SUPPORTING INFORMATION

SI Appendix for:

Population–based 3D genome structure analysis reveals driving forces in spatial genome organization

Harianto Tjong[1,#], Wenyuan Li[1,#], Reza Kalhor[1], Chao Dai[1], Shengli Hao[1], Ke Gong[1], Yonggang Zhou[1], Haochen Li[1], Xianghong Jasmine Zhou[1], Mark A. Le Gros[3,4,5], Carolyn A. Larabell[3,4,5], Lin Chen[1,2], and Frank Alber[1,*]

[1]Molecular and Computational Biology, Department of Biological Sciences, University of Southern California, 1050 Childs Way, Los Angeles, CA 90089, USA

[2]Department of Chemistry and Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, CA 90089, USA

[3]Department of Anatomy, University of California, San Francisco, CA 94148 USA

[4]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

[5]National Center for X-ray Tomography, Advanced Light Source, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

*Correspondence should be addressed to F.A. (alber@usc.edu).

#These authors contributed equally.

TABLE OF CONTENTS

A. SI MATERIALS AND METHODS

A.1. Population-based 3D genome modeling approach

Chromosomes are segmented into chromatin domains, which are represented by spherical volumes following our previously published approach (1). The population-based structural modeling approach is a probabilistic framework to generate a large number of 3D genome structures (i.e. the structure population) whose chromatin domain contacts are statistically consistent with the input experimental TCC data. Our structure population represents a deconvolution of the ensemble-averaged TCC data into a population of individual structures and represents the most likely approximation of the true structure population given all the available data. We formulated the structure optimization problem as a maximum likelihood estimation problem and designed an iterative optimization algorithm with a series of optimization strategies for efficient and scalable model estimation.

A.1.1. Defining sphere volumes

We define two types of sphere radius for each domain, the hard- and soft-core radii. The excluded volume (or hard core) radius of domain *I*, R_I^x, is proportional to the cubic root of the DNA sequence length it contains, l_I, and can be approximated as

$$R_I^x = \rho l_I^{1/3}$$

where

$$\rho^3 = \frac{O_{\text{nuc}} R_{\text{nuc}}^3}{2\sum_{I=1}^{N} l_I}$$

ρ is a coefficient that is adjusted to reproduce the nuclear volume occupancy O_{nuc}, which is the fraction of the nuclear volume with radius R_{nuc} occupied by the genome. Published data rank the level of nuclear occupancy between 10-40% (2). Thus we have chosen to model the human genome using spheres that occupy a total volume about 20% of the nuclear volume. Small variations of the volume occupancy will not change the resulting conclusions. **Table S1** contains detail information of all domain spheres. As described below spheres defined by their excluded volume radius cannot penetrate each other due to an excluded volume constraint, which ensures the minimal occupancy of the chromatin in the nucleus. To allow for interactions of chromatin regions we also define a contact radius of a domain *I*, R_I^c, which is 2 times the hard-core radius, $R_I^c = 2R_I^x$. This tolerance allows for the possibility that chromatin regions can partially loop out of their bulk domain regions to form contacts.

A.1.2. Model of sphere contacts in 3D structure

We assume a pair of spheres (i, j) have a contact in a structure m, if and only if their surface distance $d_{ijm} = \left\|\vec{x}_{im} - \vec{x}_{jm}\right\|_2 - R_i^c - R_j^c$ is equal to or smaller than zero, i.e., $d_{ijm} \le 0$, where $\vec{x}_{im}$ and R_i^c ($\vec{x}_{jm}$ and R_j^c) are center coordinates in a structure m and contact radius of sphere i (j) respectively. The surface distance d_{ijm} could allow a "soft core" overlap between two spheres and the maximal overlap cannot violate the excluded volume constraint, i.e., $d_{ijm} \ge -R_i^x - R_j^x$, where R_i^x and R_j^x are the excluded volume radius of sphere i and j respectively.

We model a sphere contact by a function of the sphere-sphere surface distance, which is essentially a mixture of a constant function and one-sided truncated Gaussian function. This function implies two features: (i) the contact does concretely take place when the surface distance between two domain spheres d_{ijm} is equal to or smaller than 0, which is modeled as a constant function; (ii) the function sharply decreases with the sphere-sphere surface distance greater than 0, which is modeled as a one-sided truncated Gaussian function. Therefore, we have defined

$$P(w_{ijm} = 1 \mid d_{ijm}) = \begin{cases} 1, & d_{ijm} \le 0 \\ \exp\left(\dfrac{d_{ijm}^2}{2\sigma^2}\right), & d_{ijm} > 0 \end{cases} \qquad [1]$$

where w_{ijm} is the latent variable (it will be introduced in the later section) indicating whether or not two spheres i and j have an assigned contact (w_{ijm}=1 represents sphere i and j have a contact in structure m, otherwise w_{ijm} =0). **Fig. S1A** shows the curve of the sphere contact function. For those sphere pairs that do not have an assigned contact (i.e. w_{ijm} =0), we have $P(w_{ijm} = 0 \mid d_{ijm}) = 1 - P(w_{ijm} = 1 \mid d_{ijm})$ (**Fig. S1B**). Please note that (i) d_{ijm} is the surface distance between two spheres and therefore is independent of sphere sizes; and (ii) the case

$d_{ijm} < -R_i^x - R_j^x$ is not allowed, because we require the "excluded volume" constraint to be satisfied in our model (shall be introduced in a later section).

Using above functions, we have $P(w_{ijm} \middle| \vec{x}_{im}, \vec{x}_{jm})$ of two spheres i and j in a structure as below:

$$P(w_{ijm} \mid \vec{x}_{im}, \vec{x}_{jm}) = P(w_{ijm} = 1 \mid d_{ijm})^{w_{ijm}} P(w_{ijm} = 0 \mid d_{ijm})^{1-w_{ijm}} \qquad [2]$$

A.1.3. Problem formulation

The chromosome conformation capture data is processed to be a contact probability matrix $\mathbf{A} = (a_{IJ})_{N\times N}$ of N domains in the genome, where $0 \le a_{IJ} \le 1$ is the contact probability of two chromosome domains I and J (will be described in section A.3.5). Note that in the human diploid genome, each domain has two homologous copies. Upper case letters (e.g., I or J) are used to denote a domain (as a chromosome region), and lower cases are used when we distinguish between the homolog copies of the domain in the diploid genome (i and i' for I, j and j' for J).

Our model, the structure population, is defined as a set of M diploid genome structures $\mathbf{X}=\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_M\}$, where the m-th structure $\mathbf{X}_m$ is a set of 3-dimensional vectors representing the center coordinates of $2N$ spheres $\mathbf{X}_m = \left\{\vec{x}_{im} : \vec{x}_{im} \in \Re^3, \quad i = 1,\ 2\ \ldots,\ 2N\right\}$.

For reconstructing a structure population $\mathbf{X}$, in principle we need the detailed information about which domain pairs are in contact in which structure of the population. The domain contact probability matrix ($\mathbf{A}$) derived from TCC data is incomplete and does not provide this information, because (i) the experimental data provides contact frequencies averaged over a population of cells without contact information at a single cell level, and (ii) the data does not distinguish between contacts from two homologous chromosome copies.

As aforementioned, we have introduced a latent variable, the "*contact indicator tensor*" $\mathbf{W} = (w_{ijm})_{2N\times 2N\times M}$, for complementing every single cell's contact information (**Fig. 1A** in the main paper). It is a binary-valued 3rd-order tensor specifying the contacts of chromatin domains for each homologous copy in each structure of the population: i.e., $w_{ijm} = 1$ indicates that the contact between domains i and j in structure m; $w_{ijm} = 0$ otherwise.

Given $\mathbf{A}=\left(a_{IJ}\right)_{N\times N}$, we aim to estimate the structure population model $\mathbf{X}$ such that the likelihood $P\left(\mathbf{A},\mathbf{W}\middle|\mathbf{X}\right)$ is maximized. The dependence relationships between these variables in an optimized structure population is: $\mathbf{X}\rightarrow\mathbf{W}\rightarrow\mathbf{A}$, because $\mathbf{W}$ is a detailed expansion of $\mathbf{A}$ at the diploid representation and single cell level and $\mathbf{X}$ is the structure population that is consistent to $\mathbf{W}$. Therefore, the likelihood $P\left(\mathbf{A},\mathbf{W}\middle|\mathbf{X}\right)$ can be expanded to $P\left(\mathbf{A}\middle|\mathbf{W}\right)P\left(\mathbf{W}\middle|\mathbf{X}\right)$.

According to the model of sphere contacts described in the previous section, $P\left(\mathbf{W}\middle|\mathbf{X}\right)$ can be expanded as $P\left(\mathbf{W}\middle|\mathbf{X}\right)=\prod_{m=1}^{M}\prod_{\substack{i,j=1\\ i\neq j}}^{2N}P\left(w_{ijm}\middle|\vec{x}_{im},\vec{x}_{jm}\right)$. Also, $P\left(\mathbf{A}\middle|\mathbf{W}\right)$ can be expanded as $P\left(\mathbf{A}\middle|\mathbf{W}\right)=\prod_{I,J}P(a_{IJ}\mid a'_{IJ})$, where a'_{IJ} is the contact probability of the domain pair I or J computed from $\mathbf{W}$. We then model each a_{IJ} as $a_{IJ}=a'_{IJ}+\varepsilon_{IJ}$, where ε_{IJ} are independent and identical normally distributed random variables with mean zero $\varepsilon_{IJ}\propto N(0,\sigma'^{2})$ (ε_{IJ} is effectively set to 0). a'_{IJ} is calculated as

$$a'_{IJ}=\frac{1}{2M}\sum_{m=1}^{M}\bar{w}_{IJm} \qquad [3]$$

where $\bar{\mathrm{W}}=(\bar{w}_{IJm})_{N\times N\times M}$ is the "projected contact tensor" (**Fig. 1A** in the main paper), which is derived from $\mathbf{W}$ by projecting its representation (with 2N homologous domains) to its counterpart without homologous domain distinction (with N domains) for domain pair I and J and is defined as below. For instance, in the projected tensor $\bar{\mathrm{W}}$, each element $\bar{w}_{IJm}=1$ indicates that any one of two homologues copies of two domains I and J have a contact in structure m, $\bar{w}_{IJm}=2$ indicates that two out of 4 possible pairs made by homologues copies of two domains I and J have contacts in structure m.

With these probabilistic models, we can maximize the log-likelihood below,

$$\log P(\mathbf{A},\mathbf{W}|\mathbf{X})=\log P\left(\mathbf{A}\middle|\mathbf{W}\right)+\log P\left(\mathbf{W}\middle|\mathbf{X}\right)=\sum_{\substack{I,J=1\\ I\neq J}}^{N}\log P(a_{IJ}\mid a'_{IJ})+\sum_{m=1}^{M}\sum_{\substack{i,j=1\\ i\neq j}}^{2N}\log P\left(w_{ijm}\middle|\vec{x}_{im},\vec{x}_{jm}\right) \qquad [4]$$

This is essentially a maximum likelihood estimation problem.

Additional constraints. In addition to the TCC data, we also include a set of spatial constraints based on additional information about the genome organization. These data are included in form of general spatial constraints acting on the 2N domain spheres: (i) a nuclear volume restraint that forces all spheres to lie inside the nuclear volume, i.e. $\left\|\vec{x}_{im}\right\|_2 < R_{\text{nuc}}$ (where R_{nuc} is the nuclear radius); (ii) excluded volume restraints that prevent the "hard core" overlap between any 2 spheres i and j, i.e., $d_{ijm} \geq -R_i^x - R_j^x$; (iii) information from 3D FISH experiment, which showed that the q-arm of chromosome 4 is tethered to the nuclear envelope (NE) (3). Accordingly we add a constraint to for the q-arm telomere domain ($\vec{x}_{\text{4qtel}}$) of chromosome 4 to be located close to the nuclear envelope ($\left\|\vec{x}_{\text{4qtel}}\right\|_2 > 0.75 R_{\text{nuc}}$). Note that, without loosing generalization, we use the origin (0,0,0) as the nuclear center, thus $\left\|\vec{x}\right\|_2$ is equivalent to the distance from the nuclear center. In summary, the maximum likelihood problem is formally expressed as follows,

$$\hat{\mathbf{X}} = \arg\max_{\mathbf{X}} \max_{\mathbf{W}} \left\{ \log \mathrm{P}(\mathbf{A}, \mathbf{W} \mid \mathbf{X}) \right\}$$

$$\text{subject to} \begin{cases} \text{spatial constraint I: nuclear volume} \\ \text{spatial constraint II: excluded volume} \\ \text{spatial constraint III: 4qtel-NE proximity} \end{cases} \quad [5]$$

Note that, in principal we could add more knowledge-based constraints into this formulation.

A.1.4. Optimization procedure

We designed an iterative optimization procedure to solve this maximum likelihood estimation problem. Since our problem does not have a closed-form solution, numerical routines and heuristic strategies are needed to efficiently approximate the solution. This is an efficient iterative solver to alternately optimize $\mathbf{W}$ and $\mathbf{X}$ while holding the other fixed. We refer to this iterative cycle as the *A/M* (*Assignment/Modeling*) steps (**Fig. 1**) and this procedure as the *A/M* algorithm, which are described as follows.

- Initialization step: an initial model estimate $\mathbf{X}^{(0)}$ is needed to start the iterative procedure. We can generate $\mathbf{X}^{(0)}$ using random domain positions, which satisfy three spatial constraints in Eq. [5] (**Fig. 1B**).
- Assignment step (*A*-step): Given the current estimated model $\mathbf{X}^{(k)}$, estimate the latent variable $\mathbf{W}$ by maximizing the log-likelihood over all possible values of $\mathbf{W}$.

$$\mathbf{W}^{(k+1)} = \arg\max_{\mathbf{W}} \left\{ \log P\left(\mathbf{A}, \mathbf{W} \middle| \mathbf{X}\right) \right\}, \quad \text{given } \mathbf{X} = \mathbf{X}^{(k)} \quad [6]$$

- Modeling step (*M*-step): Given the current estimated latent variable $\mathbf{W}^{(k+1)}$, find the model $\mathbf{X}^{(k+1)}$ that maximizes the log-likelihood of the data $\mathbf{A}$. A new structure population will be generated in which all assigned contacts in $\mathbf{W}$ will be physically present in the structure population $\mathbf{X}$.

$$\mathbf{X}^{(k+1)} = \arg\max_{\mathbf{X}} \left\{ \log P\left(\mathbf{A},\mathbf{W}|\mathbf{X}\right) \right\}, \text{ given } \mathbf{W} = \mathbf{W}^{(k+1)} \quad [7]$$

- Iterative *A/M* steps until convergence (detailed convergence criterion will be covered in section A.1.7).

We exploited the parallelism and algorithmic heuristics underlying the *A/M* steps, which can largely speed up the procedure and make the implementation scalable for the large-scale TCC data. In the next sections, we present the procedure in detail.

A.1.5. Step-wise optimization strategy

We developed a stepwise optimization strategy for the structure optimization process, based on the following considerations: (i) an initial model that already fits a portion of domain contacts in $\mathbf{A}$ can guide a more efficient search of the optimum $\mathbf{W}$ than a random structure; (ii) gradually fitting an increasing number of domain contacts (from the highest to the lowest contact probabilities $\mathbf{A}$) can effectively guide the search to the best solution. The idea of this strategy is to gradually allocate the contacts in $\mathbf{A}$ by using the optimized structure populations $\mathbf{X}$ from the previous steps to determine the contact tensor $\mathbf{W}$ for the following steps. We start the first optimization step by using only the most frequent contacts $\mathbf{A}^{\theta_1}$ (using only $a_{IJ} \geq \theta_1$ and θ_1=1.0) as input to obtain $\hat{\mathbf{X}}^{\theta_1}$, which reproduces $\mathbf{A}^{\theta_1}$ (i.e., the structure population contains all physical domain contacts according to the experimental contact probability). Then $\hat{\mathbf{X}}^{\theta_1}$ is used as the initial model of the next round of optimization for $\mathbf{A}^{\theta_2}$ which includes all domain contacts with lower contact probabilities (i.e., using only $a_{IJ} \geq \theta_2$ and $\theta_2 < \theta_1$). This in turn leads to the refined structured population $\hat{\mathbf{X}}^{\theta_2}$, which serves as input for the next step, and so on. In this way, the contacts in $\mathbf{A}$ are gradually allocated to the optimized structure population $\mathbf{X}$ and contact tensor $\mathbf{W}$. When θ is close to zero, $\mathbf{X}^{\theta}$ reproduces most elements of $\mathbf{A}$ and represents the best approximation of the true structure population given the available data. Our experience indicates that when a population starts to accumulate restraints violations then we have achieved the lowest θ we can pick. In this work, with θ below 0.01 we observed that the population started to have restraint violations, thus we decided the final θ was 0.01 where all the given restraints were still satisfied.

