Supplementary Materials for

Title: *De Novo* Prediction of Human Chromosome Structures: Epigenetic Marking Patterns Encode Genome Architecture

Authors: Michele Di Pierro[1, a, *], Ryan R. Cheng[1, a], Erez Lieberman Aiden[1, 2], Peter G. Wolynes[1, 3, 4], José N. Onuchic[1, 4, *]

Affiliations:

[1] Center for Theoretical Biological Physics, Rice University, Houston, Texas 77005.

[2] The Center for Genome Architecture, Baylor College of Medicine, Houston, TX 77030.

[3] Department of Chemistry, Rice University, Houston, Texas 77005.

[4] Department of Physics & Astronomy, Rice University, Houston, Texas 77005.

[a] These authors contributed equally.
[*] Correspondence to: Michele.DiPierro@rice.edu or jonuchic@rice.edu

This PDF file includes:

Methods

Maximum Entropy Genomic Annotations from Biomarkers Associated to Structural Ensembles (MEGABASE)

Discretization of ChIP-Seq Data Tracks

Chromatin Immunoprecipitatin (ChIP-seq) data was downloaded from ENCODE (1) for the GM12878 cell line. This data is comprised of 95 different broad and narrow peak tracks, each of which probes the enhanced presence of an epigenetic mark or nuclear binding protein at a particular locus. A full list of the targets whose ChIP-Seq tracks were used in MEGABASE can be found in Table S1. The subset of experimental tracks that probe histone modifications, used in the reduced model can be found in Table S2.

For each chromosome, the ChIP-Seq signal is re-casted into the data tracks at 50 kb resolution, i.e., loci of 50 kb in size. This is performed by integrating (summing) the ChIP-Seq signal contained within each 50 kb locus for each experiment.

Subsequently, the integrated ChIP-seq signal for each 50 kb locus is assigned a discrete state ranging from 1 (low signal) to 20 (high signal). This is performed by creating a histogram for each experiment of the integrated signal for all of the 50 kb loci in the chromosomes of GM12878. All loci belonging to the top 5% of the distribution with the highest signal are assigned the highest signal state, i.e. 20. The remaining 19 signal states are defined by partitioning the remainder of the distribution linearly with respect to the signal strength; loci are assigned to those states according to their integrated signal.

MEGABASE database: structural annotations and biochemical assays

For each 50 kb locus of human lymphoblastoid cells (cell line GM12878) sub-compartment annotations from Rao *et al* (2) were aligned with the ChIP-Seq tracks discretized as previously described. In the cited reference, these sub-compartment annotations were obtained by clustering high resolution contact maps from Hi-C as to obtain 5 main patterns of interactions – A1, A2, B1, B2, and B3. A sixth compartment, B4 was identified by Rao and coworkers exclusively in chromosome 19. Due to its limited presence, B4 is treated as B3 in MEGABASE.

This allowed for the structural and biochemical state of each chromatin locus to be described using a state vector:

$$\vec{\sigma}(l) = (C(l),\ \mathrm{Exp}_1(l), \mathrm{Exp}_2(l),\ \ldots, \mathrm{Exp}_L(l))$$

where l denotes the locus, C refers to the sub-compartment annotation at that locus (A1, A2, B1, B2, or B3), and the components labeled by Exp with subscripts ranging from 1 to L denote the discrete signals for each ChIP-Seq experiment at the same locus, which are assigned 1-20

discrete signal states. The compartment annotation can be viewed as a proxy for the chromatin structural types (CST), so that the state vector can also be interpreted as:

$$\vec{\sigma}(l) = (CST(l),\ \mathrm{Exp}_1(l), \mathrm{Exp}_2(l),\ \ldots\ , \mathrm{Exp}_L(l))$$

The probability of observing a specific CST at a locus is correlated with adjacent segments, while noise is uncorrelated. Consequently, the inclusion of non-local information could reduce error in predicting chromatin types.

To build a richer database and to improve the quality of our results, we include in the state vector of locus *l* the biochemical state of the adjacent loci (i.e., *l*-2, *l*-1, *l*+1, *l*+2). Our database is therefore consisting of the set of state vectors:

$$\vec{\sigma}(l) = (CST(l),\ \mathrm{Exp}_1(l-2),...,\mathrm{Exp}_L(l-2),\mathrm{Exp}_1(l-1),...,\mathrm{Exp}_L(l-1),\mathrm{Exp}_1(l),...,$$
$$\mathrm{Exp}_L(l),\mathrm{Exp}_1(l+1),...,\mathrm{Exp}_L(l+1),\mathrm{Exp}_1(l+2),...,\mathrm{Exp}_L(l+2))$$

for each 50 kb locus of cell line GM12878. The total length of a state vector $\vec{\sigma}$ is $N = 476$

Construction of a probabilistic model: MEGABASE

To quantify the correlations between the structural annotations and the epigenetic marking patterns, we construct a probabilistic model for the collection of *M* state vectors, $\{\vec{\sigma}^{(s)}\}_{s=1...M}$, that comprise a database described in the previous section. The probability distribution resulting from our model, $P(\vec{\sigma})$, must reproduce the single-site and pairwise frequencies of the dataset, i.e., it must satisfy the marginalization conditions $\sum_{\substack{\sigma_k \\ k\neq i}} P(\vec{\sigma}) = f_i(\sigma_i)$ and $\sum_{\substack{\sigma_k \\ k\neq i,j}} P(\vec{\sigma}) = f_{ij}(\sigma_i,\sigma_j)$, where $f_i(\sigma_i)$ and $f_{ij}(\sigma_i,\sigma_j)$ denote the single-site and pairwise frequency, respectively, and *i* and *j* are indices of the state vector.

The most general solution to the problem outlined above, i.e. the least biased model satisfying the constraints, is obtained using the Maximum Entropy Principle (3). Such a model has the form of the Boltzmann distribution,

$$P(\vec{\sigma}) = \frac{1}{Z}\exp(-H(\vec{\sigma}))$$

where

$$H(\vec{\sigma}) = -\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} J_{ij}(\sigma_i,\sigma_j) - \sum_{i=1}^{N} h_i(\sigma_i)\,.$$

$P(\vec{\sigma})$ indicates the probability of observing the state vector $\vec{\sigma}$ at any given locus l. The J_{ij} interactions capture local pairwise correlations between epigenetic markers or between markers

and chromatin types, while the h_i parameters are related to the individual frequencies of chromatin types and markers. Here, the pairwise parameters are taken to be symmetric, i.e., $J_{ij}\left(\sigma_i,\sigma_j\right)=J_{ji}\left(\sigma_j,\sigma_i\right)$ similarly to a Hopfield neural network (4).

Training MEGABASE

The *M* state vectors for the odd numbered autosomes (i.e., 1, 3, 5, 7, etc.), $\{\vec{\sigma}^{(s)}\}_{s=1\ldots M}$, were used to train MEGABASE. We adopted the iterative approach of Ekeberg et al (5), which uses the pseudo-likelihood approximation of Besag (6) to construct a probabilistic model of sequences composed of discrete labels (i.e., chromatin types or amino acid sequences). Rather than maximize the likelihood of observing data, $\{\vec{\sigma}^{(s)}\}_{s=1\ldots M}$, which can be computationally intractable, one can maximize an approximate form of the likelihood of observing $\vec{\sigma}^{(s)}$ called a pseudo-likelihood:

$$P(\vec{\sigma}=\vec{\sigma}^{(s)})\approx\prod_{i=1}^{N}P(\sigma_i=\sigma_i^{(s)}\,|\,\sigma_j=\sigma_j^{(s)}\ \text{for all}\ j\neq i)\,.$$

For a pairwise Markov random field,

$$P(\sigma_i=\sigma_i^{(s)}\,|\,\sigma_j=\sigma_j^{(s)}\ \text{for all}\ j\neq i)=\frac{\exp\left(h_i(\sigma_i^{(s)})+\sum_{\substack{j=1\\j\neq i}}^{N}J_{ij}(\sigma_i^{(s)},\sigma_j^{(s)})\right)}{\sum_{a=1}^{25}\exp\left(h_i(a)+\sum_{\substack{j=1\\j\neq i}}^{N}J_{ij}(a,\sigma_j^{(s)})\right)}$$

where the normalization in the denominator is summed over the collection of 5 labels A1, A2, B1, B2, B3 as well as the 20 signal states assigned to ChIP-Seq data.

Maximizing the pseudo-likelihood of observing the collection of *M* training state vectors, $\{\vec{\sigma}^{(s)}\}_{s=1\ldots M}$, is equivalent to minimizing the negative of the log of the pseudo-likelihoods with respect to the parameters $\mathbf{J}$ and $\mathbf{h}$:

$$\ell_{PL}=-\frac{1}{M}\sum_{s=1}^{M}\sum_{i=1}^{N}\log P(\sigma_i=\sigma_i^{(s)}\,|\,\sigma_j=\sigma_j^{(s)}\ \text{for all}\ j\neq i)$$

In practice, an L2 regularization term is added to ℓ_{PL}:

$$R=\lambda_J\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\sum_{a}^{25}\sum_{b}^{25}J_{ij}(a,b)^2+\lambda_h\sum_{i=1}^{N}\sum_{a}^{25}h_i(a)^2$$

where the parameters $\lambda_J = \lambda_h = 0.01M$ were used. The model remained robust for a wide range of parameter values ranging from 0.0001-0.1. Hence, we minimize the object function, i.e., $\ell_{PL} + R$, with respect to the parameters $\mathbf{J}$ and $\mathbf{h}$ to find the model that best represents the training set of data:

$$\{\mathbf{J},\mathbf{h}\} = \arg\min_{\{\mathbf{J},\mathbf{h}\}}(\ell_{PL}(\mathbf{J},\mathbf{h}) + R_{L2}(\mathbf{J},\mathbf{h}))$$

This procedure is equivalent to training a recurrent neural network to encode information contained in a data set (4).

Prediction of chromatin structural types from ChIP-Seq data using MEGABASE

The inferred probabilistic model can be marginalized to predict the chromatin type for a given locus l when given the experimental ChIP-Seq measurements of loci $(l-2,\ l-1,\ l,\ l+1,\ l+2)$:

$$CST(l) = \arg\max P(CST \mid \mathrm{Exp}_{1,\ldots,L}(l-2,\ l-1,\ l,\ l+1,\ l+2))$$

This is equivalent to finding the chromatin type that minimizes the inferred energy function (Potts Model) for a given set of experimental measurements. For a given new input sequence of epigenetic marks one can then find the most probable sequence of corresponding compartment annotations.

The boundaries of compartments observed in Hi-C experiments are sometimes associated to visible boundaries in ChIP-Seq tracks; in other cases, however, no noticeable change is observed in biochemical assays when transitioning from one compartment to another. As illustrated by the example in Figure S1A, MEGABASE captures the transitions between compartments even in cases for which the ChIP-Seq tracks contain no obvious boundaries. One cannot simply attribute chromatin compartmentalization to the presence or absence of any single biomarker; it is necessary to employ a multivariate classifier such as MEGABASE in order to capture compartmentalization.

A comparison of the chromatin types predicted by MEGABASE and the sub-compartment annotations from ref. *(2)* is shown in the confusion matrices of Figure S1B, which are calculated for the state vectors of the test set. Additional measures of performance are shown in Figure S30. A comparison of the compartment prediction (A/B) between MEGABASE and the ref. *(2)* annotation is shown in the confusion matrix of Figure S1C for the test set. There is high quantitative agreement between the A/B compartment classifications, while a large number of mismatches exist between the predictions and the sub-compartment annotations. While we expect most chromatin belonging to a compartment to be of similar biochemical nature, it is possible—and even expected—that mismatches should exist because of the constraints introduced by connectivity along the DNA polymer. These mismatches do not affect the quality of the final 3D structural prediction, indicating that such mismatches may, indeed, not be errors

of the algorithm but instead loci where compartment annotation and chromatin type actually differ.

Network Analysis: Markers Strongly Associated With Compartmentalization

The success achieved in reliably predicting chromosome architecture indicates that our probabilistic model captures the essential features of epigenetic marks that are associated with compartmentalization. It is therefore useful to interrogate our model to gain insight into which of the biochemical markers are most strongly associated with each of the chromatin structural types. In order to disentangle the complex network of correlations between the 95 experiments contained in our database we use the concept of mutual information. It is important to point out that statistical models like MEGABASE cannot establish causality. Strictly speaking, in our analysis we cannot establish the direction of the causality link between chromatin structural types and epigenetic markers: the marking may occur before the phase separation and drive it or equally well, once compartments form through some other mechanism they become epigenetically marked. Nevertheless, histone modifications carry much of the information necessary to predict genome architecture. Previously reported theoretical (7) and experimental studies (8-10) have shown that epigenetic modifications do indeed lead to changes in chromatin organization, suggesting that the causality link is oriented from epigenetics to structure in the same way as the information flow of our computational pipeline.

The content of mutual information shared between compartment annotations and epigenetic markings can be quantified using the Kullback-Liebler divergence between the two probabilities $P(C,\mathrm{Exp}_\varepsilon(s))$ and $P(C)P(\mathrm{Exp}_\varepsilon(s))$. $P(C,\mathrm{Exp}_\varepsilon(s))$ is the joint probability that a certain locus belongs to chromatin type C and exhibits a biochemical marker ε with a signal of s and $P(C)P(\mathrm{Exp}_\varepsilon(s))$ is the same probability calculated from a null model in which the probability for that locus to belong to chromatin type C and the probability of observing there a biochemical marker ε with signal s are independent. The Kullback-Liebler divergence can be calculated using the trained neural network to calculate the probabilities above:

$$I_{C\varepsilon} = \sum_s P(C,\mathrm{Exp}_\varepsilon(s))\log\left(\frac{P(C,\mathrm{Exp}_\varepsilon(s))}{P(C)P(\mathrm{Exp}_\varepsilon(s))}\right)$$

The resulting correlated information content between chromatin types and ChIP-Seq experiments is shown in Figure S28A. It is immediately evident that certain biochemical markers share a high content of mutual information with chromatin structural types while others do not. According to our model, histone methylations HK36me3, H3K27me3, H3K4me1, and H4K20me1 and nuclear proteins EED, ZBED1, TRIM22, and HCFC1 carry most of the information associated with identifying the chromatin types. In contrast, we see that although compartment A for example has a very high content of acethylation H3K27ac that marker is a poor predictor owing to its modest mutual information value (Figure S28A).

We further quantify the relationship between chromatin types and experiments by calculating the joint probability $P(C, \text{High Exp}_\varepsilon)$ that a locus belongs to chromatin type C and has enhanced presence of biomarker ε. To do so, we marginalize over the high signal strengths, i.e.,

$$P(C,\text{High Exp}_\varepsilon) = \sum_{s \in \text{High Signal States}} P(C,\text{Exp}_\varepsilon(s)) .$$

The five highest discrete signal states are designated as high signal states, although we observed that the results are fairly insensitive to the details of this definition. Likewise, we can similarly marginalize the null model over the high signal states, i.e.,

$$P^{\text{High Signal}}(C,\varepsilon) = P(C) \cdot \sum_{s \in \text{High Signal States}} P(\text{Exp}_\varepsilon(s)) .$$

Finally, we calculate the log ratio,

$$S = \log\left(\frac{P(C, \text{High Exp}_\varepsilon)}{P^{\text{High Signal}}(C,\varepsilon)} \right),$$

which links the chromatin types to the enhancement of specific signals in the ChIP-Seq tracks. $P(C, \text{High Exp}_\varepsilon) > P^{\text{High Signal}}(C,\varepsilon)$ denotes an increased probability of observing chromatin type C when the biochemical marker is strongly present ($S > 0$). Likewise, $P(C,\text{High Exp}_\varepsilon) < P^{\text{High Signal}}(C,\varepsilon)$ denotes that is unlikely to observe C when marker ε is present ($S < 0$). The $S = 0$ condition describes the case where C and ε are independent.

In general, loci in the A type sub-compartments, A1 and A2, exhibit a much higher degree of marking over all, by most biomarkers. While exhibiting less marking, the B compartments do display characteristic signals. B1 is characterized by enhanced probability of finding methylation H3K27me3, and to a weaker extent, enhanced probability for the nuclear protein REST and the methylation H4K20me1. B2 is generally characterized by suppression of all biochemical markers, with the notable exception of the strongly enhanced probability of displaying methylation H3K9me3. Finally, B3 appears to exhibit suppression for all markers, and a particularly strong suppression of markers H3K36me3, H3K79me2, H3K4me1, and H3K9ac, which are associated with the A compartment. These results are shown for all chromatin types and experimental markers in Figure S28B.

A representative sample of the probability density functions representing the experimental signals for the biochemical markers is shown in Figure S28C. Compartments and sub-compartments are often characterized by vastly different average content of the biochemical markers. However, the variances of the distributions are large and, consequently, the distributions are always broadly overlapping. As already stressed, these overlaps show one cannot attribute chromatin compartmentalization to the presence or absence of any single biomarker and indicate the necessity of employing a multivariate probabilistic classifier such as MEGABASE.

Full and Reduced Model

From the analysis of the MEGABASE network, it is evident that histone modifications alone carry a great amount of information about chromatin structural types. As already mentioned, previously reported theoretical (7) and experimental studies (8-10) have shown that epigenetic modifications lead to changes in chromatin organization, suggesting that histone modifications alone may carry enough information to predict the global organization of the genome.

To investigate this question, we created a reduced model by training MEGABASE using only the patterns of histone modifications. We used all the experimental ChIP-Seq tracks available for the cell line GM12878; a list of the 11 tracks used to build the database for the reduced model is reported in Table S2.

As for the full MEGABASE model, we first trained the neural network on the training set composed of odd numbered chromosomes and then made chromatin types predictions for the test set composed of the even numbered chromosomes.

The sequences of chromatin types predicted by this reduced model turn out to be only marginally different from those obtained by the full data set of ChIP-Seq tracks (Figure S29). A comparison of the results in Figure S29 (reduced MEGABASE model) with the results in Figure S1 (full MEGABASE model) indicates a large overlap between the two annotations. The results obtained from the reduced MEGABASE network indicate that it is indeed possible to predict the global organization of chromosomes using exclusively information about histone modifications.

