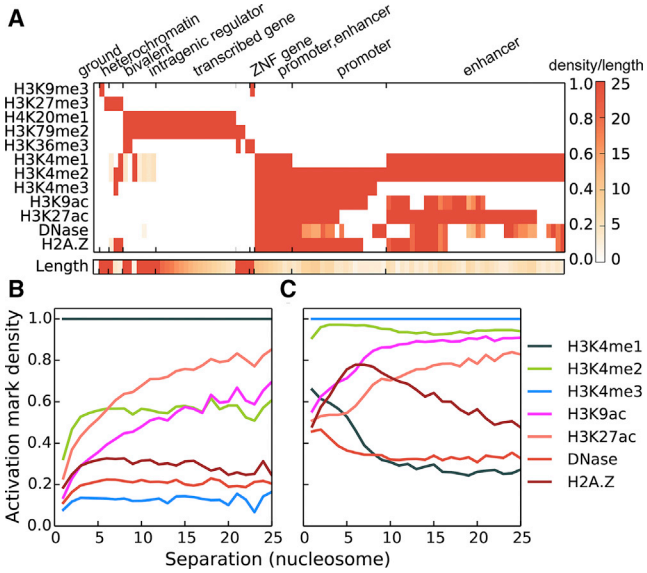

FIGURE 3 Characterization of domain-level chromatin states predicted by the information-theoretic model $E(s, L = 5)$. (*A*) Average epigenetic profiles over modified nucleosomes for the top 100 most populated states are shown. The length shown in the bottom panel represents the number of modified nucleosomes for each state, and the color bar is shown on the side. The labels for different chromatin domains shown at the top are based on known functions of their dominant epigenetic marks (see text for details). (*B* and *C*) Shown are the mean values of various epigenetic marks determined from chromatin immunoprecipitation followed by deep sequencing data for chromatin segments with different lengths of continuous H3K4me1 marks (*B*) and H3K4m3 marks (*C*). To see this figure in color, go online.

signature methylation marks H3K9me3 and H3K27me3. States with H3K4me1 or H3K4me3 are assigned as enhancer and promoter, respectively, though they often share additional activation marks that include H3K4me2, H3K9ac, H3K27ac, DNase, and H2A.Z. A distinct set of states (promoter/enhancer) with both H3K4me3 and H3K4me1 marks are found as well, supporting the fundamental similarity between promoters and enhancers (37). States marked with H3K36me3, H3K79me2, and H4K20me1 modifications are transcribed gene regions. We further label states that consist of both gene silencing marks (H3K27me3) and activation marks H3K4me1/2/3 as bivalent, states marked with transcribed gene marks and regulatory marks as intragenic regulator, and states marked with H3K9me3 and H3K36me3 as zinc finger protein gene (11).

In addition to their unique histone modification profiles, domain-level chromatin states exhibit distinct length dependence as well, as shown in the bottom plot of Fig. 3 *A*. The length of a state is defined as the number of nucleosomes with at least one epigenetic mark. For several heterochromatin and transcribed gene states, all the 25 nucleosomes share the same set of epigenetic modifications. These periodic patterns are consistent with the formation of large silent or transcribed genomic regions that span tens of kilobase pairs. Their periodicity, however, also suggests that these domain-level chromatin states have the potential to spread over the entire genome, and additional mechanisms beyond epigenetic mark interactions, such as boundary elements, must be in play to confine them at specific genomic regions (38).

On the other hand, we find a series of aperiodic promoter and enhancer states of varying lengths. We note that extended regulatory states have indeed been observed previously (18–20) and are often termed as broad H3K4me3 domains and super (stretch) enhancers. The increased appearance of additional activation marks in more extended states (see Fig. 3, *B* and *C*) may support open chromatin conformations to enhance the transcription consistency or the overall expression level of their target genes (39).

As domain-level chromatin states are local energy minima, their histone modification patterns and length dependence can be understood from the underlying energy landscape. In Fig. 4, we plot the intra- and internucleosome interaction energies between epigenetic marks. Consistent with the complex promoter and enhancer state patterns, most of the interaction energies between activation marks are negative (blue), supporting their co-existence within the same nucleosome. A notable exception is a strong

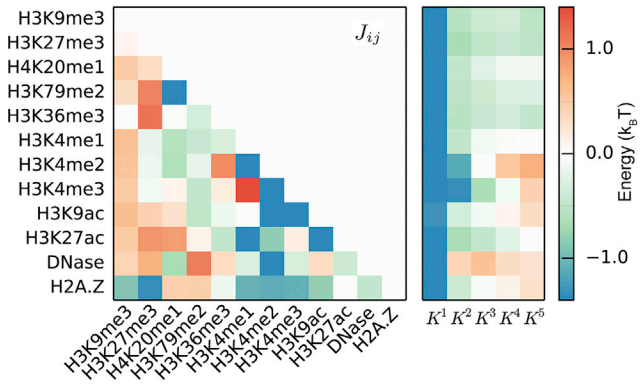

FIGURE 4 Intra- (*Left*) and inter- (*Right*) nucleosome coupling energies between epigenetic marks for the information-theoretic model $E(s, L = 5)$. To see this figure in color, go online.

repulsion between H3K4me1 and H3K4me3, two mutually exclusive marks for defining enhancers and promoters, respectively (11). We emphasize that the intranucleosome energies probe direct interactions among epigenetic marks that correspond better with physical interactions between histone modification enzymes. The correlation coefficients presented in Fig. 1, on the other hand, could arise from a transit effect as a result of indirect couplings (36,40).

As shown in the right panel of Fig. 4, all epigenetic modifications exhibit attractive interactions between the nearest neighbor nucleosomes (K^1). This attractive interaction decays rather quickly for active marks, explaining the small length of promoter and enhancer states. On the other hand, the interaction between epigenetic marks for transcribed gene regions or heterochromatin persists over an extended range to promote the formation of large domains.

Mechanism of domain-level chromatin state formation from transition path calculations

As basins of attraction, domain-level chromatin states are inherently stable and able to withstand transient fluctuations caused by the addition and removal of enzymes within a cell cycle. Thus, once established, these states provide a robust mechanism for regulating gene expression and maintaining genome stability. However, the molecular mechanism for their de novo creation as cells differentiate and their reestablishment when cells divide remains elusive. We here perform kinetic analysis of the information-theoretic model to provide mechanistic insight into domain-level chromatin state formation.

Toward that end, we first built a kinetic network to explore the dynamical transition between domain-level chromatin states (see Fig. 5 *A*). Each node in this network corresponds to one of the top 100 most populated states. A connection between two nodes is introduced if direct transitions between them were observed in a long-time enhanced simulation conducted with the generalized

