

# CTCF-mediated transcriptional regulation through cell type-specific chromosome organization in the $\beta$ -globin locus

Ivan Junier<sup>1,2,3,\*</sup>, Ryan K. Dale<sup>4</sup>, Chunhui Hou<sup>4</sup>, François Képès<sup>1,\*</sup> and Ann Dean<sup>4,\*</sup>

<sup>1</sup>Epigenomics Project and institute of Systems and Synthetic Biology, Genopole<sup>®</sup>, CNRS, University of Evry, 5 rue Henri Desbruères, Evry F-91030, <sup>2</sup>Institute of Complex Systems, Paris, France, <sup>3</sup>Centre for Genomic Regulation, C/ dr. Aiguader, 88, 08003 Barcelona, Spain and <sup>4</sup>Laboratory of Cellular and Developmental Biology, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892, USA

Received February 24, 2012; Revised May 4, 2012; Accepted May 11, 2012

## ABSTRACT

The principles underlying the architectural landscape of chromatin beyond the nucleosome level in living cells remains largely unknown despite its potential to play a role in mammalian gene regulation. We investigated the three-dimensional folding of a 1 Mbp region of human chromosome 11 containing the  $\beta$ -globin genes by integrating looping interactions of the CCCTC-binding insulator protein CTCF determined comprehensively by chromosome conformation capture (3C) into a polymer model of chromatin. We find that CTCF-mediated cell type-specific interactions in erythroid cells are organized to favor contacts known to occur *in vivo* between the  $\beta$ -globin locus control region (LCR) and genes. In these cells, the modeled  $\beta$ -globin domain folds into a globule with the LCR and the active globin genes on the periphery. In contrast, in non-erythroid cells, the globule is less compact with few but dominant CTCF interactions driving the genes away from the LCR. This leads to a decrease in contact frequencies that can exceed 1000-fold depending on the stiffness of the chromatin and the exact position of the genes. Our findings show that an ensemble of CTCF contacts functionally affects spatial distances between control elements and target genes contributing to chromosomal organization required for transcription.

## INTRODUCTION

Chromatin interactions that form loops between enhancers and target promoters over large linear distances are common in metazoans (1–3). The specificity of such interactions is thought to be influenced by DNA–protein complexes called insulators, also through the formation of looped domains (4). How enhancer–promoter proximity is established and how topological domains contribute to transcription activation is not known. Moreover, recent technological advances have revealed genome-wide networks of intra- and inter-chromosomal sites of contact (5–7). Thus, a major question that arises is to what extent nuclear proximities between sites in the genome reflect specific long-range interactions with consequences for the transcriptional activity of underlying sequences.

The  $\beta$ -globin locus is flanked by two sites, HS5 and 3'HS1, that are occupied by the insulator protein CTCF. Chromosome conformation capture (3C) experiments have shown looping together of these sites in erythroid precursor cells before the globin genes are transcribed (8). Encompassed in this loop are the  $\beta$ -globin family of genes, 5' embryonic  $\epsilon$ , fetal  $\gamma^A$  and  $\gamma^G$  and 3' adult  $\beta$  and the locus control region (LCR) that activates them sequentially in erythroid cells (9). Globin genes are subsequently found in proximity to the LCR as they are activated during development (10). These looping interactions require erythroid transcription factors GATA-1, FOG-1 and EKLF, and the ubiquitous nuclear factor Ldb1, likely as participants in multimeric protein complexes bridging the LCR and genes (11–13). In addition, the chromatin

\*To whom correspondence should be addressed. Tel: +1 301 496 6068; Fax: +1 301 496 6068; Email: anndean@helix.nih.gov

Correspondence may also be addressed to François Képès. Tel: +33 169474431; Fax: +33 169474437;

Email: francois.kepes@epigenomique.genopole.fr

Correspondence may also be addressed to Ivan Junier. Tel: +34 93 3160166; Fax: +34 93 3969983; Email: i.junier@gmail.com

Present address:

Chunhui Hou, Department of Biology, Emory University, Atlanta, GA, USA.

remodeler Brg-1 is required at an early time in the looping process (14).

CTCF has been implicated in the organization of individual chromosomes and chromosome territories (15), and a role for CTCF in both organizational and transcriptional regulation has been suggested by the correlation of CTCF-mediated loops genome wide with lamin-associated domains demarcating regions of differing epigenetic marks (6). However, a functional role for CTCF site interactions has been probed in only limited instances (16). For example, the CTCF-dependent imprinting control region (ICR) in the *Igf2/H19* locus forms an enhancer blocking loop with a site upstream of *Igf2* on maternal chromosomes, isolating the gene from enhancers it shares with H19 (17,18). Loss of CTCF or its sites results in loss of the loop and activation of *Igf2*. In contrast, individual deletion or disruption of  $\beta$ -globin HS5 or 3'HS1 CTCF sites did not affect  $\beta$ -globin transcription in mouse erythroid cells (19–22).

Recent integrative numerical frameworks have allowed the generation of spatial conformations of chromosomes *in silico* that are compatible with 3C-like high-throughput data (23–28). More causal approaches have been used to study the mechanisms that may be responsible for chromosome organization by modeling chromatin as a polymer (7,29–34). In a theoretical analysis using such a polymer model, Mukhopadhyay *et al.* suggested that long range interactions could have a crucial impact on the contact properties of genes and their enhancers (35). How chromatin configurations may influence gene expression remains unclear.