The detailed steps of the procedure are as follows (illustrated in **Fig. 1B** of the main paper).

1. Organize a list of contact probability thresholds in decreasing order of contact probability, $\Theta = \{\theta_1, \theta_2, \theta_3, ..., \theta_{\text{final}}\}$, e.g., $\Theta = \{1, 0, 0.8, 0.6, ..., 0.01\}$.
2. For each $\theta \in \Theta$ in decreasing order, we apply the following steps:
   1) From the TCC data $\mathbf{A}$, we generate a truncated data $\mathbf{A}^{\theta} = \left(a_{IJ}^{\theta}\right)_{N \times N}$, where

      $$a_{IJ}^{\theta} = \begin{cases} a_{IJ} & \text{if } a_{IJ} \geq \theta \\ 0 & \text{otherwise} \end{cases}$$

   2) Using $\mathbf{A}^{\theta}$ as input to perform the iterative *A/M* optimization algorithm and generate the solution $\hat{\mathbf{X}}^{\theta} = \arg \max_{\mathbf{X}} \max_{\mathbf{W}} \left\{ \log \mathrm{P}(\mathbf{A}^{\theta}, \mathbf{W} \mid \mathbf{X}) \right\}$.
   3) Only if the optimization in step 2 succeeds and all assigned contacts in $\mathbf{W}$ are physically present in $\mathbf{X}$, we move to the next round with a new level of θ. If it fails, we retry step 2 until successful.

The probability of observing a given contact in a structure depends on the presence of contacts in the same structure. For example, a certain chromosome contact also brings other chromosome regions into spatial proximity to each other, which in turn enhances their chances of contacting each other in the same structure rather than in a structure where the corresponding domains are far apart from each other and cannot be brought into spatial proximity. Our step-wise optimization approach naturally considers such cooperativity between domain contacts in individual structures. We assume that the more steps could lead to better "cooperativity" effects come into play. However, to speed up the whole process, we usually have several theta values. One could adopt the following recipe. After θ_1 =1 is done, find θ_2 so that $\mathbf{A}^{\theta_2}$ contains pairwise contacts as many as roughly 3 times the number of domain (3N). The next ones can be 10N, 15N, etc. number of contacts, until the population is hard to optimize. We try to keep minimal number of tunable parameters. Our experience indicates that our recipe and parameters are applicable to different data set; also our structure populations are not sensitive to the different parameter sets.

A.1.6. A-step: Parallel and efficient heuristic optimization for the contact assignment step

The *A*-step optimization problem is to "find the contact indicator tensor $\mathbf{W}$ whose derived contact probability a'_{IJ} best matches the observed a_{IJ} for every domain pair I and J". Equation [4] can be expanded as

$$\log P\left(\mathbf{A},\mathbf{W}\middle|\mathbf{X}\right)=\sum_{\substack{I,J=1\\ I\neq J}}^{N}\log P(a_{IJ}\mid a'_{IJ})+\sum_{m=1}^{M}\sum_{\substack{i,j=1\\ i\neq j}}^{2N}\left[w_{ijm}\log P(w_{ijm}=1\middle|d_{ijm})+(1-w_{ijm})\log P(w_{ijm}=0\middle|d_{ijm})\right] \quad [8]$$

To estimate $\mathbf{W}$ for a given structure population $\mathbf{X}$, a natural and intuitive strategy of maximizing Eq. [8] is to assign $w_{ijm}=1$ for which the corresponding domains are already closest in 3D space and therefore have the highest likelihood of forming a contact. That is, assignments of a given chromatin contact across the contact indicator tensor $\mathbf{W}$ are more likely realized in those genome structures in which the corresponding chromatin domains are already closer in 3D space. For a given pair of domains i and j, we utilize two mutual exclusive items (i.e. $P(w_{ijm}=1|d_{ijm})$ and $P(w_{ijm}=0|d_{ijm})$ in Eq. [8]) based on their 3D surface distance in each of the M structures: when d_{ijm} is larger than a distance threshold value (termed as d_{IJ}^{act}), let $w_{ijm}=0$ for accepting the larger log-likelihood $\log P(w_{ijm}=0|d_{ijm})$ out of two mutual exclusive items; when d_{ijm} is smaller than a distance threshold d_{IJ}^{act}, let $w_{ijm}=1$ for accepting the larger log-likelihood $\log P(w_{ijm}=1|d_{ijm})$. When I and J are domains from the same chromosome we select pairs from each homologue copy, namely {(*i*,*j*), (*i'*,*j'*)}. When I and J are domains from the different chromosomes, we select 2 pairs whose distances are the smallest two among { $d_{ijm},d_{i'jm},d_{ij'm},d_{i'j'm}$} computed from $\mathbf{X}_m$. This process is easily implemented in parallel, because the distance threshold of each domain pair can be independently calculated.

To define the distance threshold d_{IJ}^{act}, we designed a heuristic optimization procedure (i.e., distance threshold method), which is a process of determining the distance threshold d_{IJ}^{act} for each domain pair (*I*,*J*), based on the empirical distribution of all distances between their homologous copies across all structures of the population (the pairs are selected based on their distances with a procedure described below).

Distance threshold method:

Let (I, J) be a domain pair (with homologues domain copies i, i' and j, j') and $a_{IJ} > 0$:

1) The empirical distribution of domain distances between homologous copies of the domain pair (I, J) is constructed as follows. When I and J are domains from the same chromosome, we collect distances d_{ijm} and $d_{i'j'm}$ in all models (m=1, 2, …, M) which form a total set of 2M distances. When I and J are domains from different chromosomes, we collect the smallest 2 distances from the set of all possible distances $\{d_{ijm}, d_{i'jm}, d_{ij'm}, d_{i'j'm}\}$ for a total set of 2M distances.
2) The 2M distances are ranked in increasing order, and the distance threshold, d_{IJ}^{act}, is determined as the distance value at the $2a_{IJ}M$ th-quantile of all the 2M sorted distances. An illustration is shown in **Fig. S1C**.

This procedure maximizes $\log P(\mathbf{A}, \mathbf{W}|\mathbf{X})$ which have two items $\log P(\mathbf{W}|\mathbf{X})$ and $\log P(\mathbf{A}|\mathbf{W})$, because (i) it assigns contacts to those domain pairs with shortest distances, which maximizes $\log P(\mathbf{W}|\mathbf{X})$ and (ii) it uses the $2a_{IJ}M$ th-quantile of all 2M distances as the distance threshold to determine w_{ijm}, so this heuristically maximizes the first term $\log P(\mathbf{A}|\mathbf{W}) = \sum_{\substack{I,J=1 \\ I \neq J}}^{N} \log P(a_{IJ} \mid a'_{IJ})$ by making a_{IJ} exactly equal to a'_{IJ}.

Please note that in practice the predefined parameters σ and σ' in the formulation do not affect the results, if almost all probability items in the objective function are fully maximized to their extreme values (i.e., ones), which is required by our practical optimization heuristics and implementation.

A.1.7. Parallel and efficient numerical approximation for the modeling step

Given the current estimated contacts of $\mathbf{W}$, the *M*-step reconstructs the structure population $\mathbf{X}$ that matches $\mathbf{W}$. Because $\mathbf{A}$ and $\mathbf{W}$ are known in the *M*-step, the maximization problem in Eq. [4] can be reduced to $\max \log P(\mathbf{W}|\mathbf{X})$, which can be further decomposed to the sub-problem $\max \log P(\mathbf{W}_m|\mathbf{X}_m)$ for every structure m in the population, where

$P\left(\mathbf{W}_m \middle| \mathbf{X}_m\right) = \prod_{i,j} P\left(w_{ijm} \middle| \vec{x}_{im}, \vec{x}_{jm}\right)$ and $\mathbf{W}_m$ is the contact indicator matrix of structure m. Therefore, each individual structure can be independently optimized in parallel. To efficiently optimize an individual structure, we employed simulated annealing dynamics and conjugate gradient optimizations. The former is a structure modeling approach that can efficiently arrive at a stable state minimizing constraints violations; while the latter can adjust local structures in order to reach the optimum with zero constraint violations. Both are implemented in the Integrated Modeling Platform (IMP, http://www.integrativemodeling.org/)(4, 5). **Table S2** lists the parameters used in the modeling step.

According to Eq. [2], the item $\log P\left(w_{ijm} \middle| \vec{x}_{im}, \vec{x}_{jm}\right)$ in Eq. [4] can be expanded as the summation of two mutually exclusive items $w_{ijm} \log P(w_{ijm} = 1 \,|\, d_{ijm})$ and $(1 - w_{ijm}) \log P(w_{ijm} = 0 \,|\, d_{ijm})$ which describe the contact and non-contact of two spheres i and j in structure m, respectively. Our practical experiences showed that when we maximize only the contact items in the objective function, the majority of non-contact items are approximately maximized as well, due to their inherently dependent relationships. In all the 10,000 optimized structures that completely satisfied all contact items, on average (97.11 ± 0.18)% of non-contact items are also satisfied. We therefore maximize only the contact items and ignore the non-contact items for dramatically speeding up the optimization process while keeping reasonably good optimization performance. Please note that for each structure m, the *M*-step checks whether or not all contacts in $\mathbf{W}_m$ are physically present in $\mathbf{X}_m$. If not, the *M*-step will be repeatedly performed until the check is passed.

Convergence criteria. The *A/M* optimization steps are iteratively performed until each contact in $\mathbf{W}$ is physically present in $\mathbf{X}$ and the following convergence criteria is satisfied:

$$\left| a_{IJ} - \frac{1}{2M} \sum_{m=1}^{M} \bar{w}^*_{IJm} \right| \text{ ~0 for every domain pair } I \text{ and } J,$$

where a_{IJ} is the matrix element of $\mathbf{A}$, and $\mathbf{W}^* = (w^*_{ijm})_{2N \times 2N \times M}$ is defined as $w^*_{ijm} = c_{ijm} w_{ijm}$ for any $1 \le i, j \le 2N$ and any $1 \le m \le M$. $\mathbf{C} = (c_{ijm})_{2N \times 2N \times M}$ denotes the full contact tensor derived from $\mathbf{X}$, and $\bar{\mathbf{W}}^* = (\bar{w}^*_{IJm})_{N \times N \times M}$ is the projected contact tensor of $\mathbf{W}^*$ where contacts at the homologous domain level are projected to the domain level.

A.1.8. Software availability

The population-based genome modeling software and the input data used in this work are available upon request.

A.2. Tethered Conformation Capture (TCC)

The previously described HindIII-TCC library of GM12878 cells was further sequenced for this study and the new sequencing results were combined with the previously sequenced data (1) (**Table S3**). The TCC experimental procedure were described elsewhere (1). Briefly, approximately 25 million GM12878 cells were crosslinked with 1% formaldehyde, lysed and treated with Iodoacetyl-PEG2-Biotin to biotinylate chromatin. This biotinylated chromatin was then digested with HindIII and immobilized on MyOne Streptavidin T1 beads (Invitrogen) with about 100 cm^2 surface area. With chromatin immobilized, the DNA overhangs were blunted with a mixture of dATP, dTTP, dGTPαS and Biotin-14-dCTP and subjected to ligation. Afterwards, DNA was purified and treated with *E. coli* exonuclease III to remove the biotinylated residues from non-ligated DNA fragments. Ligated DNA fragments were then pulled down with streptavidin coated magnetic beads, attached to Illumina paired-end sequencing adaptors, and amplified to obtain the TCC library.

A.2.1. Assembling contact catalogue

The contact catalogue was assembled as described previously (1) with minor modifications in adjusting for ligation junctions: To increase alignment efficiency, all sequencing reads were scanned for the existence of potential ligation junction sequences (for HindIII libraries the junction sequence is “AAGCTAGCTT”). In reads with ligation junctions, all bases after the 3′ of the midpoint of the junction and the corresponding quality scores were removed. Furthermore, to adjust for star-activity of the restriction enzyme and other factors that result in ligation junctions with a sequence that slightly deviates from the expected consensus (6), we adjusted the scanning algorithm to allow for one mismatch or deletion in the entire expected junction sequence. In other words, any sequence that differed from the ligation junction by one mismatch or deletion was also considered to be a junction, and the sequence after the junction’s midpoint was removed from the corresponding read. This filtering strategy significantly improved the percentage of alignable reads.

A.2.2. Alignment to the human genome

The filtered reads were aligned against the GRCh37/hg19 reference sequence of the human genome using Bowtie-0.12.7 with a maximum of three mismatches allowed. The total number of aligned reads in each end is shown in **Table S3**. After alignment, the genomic positions of the read pairs that corresponded to the same sequencing cluster were combined to generate a catalogue of binary contacts. All pairs for which at least one of the reads could not be unambiguously aligned were removed from the contact catalogue leading to ~150 millions aligned paired reads (**Table S3**).

A.2.3. Removing non-informative pairs

Three types of pairs do not contain any information about the spatial organization of the genome and can be removed: PCR multiplications, non-ligated DNA fragments (flakes), and self-loops(1). To filter PCR multiplications, groups of read pairs that aligned to identical positions on both ends were removed, leaving only one instance per group in the catalogue. To filter flakes, pairs that aligned less than 1000 base pairs (bp) apart to opposite strands of the reference sequence were removed from the catalogue. To filter the self-loops, all pairs that aligned closer than 30,000 bp were removed.

The total number of read pairs after the above filtration steps for the binary contact catalogue of the library is about 98 Million (**Table S3)**.

A.3. TCC Data Processing

Our structural modeling approach is outlined in a flow-chart (**Fig. S1D)**.

A.3.1. Raw matrix

Interaction frequency counts were binned every 138 hindIII restriction fragment sites and a matrix registering pairwise interaction frequency between bins was constructed, we denote it as $\mathbf{C}_K \equiv \left(c_{ij}\right)_{K\times K}$, where K=6002 segments or bins.

A.3.2. Correction on the raw matrix

At the resolution of consecutive 138 hindIII sites bin matrix, we safely corrected some regions that were unusual. One type of correction was filling the 0 consecutive contacts within a chromosome. When such situation occurred for a consecutive bin pair (i, i+1) we took an average from the maximum of each bin i and i+1 to replace the 0 value. Another type of correction was to replace outstandingly high contact frequency to the EBV genome. For every bin we counted the frequency of contact to the EBV genome and calculated the average, i.e. 29.

We applied a high cutoff, i.e. 75 (about mean + 4 times standard deviations) to indicate which bins possessed high contact frequencies with EBV, and those bins usually have very high contact across genome as well. For those bins, we calculated the average of their direct neighbors contact frequencies for correcting them:

$$c_{i,j} = \min\left\{c_{i,j}, (c_{i+1,j} + c_{i-1,j})/2\right\}, j = 1..K, j \neq \{i-1, i+1\}. \qquad [9]$$

These corrections are optional since it is not generally applicable for other Hi-C data.

A.3.3. Removing contact frequency biases by iterative correction method (ICE)

We performed normalization on the smoothed matrix with an approach known as iterative correction and eigenvector decomposition (ICE) (7). Before using the ICE normalization method, we removed 1% bins with the fewest contact frequencies by merging them with their direct neighbors, following a suggested contact frequency cutoff value of 1%-2% (7), resulting a matrix with 5941 bins (**Fig. S1E**). We then applied a fast decaying power-law smoothing function on each individual chromosome submatrix:

$$\tilde{c}_{i,j} = \frac{\sum_{k=-\omega}^{\omega}\sum_{l=-\omega}^{\omega}\frac{c_{i+k,j+l}}{|sk|^p + |sl|^p + 1}}{\sum_{k=-\omega}^{\omega}\sum_{l=-\omega}^{\omega}\frac{1}{|sk|^p + |sl|^p + 1}} \qquad [10]$$

where ω, *s*, *p* we used were 3, 3, and 3, respectively.