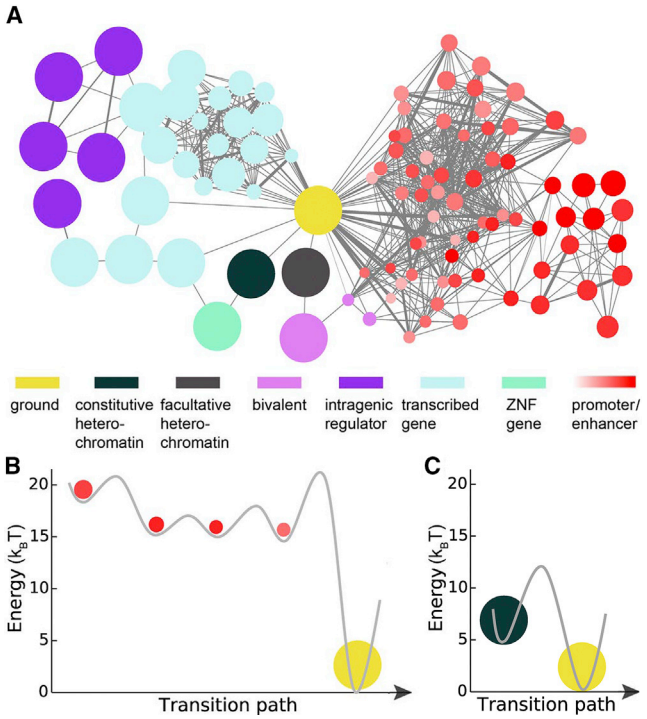

FIGURE 5 Kinetic analysis of the information-theoretic model $E(s, L = 5)$. (*A*) Shown is an illustration of the kinetic network for transitions between domain-level chromatin states. Each node represents a state, and the size of the node is proportional to the length of that state. The edge width indicates the value of the transition rates between the two connecting nodes, with thicker lines for faster transitions. The color gradient for promoter and enhancer states is used to illustrate the densities of activation marks. (*B*) The energy profile of the most probable transition path from the ground state to an enhancer state is shown. The line connecting the minima and transition states is provided as a guide for the eye. The length and density of activation marks at each minimum are indicated by the node size and color gradient as in part (*A*). (*C*) The most probable transition path from the ground state to the H3K9me3 heterochromatin is shown. To see this figure in color, go online.

Wang-Landau algorithm (25). Transition rates between domain-level chromatin states were further estimated using transition states identified in the simulation trajectory with the transition state theory (41). The detailed algorithm for rate calculations is provided in the Materials and Methods. For simplicity, only transitions with energetic barriers less than 15 k_BT are shown in Fig. 5 *A*. Network connections therefore quantify the likelihood, caused by random fluctuation, for converting a chromatin segment with a particular set of histone marks to a different state. A notable feature of this network is its apparent modularity, and domain-level chromatin states with similar biological functions naturally organize into highly connected clusters. The slow transition between clusters highlights the robustness of the overall organization of the epigenome in a given cell type to ensure stable gene expression.

Assuming Markovian dynamics, the time evolution of this kinetic network can be solved analytically to study the transition between any pairs of states without

performing computationally expensive, long timescale simulations (26). We expect the Markovian assumption to hold well, given that the network nodes are local minima in which the system will reside for a significant period to lose memory of the past and achieve dynamical decoupling.

To study the molecular mechanism for the formation of superenhancers, we solved the kinetic network and determined the most probable transition path from the ground state to an enhancer state that extends over six nucleosomes. We did not study enhancer states with much longer length as they are not included in the kinetic network because of low population. Details of the transition path calculations are provided in the Materials and Methods. As shown in Figs. 5 *B* and S3, we observe a sequential transition along which the chromatin becomes more open and more enriched with activation marks while the length of the state grows. Such a multistep, gradual transition indicates that superenhancers may not directly emerge from a genomic region along evolution or as cells differentiate but instead will undergo a maturation process. The presence of enhancer states with different potency could help cells fine-tune gene expression levels at various developmental stages. Furthermore, from detailed balance, one can show that the backward transition path from the enhancer state to the ground state is identical to the forward one. The sequential path shown in Fig. 5 *B*, therefore, also supports that superenhancers will take more steps and longer time to transit into the ground state than regular enhancers with shorter length (i.e., the red dots on the far right side of the path). Superenhancers are thus more stable with respect to perturbations and less likely to disappear from the chromatin landscape. The increased stability is in accord with their functional importance. Similar conclusions can be drawn from the transition path for the formation of a broad H3K4me3 promoter domain (see Fig. S4).

Next, we determined the transition path from the ground state to the periodic heterochromatin state with H3K9me3 marks. As shown in Fig. 5 *C*, we find that heterochromatin formation is a cooperative process, and there are no intermediate states along the pathway. To ensure the robustness of this observation, we further determined the transition path using a path deformation algorithm (42) that can search for the entire phase space and is not limited by the 100 states included in the kinetic network. As shown in Fig. S5, the new path shares the same transition barrier and does not exhibit any intermediate state either. The analysis for the heterochromatin state with the H3K27me3 mark supports similar conclusions (see Fig. S6). Cooperativity between nucleosomes will give rise to a bistable system, in which either all or none of the nucleosomes will become methylated, a phenomenon that has indeed been observed in many regulatory networks proposed for heterochromatin formation (43,44). Collective behavior between nucleosomes will significantly enhance the stability of the heterochromatin to ensure a robust inheritance of methylation marks across cell cycles. We will explore possible molecular mechanisms for such cooperativity in the next section.

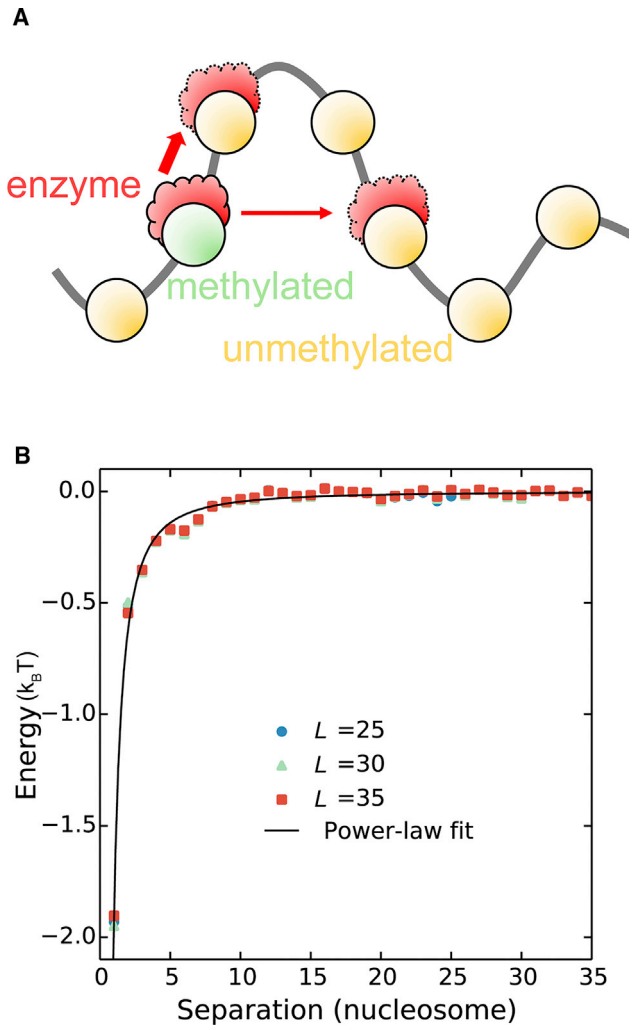

FIGURE 6 Information-theoretic energy landscape supports the globular heterochromatin conformation. (*A*) Shown is an illustration of the transfer of histone-modifying enzymes (*red*) from the methylated nucleosome (*green*) to others that are close in 3D space. This enzyme transfer is promoted by diffusion and, its rate is impacted by the distance between nucleosomes that increases at larger sequence separations. Internucleosome interaction energies arise from enzyme transfer and are therefore proportional to the contact probability between nucleosomes. (*B*) Internucleosome interaction energy for epigenetic mark H3K9me3 as a function of the genomic distance is shown. To see this figure in color, go online.

Internucleosome interactions support condensed, globular heterochromatin conformations

A striking finding from Fig. 4 is the presence of strong, attractive internucleosome interactions. These interactions measure the propensity for a pair of nucleosomes to share the same epigenetic mark and can arise if the two nucleosomes are in 3D contact to promote the transfer of corresponding

modification enzymes (43,45) or enzyme-enzyme interactions (44) (see Fig. 6 *A*). Consistent with this interpretation, all epigenetic marks exhibit a strong K^1 interaction energy as a result of the spatial proximity between the two neighboring in-sequence nucleosomes. Furthermore, the interaction energies typically decay at larger genomic separations as the nucleosomes become farther apart in space. It is therefore tempting to assume that internucleosome interaction energies are proportional to the contact probability between pairs of nucleosomes. Because different polymer configurations exhibit distinct trends in the variation of contact probabilities as a function of sequence length, studying the scaling behavior of interaction energies offers the opportunity to derive 3D chromatin domain conformations. Indeed, we find that interaction energies of the heterochromatin mark H3K9me3 decay slower than that of the activation mark H3K4me3, consistent with the fact that heterochromatin states are more condensed than euchromatin.

