Supplementary Information for:

The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules

Davide Baù[1,4], Amartya Sanyal[2,4], Bryan R. Lajoie[2,4], Emidio Capriotti[1], Meg Byron[3], Jeanne B. Lawrence[3], Job Dekker[2*], and Marc A. Marti-Renom[1*]

1. Structural Genomics Unit, Bioinformatics and Genomics Department, Centro de Investigación Príncipe Felipe, 46012 Valencia, Spain.

2. Program in Gene Function and Expression, Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School 01605-2324 Worcester MA, USA.

3. Department of Cell Biology, University of Massachusetts Medical School 01605-2324 Worcester MA, USA.

4 These authors contributed equally to the work.

* Corresponding authors:

Job Dekker
Department of Biochemistry and Molecular Pharmacology,
University of Massachusetts Medical School
Lazare Research Building room 519
364 Plantation Street
01605-2324 Worcester MA. USA
Tel +1 (508) 856-4371
Fax +1 (508) 856 4650
e-mail: Job.Dekker@umassmed.edu

Marc A. Marti-Renom
Structural Genomics Unit,
Bioinformatics and Genomics Department.
Centro de Investigación Príncipe Felipe.
Av. Autopista del Saler, 16, 46012 Valencia, Spain.
Tel: +34 96 3289680
Fax: +34 96 3289701
e-mail: mmarti@cipf.es

Version: September 20, 2010

Nature Structural & Molecular Biology: doi:10.1038/nsmb.1936

Supplementary Methods

Chromosome Conformation Capture Carbon Copy (5C) experiments result in two-dimensional tables representing the frequency of interactions between loci along a chromosome(s). To transform such two-dimensional (2D) data into a 3D conformation of higher-order chromatin folding, we used the Integrative Modeling Platform (IMP)[1]. Similar to nuclear magnetic resonance (NMR) spectroscopy, which relies on a two-dimensional (2D) representation of a molecular structure to computationally derive its 3D structure[2], the IMP approach[1] uses a 2D interaction matrix from 5C experiments to derive a set of spatial distances (proportional to the observed interactions) that will determine the 3D folding of the studied genomic domain. The conceptual aim of IMP is to determine a 3D structure of a biological molecule or complex that best satisfies diverse experimental observations.

Next sections describe in details all methods used in each of the four mains steps of our approach, including: (i) 5C data normalization, (ii) IMP model representation, scoring function and parameter optimization, (iii) model building with IMP, (iv) model ensemble analysis and de-convolution, and (v) model visualization with Chimera[3].

Expected background interactions

In the absence of specific long-range looping interactions, chromatin interactions are expected to be most frequent between sites located near each other in the linear genome, and to decrease precipitously for sites located farther apart[4]. We used the 5C data obtained for ENm008 to empirically determine this background level of interaction.

Nature Structural & Molecular Biology: doi:10.1038/nsmb.1936

We first plotted all 5C data versus genomic distance (**Supplementary Fig. 2**). Next we performed LOESS smoothing with a window size of 37 interactions (α = 0.05) to obtain a smooth curve representing the average relationship between 5C interaction counts and the genomic distance between pairs of loci. By assuming that only a small fraction of the set of 750 interactions represents specific long-range looping interactions, the LOESS curve estimates the level of expected 5C interactions in the absence of a specific looping interaction. To further estimate the variability between 5C interaction counts and the genomic distance between pairs of loci, the standard error (SE_d) was calculated as:

$$SE_d = \frac{\sigma}{\sqrt{w_d}} \quad (0)$$

Where σ was the standard deviation of interactions from the LOESS smoothing at distance *d* and w_d was the sum of the weights from the LOESS smoothing at distance *d*. Thus, the presence of a chromatin looping interaction can be inferred when the observed 5C signal obtained for a specific pair of loci is higher that its expected value. For example, the interaction between the α-globin genes and HS40 in K562 cells is ~4 times more frequent that the expected level of interaction (**Fig. 1b**). In contrast, in GM12878 cells that do not express the α-globin genes the interaction between these genes and HS40 is as frequent as expected for random collisions between sites separated by the corresponding genomic site separation, and thus we conclude that no looping interaction between these genomic loci occurs in these cells.

5C data normalization

Nature Structural & Molecular Biology: doi:10.1038/nsmb.1936

5C experimental data results in interaction counts between studied restriction fragments (*i.e.*, the quantitative determination of the number of times each specific 5C ligation product is sequenced). We applied an internal normalization by mean of *Z*-scoring the sequence counts data. The *Z*-score calculation required that all input data followed a normal distribution centered on its average. However, raw 5C data did not follow a normal distribution and values were thus transformed by applying a $\log_{10}$ to the raw data. With such normalization, the *Z*-scores of the $\log_{10}$ values of the raw frequencies for interacting fragments *i* and *j* were computed as:

$$Zscore_{i,j} = \frac{\left(\mu - f_{i,j}\right)}{\sigma} \quad (1)$$

where $f_{i,j}$ was the $\log_{10}$ 5C frequency between fragments *i* and *j*, and μ and σ were the average and standard deviation of the $\log_{10}$ frequencies of the whole 5C matrix. Such normalization allowed us to quantify the variability within the 5C matrix as well as to identify pairs of fragments that interact above or below the average interaction frequency.

Model representation and scoring function

Each restriction fragment resulting from the 5C experiment design was represented by a particle in the 3D space (that is, a point determined by its Cartesian coordinates). Thus, the 70 restriction fragments from the ENm008 region (**Supplementary Table 1**) were represented by 70 particles with an excluded volume proportional to their nucleotide length (*l*). The excluded volume was set so that two particles representing two restriction fragments did not overlap in the 3D space proportionally to their size in nucleotides Thus, a particle *i* was set to have an excluded volume of radius r_i equal to:

Nature Structural & Molecular Biology: doi:10.1038/nsmb.1936

$$r_i = 0.005 \cdot l_i \quad (2)$$

Supplementary Fig. 3 shows snapshots of the ENm008 simulations for K562 cell line using a "ball-and-stick" representation, where balls are proportional to the radius of their excluded volume and "imaginary" sticks link contiguous restriction fragments or particles. It is important to note that for IMP, there are no sticks or physical links connecting two contiguous particles and such "imaginary" sticks can cross each other during simulation.

The spatial position of each particle was determined by satisfying series of restraining oscillators (or springs) implemented between pairs particles, which aimed at maintaining them at a given equilibrium distance. In our simulations, both neighbor (*i.e.*, separated by a maximum of 1 particle) and non-neighbor particles (*i.e.*, separated by 2 or more particles) were restrained at equilibrium distances inversely proportional to their interacting 5C *Z*-scores. Three types of different restraints were used for modeling the ENm008 region: (i) harmonic oscillators ($H_{i,j}$), which ensured a pair of particles to lie at about a given equilibrium distance; (ii) lower-bound harmonic oscillators ($lbH_{i,j}$), which ensured that two particles could not get closer than a given equilibrium distance and; (iii) upper-bound harmonic oscillators ($ubH_{i,j}$), which ensured that two particles could not get separated beyond a given equilibrium distance. The exact functions of the restraints were:

$$H_{i,j} = k\left(d_{i,j} - d_{i,j}^0\right)^2 \quad (3)$$

$$\begin{cases} if\ \ d_{i,j} \le d_{i,j}^0; \quad lbH_{i,j} = k\left(d_{i,j} - d_{i,j}^0\right)^2 \\ if\ \ d_{i,j} > d_{i,j}^0; \quad lbH_{i,j} = 0 \end{cases} \quad (4)$$

Nature Structural & Molecular Biology: doi:10.1038/nsmb.1936

$$\begin{cases} if\ \ d_{i,j} \geq d^0_{i,j}; \quad ubH_{i,j} = k\left(d_{i,j} - d^0_{i,j}\right)^2 \\ if\ \ d_{i,j} < d^0_{i,j}; \quad ubH_{i,j} = 0 \end{cases} \tag{5}$$

where $d_{i,j}$ is the current distance between particles *i* and *j* during simulation, $d^0_{i,j}$ is the equilibrium distance obtained from the transformation of the 5C *Z*-scores into distances (above), and *k* is the force constant applied to the restraint, which scaled the penalty added to the IMP objective function for not satisfying it. For a pair of restrained particles, *k* was set to the square root of the absolute value of the 5C *Z*-score between them. Such setting made extreme values both for low and high raw 5C *Z*-scores to be restrained with larger *k* forces.

The type of restraint (i.e., $H_{i,j}$, $lbH_{i,j}$, or $ubH_{i,j}$) and the equilibrium distance applied to each particle were defined based on the 5C experimental data and three IMP parameters: (i) a lower-bound *Z*-score cut-off (*lZ*), (ii) a upper-bound *Z*-score cut-off (*uZ*), and (iii) a maximal proximity for two non-interacting fragments (*mP*). Identifying the optimal value for the three parameters constituted what we call "IMP calibration" and is described in detail below (section *Empirical determination of IMP parameters*). Interaction *Z*-scores between the *lZ* and *uZ* parameters, which corresponded to *Z*-scores near zero and thus with close to average interaction frequencies, were not used during modeling by IMP. IMP scoring function used then 5C data for pairs of fragments with *Z*-scores below *lZ* and above *uZ,* which corresponded to low or high interaction frequencies, respectively. Such approach allowed us to identify those pairs of interacting fragments that had either very low or very high interaction frequencies. Finally, the *mP* parameter set the closest distance between two pairs of non-interacting fragments (*i.e.*, 5C interaction frequency

Nature Structural & Molecular Biology: doi:10.1038/nsmb.1936

of zero). These three parameters were determined empirically for each cell type experiment (below).

