Supplementary Material

Gamze Gürsoy¹,†, Yun Xu¹,†, Amy L. Kenter² and Jie Liang*¹

¹Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60607, USA

²Department of Microbiology and Immunology, University of Illinois College of Medicine, Chicago, IL 60612, USA

†These authors contributed equally to this work

May 28, 2017

*jliang@uic.edu; Corresponding Author
METHODS

Normalization and calculation of propensity

The 5C interaction frequency \( f_{\text{cell}}^{SC}(i, j) \) between each nodes \( i \) and \( j \) is mapped and a total of 367 and 425 unique pairwise chromatin interactions are obtained for the K562 and GM12878 cell lines from the 5C data [1]. The interaction frequency \( f^{\text{null}}(i, j) \) between each nodes \( i \) and \( j \) that occur due to the available space in cell nucleus, excluded volume effect and the chain connectivity is calculated using the C-SAC model [2].

First, we normalized the interaction frequencies by total number of 5C interaction frequencies for each cell line and random model as following,

\[
q_{\text{cell}}^{SC}(i, j) = \frac{f_{\text{cell}}^{SC}(i, j)}{\sum_{i=1}^{\text{N}} \sum_{j=1}^{\text{N}} f_{\text{cell}}^{SC}(i, j)}
\]

\[
q^{\text{null}}(i, j) = \frac{\sum_{k} w_k I(i, j)}{\sum_{i=1}^{\text{N}} \sum_{j=1}^{\text{N}} \sum_{k} w_k I(i, j)},
\]

where \( N \) is the total number of nodes, \( w_k \) is the weight of the \( k^{th} \) chain in the ensemble and \( I(i, j) \) is an indicator function, which equals to 1 when nodes \( i \) and \( j \) interacts, equals to 0 otherwise. We calculated the propensity of each 5C interaction as,

\[
\text{prop}_{\text{cell}}^{SC}(i, j) = \frac{q_{\text{cell}}^{SC}(i, j)}{q^{\text{null}}(i, j)}
\]

Details of exclusion of non-specific physical interactions from 5C Data

Bootstrap and False Discovery Rate

**Calculation of p-value.** To test if the interaction between nodes \( i \) and \( j \) is significant, we compare how many times \( q_{\text{cell}}^{SC}(i, j) \) (the normalized interaction frequency between nodes \( i \) and \( j \) in 5C data) is less than \( q^{\text{null}}(i, j) \) by bootstrapping 1000 times of 100,000 random C-SAC chains with
replacement. $q^{null}(i, j)_m = \frac{\sum_{k'} w_{k'} I(i, j)}{\sum_i \sum_j \sum_{k'} w_{k'} I(i, j)}$ is the normalized interactions frequency of nodes in the random C-SAC ensemble, and $w_{k'}$ is the weight of $k'$th random chain from the $m$th bootstrapped 100,000 samples with replacement. The $p$-values $p_{ij}$ of interaction is:

$$p_{ij} = \frac{\sum_{m=1}^{M} I(q^{5C}_{cell}(i, j) < q^{null}(i, j)_m)}{M},$$

where $M = 1000$, and $I(\cdot)$ is a indicator function, which equals to 1 when condition is satisfied, equals to 0 otherwise. FDR correction is employed for each interactions with a genomic separation $s = |i - j|$. For each constant $s$, we sorted $p_{ij}$ ascendantly to get new $p$-value set $\{p_{ij}^{(m)}\}$, such that $p_{ij}^{(1)} \leq p_{ij}^{(2)} \leq \cdots \leq p_{ij}^{(m)}$ are ordered, where $m$ is the total number of the $p$-values in the set $\{K|K = j - i\}$. We then used Hochberg adjustment method [3] to adjust $p$-values $p_{ij}^{(m)}$,

$$\hat{p}_{ij}^{(l)} = \begin{cases} p_{ij}^{(m)} & \text{for } l = m, \\ \min(\hat{p}_{ij}^{(l+1)}, \frac{m}{l} p_{ij}^{(l)}) & \text{for } l = m - 1, \ldots, 1. \end{cases}$$

After the FDR adjustment, if the $p$-value is less than a significance level $\alpha = 5\%$, we reject the null hypothesis.