Quantitatively inferring scaling exponents from Fig. 4 can be challenging, however, because of the limited number of internucleosome couplings included. To more accurately determine the value of interaction energies at large genomic separations, we studied an additional model consisting of 201 nucleosomes. We only included the H3K9me3 mark for computational efficiency, and each nucleosome can either be methylated or free of modifications. Because H3K9me3 does not co-localize strongly with other epigenetic marks, this single mark model is expected to provide a good description for heterochromatin formation.

The single mark model adopts the same Hamiltonian as in Eq. 8. We included internucleosome interactions to a much longer range than the multimark model and determined the parameters again using the Boltzmann learning algorithm. Fig. 6 presents the resulting interaction energies as a function of nucleosome separation. These energies are robust and insensitive to the cutoff length L as well as the parameters used to process and binarize the raw data (see Fig. S7). We further fitted the interaction energies for $L = 35$ with a power law expression $E(l) = al^{\alpha}$ and obtained a value for $\alpha = -1.61$ (± 0.07). An exponent that falls in the range between -2 and -1 has been shown to support a phase transition in the one-dimensional Ising model (46), giving rise to the bistability seen in Fig. 5 *C*. Assuming that interaction energies are proportional to the contact probability $P(l)$, an exponent of -1.5 suggests that the constitutive heterochromatin conformation is consistent with an equilibrium globule (47). Globular conformations differ significantly from the rigid fibril structures with a diameter of 30 nm and can be stabilized by phase-separated liquid droplets formed by heterochromatin protein 1 (48–53).

DISCUSSION

Genome-wide histone modification profiles provide a comprehensive characterization of the chromatin landscape. Using an information-theoretic model and rigorous statistical mechanical tools, we demonstrated that epigenomics data could shed light on the mechanism of domain-level chromatin state formation as well. In particular, we found that heterochromatin states exhibit bistability and form in a highly cooperative process. On the other hand, long enhancer and promoter states grow from intermediate states of shorter length via a sequential process. These observations have significant implications on the establishment of domain-level chromatin states across cell cycles and as cells differentiate.

To support the biological relevance of transition paths predicted by the information-theoretic energy landscape, we examined the variation of epigenetic profiles as the differentiation progresses from mesodermal to IMR90 cells. Specifically, we determined, for chromatin segments from IMR90 cells that are 25 nucleosomes in length and fully marked with H3K27me3 or H3K4me1, the number of corresponding methylation marks in mesodermal cells. As shown in Fig. 7 *A*, the probability distribution for H3K27me3 numbers is bimodal, with two peaks at 0 and 25 nucleosomes, respectively. This bimodality is consistent with an all-or-none transition and supports the cooperative formation of heterochromatin. In the meantime, the probability distribution for H3K4me1 is almost uniform, with

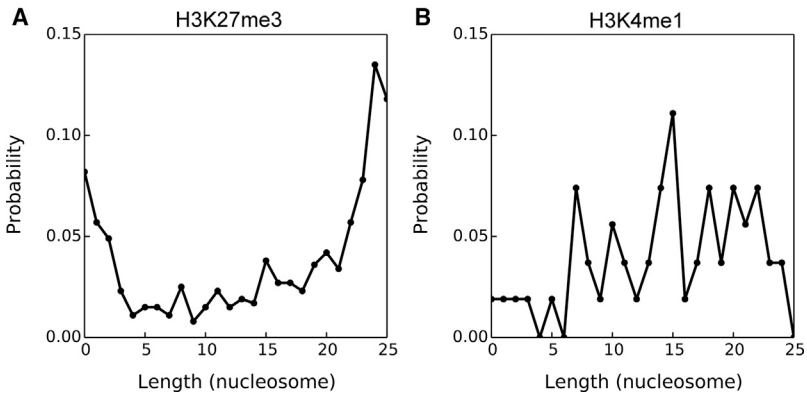

FIGURE 7 Variations of epigenomics profiles along cell differentiation are consistent with predictions from transition paths of domain-level chromatin state formation. (*A*) Shown is the probability for the number of H3K27me3 marks in 25-nucleosome-long chromatin segments from mesodermal cells given that the same segments are fully marked in IMR90 cells. (*B*) Shown is a similar plot as in part (*A*) but for the H3K4me1 mark.

significant contributions at intermediate lengths. Therefore, superenhancers in IMR90 can grow from weaker ones formed in mesodermal cells.

The information-theoretic energy landscape further predicts strong and direct interactions between nucleosomes that are far apart in sequence. These interactions support the essential role of chromatin looping in facilitating the spreading of epigenetic marks across nucleosomes. Experimentally validating the role of looping in domain-level chromatin state formation can be challenging and requires high temporal resolution to monitor the dynamics of histone modification, but some early evidence is emerging (54). The effective interaction between methylation marks also significantly constrains the corresponding chromatin organization. By studying the scaling behavior of these interactions, we argued that the 3D conformation of constitutive heterochromatin is consistent with a globular conformation but not rigid fibril structures. We note that determining the chromatin structure at the kilobase range is extremely difficult, and the existence of 30 nm chromatin fibers in vivo remains controversial (55–57), though increasing evidence argues against their presence (58–60). To our knowledge, this study is the first demonstration in deriving structural information for chromatin from epigenomics data.

SUPPORTING MATERIAL

Supporting Material can be found online at https://doi.org/10.1016/j.bpj.2019.04.006.

AUTHOR CONTRIBUTIONS

W.J.X. and B.Z. designed research, performed research, analyzed data, and wrote the manuscript.

ACKNOWLEDGMENTS

We thank A.K. Chakraborty and X. Che for helpful discussions.

This work was supported by the National Science Foundation grant MCB-1715859.