Equilibrium distances were set to be inversely proportional to the 5C *Z*-scores. Two different linear relationships were defined for neighbor (*i.e.*, *i* to *i*+1..2) and non-neighbor (*i.e.*, *i* to *i*+3..*n*) fragments. First, neighbor fragments were separated at an equilibrium distance proportional to the sum of their occupied excluded volume. For 5C experiments with K562 cells, the non-neighbor linear relationship was set to be bound by the pairs of points (3.31, 30), corresponding to the maximum *Z*-score value and the closest distance between two condensed chromatin fragments, and (-1.42, 400), corresponding to the minimum *Z*-score value and *mP* parameter optimized for the K562 5C matrix. Similarly, for 5C experiments with GM12878 cells, the non-neighbor linear relationship was set to be bound by the pairs of points (3.66, 30) and (-2.90, 500). The optimal parameters for GM12878 cells corresponded to 500 nm for *mP*, -0.2 for *lZ*, and 0.1 for *uZ* (**Supplementary Fig. 3a**). The optimal parameters for K562 cells corresponded to 400 nm, -0.1, and 0.9 for *mP*, *lZ* and *uZ*, respectively (**Supplementary Fig. 3b**).

The type of harmonic restraint applied to a pair of particles depended on whether the pairs of particles were neighbors or non-neighbors as well as on *lZ* and *uZ*. First, two neighbor particles with calculated 5C *Z*-scores were restrained by a harmonic oscillator with an equilibrium distance proportional to their 5C *Z*-score following the neighbor linear relationship. Due to the presence of repetitive elements in the genome, 15 of the 70 restriction fragments were not interrogated in the 5C analysis because no unique 5C primer could be designed (**Supplementary Table 1**). Therefore, two neighbor particles with no calculated 5C *Z*-scores were restrained by an upper-bound harmonic oscillator

Nature Structural & Molecular Biology: doi:10.1038/nsmb.1936

with an equilibrium distance corresponding to the sequence length of the intermediate fragment between their fragment centers. A *k* force of 5 was applied to ensure connectivity between neighbor fragments. Second, two non-neighbor particles with calculated 5C *Z*-scores were modeled at a distance and force proportional to their corresponding 5C *Z*-scores following the non-neighbor linear relationship described above. Pairs of particles with *Z*-scores higher than the upper-bound cut-off were restrained by a harmonic oscillator and pairs of particles with *Z*-scores lower than the lower-bound cut-off were restrained by a lower-bound harmonic oscillator. These two harmonic oscillator types aim at keeping a pair of particles at an equilibrium distance or further apart from a minimal distance, respectively. Therefore, pairs of non-neighbor particles that were observed to interact with *Z*-scores above the *uZ* parameter were kept close in space, and pairs of non-neighbor particles that were observed to interact with *Z*-scores below the *lZ* parameter were kept apart in space. The *k* force applied to these restraints was set to the square root of the absolute value of their interacting *Z*-scores. Finally, pairs of non-neighbor particles for which 5C *Z*-scores were not available were restrained based on the average 5C-*Z*-score calculated from the adjacent particles.

Model building with IMP

Following the steps described above, the ENm008 region was represented by a set of 70 particles restrained by a total of 1,049 and 1,520 harmonic oscillators for GM12878 and K562 cell lines, respectively. The next step was thus to determine an ensemble of 3D conformations that satisfied as much as possible all the imposed restraints. With that aim, IMP generates structures by simultaneously minimizing the violations of all the imposed restraints. In general, the optimization of the imposed restraints may result in

Nature Structural & Molecular Biology: doi:10.1038/nsmb.1936

different configurations with similar final IMP objective function. Therefore, to comprehensively explore the conformational space, IMP was run for a total of 50,000 independent simulations resulting in 50,000 different conformational solutions for each 5C experiment. The entire calculation took about 6 days on a 200 CPU cluster. For each individual simulation, the IMP building protocol (**Supplementary Fig. 3**) starts by assigning to all particles a set of random Cartesian coordinates within a cube of 1 μm side length, which can, however, be exceeded during the optimization protocol. The optimization is carried out by a combination of 500 Monte Carlo rounds with 5 local steps in a molecular dynamics simulation with a standard simulated annealing method[5]. At each step of the optimization, the current conformation is randomly changed and the change is accepted or rejected according to the Metropolis criteria[6]. The driving scoring function that is minimized during the optimization protocol consists of the sum of all the individual restraint scores between the 70 particles representing the ENm008 region.

Empirical determination of IMP parameters

The empirical determination of the *mP*, *lZ* and *uZ* parameters was carried out over a grid search exploring the values of 300, 400, 500, 600 and 700 nm for *mP*, -0.1 to -1.0 in bins of 0.1 for *lZ* cut-off, and 0.1 to 1.0 in bins of 0.1 for *uZ* cut-off, which were determined using the following procedure: (i) for each set of parameters, 500 models were generated using the protocol described in the previous section; (ii) from the resulting 500 conformations, a frequency contact map counting, for each solution, whether two particles were in contact (*i.e.*, within 200 nm separation) was calculated; and (iii) the correlation coefficient between the calculated frequency contact map and the 5C counts matrix used as input data in the modeling protocol was obtained. Thus, the optimal

Nature Structural & Molecular Biology: doi:10.1038/nsmb.1936

values corresponded to the grid cell with the maximum correlation coefficient between the frequency contact map calculated from a set of 500 3D models and the raw 5C counts. In other words, we selected a set of *mP*, *lZ* and *uZ* optimal parameters that resulted in the 3D models that best represented the input 5C raw data. Ideally, the correlation coefficient between the two matrices (*i.e.,* 5C counts and 3D models contact maps) would be near 1.0, indicating that the resulting ensemble of models explains all the input 5C data. However, 5C experiments capture the ensemble macroscopic state of chromatin in a population of cells and the resulting correlation coefficient is expected to be lower than 1.0. Indeed, for an optimal set of parameters the maximum correlation coefficient was 0.75 and 0.69 for GM12878 and K562 experiments, respectively. The same protocol was used to empirically determine the optimal parameters for the ensemble analysis (below).

Ensemble analysis

To make the structural analysis computationally feasible, the 10,000 solutions with the lowest IMP objective function (*i.e.,* closer to the optimal solution where all restraints are satisfied) were selected out of all the 50,000 simulations. The analysis of the selected conformations was facilitated by structurally superposing them using pair-wise rigid-body superposition that minimizes the RMSD between the superposed conformations[7]. The resulting comparison matrix, which consisted of an all-against-all equivalent position score within an empirically determined 75 nm distance cut-off, was input to the Markov Cluster Algorithm (MCL) program[8] for generating unsupervised sets of clusters of related structures. Two main parameters affect the cluster granularity in the MCL program. That is, the pre-inflation parameter (-pi) and the inflation parameter (-I). Using the

Nature Structural & Molecular Biology: doi:10.1038/nsmb.1936

protocol outlined above, we determined the optimal parameters for MCL that resulted in the highest correlation between the frequency contact map calculated from the top cluster and the input 5C count matrix. For the GM12878 experiment, the optimal parameters for MCL clustering were: 5.0 for the MCL pre-inflation parameter, and 2.0 for the MCL inflation parameter. Using these parameters, the 10,000 selected solutions resulted in 4 clusters of superposed solutions with the top cluster accounting for 29% of the 10,000 solutions. For the K562 experiment, the optimal parameters for MCL clustering were: 10.0 for the MCL pre-inflation parameter, and 2.0 for the MCL inflation parameter. Using these parameters, the 10,000 selected solutions resulted in 393 clusters of superposed solutions with the top 10 largest clusters accounting for 26% of the 10,000 solutions. It is important to note that for both cell types, the top two clusters corresponded to mirror images of each other. IMP generates solutions in Cartesian space, which however, are scored in the distance space by the degree of satisfaction of imposed restraints. Therefore, mirror solutions of an object would account for the same distances between points and thus result in the same IMP objective function.

5C de-convolution analysis

Given that 5C interaction matrices can be seen as an average state of the cell population, they are not sufficient to discern between mutually exclusive and co-occurring interactions that may take place in the diverse states (that is, in different cells) that the cell nucleus may adopt. Therefore, we de-convoluted the original 5C interaction matrix by comparing the contact frequency maps calculated from the different clustered 3D solutions. This analysis allowed us to identify specific interaction differences between clusters of solutions. Large differences in contact frequencies (*i.e.,* >25%)

Nature Structural & Molecular Biology: doi:10.1038/nsmb.1936

aided in de-convoluting the population averaged 5C interaction matrix, which provided a way of identifying fragment interactions that may partially explain the original 5C input dataset.