The ICE method was applied according to Imakaev et al. (7) with the number of iterations set to 10 (the matrix had achieved convergence before 10th iteration). Interactions of consecutive bins within a chromosome were included to allow the definition of reference frequencies ($f^{\max}$, see below). The resulting matrix is $\mathbf{F}_K \equiv \left(f_{ij}\right)_{K\times K}$.

A.3.4. Bin level contact probability

The contact probability is defined as the probability for observing a given contact in the structure population (1). For each chromatin segment (i.e. a bin in the normalized contact frequency matrix) we define a threshold value $f^{\max}$, which defines the frequency at which a contact is formed in 100% of the structure population.

For a chromosome segment (matrix bin) $f^{\max}$ is chosen based on the contact frequencies to its two adjacent bins within a chromosome.

For each bin $i, f^{\max}$ is set as

$$f_i^{\max} = \min\left\{ f_{i,i-1}, f_{i,i+1} \right\}. \qquad [11]$$

Where f_{ij}, is the normalized contact frequency between bins i and j.

For the first and last bins in a chromosome (subtelomeric regions), we set $f_i^{\max} = f_{i+1}$ and $f_i^{\max} = f_{i-1}$ respectively.

$f^{\max}$ serves as a common reference point to calibrate contact probabilities between chromatin segments. For a given pair of bins i and j the contact probability at bin level is defined as

$$p_{ij} = \min\left\{ \frac{f_{ij}}{\min\left\{ f_i^{\max}, f_j^{\max} \right\}}, 1 \right\} \qquad [12]$$

With this formulation, two consecutive bins on the same chromosome will have $p_{ij} = 1$ to guarantee the structural integrity of the chromosome. An example of this matrix is shown for chromosome 1 (**Fig. S1E**).

A.3.5. Domain level contact probability

A "coarse-grained" domain contact probability matrix, $\mathbf{A}_N$, defines the fraction of models in the population in which a given domain contact is present. It is defined as $\mathbf{A}_N = (a_{IJ})_{N\times N}$, where a_{IJ} is the contact probability between domains I and J, and $\boldsymbol{N}$ is the total number of domains in the genome. a_{IJ} is calculated from the corresponding contact probabilities at the bin level.

If $b(I)$ is the set of all bins in domain I, $b(J)$ is the set of all bins in domain J, and $p_{\alpha,\beta}$ is the contact probability set of all pairwise combinations between bins in $b(I)$ and $b(J)$, then

$$a_{IJ} \equiv \left\langle \mathrm{top}5\% \left\{ p_{\alpha,\beta} \mid \alpha \in b(I), \beta \in b(J) \right\} \right\rangle \qquad [13]$$

is the average value of the top 5% ranked contact probabilities in $p_{\alpha\beta}$ (**Fig. S1E**). If I and J are two consecutive domains in the chromosome chain the maximum $p_{\alpha,\beta}$ value is used, which naturally ensures that the two consecutive chromatin domains will always be in contact ($a_{IJ} = 1$) to guarantee the necessary structural integrity of the chromosome.

A.4. Structural Representation of the Genome

Chromosomes are segmented into chromatin domains, which are represented by spherical volumes following our previously published approach (1). The plaid pattern in the chromosome contact frequency heat maps suggests a partition of chromosomes into domains of consecutive regions with similar long-range contact behavior. We have previously determined the domain boundaries of these macro-domains using a constrained clustering algorithm combined with an automatic cluster cutoff detection (1). Regions in a domain form the vast majority of their contacts to regions within the same domain, therefore we can assume that these regions on average are closer to each other than to regions in other blocks. Chromatin regions within a domain share similar functional properties, such as similar replications timing. The long-range contacts of regions in a domain are highly correlated, indicating that the domain acts as a structural unit. Therefore, it is expected that a large fraction of the chromatin regions in a domain preferentially occupy a similar nuclear sub-territory whose spherical volume can be approximated by the domain sequence length and the total occupancy of the genome in the nucleus (1). This model does not exclude mixing of regions between blocks, but assumes that the majority of domain regions localizes in the same sphere.

To balance short and long range interaction profiles in the raw TCC matrix domain segmentation is performed on sequence distance-normalized TCC maps. We apply a distance-scaling normalization similar to our previous works (1, 9),

$$\begin{aligned}
&\overline{\mathbf{F}_{K'}} = \left(\frac{c_{ij}}{\tilde{v}_{d=|i-j|}} \right)_{K' \times K'} \\
&\tilde{v} = \lambda(v) \\
&v \equiv \left(v_d \right)_{d \in \{1,..,(K'-1)\}} \\
&v_d = \left\langle \frac{\sum_{i=1}^{K'-d} c_{i,i+d}}{K'-d} \right\rangle_{\text{all chromosome}}
\end{aligned} \qquad [14]$$

where the angle brackets denote an average over all chromosomes, d is bin distance of a $K' \times K'$ matrix, K' is the number of bins in an intrachromosome matrix, and λ is a locally weighted polynomial regression smoothing function ("*lowess*", implemented in R software). Although this normalization is carried out to individual chromosome, the scaling vector, v, is computed by averaging all chromosomes.

Following our previously published procedure (1), we used hierarchical clustering of intra-chromosomal contact frequencies using the following distance matrix: $\mathbf{D}_{K'} = \left(1-\text{PCC}(i,j)\right)_{K' \times K'}$ where PCC(i,j) is the Pearson's correlation coefficient between row vector i and j in the distance-normalized contact frequency matrix $\overline{\mathbf{F}_{K'}}$. To *automatically determine the number of clusters* (i.e. domains), we minimized an objective function to balance the number of clusters and the spread of distances inside the clusters (10). From the ~ 500 kb resolution matrix (K = 5941), the clustering method segmented the genome into domains with median size of ~3.5 Mb (**Fig. S1F**). Our hierarchical clustering method is applicable for multiple granularities of domains (at different resolution of domains).

A.5. Analyses of the structure population

A.5.1. Centromere cluster detection

To identify centromere clusters in a structure, we ran a density-based spatial clustering algorithm implemented in R (http://cran.us.r-project.org/), *DBScan* package, by calling "dbscan(d, method='dist', eps=200, MinPts=3)", where variable d is a distance matrix. The smallest cluster size and the 'reachability distance' were set to 3 and 200 nm, respectively. The surface-to-surface distance between 2 centromeric representative beads were used as the 'reachability distance'. A centromeric representative bead was called when it overlaps (covers) a centromere gap defined in hg19 reference genome. All the 46 beads were included for the non-overlapping centromeric clusters detection (see **Fig. 4C** for some cluster examples in the main paper; the analyses results are plotted in **Figs. S4A-D**). The results did not change when the distance cutoff was varied between 100 to 300 nm.

A.5.2. Correlation coefficients

A correlation of two matrices can be assessed in several ways. In most cases, we presented the element-wise or the row-based correlations. For the element-wise correlation, we construct a vector from each matrix in the same indexing order, and then assess their correlation (usually with Pearson's and Spearman's method). For the row-based assessment, a Pearson's correlation coefficient (PCC) is computed between corresponding rows from both matrices, and then the average PCC over all rows in the matrix is reported.

A.5.3. ICP

On the normalized (iterative corrected) matrix, an ICP index of bin i is calculated as described previously (1) as the fraction of interchromosomal counts across other bins,

$$\mathrm{ICP}_i = \frac{\sum\limits_{j \in \beta(i)} f_{ij}}{\sum\limits_{j \neq i}^{N} f_{ij}},$$

where $\beta(i)$ are all interchromosomal bin partners of i.

A.5.4. Epigenetic analyses

The data of histone modifications, DNase hypersensitivity, RNA polymerase II binding, and gene expression for GM12878 cell line were obtained from the Encode project (11). We downloaded the bigwig files from the following sites:

- http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHistone/
- http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeSydhTfbs/
- http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCaltechRnaSeq/

The data of DNA methylation was obtained from (12, 13). The lincRNA transcripts data (14) was downloaded from http://www.broadinstitute.org/genome_bio/human_lincrnas/?q=lincRNA_catalog.

BigwigSummary program (from USC Genome Browser) was called to extract all the data for requested regions (or the defined 639 domains) in the analyses. The epigenetic signal was computed as an average per domain, thus it did not depend on the genomic length of a domain. As for the recurrent pattern analyses of centromeric regions, we extracted the data for each chromosome from a region defined by the following. The start position of a centromeric region was chosen from either at *domain start position* of the centromere representative bead or position at *5 Mb upstream from the left border* of hg19 centromere gap, which ever was the leftmost. Likewise, the end position of a centromeric region was chosen from either at *domain end position* of the centromere representative bead or position at *5 Mb downstream from the right border* of hg19 centromere gap, which ever was the rightmost. The reason of taking this way was the recurrent structural patterns were mined by using the 46 centromere-representative beads, however some of them were completely representing the gap regions only that yield no epigenetic signals. For a detail list of the centromeric gaps and our defined centromeric regions please refer to **Table S4**.

A.6. Convergence of our results with respect to the size of the structure population

To test the convergence of our results with respect to population size, we generated 5 different populations in size of 100, 1000 5,000, 10,000, and 20,000 structures and also performed replica calculations (repeated independent calculations) for each population totaling independently calculated 10 structure populations. We first calculate the convergence of the domain-domain contact frequencies in the structure populations. At populations larger than 1,000 structures, we observe excellent convergence of contact frequencies with a good correlation value (**Fig. S7A**). Also, when analyzing the 3D structures we observe a similar behavior (**Fig. S7B).** The average radial position of each domain fully converges at population sizes larger than 1000 structures, a size much smaller than our population size reported in the paper. The average radial domain positions from a population of 20,000 structures are essentially identical to those generated with a population at 10,000 structures (PCC = 0.9998, p-value < 2.2e-16).

We also tested the reproducibility of our results in replicate calculations with increasing population sizes. All structural features and the contact frequencies are highly reproducible at population sizes larger than 1000 structures. A population of 10,000 structures has reached high reproducibility and shows fully converged structures.

Table of Pearson's correlation coefficients (PCC) when the features are compared between replica calculations

Feature	100	1k	5k	10k	20k
Average radial position: mean PCC per chromosome	-0.150	0.997	0.999	1.000	1.000
PCC of the average radial position of all domains	0.200	0.998	1.000	1.000	1.000
PCC of the pairwise contact frequencies	0.925	0.999	0.984	1.000	1.000

In summary, our results are well converged at a population size of 10,000 structures and are highly reproducible. Our results remain unchanged when calculating a larger population of 20,000 structures (**Figs. S7C-G**).

A.7. Structure population generated from a reduced data set

To assess the impact of sub-centromeric interactions on the genome structures, we generated a population of 10,000 structures from Hi-C data where interchromosomal interactions were only considered for sub-centromeric spheres (see the right most heat map shown in **Fig. S2A**). The sub-centromeric spheres were defined as those spheres containing the centromere gap (defined in the hg19 reference genome) and the closest 2 direct neighbor spheres from each side (left and right on the chromosome bead chain). The average radial positions of chromosomes in this population correlated very well with those from the "regular" structure population that was generated using the complete Hi-C data (PCC = 0.959, p-value = 6.11e-13; **Fig. 3A bottom left panel**). When these radial positions were compared to those from FISH experiments (15), the PCC is 0.650 (p-value = 0.000786).

A.8. Structure population generated with uniform centromere-centromere interaction probabilities

To test the impact of non-specific centromeric interactions on genome structures, we tested a model, in which the interactions between subcentromeric regions of all chromosomes are equally probable and therefore are not specific with respect to the chromosome identity. In this model each pair of centromeres has equal probability to interact and the total number of subcentromeric interactions is kept equal to those of the original Hi-C data. Specifically, for every pair of centromere spheres, we assigned a probability of a_{IJ}=0.030 to be in contact. No other domains were restrained to form interchromosomal contacts. This model does not produce the correct radial positioning of chromosomes as known from FISH experiments (**Fig.**

S2B; in comparison with the FISH data (15), PCC = -0.269 with p-value = 0.2147), ruling out the notion that centromere clustering based on non-specific centromere-centromere interactions lead to the correct positioning of chromosomes. Our calculations indicate that chromosome specific interactions between subcentromeric regions and not non-specific phase separation of pericentromeric heterochromatin are crucial for reproducing experimentally known chromosome positing in our model.

A.9. Structure population generated with in-situ Hi-C data

To test if our conclusions remain unchanged when using the recently published higher resolution *in-situ* Hi-C data set from the Lieberman Aiden group (16), we also generated a structure population using the *in situ* Hi-C data set. The raw matrix ($\mathbf{C}_K$, with K=5941) is constructed from mapped reads (MAPQ > 30) using 92 "…merged_nodups.txt.gz" files (GSM1551nnn) After excluding intra-chromosome pairs with distance less than 30 kilobase pairs (section A.2.3) the matrix contained more than 3.5 billion contacts. Matrix normalization and generation of the probability matrix were performed as described in sections A.3.3 to A.3.5. The contact probability matrixes from both data sets were highly correlated (0.993, p-value < 2.2e-16). We then generated a structure population of 10,000 structures with the contact probabilities from *in situ* Hi-C. The resulting structure population produces consistent results with those generated from the TCC data. In the following we compare the two structure populations.

Domain-domain contact frequencies in the model. The pairwise domain contact frequencies are highly correlated between the two structure populations (Pearson's correlation coefficient 0.9847 (p-value < 2.2e-6).

Chromosome radial positions. The average radial chromosome positions are very similar in both structure populations with a high Pearson's correlation coefficient of 0.866 (p-value = 2.67e-8; **Fig. S8B**). The radial chromosome position from both structure populations agree well with the FISH data (15) (**Fig. S8A**). Both structure populations reproduce equally well the average radial chromosome positions from the FISH experiments (Pearson's correlation coefficient of 0.762 (p-value = 2.36e-5) for the structure population from *in situ* Hi-C data and PCC = 0.747 for the structure population from the TCC data).

Domain radial positions. Also the radial positions of all the chromatin domains are very similar in both populations. The PCC of the average radial domain position in the two populations is 0.797 (p-value < 2.2e-16).

Domain-domain distances Also distances between domains are remarkably similar in both populations, (**Figs. 3B-C)**. The colocalization frequencies of four interchromosomal loci pairs (H0 with H1, H2, L1, and L2) are highly correlated in the two populations (PCC = 0.976, p-value= 0.02393). Also the cumulative distances between domains in 8 inter-chromosomal pairs of genes are reproduced remarkably well (all PCCs > 0.999 with p-values < 2.2e-16) (**Figs. S8C-D**).

Centromere clustering and positions. The structure population from *in situ* Hi-C data confirms our observations from the structure population generated with the Hi-C data. The radial position of a centromere generally decreases with increasing number of surrounding centromeres (see (**Figs. S8E** and **6A**). Also the abundance of centromere clusters per structure and the cluster size distribution is very similar between structure populations generated with *in-situ* Hi-C and TCC data (**Figs. S8F**, **Fig. S4A**, and **S4B**).

A.10. 3D DNA FISH

Based on the finding from our structural models, we designed multiple sets of 3D *fluorescence in situ hybridization* (FISH) experiments to see variation of colocalization formed by different groups of centromeres. Each group consists of 3 centromeres, they are on chromosomes {1, 9, 21}, {7, 10, 12}, and {2, 3, 6} (**Fig. 5B**). Colocalization event can be defined as all the 3 pairwise distances between centromeric domains are within a threshold. The results of these experiments are plotted in **Fig. 5C**.

To verify that centromeric regions are bridging the interaction inter-chromosomally, we performed a comparison between probes on centromeric of chromosome 1, 9, and 21 against the probes on distal regions from centromeres of the same chromosomes (the distal regions were about 56.8 Mb, 61.5 Mb, and 18.3 Mb away from centromere on chromosomes 1, 9, and 21, respectively). The contrast between the two groups is plotted in **Fig. S9B**.