**Details of nC-SAC Model**

**Obtaining Distance Constraints from Significant 5C Interaction Frequencies**

After the calculation of propensities described in the Methods section, we assume that the relationship between the propensity $prop_{ij}$ and the distance constraint $d_{ij}$ between node $i$ and $j$ follows half Gaussian distribution,

$$\frac{prop_{ij}}{\max prop_{ij}} = \exp \frac{-(d_{ij} - \mu)^2}{2\sigma^2}, \text{ and } d_{ij} > \mu.$$  

where $\sigma = \sqrt{\frac{d_{c} - \mu}{2\log \frac{\max prop_{ij}}{\min prop_{ij}}}}$. In equation 3, the entry $d_{ij}$ of the distance matrix $D$ between node $i$ and
node $j$ can be obtained as following

$$d_{ij} = \mu + \sqrt{2\sigma^2 \log \frac{\max \text{prop}_{ij}}{\text{prop}_{ij}}}, \text{ and prop}_{ij} > 0. \quad (4)$$

where $\mu$ is 30 nm which is the minimum possible distance between any node $i$ and $j$ and $\max \text{prop}_{ij}$ is the maximum propensity. The maximum possible distance between any node $i$ and $j$ is taken as 80 nm following experimentally determined threshold [4].

**Geometrical Sequential Importance Sampling Algorithm for nC-SAC Model**

Conformations that satisfy the spatial distance constraints can be generated by minimizing an error function. This function measures the deviations from the desired spatial distances, i.e. distance constraints derived from 5C frequencies.

$$\mathcal{E}(\mathbf{x}_n^{(k)}) = \frac{\sum_{(i,j) \in P_{\mathbf{x}_n}} \| x_i - x_j \| - d_{i,j}}{\sum_{(i,j) \in P_{\mathbf{x}_n}} d_{i,j}}, \quad (5)$$

where $P_{\mathbf{x}_n}$ is the list of $i$-$j$ pairs that have distance constraint $d_{ij}$ and $i,j = 1, \ldots, n$. Our objective is to generate a set of conformations satisfying all distance constraints and following target distribution $\pi(\mathbf{x}_n)$,

$$\pi(\mathbf{x}_n) = \exp(-\mathcal{E}(\mathbf{x}_n)) \quad (6)$$

Let $\mathbf{x}_t = (x_1, \ldots, x_t)$ be the vector for the three-dimensional positions of nodes from 1 to $t$. We recursively place node $t$ at position $x_t$ following a trial distribution $g_t(x_t|x_{t-1})$. The trial distribution proposes possible positions with different probabilities for node $t$ to be placed under the condition that positions $x_1, \ldots, x_{t-1}$ for node pairs 1 to $t-1$ are given. The joint trial distribution for a chain with $t$ nodes at position $x_1, \ldots, x_t$ is given by
\[ g_t(x_t) = g_1(x_1)g_2(x_2|x_1) \ldots g_t(x_t|x_{t-1}). \] (7)

Following the principle of importance sampling [5, 6, 7], the design of the trial distribution can accommodate different types of biases, which allows great flexibility for improving sampling efficiency and more accuracy for satisfying the target distribution. However, each final sample of full length chain \( x_n \) needs to be weighted to remove the biases, in a way that the original target distribution \( \pi(x_n) \) can be recovered. Specifically, we assign a weight

\[ w(x_n) = \frac{\pi(x_n)}{g_n(x_n)} \] (8)

to each conformation \( x \), where \( g_n(x) \) is the trial distribution of the full chain. Then the expected mean value of physical properties represented by a function \( h(x_n) \) of conformation \( x_n \) following the target distribution \( \pi(x_n) \) can be estimated by

\[ E_{\pi}(h(x_n)) \simeq \frac{\sum_{k=1}^{m} w(x_n^{(k)}) \cdot h(x_n^{(k)})}{\sum_{k=1}^{m} w(x_n^{(k)})}, \] (9)

where \( k=1, \ldots, m \) is the number of samples in the ensemble. We adopt the Fearnhead et al. framework [8] to generate sample conformations which minimizes the loss introduced in the sampling step when choosing a number of distinct samples from a larger sample set. It helps to maintain the diversity of the samples. The algorithm is described in Algorithm 1, where \( m_t \) is the number of samples we retain in the \( t \)th iteration and \( m_{\text{max}} \) is the maximum value of \( m_t \).