Pair-wise comparisons were performed to identify differences in long-range interactions between clusters 1 to 10 from the analysis of K562 cells (**Supplementary Fig. 5**). Such differences are likely to arise from sets of mutually excluding interactions, which cannot co-occur in a single conformation. For example, interactions occurring between the set of fragments 38, 43 and 45 and the set of fragments 49, 50 and 51 (*Z*-scores in the 5C dataset between 0.88 and 1.34) are underrepresented in cluster 2 compared to cluster 10. Conversely, interactions between fragments 11 and 35 are 30% more frequent in cluster 2 compared to cluster 10, which resulted in a similar *Z*-score of 0.98 in the original 5C analysis. Thus, whereas the 5C experiments provide only population-averaged data, our structural approach provides a means for assigning subsets of the 5C data to specific domain conformations, which is critical in identifying co-occurring and mutually excluding interactions.

Effective resolution of the ENm008 3D models

Two factors affect the precision or resolution of our models: (i) the size (bp) of 5C restriction fragments and (ii) the ensemble of solutions of the final selected cluster. To assess the effective resolution of our generated models, the actual occupancy of all particles in the selected clusters was represented by a density map calculated as a Gaussian function of variable standard deviation. The standard deviation applied to the Gaussian function that could explain at least 80% of the occupancy of the models was assessed to be the effective resolution of the ensemble of solutions representing the 3D

Nature Structural & Molecular Biology: doi:10.1038/nsmb.1936

structure of the EMm008 region. A standard deviation of 175 nm was assessed for both GM12878 and K562 cells (**Supplementary Fig. 5**). It is important to note that the 3D positions obtained by IMP correspond to points representing the center of the ligation positions designed as part of the 5C experiments. The path between points shown in our 3D models does not necessarily correspond to the path that chromatin may follow *in vivo*.

Calculation of relative abundance of restriction fragments versus radial position in globules

The following protocol was used to calculate the relative abundance of fragments containing promoters, active genes, no active genes, DNaseI hypersensitive sites, CTCF sites or H3K4me3 modifications (in **Supplementary Table 1** named as PR, AG, NA, HS, CT, and HM, respectively) at various radial positions in the globules (**Fig. 5b**). The ENCODE data for ENm008 region was obtained from the UCSC Genome Browser (http://genome.ucsc.edu/ENCODE/) tracks for: RefSeq annotated genes[9], Affymetrix/CSHL expression data (Gingeras Group at Cold Spring Harbor), Duke/NHGRI DNaseI Hypersensitivity data[10] (Crawford Group at Duke University), and Histone Modifications by Broad Institute ChIP-seq (Bernstein Group at Broad Institute of Harvard and MIT).

First, we defined chromatin globules by visually inspecting the 3D models in the selected clusters (Cluster 1 for GM12878 and cluster 2 for K562). GM12878 models showed a single globule encompassing fragments 1 to 70 and K562 models showed two globules encompassing fragments 1 to 48 and 58 to 70. Second, we calculated a center coordinates for all fragments in each globule. The analysis was carried out only to the

Nature Structural & Molecular Biology: doi:10.1038/nsmb.1936

single globule of GM12878 and the first globule of K562. The second globule in K562 was omitted due to its small size and its partial representation (*i.e.,* models reached only to the genomic coordinates 499,411 in chromosome 16). Third, we calculated the distance of each fragment to the globule center coordinates. Fourth, from the closest fragment to the center (*i.e.*, avoiding the empty globule core), we generated a series of concentric spheres of 50 nm up to 400 nm. Fifth, we calculated the number of fragments within each concentric sphere. Sixth, we calculated the relative abundance ($RA_{t,d}$) of each fragment type *t* (with *t* = PR (Promoter), AG (Active Gene) etc.) and at each distance cut-off *d* by:

$$RA_{t,d} = \frac{n_{t,d}}{n_t} \Big/ \frac{n_d}{N} \qquad (6)$$

where $n_{t,,d}$ is the number of fragments of type *t* within distance cut-off *d*, $n_{t,}$ is the number of fragments of type *t,* $n_{d,}$ is the number of fragments within distance cut-off *d,* and *N* is the total number of fragments in the globule. Thus, values of $RA_{t,d}$ larger than 1 indicate over-representation of fragments of type *t* within a distance cut-off *d* of the center of the globule. Conversely, values of $RA_{t,d}$ smaller than 1 indicate under-representation of fragments of type *t* within a distance cut-off *d* of the center of the globule.

Ensemble visualization

The UCSF Chimera package[3], a highly extensible program for interactive visualization of molecular structures, was used to produce all graphics images and to analyze the resulting ensemble of solutions. First, to visually inspect the most likely path of an ensemble of solutions (or cluster), the centroid of the cluster was calculated as the solution that best superposes the average structure of the cluster. Such selection

Nature Structural & Molecular Biology: doi:10.1038/nsmb.1936

criterion, rather than an average of the ensemble itself, warrants that the final selected path representing the ensemble solution is consistent with the input experimental data. The centroid path and the occupancy of the ensemble of solutions were represented in Chimera by using the *volume path tracer* and the *molmap* tools, respectively.

Nature Structural & Molecular Biology: doi:10.1038/nsmb.1936

Supplementary References

1. Alber, F. et al. Determining the architectures of macromolecular assemblies. *Nature* **450**, 683-94 (2007).
2. Wagner, G. et al. Protein structures in solution by nuclear magnetic resonance and distance geometry. The polypeptide fold of the basic pancreatic trypsin inhibitor determined using two different algorithms, DISGEO and DISMAN. *J Mol Biol* **196**, 611-39 (1987).
3. Pettersen, E.F. et al. UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605-12 (2004).
4. Dekker, J. The three 'C' s of chromosome conformation capture: controls, controls, controls. *Nat Methods* **3**, 17-21 (2006).
5. Kirkpatrick, S., Gelatt, C.D., Jr. & Vecchi, M.P. Optimization by Simulated Annealing. *Science* **220**, 671-680 (1983).
6. Metropolis, N. & Ulam, S. The Monte Carlo method. *J Am Stat Assoc* **44**, 335-41 (1949).
7. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702-10 (2004).
8. Enright, A.J., Van Dongen, S. & Ouzounis, C.A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575-1584 (2002).
9. Pruitt, K.D., Tatusova, T. & Maglott, D.R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, D61-5 (2007).
10. Crawford, G.E. et al. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res* **16**, 123-31 (2006).

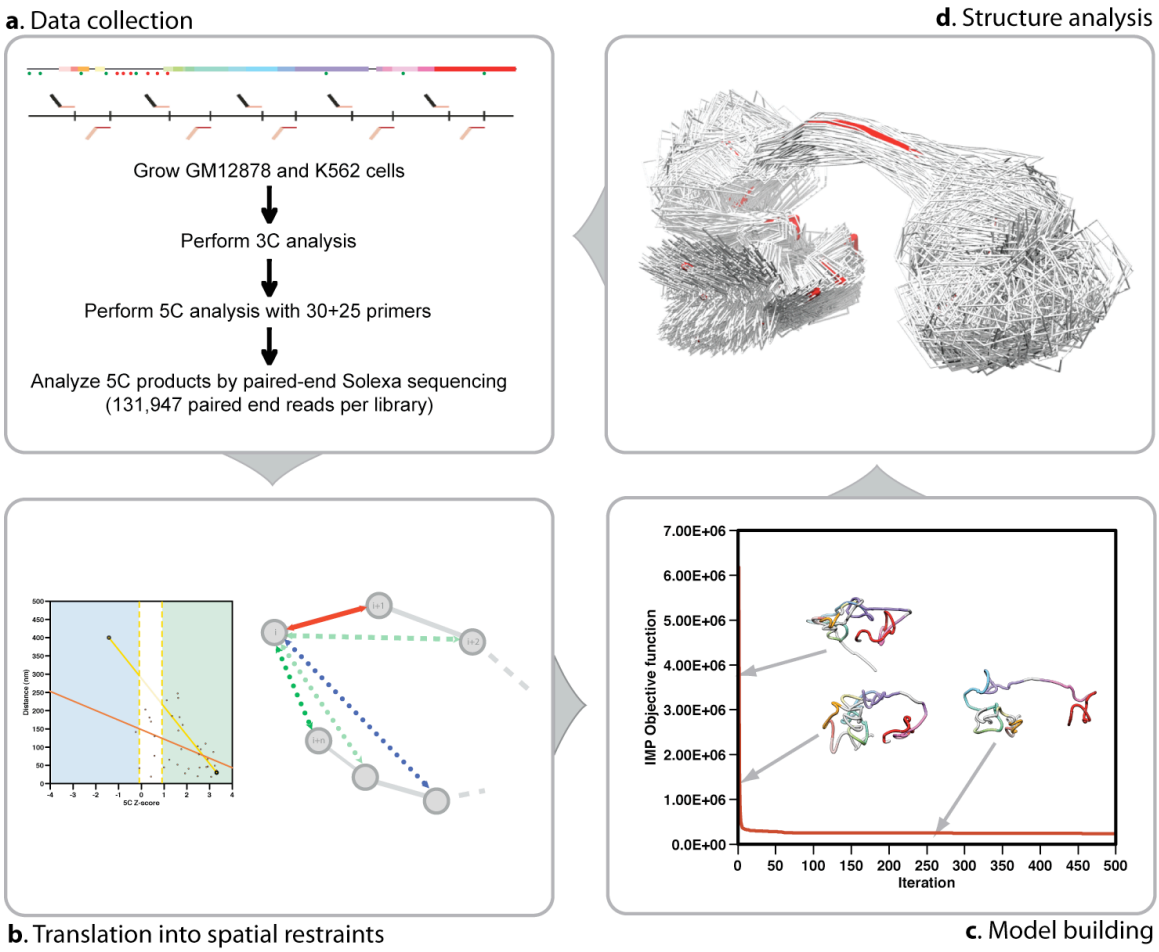
Nature Structural & Molecular Biology: doi:10.1038/nsmb.1936

Supplementary Table 1. Restriction fragment data of the ENm008 region. The table includes: the starting and ending coordinates of each fragment, nucleotide length, particle radii, FISH probe, annotated RefSeq genes, and assigned fragment type based on the ENCODE data. Fragment types are: promoters (PR), active genes (AG), no-active gene (NA), DNaseI hypersensitive site (HS), CTCF site (CT), and H3K4me3 site (HM). Fragments annotated as “Left out” were not queried during the 5C experiment. 5C counts for fragments 31 and 32 were combined because of the sequence of the corresponding 5C primers is identical.

