nature
structural &
molecular biology

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] & Marc A Marti-Renom[1]

We developed a general approach that combines chromosome conformation capture carbon copy (5C) with the Integrated Modeling Platform (IMP) to generate high-resolution three-dimensional models of chromatin at the megabase scale. We applied this approach to the ENm008 domain on human chromosome 16, containing the α-globin locus, which is expressed in K562 cells and silenced in lymphoblastoid cells (GM12878). The models accurately reproduce the known looping interactions between the α-globin genes and their distal regulatory elements. Further, we find using our approach that the domain folds into a single globular conformation in GM12878 cells, whereas two globules are formed in K562 cells. The central cores of these globules are enriched for transcribed genes, whereas nontranscribed chromatin is more peripheral. We propose that globule formation represents a higher-order folding state related to clustering of transcribed genes around shared transcription machineries, as previously observed by microscopy.

Currently, efforts are directed at producing high-resolution genome annotations in which the positions of functional elements or specific chromatin states are mapped onto the linear genome sequence[1]. However, these linear representations do not indicate functional or structural relationships between distant elements. For instance, recent insights suggest that widely spaced functional elements cooperate to regulate gene expression by engaging in long-range chromatin looping interactions. The three-dimensional (3D) organization of chromosomes is thought to facilitate compartmentalization[2,3], chromatin organization[4] and spatial sequestration of genes and their regulatory elements[5–7], all of which may modulate the output and functional state of the genome. A general approach for determining the spatial organization of chromatin can aid in the identification of long-range relationships between genes and distant regulatory elements as well as in the identification of higher-order folding principles of chromatin in general.

Chromosome conformation capture (3C)-based assays use formaldehyde cross-linking followed by restriction digestion and intramolecular ligation to study chromatin looping interactions[7–12]. 3C-based assays have been used to show that specific elements such as promoters, enhancers and insulators are involved in the formation of chromatin loops[5,7,13–16]. The frequencies at which loci interact reflect chromatin folding[7,17], and thus comprehensive chromatin interaction data sets can help researchers build spatial models of chromatin.

Previously, chromatin conformation has been modeled using polymer models[8,18] and molecular-dynamics simulations[19], which have proven valuable for understanding general features of chromatin fibers, including flexibility and compaction[20,21]. However, such methods only partially leverage the current wealth of experimental data on chromatin folding. Recently, experimentally driven approaches, in combination with computational modeling, have resulted in low-resolution models for the topological conformation of the immunoglobulin heavy chain[22], the *HoxA*[23] loci and the yeast genome[24]. However, those methods were limited by the resolution and completeness of the input experimental data[22], by insufficient model representation, scoring and optimization[23], or by limited analysis of the 3D models[24].

To overcome such limitations, we developed a new approach that couples high-throughput 5C experiments[9] with the IMP[25]. We applied this approach to determine the higher-order spatial organization of a 500-kilobase (kb) gene-dense domain located near the left telomere of human chromosome 16 (**Fig. 1a**). Embedded in this cluster of ubiquitously expressed housekeeping genes is the tissue-specific α-globin locus that is expressed only in erythroid cells. This 500-kb domain corresponds to the ENm008 region extensively studied by the ENCODE pilot project (**Fig. 1b**)[1].

The α-globin locus has been used widely as a model to study the mechanism of long-range and tissue-specific gene regulation[15,26–30]. The α-globin genes are upregulated by a set of functional elements characterized by the presence of DNase I–hypersensitive sites (HSs) located 33 to 48 kb upstream of the ζ gene. One of these elements, HS40, is considered to be of particular importance[31,32]. This element can act as an enhancer in reporter constructs and its deletion greatly affects activation of the α-globin genes[33]. HS40 is bound by several erythroid transcription factors including GATA factors and NF-E2 (ref. 34). Notably, previous 3C studies have demonstrated direct long-range

[1]Structural Genomics Unit, Bioinformatics and Genomics Department, Centro de Investigación Príncipe Felipe, Valencia, Spain. [2]Program in Gene Function and Expression, Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, Massachusetts, USA. [3]Department of Cell Biology, University of Massachusetts Medical School, Worcester, Massachusetts, USA. [4]These authors contributed equally to this work. Correspondence should be addressed to J.D. (job.dekker@umassmed.edu) or M.A.M.-R. (mmarti@cipf.es).

Received 22 November 2009; accepted 20 September 2010; published online 5 December 2010; doi:10.1038/nsmb.1936

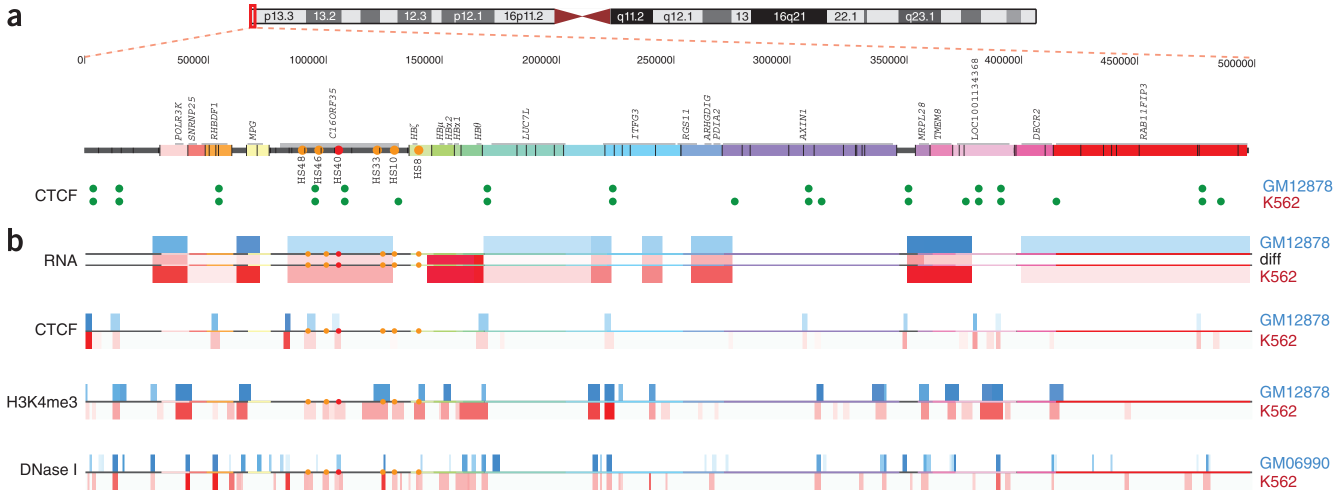


Figure 1 ENCODE region ENm008 on human chromosome 16. (**a**) Map of ENm008, including the ζ-, μ-, α2-, α1- and θ-globin genes. Genes are indicated by gray lines above the linear representation. Vertical black lines indicate HindIII restriction sites. Colored restriction fragments contain annotated genes. Red, orange and green circles mark the HS40 sites, other α-globin–related HSs and CTCF sites, respectively. (**b**) ENCODE annotations for the ENm008 region. RNA expression data, CTCF data, histone modification data (H3K4me3) and DNase I sensitivity data[56,57] are generated by the ENCODE project (http://genome.ucsc.edu/ENCODE/). Red and blue bands indicate the ENCODE track intensity for K562 and GM12878 cell lines, respectively.

looping interactions between some of these distant functional elements (HS48, HS46 and HS40) and the α-globin genes upon gene activation in mouse and human erythroid cells[15,30]. Major unanswered questions revolve around the higher-order folding of multi-gene domains such as ENm008, and how long-range interactions involved in regulation of each of the resident genes are accommodated.

We obtained comprehensive interaction maps of the α-globin locus by performing 5C analysis of the ENm008 region in GM12878 and K562 cells. These two cell lines, which differ in the expression of the α-globin genes, are studied by the ENCODE consortium, and therefore extensive chromatin structural and functional information for ENm008 is publicly available. We developed a general approach to generate 3D chromatin models on the basis of chromatin interaction data. Our models of the ENm008 domain in GM12878 cells show that it forms a single compact structure, which we refer to as a 'chromatin globule'. We find that active genes and promoters tend to be located at the center of the globule, whereas inactive genes are more peripherally positioned. Notably, in cells that express high levels of α-globin (K562), the chromatin is broken into two globules separated by an extended chromatin segment. We propose that sets of neighboring active genes cluster to form chromatin globules, perhaps analogous to transcription factories, and that a given globule can accommodate only a limited number of active genes.

RESULTS

Our approach for determining the 3D conformation of genomic domains consists of four steps (**Supplementary Fig. 1**): (i) data collection by 5C experiments, (ii) data translation into points and spatial restraints between them, (iii) model building by optimization of the imposed restraints, and (iv) ensemble analysis of the optimal 3D solutions. The following sections describe the results of each of these key steps in our approach to 3D structure determination of the ENm008 region. A summary and further details of the methods are provided in the Online Methods and **Supplementary Methods**, respectively.

5C analysis of ENm008

5C, described in detail before[9,35], uses highly multiplexed ligation-mediated amplification to detect sets of 3C ligation products. We designed 5C primers at HindIII sites using computational algorithms through our online My5C software package (http://my5C.umassmed.edu/)[36]. In total, 30 forward primers and 25 reverse primers were designed throughout the 500-kb ENm008 region, which were capable of detecting 750 unique pairwise chromatin interactions (**Supplementary Table 1**). The number of 5C ligation products, which corresponds to pairs of interacting fragments, was determined by paired-end Solexa sequencing. Consistent with previous analyses[9,37], the 5C interaction maps show prominent signals between sites located near each other. Further, GM12878 cells show more abundant long-range interactions, suggesting a more compact conformation than K562 cells (**Fig. 2**).

We determined the average relationship between genomic distance (in kilobases) and interaction probability (average read count) using the entire 5C data set (blue lines in 5C interaction profiles of **Fig. 2**). This is important because this relationship can be used as an estimate for the expected random collision frequency for pairs of loci in the absence of specific looping interactions[37] (**Supplementary Fig. 2**). In K562 cells, we detected all previously known long-range looping interactions between the active α-globin genes and the upstream distant regulatory elements (HS48, HS46 and HS40), which interacted up to six-fold more frequently than the estimated expected frequency (**Fig. 2b**). Such frequent interactions were not present in GM12878 cells, which have a repressed α-globin domain (**Fig. 2a**). Therefore, K562 can serve as a model cell line to study the conformation of the active α-globin locus, despite the fact that (i) these cells are transformed and can be variable in karyotype and gene expression profile, and (ii) primary erythroid cells could have a different chromatin conformation in this region.

Notably, novel long-range interactions were identified. For example, in both cell types, HS46 interacted very frequently with a locus located just downstream of the α-globin genes (3′ end of *LUC7L*, which encodes a RNA-binding protein similar to the yeast Luc7p). This downstream locus, in turn, interacted more frequently than expected with a region located within the more distant *AXIN1* gene. The nature of the elements involved in these interactions is currently unknown, although it is noteworthy that all the interacting fragments contain sites bound by the CTCF protein (**Fig. 1**), which is often involved in long-range interactions[13].