Alternative Models: MEGABASE A/B model and K-means Clustering

In 2009, analyzing the intra-chromosomal contact probability maps of the cell line GM06990, Lieberman *et al* (11) reported two compartments, A and B. Subsequently, analyzing the inter-chromosomal maps of the cell line GM12878, Rao *et al* (2) reported five sub-compartments (A1, A2, B1, B2, B3), plus a sixth very small one (B4). In designing MiChroM, in (12) we postulated for GM12878 the existence of a similar number of chromatin types. It possible and likely that other cell lines may exhibit a different number of compartments; it is also possible that a smaller number of discrete chromatin types may be able to generate a richer variety of long-range interaction patterns. To explore this latter possibility, we tested whether a model comprising only two types of chromatin (A and B) could reproduce the intra-chromosomal maps of GM12878. We aggregated the annotations from MEGABASE to produce an A/B classification and we simulated the conformational ensembles for the chromosomes using a simplified version of MiChroM (See the dedicated section for details about the full and simplified versions of MiChroM). We found the simulated Hi-C maps for the MEGABASE AB+MiChroM model to be slightly less accurate than the ones generated by the original five-type MEGABASE+MiChroM. This result indicates that at current resolution (50 kb) and on a single chromosome level, a simplified A/B model does indeed produce reasonable structures. On the other hand, it is clear that a model with five chromatin types generates more accurate 3D structures (See Figure S31). It is also likely that the advantage of using a 5-type model will increase when, by simulating the full nucleus of a cell, we will be able to examine the inter-chromosomal maps and compare those

to the experimental inter-chromosomal maps that were originally used to infer the existence of 6 sub-compartments.

In this manuscript, we used MEGABASE to cluster chromatin according to compartments; i.e., we used both structural and biochemical information to train a classifier for chromatin types. The existence of purely biochemical chromatin types has been previously reported (13). A final question we investigated is whether the purely biochemical clustering of chromatin produces feasible chromatin structural types. As previously discussed, higher levels of methylation and acetylation characterize type A chromatin, while type B chromatin is depleted in both. According to this observation, we clustered the ChIP-Seq tracks for GM12878 into two types using the K-means algorithm[1] and we used the simplified two-types version of MiChroM to generate conformational ensembles and contact maps for the resulting A/B annotation. We found that this clustering procedure correctly labels type B chromatin but frequently mislabels chromatin of type A (Figure S32E). As a result, the simulated contact maps are significantly deteriorated with respect to the ones obtained by the previously described A/B model (Figures S32A-D).

Interestingly, even using this purely biochemical clustering, remnants of compartmentalization are still present in the simulated chromosomes. Patterns in contact probabilities persist albeit exhibiting a significantly decreased level of segregation and much smoother transitions between compartments (See Figure S32A and S32B). This last result shows that chromosome architecture is very robust with respect to errors in the sequences of chromatin type and that the phase separation leading to compartmentalization can be induced by small differences in the biochemical nature along the DNA polymer.

Minimal Chromatin Model (MiChroM)

All molecular simulations methods in this manuscript are the same as in reference (12). For the derivations of all formulas and the tuning of all parameters please consult the cited reference.

Minimal Chromatin Model Energy Function

The reduced MiChroM energy function used in this manuscript is:

$$U_{MiChroM}(\vec{r}) = U_{HP}(\vec{r}) + \sum_{\substack{k \geq l \\ k,l \in \text{Types}}} \alpha_{kl} \sum_{\substack{i \in \{\text{Loci of Type } k\} \\ j \in \{\text{Loci of Type } l\}}} f(r_{ij}) + \sum_{d=3}^{500} \gamma(d) \sum_{i} f(r_{i,\,i+d})$$

[1] Using Euclidian distance as a metric

The full energy function of MiChroM is:

$$U_{MiChroM}(\vec{r}) = U_{HP}(\vec{r}) + \sum_{\substack{k \geq l \\ k,l \in \text{Types}}} \alpha_{kl} \sum_{\substack{i \in \{\text{Loci of Type } k\} \\ j \in \{\text{Loci of Type } l\}}} f(r_{ij}) + \chi \cdot \sum_{(i,j) \in \{\text{Loops Sites}\}} f(r_{ij}) + \sum_{d=3}^{500} \gamma(d) \sum_{i} f(r_{i,\,i+d})$$

This last version of the energy function is only used in simulation marked "with Loops" and for comparison purposes.

Homopolymer Model

The homo-polymer potential models a generic polymer. Each bead of in the polymer represents a genomic segment spanning 50 Kb of DNA.
The homo-polymer potential $U_{HP}(\vec{r})$ consists of the following five terms, U_{FENE}, U_{Angle}, U_{hc}, U_{sc} and U_c (14).

$$U_{HP}(\vec{r}) = \sum_{i \in \{\text{Loci}\}} U_{FENE}(r_{i,i+1}) + \sum_{i \in \{\text{Loci}\}} U_{hc}(r_{i,i+1}) + \sum_{i \in \{Angles\}} U_{Angle}(\theta_i) + \sum_{\substack{i,j \in \{\text{Loci}\} \\ j > i+2}} U_{sc}(r_{i,j}) + \sum_{i \in \{\text{Loci}\}} U_c(\vec{r}_i)$$

U_{FENE} (Finite Extensible Nonlinear Elastic potential) is the bonding potential applied between two consecutive monomers:

$$U_{FENE}(r_{i,j}) = \begin{cases} -\frac{1}{2} k_b R_0^2 \ln\left[1 - \left(\frac{r_{i,j}}{R_0}\right)^2\right] & \text{for } r_{i,j} \leq R_0 \\ 0 & \text{for } r_{i,j} > R_0 \end{cases}$$

A hard-core repulsive potential

$$U_{hc}(r_{i,j}) = \begin{cases} 4\varepsilon\left[\left(\frac{\sigma}{r_{i,j}}\right)^{12} - \left(\frac{\sigma}{r_{i,j}}\right)^{6} + \frac{1}{4}\right] & \text{for } r_{i,j} \leq \sigma 2^{\frac{1}{6}} \\ 0 & \text{for } r_{i,j} > \sigma 2^{\frac{1}{6}} \end{cases}$$

is added between bonded monomers to avoid overlap.
A three-body term is applied to three consecutive monomers in the following form

$$U_{Angle}(\theta_i) = k_a\left[1-\cos(\theta_i-\theta_0)\right]$$

where θ_i is the angle defined by the two vectors $\vec{r}_{i,i+1}$ and $\vec{r}_{i,i-1}$.

All non-bonded pairs interacts through a soft-core repulsive interaction (15)

$$U_{sc}(r_{i,j}) = \begin{cases} \frac{1}{2}E_{cut}\left[1+\tanh\left(\frac{2U_{LJ}(r_{i,j})}{E_{cut}}-1\right)\right] & r_{i,j} < r_0 \\ U_{LJ}(r_{i,j}) & r_0 \le r_{i,j} \le \sigma 2^{\frac{1}{6}} \\ 0 & r_{i,j} > \sigma 2^{\frac{1}{6}} \end{cases}$$

The Lennard-Jones potential $U_{LJ}(r_{i,j}) = 4\varepsilon\left[\left(\frac{\sigma}{r_{i,j}}\right)^{12}-\left(\frac{\sigma}{r_{i,j}}\right)^{6}+\frac{1}{4}\right]$ is capped off at a finite distance, allowing for chain crossing at finite energetic cost. r_0 is chosen as the distance at which $U_{LJ}(r_{i,j}) = \frac{1}{2}E_{cut}$.

The potential U_c restricts the chromosome in a spherical region, whose size is chosen to enforce a volume fraction of 0.1— corresponding to the experimentally determined density of chromatin[2] (0.012 bp/nm^3) (16). The spherical wall is included to mimic a similar confinement experienced by chromosomes inside the cell. Each monomer i of the chromosome interacts with its nearest point on the wall $\vec{r}_{np}$ through the potential $U_{hc}(r_{i,np})$.

Probability of Crosslinking

The probability of crosslinking is modeled as the function:

[2] After completion of all the simulation, we found that it was useful to recalibrate the length scale of our simulations by using available FISH data. This data being specific to human lymphoblastoid cells have greater precision than do the original data used for the set up. Using this recalibration we found that one unit of length σ in our model corresponds to 0.165μm, meaning that one bead has a radius of about 825Å. With this calibration we also found the chromatin density in our simulations to be 0.002 bp/nm^3. This density is only about 6 times smaller than the chromatin density reported in Ref: 16. Rosa A & Everaers R (2008) Structure and Dynamics of Interphase Chromosomes. *Plos Comput Biol* 4(8)., which in this case is excellent agreement considering the variability in size between cell types and even within a homogeneous cell population. Using this calibration we observe that chromosome territories are about 2-3 μm across, once again consistent with what is found in literature: 17. Cremer T & Cremer C (2001) Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet* 2(4):292-301.

$$f\left(r_{ij}\right)=\frac{1}{2}\left(1+\tanh\left[\mu\left(r_c-r_{ij}\right)\right]\right)$$

The parameters adjusted for the contact maps of GM12878 B-lymphoblastoid cells in dataset GSE63525 (2) are $\mu = 3.22$ and $r_c = 1.78$.

MiChroM Parameter Set

We consider 5 chromatin types A1, A2, B1, B2, B3 plus a non-specific type NA. Chromatin compartment B4 was detected only in chromosome 19 by Rao *et al* (2). MiChroM treats B4 as B3.

The parameters α 's governing the type-to-type interactions are:

	A1	A2	B1	B2	B3	NA
A1	-0.268028	-0.274604	-0.262513	-0.258880	-0.266760	-0.225646
A2	-0.274604	-0.299261	-0.286952	-0.281154	-0.301320	-0.245080
B1	-0.262513	-0.286952	-0.342020	-0.321726	-0.336630	-0.209919
B2	-0.258880	-0.281154	-0.321726	-0.330443	-0.329350	-0.282536
B3	-0.266760	-0.301320	-0.336630	-0.329350	-0.341230	-0.349490
NA	-0.225646	-0.245080	-0.209919	-0.282536	-0.349490	-0.255994

The parameter χ governing the loop interactions is equal to -1.612990.

The parameters α 's governing the type-to-type interactions for the reduced MiChroM A/B model are found collapsing the matrix above, resulting in:

	A	B	NA
A	-0.280631	-0.276263	-0.235363
B	-0.276263	-0.333566	-0.280648
NA	-0.235363	-0.280648	-0.255994

Ideal Chromosome Term

The Ideal Chromosome Potential is:

$$\gamma(d) = \frac{\gamma_1}{\log(d)} + \frac{\gamma_2}{d} + \frac{\gamma_3}{d^2}$$

with parameters $\gamma_1 = -0.030$, $\gamma_2 = -0.351$, $\gamma_3 = -3.727$.

Molecular Dynamics Simulations

First, we condense the polymer from an extended configuration initialized as a straight line. To condense the polymer, we perform 2x10^4 step MD simulation under the potential energy function

$$U_{Eq}(\vec{r}) = \sum_{i \in \{\text{Loci}\}} U_{FENE}(r_{i,i+1}) + \sum_{i \in \{\text{Loci}\}} U_{hc}(r_{i,i+1}) + \sum_{i \in \{Angles\}} U_{Angle}(\theta_i) + \sum_{\substack{i,j \in \{\text{Loci}\} \\ j>i+2}} U_{sc}(r_{i,j}) + \frac{1}{2} K_{Eq}\left(R_g - R_g^0\right)^2$$

which is the homopolymer potential with an additional harmonic bias on the radius of gyration R_g. We set $K_{Eq} = 200\varepsilon/\sigma^2$ and $R_g^0 = 1$. The spherical confinement is not present in this phase. Then, from this condensed polymer configuration, we perform 20 million steps equilibration with the potential energy function $U_{HP}(\vec{r})$, which also includes the confinement potential. In each chromosome, the radius of the confinement potential was set to reproduce the experimentally determined density of chromatin (16) corresponding to a volume ratio of 0.1.
All chromosome simulations were performed using the molecular dynamics package LAMMPS (18). Reduced units were used during the simulation, with

$$k_a = 2\varepsilon \qquad k_b = \frac{30\varepsilon}{\sigma^2} \qquad E_{cut} = 4\varepsilon \qquad \varepsilon = K_B T$$
$$R_0 = 1.5\sigma \qquad \sigma = 1 \qquad \theta_0 = \pi$$

Simulations were maintained at a constant temperature $T = 1.0$ via Langevin dynamics with a damping coefficient of 10.0τ, where τ is the time unit. A time step $\Delta t = 0.01\tau$ was used for all the simulations. All MD simulations were run until convergence as tested by verifying that

different replicas reported similar results. Given the different sizes of different chromosomes, this resulted in different total simulation lengths for the chromosomes (always in the order of 10^8 steps for all chromosomes). A short equilibration (10^6 time steps) at high temperature (T=10) was performed in each simulation before starting sampling chromosome conformations.

Experimental Data Sets

Contact Probabilities, Loop Locations, and Chromatin Compartment Annotations

Hi-C contact maps, chromatin compartment annotations and loops locations for GM12878 B-lymphoblastoid cells were obtained from Rao *et al.* (2). The Gene Expression Omnibus (GEO) accession number for the data sets is GSE63525 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63525).

Hi-C maps were balanced using the Knight and Ruiz (KR) vectors reported by the authors. Then, we extract a contact probability matrix $P^{\exp}$ from the stochastic matrix C by dividing each row i by its maximum entry, typically $C_{i,i+1}$.

ChIP-Seq Data

Chromatin Immunoprecipitatin (ChIP-seq) data was downloaded from ENCODE (1) for the GM12878 cell line. This data is comprised of 95 different broad and narrow peak tracks, each of which probes the enhanced presence of an epigenetic mark or nuclear binding protein at a particular locus. A full list of the targets whose ChIP-Seq tracks were used in MEGABASE can be found in Table S1. The subset of experimental tracks that probe histone modifications, used in the reduced model can be found in Table S2.

FISH Data

We compared the predictions of MEGABASE+MiChroM with two published sets of Fluorescence In Situ Hybridization (FISH) experiments. The first study published by Lieberman-Aiden et al (11) was performed on the GM06990 cell line using the hg18 assembly—Gene Expression Omnibus (GEO) accession number GSE18199. The genomic locations of the FISH probes used were mapped to the positions of GM12878 (hg19 assembly) using the Liftover software (19). The second FISH study was published by Rao et al (2) for the GM12878 cell line (GEO accession number GSE63525).

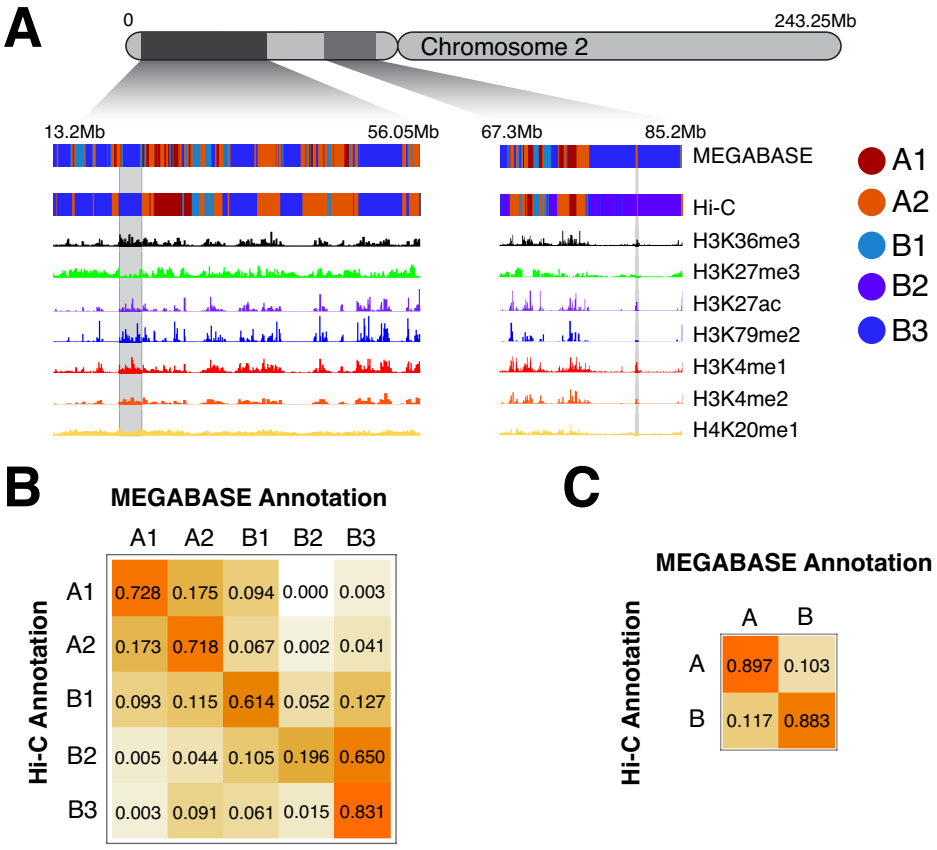


Figure S1

(A) For two representative regions of chromosome 2, we compare the sequence of chromatin types obtained from MEGABASE and the Hi-C compartment annotations. As illustrated by the region highlighted in gray on the left, MEGABASE captures the sharp changes in epigenetic markings sometimes present at the boundaries of contiguous regions of chromatin types (left boundary) while also correctly predicting less obvious transitions from one chromatin type to another (right boundary). On the right, the highlighted region shows how MEGABASE can resolve very small (50-100 kb) segments of a specific chromatin types.

(B and C) As shown by the confusion matrix[3] in figure annotations from MEGABASE largely overlap the compartments annotations from Hi-C reported in ref. (2). While we expect most

[3] For each of Hi-C compartment annotations, rows show how likely MEGABASE classifies it as each one of the 5 types.

chromatin belonging to a compartment to be of similar biochemical nature, it is possible—and even expected—that mismatches should exist because of the constraints introduced by sequence contiguity along the DNA polymer.