Nature Structural & Molecular Biology: doi:10.1038/nsmb.1936

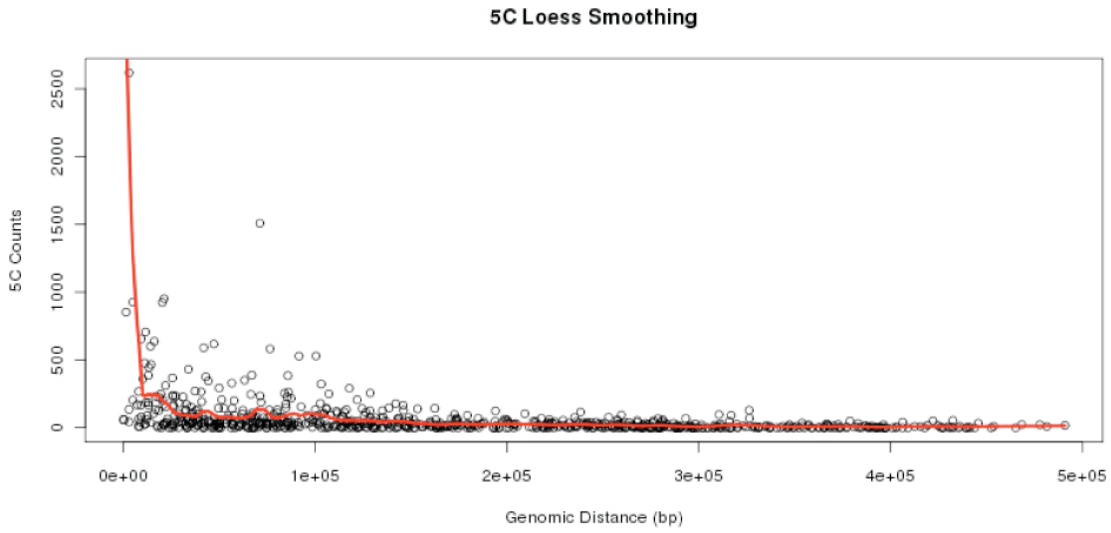
	Fragment	Start	End	Length (nt)	Radii (nm)	FISH probes	RefSeq genes	Assigned fragment types based on the ENCODE data											
								GM12878						K562 (globule 1)					
								PR	AG	NA	HS	CT	HM	PR	AG	NA	HS	CT	HM
1	Reverse 1	1	5693	5692	28.5														
2	Left out 2-5	5693	11138	5445	27.2														
3	Reverse 6	11138	15091	3953	19.8														
4	Forward 7	15091	18344	3253	16.3														
5	Forward 8	18344	29756	11412	57.1														
6	Reverse 10	29779	44231	14452	72.3		POLR3K (olymerase (RNA) III (DNA directed) polypeptide K)												
7	Reverse 11	44231	50868	6637	33.2		SNRNP25 (small nuclear ribonucleoprotein 25kDa) RHBDF1 (rhomboid 5 homolog 1)												
8	Reverse 12	50868	52846	1978	9.9		RHBDF1 (rhomboid 5 homolog 1)												
9	Reverse 13	52846	55911	3065	15.3		RHBDF1 (rhomboid 5 homolog 1)												
10	Forward 14	55911	56690	779	3.9		RHBDF1 (rhomboid 5 homolog 1)												
11	Reverse 15	56690	64056	7366	36.8		RHBDF1 (rhomboid 5 homolog 1)												
12	Forward 16	64056	65723	1667	8.3														
13	Reverse 17	65723	74448	8725	43.6		MPG (N-methylpurine-DNA glycosylase)												
14	Reverse 18	74448	86128	11680	58.4		MPG (N-methylpurine-DNA glycosylase)												
15	Forward 19	86128	86217	89	0.4														
16	Forward 20	86217	87749	1532	7.7														
17	Reverse 21	87749	90248	2499	12.5														
18	Forward 22	90248	91497	1249	6.2														
19	Forward 23	91497	95256	3759	18.8		HS48												
20	Forward 24	95256	100530	5274	26.4		HS48												
21	Forward 25	100530	104687	4157	20.8		HS40												
22	Reverse 26	104687	109838	5151	25.8														
23	Left out 27	109838	120933	11095	55.5		HS33												
24	Reverse 28	120933	131220	10287	51.4														
25	Forward 29	131220	134334	3114	15.6		HS10												
26	Left out 30	134334	147782	13448	67.2		HBZ (hemoglobin zeta) & HS8												
27	Reverse 31+32	147782	167103	19321	96.6		HBM (hemoglobin, mu) HBA2 (hemoglobin, alpha 2) HBA1 (hemoglobin, alpha 1)												
28	Reverse 33	167103	171769	4666	23.3		HBQ1 (hemoglobin, theta 1)												
29	Reverse 34	171769	185994	14225	71.1		LUC7L (LUC7-like)												
30	Forward 35	185994	189074	3080	15.4		LUC7L (LUC7-like)												
31	Left out 36	189074	192185	3111	15.6		LUC7L (LUC7-like)												
32	Reverse 37	192185	203353	11168	55.8		LUC7L (LUC7-like)												
33	Reverse 38	203353	217802	14449	72.2		LUC7L (LUC7-like)												
34	Reverse 39	217802	225341	7539	37.7		LUC7L (LUC7-like) ITFG3 (integrin alpha FG-GAP repeat)												
35	Reverse 40	225341	235762	10421	52.1		ITFG3 (integrin alpha FG-GAP repeat)												
36	Left out 41	235762	237363	1601	8.0		ITFG3 (integrin alpha FG-GAP repeat)												
37	Forward 42	237363	239095	1732	8.7		ITFG3 (integrin alpha FG-GAP repeat)												
38	Reverse 43	239095	247214	8119	40.6		ITFG3 (integrin alpha FG-GAP repeat)												
39	Left out 44	247214	260279	13065	65.3		ITFG3 (integrin alpha FG-GAP repeat) RGS11 (regulator of G-protein signaling 11)												
40	Reverse 45	260279	277942	17663	88.3		RGS11 (regulator of G-protein signaling 11) ARHGDIG (Rho GDP dissociation inhibitor gamma) PDIA2 (protein disulfide isomerase family A, member 2)												
41	Left out 46	277942	286059	8117	40.6		AXIN1 (axin 1)												
42	Forward 47	286059	289198	3139	15.7		AXIN1 (axin 1)												
43	Forward 48	289198	303867	14669	73.3		AXIN1 (axin 1)												
44	Forward 49	303867	310528	6661	33.3		AXIN1 (axin 1)												
45	Forward 50	310528	314538	4010	20.1		AXIN1 (axin 1)												
46	Forward 51	314538	327240	12702	63.5		AXIN1 (axin 1)												
47	Left out 52	327240	331940	4700	23.5		AXIN1 (axin 1)												
48	Forward 53	331940	332109	169	0.8		AXIN1 (axin 1)												
49	Forward 54	332109	334889	2780	13.9		AXIN1 (axin 1)												
50	Forward 55	334889	335401	512	2.6		AXIN1 (axin 1)												
51	Left out 56	335401	346310	10909	54.5		AXIN1 (axin 1)												
52	Forward 57	346310	352385	6075	30.4														
53	Forward 58	352385	353009	624	3.1														
54	Reverse 59	353009	360924	7915	39.6		MRPL28 (mitochondrial ribosomal protein L28)												
55	Reverse 60	360924	372670	11746	58.7		TMEM8 (transmembrane protein 8)												
56	Left out 61	372670	376971	4301	21.5		LOC100134368 (similar to hCG1644121 ncRNA)												
57	Reverse 62	376971	380160	3189	15.9		LOC100134368 (similar to hCG1644121 ncRNA)												
58	Reverse 63	380160	401140	20980	104.9		LOC100134368 (similar to hCG1644121 ncRNA)												
59	Left out 64	401140	402437	1297	6.5		DECR2 (2,4-dienoyl CoA reductase 2)												
60	Reverse 65	402437	418222	15785	78.9		RAB11FIP3 (RAB11 family interacting protein 3, class II)												
61	Forward 66	418222	421291	3069	15.3		RAB11FIP3 (RAB11 family interacting protein 3, class II)												
62	Forward 67	421291	433293	12002	60.0		RAB11FIP3 (RAB11 family interacting protein 3, class II)												
63	Left out 68	433293	441045	7752	38.8		RAB11FIP3 (RAB11 family interacting protein 3, class II)												
64	Forward 69	441045	445126	4081	20.4		RAB11FIP3 (RAB11 family interacting protein 3, class II)												
65	Forward 70	445126	447073	1947	9.7		RAB11FIP3 (RAB11 family interacting protein 3, class II)												
66	Forward 71	447073	454365	7292	36.5		RAB11FIP3 (RAB11 family interacting protein 3, class II)												
67	Reverse 72	454365	483412	29047	145.2		RAB11FIP3 (RAB11 family interacting protein 3, class II)												
68	Forward 73	483412	483472	60	0.3		RAB11FIP3 (RAB11 family interacting protein 3, class II)												
69	Reverse 74	483472	496513	13041	65.2		RAB11FIP3 (RAB11 family interacting protein 3, class II)												
70	Forward 75	496513	499411	2898	14.5		RAB11FIP3 (RAB11 family interacting protein 3, class II)												