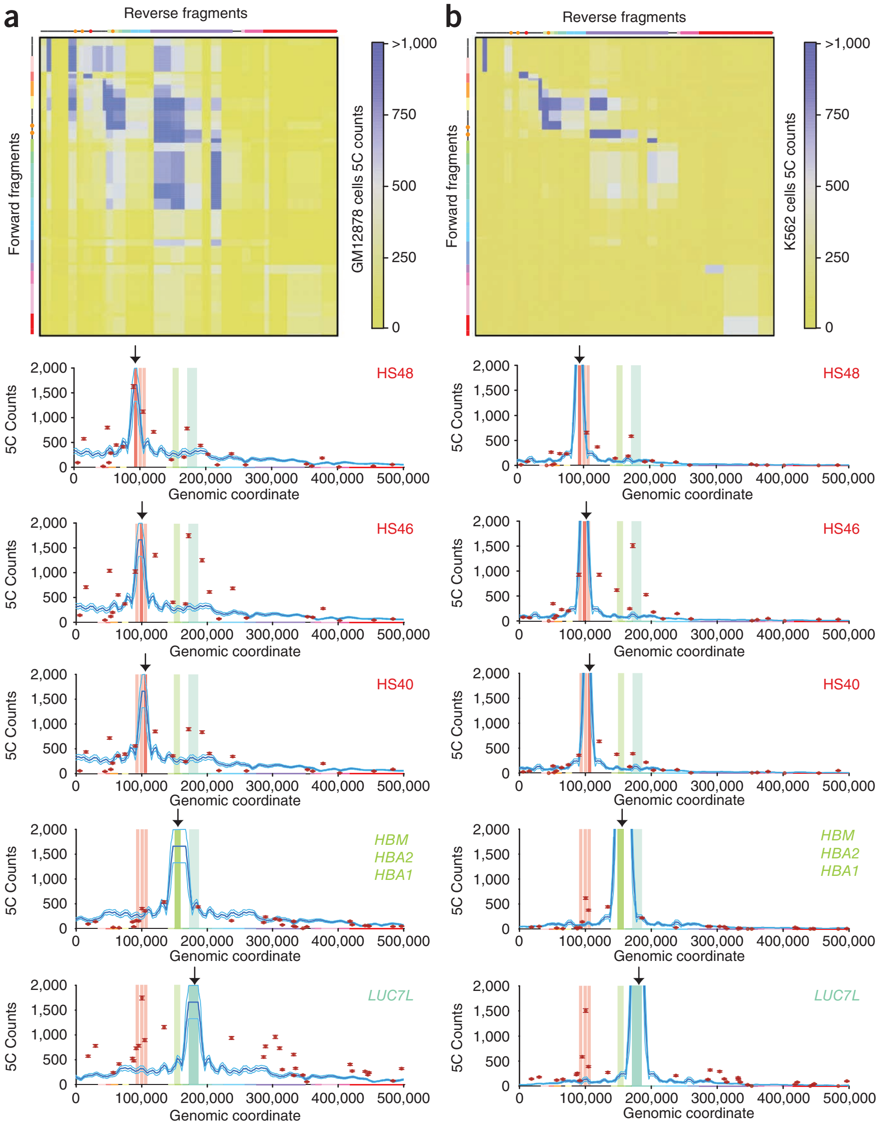


Figure 2 5C analysis of the 500-kb ENCODE region ENm008. (**a**) 5C experimental data for GM12878 cell lines. Upper plot shows 5C count matrix colored yellow to blue to indicate low to high counts. For easy inspection, the axis labels are replaced by the linear representation of the forward and reverse fragments of the ENm008 region. Lower plots show 5C interaction profiles for fragments containing HS48, HS46, HS40, *HBM*, *HBA2*, *HBA1* and 3′ end of *LUC7L*, respectively. The plots show the 5C counts and their associated s.e.m. of interactions between the anchor fragment (indicated by vertical arrows) and the rest of the queried fragments in the ENm008 region; colored bars indicate the positions of HS elements (red), globin genes (green) and *LUC7L* gene (blue). Blue solid lines show the average and s.e.m. of the expected relationship between interaction frequency (5C counts) and genomic distance (kb), determined by LOESS smoothing of the complete data set (**Supplementary Fig. 2**). Red circles show the observed 5C counts for each of the queried fragments. (**b**) 5C experimental data for K562 cell lines. Data are represented as in **a**.

From 5C data to points and restraints

Chromatin interaction frequencies can be used as a proxy for spatial distance between interacting fragments[12]. Thus, our first step was to translate the 5C experimental data into a set of distances dependent on the observed interactions. The IMP represents a genomic domain as a set of points (one per restriction fragment) and the spatial restraints (or springs) between them, with equilibrium distances proportional to the observed frequency of interaction. The type and force of the restraints that place each of the 70 points representing the ENm008 region were defined by the IMP calibration, which was carried out in two steps. First, 5C counts were normalized by $\log_{10}$ transformation and *Z*-score computation from the average and s.d. of all $\log_{10}$ values in the interaction matrix. A *Z*-score indicates how many s.d. a measure is above or below the mean of the measure. Second, two linear relationships were defined linking 5C *Z*-scores to spatial distances for restraining pairs of fragments: (i) two neighbor fragments (i to $i + 1…2$) were restrained on the basis of the linear relationship between the 5C *Z*-scores and the sum of the excluded volume occupied by the nucleotides between the centers of the two fragments (**Supplementary Table 1**), and (ii) two non-neighbor fragments (i to $i + 3…n$) were restrained on the basis of the relationship bound by an empirically determined closest possible distance between two non-interacting fragments and the excluded volume of a canonical 30-nm fiber (**Supplementary Fig. 3**). These two linear relationships between 5C *Z*-score and spatial distances relied on the following assumptions: (i) the different 5C *Z*-score distributions between neighbor and non-neighbor fragments reflected their different response in 5C experiments[37]; (ii) consecutive fragments were spatially restrained proportionally to the occupancy of their chromatin fragments, with a relationship of 0.01 nm per base pair (bp), assuming a canonical 30-nm fiber[38]; and (iii) two non-neighbor fragments could not get closer in space than 30 nm, which corresponds to the diameter of the chromatin fiber. Even though the precise diameter of the chromatin fiber *in vivo* is unknown and probably fluctuates, it has been shown that the observed looping frequencies from 5C experiments in human cells are consistent with a 30-nm fiber[39]. Moreover, the assumption that chromatin adopts a 30-nm fiber affects only the final scale of the resulting 3D models, which is controlled by the excluded volume assigned to the fragments. The results from our FISH experiments (below) indicated that the use of 0.01 nm per bp resulted in models of the appropriate scale. Finally, the values of two *Z*-score cutoffs were also optimized and defined the type of restraint imposed between two non-neighbor fragments. The optimal parameters found were as follows: for GM12878 cells, 500 nm for the lowest *Z*-score, a *Z*-score of –0.2 for the lower-bound cutoff and a *Z*-score of 0.1 for the upper-bound cutoff; for K562 cells, 400 nm for the lowest *Z*-score, a *Z*-score of –0.1 for the lower-bound cutoff, and a *Z*-score of 0.9 for the upper-bound cutoff (**Supplementary Methods**).

All 70 fragments representing the studied region were restrained with a total of 1,520 and 1,049 restraints for GM12878 and K562 cells, respectively (**Supplementary Fig. 3**). The forces applied to the defined restraints were also set proportional to the absolute value of the 5C *Z*-score observed between a pair of fragments. That is, the more extreme the *Z*-score, the stronger the force constant applied to the restraint. By making the harmonic forces proportional to

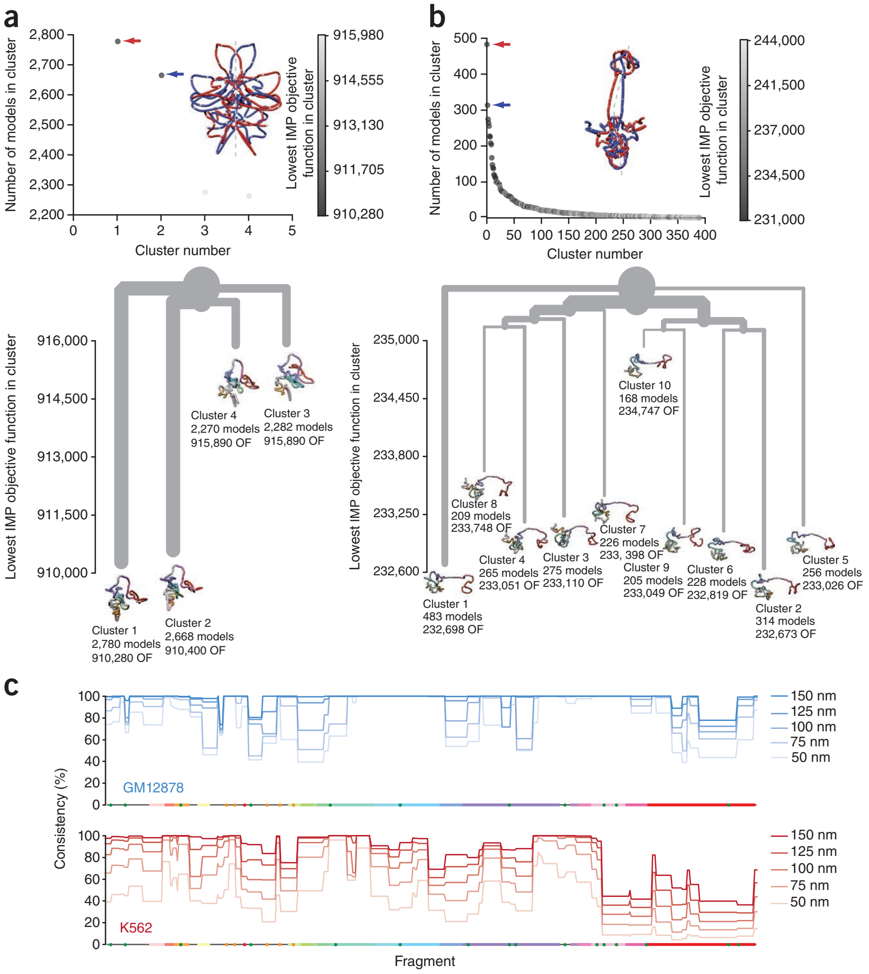


Figure 3 Ensemble of solutions. (**a**) Cluster analysis for the selected 10,000 GM12878 models. Upper plot shows the number of models per cluster plotted against the cluster number. Points are colored on a grayscale proportional to the lowest IMP objective function in the cluster. IMP mirroring is illustrated by the superimposition of cluster 1 (red arrow) and cluster 2 (blue arrow) centroids (that is, the solution closest to the center of each cluster). Lower plot shows the structural relationship between the top cluster centroids. The tree was generated on the basis of the structural similarity between each of the centroids; branch thickness is proportional to the number of solutions at each branch point. A structure of each centroid, colored as in its linear representation (**Fig. 1a**), is placed on a vertical scale proportional to the lowest IMP objective function within the cluster it represents. (**b**) Cluster analysis for the selected 10,000 K562 models. Data are represented as in panel **a**. (**c**) Model consistency for the ensemble of solutions in cluster 1 of GM12878 models (blue) and cluster 2 of K562 models (red).

the variability of the *Z*-score, we ensured that restraints between pairs of points with extreme *Z*-score values were stronger than those between pairs of points with average 5C interaction frequencies. An exception to this rule was applied to neighbor fragments. In such cases, the forces were set to a value of 5.0, which was large enough to maintain connectivity between neighbor fragments.

Generation of spatial models of ENm008

Once the restraints had been defined, IMP generated a 3D model of the ENm008 region by searching for a spatial arrangement of all points that minimized the violation of the imposed restraints (**Supplementary Fig. 3c,d**). Thus, IMP expressed the problem of determining the chromatin structure as an optimization problem, assuming that the conformation of the locus is largely determined by chromatin interactions within the locus. The absence of strong interactions outside the locus comparable in frequency to the ones we observed within the locus was recently confirmed by Hi-C, a method that couples proximity-based ligation with massively parallel sequencing to probe the three-dimensional architecture of whole genomes[12].