Here, we investigate whether regional long-range CTCF contacts contribute to transcriptional control at the  $\beta$ -globin locus. To this end, we constructed chromosome models that satisfy the constraints of long-range CTCF dependent interactions over 1 Mb of human chromosome 11 encompassing the  $\beta$ -globin locus (36). The models consist of a single, self-avoiding polymer chain along which specific sites (the CTCF sites here) are able to interact (30). Using 3C data as input, simulations of the best-fit models make it possible to quantify the spatial proximity between the globin genes and their LCR in erythroid K562 cells and non-erythroid 293T cells. The results show that the set of regional CTCF site interactions drive the LCR and globin genes closer together in expressing cells than in silent cells and nucleate a globule with the contacting LCR/globin gene pair deployed away from the center.

## MATERIALS AND METHODS

### Modeling the chromatin around the $\beta$ -globin locus: a self-avoiding WLC with designated interacting sites

We simulated 1 Mbp long chromosome segments corresponding to the set of interacting CTCF sites in each cell type (Figure 1), using a polymer model of the chromatin fiber (37) (Supplementary Figure S1). In this worm-like chain (WLC) model, the chromatin fiber behaves as a flexible chain (37–39). We further excluded the possibility of the chromatin fiber overlapping itself (self-avoidance

effect) by specifying the chain diameter to be 30 nm. The base-pair density along the fiber was fixed at 150 bp/nm (37). In order to prevent the formation of knots, simulated chromosomes were circularized by adding a 0.2 Mbp long fiber free of CTCF sites (not represented in the figures), which is long enough (1.3  $\mu$ m) to prevent looping of the ends of the 1 Mbp chromosome. This led to a final 1.2 Mbp chromosome segment, which was confined to a spherical volume of 1  $\mu$ m diameter.

Along the chromatin, designated sites akin to CTCF sites interact with each other according to a harmonic potential  $V(x) = \frac{k}{2}x^2$  whose value is lowest for a spatial distance between the sites,  $x$ , that is equal to zero. For any given pair of known interacting sites, the spatial distance,  $x$ , separating them varies according to the chromosome conformation and the larger the intensity,  $k$ , of their potential, the more effective the potential  $V(x)$ , that is, the more frequently the sites interact. Unlike models where interactions are short-range (30), long-range potentials such as harmonic ones prevent interacting sites from being spatially distant from each other. This allowed us to comprehensively study the statistical properties of the models by running numerous short (i.e. <1 day) simulations.

### Fitting the polymer models: an iterative procedure

3C experiments quantify interaction frequency between two sites. For input to the models, these values were converted to potential intensities. To this end, the potential intensities for a given set of 3C contact frequencies were iteratively updated to obtain contact frequencies *in silico* as close as possible to contact frequencies *in vivo* (Supplementary Material). In addition, we do not consider interactions between sites that are in contact with a common third site in order to avoid non-linear effects on the uniqueness of the solution of the iterative algorithm (Supplementary Figure S2).

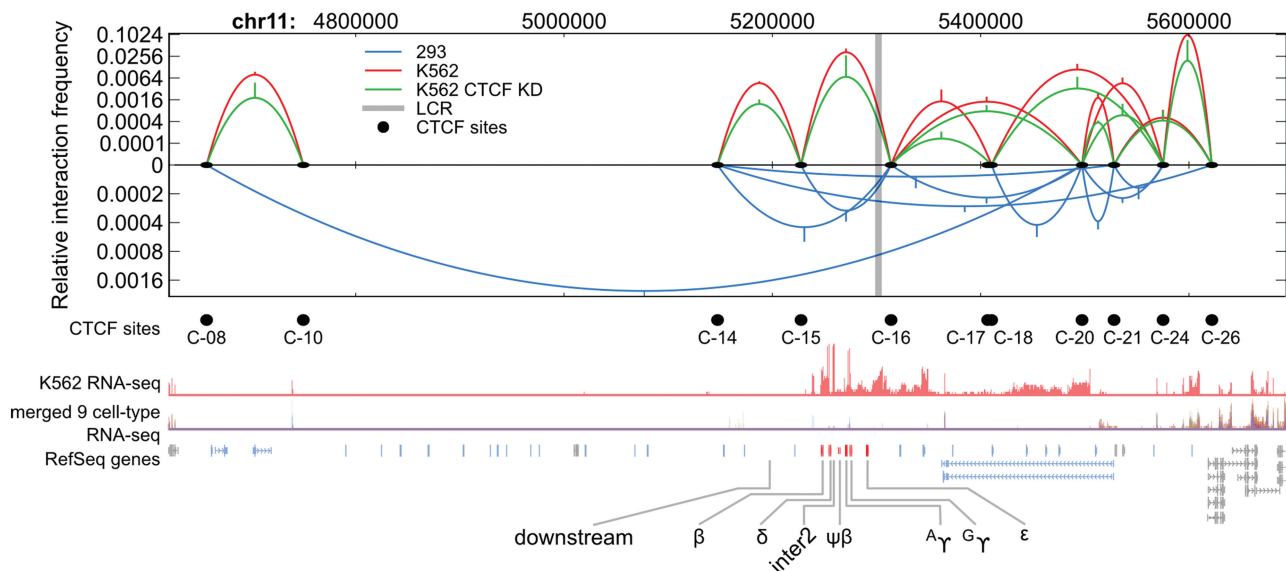
### Computing statistical properties of best-fit polymer models

Given a best-fit polymer model, we used distributed computer facilities in order to sample an equilibrium set of conformations as large as possible. An equivalent computation time of 3100 days (~8.5 years) was run on the European Grid Infrastructure (EGI) using two virtual organizations (VO): biomedical computing (biomed) and complex systems computing (vo.iscpif.fr, see <http://www.iscpif.fr/tiki-index.php?page=csgid>). The use of the distributed facilities was accomplished through the OpenMOLE project (<http://www.openmole.org/>) (40). For a more detailed description of the methods, see Supplementary Data.