REFERENCES

1. Bernstein, B. E., A. Meissner, and E. S. Lander. 2007. The mammalian epigenome. *Cell.* 128:669–681.
2. Goldberg, A. D., C. D. Allis, and E. Bernstein. 2007. Epigenetics: a landscape takes shape. *Cell.* 128:635–638.
3. Bickmore, W. A., and B. van Steensel. 2013. Genome architecture: domain organization of interphase chromosomes. *Cell.* 152:1270–1284.
4. Turner, B. M. 2000. Histone acetylation and an epigenetic code. *BioEssays.* 22:836–845.
5. Cuvier, O., and B. Fierz. 2017. Dynamic chromatin technologies: from individual molecules to epigenomic regulation in cells. *Nat. Rev. Genet.* 18:457–472.
6. Probst, A. V., E. Dunleavy, and G. Almouzni. 2009. Epigenetic inheritance during the cell cycle. *Nat. Rev. Mol. Cell Biol.* 10:192–206.
7. Margueron, R., and D. Reinberg. 2010. Chromatin structure and the inheritance of epigenetic information. *Nat. Rev. Genet.* 11:285–296.
8. Whitehouse, I., and D. J. Smith. 2013. Chromatin dynamics at the replication fork: there's more to life than histones. *Curr. Opin. Genet. Dev.* 23:140–146.
9. Grewal, S. I., and S. Jia. 2007. Heterochromatin revisited. *Nat. Rev. Genet.* 8:35–46.
10. Soshnev, A. A., S. Z. Josefowicz, and C. D. Allis. 2016. Greater than the sum of parts: complexity of the dynamic epigenome. *Mol. Cell.* 62:681–694.
11. Kundaje, A., W. Meuleman, …, M. Kellis; Roadmap Epigenomics Consortium.. 2015. Integrative analysis of 111 reference human epigenomes. *Nature.* 518:317–330.
12. Ernst, J., and M. Kellis. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods.* 9:215–216.
13. Hoffman, M. M., J. Ernst, …, W. S. Noble. 2013. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.* 41:827–841.
14. Zhou, J., and O. G. Troyanskaya. 2016. Probabilistic modelling of chromatin code landscape reveals functional diversity of enhancer-like chromatin states. *Nat. Commun.* 7:10528.
15. Strahl, B. D., and C. D. Allis. 2000. The language of covalent histone modifications. *Nature.* 403:41–45.
16. Rando, O. J. 2012. Combinatorial complexity in chromatin structure and function: revisiting the histone code. *Curr. Opin. Genet. Dev.* 22:148–155.
17. Marco, E., W. Meuleman, …, G. C. Yuan. 2017. Multi-scale chromatin state annotation using a hierarchical hidden Markov model. *Nat. Commun.* 8:15011.
18. Parker, S. C., M. L. Stitzel, …, F. S. Collins; NISC Comparative Sequencing Program; National Institutes of Health Intramural Sequencing Center Comparative Sequencing Program Authors; NISC Comparative Sequencing Program Authors. 2013. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc. Natl. Acad. Sci. USA.* 110:17921–17926.
19. Whyte, W. A., D. A. Orlando, …, R. A. Young. 2013. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell.* 153:307–319.
20. Benayoun, B. A., E. A. Pollina, …, A. Brunet. 2014. H3K4me3 breadth is linked to cell identity and transcriptional consistency. *Cell.* 158:673–688.
21. Kuan, P. F., D. Chung, …, S. Keleş. 2011. A statistical framework for the analysis of ChIP-Seq data. *J. Am. Stat. Assoc.* 106:891–903.
22. Rotem, A., O. Ram, …, B. E. Bernstein. 2015. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.* 33:1165–1172.
23. Ackley, D. H., G. E. Hinton, and T. J. Sejnowski. 1985. A learning algorithm for boltzmann machines. *Cogn. Sci.* 9:147–169.
24. Barton, J. P., M. Kardar, and A. K. Chakraborty. 2015. Scaling laws describe memories of host-pathogen riposte in the HIV population. *Proc. Natl. Acad. Sci. USA.* 112:1965–1970.
25. Zhou, Q. 2011. Random walk over basins of attraction to construct ising energy landscapes. *Phys. Rev. Lett.* 106:180602.
26. E, W., and E. Vanden-Eijnden. 2010. Transition-path theory and path-finding algorithms for the study of rare events. *Annu. Rev. Phys. Chem.* 61:391–420.
27. Feingold, E. A., P. J. Good, …, S. C. Harvey; ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia of DNA elements) project. *Science.* 306:636–640.
28. Schneidman, E., M. J. Berry, II, …, W. Bialek. 2006. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature.* 440:1007–1012.

29. Morcos, F., A. Pagnani, …, M. Weigt. 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci.* 108:E1293–E1301.

30. Ferguson, A. L., J. K. Mann, …, A. K. Chakraborty. 2013. Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. *Immunity.* 38:606–617.

31. Zhang, B., and P. G. Wolynes. 2015. Topology, structures, and energy landscapes of human chromosomes. *Proc. Natl. Acad. Sci. USA.* 112:6062–6067.

32. Zhang, B., and P. G. Wolynes. 2016. Shape transitions and chiral symmetry breaking in the energy landscape of the mitotic chromosome. *Phys. Rev. Lett.* 116:248101.

33. Zhang, B., and P. G. Wolynes. 2017. Genomic energy landscapes. *Biophys. J.* 112:427–433.

34. Di Pierro, M., B. Zhang, …, J. N. Onuchic. 2016. Transferable model for chromosome architecture. *Proc. Natl. Acad. Sci. USA.* 113:12168–12173.

35. Mora, T., A. M. Walczak, …, C. G. Callan, Jr. 2010. Maximum entropy models for antibody diversity. *Proc. Natl. Acad. Sci. USA.* 107:5405–5410.

36. Zhou, J., and O. G. Troyanskaya. 2014. Global quantitative modeling of chromatin factor interactions. *PLoS Comput. Biol.* 10:e1003525.

37. Andersson, R., A. Sandelin, and C. G. Danko. 2015. A unified architecture of transcriptional regulatory elements. *Trends Genet.* 31:426–433.

38. Wang, J., S. T. Lawry, …, S. Jia. 2014. Chromosome boundary elements and regulation of heterochromatin spreading. *Cell. Mol. Life Sci.* 71:4841–4852.

39. Jin, W., Q. Tang, …, K. Zhao. 2015. Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. *Nature.* 528:142–146.

40. Nguyen, H. C., R. Zecchina, and J. Berg. 2017. Inverse statistical problems: from the inverse Ising problem to data science. *Adv. Phys.* 66:197–261.

41. Pechukas, P. 1981. Transition state theory. *Annu. Rev. Phys. Chem.* 32:159–177.

42. Nemoto, K. 1988. Metastable states of the SK spin glass model. *J. Phys. Math. Gen.* 21:L287–L294.

43. Dodd, I. B., M. A. Micheelsen, …, G. Thon. 2007. Theoretical analysis of epigenetic cell memory by nucleosome modification. *Cell.* 129:813–822.

44. Zhang, H., X. J. Tian, …, J. Xing. 2014. Statistical mechanics model for the dynamics of collective epigenetic histone modification. *Phys. Rev. Lett.* 112:068101.

45. Erdel, F., K. Müller-Ott, and K. Rippe. 2013. Establishing epigenetic domains via chromatin-bound histone modifiers. *Ann. N. Y. Acad. Sci.* 1305:29–43.

46. Dyson, F. J. 1969. Existence of a phase-transition in a one-dimensional Ising ferromagnet. *Commun. Math. Phys.* 12:91–107.

47. Mirny, L. A. 2011. The fractal globule as a model of chromatin architecture in the cell. *Chromosome Res.* 19:37–51.

48. Strom, A. R., A. V. Emelyanov, …, G. H. Karpen. 2017. Phase separation drives heterochromatin domain formation. *Nature.* 547:241–245.

49. Larson, A. G., D. Elnatan, …, G. J. Narlikar. 2017. Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin. *Nature.* 547:236–240.

50. Klosin, A., and A. A. Hyman. 2017. Molecular biology: a liquid reservoir for silent chromatin. *Nature.* 547:168–170.

51. Hnisz, D., K. Shrinivas, …, P. A. Sharp. 2017. A phase separation model for transcriptional control. *Cell.* 169:13–23.

52. Sabari, B. R., A. Dall'Agnese, …, R. A. Young. 2018. Coactivator condensation at super-enhancers links phase separation and gene control. *Science.* 361:eaar3958.

53. Shi, G., L. Liu, …, D. Thirumalai. 2018. Interphase human chromosome exhibits out of equilibrium glassy dynamics. *Nat. Commun.* 9:3161.

54. Oksuz, O., V. Narendra, …, D. Reinberg. 2018. Capturing the onset of PRC2-mediated repressive domain formation. *Mol. Cell.* 70:1149–1162.e5.

55. Luger, K., M. L. Dechassa, and D. J. Tremethick. 2012. New insights into nucleosome and chromatin structure: an ordered state or a disordered affair? *Nat. Rev. Mol. Cell Biol.* 13:436–447.

56. Schlick, T., J. Hayes, and S. Grigoryev. 2012. Toward convergence of experimental studies and theoretical modeling of the chromatin fiber. *J. Biol. Chem.* 287:5183–5191.