Starting from randomly positioned points within a cube of side length 1 μm, IMP iteratively moves all points so as to force them to a conformation that minimally violates the imposed restraints. Given that the 5C analyses are population averaged, the 3D models generated by IMP can represent only the macroscopic state of the system and thus result in an ensemble of solutions reflecting the variability of chromatin conformation[40]. 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*.

A total of 50,000 models were generated for each cell type, which ensured a fair coverage of the search space. We then selected the 10,000 models with the least number of violated restraints to be clustered according to their structural similarity (**Supplementary Methods**). GM12878 models clustered in a total of four different conformations (**Fig. 3a**). The first- and second-most-populated clusters contained the conformations with the lowest IMP objective function, indicating that a minimum in the search space was found for most of the independent runs. This shows that (i) 5C data are sufficient for uniquely identifying a set of dominant conformations, and (ii) the top two clusters represent topological mirror solutions, providing further confidence in the results.

Models obtained for K562 cells formed a more variable set of solutions, with a total of 393 different structure clusters, including ten large clusters with more than 150 solutions each and 194 clusters with less than ten solutions each (**Fig. 3b**). The large number of clusters with few members may represent a diverse set of local minima conformations that partly satisfy the K562 5C interaction data. Such diverse solutions could reflect a more variable chromatin conformation of the domain in K562 cells, perhaps related to variable karyotypes and gene expression in individual cells in this cancer-derived cell line. It is important to note that even though we selected representative clusters to describe key properties of the α-globin locus structure (below), only the ensemble of all solutions from the top clusters reflected the range of multiple distinct conformations that may be present in the cell population.

We studied whether the different conformations we observed for individual models within a cluster of solutions could be considered locally consistent (**Fig. 3c**). Such analysis allowed us to identify local regions in the structures that were conserved for most of the pairwise structure alignments between the models in the selected

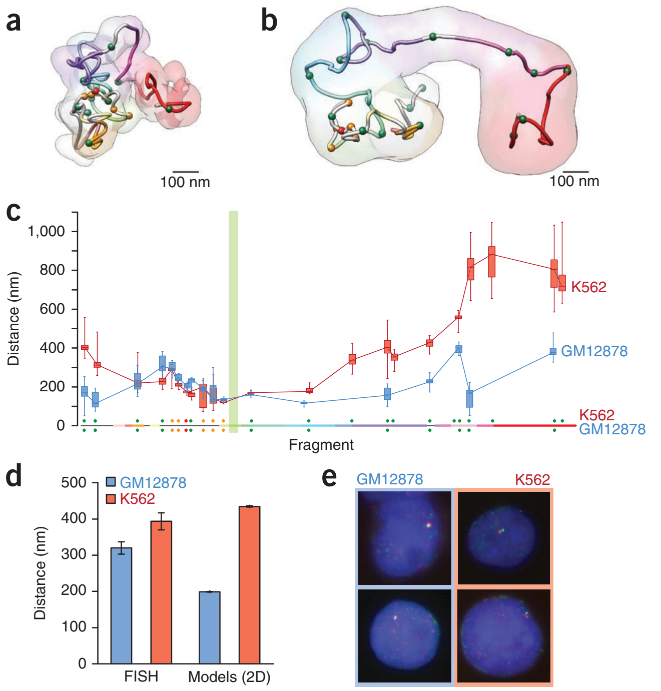


Figure 4 3D models of the ENm008 ENCODE region containing the α-globin locus. (**a**) 3D structure of the GM12878 models represented by the centroid of cluster 1. The 3D model is colored as in its linear representation (**Fig. 1a**). Regulatory elements are represented as spheres colored red (HS40), orange (other HSs) and green (CTCF). (**b**) 3D structure of the K562 models represented by the centroid of cluster 2. Data are represented as in panel **a**. (**c**) Distances between the α-globin genes (restriction fragments 31 and 32) and other restriction fragments in ENm008. The plot shows the distribution and s.d. of the mean of distances for GM12878 models in cluster 1 (blue) and K562 models in cluster 2 (red). (**d**) Average distances (and their s.e.m.) between a pair of loci located on either end of the ENm008 domain, as determined by FISH with two fosmid probes (see Online Methods) and from a 2D representation of the IMP-generated models in both cell lines. (**e**) Example images obtained with FISH of GM12878 and K562 cell lines. The images show smaller distances between the probes in GM12878 than in K562 cell lines.

cluster. GM12878 models were locally consistent; only one fragment (reverse 21) of these models did not have a consistent local conformation (that is, not superimposable within 150 nm for more than 75% of the models). In K562 cells, as many as 82% of the fragments were consistent across the models. This analysis shows that even in the more variable K562 models most of the region contains conserved local features, and that the diversity is the result of variable positioning of only a small minority of fragments (18%).

Models reproduce known long-range interactions

We determined whether the 3D models reflected the known long-range interactions involving the α-globin genes (**Fig. 4**). We used the selected cluster of models to calculate the average distance between the restriction fragment containing the α-globin genes and other restriction fragments in ENm008 in both GM12878 and K562 cells. Restriction fragments containing the enhancer (HS40) and α-globin genes were closely juxtaposed in K562 cells (159.1 ± 13.3 nm). In contrast, HS40 was the only fragment that was located farther from the α-globin genes in the inactive GM12878 cells (228.2 ± 17.3 nm) than in K562 cells; all other fragments in GM12878 cells were located closer to the α-globin genes (**Fig. 4c**). These observations are consistent with previous 3C experiments showing that strong interaction between HS40 and the α-globin genes is evident only when the genes are expressed.

Validation by fluorescence *in situ* hybridization

We used an independent method, fluorescence *in situ* hybridization (FISH), to validate a particular aspect of our 3D models for the ENm008 region. For small genomic domains such as the one studied here, determining the spatial positions of individual restriction fragments within the domain by FISH is not straightforward given the resolution of light microscopy, which is limited to ~200 nm. However, the models of the ENm008 domain predict that the locus is in a more extended conformation in K562 cells than in GM12878 cells, which would lead to a greater average 2D interphase distance between the ends of the 500-kb locus. Prior work has demonstrated that this distance is large enough to be measured by interphase mapping with FISH[41].

We found that in GM12878 these loci were on average 318.8 ± 17.0 nm apart, whereas in K562 cells they were 391.9 ± 23.4 nm apart. These differences, which are statistically significant ($P < 0.011$), show that in K562 cells the locus is in a more extended conformation, consistent with the models generated by IMP, in which the 2D distances (that is, without considering the orientation of the model) were 198.9 ± 0.7 nm and 434.6 ± 1.4 nm for GM12878 and K562 models, respectively (**Fig. 4d**,**e**).

Formation of chromatin globules

A noteworthy feature observed in both cell lines was the formation of compact chromatin clusters, which we termed chromatin globules. In GM12878 cells, the ENm008 region forms a single chromatin globule, whereas in K562 cells, the locus forms two chromatin globules (**Fig. 4a**,**b** and **Supplementary Videos 1** and **2**). This large-scale difference in conformation between the two cell lines is also evidenced by the contact-map differences between GM12878 and K562 models (**Fig. 5a**). The heat map shows that most distances in GM12878 are smaller than in K562 cells, consistent with the formation of a single compact chromatin globule. However, also consistent with the 5C data, the α-globin genes and the distant regulatory elements are closer in space in K562 cells than in GM12878 cells (red areas in **Fig. 5a**).

To explore whether these globules have some degree of internal organization, we determined the locations of genes and putative regulatory elements within the chromatin globules. We measured the radial positions of active genes, gene promoters, HSs, sites bound by CTCF and sites marked with trimethylated histone H3 Lys4 (H3K4me3) by calculating the average distance between each corresponding restriction fragment and the geometrical center of the globules. Notably, we found that in the IMP models from both cell types, active genes and gene promoters are enriched near the center of the globule, whereas inactive genes and restriction fragments that do not contain genes are more peripheral (**Fig. 5b**). In contrast, HSs, CTCF-bound sites and sites marked by H3K4me3 are not preferentially located in the center, but are found throughout the globules.

In GM12878 cells, we visually identified nine loops ranging from about 20 to 70 kb long, with an average length of ~50 kb, an average distance between anchors of 102.8 ± 5.1 nm and an average path length of 547.9 ± 96.9 nm (**Fig. 5c**). In K562 cells, the locus forms two chromatin globules (five loops and two loops, respectively) ranging from about 30 to 70 kb, with an average length of ~60 kb, an average distance between anchors of 231.2 ± 129.2 nm (190.6 ± 43.5 nm not considering loop 6 connecting the two globular domains) and an average path length of 600.1 ± 90.2 nm. Because our experiments covered only the ENm008 region, we were not able to determine whether the second chromatin globule observed in K562 cells contained additional

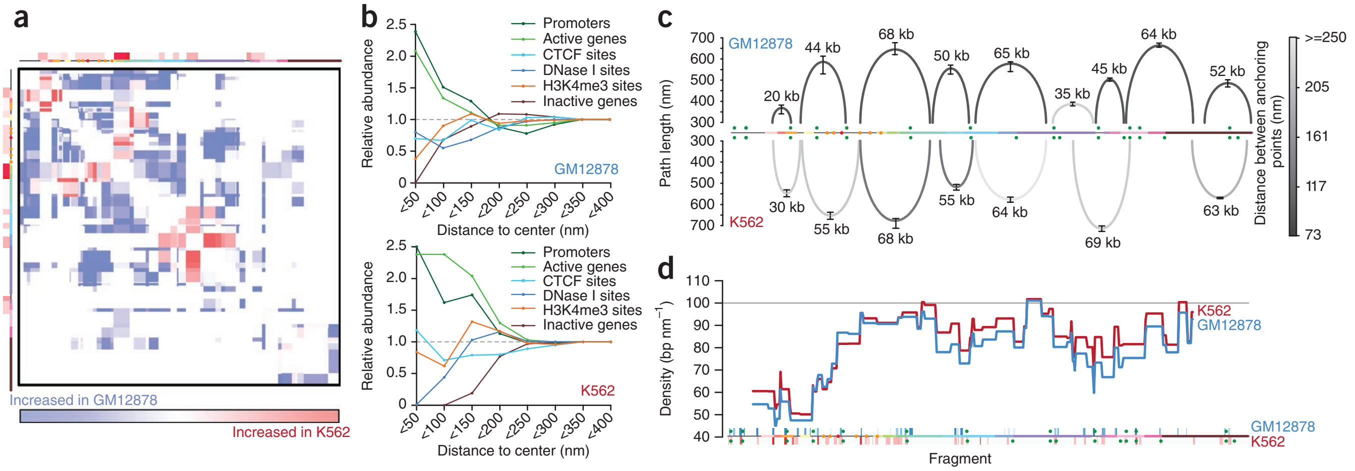


Figure 5 Analysis of chromatin globules. (**a**) Frequency contact map differences between models in cluster 1 of GM12878 cells and cluster 2 of K562 cells. Differential expression levels are shown next to the 1D representation of the ENm008 on each axis of the plot. (**b**) Relative abundance of different ENm008 fragment types, plotted against the center of the chromatin globule, for GM12878 (upper plot) and K562 (lower plot). Plots show cumulative relative abundance of annotations versus radial position in the globule. Active genes and promoters are enriched in the center. (**c**) Observed loops in the centroids of selected clusters for GM12878 (top) and K562 (bottom) models. The loops are placed over the 1D representation of the ENm008 region (**Fig. 1a**). Loop height is proportional to the path length of the loop. Loops are colored according to the distance between the anchor points (dark, near; light, far). Loop sizes in kilobases (kb) are indicated at the tip of each loop. (**d**) Chromatin density for the ensemble of solutions in cluster 1 of GM12878 models (blue) and cluster 2 of K562 models (red). HSs are shown next to the 1D representation of the ENm008 on the *x* axis of the plot.

genes beyond the LOC100134368, *DECR2* and *RAB11FIP3* genes. Overall, the models suggest that chromatin is organized around chromatin globules, with rosettes of 50- to 60-kb chromatin loops and centers enriched with active genes and their promoters.