## RESULTS

### CTCF site long-range interactions surrounding the $\beta$ -globin locus are primarily cell type specific

In a previous study, we used chromatin immunoprecipitation combined with microarray analysis (ChIP-chip) and ChIP and qPCR to survey and validate CTCF



**Figure 1.** CTCF-based chromatin loops differ between K562 cells and 293T cells. CTCF site interactions from Hou *et al.* (36) are depicted in a schematic diagram of the globin locus. Pairwise interaction frequencies detected by 3C are shown by arcs for K562 (red), K562 CTCF RNAi knockdown (green) and 293T (blue) cells. Black dots below the *x*-axis indicate CTCF-occupied sites confirmed by qPCR used in this analysis. The LCR is indicated by a gray vertical line. RNA seq data from the ENCODE Consortium (CalTech data set from Wold and Myers groups) for K562 cells (red track) and nine other non-erythroid cell types combined (multicolor track) are shown as normalized read density (reads per million, RPM) on a  $\ln(x+1)$  scale. RefSeq gene models for the locus are shown, including globin genes (orange), odorant receptor genes (blue) and other genes (gray). All coordinates and data are for the hg19 assembly.

sites up- and down-stream of the  $\beta$ -globin locus on human chromosome 11 in K562 cells, where the fetal  $\gamma$ -globin gene is transcribed and in 293T cells where the globin locus is silent (36). A wealth of structural information is available for K562 cells through the ENCODE consortium and these cells demonstrate the appropriate LCR- $\gamma$ -globin chromatin looping while 293T cells do not (41) (Kiefer, C.M. and Dean, A., unpublished data). All sites analyzed were similarly occupied by CTCF in the two cell types and all were co-occupied by cohesin complex members Rad21, SMC1 and SMC3. We then used 3C to determine, in these two cell types and in K562 cells in which CTCF had been reduced using RNAi (K562kd), the frequencies of interaction among all occupied CTCF sites surrounding the globin locus over a chromosomal range of 1.4 Mbp. 3C quantifies the interaction frequency between two sites in chromatin after formaldehyde cross-linking of protein and DNA, restriction enzyme digestion, re-ligation and detection of novel sequence junctions by qPCR (42).

A representation of the interactions previously observed (36) is presented in Figure 1. Each black dot represents an occupied CTCF site. Each parabola represents a pairwise CTCF site interaction where the height of the curve indicates the relative frequency of interaction on a log scale. Also represented are the decreased interactions between these sites when CTCF is knocked down in K562 cells by RNAi, reducing  $\gamma$ -globin transcription and allowing incursion into the locus of the repressing histone modification H3k9me2. Although CTCF was bound at nearly identical locations throughout the locus in the different cell types, cell type-specific patterns of long-range

interactions distinguish K562 cells (red) from 293T cells (blue) with 60% of interactions being cell type-specific. Interestingly, the loop between  $\beta$ -globin 3'HS1 and HS5 (C15 and C16, respectively) was observed in both cell types. In K562kd cells, all interaction frequencies were significantly reduced compared to CTCF-replete cells but the pattern of interactions was unchanged. These data raise the possibility that CTCF organizes the chromosomal domain structure surrounding the globin locus to influence transcription.

#### Modeling cell type-specific interactions between CTCF sites

To address the issue of whether differential cell type-specific CTCF contacts affect the conformation of the globin region, a polymer model of the 30 nm chromatin fiber was designed for each of the three cell types under investigation: K562, K562kd and 293T. Disregarding its internal structure, chromatin behaves as a flexible chain that self-avoids. We used the simplest continuous model, known as the WLC, which includes flexibility properties of a single chain (37–39) and considers the self-avoiding nature of a 30 nm thick fiber. Each WLC model represents a 1 Mbp region of human chromosome 11 containing the interacting CTCF sites and the  $\beta$ -globin locus. Interaction forces were determined by iterating the DNA-folding algorithm until the frequency of contacts between the sites corresponded to the experimentally measured CTCF-dependent chromatin contacts (Figure 1) (36). Importantly, CTCF interaction frequencies were the sole input to the polymer models and the inputs did not include the LCR or any gene promoter sites. This allowed us to

make independent *in silico* measurements of LCR–promoter proximity during the model runs using these proximities as a ‘read-out’ of the potential transcriptional state of the locus.