Nature Structural & Molecular Biology: doi:10.1038/nsmb.1936



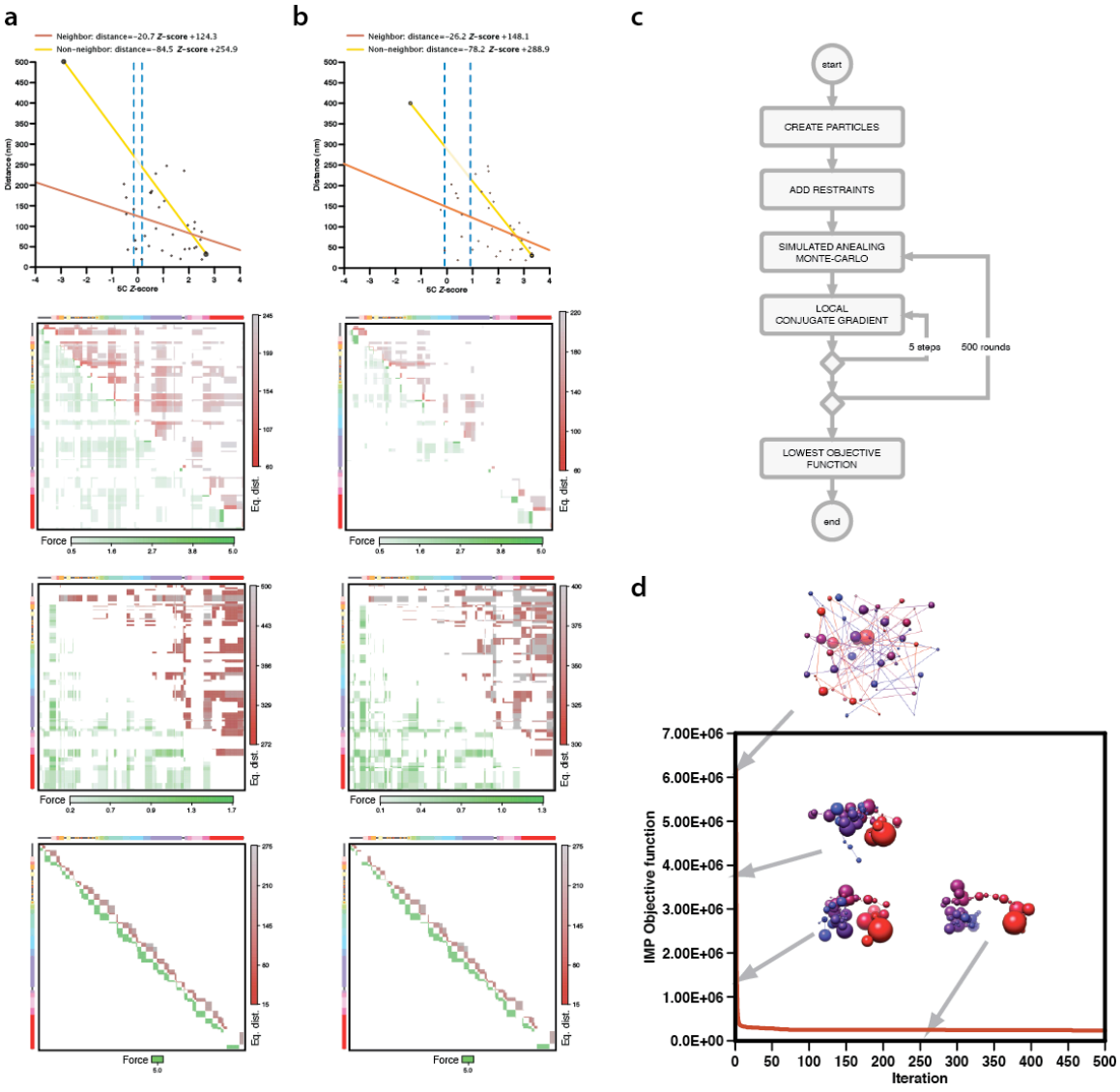
Supplementary Figure 1. General approach for determining the 3D structure of genomic domains. (**a**) 5C data collection. (**b**) Translation of experimental 5C counts into spatial points and restraints between them. (**c**) Model building by minimizing the imposed restraints. (**d**) Model ensemble analysis.

Nature Structural & Molecular Biology: doi:10.1038/nsmb.1936



Supplementary Figure 2. 5C counts for all 750 interactions detected in K562 cells within ENm008 were plotted against the genomic distance between the corresponding restriction fragments. The average expected level of interaction was determined using LOESS smoothing (α = 0.05) (red line). The average profile provides an estimate for the level of interaction expected when no specific chromatin looping interactions occur. Expected interaction frequencies decrease for loci located farther from the anchor element.

Nature Structural & Molecular Biology: doi:10.1038/nsmb.1936



Supplementary Figure 3. IMP calibration and optimization. (**a**) IMP calibration for GM12878 cells. Upper plot shows the linear relationship between 5C *Z*-scores and equilibrium distance between neighbor (red linear fitting) and non-neighbor fragments (yellow line). Two vertical dashed blue lines indicate upper- and lower-*Z*-scores cut-offs. Lower plots show harmonic, lower-bound harmonic and upper-bound harmonic equilibrium distances and forces applied to pairs of restrained fragments during simulation, respectively. Upper-right corner, red to grey indicates short to large equilibrium distances. Lower-left corner, green to grey indicates strong to weak force constants. For easy inspection, the axis labels are substituted by the linear

Nature Structural & Molecular Biology: doi:10.1038/nsmb.1936

representation of the ENm008 region. (**b**) IMP calibration for K562 cell 5C data. Data are represented as in panel **a**. (**c**) Flowchart of the IMP optimization protocol used to model the ENm008 region. (**d**) Schematic representation of a typical optimization process for a single simulation corresponding to the centroid of K562 cluster 2. The modeling starts with a randomized configuration and ends with an optimal configuration after the minimization of the IMP objective function accounting for all violated restraints. Models are show for four different snapshots during the optimization. Each restriction fragment is represented as a single point of radius proportional to their excluded volume (Supplementary Table 1). Straight lines (or sticks) connect adjacent restriction fragments, which are colored from blue (starting coordinate of chromosome 16) to red (499,411 nt in chromosome 16).