Estimates of chromatin compaction based on spatial models

Chromatin across the ENm008 region was not uniformly dense, as determined by the contour length of the chromatin fiber (**Fig. 5d**). As expected, the average chromatin path was much denser than that of naked DNA, which is about 3 bp nm^{-1}. We found that the telomere-proximal end of ENm008, which contains the highest density of active genes as well as most of the regulatory elements (as estimated from the density of HSs; **Fig. 1b**), has chromatin fiber compaction that corresponds to ~50 bp nm^{-1}. In contrast, the telomere-distal region has a denser chromatin region (~100 bp nm^{-1}). Notably, GM12878 cell models result in a less dense chromatin fiber, on average, despite folding into a single chromatin globule. However, the region containing the HS40 enhancer of α-globin genes is more compact, on average, in GM12878 cells than in K562 cells, consistent with the predicted relationship between transcription and formation of more open chromatin.

Local chromatin features and three-dimensional folding

The analysis of chromatin compaction shows how our models can reveal new insights into spatial relationships between distant 1D annotations and their 3D conformation. To further demonstrate this, we have generated tracks for the UCSC Genome Browser[42] showing the interaction frequency maps resulting from our 5C experiments and 3D models (**Supplementary Fig. 4**). These tracks allow direct visualization of spatial relationships between widely spaced genomic elements in the context of all publicly available 1D genome annotations. For instance, we find that the α-globin genes are spatially close to a region containing the genes *POLR3K* and *MPG* near the left end of the region. Both interacting regions are transcriptionally active and marked by histone modifications associated with open chromatin (for example, di- and trimethylated histone H3 Lys4 (H3K4me2, H3K4me3) and acetylated H3 and H4 (H3ac and H4ac)). This is consistent with our observation that active genes tend to form the cores of the chromatin globules, a pattern that has also been suggested in previous work showing association between active genes[10,15,43].

DISCUSSION

Here we have combined high-throughput *in vivo* chromatin interaction mapping with the IMP to characterize the higher-order chromatin conformation of the ENm008 region containing the α-globin domain in cells that do or do not express the globin locus. The 5C data and the 3D models derived from them accurately reflect the known long-range interactions between the α-globin genes and their distant regulatory elements, validating our approach. Furthermore, we have identified a higher-order chromatin folding motif in which groups of adjacent genes cluster to form chromatin globules. Analysis of the internal architecture of the globules revealed that active genes are enriched in the cores of these structures. These observations suggest that chromatin globules may represent subnuclear structures dedicated to gene expression, perhaps related to the clustering of shared transcription machineries.

Formation of chromatin globules

A chromatin globule forms a rosette-like structure with loops of ~50–60 kb, an average path length of 500–600 nm and a distance between anchors of 100–200 nm. Such spatial organization is consistent with the 'multi-loop subcompartment' model, which proposes that chromatin is folded into rosettes of small loops, with the rosettes connected by linkers of variable size[22,44]. Notably, FISH experiments have also revealed that chromatin can form strings of globular domains of around the same size (in kilobases) as the globules identified here[45]. The type and function of proteins involved in maintaining these chromatin globules are unknown.

It has been proposed that active genes interact at discrete sites (also called transcription factories) where several RNA polymerases are concentrated[46]. It is still unresolved whether such transcription machineries are a consequence or cause of transcription of gene clusters in the nucleus[47]. However, it has been observed by EM that these sites of transcription can range from 45 to 100 nm in diameter[46,48–51],

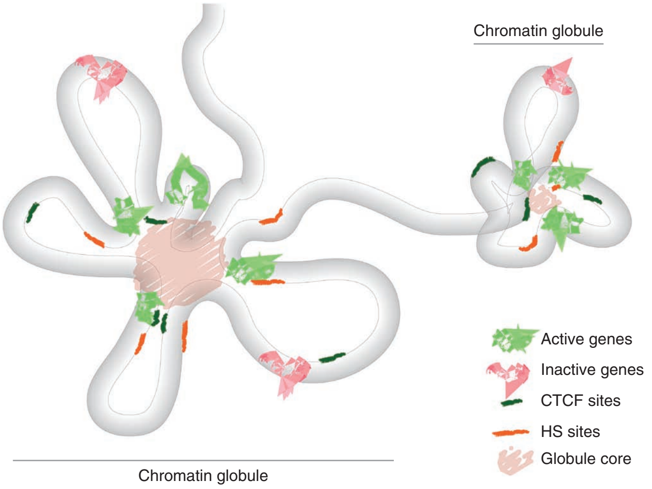


Figure 6 Diagram of the proposed chromatin-globule model for higher-order chromatin folding of actively transcribed genomic regions.

and they include a limited number of active RNA polymerases (about eight) as estimated from the number of nascent RNA molecules[46]. Our models agree with these estimates. The first chromatin globule in our K562 models, which includes the α-globin genes as well as other nearby housekeeping genes, wraps around a cavity with an average diameter of ~100–110 nm, which would fit a hypothetical transcription factory.

Of particular interest is our observation that the ENm008 region forms a single large chromatin globule in GM12878 cells, but two smaller globules in K562 cells. The major difference between these two cell lines is the expression of the α-globin gene cluster, which is actively transcribed in K562 cells. In GM12878 cells, only six or seven genes are actively transcribed in the ENm008 region, which would all fit within one transcription factory. In contrast, in K562 cells, the additional activity of the α-globin genes seems to exceed the capacity of a single transcription factory, with about ten genes being actively transcribed. We entertain the idea that the number of active genes that can cluster to form a chromatin globule may be limited to only around eight genes, which would be consistent with the elongated-beaded structures of active chromatin regions observed by light microscopy[45,52]. This is a highly speculative idea, however, and it is also possible that the extended conformation in K562 cells is related to the transformed state of this cell line. Further experiments are needed to shed light on the determinants of globule formation.

From our models, we cannot say whether these chromatin globules self-assemble around genes sharing common transcription machineries, actively assemble on demand or already exist as a complex fixed to an as-yet-unknown underlying nuclear substructure. It has been proposed that transcriptionally active regions may attain increased chromatin mobility[6]. It is noteworthy that the K562 models have higher variability and lower consistency than models from GM12878 cells, which could relate either to the region being broken into two globules or to the fact that the region is more transcriptionally active overall.

Chromatin density

Even for transcriptionally active regions, chromatin is about 400- to 1,000-fold more compact than the 30-nm fiber[53]. Therefore, decondensation of chromatin may be transient[54]. We observed for both cell lines that transcriptionally inactive regions were, on average, about twice as dense as regions containing either transcribed genes or their regulatory elements. Notably, the region including HS40, HS46 and HS48 was denser, on average, in GM12878 than in K562 cells, whereas the other regions we studied were denser, on average, in K562 than in GM12878 cells. Our results indicate that chromatin undergoes a certain amount of decondensation when genes are expressed[55]. This shows that 5C experiments, reflected by our models, are able to capture such subtle differences.

A chromatin-globule model

Our 3D structures suggest a model for higher-order chromatin folding based on the formation of chromatin globules (**Fig. 6**). Chromatin globules would be spatially separated and would form by clustering of a limited number of actively transcribed genes. Within the context of the chromatin globules, our analysis identified specific long-range interactions between genes and their regulatory elements, as well as novel interactions between sites bound by CTCF. The potential roles of such regulatory elements in globule formation are currently unknown.

The discovery of chromatin globules shows that our models can reveal novel higher-order features of chromosome architecture. Our 3D models for the ENm008 region are in agreement with (i) our own FISH experiments validating the models' overall size and shape, (ii) previously described biological phenomena such as the clustering of active genes, and (iii) local chromatin structural features from the ENCODE consortium such as DNase I sensitivity. By revealing the relative spatial arrangements of genes and their regulatory elements, our approach could further leverage large-scale efforts to annotate genes and regulatory elements along the linear genome.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/nsmb/.

Note: Supplementary information is available on the Nature Structural & Molecular Biology website.

ACKNOWLEDGMENTS

We thank the IMP community (http://www.integrativemodeling.org), especially D. Russell, B. Webb and A. Sali, as well as the Chimera developers (http://www.cgl.ucsf.edu/chimera/), especially T. Goddard and T. Ferrin. We also thank M. Umbarger, M. Wright, G. Church, M.S. Madhusudhan, M. Walhout and Dekker laboratory members for fruitful discussions. We acknowledge support from the Spanish Ministerio de Ciencia e Innovación (BIO2007/66670 and BFU2010/19310 to M.A.M.-R.), the US National Institutes of Health (NIH; HG003143 to J.D. and GM053234 to J.B.L.) and the Keck Foundation (J.D.). Finally, we are grateful to the ENCODE project (funded by the NIH and the US National Human Genome Research Institute) for providing annotations of the ENm008 region. In particular, we thank the ENCODE groups led by T. Gingeras (expression data, Cold Spring Harbor Laboratory), G. Crawford (DNase I data, Duke University) and B. Bernstein (CTCF data, H3K4me3 data, Broad Institute of Harvard and MIT). ENCODE data are publicly available through the ENCODE Data Coordination Center at the University of California, Santa Cruz (http://genome.ucsc.edu/ENCODE/).

AUTHOR CONTRIBUTIONS

B.R.L. performed the bioinformatics design and analysis of the 5C experiments. A.S. performed the 5C experiments. D.B., E.C. and M.A.M.-R. carried out the IMP computational modeling. M.B., A.S. and J.B.L. performed the FISH experiments. D.B., B.R.L., A.S., J.D. and M.A.M.-R. wrote the manuscript. J.D. and M.A.M.-R. conceived the work.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/nsmb/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).