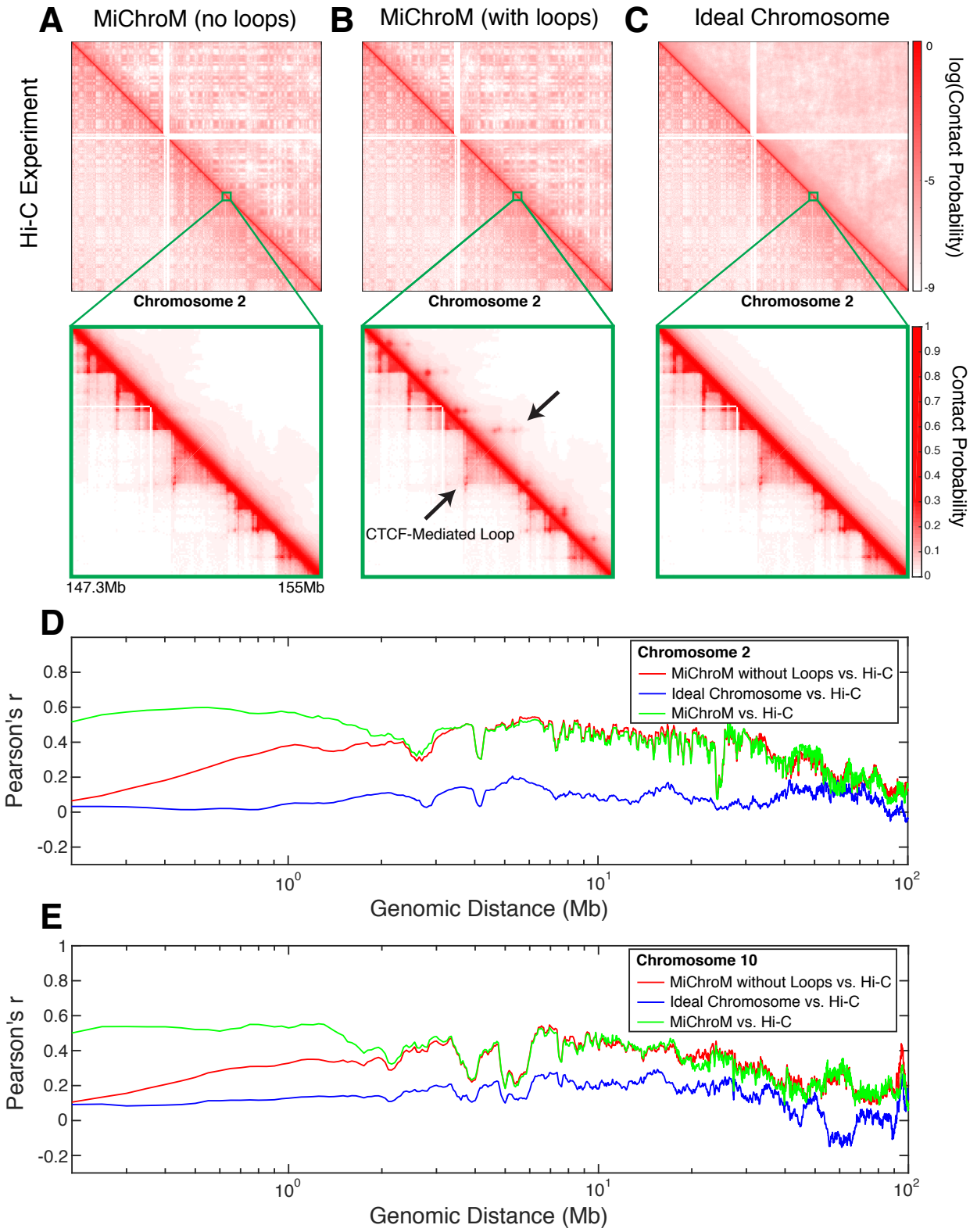


Figure S2

Chromatin loops are created when two loci, typically at a genomic distance extending from hundreds of kilobases (kb) to a few megabases (Mb) of DNA, form a particularly strong contact. These strong contacts manifest themselves as local peaks in the contact probability maps

obtained by Hi-C (2). The majority of chromatin loops are associated with the presence of CCCTC-binding factor (CTCF) and cohesin. By binding to specific sequences, CTCF defines the contact points for cohesin-mediated looping interactions. It has been suggested that cohesin is loaded on DNA elsewhere and then recruited to the loop anchors by a process of extrusion of the 10 nm fiber (20, 21).

In order to highlight the relationship between chromatin types and compartmentalization, we use the MiChroM Hamiltonian omitting the term in that energy function that models the CTCF-mediated looping interactions. These looping interactions seem to arise from a distinct process from compartmentalization and omitting such looping interactions does not affect the large-scale architecture of chromosomes (12, 22, 23). For completeness, we also performed simulations using the full MiChroM Hamiltonian including looping interactions. While including the effect of looping interactions (i.e. using the full MiChroM Hamiltonian) clearly produces more accurate chromosome structural ensembles, using this model would come with the tradeoff of necessitating further experimental input beyond genome sequence and ChIP-Seq since it is not yet clear how to predict all the locations of interacting loop anchors. To obviate this issue, in the full simulations we used the loops anchors annotations from Hi-C already found in (2). When comparing the results from the two versions of MiChroM, it becomes evident that compartmentalization remains unaffected: the effect of looping interactions is limited to genomic distances up to only few megabases.

(A) For chromosome 2 of GM12878 we show a comparison between the experimental contacts map in ref. (2) (lower diagonal region of the matrix) and the map generated by using MiChroM without CTCF-mediated looping interactions (upper diagonal region).

(B) For chromosome 2 of GM12878 we show a comparison between the experimental contacts map in ref. (2) (lower diagonal region of the matrix) and the map generated by using the full MiChroM Hamiltonian, i.e. including CTCF-mediated looping interactions (upper diagonal region). In the magnified inset the local probability peaks generated by the CTCF-mediated looping interactions are clearly visible in both experimental and predicted maps.

(C) For chromosome 2 of GM12878 we show a comparison between the experimental contacts map in ref. (2) (lower diagonal region of the matrix) and the map generated by using the Ideal Chromosome potential (upper diagonal region). The Ideal Chromosome potential models the local order in chromatin and is a translationally invariant energy term. The Ideal Chromosome reproduces the probability of observing a contact as a function of the genomic distance separating the two loci forming the contact.

(D) Pearson's correlation as a function of the genomic distance between experimental the contacts map of chromosome 2 of GM12878 in ref. (2) and the contact map generated by using MEGABASE. Contact maps generated using the full MiChroM Hamiltonian, MiChroM without CTCF-mediated looping interactions, and the Ideal Chromosome are shown in green, red, and blue respectively. The full MiChroM Hamiltonian generates better results up to distances of about 2 Mb. For all genomic distances exceeding 2 Mb the effect of omitted looping interactions is marginal and the maps produced by using MiChroM with and without CTCF-mediated looping interactions are very similar. Maps produced using the Ideal Chromosome potential are

shown as reference. While producing correct average contact probabilities, these maps never correlate with the experimental ones.

(E) The analysis in panel D is repeated here for chromosome 10. Results are consistent.

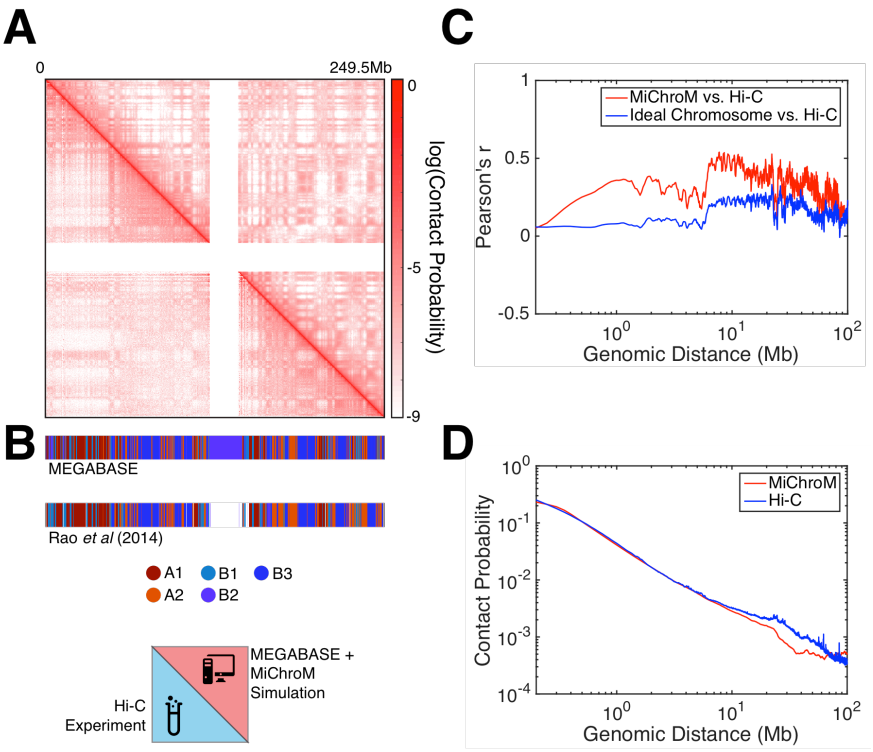


Figure S3
For all autosomal chromosomes MEGABASE+MiChroM generates conformational ensembles that accurately predict the results of DNA-DNA ligation assays.

(A) Contact map of chromosomes 1 (belonging to the training set) represented in log scale. Upper diagonal region shows the predicted map; MEGABASE+MiChroM generated this map *in silico* from ChIP-Seq input. The lower diagonal region shows maps from Hi-C (2). The quality of the predicted contact map is high, as shown by the symmetry of the map. Pearson's correlation between the two datasets is shown in Table S3.

(B) Comparison between the compartment annotations obtained by Hi-C (2) and MEGABASE structural type annotations.

(C) Pearson's correlation between experimental and simulated contact maps as a function of the genomic distance. MEGABASE+MiChroM generates contact maps that are well correlated with the experimental ones for distances exceeding the hundreds of Mb. As term of comparison, we show in blue the correlation between experimental maps and maps obtained using a homo-polymeric model including the Ideal Chromosome Potential (i.e. MiChroM without Type-to-Type interactions).

(D) The probability of contacts as a function of genomic distance in both measured and predicted maps.

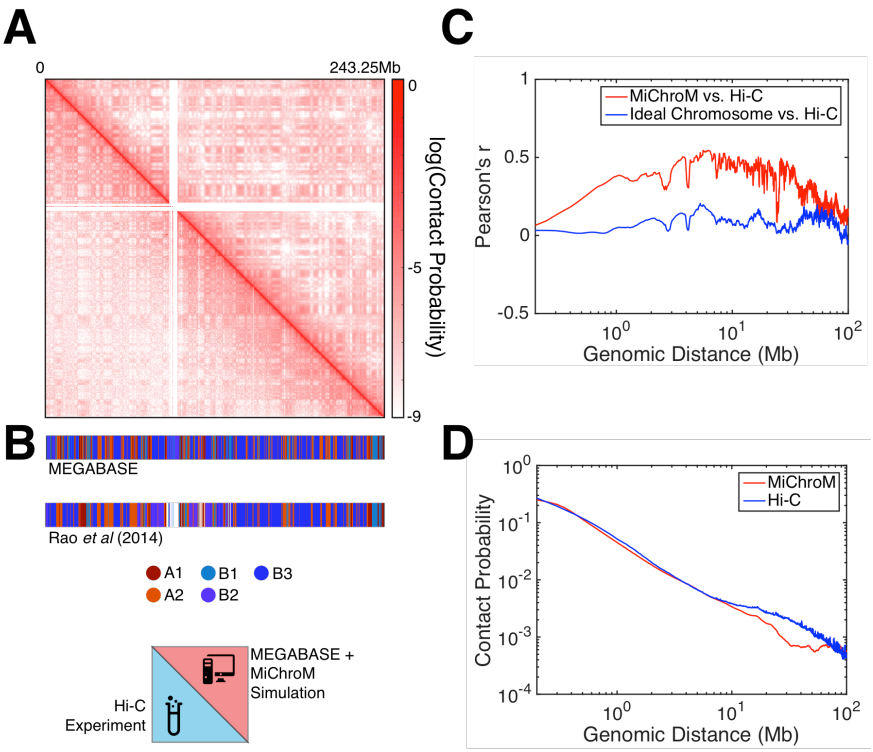


Figure S4
For all autosomal chromosomes MEGABASE+MiChroM generates conformational ensembles that accurately predict the results of DNA-DNA ligation assays.

(A) Contact map of chromosomes 2 (belonging to the test set) represented in log scale. Upper diagonal region shows the predicted map; MEGABASE+MiChroM generated this map *in silico* from ChIP-Seq input. The lower diagonal region shows maps from Hi-C (2). The quality of the predicted contact map is high, as shown by the symmetry of the map. Pearson's correlation between the two datasets is shown in Table S3.

(B) Comparison between the compartment annotations obtained by Hi-C (2) and MEGABASE structural type annotations.

(C) Pearson's correlation between experimental and simulated contact maps as a function of the genomic distance. MEGABASE+MiChroM generates contact maps that are well correlated with the experimental ones for distances exceeding the hundreds of Mb. As term of comparison, we show in blue the correlation between experimental maps and maps obtained using a homo-polymeric model including the Ideal Chromosome Potential (i.e. MiChroM without Type-to-Type interactions).

(D) The probability of contacts as a function of genomic distance in both measured and predicted maps.

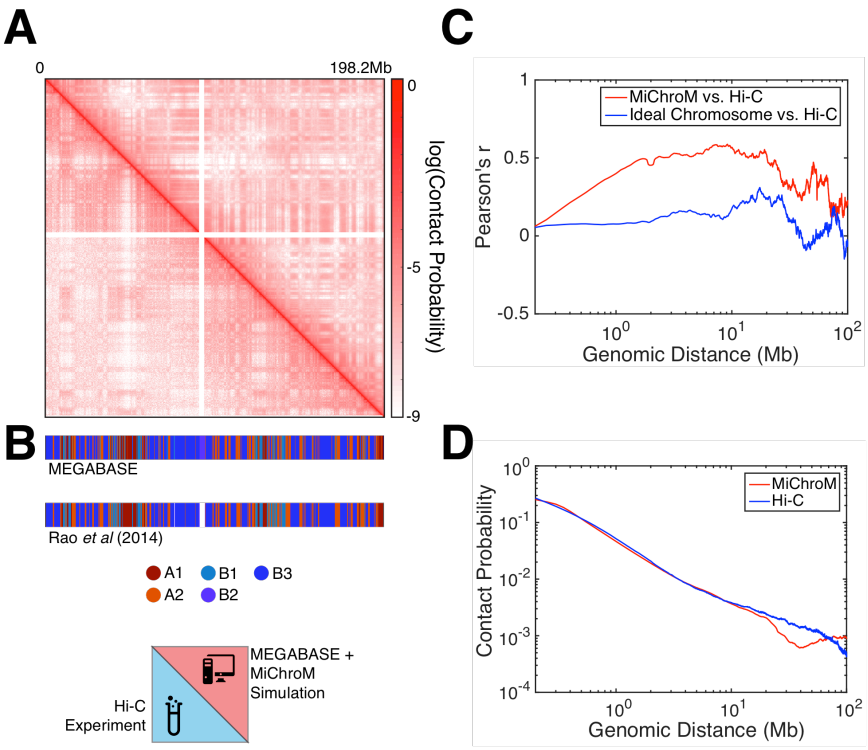


Figure S5

For all autosomal chromosomes MEGABASE+MiChroM generates conformational ensembles that accurately predict the results of DNA-DNA ligation assays.

(A) Contact map of chromosomes 3 (belonging to the training set) represented in log scale. Upper diagonal region shows the predicted map; MEGABASE+MiChroM generated this map *in silico* from ChIP-Seq input. The lower diagonal region shows maps from Hi-C (2). The quality of the predicted contact map is high, as shown by the symmetry of the map. Pearson's correlation between the two datasets is shown in Table S3.

(B) Comparison between the compartment annotations obtained by Hi-C (2) and MEGABASE structural type annotations.

(C) Pearson's correlation between experimental and simulated contact maps as a function of the genomic distance. MEGABASE+MiChroM generates contact maps that are well correlated with the experimental ones for distances exceeding the hundreds of Mb. As term of comparison, we show in blue the correlation between experimental maps and maps obtained using a homo-polymeric model including the Ideal Chromosome Potential (i.e. MiChroM without Type-to-Type interactions).

(D) The probability of contacts as a function of genomic distance in both measured and predicted maps.

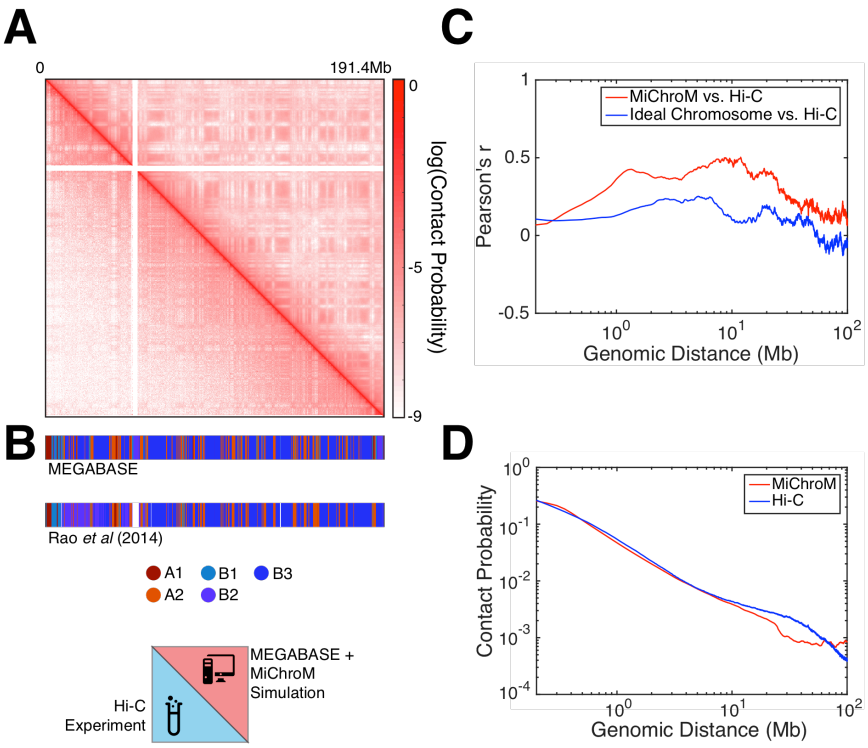


Figure S6

For all autosomal chromosomes MEGABASE+MiChroM generates conformational ensembles that accurately predict the results of DNA-DNA ligation assays.