3C experiments measure interactions in a population of cells. We assume that the measured population-averaged frequencies are analogous to time-averaged frequencies at equilibrium in the models—that is, averaging over populations is equivalent to averaging over time. Two parameters of the model are adjustable, namely, the strength of interaction between the CTCF sites (interaction potential) and the persistence length, which describes the flexibility status of the chromatin fiber—the larger the persistence length, the less flexible the chromatin fiber (Supplementary Figure S1). Since the persistence length is not precisely known *in vivo*, we studied three values (100, 200 and 300 nm) typical of those reported in the literature (38). We then iteratively and independently adjusted intensities of the interaction potentials so that equilibrium contact frequencies between CTCF sites best matched experimentally measured frequencies (36). These best-fit interaction potentials were subsequently used in a benchmark polymer model to compute statistical properties, such as equilibrium distributions of the spatial distances between the globin genes and the LCR and typical chromatin conformations.

The pattern of intensities obtained for the interaction potentials in K562 and K562kd cells were comparable (Supplementary Table S1), with intensities reduced in K562kd cells compared to K562 cells. This indicates that interactions are based on a similar but weakened mechanism, in accordance with the incomplete knockdown by RNAi which allows some CTCF protein molecules to remain associated with chromatin (36). In contrast, intensities obtained for the potentials in 293T cells were globally much lower than those in K562 cells, consistent with the overall lower frequency of 3C interactions observed in 293T cells. More specifically, two categories of interactions of qualitatively different natures can be distinguished in 293T cells. The first category consists of the pairs C-08/C-20 and C-20/C-21 (Figure 1) for which the interaction intensities were relatively strong. The second category contains all the other pairs of interacting sites for which the intensities were at least two orders of magnitude lower (Supplementary Table S1).

#### Sites within the globin locus domain are in closer proximity to the LCR in erythroid cells than in non-erythroid cells

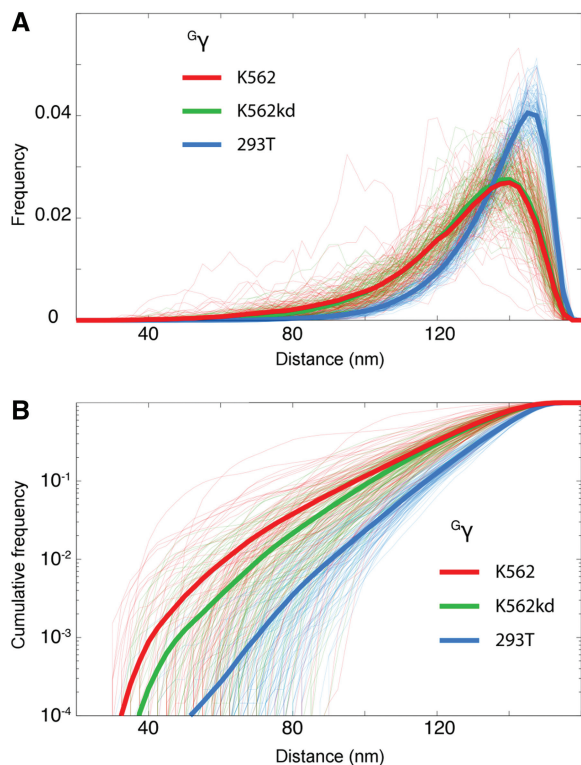
Next, we sought to tie together the modeled interactions in these cells with known expression patterns. Since the primary difference between the cell types is the expression of globin genes, and interaction with the LCR is required for transcribing these genes, we chose to use the proximity of the LCR to a series of targets in the locus as a proxy for transcriptional activity. The primary set of targets consisted of the transcription start site (TSS) of each of the globin genes in the locus. In addition, we included the TSS for the  $\beta$ -globin pseudogene, two intergenic sites and a long-distance site downstream from  $\beta$ -globin (Figure 1).

The distributions of 3-D spatial distances between the LCR and each gene were measured for each cell type over the course of 100 simulations of the corresponding best-fit model. Frequencies and cumulative frequencies show the time fraction two sites spend at a given distance or at a distance lower than a given distance, respectively. These are displayed in Figure 2 for a representative target, the  $G_\gamma$ -globin gene, for all simulations (thin curves) and for the corresponding averages (thick curves). Figure 2A demonstrates a clear shift of the distributions toward large distances for 293T cells compared to K562 cells. Looking across all targets (Supplementary Figure S3), as genomic distance from the LCR increases, so does the three-dimensional spatial distance. This effect is stronger in 293T cells than in K562 cells so that at the largest genomic distances, there is the greatest difference between cell types. Cumulative frequencies, shown on a log scale, reveal a tendency for the globin genes to be closer to the LCR in K562 than in K562kd (Figure 2B and Supplementary Figure S3). Importantly, the results do not qualitatively depend on microscopic parameters (the persistence length) of the chromatin fiber. Taken together, these results show that in K562 cells, regional CTCF sites interact in such a way that globin genes are closer to the LCR than in 293T cells, and to a lesser extent, than in K562kd cells.

#### Effects of chromatin features and genomic organization on LCR contacts

We next studied how the physical features of the chromatin fiber affect the capacity of globin genes to make contact with the LCR. We compared results obtained for three levels of stiffness of the chromatin fiber typical of those reported *in vivo*, which are defined by persistence lengths equal to 100, 200 and 300 nm, respectively (Figure 3). Since sites are likely to interact through protein complexes (e.g. transcription factors, RNA polymerase), we chose a value of 10 nm to represent the space taken up by these complexes, and defined an interaction event to exist when the LCR was within 10 nm of a target site, i.e. when there was  $\leq 40$  nm between the centers of the two 30-nm chromatin fibers. We then quantified the frequency of LCR–target interaction events across the locus.