**Trial Distribution.** Priority score \( \beta_t^{(l)} \) in Algorithm 1 is the trial distribution \( g_t(x_t|x_{t-1}) \) for the partial chain \( x_{t-1} \). It biases the chain to grow towards to the regions that will potentially satisfy the target distribution, such that the full chain will eventually satisfy all the distance constrains. A growth function is required because the distance constraints of the future nodes cannot be directly used during the growth of a partial chain. Our priority score consists of three components:
Algorithm 1: Sequential Importance Sampling algorithm for nC-SAC model

1: Set $m_1 = 1$, $w_1^{(1)} = 1.0$ and place the first residue at fixed $x_1^{(1)}$
2: for $t = 2 \to n$ do
3: \hspace{1em} $L_t = 0$
4: \hspace{2em} $// L_t$: number of length $t$ chains that can be obtained from samples obtained at step $t-1$. \hspace{2em} \textbf{for} sample $j = 1 \to m_{t-1}$ do
5: \hspace{3em} Find all of the valid sites $x_{t-1}^{(i,j)}$, $i = 1, \cdots, l_t^{(j)}$ for placing $x_t$ next to partial chain $x_{t-1}^{(j)}$
6: \hspace{3em} $// l_t^{(j)}$: number of available sites to place $x_t$ next to partial chain $x_{t-1}^{(j)}$. \hspace{2em} Generate $l_t^{(j)}$ number of $t$-length chain $\tilde{x}_t^{(L_t+i)} = (x_{t-1}^{(j)}, x_t^{(i,j)})$
7: \hspace{3em} $\tilde{w}_t^{(L_t+i)} = w_{t-1}^{(j)}$
8: \hspace{3em} $//$ Temporary weights for uniform distribution. \hspace{1em} \textbf{if} $L_t \leq m_{\text{max}}$ then
9: \hspace{4em} Let $m_t = L_t$ and \{(x_{t}^{(j)}, w_{t}^{(j)})\}_{j=1}^{m_t} = \{(\tilde{x}_t^{(l)}, \tilde{w}_t^{(l)})\}_{l=1}^{L_t}$
10: \hspace{4em} else \hspace{1em} for $l = 1 \to L_t$ do
11: \hspace{5em} Assign a priority score $\beta_t^{(l)}$ for chain $\tilde{x}_t^{(l)}$ according to the constraints
12: \hspace{5em} \textbf{end for}
13: \hspace{4em} end if
14: \hspace{1em} Draw $r$ from uniform distribution $\mathcal{U}[0, 1)$
15: for $j = 1 \to m_{\text{max}}$ do
16: \hspace{2em} Find integer $J_j$ such that $\sum_{l=1}^{L_t} \min\{c\beta_t^{(l)}, 1\} < r_j \leq \sum_{l=1}^{J_j} \min\{c\beta_t^{(l)}, 1\}$
17: \hspace{2em} Select sample $x_{t}^{(j)} = \tilde{x}_t^{(J_j)}$
18: \hspace{2em} Set weight $w_t^{(j)} = \tilde{w}_t^{(J_j)} \cdot (\gamma_t^{(J_j)}/\beta_t^{(J_j)})$
19: \hspace{2em} \textbf{end for}
20: \hspace{1em} \textbf{end for}
21: for $j = 1 \to m_n$ do
22: \hspace{2em} Calculate importance weight $w(x_n^{(j)}) \propto w_n^{(j)} \pi(x_n^{(j)})$
23: \hspace{2em} \textbf{end for}
24: \textbf{end for}
growth potential from collision constraints, growth potential from distance constraints and growth potential from loop consideration.

The priority score $\beta^{(l)}_t$ for chain $\tilde{x}^{(l)}_t$ is set as

$$\beta^{(l)}_t = \exp \left[ -\frac{\rho_1 f_1(\tilde{x}^{(l)}_t) + \rho_2 f_2(\tilde{x}^{(l)}_t) + \rho_3 f_3(\tilde{x}^{(l)}_t)}{\tau_t} \right]$$

(10)

where $\rho_1, \rho_2,$ and $\rho_3$ are coefficients of the three growth potential functions, $\tau_t$ is a temperature like variable. In this study, we set $\rho_1 = \rho_2 = \rho_3 = \tau_t = 1$.

(1) Growth potential from collision constraint. This potential is designed to maintain the self-avoiding property of a 30 nm chromatin fiber. Let $f_1(x_t)$ be the growth potential from collision constraint,

$$f_1(x_t) = \sum_{B_{\tilde{x}_{t-1}}} h_1(x_t, B_{\tilde{x}_{t-1}}, r_0),$$

(11)

where $h_1$ is the loss function to measure the violation of constraint of $x_t$ with its previous partial chain $B_{\tilde{x}_{t-1}}$.

$$h_1(x_t, B_{\tilde{x}_{t-1}}, r_0) = I(||x_t - \tilde{x}_i|| < r_0), \text{ any } \tilde{x}_i \in B_{\tilde{x}_{t-1}},$$

(12)

where $I(\cdot)$ is an indicator function, such that $h_1(x_t, B_{\tilde{x}_{t-1}}, r_0) = 0$, when $||x_t - \tilde{x}_i|| \geq r_0$, and $h_1(x_t, B_{\tilde{x}_{t-1}}, r_0) = 1$, when $||x_t - \tilde{x}_i|| \leq r_0$, and $r_0 = 30$ nm, adapted from experimentally verified threshold [4].