(A) Contact map of chromosomes 4 (belonging to the test set) represented in log scale. Upper diagonal region shows the predicted map; MEGABASE+MiChroM generated this map *in silico* from ChIP-Seq input. The lower diagonal region shows maps from Hi-C (2). The quality of the predicted contact map is high, as shown by the symmetry of the map. Pearson's correlation between the two datasets is shown in Table S3.

(B) Comparison between the compartment annotations obtained by Hi-C (2) and MEGABASE structural type annotations.

(C) Pearson's correlation between experimental and simulated contact maps as a function of the genomic distance. MEGABASE+MiChroM generates contact maps that are well correlated with the experimental ones for distances exceeding the hundreds of Mb. As term of comparison, we show in blue the correlation between experimental maps and maps obtained using a homo-polymeric model including the Ideal Chromosome Potential (i.e. MiChroM without Type-to-Type interactions).

(D) The probability of contacts as a function of genomic distance in both measured and predicted maps.

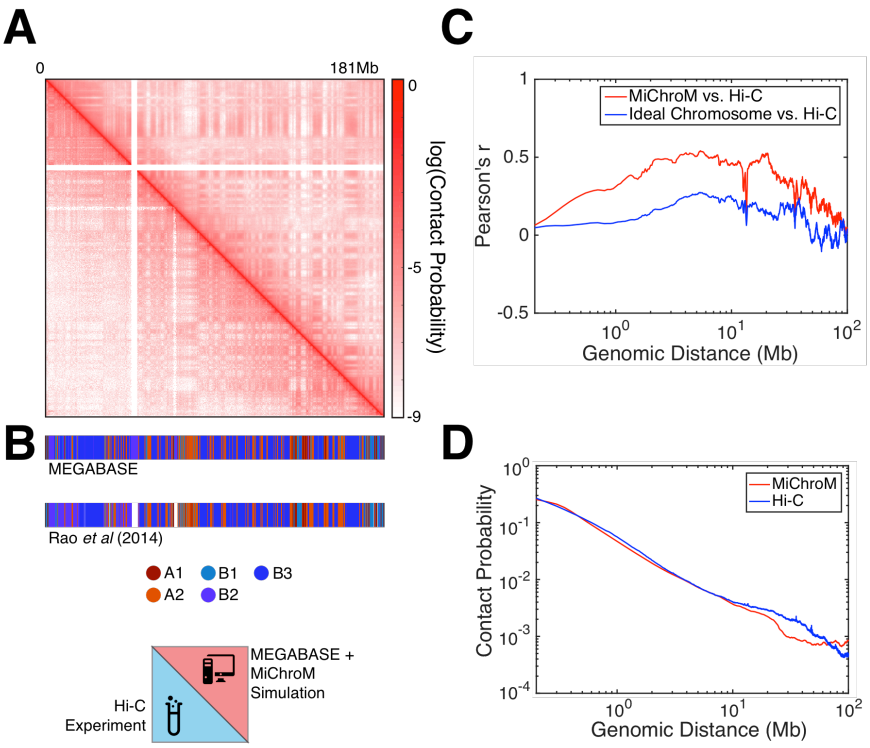


Figure S7
For all autosomal chromosomes MEGABASE+MiChroM generates conformational ensembles that accurately predict the results of DNA-DNA ligation assays.

(A) Contact map of chromosomes 5 (belonging to the training set) represented in log scale. Upper diagonal region shows the predicted map; MEGABASE+MiChroM generated this map *in silico* from ChIP-Seq input. The lower diagonal region shows maps from Hi-C (2). The quality of the predicted contact map is high, as shown by the symmetry of the map. Pearson's correlation between the two datasets is shown in Table S3.

(B) Comparison between the compartment annotations obtained by Hi-C (2) and MEGABASE structural type annotations.

(C) Pearson's correlation between experimental and simulated contact maps as a function of the genomic distance. MEGABASE+MiChroM generates contact maps that are well correlated with the experimental ones for distances exceeding the hundreds of Mb. As term of comparison, we show in blue the correlation between experimental maps and maps obtained using a homo-polymeric model including the Ideal Chromosome Potential (i.e. MiChroM without Type-to-Type interactions).

(D) The probability of contacts as a function of genomic distance in both measured and predicted maps.

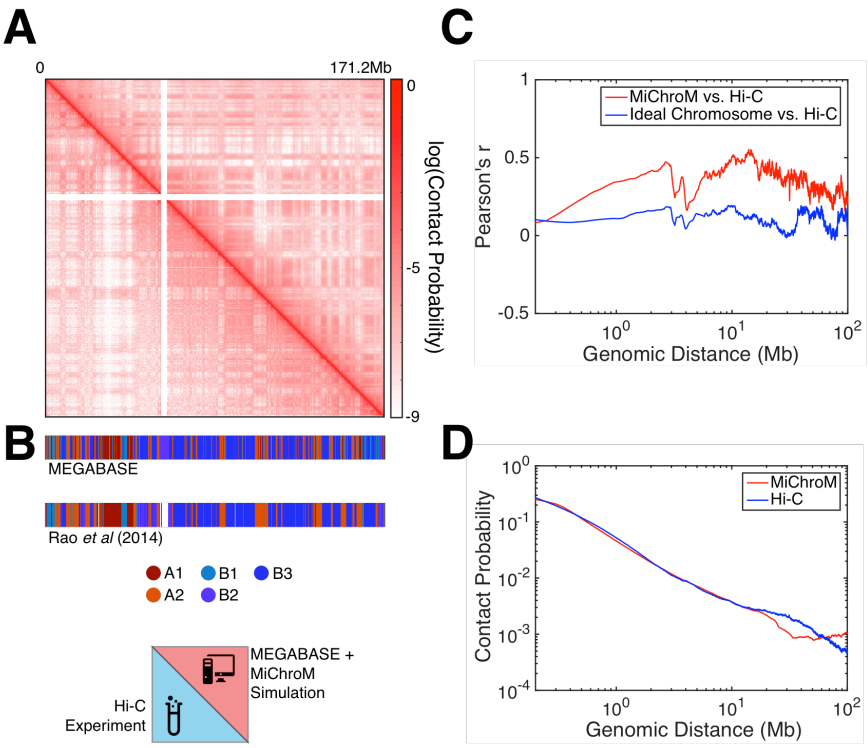


Figure S8

For all autosomal chromosomes MEGABASE+MiChroM generates conformational ensembles that accurately predict the results of DNA-DNA ligation assays.

(A) Contact map of chromosomes 6 (belonging to the test set) represented in log scale. Upper diagonal region shows the predicted map; MEGABASE+MiChroM generated this map *in silico* from ChIP-Seq input. The lower diagonal region shows maps from Hi-C (2). The quality of the predicted contact map is high, as shown by the symmetry of the map. Pearson's correlation between the two datasets is shown in Table S3.

(B) Comparison between the compartment annotations obtained by Hi-C (2) and MEGABASE structural type annotations.

(C) Pearson's correlation between experimental and simulated contact maps as a function of the genomic distance. MEGABASE+MiChroM generates contact maps that are well correlated with the experimental ones for distances exceeding the hundreds of Mb. As term of comparison, we show in blue the correlation between experimental maps and maps obtained using a homo-polymeric model including the Ideal Chromosome Potential (i.e. MiChroM without Type-to-Type interactions).

(D) The probability of contacts as a function of genomic distance in both measured and predicted maps.

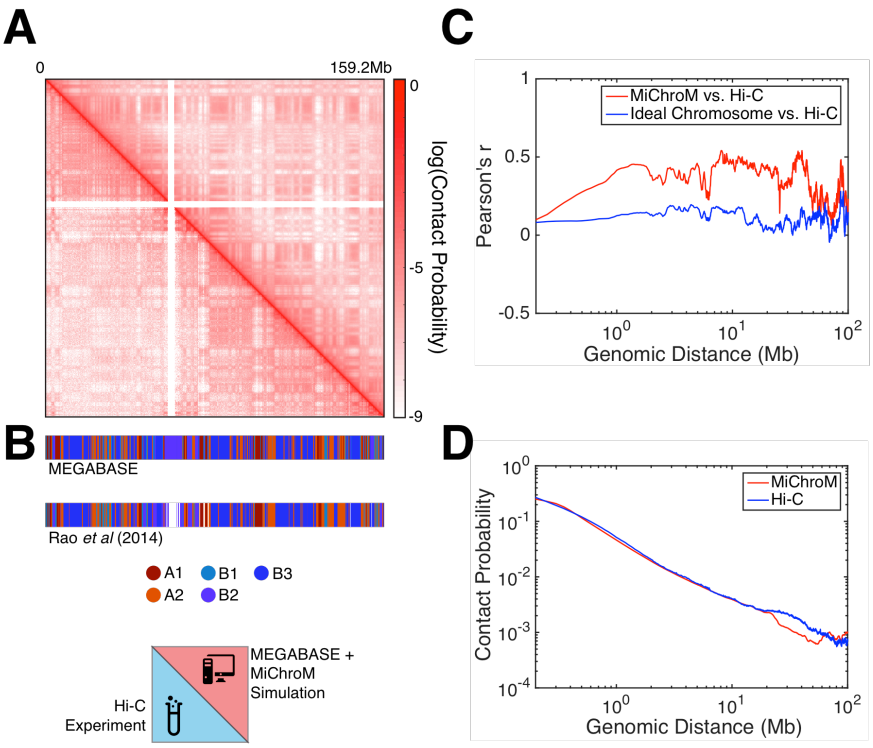


Figure S9

For all autosomal chromosomes MEGABASE+MiChroM generates conformational ensembles that accurately predict the results of DNA-DNA ligation assays.

(A) Contact map of chromosomes 7 (belonging to the training set) represented in log scale. Upper diagonal region shows the predicted map; MEGABASE+MiChroM generated this map *in silico* from ChIP-Seq input. The lower diagonal region shows maps from Hi-C (2). The quality of the predicted contact map is high, as shown by the symmetry of the map. Pearson's correlation between the two datasets is shown in Table S3.

(B) Comparison between the compartment annotations obtained by Hi-C (2) and MEGABASE structural type annotations.

(C) Pearson's correlation between experimental and simulated contact maps as a function of the genomic distance. MEGABASE+MiChroM generates contact maps that are well correlated with the experimental ones for distances exceeding the hundreds of Mb. As term of comparison, we show in blue the correlation between experimental maps and maps obtained using a homo-polymeric model including the Ideal Chromosome Potential (i.e. MiChroM without Type-to-Type interactions).

(D) The probability of contacts as a function of genomic distance in both measured and predicted maps.

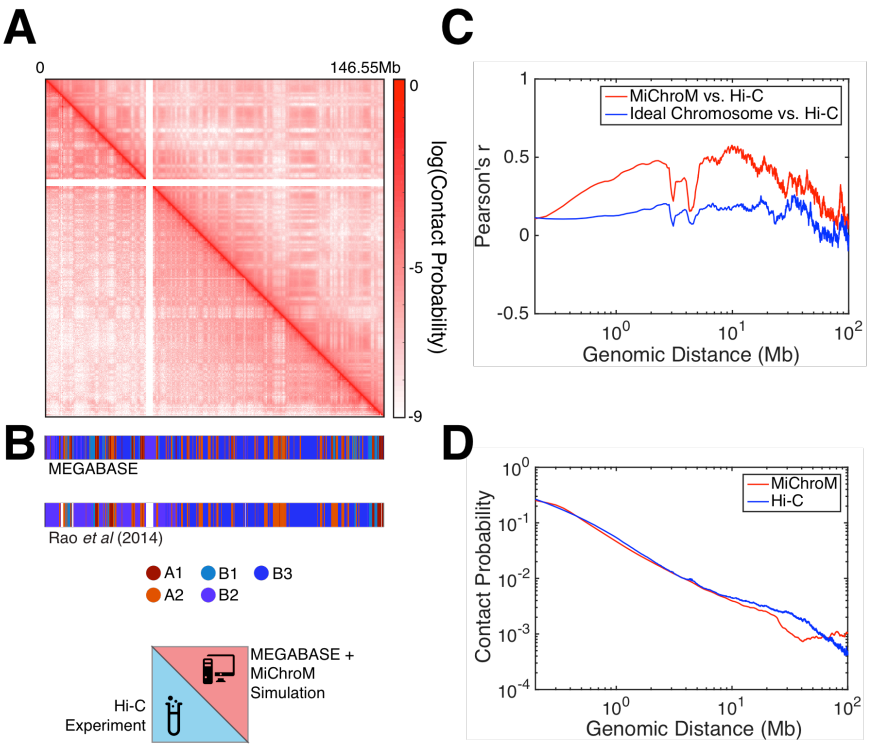


Figure S10
For all autosomal chromosomes MEGABASE+MiChroM generates conformational ensembles that accurately predict the results of DNA-DNA ligation assays.

(A) Contact map of chromosomes 8 (belonging to the test set) represented in log scale. Upper diagonal region shows the predicted map; MEGABASE+MiChroM generated this map *in silico* from ChIP-Seq input. The lower diagonal region shows maps from Hi-C (2). The quality of the predicted contact map is high, as shown by the symmetry of the map. Pearson's correlation between the two datasets is shown in Table S3.

(B) Comparison between the compartment annotations obtained by Hi-C (2) and MEGABASE structural type annotations.

(C) Pearson's correlation between experimental and simulated contact maps as a function of the genomic distance. MEGABASE+MiChroM generates contact maps that are well correlated with the experimental ones for distances exceeding the hundreds of Mb. As term of comparison, we show in blue the correlation between experimental maps and maps obtained using a homo-polymeric model including the Ideal Chromosome Potential (i.e. MiChroM without Type-to-Type interactions).

(D) The probability of contacts as a function of genomic distance in both measured and predicted maps.

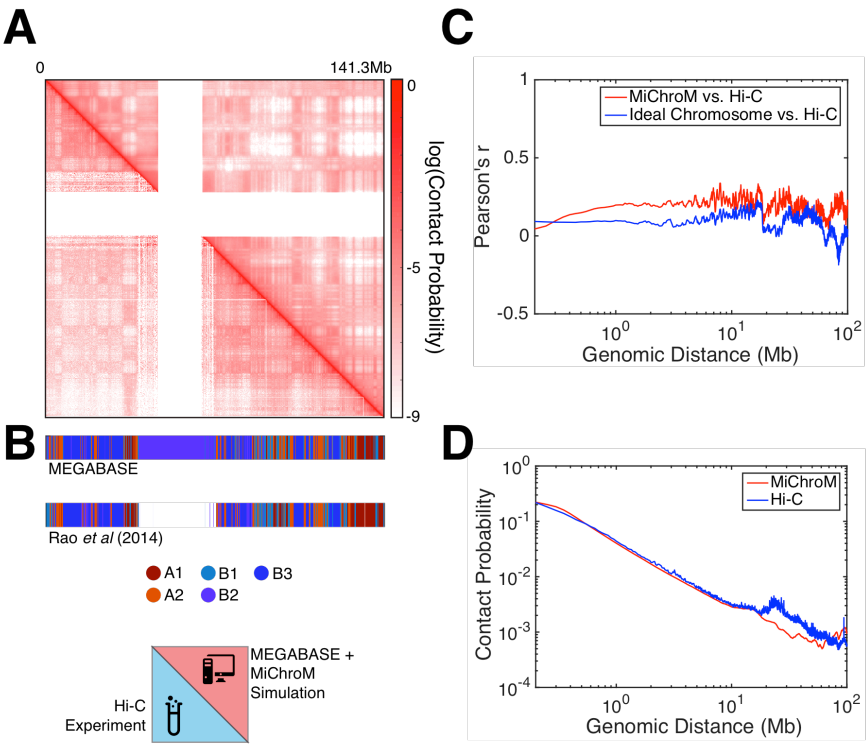


Figure S11
For all autosomal chromosomes MEGABASE+MiChroM generates conformational ensembles that accurately predict the results of DNA-DNA ligation assays.

(A) Contact map of chromosomes 9 (belonging to the training set) represented in log scale. Upper diagonal region shows the predicted map; MEGABASE+MiChroM generated this map *in silico* from ChIP-Seq input. The lower diagonal region shows maps from Hi-C (2). The quality of the predicted contact map is high, as shown by the symmetry of the map. Pearson's correlation between the two datasets is shown in Table S3. For this chromosome, the Pearson's correlation is relatively lower due to imperfect coverage in the experimental Hi-C map, which is visible in figure.

(B) Comparison between the compartment annotations obtained by Hi-C (2) and MEGABASE structural type annotations.

(C) Pearson's correlation between experimental and simulated contact maps as a function of the genomic distance. MEGABASE+MiChroM generates contact maps that are well correlated with the experimental ones for distances exceeding the hundreds of Mb. As term of comparison, we show in blue the correlation between experimental maps and maps obtained using a homo-polymeric model including the Ideal Chromosome Potential (i.e. MiChroM without Type-to-Type interactions).

(D) The probability of contacts as a function of genomic distance in both measured and predicted maps.