A.10.1. DNA FISH probe design

For any particular chromosome domains (regions), multiple BAC clones were chosen and synthesized by Empire Genomics and tested individually for their specificity. The probe of a pericentromeric region on chromosome 1 lies at region 1q11, a standard designed by the company. Pericentromeric probes are RP11-831B17 (chr2: 96,977,476-97,209,012), RP11-1082I19 (chr3: 87,569,564-87,784,200), RP11-973P24 (chr6: 65,112,008-65,298,120), RP11-144H20 (chr7: 61,968,709-62,155,949), RP11-912B9 (chr9: 40,475,859-40,773,473), RP11-300L24 (chr10: 42,965,138-43,178,326), RP11-349I21 (chr12: 38,143,496-38,364,968), RP11-1089A22 (chr21: 10,949,929-11,145,665). The probes that are far from centromeric regions are RP11-57D16 (chr1: 179,836,951-179,993,708), RP11-760E14 (chr9: 110,383,457-110, 589, 787), and RP11-655P6 (chr21: 31,137,227–31, 348, 263). **Fig. 5B** illustrates the probes locations on the chromosomes.

A.10.2. FISH experimental procedure

The experiment was performed following the previous protocols (1, 17, 18) with slight modification. GM12878 cells were cultured in DMEM medium supplemented with 15% FBS, glutamine and penicillin/streptomycin as suggested by ENCODE. Two days before the experiment, 22mm x 22mm coverslip were cleaned and coated with L-poly-lysine (1mg/ml) at room temperature for 1-2 hours, and dried in tissue culture hood after brief rinse with sterile MilliQ water. On the day of the experiment, 10 million GM12878 cells were harvested by centrifugation at 100g for 10 minutes, resuspended in fresh culture medium ($3x10^6$ cells/ml), and seeded evenly on the coverslip in a 6-well tissue culture plate. After incubating at 37°C for 1 hour and briefly washing with PBS, the cells (on coverslips) were fixed with 4% freshly made paraformaldehyde (in 0.4x PBS) at room temperature for 10 minutes. The cell membrane was permeabilized firstly with 0.5% triton X100/1xPBS at room temperature for 20 minutes, and then through 4-5 freeze-and-thaw cycles (by dipping in liquid nitrogen and then thaw in room temperature) in the next day after pretreatment overnight with 20% glycerol/1xPBS and also before each dip. To facilitate access of the FISH probe to the chromatin DNA, the samples were washed with 0.05% triton X100/1xPBS 5 minutes each for two times, and then treated with 0.1N HCl at room temperature for 5-10 minutes to remove basic nuclear proteins. The HCl was removed from the sample followed by two washes with 0.05% triton X100/PBS and one wash with 2x SSC (diluted from 20xSSC: 3M NaCl, 0.3M sodium citrate, pH 7.0) 5-10 minutes each wash. The coverslips were then stored in 50% formamide/2x SSC at 4°C and were ready for the next step (good for 2 days to 2 months).

The denaturation and hybridization steps were performed according to the protocols suggested by the manufacturer (https://www.empiregenomics.com/files/store/products/FISH_probes/FISH_Protocol.pdf). The coverslips were brought to room temperature for 24 hours in advance before denaturing. On the day of experiment, fresh 70% formamide/2x SSC was prepared and pre-warmed at 73°C for 30 minutes. Cells on the coverslips were denatured in this solution (73°C) for 5 minutes, and dried through sequentially dipping into 70%, 85% and 100% ethanol 1 minute each at room temperature, and finally through evaporation at 45°C for 20 minutes. FISH probes were denatured similarly in 70% formamide/2x SSC for 5 minutes at 73°C and then quickly cooled down on ice. After incubating at 37°C for 10-20 minutes, three probes (150 ng each) for either targeted regions or for the three control regions were mixed thoroughly with 18 μl hybridization buffer (provided by the manufacturer), and applied evenly with the sample on a microscope slide. Hybridization of FISH probes with the samples occurred in a humidified chamber containing a paper towel soaked with 50% fomamide/ 2x SSC in dark at 37°C for 18-20 hours. Unbound FISH probes were removed by a series of washes, three times with 0.3% NP-40/0.4x SSC at 73°C for 2 minutes, each followed by a wash with 0.1% NP40/2x SSC at room temperature for 1 minute. After air-drying for 5 minutes in dark, the coverslips were mounted on microscope slide with 10 μl DAPI mounting solution and kept in dark at 4°C (ready for imaging).

A.10.3. FISH image acquisition

The FISH images were acquired with Zeiss Laser Scanning Confocal microscope (LSC780) with 63x magnification oil immersion objective lenses. Cells are randomly chosen (each vision field contains 6-15 cells). Signals from four different fluorophores were obtained with two alternative frame scans for best separation: first scan with two laser beams of 488 nm and 594 nm, followed by the second scan of 405 nm (for DAPI) and 532 nm laser beams (for the yellow probe used in targeted group) or 405 nm and 555 nm laser beams (for orange probe used in control group). The minimal laser power was used in combination with appropriate filter settings (MBS 488/594 and MBS 458/514/561/633) to greatly reduce the signal bleed through between channels. Images of cells with optical Z sections from the bottom to the top with 0.25 μm or 0.3 μm intervals were acquired one section after another (frame scanning) with the software Zen provided by the manufacturer. Signals of each probe were stored in separate channels (4 channels for 3 chromosomal regions plus DAPI staining of the whole chromosomal DNA).

A.10.4.Image data analysis procedures

The nucleus detection and distance measurements between probes were performed using the Nemo software for FISH image analyses (19). Automated nucleus detection mode was used as the standard procedure, and additional manual selection followed when needed. Each of the cells was subject to manual inspection and validated for containing at least the expected six bright spots corresponding to the location of the three FISH probes in the diploid nucleus (2 FISH signals per probe/marker are expected). The pairwise distances between probes were extracted to get the average distance among the 3 loci on different chromosomes or to determine whether the 3 loci were colocalized. We refer to the three loci as a triplet. Each of the three probes A, B, C is located on two homologues chromosome copies and the corresponding probe copies are labeled A and A'; B and B'; C and C'. Given 3 probes, there are a total of 8 possible triplett probe combinations per cell that can define a co-localization cluster: {*A,B,C*}, {*A,B,C'*}, {*A,B',C*}, {*A,B',C'*}, {*A',B,C*}, {*A',B,C'*}, {*A',B',C*}, {*A',B',C'*}. For each possible triplet we calculate the distance between the corresponding loci.

For each triplet, we calculate the "*triplet distance"*, which is defined as the average distances between probes in a triplet (**Fig. 5B**). To determine the cumulative frequency distribution we select the triplet with smallest triplet distance among all the possible triplet combinations per cell (**Figs. 5B** and **S9B)**.

A.10.5. Image data analysis results

To test the chromosome-specific nature of our predicted centromere clusters, we specifically performed 3D FISH experiments for three centromere clusters that are predicted with largely different frequencies in the population (**Fig. 5B**). We analyzed a total of more than 1500 3D FISH images. Since the chromatin domains used in our structural model are of a median size ~3.5 Mb, and the FISH probe is of the size ~200 kb, we expect that the cluster occurrence frequencies observed in FISH images shall be generally much lower than those from our structural population. Therefore, we are not using FISH to validate the absolute frequencies of individual clusters, but the relative frequency order among the clusters.

In our models, the centromere cluster of chromosomes 7, 10, and 12 (cluster 7-10-12) occurs substantially more frequently than the cluster of chromosomes 2, 3, and 6 (cluster 2-3-6), but less frequently than the cluster formed by chromosomes 1, 9, and 21 (cluster 1-9-21) (**Fig. 5C**). Because the cluster 1-9-21 is observed most frequently, its frequency serves as the reference. In our model the frequency of cluster 1-9-21 is about 1.2 fold larger than the frequency of cluster

7-10-12, and about 24 fold larger than the frequency of cluster 2-3-6. The ranked order of cumulative cluster occurrences and also the relative frequencies of the clusters in the FISH experiments were determined and compared with our model as follows.

1) Cumulative percentage of cells with respect to probe distances. To compare the colocalization propensity of centromeres in the three clusters without the use of a specific distance cutoff, we calculated for each cluster the cumulative percentage of cells with respect to the average distance among the triplet probes. In a diploid genome, there are 8 possible triplet combinations of the three probes and we determined for each cell the triplet with the smallest average distance. The cumulative percentage of cells for the smallest triplet distance for all three clusters demonstrates that centromeres 1, 9, and 21 are indeed consistently more frequently in proximity to each other than centromeres 7, 10, and 12, while the centromeres of chromosomes 2-3-6 are the least frequent to be in proximity (**Fig. 5B** lower right panel).

2) Relative cluster frequencies. We then compared the relative frequencies of clusters in the population. Given a specific distance cutoff, we define the three centromeres to form a cluster if all three pairwise probe distances are simultaneously smaller than a cutoff in the same cell image.

Selecting a distance cutoff: Human centromeres contain extensive tandem repeat arrays (1,500 to >30,000 copies) and their actual size can span up to 5Mb of DNA (20). It is fair to assume an average size of ~3Mb of centromeric DNA per chromosome. Because the FISH probes are adjacent to the centromeres, we need to consider the centromeric DNA when choosing the probe distance cutoff. We chose a distance cutoff of 1.5 micron, to allow for a consideration of the total expected ~9 Mb of centromeric DNA for three chromosomes.

Using a distance cutoff of 1.5 micron for both experiments and models, the absolute frequencies of the clusters 1-9-21, 7-10-12 and 2-3-6 are 6.5%, 4.4%, and 1.5% from the 3D FISH experiments, versus 34%, 29.4%, and 1.4% from our structure population, respectively. As expected, the cluster frequencies observed in FISH are lower than those from our structural population, because of the much smaller sizes of the FISH probes (~200kb) compared to those of the domain (~3.5Mb) used in the structure modeling. Therefore, FISH cannot validate the absolute frequencies of individual clusters, but the relative frequency order among the clusters. Indeed, **Fig. 5C** shows that our model predicts very well the relative cluster frequencies in FISH experiments. Also in FISH experiments, the centromere cluster 1-9-21 shows a substantially higher relative frequency than all the other clusters. For example, in FISH experiments the observed frequency value of the cluster 7-10-12 is only 67% of the frequency of the cluster 1-9-21. The frequency of the cluster 2-3-6 is only 23% of the frequency of the cluster 1-9-21. In the model, the rank order of frequencies is identical. The highest frequency is observed for cluster 1-9-21. The frequency of cluster 7-10-12 is only 86%, and the frequency of cluster 2-3-6 is only 4% of the frequency observed for cluster 1-9-21, respectively. The results are essentially unchanged when using a distance cutoff of 1.25 micron.

3) Centromere as contact points for chromosome cluster formation. Additionally, we tested if the centromeres are the main points of interactions for the chromosome cluster 1-9-21. We found that the three markers located in the pericentromeric regions of chromosomes 1, 9, and 21 showed substantially higher co-localization frequency (~3 fold at distance threshold 1.5 micron; **Fig. 5D**) than a control group of markers located at more distal regions from centromeres on the same chromosomes (56.8 Mb, 61.5 Mb, and 18.3 Mb away from centromere on chromosomes 1, 9, and 21, respectively). The cumulative percentage of the average probe triplet distances is consistently smaller for the subcentromeric probe cluster than for the control probes at more distant locations from the centromere (**Fig. S9**).

In summary, we experimentally validated the finding from our structural modeling that individual chromosomes differ substantially in their propensity to form centromere clusters. The observed rank order of centromere cluster frequencies is consistent with the predictions: cluster 1-9-21 is more frequent than cluster 7-10-12, while cluster 2-3-6 is the least frequent among the 3 clusters. We conclude that centromere cluster formation is highly chromosome specific in nature.

A.11. Cryo-X-ray Tomography

Lymphoblastoid cells (GM12878) were obtained from the Coriell Cell Repositories at the Coriell Institute for Medical Research and cultured in RPMI 1640 Basal Media (Life Technologies # 12633-012; No HEPES, No L-Glutamine, + Non-Essential Amino Acids, + 110 mg/L Sodium Pyruvate) plus 2mM L-Glutamine (1:100, Life Technologies # 25030-149), 0.4g/100mL (0.4% w/v), Pen Strep (1:100, Life Technologies #15140-148), and 15% Fetal Bovine Serum (ATCC # 30-2020). Cells were grown in 10 mL of complete growth media in an upright T25 flask. For soft x-ray tomography, cells were loaded into thin walled (200 nm) glass capillaries (in growth medium), and rapidly frozen by mechanically plunging, at 2 m/sec, into liquid nitrogen cooled propane. Projection images were collected at 517 eV using XM-2, the National Center for X-ray Tomography soft X-ray microscope at the Advanced Light Source of Lawrence Berkeley National Laboratory; the microscope was equipped with a resolution defining 50-nm objective lens. During data collection, the cells were maintained in a stream of helium gas that had been cooled to liquid nitrogen temperatures (21, 22). Cooling the specimen allows collection of projection images while mitigating the effects of exposure to radiation. For each dataset, 180 projection images were collected sequentially around a rotation axis in 1° increments to give a total rotation of 180°. An exposure time of between 150 and 300 ms was used (depending on synchrotron ring current). Projection images were manually aligned using IMOD software by tracking gold fiducial markers on adjacent images (23) and tomographic reconstructions were calculated using the iterative reconstruction method (24, 25). LAC values were determined as described previously (26).

B. SI FIGURES

B.1. Figure S1

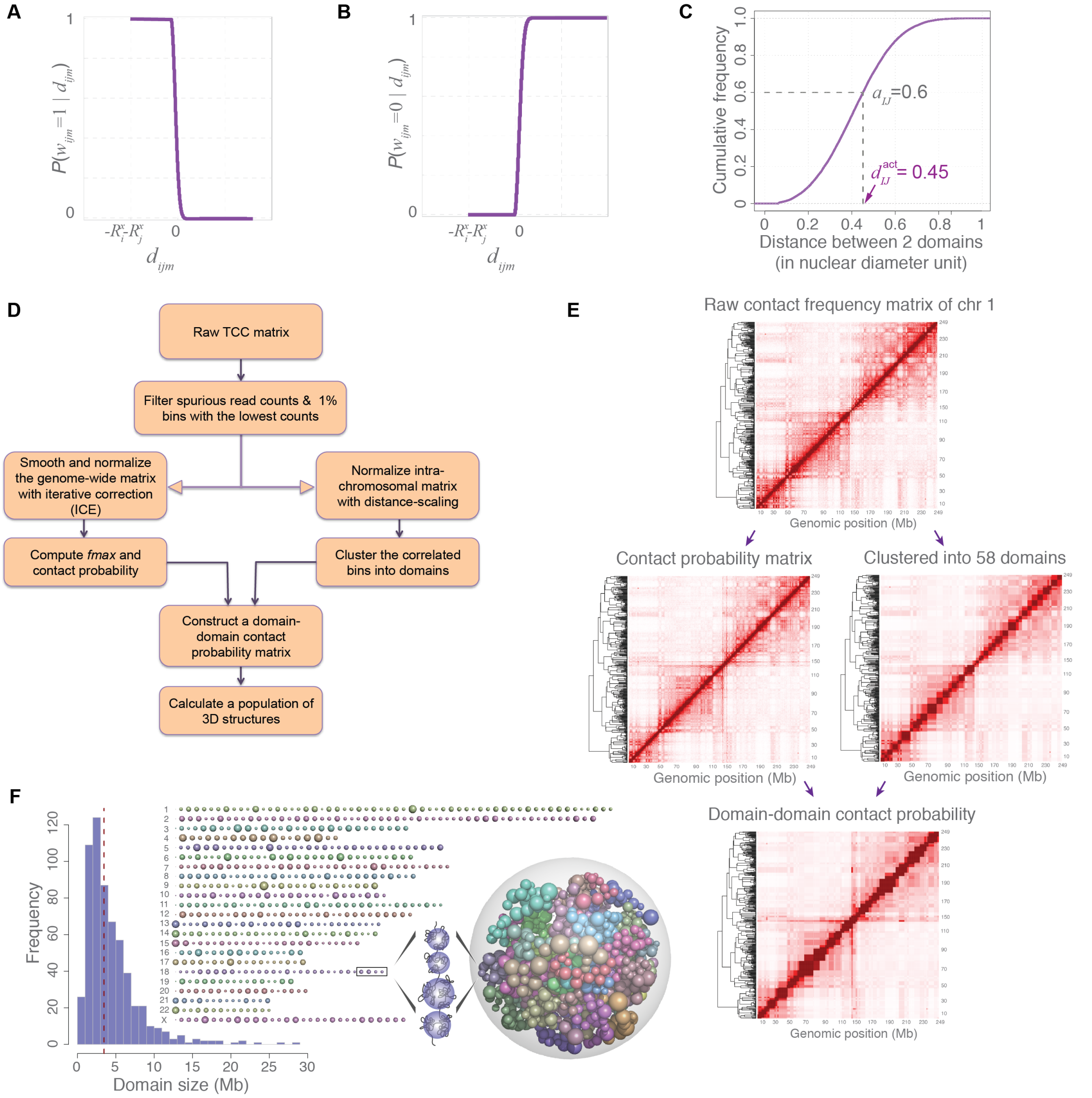

A
1
0
P(w_ijm =1 | d_ijm)
-R_i^x-R_j^x
0
d_ijm
B
1
0
P(w_ijm =0 | d_ijm)
-R_i^x-R_j^x
0
d_ijm
C
Cumulative frequency
0
0.2
0.4
0.6
0.8
1.0
a_IJ=0.6
d_IJ^act= 0.45
0
0.2
0.4
0.6
0.8
1
Distance between 2 domains
(in nuclear diameter unit)
D
Raw TCC matrix
Filter spurious read counts & 1% bins with the lowest counts
Smooth and normalize the genome-wide matrix with iterative correction (ICE)
Normalize intra-chromosomal matrix with distance-scaling
Compute fmax and contact probability
Cluster the correlated bins into domains
Construct a domain-domain contact probability matrix
Calculate a population of 3D structures
E
Raw contact frequency matrix of chr 1
Genomic position (Mb)
Contact probability matrix
Clustered into 58 domains
Genomic position (Mb)
Genomic position (Mb)
Domain-domain contact probability
Genomic position (Mb)
F
Frequency
0
20
40
60
80
100
120
0
5
10
15
20
25
30
Domain size (Mb)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X