57. Boulé, J. B., J. Mozziconacci, and C. Lavelle. 2015. The polymorphisms of the chromatin fiber. *J. Phys. Condens. Matter.* 27:033101.

58. Joti, Y., T. Hikima, …, K. Maeshima. 2012. Chromosomes without a 30-nm chromatin fiber. *Nucleus.* 3:404–410.

59. Sanborn, A. L., S. S. Rao, …, E. L. Aiden. 2015. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci. USA.* 112:E6456–E6465.

60. Ou, H. D., S. Phan, …, C. C. O'Shea. 2017. ChromEMT: visualizing 3D chromatin structure and compaction in interphase and mitotic cells. *Science.* 357:eaag0025.

Biophysical Journal, Volume *116*

Supplemental Information

Learning the Formation Mechanism of Domain-Level Chromatin States with Epigenomics Data

Wen Jun Xie and Bin Zhang

Supplementary Information for "Learning the Formation Mechanism of Domain-Level Chromatin States with Epigenomics Data"

Wen Jun Xie and Bin Zhang
Departments of Chemistry, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139-4307

1. Estimating inter-nucleosome correlation length of epigenetic marks

To quantitatively probe the strength of inter-nucleosome interactions, we fitted the correlation coefficients between epigenetic marks as a function of nucleosome separation (l) with the expression $a * \exp\left(-\frac{l}{b}\right) + c$. The fitting parameter b is used to measure the correlation length of epigenetic marks. Fitting results for self-correlation of the same mark are shown in Table S1.

Most of the correlation lengths are smaller than 25 nucleosomes, supporting our use of $N = 25$ as the system size. For marks with self-correlation lengths longer than 25, they often support the formation of large, homogeneous chromatin domains. The effect of these marks will therefore be well captured by the periodic boundary condition.

2. Deriving information-theoretic models from the maximum entropy principle

As mentioned in the main text, $E(\boldsymbol{s}, L)$ is the most probable model to reproduce experimental constraints. Here we provide a detailed derivation for its expression based on the maximum entropy principle.

To start, we define the system of interest as a chain of N nucleosomes, each one of which is characterized by M epigenetic marks. A configuration, or state, of this system is denoted as $\boldsymbol{s} = \{[s_1^1, \dots, s_1^N], \dots, [s_M^1, \dots, s_M^N]\}$, which can be viewed as a $M \times N$ matrix and s_i^n denotes whether mark i is present $(= 1)$ at the n-th nucleosome or not $(= 0)$. Next, we seek for a probability distribution function of state $\boldsymbol{s}$ that maximizes the information entropy

$$S = -\sum_{\boldsymbol{s}} P(\boldsymbol{s}) \ln P(\boldsymbol{s})$$

while reproducing the mean occupancy of each epigenetic mark i,

$$\frac{1}{N}\sum_{\boldsymbol{s}}\sum_{n} P(\boldsymbol{s}) s_i^n = \langle s_i \rangle_{\text{exp}}, \qquad \text{for } i = 1, \dots, M$$

and the correlation between two epigenetic marks i and j separated by l nucleosomes, $\langle s_i s_j \rangle_{\text{exp}}^l$,

$$\frac{1}{N}\sum_{\boldsymbol{s}}\sum_{n} P(\boldsymbol{s}) s_i^n s_j^{n+l} = \langle s_i s_j \rangle_{\text{exp}}^l, \qquad \text{for } l = 0, \dots, L, \;\; 1 \le i \le j \le M.$$

The solution to this constrained optimization problem can be found by minimizing the following Lagrangian

$$L = -\sum_{\boldsymbol{s}} P(\boldsymbol{s})\ln P(\boldsymbol{s}) + \sum_{n}\sum_{i}\left[h_i\left(\langle s_i\rangle_{\exp} - \sum_{\boldsymbol{s}} P(\boldsymbol{s}) s_i^n\right) + \sum_{\substack{j\geq i,\ \text{for } l>0 \\ j>i,\text{for } l=0}}^{M}\sum_{l=0}^{L} K_{ij}^{l}\left(\langle s_i s_j\rangle_{\exp}^{l} - \sum_{\boldsymbol{s}} P(\boldsymbol{s}) s_i^n s_j^{n+l}\right)\right]$$

where h_i and K_{ij}^{l} are Lagrangian multipliers.

The solution for $\frac{\partial L}{\partial P(\boldsymbol{s})} = 0$ provides the least structured $P(\boldsymbol{s})$ in the form of a Boltzmann distribution

$$P(\boldsymbol{s}) = e^{-E(\boldsymbol{s},L)}/Z,$$

where the energy function is defined as

$$E(\boldsymbol{s},L) = \sum_{n}\sum_{i}(h_i s_i^n + \sum_{j>i} K_{ij}^{0} s_i^n\, s_j^n + \sum_{j\geq i}^{M}\sum_{l=1}^{L} K_{ij}^{l} s_i^n s_j^{n+l}),$$

and Z is the partition function. The unit of energy is $\mathrm{k_B}T$.

Following the argument above, we can arrive at a series of information theoretic models presented in the main text by adjusting the input experimental constraints. For example, if only $\langle s_i\rangle_{\exp}$ and $\langle s_i s_j\rangle_{\exp}^{0}$, for $1 \leq i < j \leq M$ are provided as constraints, then we arrive at the *intra-nucleosome model*,

$$E(\boldsymbol{s},L) = \sum_{n}\sum_{i}(h_i s_i^n + \sum_{j>i} J_{ij} s_i^n\, s_j^n).$$

If $L \times M$ additional self-correlation coefficients $\langle s_i s_i\rangle_{\exp}^{l}$, for $i = 1,\cdots,M$, and $l = 1,\cdots,L$ are provided as constraints, then we arrive at the *inter-nucleosome model with only self-interactions*,

$$E(\boldsymbol{s},L) = \sum_{n}\sum_{i}(h_i s_i^n + \sum_{j>i} J_{ij} s_i^n\, s_j^n + \sum_{l=1}^{L} K_i^{l} s_i^n s_i^{n+l}).$$

Finally, if we further introduce cross-mark correlation between neighboring nucleosomes $\langle s_i s_j\rangle_{\exp}^{1}$, for $1 \leq i < j \leq M$ as constraints, we arrive at the *inter-nucleosome model with both self- and cross-mark interactions*,

$$E(\boldsymbol{s},L) = \sum_{n}\sum_{i}(h_i s_i^n + \sum_{j>i} J_{ij} s_i^n\, s_j^n + \sum_{l=1}^{L} K_i^{l} s_i^n s_i^{n+l} + \sum_{j>i} K_{ij}^{1} s_i^n\, s_j^{n+1})$$

3. Estimating transition barriers using Wang-Landau sampling

We used the following approach to determine the transition barriers between pairs of chromatin domains.

First, we collected an ensemble of transition paths connecting the two domains from a simulation trajectory conducted using the Metropolis Monte Carlo algorithm coupled with the Wang-Landau algorithm (1). The Wang-Landau algorithm is an enhanced sampling technique that aims at producing a random walk over basins of attraction, thus greatly accelerating the transition between local minima. It estimates the density of states of the system along the simulation and use it to bias the sampling. In the following, we briefly summarize the key steps of this algorithm when applied to a spin glass model.

Suppose that we are interested in the top N most populated minima of the spin glass model. By assigning spin configurations to their closest local minimum using the steepest descent algorithm outlined in the previous section, the phase space thus can be partitioned into $N+1$ basins $B_0, B_1, B_2, \dots, B_N$. B_0 includes all the configurations that cannot be assigned to the top N minima. To estimate the density of state for each basin, we further partition the energy space $u_1 < u_2 < \cdots < u_L = \infty$ into L intervals, where u_1 is the lower bound of the model Hamiltonian, and is set to zero in our case. Therefore, the phase space is now partitioned into $(N+1) \times L$ regions, $B_{im} = \{\boldsymbol{s} \in B_i : E(\boldsymbol{s}) \in [u_m, u_{m+1})\}$, for $i = 0, \cdots, N$ and $m = 1, \cdots L$. The statistical weight (density of states), g_{im}, for each region is set as one initially.