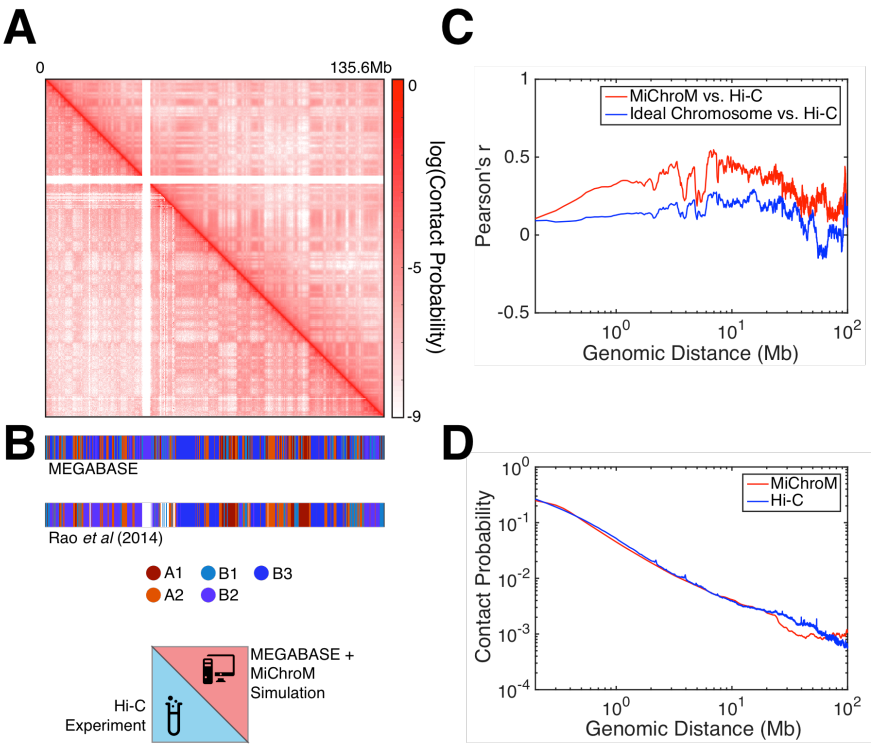


Figure S12
For all autosomal chromosomes MEGABASE+MiChroM generates conformational ensembles that accurately predict the results of DNA-DNA ligation assays.

(A) Contact map of chromosomes 10 (belonging to the test set[4]) represented in log scale. Upper diagonal region shows the predicted map; MEGABASE+MiChroM generated this map *in silico* from ChIP-Seq input. The lower diagonal region shows maps from Hi-C (2). The quality of the predicted contact map is high, as shown by the symmetry of the map. Pearson's correlation between the two datasets is shown in Table S3.

(B) Comparison between the compartment annotations obtained by Hi-C (2) and MEGABASE structural type annotations.

(C) Pearson's correlation between experimental and simulated contact maps as a function of the genomic distance. MEGABASE+MiChroM generates contact maps that are well correlated with the experimental ones for distances exceeding the hundreds of Mb. As term of comparison, we show in blue the correlation between experimental maps and maps obtained using a homo-polymeric model including the Ideal Chromosome Potential (i.e. MiChroM without Type-to-Type interactions).

[4] The Hi-C map of Chromosome 10 (Ref: 2. Rao SSP, *et al.* (2014) A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* 159(7):1665-1680.) was used to train the parameters of the MiChroM Hamiltonian (Ref: 12. Di Pierro M, Zhang B, Aiden EL, Wolynes PG, & Onuchic JN (2016) Transferable model for chromosome architecture. *Proceedings of the National Academy of Sciences of the United States of America* 113(43):12168-12173.).

(D) The probability of contacts as a function of genomic distance in both measured and predicted maps.

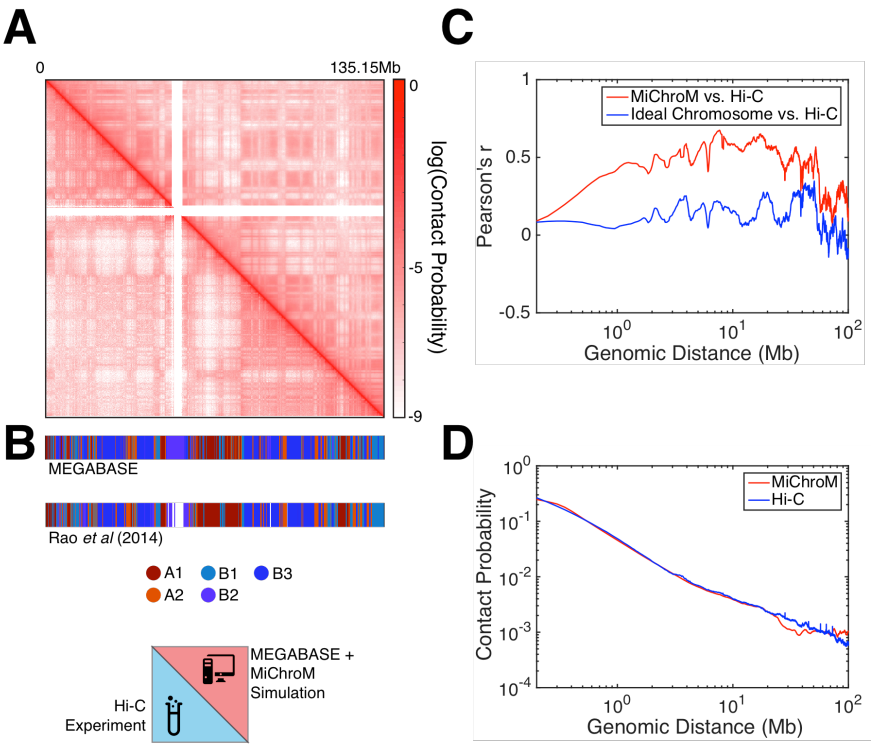


Figure S13
For all autosomal chromosomes MEGABASE+MiChroM generates conformational ensembles that accurately predict the results of DNA-DNA ligation assays.

(A) Contact map of chromosomes 11 (belonging to the training set) represented in log scale. Upper diagonal region shows the predicted map; MEGABASE+MiChroM generated this map *in silico* from ChIP-Seq input. The lower diagonal region shows maps from Hi-C (2). The quality of the predicted contact map is high, as shown by the symmetry of the map. Pearson's correlation between the two datasets is shown in Table S3.

(B) Comparison between the compartment annotations obtained by Hi-C (2) and MEGABASE structural type annotations.

(C) Pearson's correlation between experimental and simulated contact maps as a function of the genomic distance. MEGABASE+MiChroM generates contact maps that are well correlated with the experimental ones for distances exceeding the hundreds of Mb. As term of comparison, we show in blue the correlation between experimental maps and maps obtained using a homo-polymeric model including the Ideal Chromosome Potential (i.e. MiChroM without Type-to-Type interactions).

(D) The probability of contacts as a function of genomic distance in both measured and predicted maps.

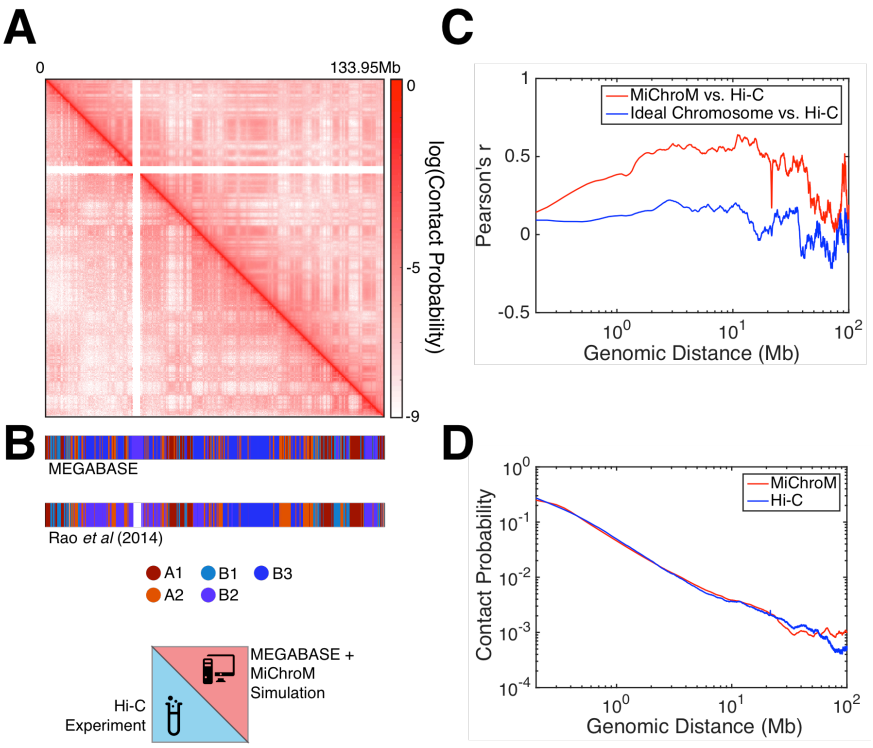


Figure S14

For all autosomal chromosomes MEGABASE+MiChroM generates conformational ensembles that accurately predict the results of DNA-DNA ligation assays.

(A) Contact map of chromosomes 12 (belonging to the test set) represented in log scale. Upper diagonal region shows the predicted map; MEGABASE+MiChroM generated this map *in silico* from ChIP-Seq input. The lower diagonal region shows maps from Hi-C (2). The quality of the predicted contact map is high, as shown by the symmetry of the map. Pearson's correlation between the two datasets is shown in Table S3.

(B) Comparison between the compartment annotations obtained by Hi-C (2) and MEGABASE structural type annotations.

(C) Pearson's correlation between experimental and simulated contact maps as a function of the genomic distance. MEGABASE+MiChroM generates contact maps that are well correlated with the experimental ones for distances exceeding the hundreds of Mb. As term of comparison, we show in blue the correlation between experimental maps and maps obtained using a homo-polymeric model including the Ideal Chromosome Potential (i.e. MiChroM without Type-to-Type interactions).

(D) The probability of contacts as a function of genomic distance in both measured and predicted maps.

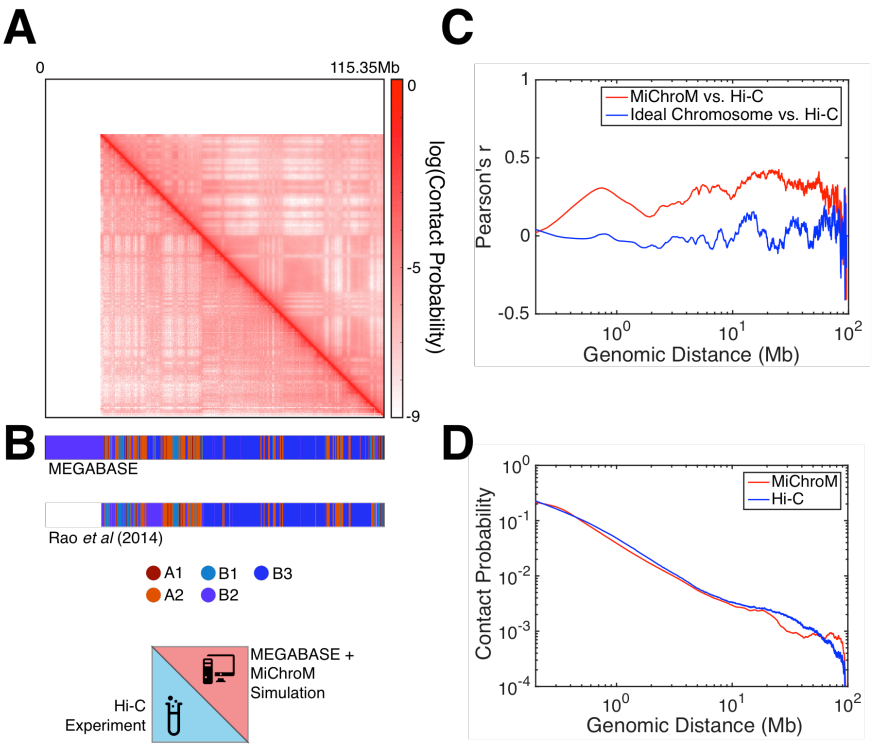


Figure S15

For all autosomal chromosomes MEGABASE+MiChroM generates conformational ensembles that accurately predict the results of DNA-DNA ligation assays.

(A) Contact map of chromosomes 13 (belonging to the training set) represented in log scale. Upper diagonal region shows the predicted map; MEGABASE+MiChroM generated this map *in silico* from ChIP-Seq input. The lower diagonal region shows maps from Hi-C (2). The quality of the predicted contact map is high, as shown by the symmetry of the map. Pearson's correlation between the two datasets is shown in Table S3.

(B) Comparison between the compartment annotations obtained by Hi-C (2) and MEGABASE structural type annotations.

(C) Pearson's correlation between experimental and simulated contact maps as a function of the genomic distance. MEGABASE+MiChroM generates contact maps that are well correlated with the experimental ones for distances exceeding the hundreds of Mb. As term of comparison, we show in blue the correlation between experimental maps and maps obtained using a homo-polymeric model including the Ideal Chromosome Potential (i.e. MiChroM without Type-to-Type interactions).

(D) The probability of contacts as a function of genomic distance in both measured and predicted maps.

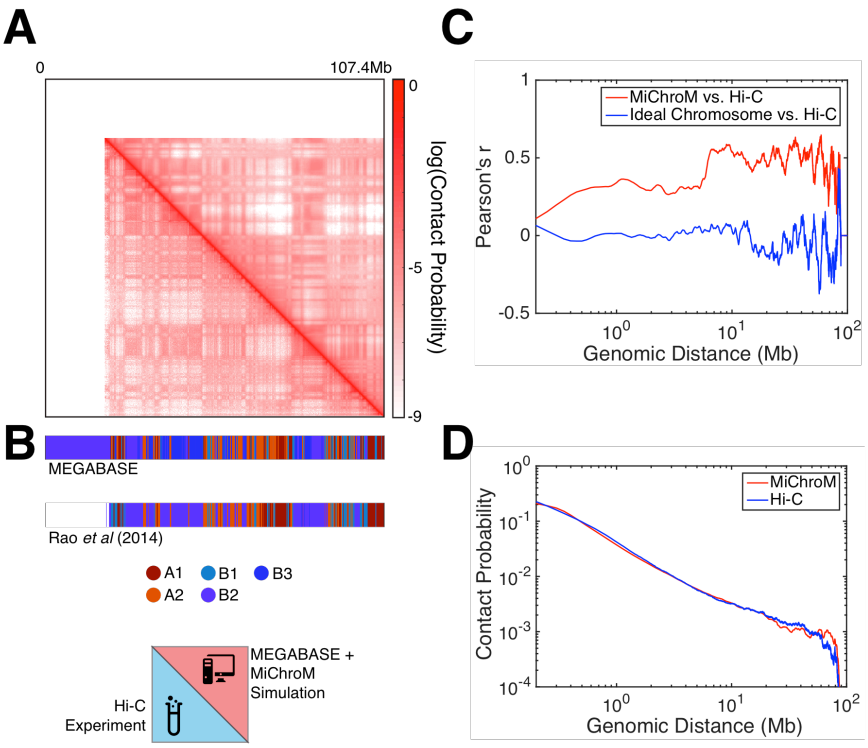


Figure S16

For all autosomal chromosomes MEGABASE+MiChroM generates conformational ensembles that accurately predict the results of DNA-DNA ligation assays.

(A) Contact map of chromosomes 14 (belonging to the test set) represented in log scale. Upper diagonal region shows the predicted map; MEGABASE+MiChroM generated this map *in silico* from ChIP-Seq input. The lower diagonal region shows maps from Hi-C (2). The quality of the predicted contact map is high, as shown by the symmetry of the map. Pearson's correlation between the two datasets is shown in Table S3.

(B) Comparison between the compartment annotations obtained by Hi-C (2) and MEGABASE structural type annotations.

(C) Pearson's correlation between experimental and simulated contact maps as a function of the genomic distance. MEGABASE+MiChroM generates contact maps that are well correlated with the experimental ones for distances exceeding the hundreds of Mb. As term of comparison, we show in blue the correlation between experimental maps and maps obtained using a homo-polymeric model including the Ideal Chromosome Potential (i.e. MiChroM without Type-to-Type interactions).

(D) The probability of contacts as a function of genomic distance in both measured and predicted maps.

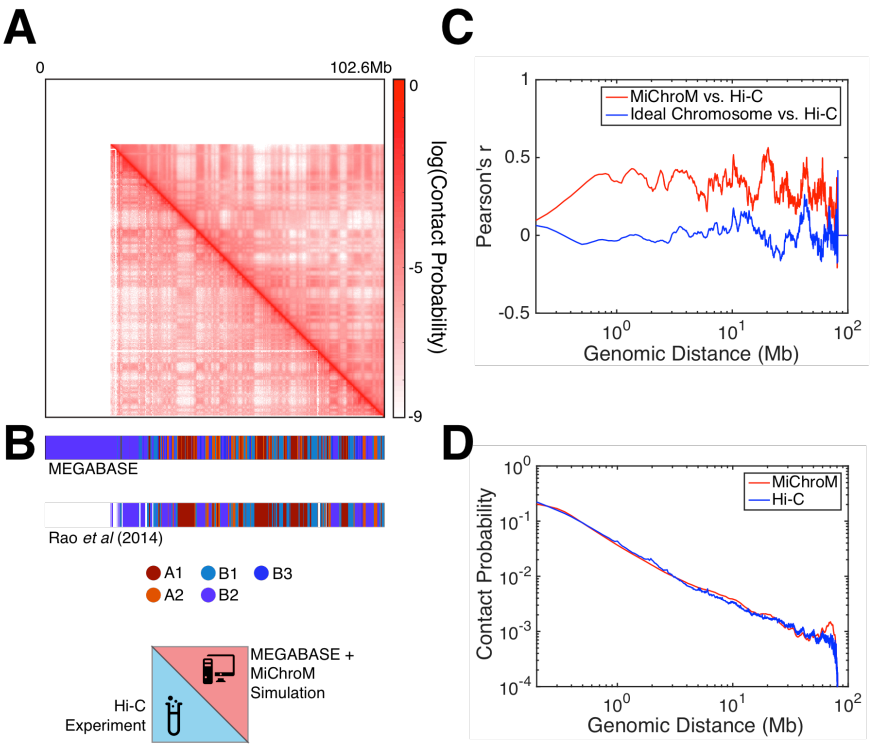


Figure S17
For all autosomal chromosomes MEGABASE+MiChroM generates conformational ensembles that accurately predict the results of DNA-DNA ligation assays.

(A) Contact map of chromosomes 15 (belonging to the training set) represented in log scale. Upper diagonal region shows the predicted map; MEGABASE+MiChroM generated this map *in silico* from ChIP-Seq input. The lower diagonal region shows maps from Hi-C (2). The quality of the predicted contact map is high, as shown by the symmetry of the map. Pearson's correlation between the two datasets is shown in Table S3.

(B) Comparison between the compartment annotations obtained by Hi-C (2) and MEGABASE structural type annotations.