Figure S1 Descriptions of methods. (**A**) The function for spheres i and j in structure m that have a contact (defined in Eq. [1]). (**B**) The function for spheres i and j in structure m that do not have a contact. (**C**) Illustration of obtaining activation distance d_{IJ}^{act} for a given probability of interaction a_{IJ} that is posited to a pair of domains I and J. The curve is cumulative frequency of the pair in the optimized structure population. As an example, sphere pairs whose current distances are within the smallest 60% distances (e.g. ~0.45 nuclear diameter unit in the figure) will be restrained to be in contact in the next stage of optimization. **(D)** Flow chart of our input data processing for 3D modeling. (**E**) The flow of matrix transformations from the raw TC matrix to coarse contact probability matrix. As described in **D**, a contact probability and domain matrix for 3D representation are generated (see sections A.3-A.4). Shown here is chromosome 1, the hierarchical clustering approach resulted in 58 domains. The total number of genome-wide domains is 639. (**F**) The histogram of size of the 639 domains determined for our structural models (see **Table S1**). The vertical dash line marks the median size of domain (3.5 Mb). Spheres representations are illustrated where the volume of each sphere is proportional to the genomic length of represented domain. The spheres also include 27 "dummy beads", filling the genomic gaps and serve as excluded volumes, totaling 666 beads. In total, there are 1332 beads within each nucleus representing female human lymphoblastoid genome model.

B.2. Figure S2

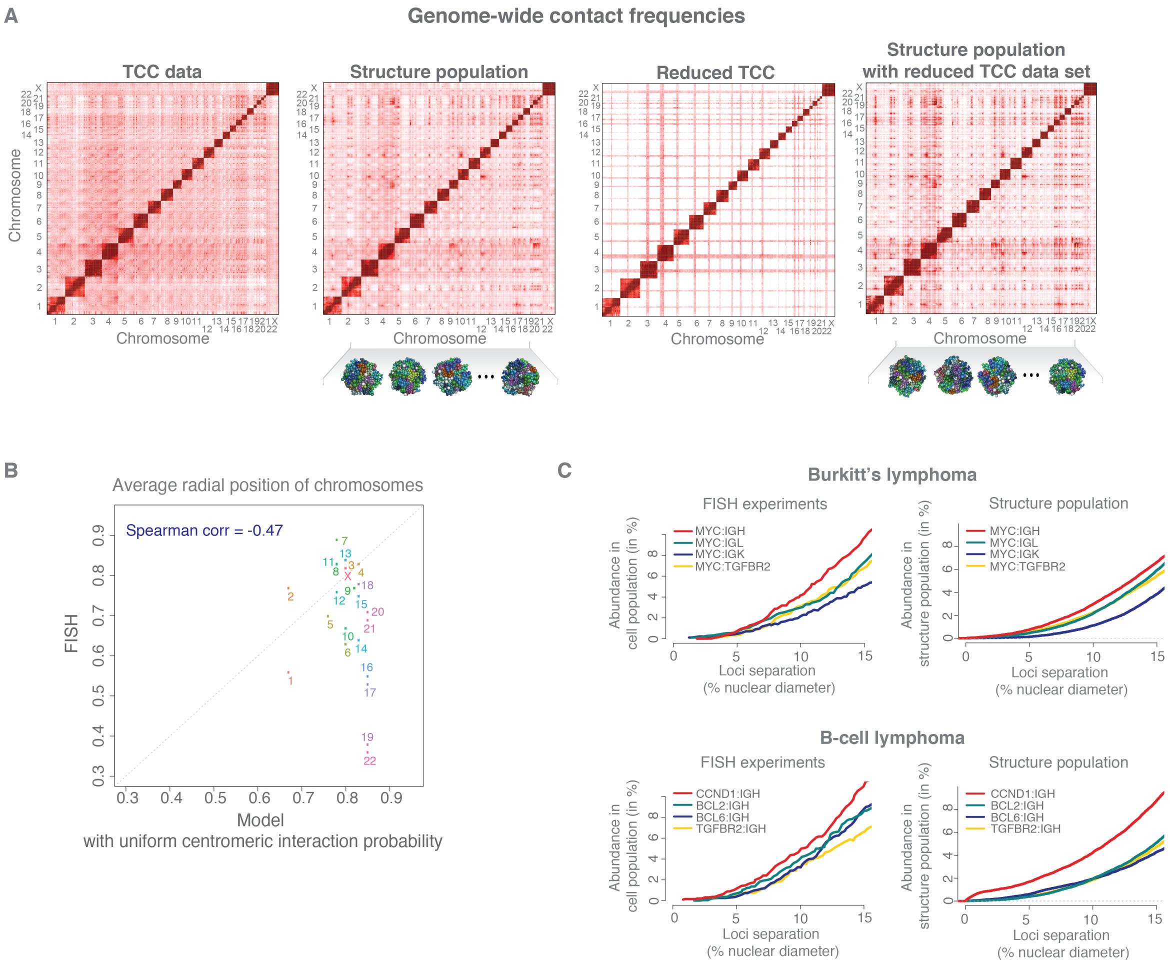


Figure S2 Comparison between models and experiments. (**A**) Genome-wide contact probability heat maps. From left to right: TCC data, structure population, reduced TCC data, structure population from the reduced TCC data. The reduced TCC data only includes intrachromosomal and subcentromeric interchromosomal interactions. To visualize the heat maps, the relative bin size reflects the corresponding domain size. The color scale ranges from white to dark red, for low to high contact frequencies respectively. (**B**) Comparison of the chromosomes' average radial positions from FISH data and a structure population with uniform centromere-centromere interaction probabilities. This model does not reflect the correct chromosomes positions in nucleus. (**C**) Cumulative frequency distribution of gene pair's distances replotted in the same way as in the original Roix et al. paper (27), to compare results from FISH experiments (left panels) and structure population (right panels). Four pairs of genes associated with Burkitt's lymphoma (top panels), and B-cell lymphoma (bottom panels).

B.3. Figure S3

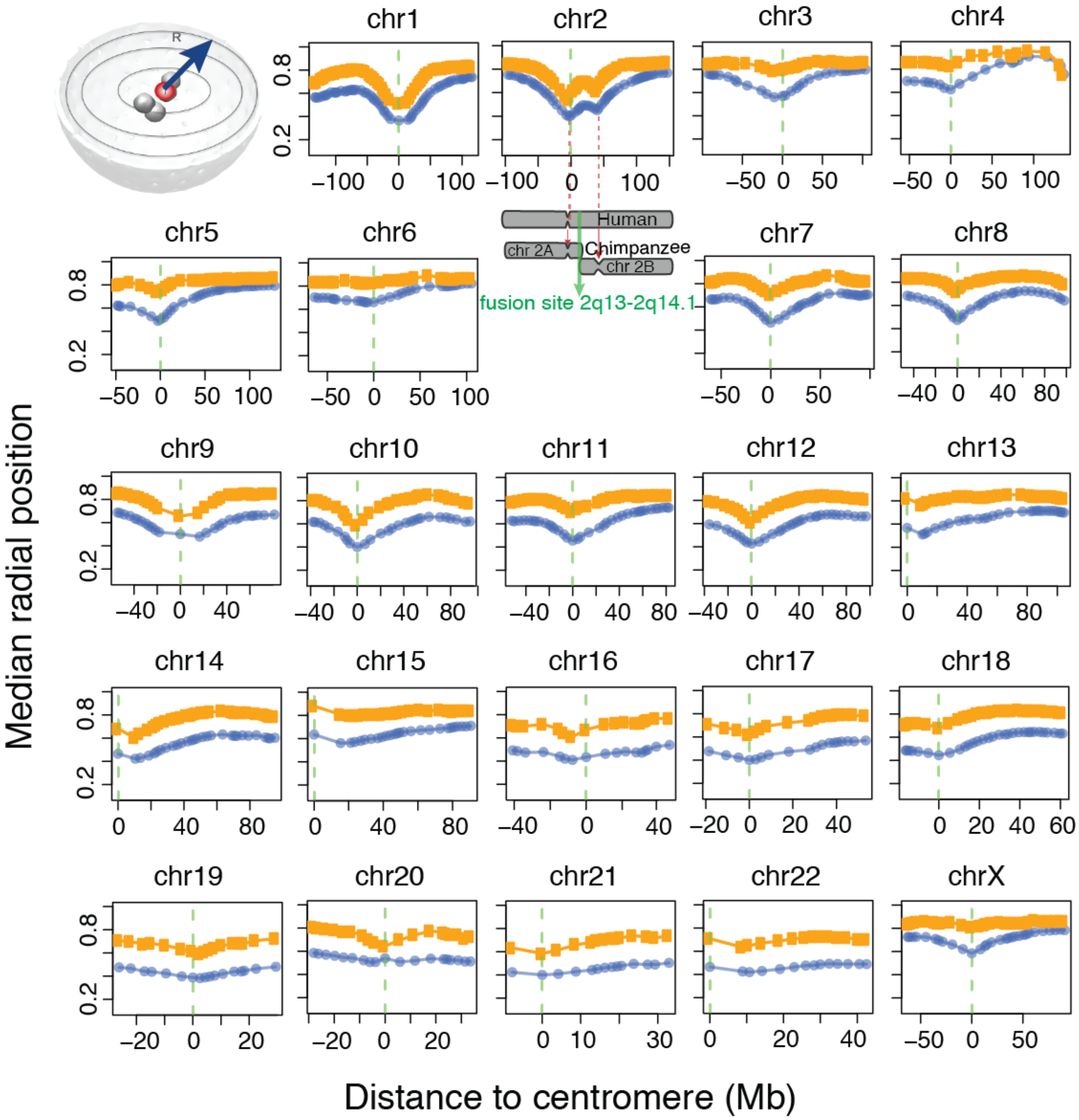


Figure S3 Radial positions**.** The median radial position of each chromosome domain is plotted against its sequence position (related to **Fig. 4A**). Blue and orange curves correspond to the domain positions calculated from the radially inner- and outer-most chromosome copy in each structure, respectively. Centromere is at position 0, marked with the vertical green dashed line. For most chromosomes, a dip is apparent near centromere. Below the box of chromosome 2, we illustrate the hypothetical head-to-head fusion of ape's chromosomes 2A and 2B. The second dip in the plot for chromosome 2 could be related to the vestigial centromere. The q-telomere of chromosome 4 is known to locate nearby the nuclear envelope(3), thus we include this information in the model

B.4. Figure S4

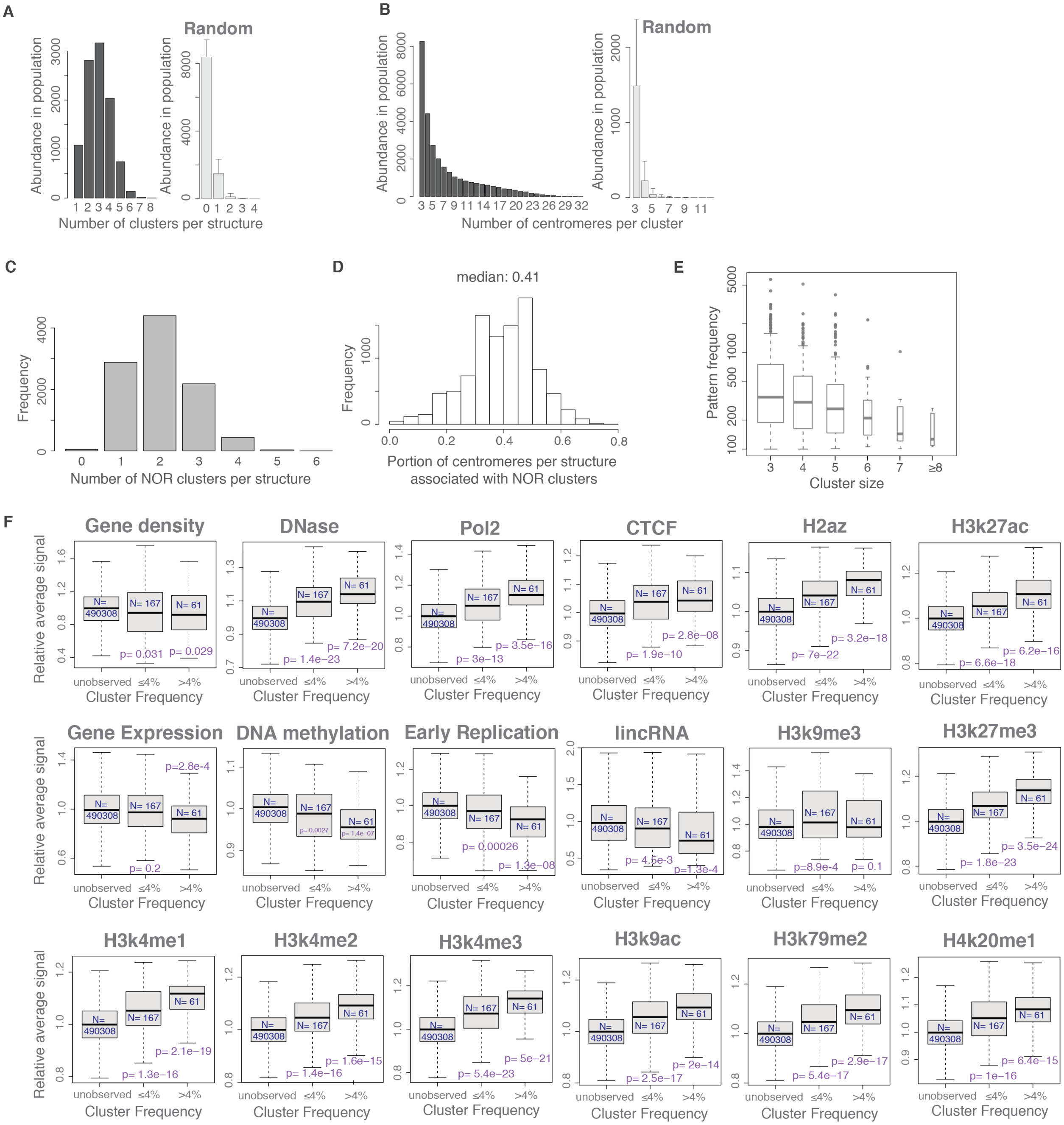


Figure S4 Centromere clusters. (**A**) (left panel) Histogram showing the abundance of centromere clusters per structure in the population. The median and most abundant number is 3**.** (Right panel) Random control. Histogram showing the abundance of clusters when one domain per chromosome is randomly picked in the structure population; error bars are standard deviations from 1,000 randomizations. The random control shows a dramatically reduced interchromosomal clustering and the majority of structures show no clusters. (**B**) (left panel) Histogram of the size distribution of centromere clusters observed in the structure population. The number of observed clusters decreases sharply with the increase of cluster size. (right panel) random control defined as in **A**. (**C**) The histogram of the number of NOR (nucleolus organizing region) clusters if defined as clustered centromeres that contain at least one acrocentric centromere. With this definition, roughly 66.7% of centromere clusters contain NOR. (**D**) Histogram of fraction out of 46 centromeres that participate in NOR clusters. About 40% of centromeres are associated with NOR clusters. The fraction is computed for each structure, thus the total of data point in the histogram is 10,000. Data

shown in panels A-D were generated with the method described in section A.5.1. (**E**) Box-and-whisker plots showing the relationship of frequency and size of the 798 centromeric frequent pattern clusters (above 1% abundance in structure population). The more centromeres involve the less frequent such patterns occur in structure population. (**F**) Epigenetic signals of pericentromeric regions from the frequent pattern clusters grouped in different frequencies: unobserved (below 1% cutoff; see **Material and Methods** section “Detection of centromere cluster recurrent patterns”), infrequent (1%-4%), and frequent (>4%). The p-values shown are calculated with one-sided Wilcoxon test against the ‘unobserved’ group. As an example, clusters formed by more than 4 centromeres are selected; the conclusion still holds as well using different cluster size.