2. Lamond, A.I. & Spector, D.L. Nuclear speckles: a model for nuclear organelles. *Nat. Rev. Mol. Cell Biol.* **4**, 605–612 (2003).
3. Misteli, T. Beyond the sequence: cellular organization of genome function. *Cell* **128**, 787–800 (2007).
4. Fraser, P. Transcriptional control thrown for a loop. *Curr. Opin. Genet. Dev.* **16**, 490–495 (2006).
5. de Laat, W. & Grosveld, F. Spatial organization of gene expression: the active chromatin hub. *Chromosome Res.* **11**, 447–459 (2003).
6. Fraser, P. & Bickmore, W. Nuclear organization of the genome and the potential for gene regulation. *Nature* **447**, 413–417 (2007).
7. Dekker, J. Gene regulation in the third dimension. *Science* **319**, 1793–1794 (2008).
8. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
9. Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* **16**, 1299–1309 (2006).
10. Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.* **38**, 1348–1354 (2006).
11. Zhao, Z. *et al.* Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat. Genet.* **38**, 1341–1347 (2006).
12. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
13. Phillips, J.E. & Corces, V.G. CTCF: master weaver of the genome. *Cell* **137**, 1194–1211 (2009).
14. Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F. & de Laat, W. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Mol. Cell* **10**, 1453–1465 (2002).
15. Zhou, G.L. *et al.* Active chromatin hub of the mouse alpha-globin locus forms in a transcription factory of clustered housekeeping genes. *Mol. Cell. Biol.* **26**, 5096–5105 (2006).
16. Murrell, A., Heeson, S. & Reik, W. Interaction between differentially methylated regions partitions the imprinted genes Igf2 and H19 into parent-specific chromatin loops. *Nat. Genet.* **36**, 889–893 (2004).
17. Ohlsson, R. & Gondor, A. The 4C technique: the 'Rosetta stone' for genome biology in 3D? *Curr. Opin. Cell Biol.* **19**, 321–325 (2007).
18. Mateos-Langerak, J. *et al.* Spatially confined folding of chromatin in the interphase nucleus. *Proc. Natl. Acad. Sci. USA* **106**, 3812–3817 (2009).
19. Wedemann, G. & Langowski, J. Computer simulation of the 30-nanometer chromatin fiber. *Biophys. J.* **82**, 2847–2859 (2002).
20. Dekker, J. Mapping in vivo chromatin interactions in yeast suggests an extended chromatin fiber with regional variation in compaction. *J. Biol. Chem.* **283**, 34532–34540 (2008).
21. Wachsmuth, M., Caudron-Herger, M. & Rippe, K. Genome organization: balancing stability and plasticity. *Biochim. Biophys. Acta* **1783**, 2061–2079 (2008).
22. Jhunjhunwala, S. *et al.* The 3D structure of the immunoglobulin heavy-chain locus: implications for long-range genomic interactions. *Cell* **133**, 265–279 (2008).
23. Fraser, J. *et al.* Chromatin conformation signatures of cellular differentiation. *Genome Biol.* **10**, R37 (2009).
24. Duan, Z. *et al.* A three-dimensional model of the yeast genome. *Nature* **465**, 363–367 (2010).
25. Alber, F. *et al.* Determining the architectures of macromolecular assemblies. *Nature* **450**, 683–694 (2007).
26. Hughes, J.R. *et al.* Annotation of cis-regulatory elements by identification, subclassification, and functional assessment of multispecies conserved sequences. *Proc. Natl. Acad. Sci. USA* **102**, 9830–9835 (2005).
27. Higgs, D.R., Vernimmen, D., Hughes, J. & Gibbons, R. Using genomics to study how chromatin influences gene expression. *Annu. Rev. Genomics Hum. Genet.* **8**, 299–325 (2007).
28. Higgs, D.R. & Wood, W.G. Long-range regulation of alpha globin gene expression during erythropoiesis. *Curr. Opin. Hematol.* **15**, 176–183 (2008).
29. Lower, K.M. *et al.* Adventitious changes in long-range gene expression caused by polymorphic structural variation and promoter competition. *Proc. Natl. Acad. Sci. USA* **106**, 21771–21776 (2009).
30. Vernimmen, D., De Gobbi, M., Sloane-Stanley, J.A., Wood, W.G. & Higgs, D.R. Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression. *EMBO J.* **26**, 2041–2051 (2007).
31. Higgs, D.R. *et al.* A major positive regulatory region located far upstream of the human alpha-globin gene locus. *Genes Dev.* **4**, 1588–1601 (1990).
32. Chen, H., Lowrey, C.H. & Stamatoyannopoulos, G. Analysis of enhancer function of the HS-40 core sequence of the human alpha-globin cluster. *Nucleic Acids Res.* **25**, 2917–2922 (1997).
33. Bernet, A. *et al.* Targeted inactivation of the major positive regulatory element (HS-40) of the human alpha-globin gene locus. *Blood* **86**, 1202–1211 (1995).
34. De Gobbi, M. *et al.* Tissue-specific histone modification and transcription factor binding in alpha globin gene expression. *Blood* **110**, 4503–4510 (2007).
35. Dostie, J. & Dekker, J. Mapping networks of physical interactions between genomic elements using 5C technology. *Nat. Protoc.* **2**, 988–1002 (2007).
36. Lajoie, B.R., van Berkum, N.L., Sanyal, A. & Dekker, J. My5C: web tools for chromosome conformation capture studies. *Nat. Methods* **6**, 690–691 (2009).
37. Dekker, J. The three 'C' s of chromosome conformation capture: controls, controls, controls. *Nat. Methods* **3**, 17–21 (2006).
38. Gerchman, S.E. & Ramakrishnan, V. Chromatin higher-order structure studied by neutron scattering and scanning transmission electron microscopy. *Proc. Natl. Acad. Sci. USA* **84**, 7802–7806 (1987).
39. Rosa, A., Becker, N.B. & Everaers, R. Looping probabilities in model interphase chromosomes. *Biophys. J.* **98**, 2410–2419 (2010).
40. Voss, T.C. & Hager, G.L. Visualizing chromatin dynamics in intact cells. *Biochim. Biophys. Acta* **1783**, 2044–2051 (2008).
41. Lawrence, J.B., Singer, R.H. & McNeil, J.A. Interphase and metaphase resolution of different distances within the human dystrophin gene. *Science* **249**, 928–932 (1990).
42. Kuhn, R.M. *et al.* The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res.* **37**, D755–D761 (2009).
43. Osborne, C.S. *et al.* Myc dynamically and preferentially relocates to a transcription factory occupied by Igh. *PLoS Biol.* **5**, e192 (2007).
44. Münkel, C. *et al.* Compartmentalization of interphase chromosomes observed in simulation and experiment. *J. Mol. Biol.* **285**, 1053–1065 (1999).
45. Müller, W.G. *et al.* Generic features of tertiary chromatin structure as detected in natural chromosomes. *Mol. Cell. Biol.* **24**, 9359–9370 (2004).
46. Martin, S. & Pombo, A. Transcription factories: quantitative studies of nanostructures in the mammalian nucleus. *Chromosome Res.* **11**, 461–470 (2003).
47. Sutherland, H. & Bickmore, W.A. Transcription factories: gene expression in unions? *Nat. Rev. Genet.* **10**, 457–466 (2009).
48. Iborra, F.J., Pombo, A., Jackson, D.A. & Cook, P.R. Active RNA polymerases are localized within discrete transcription 'factories' in human nuclei. *J. Cell Sci.* **109**, 1427–1436 (1996).
49. Pombo, A. *et al.* Regional specialization in human nuclei: visualization of discrete sites of transcription by RNA polymerase III. *EMBO J.* **18**, 2241–2253 (1999).
50. Eskiw, C.H., Rapp, A., Carter, D.R. & Cook, P.R. RNA polymerase II activity is located on the surface of protein-rich transcription factories. *J. Cell Sci.* **121**, 1999–2007 (2008).
51. Carter, D.R., Eskiw, C. & Cook, P.R. Transcription factories. *Biochem. Soc. Trans.* **36**, 585–589 (2008).
52. Goetze, S. *et al.* The 3D structure of human interphase chromosomes is related to the transcriptome map. *Mol. Cell. Biol.* **27**, 4475–4487 (2007).
53. Hu, Y., Kireev, I., Plutz, M., Ashourian, N. & Belmont, A.S. Large-scale chromatin structure of inducible genes: transcription on a condensed, linear template. *J. Cell Biol.* **185**, 87–100 (2009).
54. Boeger, H., Griesenbeck, J. & Kornberg, R.D. Nucleosome retention and the stochastic nature of promoter chromatin remodeling for transcription. *Cell* **133**, 716–726 (2008).
55. Gheldof, N., Tabuchi, T.M. & Dekker, J. The active FMR1 promoter is associated with a large domain of altered chromatin conformation with embedded local histone modifications. *Proc. Natl. Acad. Sci. USA* **103**, 12463–12468 (2006).
56. Xi, H. *et al.* Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome. *PLoS Genet.* **3**, e136 (2007).
57. Crawford, G.E. *et al.* DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nat. Methods* **3**, 503–509 (2006).

ONLINE METHODS

5C analysis of the ENm008 region. 5C primers were designed at HindIII restriction sites using 5C primer design tools we previously developed[9] and made available online at http://my5C.umassmed.edu/ (ref. 36). Reverse primers were designed for fragments overlapping a known transcription start site from GENCODE transcripts[58], or overlapping a start site experimentally determined by CAGE-tag data of the ENCODE pilot project[1]. Forward primers were designed for all other HindIII restriction fragments. Primers were excluded if highly repetitive sequences prevented the design of a sufficiently unique 5C primer. Primer design thresholds were as follows: U-BLAST, 3; S-BLAST, 130: 15-MER, 1320; MIN_FSIZE, 40; MAX_FSIZE, 50000; OPT_TM, 65; OPT_PSIZE, 40. The DNA sequence of the universal tails of forward primers was 5′-CCTCTCTATGGGCAGTCGGTGAT-3′; the DNA sequence for the universal tails of reverse primers was 5′-AGAGAATGAGGAACCCGGGGCAG-3′. A six-base barcode was included between the specific part of the primers and the universal tail. In total, 26 reverse primers and 30 forward primers were designed (**Supplementary Table 1**). Note that reverse primers 31 and 32 recognized identical sequences corresponding to the two HindIII fragments that contained the identical *HBA1* and *HBA2* genes. Therefore, we combined 5C data obtained with these two primers. As a result, the effective number of reverse primers was 25. Forward and reverse 5C primers were distributed homogeneously along the ~500-kb ENm008 region. This resulted in a 5C interaction map that is evenly populated, which made the data particularly well suited for 3D modeling approaches. Primer sequences are available as supplementary information (**Supplementary Data 1**).

We performed 3C with HindIII as described[35], using exponentially growing GM12878 and K562 cells. We then performed 5C in 20 reactions, each containing an amount of 3C library that represents 200,000 genome equivalents. We amplified 5C ligation products using a pair of universal primers that recognize the common tails of the 5C forward and reverse primers. To facilitate paired-end DNA sequence analysis on the Illumina GA2 platform, we ligated paired-end adapters to the 5C library using the Illumina PE protocol. The ENm008 5C library was then sequenced on the Illumina GA2 platform together with 14 other (unrelated) 5C libraries. For K562 cells, we obtained 31,331,096 paired-end reads of 36 bases each, of which 25,452,766 could be mapped back to specific libraries using Novoalign (http://www.novocraft.com/). A total of 131,947 paired-end sequences were specifically mapped to interactions within ENm008 (**Supplementary Data 2**). For GM12878 cells, we obtained 29,222,267 paired-end reads of 36 bases each, of which 25,081,876 could be mapped back to specific libraries. A total of 182,989 paired-end sequences were specifically mapped to interactions within ENm008 (**Supplementary Data 3**). The 5C data for K562 cells represent the spatial conformation of 19,589 genome equivalents (the largest number of reads obtained for a single restriction fragment) up to 131,947 genome equivalents (the total number of reads obtained for all fragments in K562 cells). The 5C data for GM12878 cells represent the spatial conformation of 20,314 genome equivalents (the largest number of reads obtained for a single restriction fragment) up to 182,989 genome equivalents (the total number of reads obtained for all fragments in GM12878 cells).

Fluorescence *in situ* hybridization analysis of the ENm008 region. Our standard protocols for cell fixation and *in situ* hybridization have been described[59], as have the measurement and analysis of FISH signals to determine interphase distances[41]. Briefly, for fixation of the cell suspensions studied here, cells were attached to coverslips coated in either polylysine or BD Cell-TaK cell and tissue adhesive (BD Biosciences), extracted in CSK buffer with 5% (v/v) Triton for 3 min and fixed in 4% (w/v) paraformaldehyde for 10 min, then stored in 70% (v/v) ethanol. Probes used for FISH analysis were globin-1 and globin-2 genomic sequences, which covered fragments 6 to 14 and 58 to 62, respectively. Probes were nick-translated with either biotin-11-dUTP or digoxigenin-16-dUTP (Roche Diagnostics). After being denatured in 70% (w/v) formamide and 2× saline sodium citrate (SSC; EMD Chemicals) at 75 °C for 2 min, cellular DNA was hybridized with 2.5 μg ml^{-1} of probe overnight at 37 °C in 2× SSC and 50% (w/v) formamide. Cells were washed three times for 20 min each, and then probes were detected using either anti-digoxigenin bound to fluorescein (Roche Diagnostics) or Alexa Fluor 594 Streptavidin (Invitrogen) in 1% (w/v) BSA and 4× SSC for 1 h at 37 °C; this was followed by three 10-min rinses. Cells were counterstained with the DNA dye DAPI.