Nature Structural & Molecular Biology: doi:10.1038/nsmb.1936

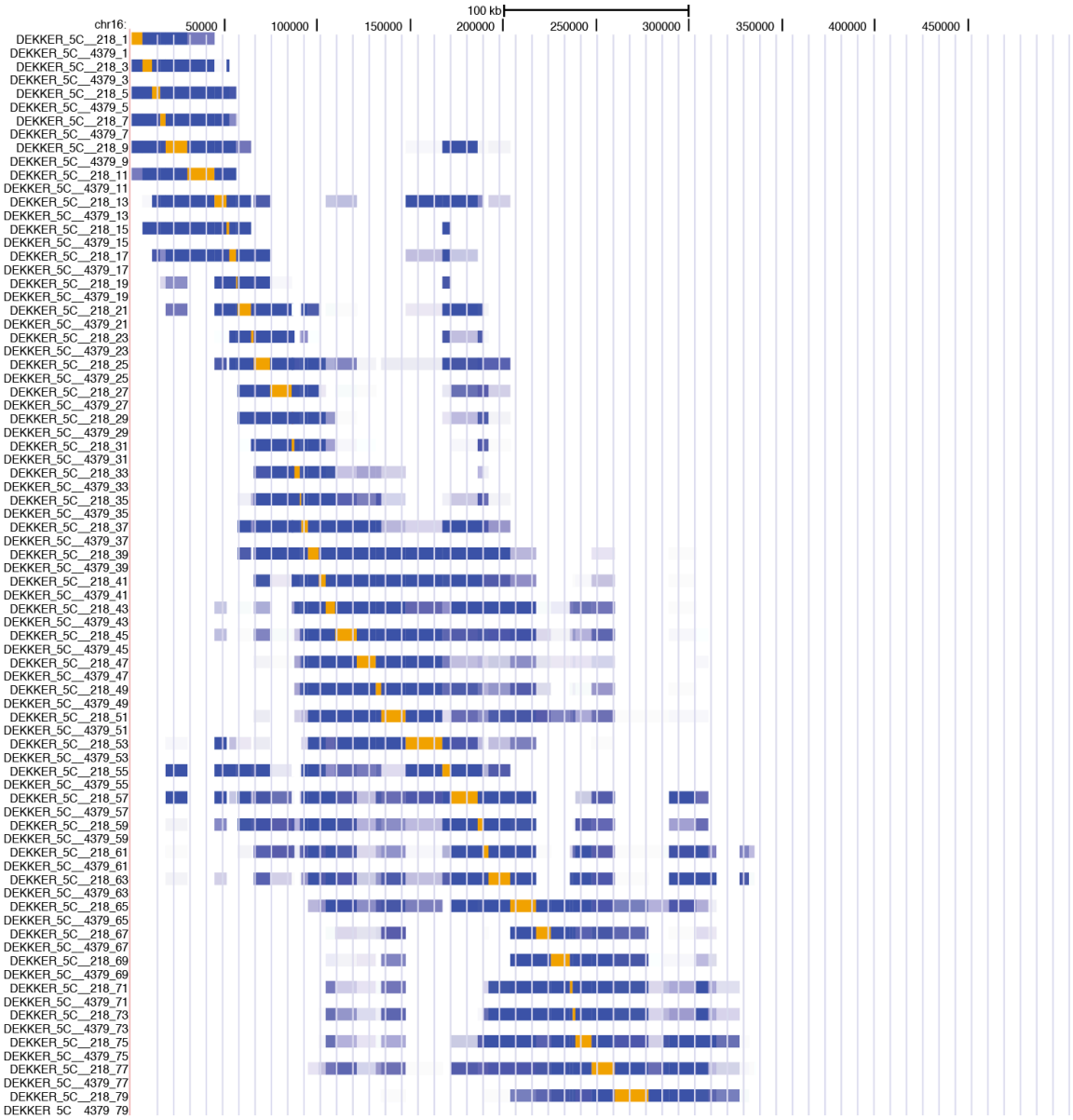
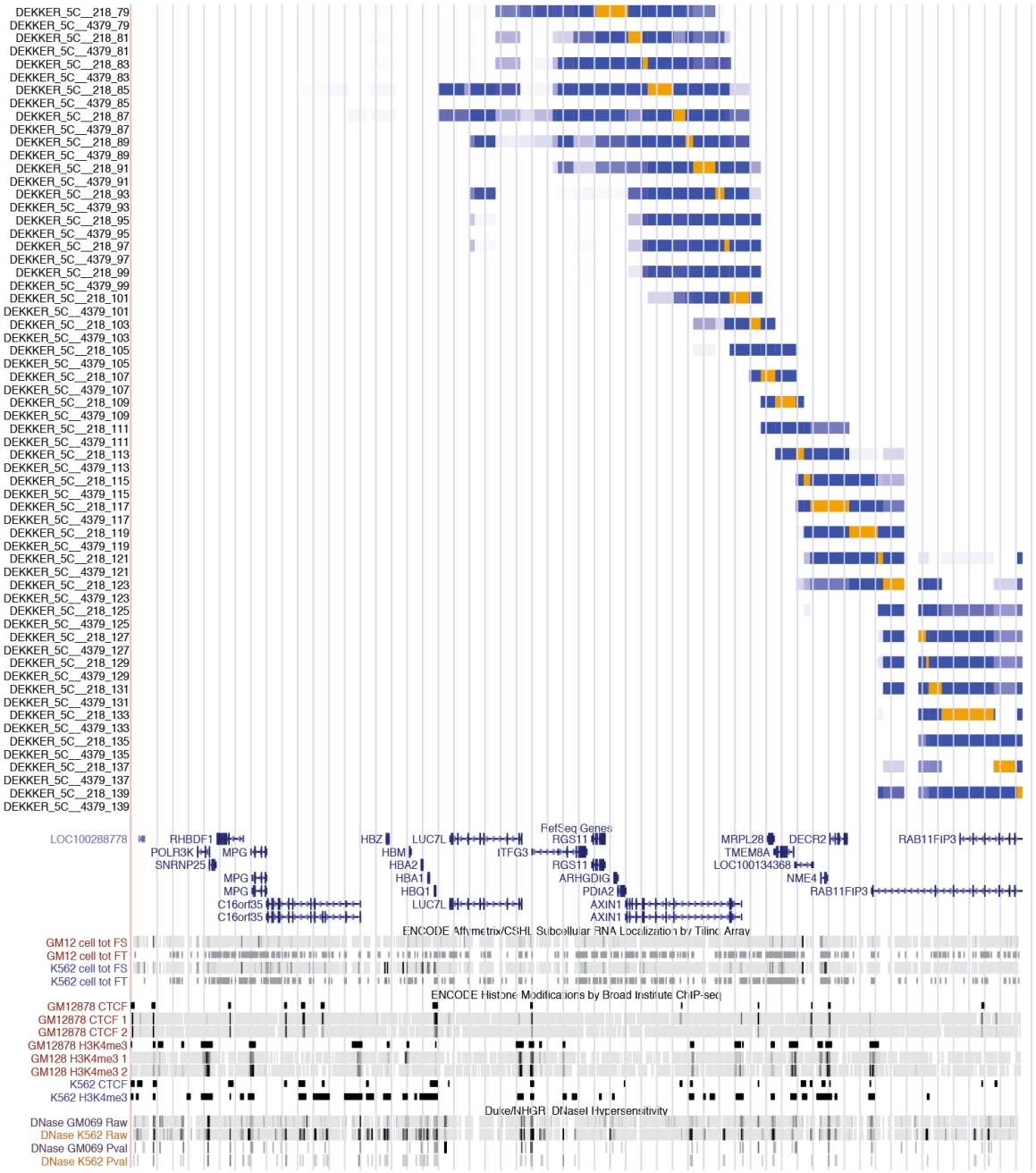