B.5. Figure S5

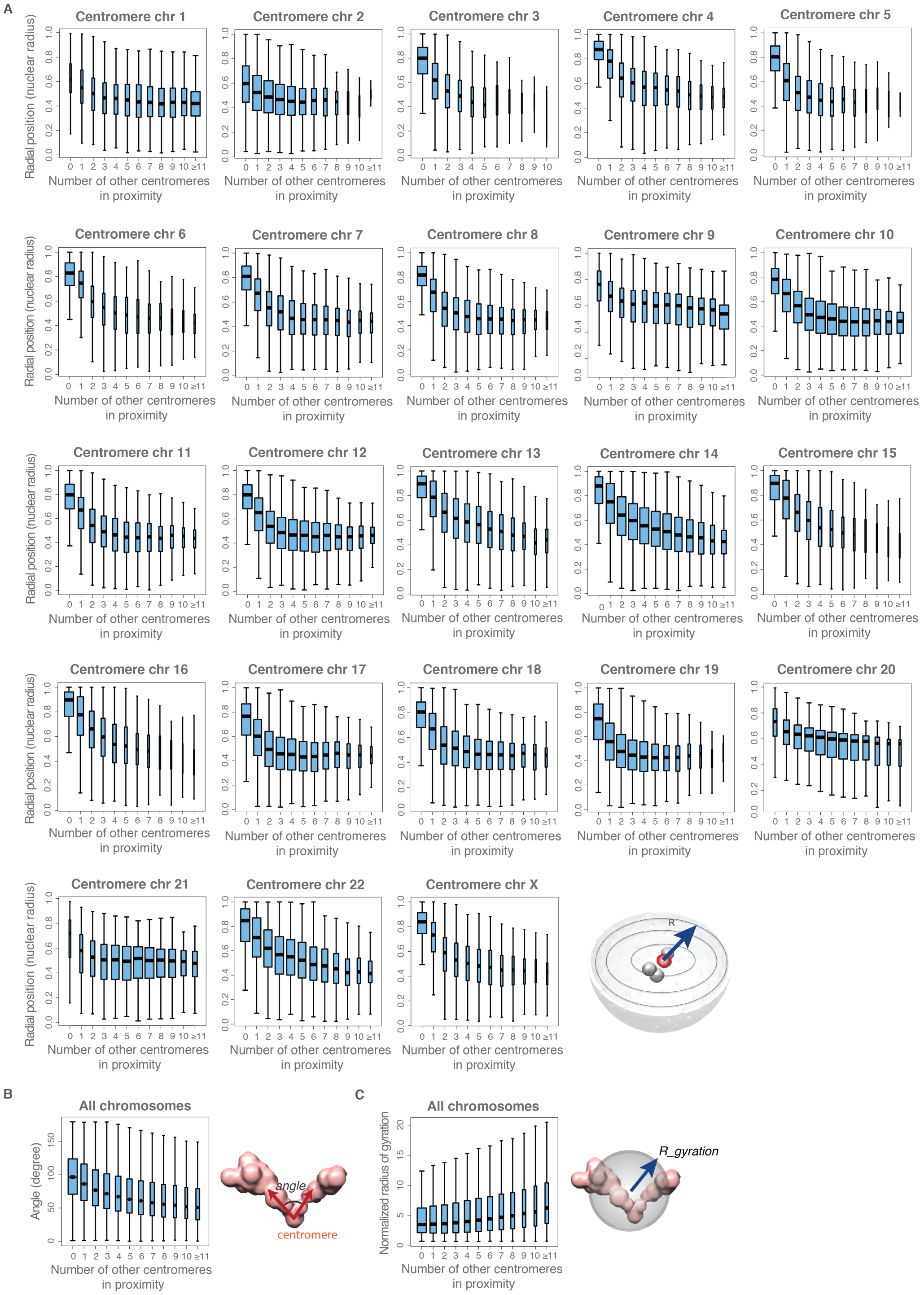


Figure S5 (A) Box-and-whisker plots of radial position of centromeres for each chromosome as a function of cluster size (related to **Fig. 6A**). The whiskers mark the range of data group within each box. The outliers are not shown. **(B)** Box-and-whisker plots of angle formed between chromosome arms within 30 Mb from centromeres as a function of cluster size. **(C)** Box-and-whisker plots showing radius of gyration normalized by the chromosome size (genomic length) as a function of other centromeres in contact. In general, there is a trend of chromosomes to be more elongated when the centromeres have more other centromeres nearby.

B.6. Figure S6

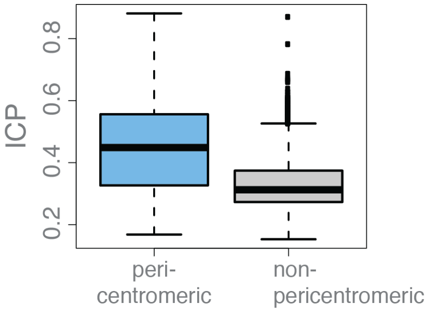


Figure S6 Comparison of Inter-chromosomal interaction fraction (ICP) between pericentromeric regions and the rest of the genome. The ICP was calculated from a genome-wide matrix binned every 138-hindIII sites (the pericentromeric regions were taken up to 2 bins from the right and left centromeric gaps). See section A.5.3 for detail on ICP.

B.7. Figure S7

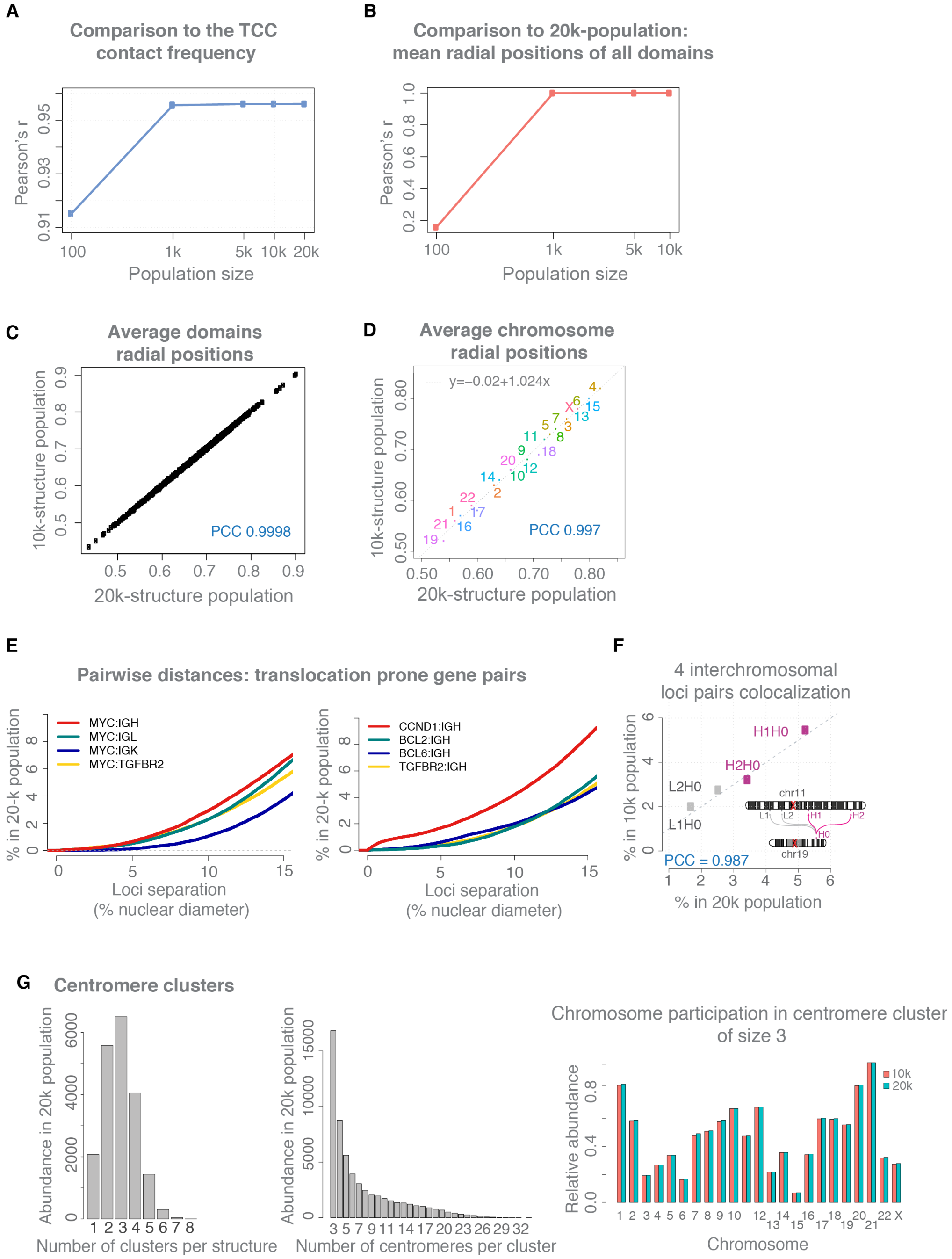


Figure S7 Convergence tests. (**A**) Pearson's correlation coefficients between normalized contact frequency matrices from the experimental TCC data and structure populations with increasing population sizes (section A.5.2). (**B**) Pearson's correlation coefficients of the mean radial positions of all domains in a population of 20,000 structures compared with those from populations ranging in size between 100 and 10,000 structures. (**C**) Comparison of the average radial positions of each domain in populations of 10,000 and 20,000 structures. (**D**) Comparison of the average radial position of chromosomes from populations of 10,000 and 20,000 structures. (**E**) Cumulative frequency plots of distances between pairs of domains in the population of 20,000 structures (related to **Fig 3C** or **Fig. S2C)**. (**F**) Comparison of the colocalization frequency for 4 interchromosomal loci pairs in a population of 10,000 structures (presented in the main text) and a population of 20,000 structures. (**G**) (**Left and middle panels**) The statistic of centromere clusters found in the population of 20,000 structures (related to **Figs. S4A and S4B**). (Right panel)

Chromosome propensity to participate in centromere clusters containing 3 centromeres in the populations with 10,000 and 20,000 structures, respectively. For each structure, a colocalization event of a triplet (3 centromeres form different chromosomes) is called if all pairwise distances connecting the triplet are less than a threshold (1.5 micron) (**see Fig. 5B** bottom mid panel). The total number of detected triplet combinations is 1771. The propensity is calculated as the total frequency of the chromosome to be found in all clusters. The maximum height of the bar is normalized to 1. The bar plots show that the results from both population are essentially identical.

B.8. Figure S8

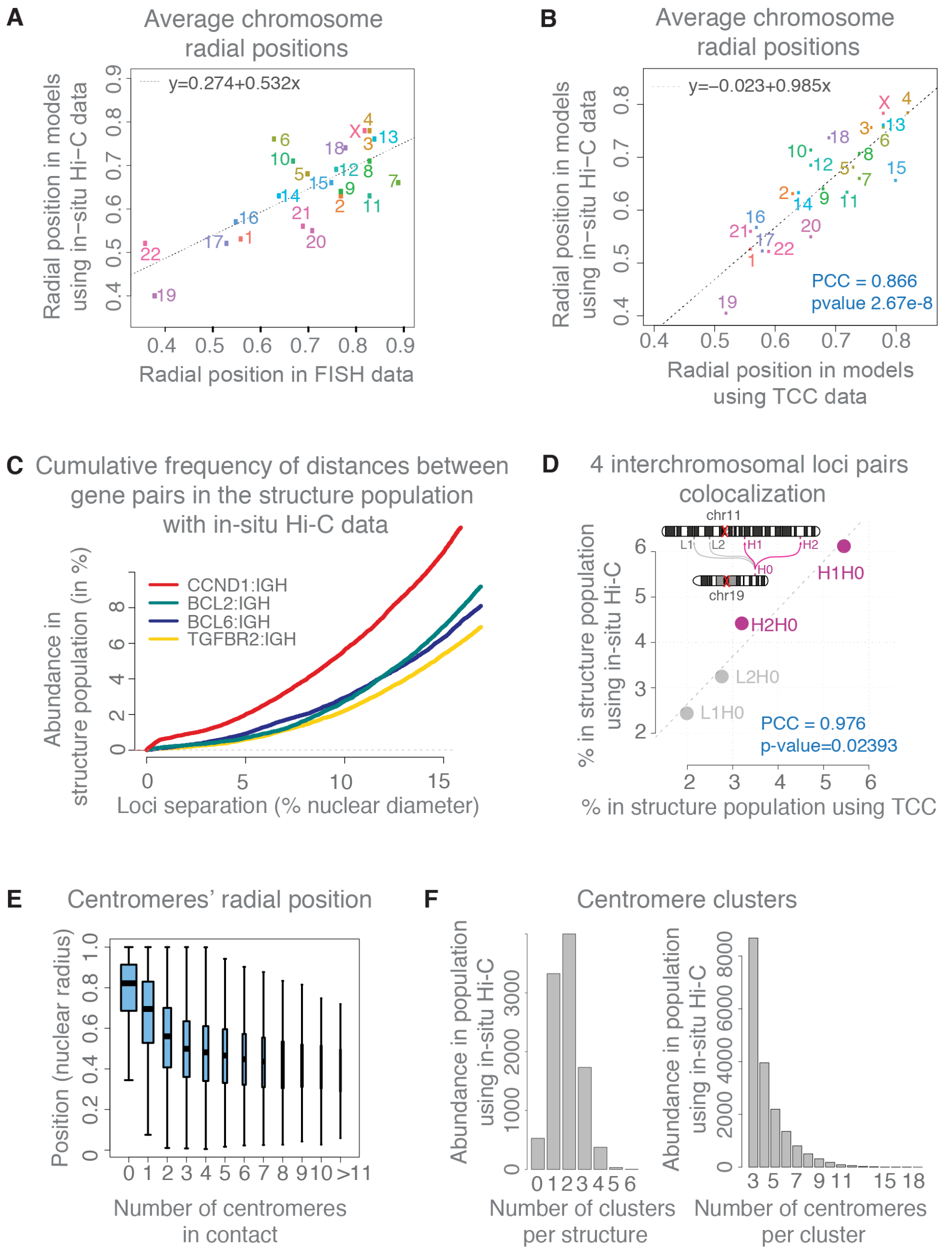


Figure S8 Results from the structure population generated with high-resolution in-situ Hi-C data (16). (**A**) Comparison of chromosomes' average radial positions between the structure population generated with the in-situ Hi-C data and FISH data (15). (**B**) Comparison of chromosomes' average radial positions in structure populations generated with in-situ Hi-C and TCC. (**C**) Cumulative frequency of distances between gene pairs in the structure population generated with the in-situ Hi-C data. The trend is very similar to that in the structure population generated with TCC data (**Fig. 3C**). (**D**) Comparison of colocalization counts for the 4 interchromosomal loci pairs between structure populations generated with in-situ Hi-C and TCC data (related to **Fig. 3B**). (**E**) Radial position of any centromere as a function of the number of other centromeres around it in the structure population generated with in-situ Hi-C data (related to **Fig. 6A**). (**F**) Histogram showing the abundance of centromere clusters per structure (left panel), and the cluster size distribution in the population generated with in-situ Hi-C data (related to **Figs. S4A** and **S4B**).