We then update the density of states by performing the following Monte Carlo sampling. Starting from a random configuration $\boldsymbol{s}$ in the basin region B_{im}, we perform a Monte Carlo move with a single-spin flip. This move will lead to a new configuration $\boldsymbol{s}'$ in the region B_{jn}. The new configuration will be accepted with the following probability

$$p(\boldsymbol{s} \to \boldsymbol{s}') = \min\{1, e^{\beta\left(E(\boldsymbol{s}) - E(\boldsymbol{s}')\right)} g_{im}/g_{jn}\},$$

where β is an effective temperature.

Along the simulation, we keep track the history of the simulation trajectory with a histogram. Specifically, we will update the number of times a region has been visited H_{jn} by 1 if the new configuration $\boldsymbol{s}'$ is accepted. In the meantime, the corresponding density of states g_{jn} will be updated as $g_{jn} f$, where f is a scaling factor larger than 1. The goal of the Wang-Landau sampling is to achieve a flat histogram, indicating a random walk in the energy basins.

The above sampling will be conducted iteratively. For example, if the flatness of the histogram becomes acceptable (maximal fluctuation is less than 25%.), f will be scaled down to $\sqrt{f}$. The iteration completes when f is close to 1.

When applying the Wang-Landau algorithm to the informational theoretic model, we limited the sampling to the top 100 chromatin domains that cover more than 94% of the nucleosomes. The modification factor f of density of states is set as $e \approx 2.71828$ at the beginning of the sampling. The effective temperature β is set as $1/(2\mathrm{k_B}T)$ to accelerate the

sampling, resulting the Metropolis acceptance ratio of ~15%. The energy range of interest is [0,120 $k_B T$] and the energy width is 10 $k_B T$. We stopped the sampling when f<1.000001.

Second, from the ensemble of transition paths $\mathbb{P}_{AB}$ collected from Wang-Landau sampling, the energy for the transition state connecting chromatin domains A and B is identified as

$$E_{AB} = \min_{p \in \mathbb{P}_{AB}} \max_{s \in p} E(\boldsymbol{s}),$$

where $E(\boldsymbol{s})$ is the energy of configuration $\boldsymbol{s}$. The transition barrier is then obtained as $\Delta E_{AB} = E_{AB} - E_A$ for the transition from A and B, and $\Delta E_{BA} = E_{AB} - E_B$ for the reverse transition.

4. Calculating the minimum energy barrier path between chromatin domains using a path deformation approach

As an independent validation of the transition paths determined in the main text, we further determined the transition barrier and minimum energy path between the ground state and the H3K9me3 heterochromatin using a path deformation approach (2)(3). This approach explores the entire phase space when searching for the transition state. Therefore, unlike the method presented in the main text, it is not restricted by the number of chromatin domains included in the kinetic network.

For the convenience of discussion, we define the mark-flipping operator O that turns a particular epigenetic mark on or off. Since a continuous path connecting two states A and B consists of a set of configurations that differ only by one mark, it can be described by a product of operators, $B = O_1 O_2 \ldots O_n A$.

The path deformation algorithm starts with a randomly generated minimum length path, for which the number of operators equals to the total number of different mark configurations in the two states. Next, it aims to find the minimum transition barrier connecting the two states by iteratively deforming the highest energy configuration along the path. We used two types of moves to deform the path. The first one swaps the operator that leads to the highest energy configuration with the one leaving it. In the second move, two operators are added before and after the operator that leads to the highest energy configuration. These two operators modify the same mark randomly selected from the 12 epigenetic marks. They have opposite effects: if one switches a mark on, the other will turn it off. Newly deformed paths are only accepted if they lead to a decrease of the transition barrier.

We carried out a total of 1000 independent calculations to search for the transition barrier between the ground state and the heterochromatin domain. For each calculation, the iteration terminates when the total length of the path exceeds 20000 configurations. Remarkably, all these paths give rise to the same barrier as the one presented in the main text (Fig. 5C). From the transition state, we further constructed the minimum energy path using the steepest gradient descent algorithm. As shown in Fig. S5, the path exhibits no intermediate states, supporting a cooperative transition for the formation of constitutive heterochromatin.

5. Evaluating the sensitivity of interaction energy strength with respect to the Poisson threshold used in the binarization of experimental data

Genome wide histone modification profiles were processed with a Poisson background model to remove biases that might arise from the DNA sequence (4). A Poisson threshold of 0.0001, a typical number used in the widely popular software ChromHMM (5), was used to detect significant experimental signals compared to the control data. A variation of the threshold value could significantly affect mark occupancy, and the correlation coefficients. Indeed, as shown in Fig. S7A, the inter-nucleosome correlation changes significantly as we vary the threshold by a factor of two.

To investigate the sensitivity of the derived interaction energies with respect to the Poisson threshold, we reparametrized the simplified heterochromatin domain shown in Fig. 6 of the main text using these new correlation coefficients. It is reassuring to see that the inter-nucleosome interaction energies with different thresholds are quite close to each other (see Fig. S7B). Fitting these interaction energies using a power-law expression $E(l) = al^{\alpha}$ provides scaling exponents of -1.59 ± 0.06 and -1.45 ± 0.06 for the threshold 0.00005 and 0.0002, respectively. These number are very close to the value $-1.61(\pm 0.07)$ provided in the main text.

References:

1. Zhou Q (2011) Random walk over basins of attraction to construct ising energy landscapes. *Phys Rev Lett* 106:180602.
2. Nemoto K (1988) Metastable states of the SK spin glass model. *J Phys A Math Gen* 21(5):L287–L294.
3. Vertechi D, Virasoro MA (1989) Energy barriers in SK spin-glass model. *J Phys Fr* 50(17):2325–2332.
4. Kuan PF, et al. (2011) A statistical framework for the analysis of ChIP-Seq data. *J Am Stat Assoc* 106:891–903.
5. Ernst J, Kellis M (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* 9:215.

Table S1. Correlation length of epigenetic marks. See text *Section: Estimating inter-nucleosome correlation length of epigenetic marks* for a detailed discussion.

Epigenetic mark	Correlation length (nucleosome)
H3K9me3	5.6
H3K27me3	24.1
H4K20me1	53.1
H3K79me2	45.1
H3K36me3	37.1
H3K4me1	5.8
H3K4me2	5.9
H3K4me3	6.2
H3K9ac	6.9
H3K27ac	8.0
DNase	1.0
H2A.Z	3.6

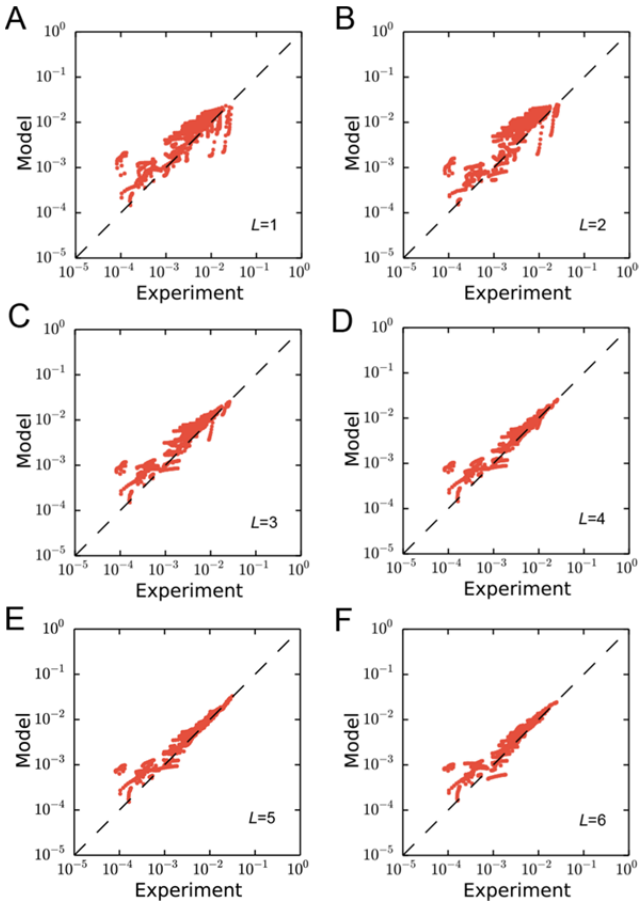

Fig. S1. Comparison between experimental inter-nucleosome correlations and predictions from information-theoretic models $E(s, L)$ with different L. In making these plots, the same experimental data as in Fig. 2B of the main text was used.