The frequency of an interaction event can be shown on a negative log scale so that higher values indicate greater isolation of the LCR from targets. This analysis highlights two important points (Figure 3). First, the stiffer the chromatin fiber, the larger the difference between 293T cell and K562 cell contact frequencies. For example, the isolation of  $G_\gamma$  from the LCR is  $\sim 20$ -fold larger in 293T than in K562 at a persistence length of 100 nm; it is at least 1000-fold larger at a persistence length of 300 nm, which is within the range of the expected persistence lengths *in vivo* (38). Second, the smaller the genomic distance to the LCR, the larger the difference in contact frequencies between K562 and K562kd cells. For example, at a persistence length equal to 200 nm, LCR contacts with  $\beta$  (genomic distance = 53 kbp) are only  $\sim 1.5$ -fold less frequent in K562kd cells than in K562 cells; contacts with  $\gamma$  (genomic distance = 26 kbp) are  $\sim 4.5$ -fold less



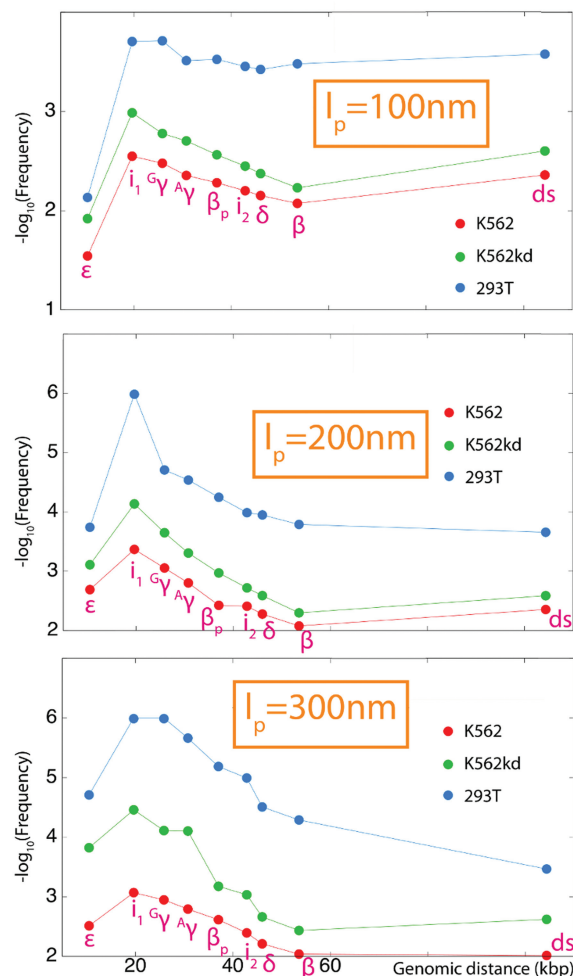
**Figure 2.** Distribution of distances between  $G_\gamma$  and the LCR using a best-fit polymer model of the  $\beta$ -globin locus. The curves show the distributions of the spatial distances (in nm) separating the LCR from  $G_\gamma$  in the three cell types K562, K562kd and 293T, for a chromatin fiber with a 200 nm persistence length. Thin dotted curves are distributions obtained from 100 different runs, which were used to compute the average curves (thick solid curves). (A) The y-axis reports the time fraction the two sites spend at the distance reported along the x-axis. (B) The y-axis reports the time fraction the two sites spend at a distance lower than that reported along the x-axis. Note that at distances lower than 100 nm, the cumulative frequency is 10–100 times larger in K562 cells than in 293T cells.

frequent. The observation that the  $G_\gamma$ -globin gene is more isolated from the LCR in K562kd cells than in CTCF-replete K562 cells is consistent with the reduction of  $G_\gamma$ -globin expression by RNAi-mediated CTCF knock down (36).

### Spatial conformations of the globin locus during active transcription of the $\gamma$ -globin genes

Compared to 293T cells, CTCF interactions in K562 cells provide an overall modeled conformation of the globin locus that favors contact between the LCR and globin genes (Figure 3). We ran 100 additional simulations of a K562 polymer model to study the topology of the chromatin conformations that involved LCR–globin gene interactions. For each simulation, snapshots were stored of the first configuration that was reached when the LCR was in contact with a globin gene of interest. Figure 4A shows one such configuration for  $G_\gamma$  (see Supplementary Figure S4 for other genes).