B.9. Figure S9

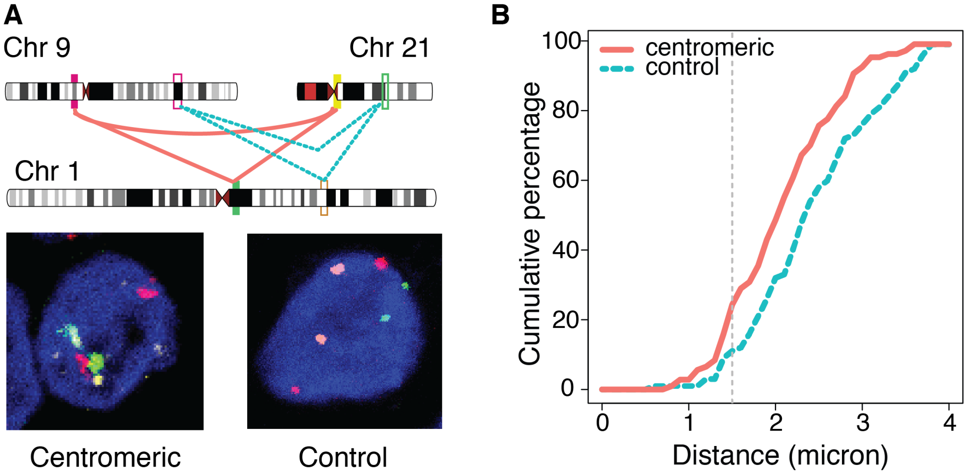


Figure S9 Centromeres as contact points for interchromosomal interactions (**A**).FISH experimental set up to verify that centromeric regions are bridging the interaction inter-chromosomally. A group of pericentromeric and distal regions from the respective centromere locations are referred to as the “centromeric” and “control” domains, respectively. (**B**) The cumulative frequency plots of the shortest triplet distance in each cell for the centromeric and control experiments. The triplet distance formed by pericentromeric regions is on average smaller than that formed by the control regions.

C. SI TABLES

C.1. Table S1

Bead	CHR	Start	End	Bead	CHR	Start	End	Bead	CHR	Start	End
1	chr1	10000	2292301	214	chr6	142711203	144961557	427	chr13	96864394	97782307
2	chr1	2292302	8036418	215	chr6	144961558	148916748	428	chr13	97782308	101296998
3	chr1	8036419	12563687	216	chr6	148916749	151974500	429	chr13	101296999	103761152
4	chr1	12563688	15763814	217	chr6	151974501	154585335	430	chr13	103761153	105998830
5	chr1	15763815	17884031	218	chr6	154585336	161620234	431	chr13	105998831	108634564
6	chr1	17884032	20903496	219	chr6	161620235	166688690	432	chr13	108634565	110562306
7	chr1	20903497	29327009	220	chr6	166688691	171055066	433	chr13	110562307	112162435
8	chr1	29327010	31287679	221	chr7	10000	3413224	434	chr13	112162436	112895211
9	chr1	31287680	33977837	222	chr7	3413225	5108971	435	chr13	112895212	115109877
10	chr1	33977838	35326034	223	chr7	5108972	8540956	436	chr14	1.90E+07	20867022
11	chr1	35326035	47176371	224	chr7	8540957	16588484	437	chr14	20867023	25169211
12	chr1	47176372	48308192	225	chr7	16588485	21242923	438	chr14	25169212	31133962
13	chr1	48308193	51059657	226	chr7	21242924	31078487	439	chr14	31133963	32093020
14	chr1	51059658	55878784	227	chr7	31078488	35626107	440	chr14	32093021	34174083
15	chr1	55878785	58791228	228	chr7	35626108	40381214	441	chr14	34174084	36303394
16	chr1	58791229	60196107	229	chr7	40381215	43531536	442	chr14	36303395	40748637
17	chr1	60196108	61946489	230	chr7	43531537	45498612	443	chr14	40748638	45076768
18	chr1	61946490	68011294	231	chr7	45498613	51348649	444	chr14	45076769	45928473
19	chr1	68011295	77903221	232	chr7	51348650	54921972	445	chr14	45928474	49802152
20	chr1	77903222	78431936	233	chr7	54921973	56446575	446	chr14	49802153	53529383
21	chr1	78431937	84609747	234	chr7	56446576	64149744	447	chr14	53529384	54837017
22	chr1	84609748	90330435	235	chr7	64149745	72488850	448	chr14	54837018	56261152
23	chr1	90330436	92049028	236	chr7	72488851	77692166	449	chr14	56261153	58421803
24	chr1	92049029	94472462	237	chr7	77692167	86205763	450	chr14	58421804	62700712
25	chr1	94472463	99948170	238	chr7	86205764	87999938	451	chr14	62700713	63854796
26	chr1	99948171	102180488	239	chr7	87999939	89736806	452	chr14	63854797	67873340
27	chr1	102180489	107381081	240	chr7	89736807	92825661	453	chr14	67873341	78074181
28	chr1	107381082	109336431	241	chr7	92825662	97567992	454	chr14	78074182	79471731
29	chr1	109336432	112250370	242	chr7	97567993	102990222	455	chr14	79471732	81156445
30	chr1	112250371	118315245	243	chr7	102990223	104319207	456	chr14	81156446	82054127
31	chr1	118315246	120471506	244	chr7	104319208	108463335	457	chr14	82054128	88005284
32	chr1	120471507	144357570	245	chr7	108463336	110500122	458	chr14	88005285	89296801
33	chr1	144357571	149681738	246	chr7	110500123	127131493	459	chr14	89296802	97678848
34	chr1	149681739	151984210	247	chr7	127131494	135822990	460	chr14	97678849	100107611
35	chr1	151984211	153375986	248	chr7	135822991	137479447	461	chr14	100107612	101396118
36	chr1	153375987	156934172	249	chr7	137479448	140720690	462	chr14	101396119	102089936
37	chr1	156934173	162817133	250	chr7	140720691	144617718	463	chr14	102089937	107289539
38	chr1	162817134	166717331	251	chr7	144617719	147661352	464	chr15	2.00E+07	30675812
39	chr1	166717332	169852119	252	chr7	147661353	148493161	465	chr15	30675813	31786353
40	chr1	169852120	171213428	253	chr7	148493162	152621180	466	chr15	31786354	35333070
41	chr1	171213429	175506417	254	chr7	152621181	154649569	467	chr15	35333071	38344599
42	chr1	175506418	178496366	255	chr7	154649570	157116672	468	chr15	38344600	40056098
43	chr1	178496367	185216905	256	chr7	157116673	159128662	469	chr15	40056099	45727969
44	chr1	185216906	193726191	257	chr8	10000	6279937	470	chr15	45727970	47861282
45	chr1	193726192	196900464	258	chr8	6279938	12380520	471	chr15	47861283	50521862
46	chr1	196900465	200399115	259	chr8	12380521	19074750	472	chr15	50521863	52791430
47	chr1	200399116	208145565	260	chr8	19074751	25915341	473	chr15	52791431	55427796
48	chr1	208145566	211357325	261	chr8	25915342	31270082	474	chr15	55427797	57769091
49	chr1	211357326	213422329	262	chr8	31270083	37630017	475	chr15	57769092	58618489
50	chr1	213422330	219192014	263	chr8	37630018	38794534	476	chr15	58618490	60104879
51	chr1	219192015	225795054	264	chr8	38794535	40972012	477	chr15	60104880	63292566
52	chr1	225795055	231620935	265	chr8	40972013	43214214	478	chr15	63292567	68678921
53	chr1	231620936	234446264	266	chr8	43214215	48204267	479	chr15	68678922	74463339
54	chr1	234446265	235497677	267	chr8	48204268	50091197	480	chr15	74463340	79065139
55	chr1	235497678	237072546	268	chr8	50091198	52659153	481	chr15	79065140	86446928
56	chr1	237072547	243273492	269	chr8	52659154	62795411	482	chr15	86446929	88694111
57	chr1	243273493	247243146	270	chr8	62795412	66539974	483	chr15	88694112	89819927
58	chr1	247243147	249240620	271	chr8	66539975	68478084	484	chr15	89819928	91663429
59	chr2	10000	7912767	272	chr8	68478085	70677272	485	chr15	91663430	93694574
60	chr2	7912768	13048721	273	chr8	70677273	75355548	486	chr15	93694575	96937954
61	chr2	13048722	20164452	274	chr8	75355549	80633630	487	chr15	96937955	98754370
62	chr2	20164453	21217470	275	chr8	80633631	82828326	488	chr15	98754371	102521391
63	chr2	21217471	23450099	276	chr8	82828327	94851654	489	chr16	60000	5200686
64	chr2	23450100	33814093	277	chr8	94851655	96305773	490	chr16	5200687	8853995
65	chr2	33814094	36961020	278	chr8	96305774	98397828	491	chr16	8853996	25449026
66	chr2	36961021	39833365	279	chr8	98397829	104500658	492	chr16	25449027	26975263
67	chr2	39833366	42267643	280	chr8	104500659	110956845	493	chr16	26975264	31209757
68	chr2	42267644	48597930	281	chr8	110956846	116197966	494	chr16	31209758	35285800
69	chr2	48597931	53705957	282	chr8	116197967	123597573	495	chr16	46385801	56544685
70	chr2	53705958	55918399	283	chr8	123597574	126993902	496	chr16	56544686	58972957
71	chr2	55918400	60147314	284	chr8	126993903	136061671	497	chr16	58972958	66554287
72	chr2	60147315	65814188	285	chr8	136061672	139628041	498	chr16	66554288	70511979
73	chr2	65814189	68423369	286	chr8	139628042	141159208	499	chr16	70511980	72955167
74	chr2	68423370	76111430	287	chr8	141159209	146304021	500	chr16	72955168	74389111
75	chr2	76111431	84826496	288	chr9	10000	4476839	501	chr16	74389112	75590712
76	chr2	84826497	89916825	289	chr9	4476840	7217320	502	chr16	75590713	83853261
77	chr2	89916826	92326170	290	chr9	7217321	14074501	503	chr16	83853262	90294752
78	chr2	95326171	103322354	291	chr9	14074502	15826949	504	chr17	0	9544714
79	chr2	103322355	105621881	292	chr9	15826950	18821757	505	chr17	9544715	15843693
80	chr2	105621882	107208384	293	chr9	18821758	22484837	506	chr17	15843694	20942908
81	chr2	107208385	108955285	294	chr9	22484838	26734942	507	chr17	20942909	25799526
82	chr2	108955286	114992986	295	chr9	26734943	27566443	508	chr17	25799527	26854162

83	chr2	114992987	118445358	296	chr9	27566444	32161104	509	chr17	26854163	30951213
84	chr2	118445359	122968871	297	chr9	32161105	33159751	510	chr17	30951214	33511358
85	chr2	122968872	127294885	298	chr9	33159752	38640739	511	chr17	33511359	49492303
86	chr2	127294886	132310754	299	chr9	38640740	69684409	512	chr17	49492304	52856107
87	chr2	132310755	134809692	300	chr9	69684410	71913384	513	chr17	52856108	55165101
88	chr2	134809693	137282921	301	chr9	71913385	81144361	514	chr17	55165102	58531744
89	chr2	137282922	143571697	302	chr9	81144362	86119319	515	chr17	58531745	61815825
90	chr2	143571698	145616066	303	chr9	86119320	89311621	516	chr17	61815826	62878155
91	chr2	145616067	147739715	304	chr9	89311622	91878774	517	chr17	62878156	64867407
92	chr2	147739716	150276212	305	chr9	91878775	101583010	518	chr17	64867408	66565213
93	chr2	150276213	151848121	306	chr9	101583011	103413042	519	chr17	66565214	70618570
94	chr2	151848122	153668628	307	chr9	103413043	106816661	520	chr17	70618571	72475813
95	chr2	153668629	157003426	308	chr9	106816662	114287491	521	chr17	72475814	81195208
96	chr2	157003427	162232014	309	chr9	114287492	116316334	522	chr18	10000	1019054
97	chr2	162232015	169032739	310	chr9	116316335	118036583	523	chr18	1019055	2350616
98	chr2	169032740	180118286	311	chr9	118036584	123308312	524	chr18	2350617	3806094
99	chr2	180118287	190457929	312	chr9	123308313	129320044	525	chr18	3806095	8586119
100	chr2	190457930	192180249	313	chr9	129320045	141153430	526	chr18	8586120	13580867
101	chr2	192180250	196907816	314	chr10	60000	3246049	527	chr18	13580868	19065138
102	chr2	196907817	198673366	315	chr10	3246050	8122613	528	chr18	19065139	24448814
103	chr2	198673367	200936963	316	chr10	8122614	11320890	529	chr18	24448815	28818512
104	chr2	200936964	204823242	317	chr10	11320891	15566693	530	chr18	28818513	29827991
105	chr2	204823243	206599012	318	chr10	15566694	26627102	531	chr18	29827992	32387856
106	chr2	206599013	209377842	319	chr10	26627103	27667212	532	chr18	32387857	34197321
107	chr2	209377843	213378921	320	chr10	27667213	30596559	533	chr18	34197322	36568817
108	chr2	213378922	216733381	321	chr10	30596560	32640253	534	chr18	36568818	39476079
109	chr2	216733382	220445309	322	chr10	32640254	43087186	535	chr18	39476080	42230018
110	chr2	220445310	230716288	323	chr10	43087187	45703881	536	chr18	42230019	43487703
111	chr2	230716289	234749669	324	chr10	45703882	52418690	537	chr18	43487704	48554641
112	chr2	234749670	243189372	325	chr10	52418691	63655455	538	chr18	48554642	49431976
113	chr3	60000	4343841	326	chr10	63655456	65613824	539	chr18	49431977	51586034
114	chr3	4343842	5328019	327	chr10	65613825	69594819	540	chr18	51586035	54346622
115	chr3	5328020	8716280	328	chr10	69594820	76794103	541	chr18	54346623	57104483
116	chr3	8716281	17817927	329	chr10	76794104	82691636	542	chr18	57104484	59845134
117	chr3	17817928	31046121	330	chr10	82691637	85365512	543	chr18	59845135	61220395
118	chr3	31046122	34369923	331	chr10	85365513	88232624	544	chr18	61220396	61989811
119	chr3	34369924	36545400	332	chr10	88232625	91782482	545	chr18	61989812	64607414
120	chr3	36545401	59858957	333	chr10	91782483	92521081	546	chr18	64607415	67415053
121	chr3	59858958	71822583	334	chr10	92521082	106234539	547	chr18	67415054	68269319
122	chr3	71822584	73206710	335	chr10	106234540	111385726	548	chr18	68269320	71424614
123	chr3	73206711	90504853	336	chr10	111385727	116812297	549	chr18	71424615	75291183
124	chr3	93504854	97472507	337	chr10	116812298	118537187	550	chr18	75291184	76720473
125	chr3	97472508	108489046	338	chr10	118537188	127551854	551	chr18	76720474	78017247
126	chr3	108489047	111173469	339	chr10	127551855	133668668	552	chr19	60000	8310778
127	chr3	111173470	114950403	340	chr10	133668669	135524746	553	chr19	8310779	9821388
128	chr3	114950404	121025742	341	chr11	60000	4067596	554	chr19	9821389	13446545
129	chr3	121025743	130033177	342	chr11	4067597	8261325	555	chr19	13446546	19947450
130	chr3	130033178	144119033	343	chr11	8261326	12501925	556	chr19	19947451	24631781
131	chr3	144119034	148685747	344	chr11	12501926	16751030	557	chr19	27731782	29742907
132	chr3	148685748	161185247	345	chr11	16751031	19620497	558	chr19	29742908	31412328
133	chr3	161185248	169211143	346	chr11	19620498	20945397	559	chr19	31412329	32945866
134	chr3	169211144	172415138	347	chr11	20945398	25679328	560	chr19	32945867	34922514
135	chr3	172415139	176709233	348	chr11	25679329	31536648	561	chr19	34922515	38789995
136	chr3	176709234	182559384	349	chr11	31536649	32896635	562	chr19	38789996	42942910
137	chr3	182559385	186823935	350	chr11	32896636	35587894	563	chr19	42942911	43847288
138	chr3	186823936	193216754	351	chr11	35587895	36693480	564	chr19	43847289	44996141
139	chr3	193216755	197962429	352	chr11	36693481	43050472	565	chr19	44996142	52213506
140	chr4	10000	18204216	353	chr11	43050473	46124382	566	chr19	52213507	59118982
141	chr4	18204217	24427544	354	chr11	46124383	48410901	567	chr20	60000	3769876
142	chr4	24427545	27232875	355	chr11	48410902	55662028	568	chr20	3769877	5976008
143	chr4	27232876	35962220	356	chr11	55662029	57151097	569	chr20	5976009	10717615
144	chr4	35962221	43098675	357	chr11	57151098	60005272	570	chr20	10717616	12822728
145	chr4	43098676	47099934	358	chr11	60005273	71465508	571	chr20	12822729	17683573
146	chr4	47099935	58329053	359	chr11	71465509	78230188	572	chr20	17683574	21722056
147	chr4	58329054	68146028	360	chr11	78230189	79520266	573	chr20	21722057	22634161
148	chr4	68146029	90096317	361	chr11	79520267	82604970	574	chr20	22634162	26104674
149	chr4	90096318	98817301	362	chr11	82604971	88461835	575	chr20	26104675	30158120
150	chr4	98817302	124843957	363	chr11	88461836	92675826	576	chr20	30158121	37821757
151	chr4	124843958	128573749	364	chr11	92675827	96278942	577	chr20	37821758	41927634
152	chr4	128573750	130280418	365	chr11	96278943	101898380	578	chr20	41927635	50292932
153	chr4	130280419	138953305	366	chr11	101898381	107490223	579	chr20	50292933	52122155
154	chr4	138953306	154686075	367	chr11	107490224	108782556	580	chr20	52122156	52679092
155	chr4	154686076	183627180	368	chr11	108782557	111025989	581	chr20	52679093	55027841
156	chr4	183627181	187392839	369	chr11	111025990	111980495	582	chr20	55027842	58785222
157	chr4	187392840	191044275	370	chr11	111980496	116657533	583	chr20	58785223	60479005
158	chr5	10000	2165693	371	chr11	116657534	121654506	584	chr20	60479006	62965519
159	chr5	2165694	5072780	372	chr11	121654507	124403456	585	chr21	9411193	15606100
160	chr5	5072781	23414161	373	chr11	124403457	126369943	586	chr21	15606101	17784672
161	chr5	23414162	31374782	374	chr11	126369944	131790733	587	chr21	17784673	24708071
162	chr5	31374783	43669756	375	chr11	131790734	133611049	588	chr21	24708072	26168115
163	chr5	43669757	46405640	376	chr11	133611050	134946515	589	chr21	26168116	27526368
164	chr5	49405641	54531362	377	chr12	60000	13156448	590	chr21	27526369	30037254
165	chr5	54531363	56556576	378	chr12	13156449	16324416	591	chr21	30037255	31337602
166	chr5	56556577	62242200	379	chr12	16324417	21586872	592	chr21	31337603	32595686
167	chr5	62242201	63964519	380	chr12	21586873	24670697	593	chr21	32595687	33586838
168	chr5	63964520	81583336	381	chr12	24670698	28130993	594	chr21	33586839	38927594
169	chr5	81583337	87940849	382	chr12	28130994	30959954	595	chr21	38927595	40876745
170	chr5	87940850	90009673	383	chr12	30959955	32949874	596	chr21	40876746	42747926
171	chr5	90009674	94801806	384	chr12	32949875	38939960	597	chr21	42747927	48119894