(2) Growth potential from distance constraints. Given a partial chain $x_{t-1}$, the position of the current node $t$ ($x_t \notin x_{t-1}$) is determined according to the distance constraints derived from 5C interactions. Let $f_2(x_t)$ be the growth potential from distance constraints,

$$f_2(x_t) = h_2(||x_1 - x_t||, \ldots, ||x_K - x_t||, (d_{1,t}, \ldots, d_{K,t})),$$

(13)
where $i_k$ is the $k^{th}$ node that has distance constraint $d_{ik,t}$ with the node $t$ and $K$ is the total number of nodes that have distance constraints $d_{ik,t}$ with the current node $t$ in the partial chain $x_{t-1}$. $h_2$ is the loss function to measure the error between distances between the nodes in the chain and their corresponding distance constraints.

$$h_2((\|x_{i_1} - x_t\|, \ldots, \|x_{i_K} - x_t\|), (d_{i_1,t}, \ldots, d_{i_K,t})) = \frac{\sum_{i_k \in P_t} \|x_{i_k} - x_t\| - d_{i_k,t}}{\sum_{i_k \in P_t} d_{i_k,t}}, \quad (14)$$

(3) Growth potential from loop constraints. Due to the sparseness of 5C interactions, there are several nodes that do not have any distance constraints. For a node $t$ with no distance constraints from 5C data in the partial chain $x_{t-1}$, we employ a distance constraint to enforce node $t$ to follow triangle inequality. Let $f_3(x_t)$ be the potential from loop constraints,

$$f_3(x_t) = h_3(x_t, O_t), \quad (15)$$

where $O_t = \{(t_{ik}, t_{jk}) | \text{interaction pair } t_{ik} \text{ and } t_{jk} \text{ and } t_{ik} < t < t_{jk}\}$, and, $h_3$ is a loss function to measure the triangle inequality,

$$h_3(x_t, O_t) = \sum_{(t_{ik}, t_{jk}) \in O_t} I\left(\|x_t - x_{t_{ik}}\| - d_{t_{ij},t_{ik}} > \sum_{l=t}^{t_{jk}-1} d_{l,l+1}\right), \quad (16)$$

d_{l,l+1} is the length of segment between node $l$ and $l+1$, $l$ from $t$ to $t_{jk} - 1$, and $I(\cdot)$ is an indicator function such that it is equal to 1 when the distance between node $t$ and node $t_{ik}$ is greater than the sum of the rest of the segment length between the node $t$ and node $t_{jk}$, it is equal to 0 otherwise.

**Target distribution.** The target score $\beta_t^{(l)}$ that represents the target distribution for chain $\tilde{x}_t^{(l)}$ is set as

$$\beta_t^{(l)} = \exp\left[-\frac{\rho_1 f_1(\tilde{x}_t^{(l)}) + \rho_2 f_2(\tilde{x}_t^{(l)})}{\tau_t^{(l)}}\right], \quad (17)$$
where $\rho_1, \rho_2$ are coefficients of the two growth potential functions: growth potential from collision constraint and growth potential from distance constraint. $\tau'_t$ is a temperature-like variable, and $\tau'_t = \frac{1}{\tau_t}$.

**Analysis of Ensemble of nC-SAC Chains**

**Root mean square deviation (RMSD) of distance**

\[
\text{RMSD}(c_m, c_n) = \sqrt{\frac{\sum_{k=1}^{l} |c_m^k - c_n^k|^2}{l}},
\]

(18)

where $c_m = \{d(i, j)| (i, j)\}$ is the collection of spatial distances between nodes of $m$th predicted conformation, $l$ is the total number of significant interactions and $c_m^k$ is the $k$th distance element in the set $c_m$.

**Probability $q_{ij}^{\text{pred}}$ of interactions between $i$ and $j$ in prediction model**

\[
q_{ij}^{\text{pred}} = \frac{\sum_k w_k I(i, j)}{\sum_i \sum_j \sum_k w_k I(i, j)},
\]

(19)

where $I(i, j)$ is indicator function. If the distance between nodes $i$ and $j$ in the $k'$th conformation is less than the threshold $d_c$, $I(i, j) = 1$, otherwise 0. $k'$ is range from 1 to 10,000 conformations.