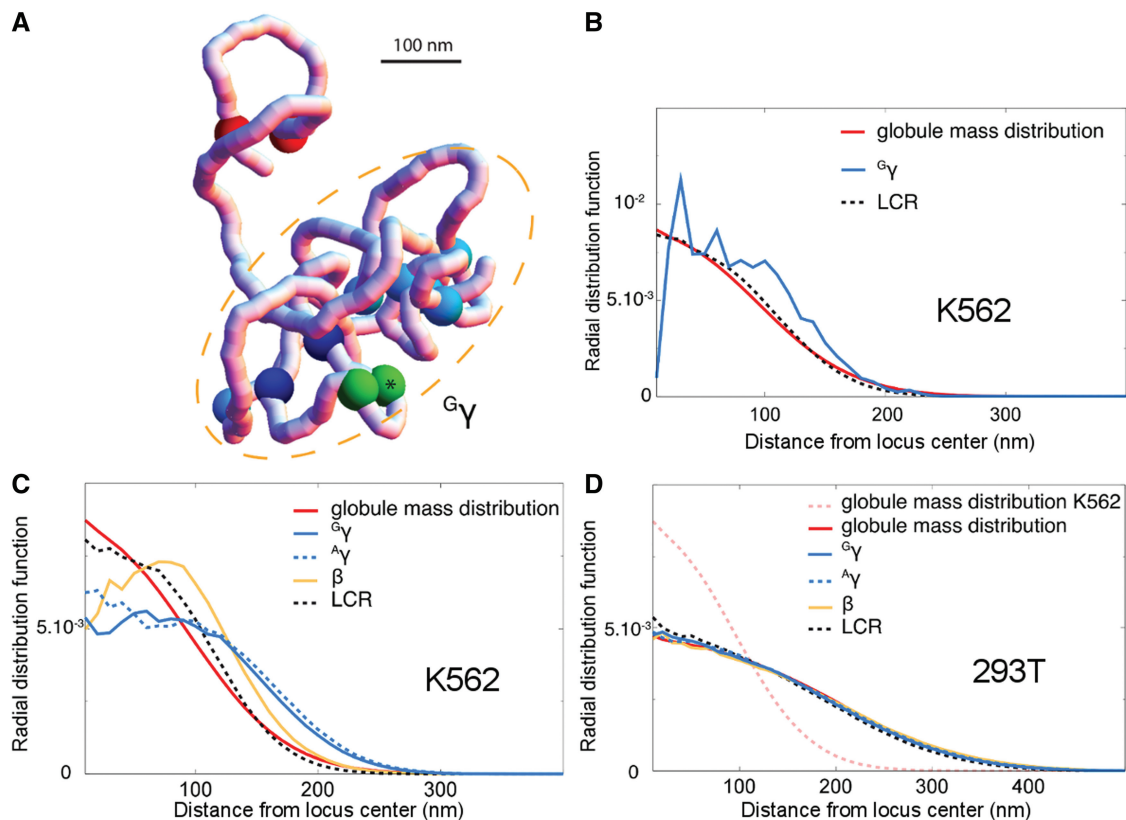
Like the 3C contact maps to which interaction potentials were fitted (Figure 1), the snapshots reveal conformations that are composed of one globule, where many



**Figure 3.** Interaction frequency as a function of the genomic distance to the LCR. The panels show the co-logarithm of the contact frequency between the LCR and  $\beta$ -globin genes, the intergenic sites ( $i_1$ ,  $i_2$ ) and the downstream site (ds) for three persistence lengths. Higher values indicate increased isolation of LCR from target. A contact is defined whenever the distance between the sites along the polymer and the LCR is lower than 40 nm. This figure shows that the stiffer the chromatin, the lower the contact frequency in 293T cells. Note that due to numerical limitations, frequencies lower than  $10^{-6}$  cannot be detected: they are reported to be equal to  $10^{-6}$  (e.g., for  $G_\gamma$  at a persistence length of 300 nm).

contacts are present (blue sites), plus one loop where only two sites interact (red sites). These two structures tend to repel each other, due to volume exclusion effects (43). Interestingly, the contacts between the LCR and  $G_\gamma$  gene (green sites) occur most often on the periphery of the globule. We quantified this with a radial distribution function, which shows the probability that a site is found at a certain distance away from the center of mass of the globule.

Whereas the LCR does not have, on average, a preferred location with respect to the center of mass of the globule (compare red and black dotted curves in Figure 4B), interaction events between  $G_\gamma$  and the LCR tend to happen farther out from the center of the globule than is typical for the locus as a whole (Figure 4B,



**Figure 4.** Chromatin conformations favoring contacts between the  $\beta$ -globin genes and LCR in K562 cells. (A) Typical conformation of the 1 Mbp regions around the  $\beta$ -globin locus during a contact between LCR (green + star) and  $G_\gamma$  (green). Blue sites: CTCF sites that form a connected network of interaction (Supplementary Figure S1). Darkest blue sites: CTCF sites that surround the  $\beta$ -globin locus. Red sites: the isolated interaction between C-08 and C-10. The conformation can be divided into a loop (stabilized by the red sites) and a compact globule (dashed orange ellipse) encompassing the region from C-03 to C-10. (B) Spatial location of the contact: using 1000 equilibrium simulations of the same best-fit polymer as in A, we report (i) the radial mass distribution of the compact globule, i.e. the average probability density for the location of the C-03 to C-10 region with respect to the globule center of mass; (ii) the radial distribution of  $G_\gamma$  and LCR during contacts and (iii) the radial distribution of the LCR (no matter the position of  $G_\gamma$ ). One can see that the  $G_\gamma$ /LCR contacts tend to occur away from the globule center. (C) Spatial location of the globin genes in K562 (obtained from 100 simulations of the best-fit polymer). Genes tend to be located away from the center regardless of LCR contact. Large distances are particularly enhanced in the case of the  $\gamma$  genes. (D) Same as in C but for 293T cells. No particular location can be observed for any of the genes.