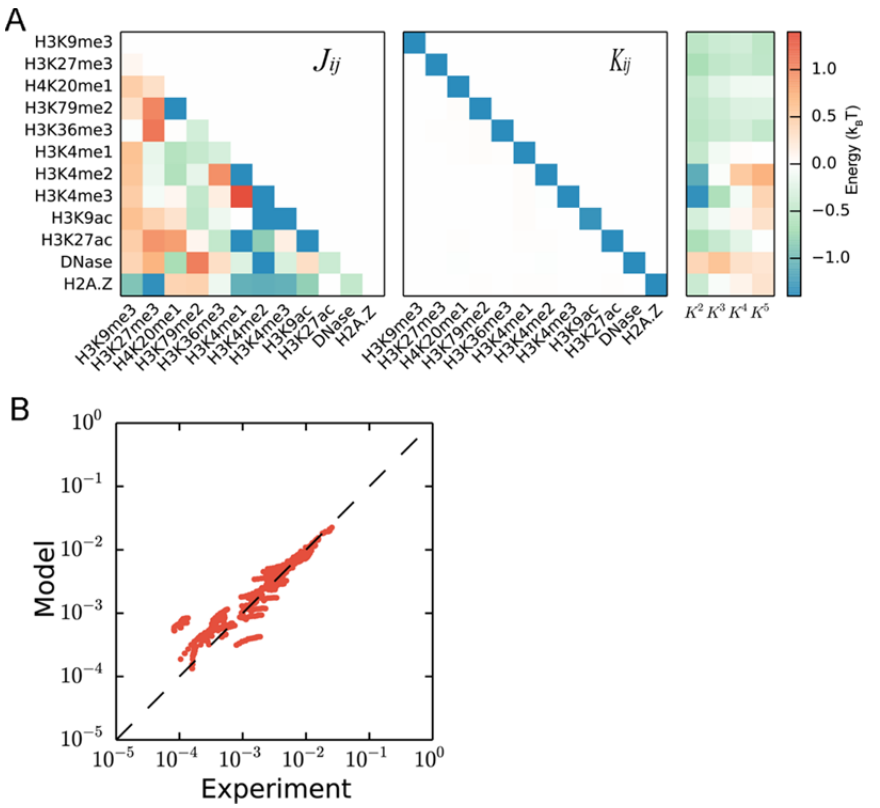

Fig. S2. Performance of an information theoretic model with explicit coupling between all pairs of epigenetic marks on neighboring nucleosomes. (A) Intra-nucleosome interaction (*left*), nearest neighbor interaction (*middle*), and long-range interaction (right) energies of the model. An explicit expression of the model Hamiltonian is provided in the *Section: Deriving information-theoretic models from the maximum entropy principle*. The cross-mark interaction energies are significantly smaller compared to the self-interaction, and are barely visible in the middle panel. (B) Comparison between experimental and simulated inter-nucleosome correlations. The experimental data is identical to those shown in Fig. 2B of the main text. The relative error between simulated and experimental data is 17.6%. The slope and R-squared of the linear fit are 0.93 and 0.97, respectively. We therefore conclude that including the cross-mark coupling on the nearest neighbor nucleosomes doesn't improve the model performance significantly. For simplicity, these terms were not included in the models presented in the main text.

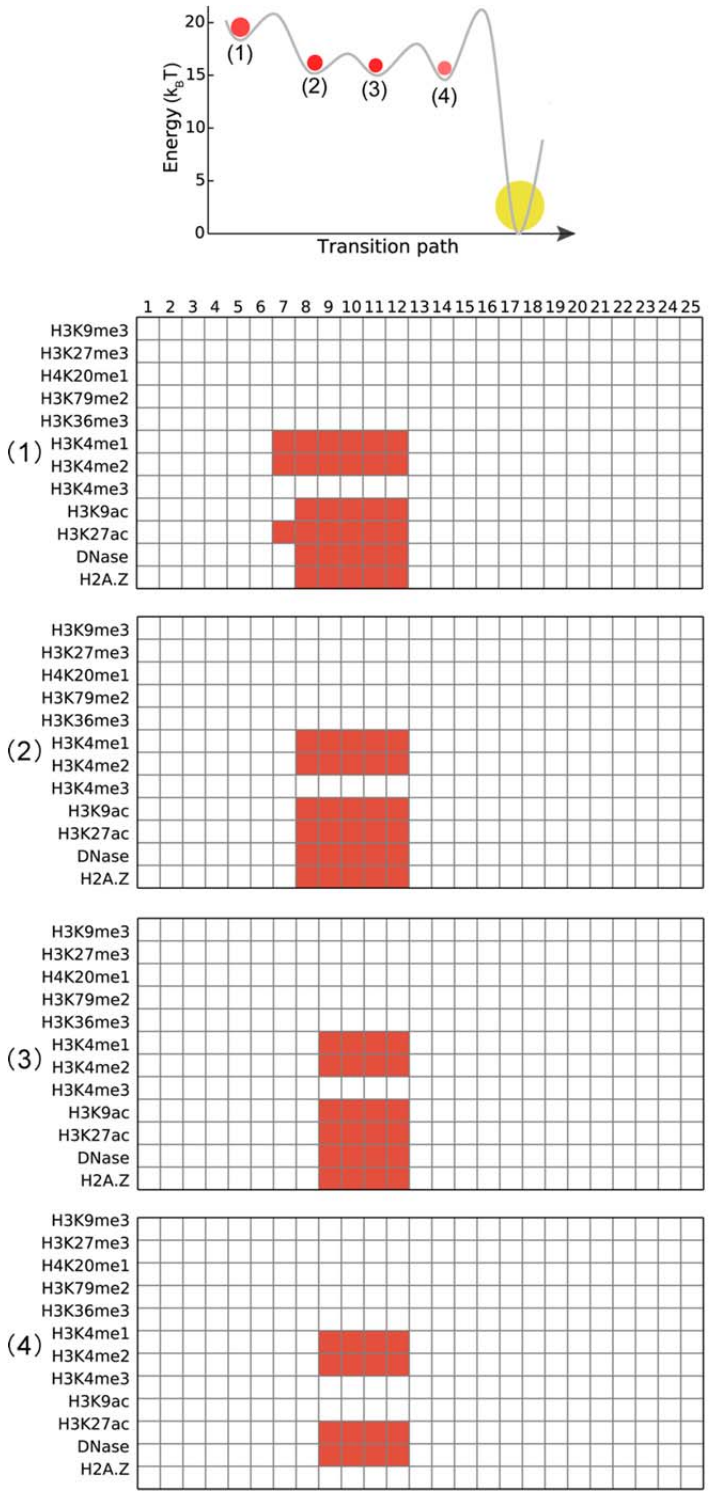

Fig. S3. Configurations for intermediate states along the most probable transition path from the ground state to an enhancer domain. The same path as in Fig. 5B is shown at the top for reference. The rows in each configuration correspond to different epigenetic marks, and the columns represent different nucleosomes. A grid is marked with red if the epigenetic mark is turned on for that nucleosome.