We captured digital images with an Axiovert 200 or an Axiophot Zeiss microscope equipped with a ×100 PlanApo objective (NA 1.4) and Chroma 83000 multi-band-pass dichroic and emission filter sets, set up in a wheel to prevent optical shift. Images were captured with the Zeiss AxioVision software and either an Orca-ER camera or a 200-series Photometrics cooled charge-coupled device (CCD) camera. Measurements were taken using the length measurement tool within the AxioVision software, version 4.7. The distances between at least 100 red-green signal pairs for each cell line, measured from the center of red signal to that of the green signal, were obtained as described[41]. Presumptive G2 cells, with closely spaced doublet spots of the same color signal, were not measured.

We used two fosmid probes that recognize either end of ENm008 to determine the 2D spatial distance between the corresponding positions in both cell lines. The fosmid probes map to positions 34,512–77,058 and 386,139–425,502 on chromosome 16 (**Supplementary Table 1**).

Integrated modeling platform. Structure determination by IMP consists of four main steps: data generation by experiment, translation of the data into spatial restraints, building of an ensemble of structures by satisfaction of these restraints, and analysis of the ensemble to produce the final structure. The **Supplementary Methods** and **Supplementary Figures 5** and **6** describe in detail 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 deconvolution, and (v) model visualization with Chimera[60]. Tabulated text files of data used to construct contact maps are provided in **Supplementary Data 4**–**7**

58. Harrow, J. *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome Biol* **7** (Suppl. 1), S4 (2006).
59. Tam, R., Shopland, L.S., Johnson, C.V., McNeil, J. & Lawrence, J.B. Applications of RNA FISH for visualizing gene expression and nuclear architecture in *FISH: A Practical Approach* (eds. Beatty, B.G., Mai, S. & Squire, J.) 93–118 (Oxford University Press, New York, 2002).
60. Pettersen, E.F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).

doi:10.1038/nsmb.1936

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

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.

We first plotted all 5C data versus genomic distance (**Supplementary Fig. 2**). Next we performed LOESS smoothing with a window size of 37 interactions ($\alpha = 0.05$) 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:

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

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

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:

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

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:

$$r_i = 0.005 \cdot l_i \tag{2}$$

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:

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

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

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

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 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 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 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 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 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%)

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 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 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:

$$RA_{t,d} = \frac{n_{t,d}}{n_t} \bigg/ \frac{n_d}{N} \tag{6}$$

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

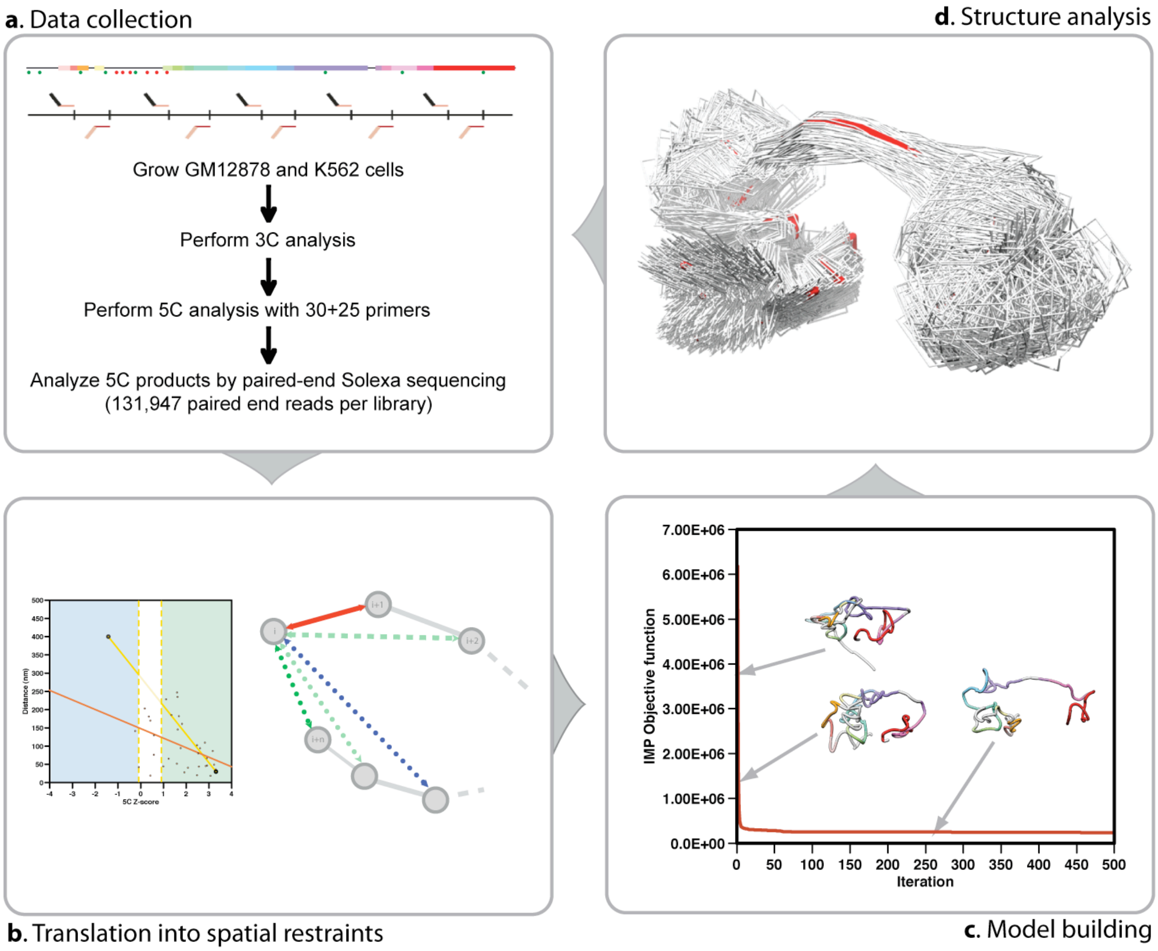
The UCSF Chimera package[3], a highly extensible program for interactive visualization of molecular structures, was used to produce all graphics images and to analyze the resulting ensemble of solutions. First, to visually inspect the most likely path of an ensemble of solutions (or cluster), the centroid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

Supplementary References

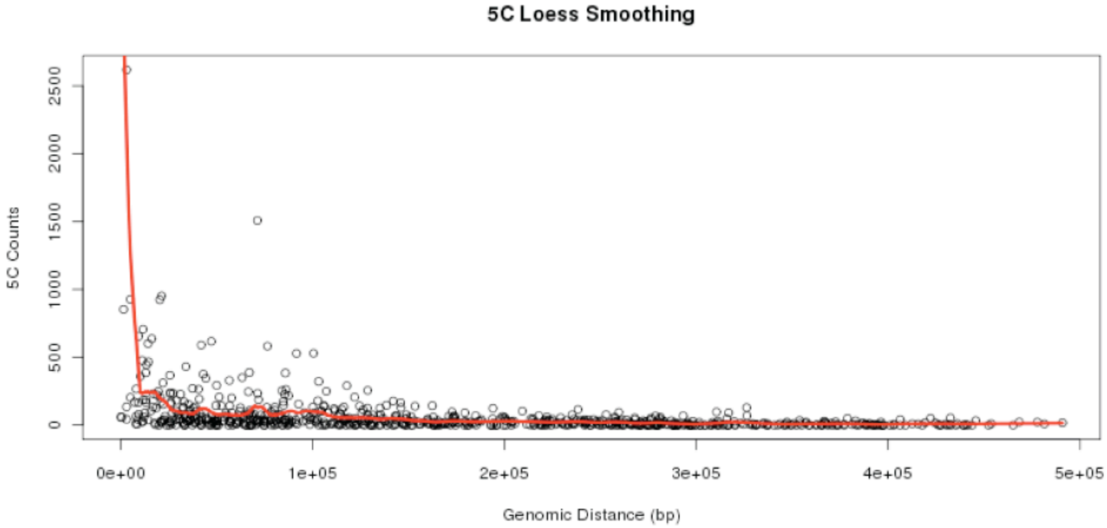
1. Alber, F. et al. Determining the architectures of macromolecular assemblies. *Nature* **450**, 683-94 (2007).
2. Wagner, G. et al. Protein structures in solution by nuclear magnetic resonance and distance geometry. The polypeptide fold of the basic pancreatic trypsin inhibitor determined using two 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).

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.

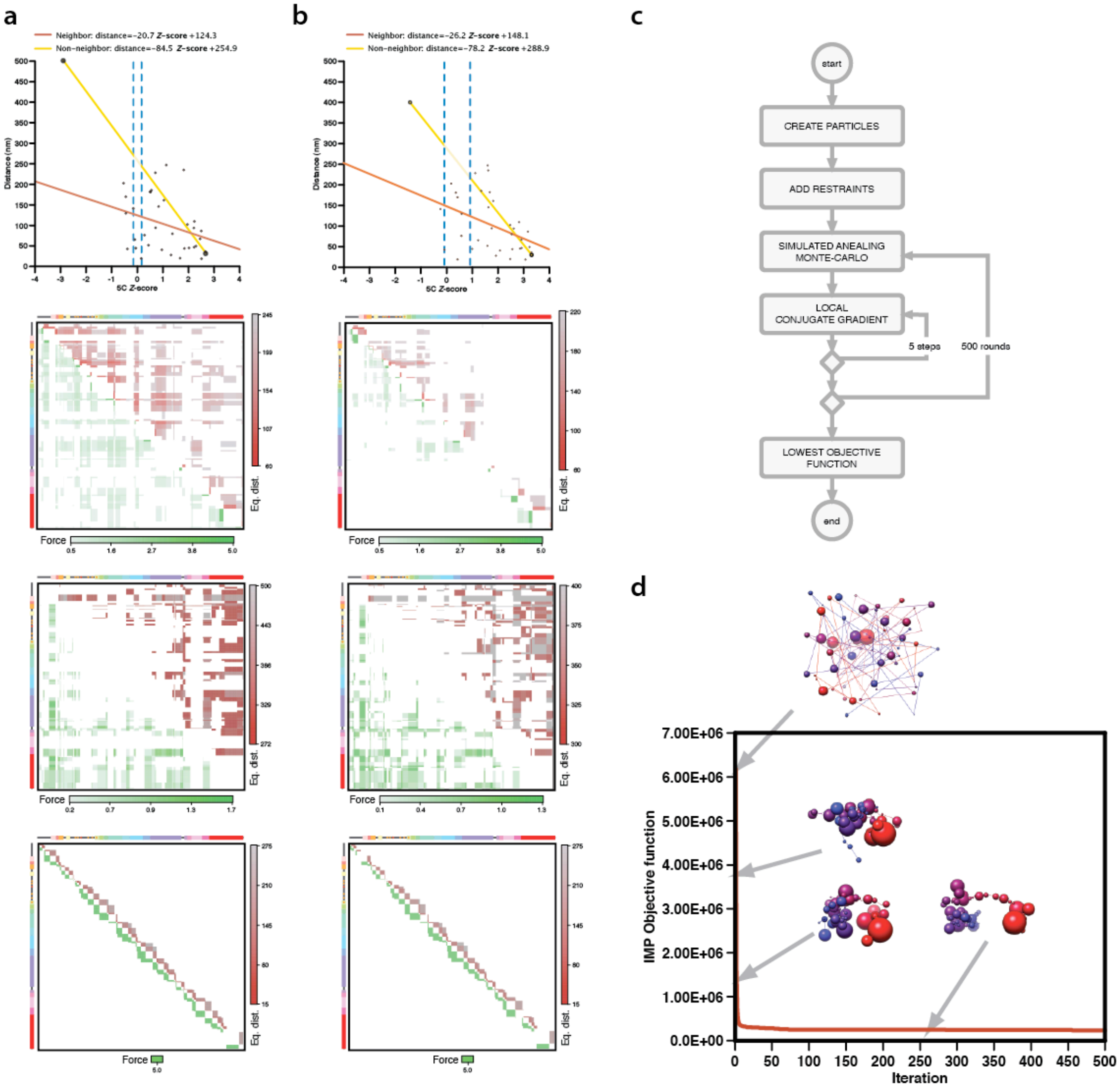
	Fragment	Start	End	Length (nt)	Radii (nm)	FISH probes	RefSeq genes	Assigned fragment types based on the ENCODE data											
								GM12878						K562 (globule 1)					
								PR	AG	NA	HS	CT	HM	PR	AG	NA	HS	CT	HM
1	Reverse 1	1	5693	5692	28.5					■	■	■	■			■		■	■
2	Left out 2-5	5693	11138	5445	27.2					■						■			
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)		■										



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.