172	chr5	94801807	96487064	385	chr12	38939961	45442952	598	chr22	16050000	17779305
173	chr5	96487065	102956932	386	chr12	45442953	48849895	599	chr22	17779306	19819434
174	chr5	102956933	106586805	387	chr12	48849896	58107253	600	chr22	19819435	24548561
175	chr5	106586806	112774600	388	chr12	58107254	60236681	601	chr22	24548562	29546412
176	chr5	112774601	114319751	389	chr12	60236682	62355762	602	chr22	29546413	32261721
177	chr5	114319752	116517298	390	chr12	62355763	67847079	603	chr22	32261722	35829153
178	chr5	116517299	121317771	391	chr12	67847080	70171457	604	chr22	35829154	36889692
179	chr5	121317772	127610500	392	chr12	70171458	72525591	605	chr22	36889693	39148469
180	chr5	127610501	130646898	393	chr12	72525592	75913586	606	chr22	39148470	41692258
181	chr5	130646899	134337915	394	chr12	75913587	77257973	607	chr22	41692259	43741167
182	chr5	134337916	137126622	395	chr12	77257974	81328268	608	chr22	43741168	47287066
183	chr5	137126623	143379965	396	chr12	81328269	89756495	609	chr22	47287067	49927446
184	chr5	143379966	148580161	397	chr12	89756496	90569009	610	chr22	49927447	51244565
185	chr5	148580162	151211499	398	chr12	90569010	92183896	611	chrX	60000	2834822
186	chr5	151211500	156063324	399	chr12	92183897	97181864	612	chrX	2834823	6614453
187	chr5	156063325	159655616	400	chr12	97181865	101887789	613	chrX	6614454	10007596
188	chr5	159655617	166994349	401	chr12	101887790	105727373	614	chrX	10007597	25413047
189	chr5	166994350	170666438	402	chr12	105727374	108900562	615	chrX	25413048	28860834
190	chr5	170666439	180905259	403	chr12	108900563	114227242	616	chrX	28860835	31457611
191	chr6	60000	8069471	404	chr12	114227243	120409337	617	chrX	31457612	36730613
192	chr6	8069472	10210449	405	chr12	120409338	125962433	618	chrX	36730614	50185810
193	chr6	10210450	18423275	406	chr12	125962434	131371004	619	chrX	50185811	57807014
194	chr6	18423276	20159475	407	chr12	131371005	133841894	620	chrX	57807015	67500977
195	chr6	20159476	21548362	408	chr13	19020000	20293983	621	chrX	67500978	74620720
196	chr6	21548363	24230912	409	chr13	20293984	22047884	622	chrX	74620721	76466498
197	chr6	24230913	44458000	410	chr13	22047885	23991532	623	chrX	76466499	78556126
198	chr6	44458001	47975922	411	chr13	23991533	32406542	624	chrX	78556127	85514782
199	chr6	47975923	51622205	412	chr13	32406543	34674298	625	chrX	85514783	95616814
200	chr6	51622206	53819516	413	chr13	34674299	40150597	626	chrX	95616815	97343516
201	chr6	53819517	56331401	414	chr13	40150598	47614184	627	chrX	97343517	99918171
202	chr6	56331402	57126069	415	chr13	47614185	48459863	628	chrX	99918172	103464238
203	chr6	57126070	70111157	416	chr13	48459864	53477849	629	chrX	103464239	105726332
204	chr6	70111158	74655914	417	chr13	53477850	60296281	630	chrX	105726333	112341341
205	chr6	74655915	87677350	418	chr13	60296282	61103033	631	chrX	112341342	115150654
206	chr6	87677351	91323208	419	chr13	61103034	64599469	632	chrX	115150655	117035759
207	chr6	91323209	97041479	420	chr13	64599470	68601651	633	chrX	117035760	119946726
208	chr6	97041480	101316527	421	chr13	68601652	72957380	634	chrX	119946727	128504492
209	chr6	101316528	105182394	422	chr13	72957381	81351881	635	chrX	128504493	136300266
210	chr6	105182395	112572391	423	chr13	81351882	91452495	636	chrX	136300267	142251119
211	chr6	112572392	134195186	424	chr13	91452496	92347230	637	chrX	142251120	146863673
212	chr6	134195187	139963608	425	chr13	92347231	94916114	638	chrX	146863674	151979450
213	chr6	139963609	142711202	426	chr13	94916115	96864393	639	chrX	151979451	155260559

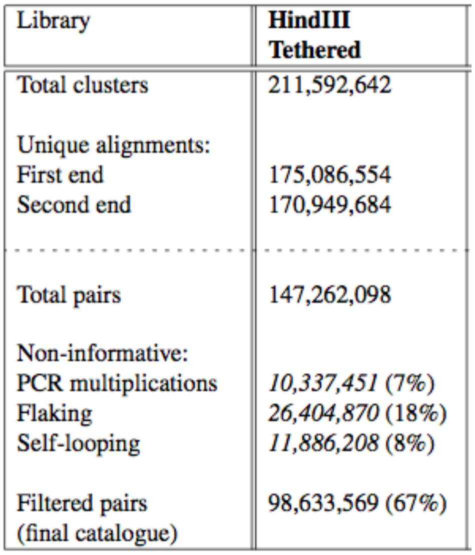
Table S1 List of 639 domains from the constrained hierarchical clustering method (see section A.4).

C.2. Table S2

Description	Symbol	Value	Unit
Number of spheres (domains)	$2N$	2x666	N/A
Number of structures	M	10,000	N/A
Stepwise optimization in probability threshold	Θ	{1, 0.4, 0.1, 0.01}	N/A
Nuclear occupancy	O_{nuc}	0.2	N/A
Radius of nucleus	R_{nuc}	5,000	IMP length
Harmonic constant	k	1	IMP unit
Sphere mass	$mass$	1	IMP mass
Temperatures, simulated annealing	T	Vary 300-500,000	IMP unit

Table S2 Modeling parameters used in this paper.

C.3. Table S3



Library	**HindIII Tethered**
Total clusters	211,592,642
Unique alignments:	
First end	175,086,554
Second end	170,949,684
Total pairs	147,262,098
Non-informative:	
PCR multiplications	*10,337,451* (7%)
Flaking	*26,404,870* (18%)
Self-looping	*11,886,208* (8%)
Filtered pairs (final catalogue)	98,633,569 (67%)

Table S3 The sequencing, alignment, pairing, and filtering statistics for the library. The italicized numbers for PCR multiplication, flaking, and self-looping mark the pairs that were filtered out of the initial catalogue in order to obtain the final catalogue. Numbers in parentheses are percentage values of each category compared to the “Total pairs” row. The last row (“Filtered pairs”) represents the catalogues that were used for all later analyses.

C.4. Table S4

Centromere Position	Gap in hg19 genome		Centromere bead domain		Centromere region for epigenetic	
	Start	End	Start	End	Start	End
chr1	121535434	124535433	120471507	144357570	116535421	144357571
chr2	92326171	95326170	92326171	95326170	87326158	100326183
chr3	90504854	93504853	90504854	93504853	85504841	98504866
chr4	49660117	52660116	47099935	58329053	44660104	58329054
chr5	46405641	49405640	46405641	49405640	41405628	54405653
chr6	58830167	61830165	57126070	70111157	53830154	70111158
chr7	58054331	61054330	56446576	64149744	53054318	66054343
chr8	43838887	46838886	43214215	48204267	38838874	51838899
chr9	47367680	50367678	38640740	69684409	38640739	69684410
chr10	39254936	42254934	32640254	43087186	32640253	47254947
chr11	51644206	54644204	48410902	55662028	46644193	59644217
chr12	34856694	37856693	32949875	38939960	29856681	42856706
chr13	16000000	18999999	0	19019999	15999999	24000012
chr14	16000000	18999999	0	18999999	15999999	24000012
chr15	17000000	19999999	0	19999999	16999999	25000012
chr16	35335801	38335800	35285801	46385800	30335788	46385801
chr17	22263006	25263005	20942909	25799526	17262993	30263018
chr18	15460898	18460897	13580868	19065138	10460885	23460910
chr19	24681782	27681781	24631782	27731781	19681769	32681794
chr20	26369569	29369568	26104675	30158120	21369556	34369581
chr21	11288129	14288128	9411193	15606100	6288116	19288141
chr22	13000000	15999999	0	16049999	12999999	21000012
chrX	58632012	61632011	57807015	67500977	53631999	67500978

Table S4 Positions in hg19 genome that flank centromere gaps, domain borders represented by centromeric beads in our model, and centromeric regions that were used for epigenetic data extraction of the recurrent structural patterns.

SI References

1. Kalhor R, Tjong H, Jayathilaka N, Alber F, & Chen L (2012) Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat Biotechnol* 30(1):90-98.
2. de Nooijer S, Wellink J, Mulder B, & Bisseling T (2009) Non-specific interactions are sufficient to explain the position of heterochromatic chromocenters and nucleoli in interphase nuclei. *Nucleic Acids Res* 37(11):3558-3568.
3. Tam R, Smith KP, & Lawrence JB (2004) The 4q subtelomere harboring the FSHD locus is specifically anchored with peripheral heterochromatin unlike most human telomeres. *The Journal of cell biology* 167(2):269-279.
4. Russel D*, et al.* (2012) Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol* 10(1):e1001244.
5. Alber F*, et al.* (2007) Determining the architectures of macromolecular assemblies. *Nature* 450(7170):683-694.
6. Kalhor R (2012) Exploring three-dimensional organization of the genome by mapping chromatin contacts and population modeling. Doctoral Dissertation (University of Southern California).
7. Imakaev M*, et al.* (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature methods* 9(10):999-1003.
8. Yaffe E & Tanay A (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet* 43(11):1059-1065.
9. Lieberman-Aiden E*, et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326(5950):289-293.
10. Kelley LA, Gardner SP, & Sutcliffe MJ (1996) An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies. *Protein engineering* 9(11):1063-1065.
11. Consortium EP*, et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57-74.
12. Schlesinger F, Smith AD, Gingeras TR, Hannon GJ, & Hodges E (2013) De novo DNA demethylation and noncoding transcription define active intergenic regulatory elements. *Genome research* 23(10):1601-1614.
13. Song Q*, et al.* (2013) A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PloS one* 8(12):e81148.
14. Cabili MN*, et al.* (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* 25(18):1915-1927.
15. Boyle S*, et al.* (2001) The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Hum Mol Genet* 10(3):211-219.
16. Rao SS*, et al.* (2014) A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* 159(7):1665-1680.
17. Solovei I & Cremer M (2010) 3D-FISH on cultured cells combined with immunostaining. *Methods in molecular biology* 659:117-126.
18. Bienko M*, et al.* (2013) A versatile genome-scale PCR-based pipeline for high-definition DNA FISH. *Nature methods* 10(2):122-124.
19. Iannuccelli E*, et al.* (2010) NEMO: a tool for analyzing gene and chromosome territory distributions from 3D-FISH experiments. *Bioinformatics* 26(5):696-697.
20. Cleveland DW, Mao Y, & Sullivan KF (2003) Centromeres and kinetochores: from epigenetics to mitotic checkpoint signaling. *Cell* 112(4):407-421.

21. Le Gros MA, McDermott G, & Larabell CA (2005) X-ray tomography of whole cells. *Current opinion in structural biology* 15(5):593-600.
22. McDermott G, Le Gros MA, Knoechel CG, Uchida M, & Larabell CA (2009) Soft X-ray tomography and cryogenic light microscopy: the cool combination in cellular imaging. *Trends in cell biology* 19(11):587-595.
23. Kremer JR, Mastronarde DN, & McIntosh JR (1996) Computer visualization of three-dimensional image data using IMOD. *Journal of structural biology* 116(1):71-76.
24. Mastronarde DN (2005) Automated electron microscope tomography using robust prediction of specimen movements. *Journal of structural biology* 152(1):36-51.
25. Stayman JW & Fessler JA (2004) Compensation for nonuniform resolution using penalized-likelihood reconstruction in space-variant imaging systems. *IEEE transactions on medical imaging* 23(3):269-284.
26. Weiss D*, et al.* (2001) Tomographic imaging of biological specimens with the cryo transmission X-ray microscope. *Nucl Instrum Meth A* 467:1308-1311.
27. Roix JJ, McQueen PG, Munson PJ, Parada LA, & Misteli T (2003) Spatial proximity of translocation-prone gene loci in human lymphomas. *Nat Genet* 34(3):287-291.