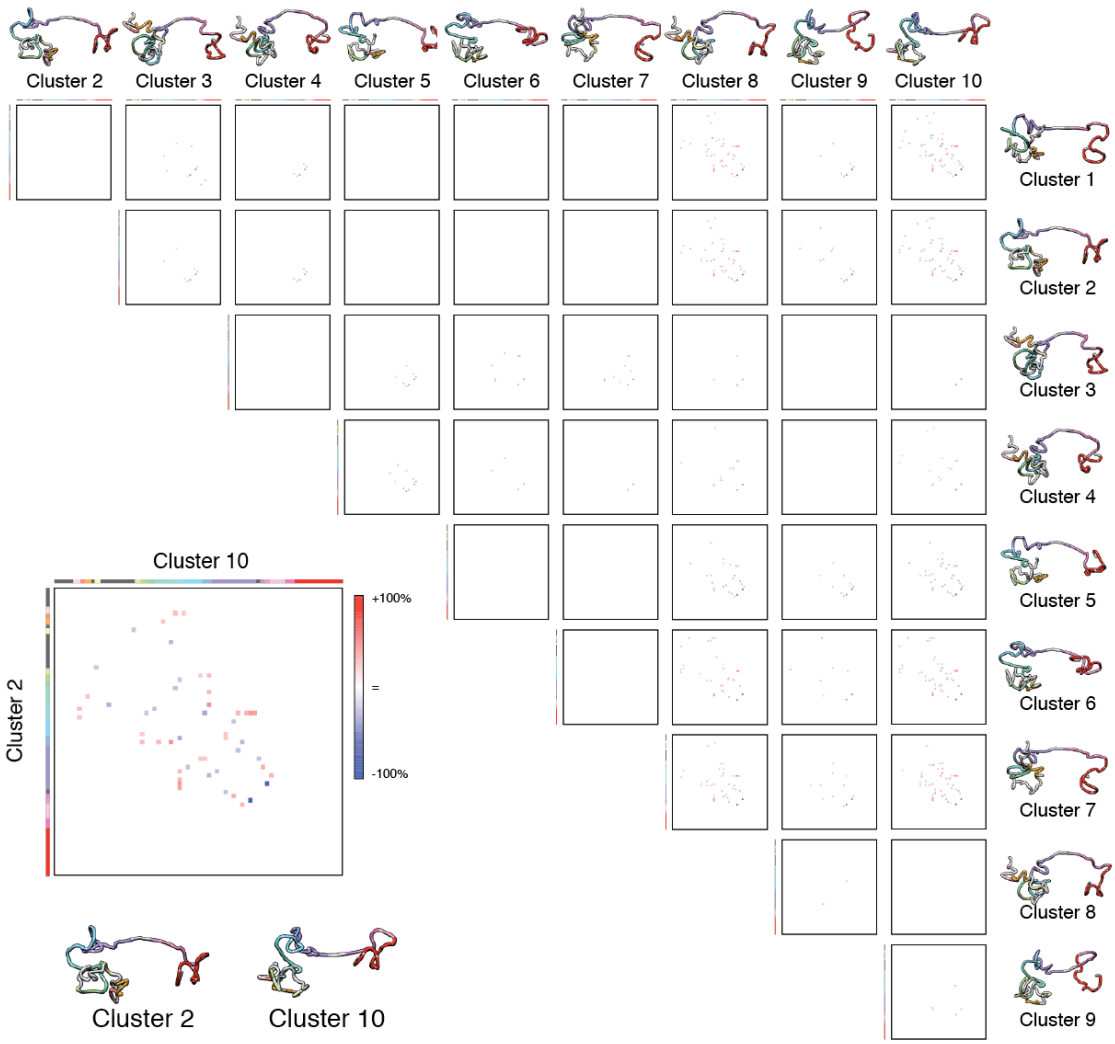
Figure 4 continues in next page…

Nature Structural & Molecular Biology: doi:10.1038/nsmb.1936



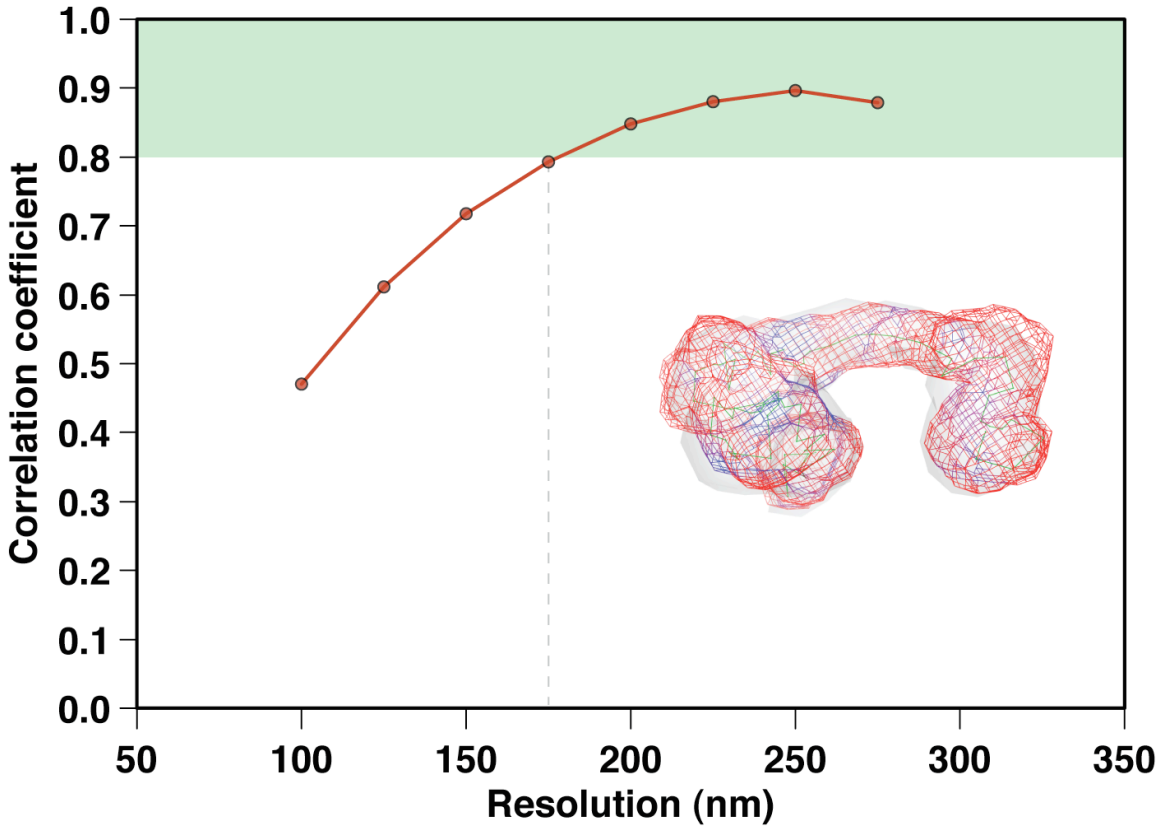
Supplementary Figure 4. 1D annotation enhanced by 3D models. UCSC Genome Browser representation of the frequency contact map calculated from the ensemble of solutions in cluster 2 of K562 models. Each track displays the long-range contacts (white to blue indicate low to high contact frequency) observed for a single restriction fragment (orange). The panel also shows the UCSC tracks used in **Fig. 1b**.

Nature Structural & Molecular Biology: doi:10.1038/nsmb.1936



Supplementary Figure 5. 5C de-convolution analysis using solution ensembles for K562 cells. Frequency contact map comparison of the top ten clusters of solutions. Red to blue dots indicates increased or decreased interacting frequencies between the compared ensembles of solutions for each cluster, respectively. Inner plot shows a detailed analysis of the comparison between cluster 2 and cluster 10 in K562 cells experiment.

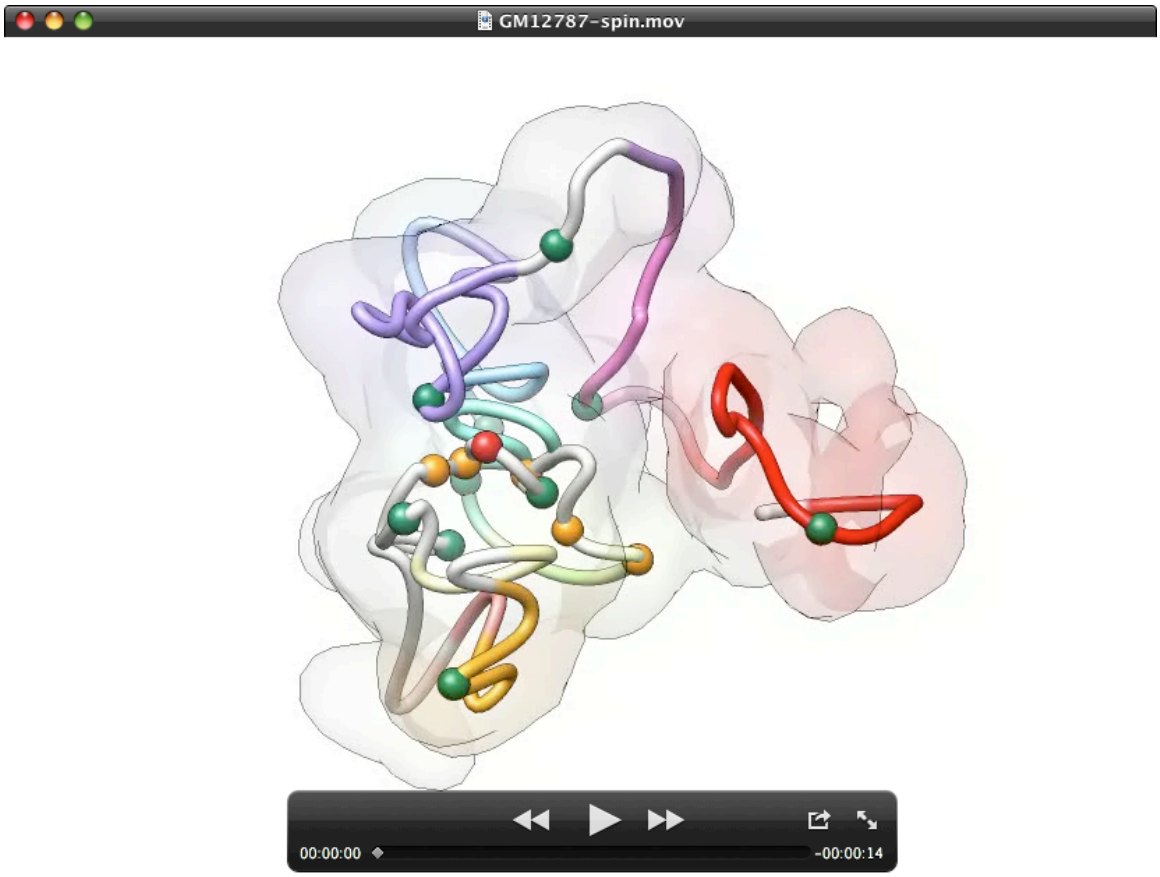
Nature Structural & Molecular Biology: doi:10.1038/nsmb.1936



Supplementary Figure 6. Model resolution. The standard deviation of the applied Gaussian to the ensemble of solutions in cluster 2 of K562 models is plotted against the correlation coefficient of the Gaussian against the actual occupancy of the models. Green background defines a similarity area where the resolution of the Gaussian covers most of the particles in the ensemble of solutions (i.e., correlation coefficient above 0.8). Inner image corresponds to the fitting of the actual model occupancy and a calculated Gaussian of 175 nm resolution.