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 ($\alpha = 0.05$) (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.



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 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).

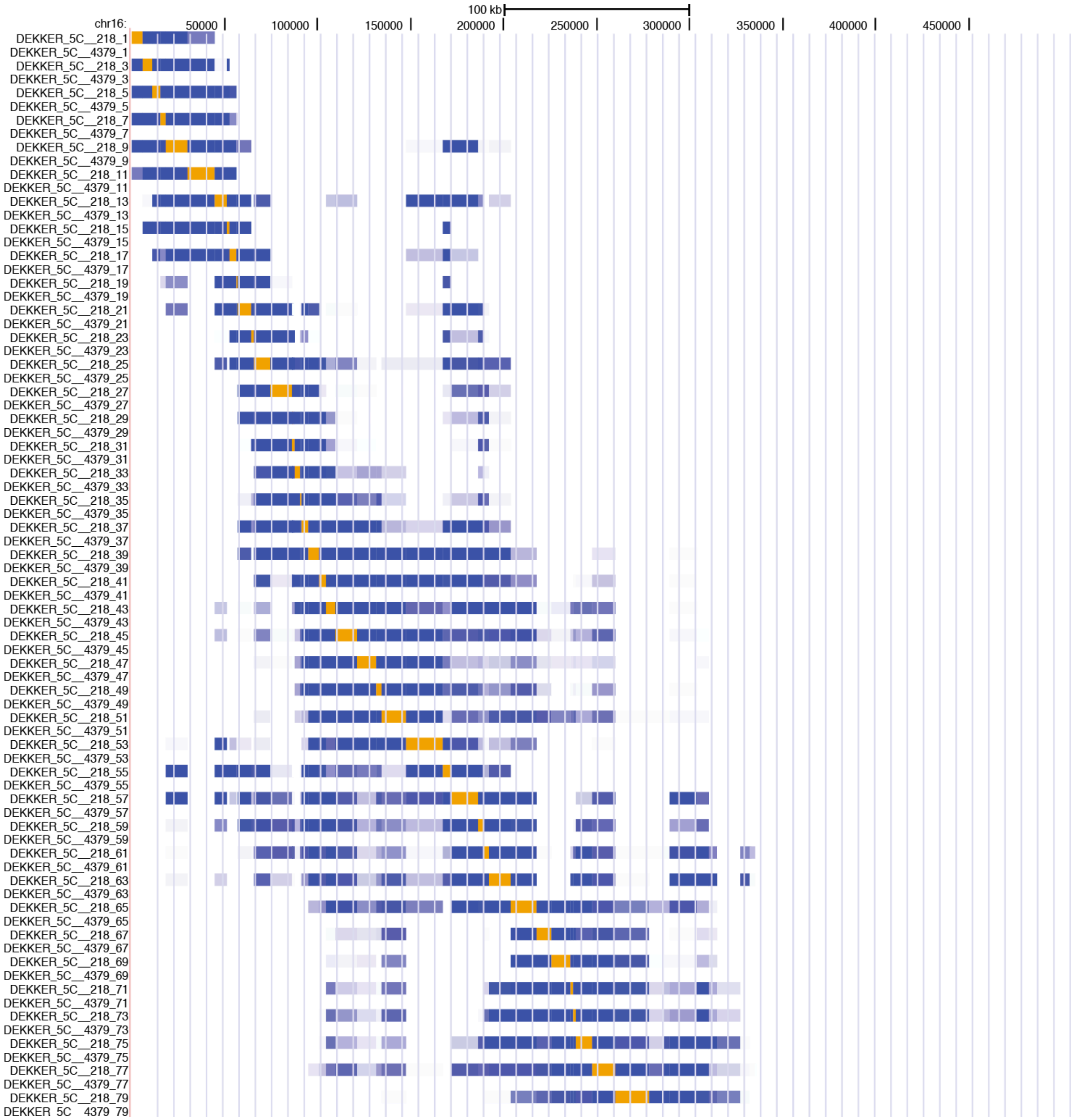
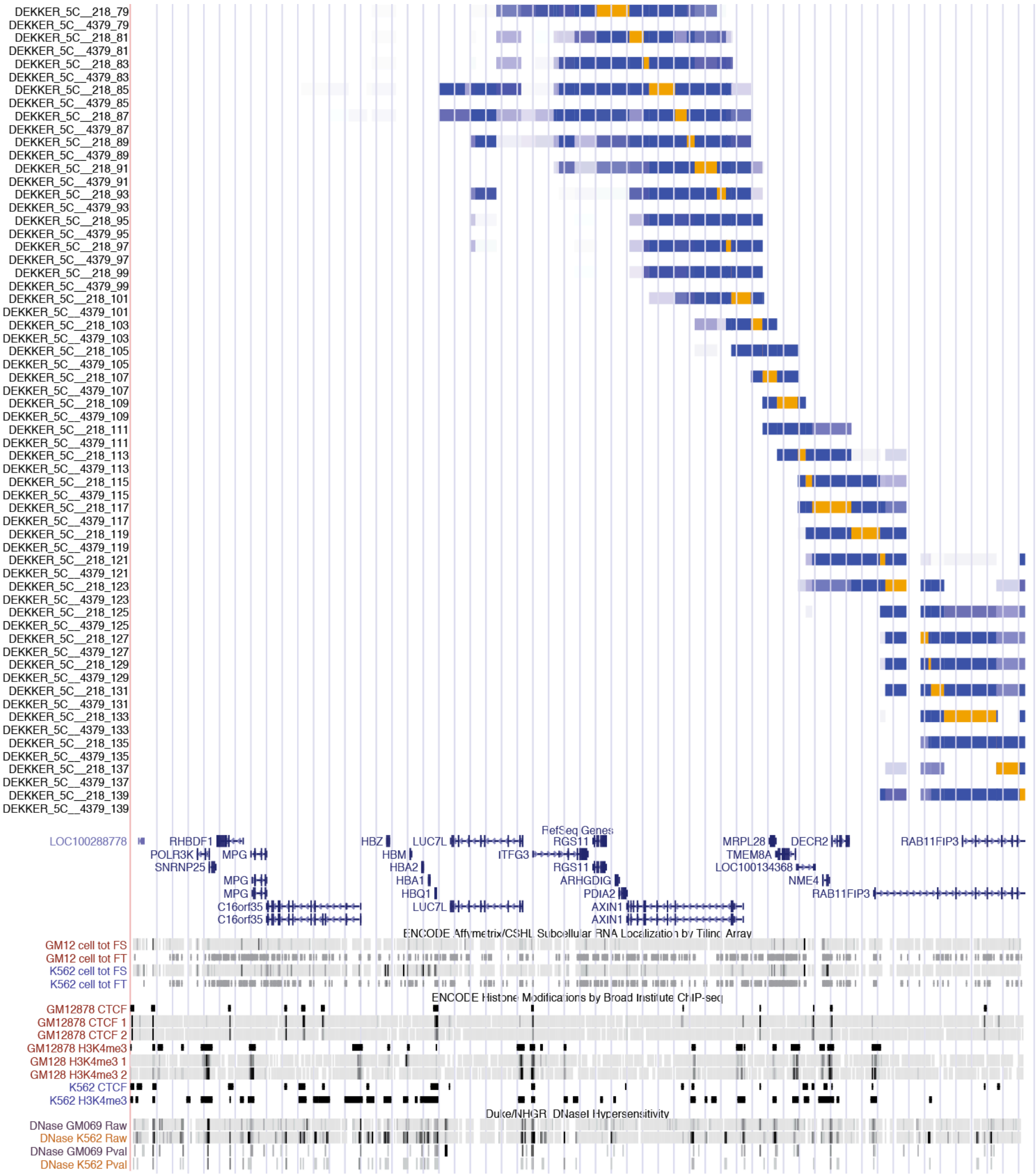
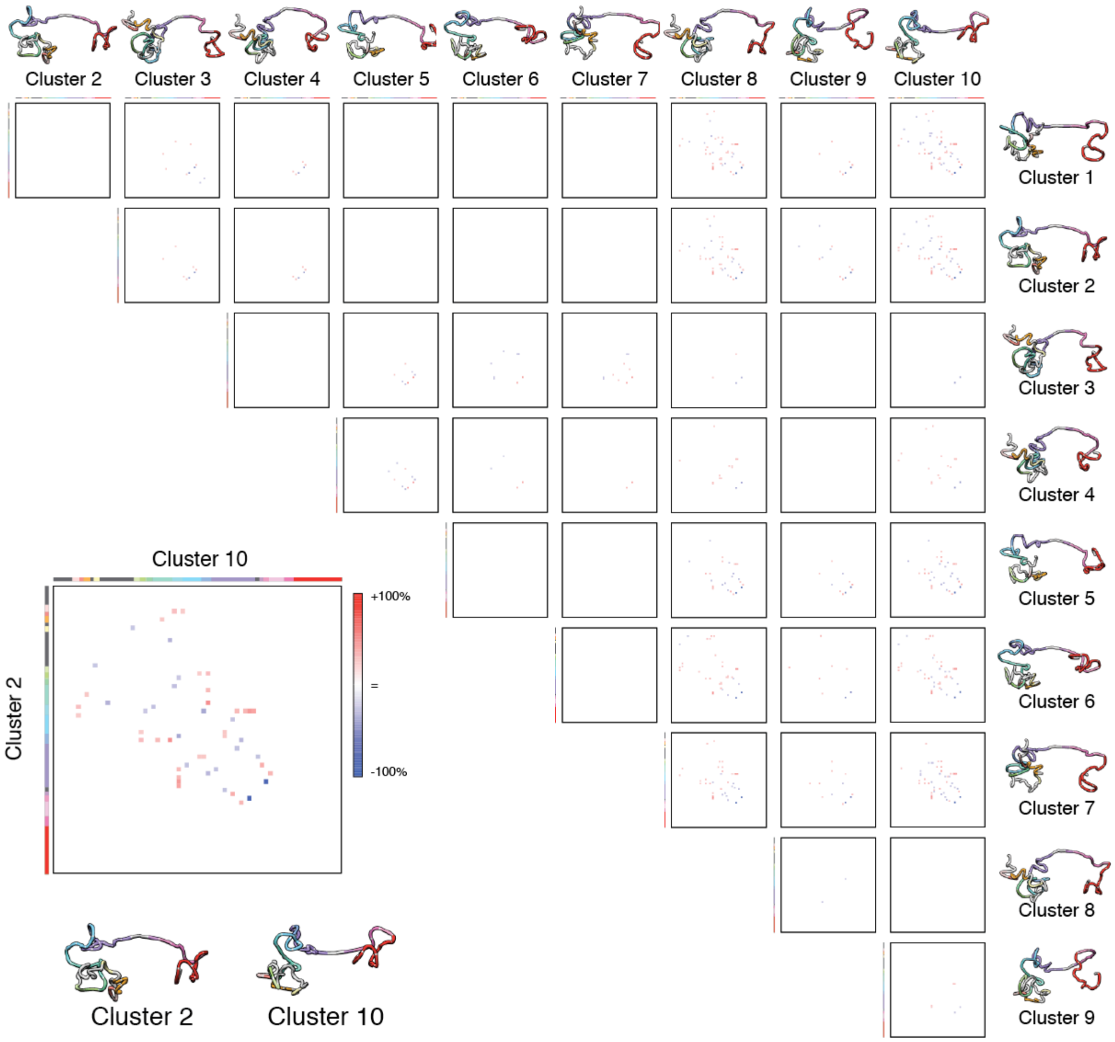