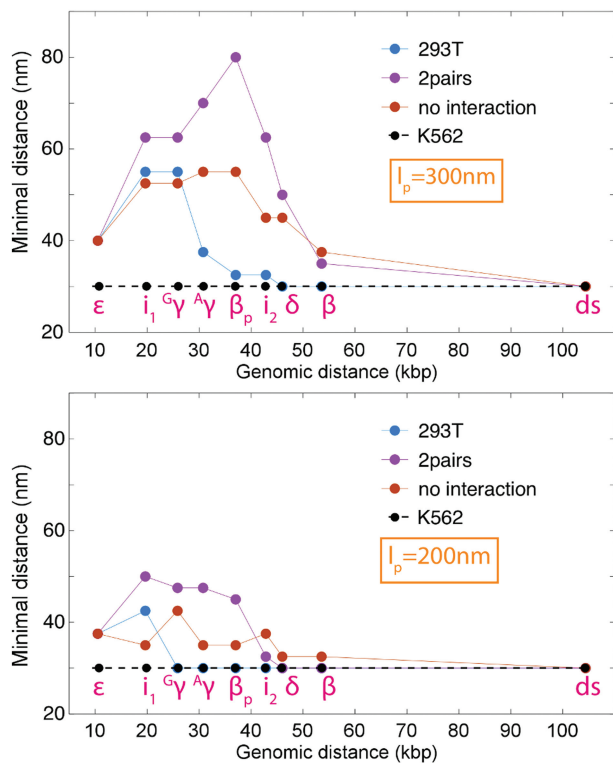
blue; note that the blue line is less smooth since LCR- $G_\gamma$  interactions only occur during a subset of all time steps represented by the red and black curves). Extending this analysis to  $A_\gamma$  and  $\beta$  indicates that all globin genes, but particularly  $\gamma$ -globin genes, tend to be located more peripherally to the globule regardless of LCR contact (Figure 4C). In contrast, in 293T cells, where the globule is less compact, no preferential location is observed for any of the locus sites of interest (Figure 4D). These findings suggest that, in addition to favoring contacts with the LCR, the CTCF-driven globule in K562 cells tends to displace the genes to be activated, i.e. the  $\gamma$ -globin genes here, away from the surrounding chromatin.

#### Dominant CTCF interactions and stiff chromatin prevent contacts between the LCR and globin genes in 293T cells

The interaction potentials observed in 293T cells can be divided into two categories based on strength (Supplementary Table S1). The strongest potentials are between C-08 and C-20 and between C-20 and C-21.

A polymer model where these interactions alone are present leads to a reduction of the tendency for globin genes to be spatially close to the LCR when the chromatin fiber is stiff (Supplementary Figure S5). To investigate the influence of these interactions, in particular whether the strongest interactions found in 293T cells are sufficient to decrease LCR-gene interactions compared to K562 cells, we used two additional models: one where only the two strongly interacting sites are present (ignoring all other interactions measured by 3C in 293T cells) and another using chromatin with no interacting sites. Since the interaction events we defined earlier (40 nm between chromatin fiber centers) do not always occur in 293T cells as they do in K562 cells, we used the minimal distance obtained in 100 simulations as an alternative metric to represent LCR-target proximities.

The model with no interacting sites serves as a baseline (red lines, Figure 5). One might hypothesize that introducing any interacting sites in this locus would bring the LCR closer to targets on average. However, interestingly, the model with just two pairs of strongly



**Figure 5.** Minimal distances between the LCR and the  $\beta$ -globin genes. Each point represents the minimal spatial distance obtained in our simulations between a globin gene and the LCR at persistence lengths 200 and 300 nm. In K562 cells, the minimal distance is 30 nm, indicating that physical contacts are always possible. The situation is different for 293T cells. In particular, when only the two strongest potential pairs in 293T cells are considered ('2pairs'), the minimal distance can be as large as 80 nm (purple points). The red points show the results where no interaction is considered. By taking into account all the interactions involved in 293T cells, one can observe for persistence lengths of the chromatin around 300 nm a complete isolation specific to  $\gamma$  and points linearly closer to the LCR.

interacting sites results in a substantial isolation of the LCR from the targets over most of the locus compared to the model with no interacting sites (Figure 5, compare purple and red lines). This strong isolating effect decreases when the rest of the 293T interaction potentials are included (Figure 5, blue lines), even though the short minimal distances reported here remain very infrequent (Figure 3). These additional interactions tend to re-establish contacts of a subset of the globin genes, e.g.  $\beta$  (Figure 5, upper panel). The isolating effect of the two pairs of strongly interacting sites suggests that they may play an important role in cell type-specific chromosome architecture surrounding the  $\beta$ -globin locus.

## DISCUSSION

Above the megabase scale, Hi-C studies have revealed that contact maps of individual human chromosomes can be divided into two or three classes of regions, with one class that is more transcriptionally active than the other(s) (7,44). Novel conformation capture techniques have

further revealed that inter-chromosomal contacts not only correlate with transcriptional activity but also with CTCF-bound regions (26,44). Between the megabase scale and the nucleosome level of chromatin folding little is known about chromatin organization. In this study, we use a thermodynamic framework based on a WLC model of chromatin where designated sites along the DNA, here CTCF sites over 1 Mbp around the  $\beta$ -globin locus, are able to form pairwise interactions (30,36). The results indicate that (i) proximity within the globin locus between the LCR and genes and (ii) displacement of the active LCR/globin pair away from surrounding chromatin are both facilitated by overall erythroid cell type-specific CTCF chromosome organization. Thus, providing a thermodynamic topological model with experimental measurements of interaction frequencies of CTCF sites alone predicts a chromosome conformation pattern consistent with the known biology of the human  $\beta$ -globin locus.