**Propensity $\text{Prop}_{ij}^{\text{pred}}$ of interaction between $i$ and $j$ in prediction model**

\[
\text{Prop}_{ij}^{\text{pred}} = \frac{q_{ij}^{\text{pred}}}{q_{ij}^{R}} = \frac{\sum_k w_k I(i, j)}{\sum_i \sum_j \sum_k w_k I(i, j)}
\]

(20)

where $I(i, j)$ is indicator function, if the distance between nodes $i$ and $j$ in the $k'$th conformation less than the threshold $d_c$, $I(i, j) = 1$, otherwise 0. $k$ is range from 1 to 100,000 conformations for the random model we generated. $k'$ is range from 1 to 10,000 predicted conformations.
Percentage $p_{ij}^{pred}$ of interactions between nodes $i$ and $j$ in the constructed model

$$p_{ij}^{pred} = \frac{\sum_{k'} w_{k'} I(i, j)}{\sum_{k'} w_{k'}}$$

where $p_{ij}^{pred}$ is to measure the fraction of the frequency of the interaction between nodes $i$ and $j$ in the ensemble of predicted conformations.

**Density-based algorithm for clustering**

In this study, we adapt a density-based clustering method to cluster all the predicted conformations according to their similarities for each cell line [9]. We measured the radius of gyration and RMSD (Eq. 18) for each pair of conformations, used them as the similarity measure in the clustering algorithm. The threshold square root difference in the radius of gyration is taken as 34 nm. This choice is arbitrary, but facilitate fast clustering while reflect the appropriate spatial resolution important for biological interactions. Following [9], we set the number of conformations $\text{min confs} = 5$ to determine the membership of a chromatin chain to ensure $\text{min confs} \geq D + 1$, where $D = 3$ for chromatin chains in three-dimensional space. We consider a conformation $i$ belongs to the cluster $c_k$, if $i$ has more than 5 similar conformations in the cluster $c_k$. The threshold RMSD is taken as 34 nm and we consider a conformation $i$ belongs to the cluster $c_k$, if $i$ has more than 5 similar conformations in the cluster $c_k$. 

10
**Chromatin properties.** With $m$ successfully generated chromatin chains (random C-SAC chains or nC-SAC chains), we can calculate the physical properties of the population of chromatin fibers. If the configurations of the $j$-th successfully generated chromatin chain is $x^{(j)} = (x_1^{(j)}, \ldots, x_n^{(j)})$, and its associated weight $w^{(j)}$, then we can calculate the mean value of a physical property $\tilde{h}(x)$ such as the presence of an $(i, j)$ interaction between nodes as:

$$\tilde{h}(x) = \mathbb{E}_{\pi(x)}[h(x)] = \frac{\sum_{j=1}^{m} h(x^{(j)}) \cdot w^{(j)}}{\sum_{j=1}^{m} w^{(j)}}.$$