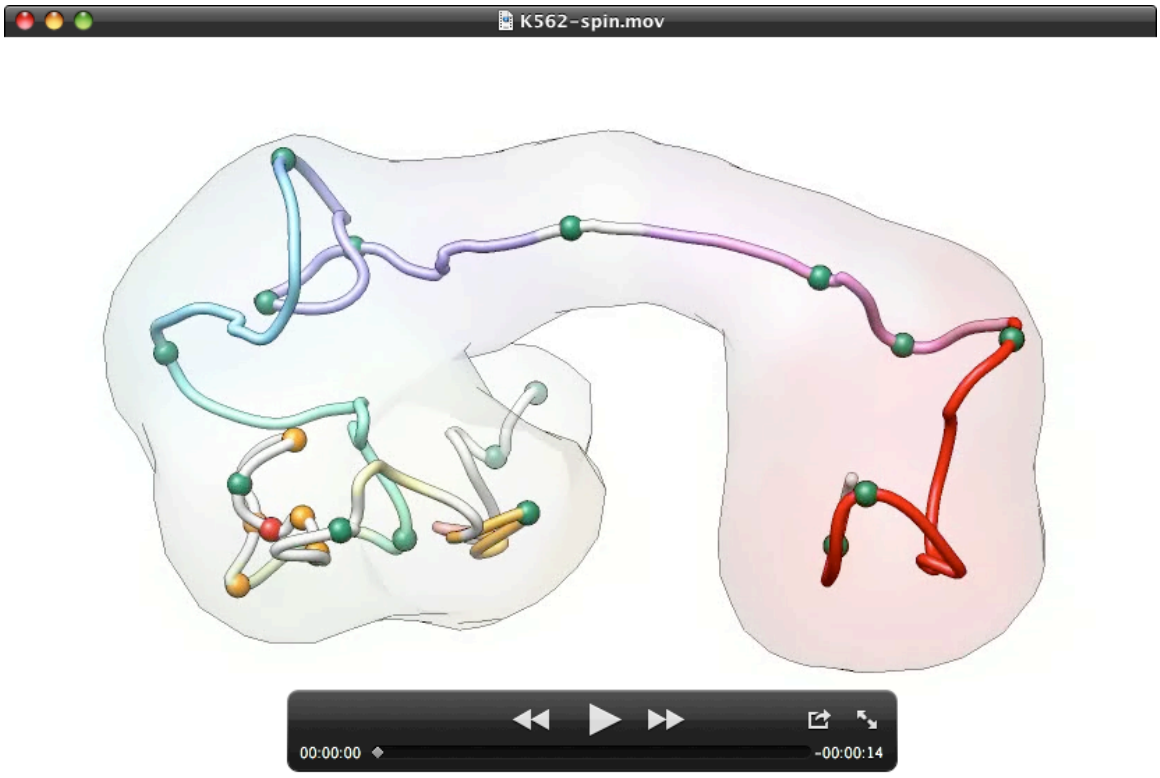
Nature Structural & Molecular Biology: doi:10.1038/nsmb.1936

Supplementary Video 1. Video of the spinning 3D structure for the ENm008 region in GM12878 cell lines. The region includes the α-globin locus, which contains, from telomere to centromere, the ζ, μ (also known as α^D), α2, α1, and θ globin genes. Colored fragments contain annotated genes. Red (HS40), orange (other HSs) and green (CTCF-bound elements) spheres localize regulatory elements.



Nature Structural & Molecular Biology: doi:10.1038/nsmb.1936

Supplementary Video 2. Video of the spinning 3D structure for the ENm008 region in K562 cell lines. The region includes the α-globin locus, which contains, from telomere to centromere, the ζ, μ (also known as α^D), α2, α1, and θ globin genes. Colored fragments contain annotated genes. Red (HS40), orange (other HSs) and green (CTCF-bound elements) spheres localize regulatory elements.



Nature Structural & Molecular Biology: doi:10.1038/nsmb.1936

Supplementary Data 1. 5C primer sequences in a tabulated text file. DNA sequences of 5C primers used for analysis of the conformation of ENm008. This is the standard output of the My5C.primers program. Columns in the tabulated file indicate:

Column 1: Primer name. The name shows whether the primer is Forward (FOR) primer or a Reverse primer (REV). The nomenclature is as follows: the name of the first forward primer is: 5C_305_ENm008_FOR_7. "5C_305" is a number that refers to the particular primer design in the My5C.primers database. "Enm008" is the name of the genomic region. "FOR_7" indicates that the primer is a forward primer and the number is the number of the *Hin*dIII fragment (numbered from the beginning of ENm008).

Column 2: Name of the genome region.

Column 3: Primer type (FOR = forward, REV = reverse).

Column 4: Genome assembly.

Column 5: The chromosome number the corresponding restriction fragment is on.

Column 6: Fragment_ID corresponds to the number of the restriction fragment, numbering starts at the beginning (5' end) of the genomic region.

Column 7: Primer_ID (1 or 2) corresponds to FOR and REV primers.

Column 8: Start position of the 5C primer (genomic coordinates).

Column 9: End position of the 5C primer (genomic coordinates).

Column 10: DNA sequence of the specific part of the 5C primer that anneals to the 3C library.

Column 11: Length (bp) of the specific part of the primer.

Column 12: DNA sequence added to the 5' end of the specific part of Forward primers or 3' end of the specific part of reverse primers (filler sequence). This DNA sequence is added to equalize the length of all 5C primers.

Column 13: Length (bp) of the filler sequence shown in Column 12.

Column 14: The melting temperature (Tm) of the specific part of the 5C primer.

Column 15: The GC percentage of the specific part of the 5C primers (sequence in column 10).

Nature Structural & Molecular Biology: doi:10.1038/nsmb.1936

Column 16: Start position of the corresponding restriction fragment (genomic coordinates).

Column 17: End position of the corresponding restriction fragment (genomic coordinates).

Column 18: Size of the corresponding restriction fragment (base pairs).

Column 19: ELEMENTID is a number that identifies any list of elements of interest the user had uploaded to My5C.primers and for which the specific 5C primer was designed.

Column 20: INTERSECTIONID is a number that identifies a specific element in the list of elements referenced in column 19.

Column 21: E_NAME is the name of the specific element (referred to in Column 20) that has intersected with this fragment.

Column 22: The 15-mer frequency of the specific part of the primer + the filler sequence. High 15-mer frequencies indicate a reduced uniqueness of the primer.

Column 23: BLAST count for the sequence of the primer containing the specific part + filler sequence (only 'exact' hits; exact means at least 20/23 bases align).

Column 24: BLAST count for the sequence of the primer containing the specific part + filler (exact+ similar hits; similar means any blast alignment).

Column 25: DNA sequence of the universal tail of the primer.

Column 26: Barcode sequence inserted at the 3' end of the universal tail (for Forward primers) or at the 5' end of the universal tail (for Reverse primers). Note that My5C.primers currently does not have the option to include barcodes. In this experiment 6-base barcodes were added to the 5C primers to facilitate mapping of DNA sequences.

Column 27: Barcode numerical code.

Column 28: Complete DNA sequence of the 5C primer.

Supplementary Data 2. 5C frequency counts matrix for ENm008 in GM12878 cells in a tabulated text file. The dataset corresponds to the data shown in **Fig. 1**. The numbers in the matrix correspond to the DNA sequence counts that were mapped to pairs of 5C

Nature Structural & Molecular Biology: doi:10.1038/nsmb.1936

primers within the ENm008 region. Columns are for reverse primers while rows are for forward primers. The names of the columns and rows (*e.g.* 5C_305_ENm008_FOR_7|hg18|chr16:15091-18344) indicate the primer name (5C_305_ENm008_FOR_7); the genome that the primer recognized (hg18 represents the human genome assembly 18); and the chromosome number and genomic coordinates (chr16:15091-18344).

Supplementary Data 3. 5C frequency counts matrix for ENm008 in K562 cells in a tabulated text file. The dataset corresponds to the data shown in **Fig. 1**. The numbers in the matrix correspond to the DNA sequence counts that were mapped to pairs of 5C primers within the ENm008 region. Columns are for reverse primers while rows are for forward primers. The names of the columns and rows are described in the legend for **Supplementary File 2**.

Supplementary Data 4. Contact map for ENm008 in GM12878 cells in a tabulated text file. 5C frequency contact maps for the ENm008 region were calculated using the 2,780 models in cluster number 1. The numbers in the matrix correspond to the number of times a particular pair of fragments interacted (*i.e.,* were separated by a distance within 200 nm) for each model. Columns are for reverse primers while rows are for forward primers. The names of the columns and rows are described in the legend for **Supplementary File 2**.

Nature Structural & Molecular Biology: doi:10.1038/nsmb.1936

Supplementary Data 5. Contact map for ENm008 in K562 cells in a tabulated text file. 5C frequency contact maps for the ENm008 region were calculated using the 314 models in cluster number 2. The numbers in the matrix correspond to the number of times a particular pair of fragments interacted (*i.e.,* were separated by a distance within 200 nm) for each model. Columns are for reverse primers while rows are for forward primers. The names of the columns and rows are described in the legend for **Supplementary File 2**.

Supplementary Data 6. Contact map for ENm008 in GM12878 cells as BED formatted file for direct upload into the UCSC Genome Browser. Such file includes all needed tracks to reproduce the long-range annotation of the ENm008 region shown in **Supplementary Fig. 6**. The names of the tracks are described in the legend for **Supplementary Data 2**.

Supplementary Data 7. Contact map for ENm008 in K562 cells as BED formatted file for direct upload into the UCSC Genome Browser. Such file includes all needed tracks to reproduce the long-range annotation of the ENm008 region shown in **Supplementary Fig. 6**. The names of the tracks are described in the legend for **Supplementary Data 2**.

Nature Structural & Molecular Biology: doi:10.1038/nsmb.1936