(C) Pearson's correlation between experimental and simulated contact maps as a function of the genomic distance. MEGABASE+MiChroM generates contact maps that are well correlated with the experimental ones for distances exceeding the hundreds of Mb. As term of comparison, we show in blue the correlation between experimental maps and maps obtained using a homo-polymeric model including the Ideal Chromosome Potential (i.e. MiChroM without Type-to-Type interactions).

(D) The probability of contacts as a function of genomic distance in both measured and predicted maps.

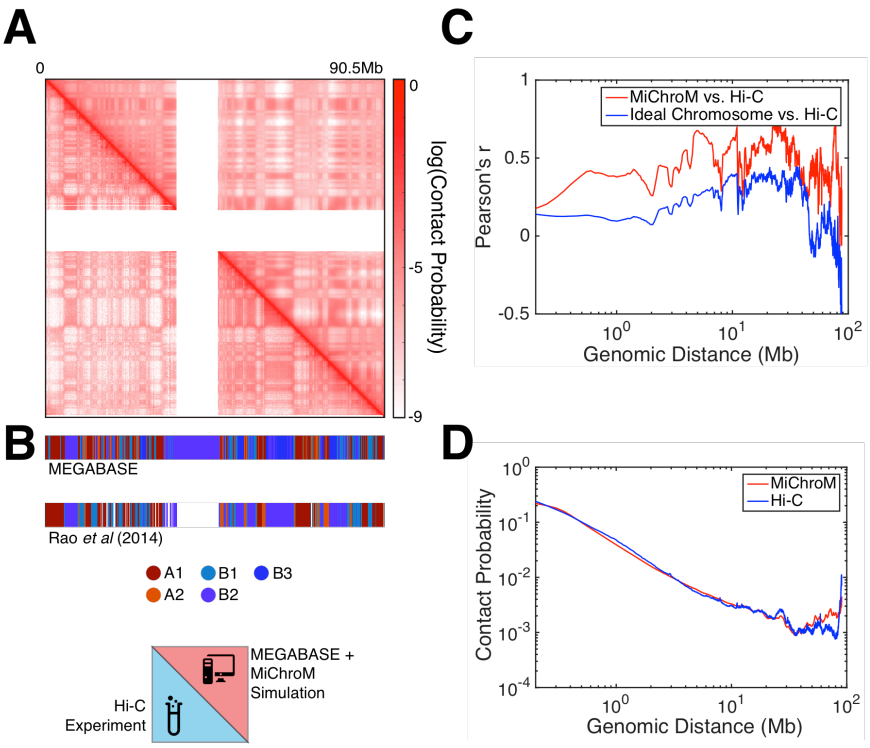


Figure S18

For all autosomal chromosomes MEGABASE+MiChroM generates conformational ensembles that accurately predict the results of DNA-DNA ligation assays.

(A) Contact map of chromosomes 16 (belonging to the test set) represented in log scale. Upper diagonal region shows the predicted map; MEGABASE+MiChroM generated this map *in silico* from ChIP-Seq input. The lower diagonal region shows maps from Hi-C (2). The quality of the predicted contact map is high, as shown by the symmetry of the map. Pearson's correlation between the two datasets is shown in Table S3.

(B) Comparison between the compartment annotations obtained by Hi-C (2) and MEGABASE structural type annotations.

(C) Pearson's correlation between experimental and simulated contact maps as a function of the genomic distance. MEGABASE+MiChroM generates contact maps that are well correlated with the experimental ones for distances exceeding the hundreds of Mb. As term of comparison, we show in blue the correlation between experimental maps and maps obtained using a homo-polymeric model including the Ideal Chromosome Potential (i.e. MiChroM without Type-to-Type interactions).

(D) The probability of contacts as a function of genomic distance in both measured and predicted maps.

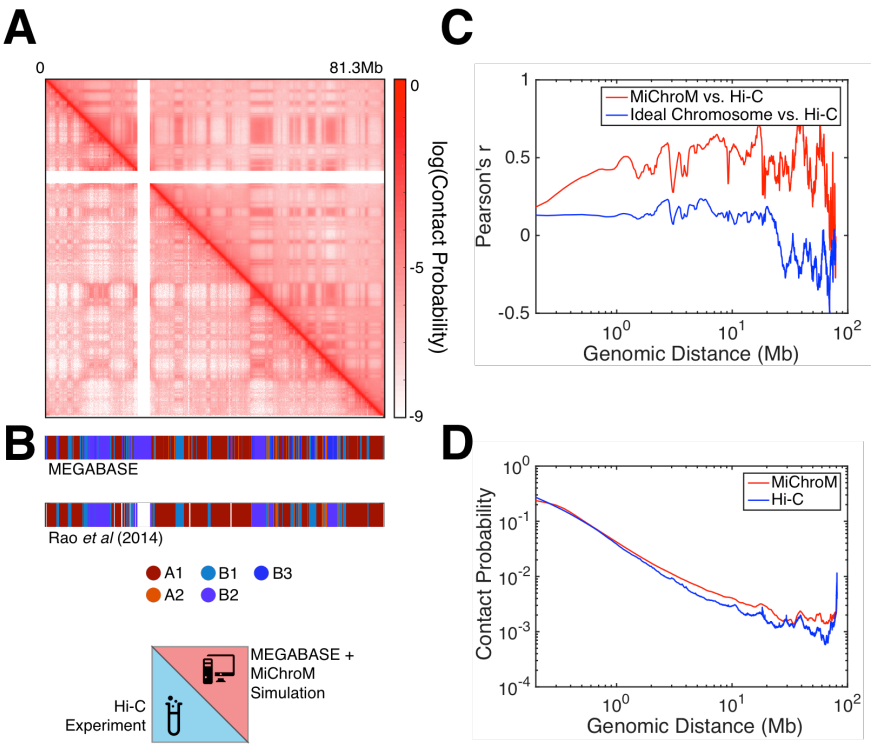


Figure S19
For all autosomal chromosomes MEGABASE+MiChroM generates conformational ensembles that accurately predict the results of DNA-DNA ligation assays.

(A) Contact map of chromosomes 17 (belonging to the training set) represented in log scale. Upper diagonal region shows the predicted map; MEGABASE+MiChroM generated this map *in silico* from ChIP-Seq input. The lower diagonal region shows maps from Hi-C (2). The quality of the predicted contact map is high, as shown by the symmetry of the map. Pearson's correlation between the two datasets is shown in Table S3.

(B) Comparison between the compartment annotations obtained by Hi-C (2) and MEGABASE structural type annotations.

(C) Pearson's correlation between experimental and simulated contact maps as a function of the genomic distance. MEGABASE+MiChroM generates contact maps that are well correlated with the experimental ones for distances exceeding the hundreds of Mb. As term of comparison, we show in blue the correlation between experimental maps and maps obtained using a homo-polymeric model including the Ideal Chromosome Potential (i.e. MiChroM without Type-to-Type interactions).

(D) The probability of contacts as a function of genomic distance in both measured and predicted maps.

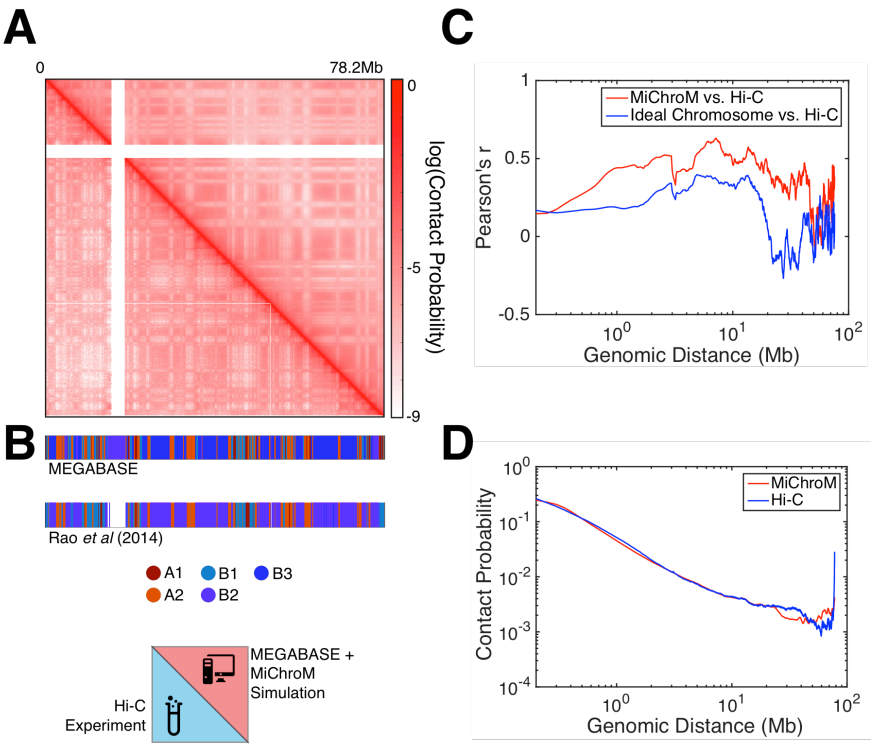


Figure S20
For all autosomal chromosomes MEGABASE+MiChroM generates conformational ensembles that accurately predict the results of DNA-DNA ligation assays.

(A) Contact map of chromosomes 18 (belonging to the test set) represented in log scale. Upper diagonal region shows the predicted map; MEGABASE+MiChroM generated this map *in silico* from ChIP-Seq input. The lower diagonal region shows maps from Hi-C (2). The quality of the predicted contact map is high, as shown by the symmetry of the map. Pearson's correlation between the two datasets is shown in Table S3.

(B) Comparison between the compartment annotations obtained by Hi-C (2) and MEGABASE structural type annotations.

(C) Pearson's correlation between experimental and simulated contact maps as a function of the genomic distance. MEGABASE+MiChroM generates contact maps that are well correlated with the experimental ones for distances exceeding the hundreds of Mb. As term of comparison, we show in blue the correlation between experimental maps and maps obtained using a homo-polymeric model including the Ideal Chromosome Potential (i.e. MiChroM without Type-to-Type interactions).

(D) The probability of contacts as a function of genomic distance in both measured and predicted maps.

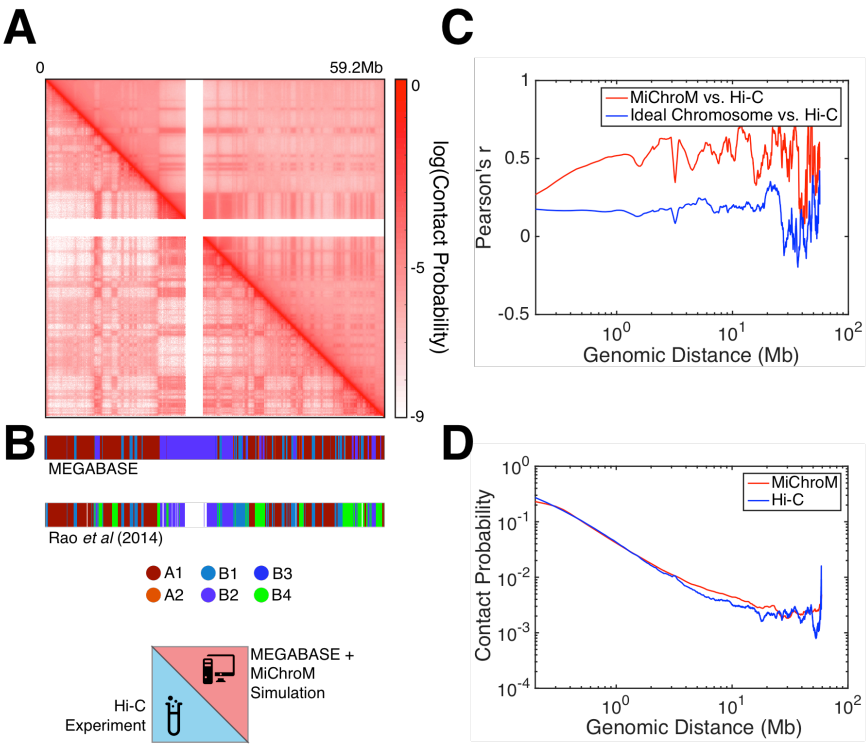


Figure S21

For all autosomal chromosomes MEGABASE+MiChroM generates conformational ensembles that accurately predict the results of DNA-DNA ligation assays.

(A) Contact map of chromosomes 19 (belonging to the training set) represented in log scale. Upper diagonal region shows the predicted map; MEGABASE+MiChroM generated this map *in silico* from ChIP-Seq input. The lower diagonal region shows maps from Hi-C (2). The quality of the predicted contact map is high, as shown by the symmetry of the map. Pearson's correlation between the two datasets is shown in Table S3.

(B) Comparison between the compartment annotations obtained by Hi-C (2) and MEGABASE structural type annotations.

(C) Pearson's correlation between experimental and simulated contact maps as a function of the genomic distance. MEGABASE+MiChroM generates contact maps that are well correlated with the experimental ones for distances exceeding the hundreds of Mb. As term of comparison, we show in blue the correlation between experimental maps and maps obtained using a homo-polymeric model including the Ideal Chromosome Potential (i.e. MiChroM without Type-to-Type interactions).

(D) The probability of contacts as a function of genomic distance in both measured and predicted maps.

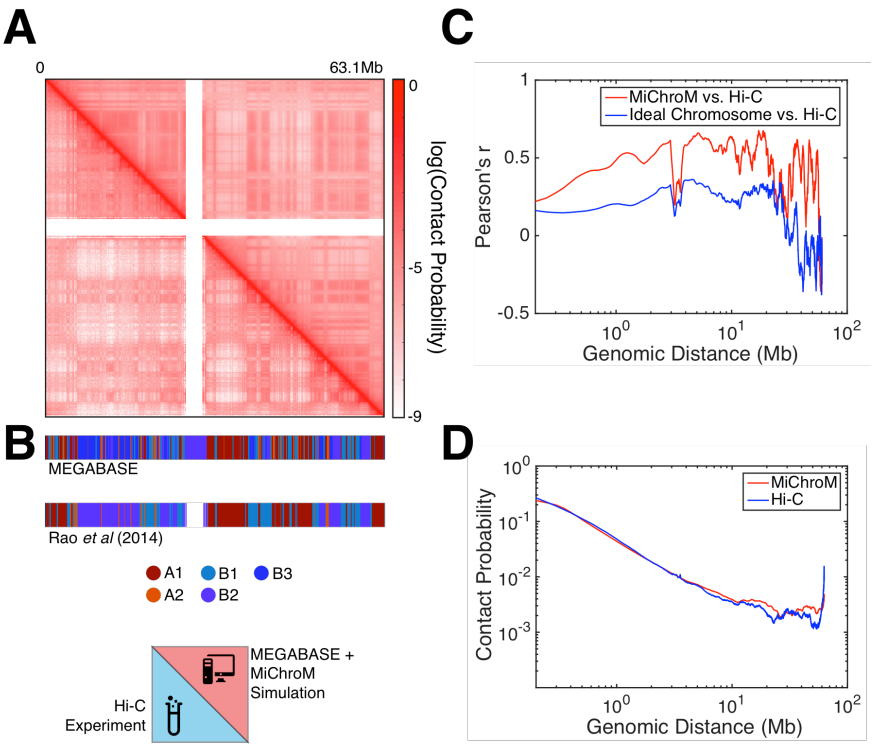


Figure S22

For all autosomal chromosomes MEGABASE+MiChroM generates conformational ensembles that accurately predict the results of DNA-DNA ligation assays.

(A) Contact map of chromosomes 20 (belonging to the test set) represented in log scale. Upper diagonal region shows the predicted map; MEGABASE+MiChroM generated this map *in silico* from ChIP-Seq input. The lower diagonal region shows maps from Hi-C (2). The quality of the predicted contact map is high, as shown by the symmetry of the map. Pearson's correlation between the two datasets is shown in Table S3.

(B) Comparison between the compartment annotations obtained by Hi-C (2) and MEGABASE structural type annotations.

(C) Pearson's correlation between experimental and simulated contact maps as a function of the genomic distance. MEGABASE+MiChroM generates contact maps that are well correlated with the experimental ones for distances exceeding the hundreds of Mb. As term of comparison, we show in blue the correlation between experimental maps and maps obtained using a homo-polymeric model including the Ideal Chromosome Potential (i.e. MiChroM without Type-to-Type interactions).

(D) The probability of contacts as a function of genomic distance in both measured and predicted maps.

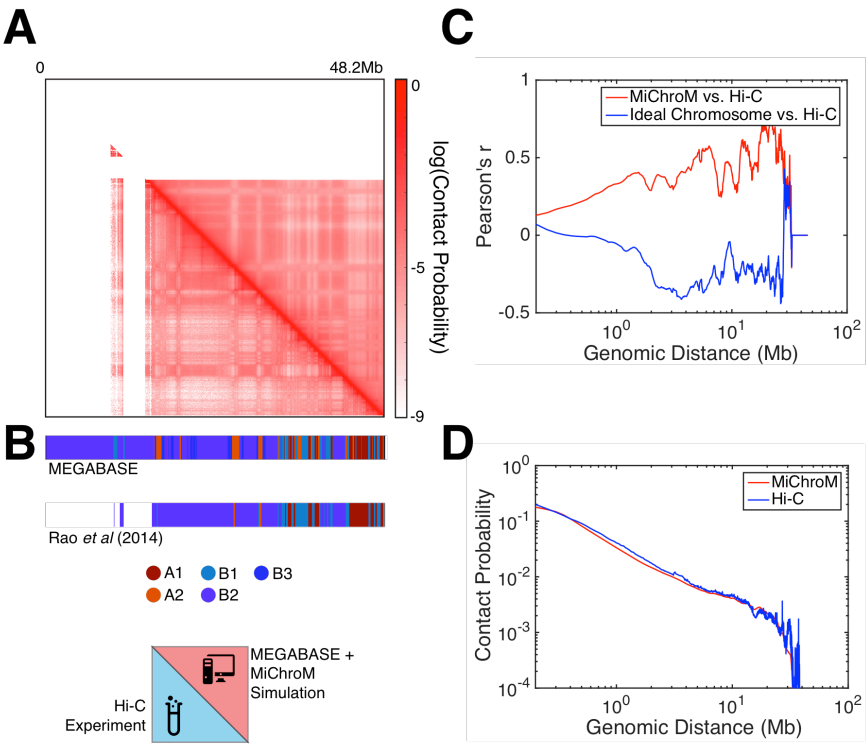


Figure S23
For all autosomal chromosomes MEGABASE+MiChroM generates conformational ensembles that accurately predict the results of DNA-DNA ligation assays.