**Figures**

Supplementary Figure 1: **Interaction profiles of selected nodes in both cell lines and in the random ensemble.** $x$-axis denotes the nodes and $y$-axis is the propensity of interaction between the anchor node and the rest of the locus. The interactions between $\alpha$-globin gene and the nodes 12, 13 and 14 (enhancers HS40/46/48) are highlighted in the dotted box.
Supplementary Figure 2: Conversion of ensemble by the sampled number of chains. (a) The Pearson Correlation between the interaction frequencies of full ensemble and partial ensembles with different sampling rates. (b) The heatmaps of interaction frequencies of a partial ensemble where only 10% of 10,000 chains were sampled and the heatmap of interaction frequencies of the full ensemble.
Supplementary Figure 3: Histogram of number of interactions that overlap with nC-SAC prediction when 89 ChIA-PET functional interactions are shuffled. *p*-value of obtaining 52 or more overlap with nC-SAC predicted interactions is $< 10^{-5}$, when interactions are shuffled for 100,000 times.
Supplementary Figure 4: **Interaction profiles of enhancers, α-globin gene and POL3RK gene in the constructed 3D ensembles.** x-axis denotes the nodes and y-axis is the propensity of interaction between the anchor node and the rest of the locus. (a) The interactions between enhancers and the rest of the locus in the GM12878 cell line, K562 cell line and the difference between these cell lines are depicted, respectively. (b) The interactions between α-globin gene and the rest of the locus in the GM12878 cell line, K562 cell line and the difference between these cell lines are depicted, respectively. (c) The interactions between POL3RK gene and the rest of the locus in the GM12878 cell line, K562 cell line and the difference between these cell lines are depicted, respectively.
Supplementary Figure 5
Supplementary Figure 5 (previous page): Chip-Seq enrichment peaks obtained from ENCODE database [10]. Red bars denote the enrichments in the silent GM12878 cell line and the blue bars denote the enrichments in the active K562 cell line. (a) Transcription factor Pu.1 enrichment peaks per node. Enrichment for transcription factor Pu.1 on node 5 (POL3RK) is highlighted in dashed box, where more than 2 fold increase is measured in the silent GM12878 cell line compared to the active K562 cell line. (b) Transcription factor Sp.1 enrichment peaks per node. Enrichment for transcription factor Sp.1 on the node 5 (POL3RK) is highlighted in dashed box, where no enrichment exists in the active K562 cell line. (c) Histone modification H3kme2 enrichment peaks per node. Enrichments for H3kme2 on the nodes 12/13/14 (HS40/46/48) are highlighted in dashed box, where it only exists in the active K562 cell line. (d) H2A.Z enrichment peaks per node. The nodes (node 5:POL3RK, nodes 12/13/14:HS40/46/48 and node 21:α-globin gene) that effect the expression level α-globin gene are shown in dashed boxes. (e) RNAPII enrichment peaks per node. Enrichments for RNAPII on the nodes 12/13/14 (HS40/46/48) and node 21 (α-globin gene) are highlighted in dashed box, where they exist only in the active K562 cell line.
Supplementary Table 1: The interactions that are captured by ChIA-PET study in K562 cell line. Genomic locations are in basepairs and reference genome is hg18.

| Genomic Location 1 | Genomic Location 2 | Node 1 | Node 2 | Factor     | Status                  |
|--------------------|--------------------|--------|--------|------------|-------------------------|
| 55,200-56,599      | 169,200-171,999    | 7      | 21     | H4K4me3    | New prediction          |
| 55,000-56,999      | 351,800-354,999    | 7      | 39     | H4K4me1    | Not predicted / not in 5C |
| 96,800-98,199      | 124,000-130,599    | 12     | 16     | H4K4me3    | New prediction          |
| 96,600-98,599      | 132,000-135,599    | 12     | 18     | H4K4me2    | New prediction          |
| 96,600-98,599      | 167,200-168,799    | 12     | 21     | H4K427ac   | Predicted / in 5C       |
| 94,200-96,199      | 169,200-172,199    | 12     | 21     | H4K4me1    | Predicted / in 5C       |
| 96,800-98,199      | 169,200-171,999    | 12     | 21     | H4K4me3    | Predicted / in 5C       |
| 94,897-95,586      | 170,079-171,864    | 12     | 21     | PolII      | Predicted / in 5C       |
| 96,798-97,306      | 170,079-171,864    | 12     | 21     | PolII      | Predicted / in 5C       |
| 96,600-98,599      | 169,200-172,199    | 12     | 22     | H4K4me1    | Predicted / in 5C       |
| 96,600-98,599      | 169,200-173,399    | 12     | 22     | H4K427ac   | Predicted / in 5C       |
| 96,600-98,599      | 167,200-172,199    | 12     | 22     | H4K4me2    | Predicted / in 5C       |
| 100,400-101,199    | 167,200-168,799    | 13     | 21     | H4K427ac   | Predicted / in 5C       |
Supplementary Table 2: Details of the predictions and comparison with ChIA-PET RNAPII data in K562 cell line