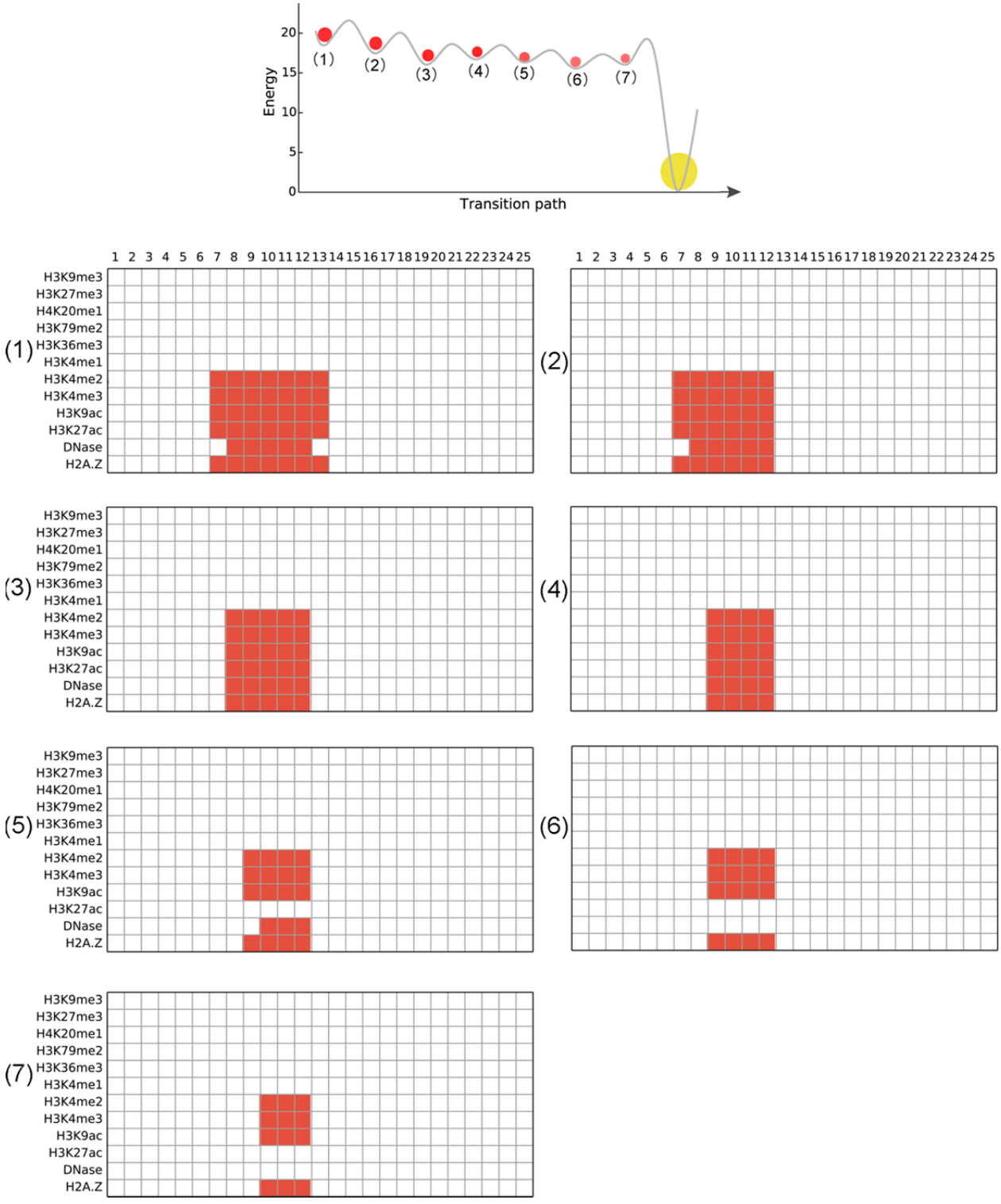

Fig. S4. Configurations for intermediate states along the most probable transition path from the ground state to a promoter domain. Similar to the path shown in Fig. 5B of the main text, this path exhibits multiple intermediate states as well.

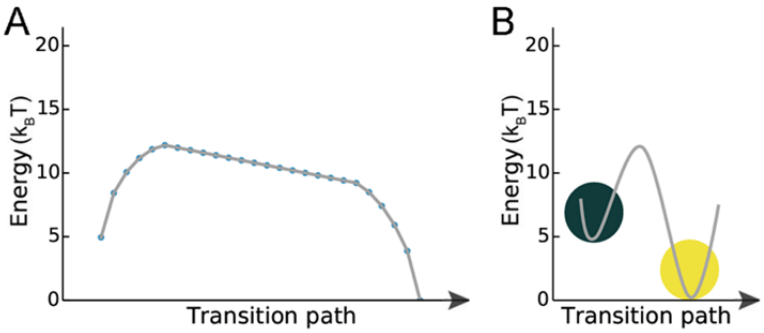

Fig. S5. Transition paths from the ground state to the H3K9me3 heterochromatin domain. (A) Transition path constructed using the path deformation approach. The states along the transition path are shown in dots. Details for this calculation are provided in the *Section: Calculating the minimum energy barrier path between chromatin domains using a path deformation approach.* (B) Transition path constructed from the Wang-Landau sampling. This figure is the same as Fig. 5C in the main text.

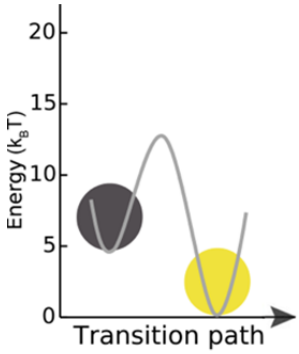

Fig. S6. The most probable transition path from the ground state to the H3K27me3 heterochromatin domain. Similar to the path shown in Fig. 5C of the main text for the H3K9me3 heterochromatin, there are no intermediate states along the path, and the transition is a cooperative process.

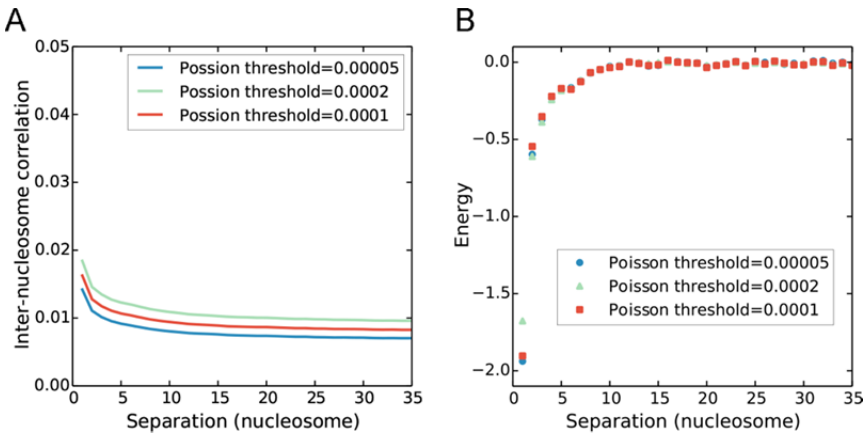

Fig. S7. Inter-nucleosome (A) correlation and (B) interaction energies for H3K9me3 as a function of the genomic distance derived from data processed with different Poisson thresholds. See the text *Section: Evaluating the sensitivity of interaction energy strength with respect to the Poisson threshold used in the binarization of experimental data* for details.