(A) Contact map of chromosomes 21 (belonging to the training set) represented in log scale. Upper diagonal region shows the predicted map; MEGABASE+MiChroM generated this map *in silico* from ChIP-Seq input. The lower diagonal region shows maps from Hi-C (2). The quality of the predicted contact map is high, as shown by the symmetry of the map. Pearson's correlation between the two datasets is shown in Table S3.

(B) Comparison between the compartment annotations obtained by Hi-C (2) and MEGABASE structural type annotations.

(C) Pearson's correlation between experimental and simulated contact maps as a function of the genomic distance. MEGABASE+MiChroM generates contact maps that are well correlated with the experimental ones for distances exceeding the hundreds of Mb. As term of comparison, we show in blue the correlation between experimental maps and maps obtained using a homo-polymeric model including the Ideal Chromosome Potential (i.e. MiChroM without Type-to-Type interactions).

(D) The probability of contacts as a function of genomic distance in both measured and predicted maps.

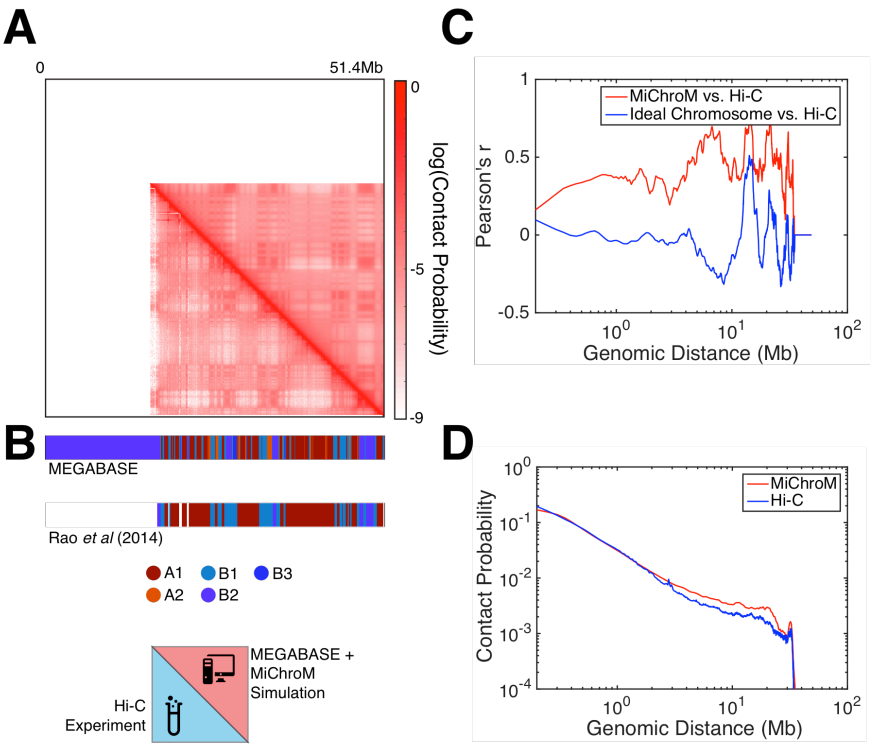


Figure S24
For all autosomal chromosomes MEGABASE+MiChroM generates conformational ensembles that accurately predict the results of DNA-DNA ligation assays.

(A) Contact map of chromosomes 22 (belonging to the test set) represented in log scale. Upper diagonal region shows the predicted map; MEGABASE+MiChroM generated this map *in silico* from ChIP-Seq input. The lower diagonal region shows maps from Hi-C (2). The quality of the predicted contact map is high, as shown by the symmetry of the map. Pearson's correlation between the two datasets is shown in Table S3.

(B) Comparison between the compartment annotations obtained by Hi-C (2) and MEGABASE structural type annotations.

(C) Pearson's correlation between experimental and simulated contact maps as a function of the genomic distance. MEGABASE+MiChroM generates contact maps that are well correlated with the experimental ones for distances exceeding the hundreds of Mb. As term of comparison, we show in blue the correlation between experimental maps and maps obtained using a homo-polymeric model including the Ideal Chromosome Potential (i.e. MiChroM without Type-to-Type interactions).

(D) The probability of contacts as a function of genomic distance in both measured and predicted maps.

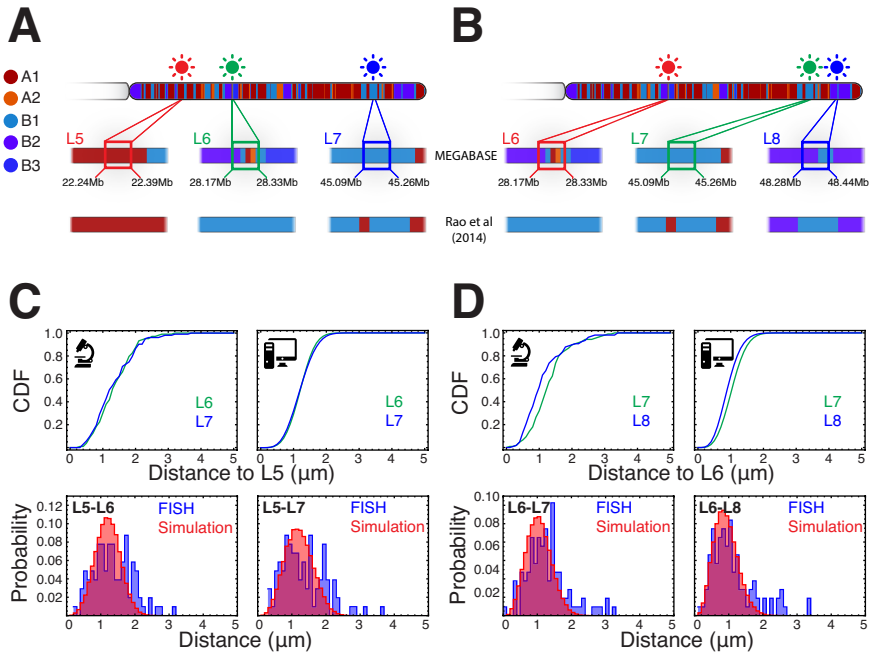


Figure S25

Simulations and 3D Fluorescence In Situ Hybridization (FISH) experiments support the idea that compartmentalization observed in Hi-C maps emerges from the phase separation of chromatin structural types. MEGABASE+MiChroM simulations of chromosome 22 of human B-Lymphocyte cells (GM12878) are compared with results of FISH experiments (for closely related human B-Lymphocyte cell line GM06990). The average ratio between simulated distances and FISH-measured distances has been used to calibrate the length scale of simulation. For this cell line and this experimental set up, one unit of length in simulation corresponds to a length of $0.17\mu m$, which also implies the size of a simulated chromosomal territory being approximately 2-3 μm across— consistent with what was previously reported in (17).

(A and B) The positions of the fluorescent probes along the chromosome are illustrated together with the annotations from MEGABASE and with the aligned compartment annotations from ref. (2). In 2009, the authors of ref. (11), assigned the four loci to alternating compartments— L5 and L7 in compartment A, and L6 and L8 in compartment B. Subsequently, in 2014 higher resolutions Hi-C experiments resulted in finer compartments annotations in partial agreement with the previously reported annotations. MEGABASE is also in partial agreement with both preexisting annotations.

(C and D) The Cartesian distances between four loci (L5, L6, L7, and L8) in chromosome 22 were measured in two distinct 3D FISH experiments reported in ref. (11). The same distances were measured using the MEGABASE+MiChroM pipeline. The cumulative distribution functions show that loci composed of chromatin belonging to the same type tend to be closer in space than otherwise, despite the interlaced order and despite lying at greater genomic distances.

This phenomenon is observed in FISH experiments and it is correctly predicted by our ChIP-Seq based modeling. The comparison between the predicted and the measured probability distributions shows excellent agreement for both the average distance and the distance fluctuations.

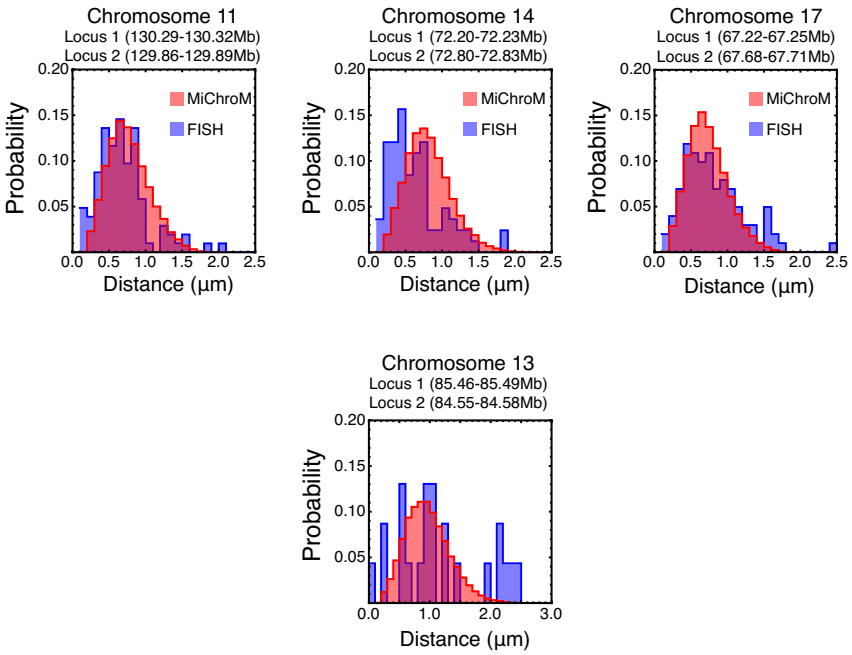


Figure S26

The predicted distribution of distances between selected chromatin loci well reproduces those measured by fluorescence in situ hybridization (FISH) in cell line GM12878 and reported in ref. (2). The pair of loci from chromosome 11 is used to calibrate the unit of length in the simulations. After calibration, the average distance between the pairs of loci in chromosomes 14 and 17 is correctly predicted. For this cell line and this experimental set up, one unit of length in simulation corresponds to $L = 0.24 \mu m$ measured at the microscope. This conversion implies the size of a simulated chromosomal territory being approximately 2-3 μm across— consistent with what was previously reported in (17) The simulated ensemble is also consistent with experimental data collected from chromosome 13, even though in this case the sparsity of the experimental data set does not allow strong conclusions. The fluctuations of the distances for all the four pairs are also correctly predicted in simulation.

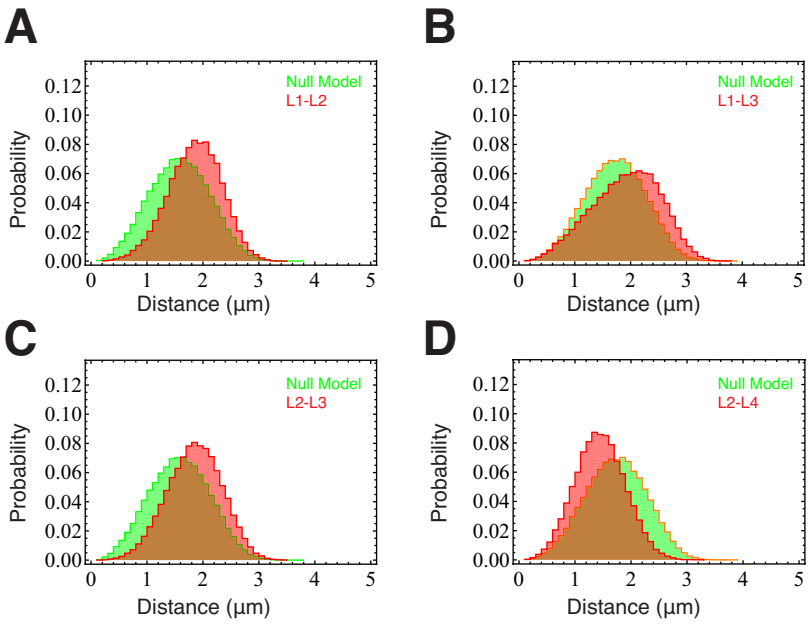


Figure S27.
Comparison of simulated FISH distances with a null model.
The simulated probability distributions of distances between the probes shown in Figure 3 are compared with a null model. The null model (green) is the distribution of distances between all pairs of loci separated by a fixed genomic distance, chosen to match the ones between the probes (A) L1-L2, (B) L1-L3, (C) L2-L3, and (D) L2-L4. The simulated distances between the probes (red distributions) are the same as the ones in Figure 3. The distribution of distances resulting from our model is specific to the pair of loci according to the topology of the chromosome and the interactions in the energy function. These loci-specific distributions deviate significantly from the ones characteristic of the null model.

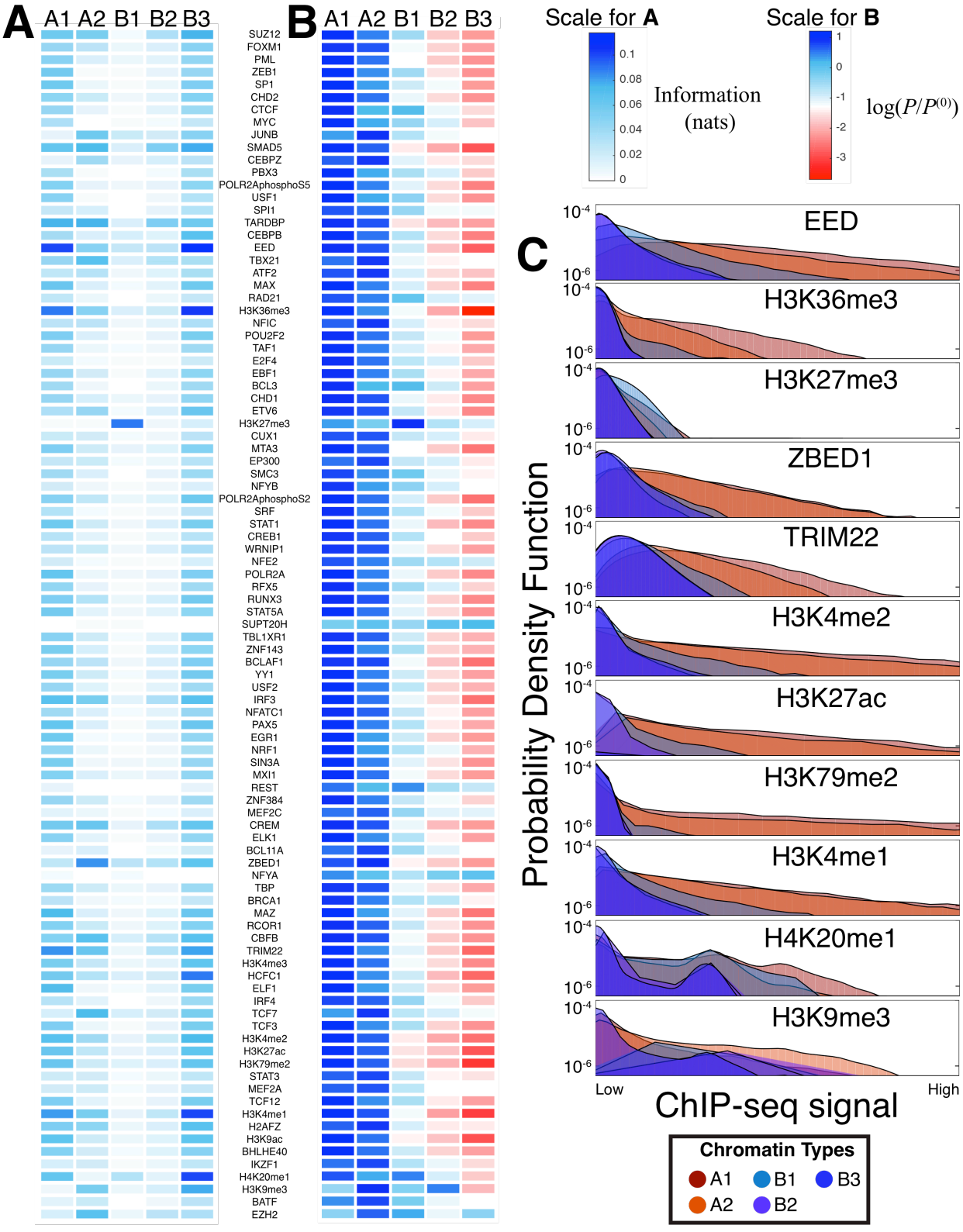

A
A1 A2 B1 B2 B3
B
A1 A2 B1 B2 B3
SUZ12
FOXM1
PML
ZEB1
SP1
CHD2
CTCF
MYC
JUNB
SMAD5
CEBPZ
PBX3
POLR2AphosphoS5
USF1
SPI1
TARDBP
CEBPB
EED
TBX21
ATF2
MAX
RAD21
H3K36me3
NFIC
POU2F2
TAF1
E2F4
EBF1
BCL3
CHD1
ETV6
H3K27me3
CUX1
MTA3
EP300
SMC3
NFYB
POLR2AphosphoS2
SRF
STAT1
CREB1
WRNIP1
NFE2
POLR2A
RFX5
RUNX3
STAT5A
SUPT20H
TBL1XR1
ZNF143
BCLAF1
YY1
USF2
IRF3
NFATC1
PAX5
EGR1
NRF1
SIN3A
MXI1
REST
ZNF384
MEF2C
CREM
ELK1
BCL11A
ZBED1
NFYA
TBP
BRCA1
MAZ
RCOR1
CBFB
TRIM22
H3K4me3
HCFC1
ELF1
IRF4
TCF7
TCF3
H3K4me2
H3K27ac
H3K79me2
STAT3
MEF2A
TCF12
H3K4me1
H2AFZ
H3K9ac
BHLHE40
IKZF1
H4K20me1
H3K9me3
BATF
EZH2
Scale for A
0.1
0.08
0.06
0.04
0.02
0
Information (nats)
Scale for B
1
0
-1
-2
-3
log(P/P(0))
C
Probability Density Function
EED
H3K36me3
H3K27me3
ZBED1
TRIM22
H3K4me2
H3K27ac
H3K79me2
H3K4me1
H4K20me1
H3K9me3
10-4
10-6
Low
High
ChIP-seq signal
Chromatin Types
A1
B1
B3
A2
B2

Figure S28
Insights on the relationship between compartmentalization and biochemical markers

(A) A select number of biochemical markers are strongly associated with the compartment annotations. The information content between a compartment annotation and a biochemical marker is plotted for our probabilistic model. High information content indicates a strong statistical coupling between a compartment annotation and a biochemical marker.