| Node | New Prediction | Already in 5C Data | No Primer Site | No Record in 5C | 5C Count |
|------|----------------|--------------------|----------------|-----------------|----------|
| 5    | 11             | $                 |                |                 |          |
| 6    | 16             |                    | $              |                 |          |
| 6    | 26             | +                  |                |                 |          |
| 6    | 54             | +                  |                |                 |          |
| 8    | 22             | +                  |                |                 | 0        |
| 8    | 38             | +                  |                |                 | 0        |
| 8    | 40             |                    |                |                 | 33       |
| 8    | 54             |                    | $              |                 | 5        |
| 9    | 27             | +                  |                | #               |          |
| 9    | 40             | +                  |                |                 |          |
| 9    | 54             |                    | $              |                 | 37       |
| 11   | 22             | +                  |                |                 | 18       |
| 12   | 22             |                    |                |                 | 37       |
| 13   | 21             |                    |                |                 | 6        |
| 14   | 22             |                    |                |                 |          |
| 15   | 20             | +                  |                |                 |          |
| 15   | 22             | +                  |                |                 |          |
| 16   | 18             |                    | $              |                 |          |
| 17   | 24             | +                  |                |                 |          |
| 18   | 39             |                   | $              |                 |          |
| 20   | 22             |                    | $              |                 |          |
| 21   | 32             |                    | $              |                 |          |
| 21   | 36             | +                  |                |                 |          |
| 22   | 24             |                    |                |                 |          |
| 22   | 27             |                    |                |                 |          |
| 22   | 54             |                    | $              |                 |          |
| 23   | 26             | +                  |                |                 |          |
| 26   | 28             | +                  |                |                 |          |
| 26   | 31             |                    | $              |                 |          |
| 26   | 38             | +                  |                |                 |          |
| 26   | 39             |                    | $              |                 |          |
| 27   | 36             | +                  |                |                 |          |
| 27   | 39             |                    | $              |                 |          |
| 28   | 31             | +                  |                |                 |          |
| 29   | 31             | +                  |                |                 |          |
| 30   | 39             |                    | $              |                 |          |
| 31   | 34             | +                  | 18             |                 |          |
| 31   | 38             |                    | $              |                 |          |
| 31   | 39             |                    | $              |                 |          |
Supplementary Table 3: Details of the predictions and comparison with ChIA-PET RNAPII data

| Node | Node | New Prediction | Already in 5C Data | No Primer Site | No Record in 5C | 5C Count |
|------|------|----------------|-------------------|----------------|-----------------|----------|
| 34   | 38   | +              |                   |                |                 |          |
| 35   | 37   |                |                   |                |                 |          |
| 35   | 39   |                |                   | $              |                 |          |
| 35   | 44   |                |                   | $              |                 |          |
| 36   | 39   |                |                   | $              |                 |          |
| 37   | 41   |                |                   | $              |                 | 27       |
| 38   | 40   |                |                   | $              |                 |          |
| 38   | 44   |                |                   | $              |                 |          |
| 39   | 42   |                |                   | $              |                 |          |
| 39   | 44   |                |                   | $              |                 |          |
| 39   | 54   |                |                   | $              |                 |          |
| 40   | 43   |                |                   | $              |                 | 23       |
| 40   | 54   |                |                   | $              |                 |          |
| 41   | 44   |                |                   | $              |                 |          |
| 41   | 54   |                |                   | $              |                 | 2        |
| 42   | 52   |                |                   | +              |                 |          |
| 42   | 54   |                |                   | +              |                 |          |
| 43   | 54   |                |                   | -              |                 |          |
| 44   | 50   |                |                   | $              |                 |          |
| 44   | 52   |                |                   | $              |                 |          |
| 44   | 54   |                |                   | $              |                 |          |
| 46   | 51   |                |                   | $              |                 |          |
| 46   | 54   |                |                   | +              |                 |          |
| 49   | 54   |                |                   | +              |                 |          |
| 50   | 54   |                |                   | +              |                 |          |
### Supplementary Table 4: Details of the predictions and comparison with ChIA-PET CTCF data in K562 cell line

| Node | Node | New Prediction | Already in 5C Data | No Primer Site | No Record in 5C | 5C Count |
|------|------|----------------|--------------------|----------------|-----------------|----------|
| 8    | 40   | +              |                    |                |                 | 0        |
| 8    | 11   | +              |                    |                |                 | #        |
| 8    | 22   | +              |                    |                |                 | $        |
| 9    | 11   | +              |                    |                |                 | #        |
| 9    | 22   |                |                    |                |                 | $        |
| 11   | 40   |                |                    |                |                 | #        |
| 11   | 22   | +              |                    |                |                 | #        |
| 12   | 22   |                |                    |                |                 | $        |
| 13   | 22   | -              |                    |                |                 | $        |
| 15   | 22   | +              |                    |                |                 | #        |
| 43   | 52   | +              |                    |                |                 | #        |

### Supplementary Table 5: Details of the predictions and comparison with ChIA-PET RAD21 data in K562 cell line

| Node | Node | New Prediction | Already in 5C Data | No Primer Site | No Record in 5C | 5C Count |
|------|------|----------------|--------------------|----------------|-----------------|----------|
| 7    | 21   | +              |                    |                |                 | 0        |
| 7    | 39   |                |                    |                |                 |          |
| 11   | 21   | +              |                    |                |                 |          |
| 12   | 21   |                |                    |                |                 |          |
| 26   | 38   |                |                    |                |                 |          |
| 33   | 38   | +              |                    |                |                 | $        |
| 43   | 51   |                |                    |                |                 | $        |
| 43   | 52   | +              |                    |                |                 | #        |
Supplementary Table 6: Details of the predictions and comparison with ChIA-PET RAD21 data in GM12878 cell line

| Node | Node | New Prediction | Already in 5C Data | No Primer Site | No Record in 5C | 5C Count |
|------|------|----------------|--------------------|----------------|-----------------|---------|
| 7    | 10   | +              |                    |                |                 | 0       |
| 12   | 21   |                | -                  |                |                 |         |
| 14   | 21   |                | -                  |                |                 |         |
| 26   | 39   |                |                    |                |                 | #       |