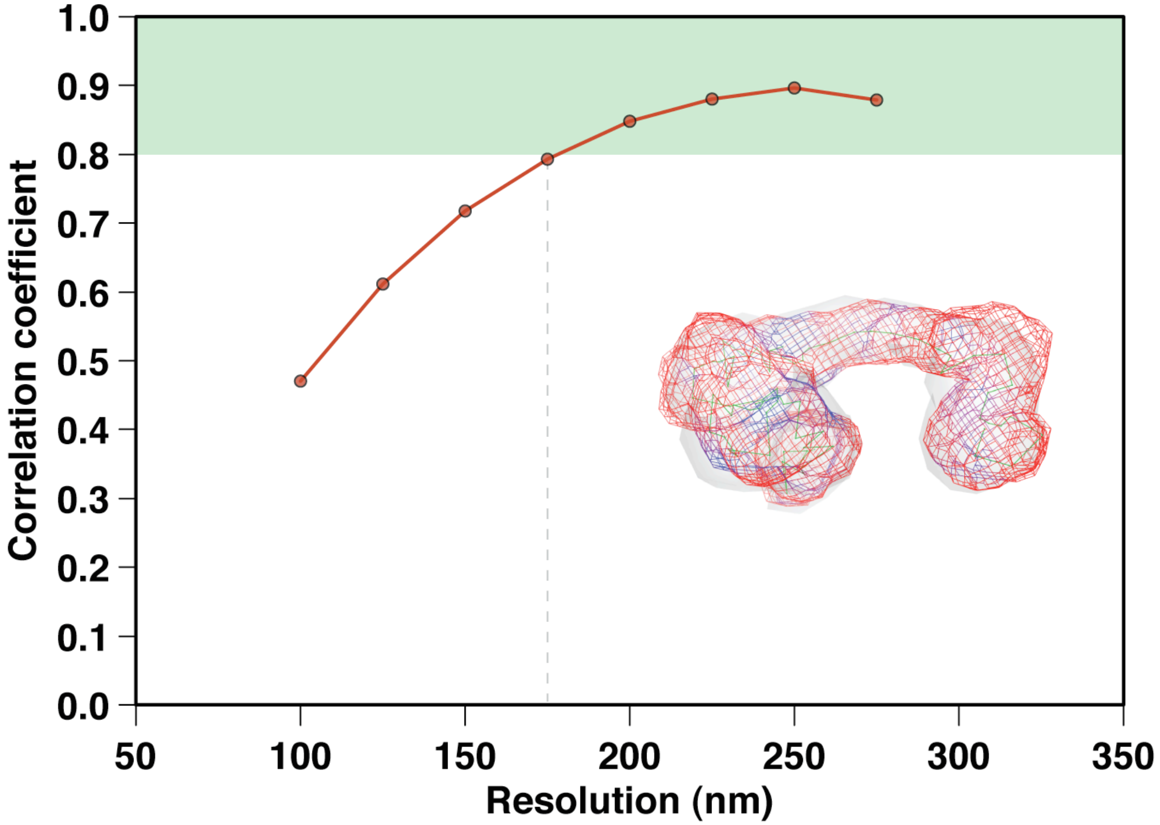
Figure 4 continues in next page…



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**.

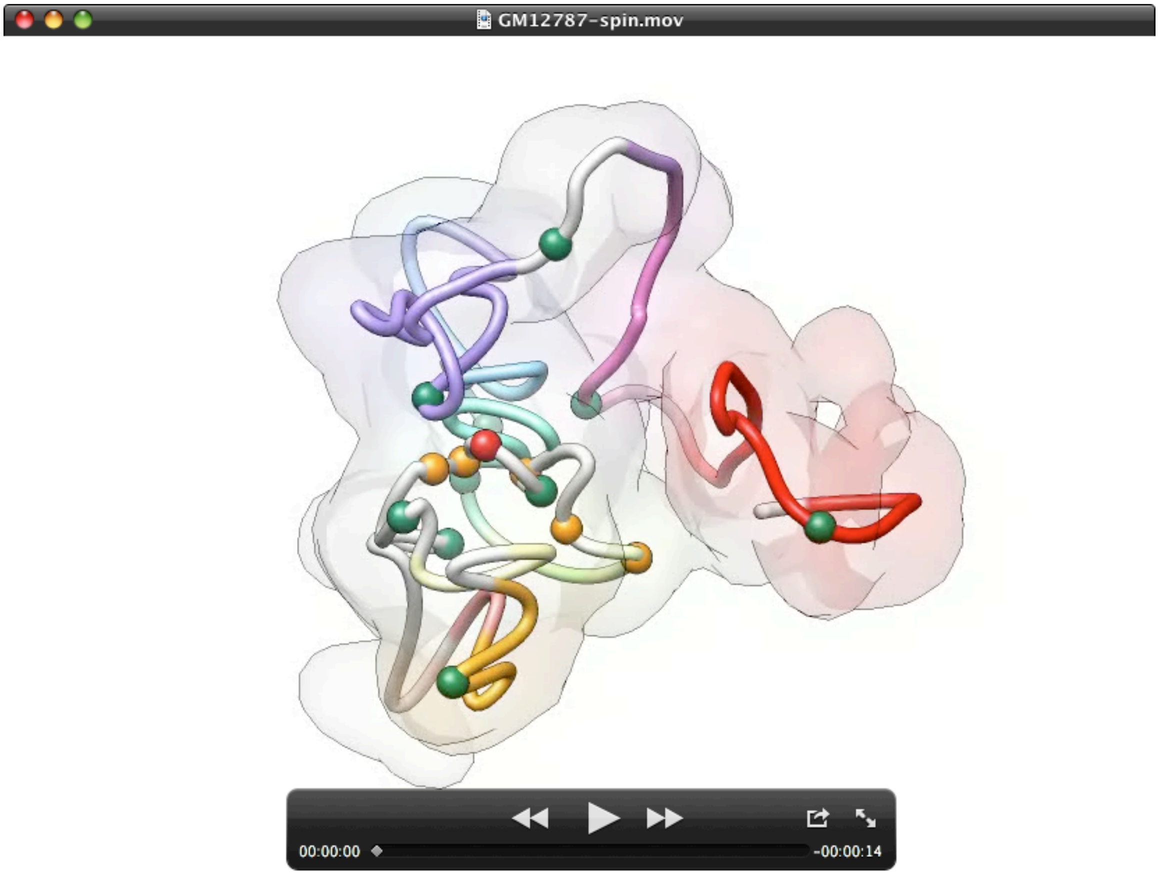


Supplementary Figure 5. 5C de-convolution analysis using solution ensembles for K562 cells. Frequency contact map comparison of the top ten clusters of solutions. Red to blue dots indicates increased or decreased interacting frequencies between the compared ensembles of solutions for each cluster, respectively.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

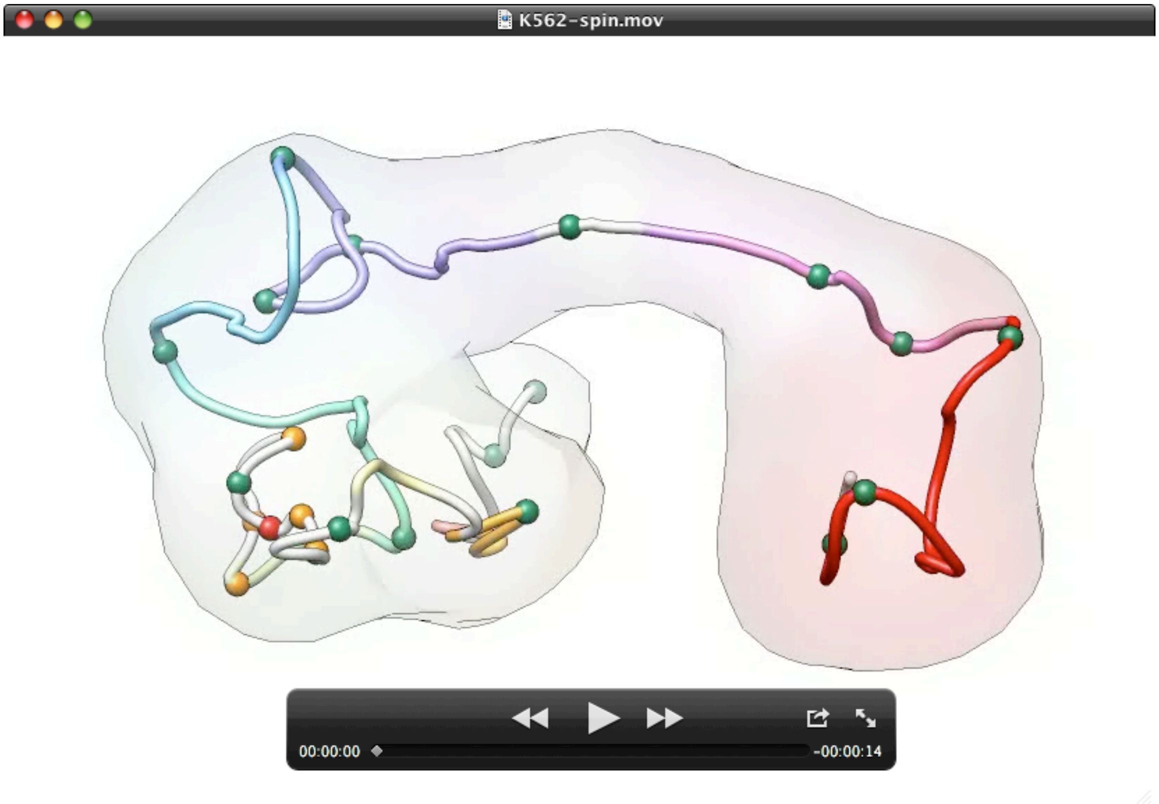


Supplementary Figure 6. Model resolution. The standard deviation of the applied Gaussian to the ensemble of solutions in cluster 2 of K562 models is plotted against the correlation coefficient of the Gaussian against the actual occupancy of the models. Green background defines a similarity area where 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.

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.



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.



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).

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 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_7lhg18lchr16:15091-18344) 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**.

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**.