(B) Sub-compartments A and B exhibit probabilistic enhancement and depletion, respectively, of almost all biochemical markers. The log ratio is plotted between the joint probability of observing a compartment type and the presence of the indicated biochemical marker with respect to a model for which the type and marker are independent (null model). Blue indicates the enhancement of the joint probability with respect to the null model, whereas Red indicates depletion.

(C) The probability density function of the measured signal is plotted for each chromatin structural type (A1, A2, B1, B2, B3; colors defined in the legend) for selected ChIP-Seq assays. The PDFs are plotted on a log-linear scale.

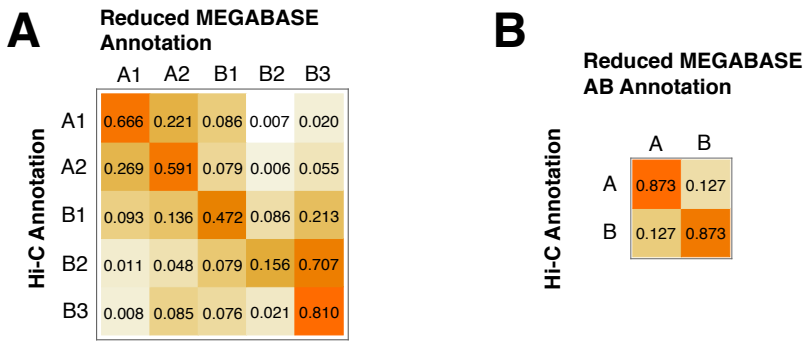
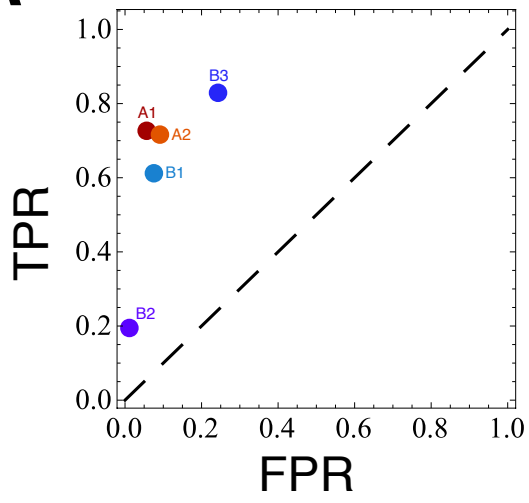


Figure S29

(A) A confusion matrix showing the chromatin type annotation of the reduced MEGABASE model (using only the 11 histone modifications) for each of the chromatin types classified in ref. (2). Likewise, (B) shows the same comparison for A/B compartments.

A



B

	A1	A2	B1	B2	B3
Recall (TPR)	0.7281	0.7178	0.6135	0.1962	0.8306
Precision (PPV)	0.6283	0.6922	0.5037	0.7983	0.6093
F1-Score	0.6745	0.7048	0.5532	0.3150	0.7030

Figure S30. Selected performance measures for MEGABASE.
To further assess the quality of MEGABASE, we recast it as 5 binary classifiers. Does the segment of chromatin belong to type X? Yes or No. (A) The 5 virtual binary classifiers are shown in the Receiver Operating Characteristic (ROC) Space. (B) The Recall, Precision, and F1-Score are reported in a Table for all binary classifiers.

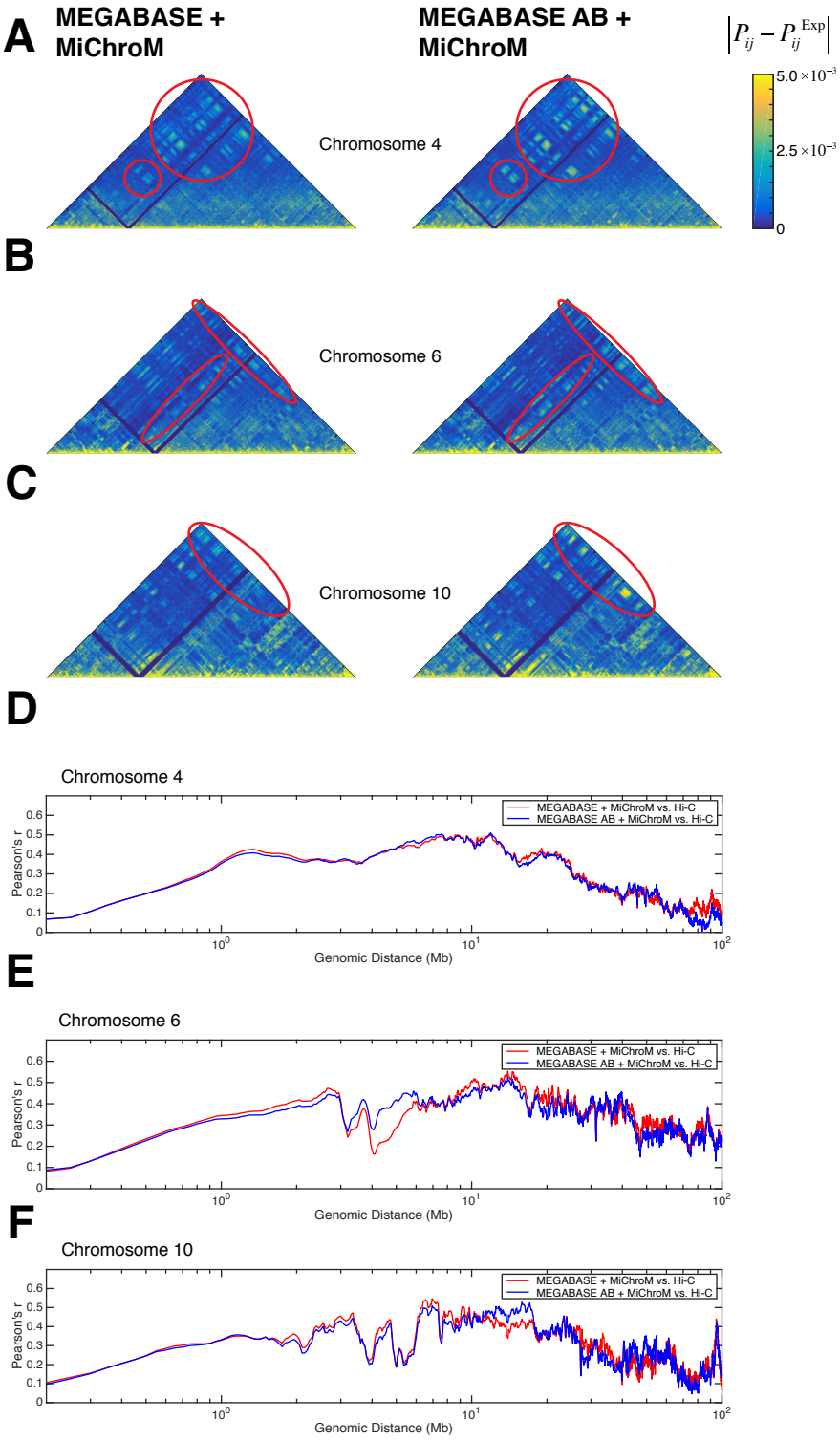


Figure S31.
The intra-chromosomal contact maps produced by MEGABASE + MiChroM are more accurate then the ones produced by MEGABASE AB + MiChroM, even though the differences are subtle.

(A, B, and C) The error in the simulated contact maps with respect to the experimental Hi-C maps of ref. *(2)* is shown for representative chromosomes 4, 6 and 10, respectively. The differences with respect to experiment are larger for the MEGABASE AB + MiChroM compared to the simulated maps of MEGABASE + MiChroM, showing the advantage of a model with 5-types compared to a model with 2-types. **(D, E, and F)** The Pearson's coefficient between the experimental Hi-C data and simulated contact probabilities are plotted as a function of genomic distance.

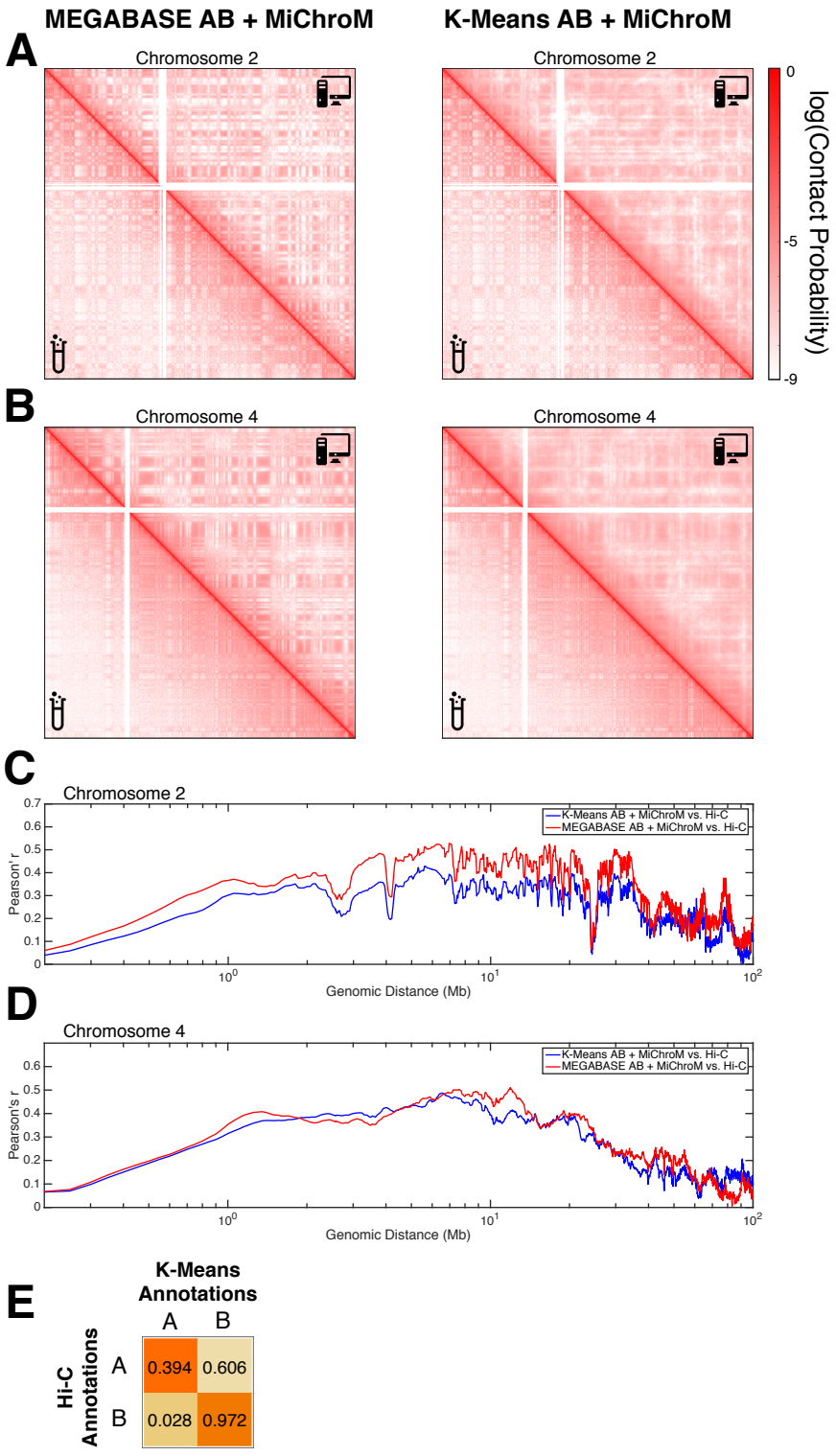


Figure S32. Purely biochemical clustering is unable to correctly classify structural types. (A and B) For chromosome 2 and 4, a comparison between simulated Hi-C maps generated using MEGABASE AB + MiChroM (left) and K-Means clustering of ChIP-Seq data + MiChroM is shown. (C and D) Pearson's correlation vs. genomic distance for the maps in A and B. (E)

Confusion matrix between the annotations obtained by MEGABASE AB and K-means of ChIP-Seq for the test set.

Table S1. List of ChIP-seq experiments obtained from ENCODE for GM12878

1. SUZ12
2. FOXM1
3. PML
4. ZEB1
5. SP1
6. CHD2
7. CTCF
8. MYC
9. JUNB
10. SMAD5
11. CEBPZ
12. PBX3
13. POLR2AphosphoS5
14. USF1
15. SPI1
16. TARDBP
17. CEBPB
18. EED
19. TBX21
20. ATF2
21. MAX
22. RAD21
23. H3K36me3
24. NFIC
25. POU2F2
26. TAF1
27. E2F4
28. EBF1
29. BCL3
30. CHD1
31. ETV6
32. H3K27me3
33. CUX1
34. MTA3
35. EP300
36. SMC3
37. NFYB
38. POLR2AphosphoS2
39. SRF
40. STAT1
41. CREB1
42. WRNIP1
43. NFE2
44. POLR2A
45. RFX5
46. RUNX3
47. STAT5A

48. SUPT20H
49. TBL1XR1
50. ZNF143
51. BCLAF1
52. YY1
53. USF2
54. IRF3
55. NFATC1
56. PAX5
57. EGR1
58. NRF1
59. SIN3A
60. MXI1
61. REST
62. ZNF384
63. MEF2C
64. CREM
65. ELK1
66. BCL11A
67. ZBED1
68. NFYA
69. TBP
70. BRCA1
71. MAZ
72. RCOR1
73. CBFB
74. TRIM22
75. H3K4me3
76. HCFC1
77. ELF1
78. IRF4
79. TCF7
80. TCF3
81. H3K4me2
82. H3K27ac
83. H3K79me2
84. STAT3
85. MEF2A
86. TCF12
87. H3K4me1
88. H2AFZ
89. H3K9ac
90. BHLHE40
91. IKZF1
92. H4K20me1
93. H3K9me3
94. BATF
95. EZH2

Table S2. List of epigenetic marks obtained from ENCODE for GM12878.

1. H2AFZ
2. H3K27ac
3. H3K27me3
4. H3K36me3
5. H3K4me1

6. H3K4me2
7. H3K4me3
8. H3K79me2
9. H3K9ac
10. H3K9me3
11. H4K20me1

Table S3 Pearson correlation of between Hi-C map and simulated contact probabilities from MEGABASE+MiChroM

Chromosome	Pearson's *r*
1	0.941
2	0.952
3	0.962
4	0.963
5	0.956
6	0.956
7	0.954
8	0.955
9	0.879*
10	0.944
11	0.955
12	0.959
13	0.961
14	0.960
15	0.935
16	0.946
17	0.953
18	0.963
19	0.960
20	0.956
21	0.924
22	0.942

* Experimental dataset for chromosome 9 has imperfect coverage

References

1. Dunham I, *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57-74.
2. Rao SSP, *et al.* (2014) A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* 159(7):1665-1680.
3. Jaynes ET (1957) Information Theory and Statistical Mechanics. *Physical Review* 106(4):620-630.
4. Hopfield JJ (1982) Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *P Natl Acad Sci-Biol* 79(8):2554-2558.
5. Ekeberg M, Lovkvist C, Lan YH, Weigt M, & Aurell E (2013) Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Physical Review E* 87(1).
6. Besag J (1975) Statistical-Analysis of Non-Lattice Data. *Statistician* 24(3):179-195.
7. Potoyan DA & Papoian GA (2012) Regulation of the H4 tail binding and folding landscapes via Lys-16 acetylation. *P Natl Acad Sci USA* 109(44):17857-17862.
8. Shogren-Knaak M, *et al.* (2006) Histone H4-K16 acetylation controls chromatin structure and protein interactions. *Science* 311(5762):844-847.
9. Tessarz P & Kouzarides T (2014) Histone core modifications regulating nucleosome structure and dynamics. *Nat Rev Mol Cell Bio* 15(11):703-708.
10. Wilkins BJ, *et al.* (2014) A Cascade of Histone Modifications Induces Chromatin Condensation in Mitosis. *Science* 343(6166):77-80.
11. Lieberman-Aiden E, *et al.* (2009) Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* 326(5950):289-293.
12. Di Pierro M, Zhang B, Aiden EL, Wolynes PG, & Onuchic JN (2016) Transferable model for chromosome architecture. *Proceedings of the National Academy of Sciences of the United States of America* 113(43):12168-12173.
13. Zhou J & Troyanskaya OG (2016) Probabilistic modelling of chromatin code landscape reveals functional diversity of enhancer-like chromatin states. *Nat Commun* 7.
14. Kremer K & Grest GS (1990) Dynamics of Entangled Linear Polymer Melts - a Molecular-Dynamics Simulation. *Journal of Chemical Physics* 92(8):5057-5086.
15. Naumova N, *et al.* (2013) Organization of the Mitotic Chromosome. *Science* 342(6161):948-953.
16. Rosa A & Everaers R (2008) Structure and Dynamics of Interphase Chromosomes. *Plos Comput Biol* 4(8).
17. Cremer T & Cremer C (2001) Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet* 2(4):292-301.
18. Plimpton S (1995) Fast Parallel Algorithms for Short-Range Molecular-Dynamics. *Journal of Computational Physics* 117(1):1-19.
19. Rhead B, *et al.* (2010) The UCSC Genome Browser database: update 2010. *Nucleic Acids Res* 38:D613-D619.
20. Alipour E & Marko JF (2012) Self-organization of domain structures by DNA-loop-extruding enzymes. *Nucleic Acids Res* 40(22):11202-11212.
21. Sanborn AL, *et al.* (2015) Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *P Natl Acad Sci USA* 112(47):E6456-E6465.
22. Rao S, *et al.* (2017) Cohesin Loss Eliminates All Loop Domains, Leading To Links Among Superenhancers And Downregulation Of Nearby Genes. *bioRxiv*.
23. Nora EP, *et al.* (2017) Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell* 169(5):930-+.