Supplementary Table 7: The average spatial distances between nodes in the chains with three-way interaction in GM12878 cell line and in the chains without three-way interaction in K562 cell line

| Node(s)          | Node                          | Distance in GM12878 (nm) | Distance in K562 (nm) |
|------------------|-------------------------------|--------------------------|-----------------------|
| 12/13/14 (HS40/46/48) | 21 (α-globin gene)            | 80±5                     | 74.9±5.1              |
| 12/13/14 (HS40/46/48) | 5 (POL3RK)                    | 50.1±20.0                | 134.6±30.2            |
| 21 (α-globin gene)   | 5 (POL3RK)                    | 62.4±18.5                | 140.1±28.2            |
References

[1] D. et al. Baù. The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nature structural & molecular biology*, 18(1):107–114, 2010.

[2] Xu Y. Kenter A.L. Gürsoy, G. and J. Liang. Spatial confinement is a major determinant of folding landscape of human chromosomes and genetic programming of cell. *Nucleic Acids Research*, 42(13):8223–8230, 2014.

[3] B. Yoav and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 57(1):289–300, 1995.

[4] Nora EP Piolot T Lam F Dekker J Tiana G Heard E. Giorgetti L, Galupa R. Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell*, 157(4):950–963, 2014.

[5] A.W. Marshall. The use of multistage sampling schemes in monte carlo computations. In *Symposium on Monte Carlo methods, University of Florida, 1954*, pages 123–140. John Wiley and Sons, Inc., New York, 1956.

[6] Zhang J. Liang, J. and R. Chen. Statistical geometry of packing defects of lattice chain polymer from enumeration and sequential monte carlo method. *The Journal of Chemical Physics*, 117:3511–3521, 2002.

[7] J.S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, 2008.

[8] P. Fearnhead and P. Clifford. On-line inference for hidden markov models via particle filters. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 65(4):887–899, 2003.
[9] Kriegel H-P Sander J. Ester, M. and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining*, pages 226–231. Kdd, 1996.

[10] Long J Kuchenbaecker KB Droit A Michailidou K Ghoussaini M Kar S Freeman A Hopper JL Milne RL Bolla MK Wang Q Dennis J Agata S Ahmed S Aittomki K Andrulis IL Anton-Culver H Antonenkova NN Arason A Arndt V Arun BK Arver B Bacot F Barrowdale D Baynes C Beeghly-Fadel A Benitez J Bermisheva M Blomqvist C Blot WJ Bogdanova NV Bojesen SE Bonanni B Borresen-Dale AL Brand JS Brauch H Brennan P Brenner H Broeks A Brning T Burwinkel B Buys SS Cai Q Caldes T Campbell I Carpenter J Chang-Claude J Choi JY Claes KB Clarke C Cox A Cross SS Czene K Daly MB de la Hoya M De Leeneer K Devilee P Diez O Domchek SM Doody M Dorfling CM Drk T Dos-Santos-Silva I Dumont M Dwek M Dworniczak B Egan K Eilber U Einbeigi Z Ejlertsen B Ellis S Frost D Laloo F; EMBRACE Fasching PA Figueroa J Flyger H Friedlander M Friedman G Gambino G Gao YT Garber J Garca-Closas M Gehrig A Damiola F Lesueur F Mazoyer S Stoppa-Lyonnet D; behalf of GEMO Study Collaborators Giles GG Godwin AK Goldgar DE Gonzalez-Neira A Greene MH Gunel P Haeberle L Haiman CA Hallberg E Hamann U Hansen TV Hart S Hartikainen JM Hartman M Hassan N Healey S Hogervorst FB Verhoef S; HEBON Hendricks CB Hillemanns P Hollestelle A Hulick PJ Hunter DJ Imyanitov EN Isaacs C Ito H Jakubowska A Janavicius R Jaworska-Bieniek K Jensen UB John EM Joly Beauparlant C et. al. Zeng C, Guo X. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012.