Our analysis uses long-range interaction potentials to efficiently sample conformations that are consistent with the CTCF contact map. The main advantage of this choice is that it allows a comprehensive statistical analysis of chromosome folding. The disadvantage is that real interactions have a shorter range in the cell, and the choice of long-range interactions may alter the statistical properties of the simulated chromosome conformations. Nevertheless, long-range interactions might mimic the real situation in cells, where non-functional chromosome conformations appear to be dis-favored by mechanisms that remain elusive.

CTCF sites within 1 Mbp of the  $\beta$ -globin locus were occupied in both K562 (active locus) and 293T (silent locus) cells, however, their pairwise interactions were primarily cell type specific, suggesting that factors other than CTCF are important for the specificity of loop formation. In some loci, cohesin co-occupancy is the determinant of CTCF insulator function, however, cohesin was present at CTCF sites we studied in both K562 and 293T cells (45). Modifications to these proteins or recruitment of additional components may explain the difference in loop formation between the cell types. In addition, looping interactions beyond those we studied are likely to impinge on chromosome organization in the  $\beta$ -globin region, including those mediated by different protein factors (46,47). Nonetheless, our simulations indicate that consideration of CTCF looping contacts alone predicts a cell type-specific chromosome organization in which the probability of intra-loop LCR/globin gene contacts is favored for K562 cells and hindered in 293T cells. Interestingly, numerous interacting sites around the  $\alpha$ -globin locus were found to be bound by CTCF protein and our previous data revealed that  $\alpha$ -globin transcription was reduced upon CTCF knock down (36,48). This further reinforces the idea of an important role of regional CTCF sites in the interplay between the architecture of chromosomes and their function.

The modeled distances from the LCR to globin genes as a whole are smaller in K562 cells, where the locus is active, than in 293T cells where the locus is inactive. Moreover, reduction of CTCF in K562 cells using RNAi, which leads

to a reduction of  $\gamma$ -globin transcription (36), increases these modeled distances compared to cells replete with CTCF. Thus, CTCF site contacts in K562 cells may provide a constraint such that the collision dynamics between globin genes and the LCR are the most appropriate for LCR–globin gene interactions. We suggest that the relative closeness of the LCR to the coding regions of the locus modeled in K562 cells, compared to 293T cells, favors ‘sampling’ of contacts between the LCR and genes that, if stabilized by protein–protein interactions, allow looping and transcription activation. Our results indicate gains in sampling time that can exceed 1000-fold, depending on the stiffness of the chromatin and the position of the gene. In this view, the active K562  $\gamma$  genes would successfully establish LCR contact, whereas, the silent  $\beta$  gene would not, due to the modulation of required developmental stage-specific proteins or protein complexes. The data do not exclude the possibility that LCR/globin gene contact is the initial organizing event for the locus during erythroid differentiation and that the CTCF-looping ensemble evolves to stabilize this interaction. Reduction of the LCR/globin gene interaction when CTCF is decreased by RNAi shows the importance of the CTCF-mediated loop ensemble to maintain contact.

For transcription to occur, the LCR and gene to be activated must be accessible to the transcription machinery. Our studies highlight in K562 cells a modeled preferential exposure of the LCR and  $\gamma$  genes to the nuclear milieu at the surface of a CTCF-driven globule. This positioning is reminiscent of extrusion from a chromosome territory and might facilitate localization of the  $\gamma$ -globin/LCR ensemble within nuclear foci of transcription to which globin genes are known to migrate to achieve high levels of transcription (49–51). Interestingly, modeling of chromosome conformation capture carbon copy (5C) interactions in K562 cells suggested that the  $\alpha$ -globin genes cluster together with neighboring active housekeeping genes that surround them in the interior of a globule that might correspond to a transcription factory (48). In contrast to the  $\alpha$ -globin locus, the  $\beta$ -globin locus is embedded among odorant receptor genes that are silent in erythroid cells. Thus, the tendency for the active  $\gamma$ -globin gene and LCR to be on the periphery of a globule might be related to the difference in transcriptional state compared to neighboring genes.

This study highlights the interplay of numerous CTCF sites within a chromosomal region that interact to form multiple chromatin loops. The modeled conformation of the region reveals an ensemble of interactions (the globule), whose overall distribution may have more subtle biological implications (e.g. the propensity for  $\gamma$  to be on the periphery) than suggested by consideration of individual loops. This may help to explain why individual deletion or disruption of HS5 or 3'HS1 (C-16 and C-15, respectively, in Figure 1) did not affect  $\beta$ -globin transcription in mouse erythroid cells (19–22). Likewise, in this view, the insertion of an extra HS5 site (52) that creates two new loops and limits access of the LCR to globin genes could be interpreted as introducing an imbalance to the ensemble of interactions, similar to the two pairs of strong interacting CTCF sites in 293T cells that isolate

the LCR from the locus. These ideas suggest specific, testable biological hypotheses. For example, disrupting the strongly interacting CTCF site pairs in 293T cells should reduce LCR isolation from globin genes, although such manipulation is unlikely to affect transcriptional activity due to the absence of erythroid transcriptional activators in these cells. Alternatively, incremental deletion of CTCF sites in the K562 computational model could suggest which sites are most important for LCR/globin gene contact, and the predictions could then be tested experimentally.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1, Supplementary Figures 1–7 and Supplementary Material and Methods.

## ACKNOWLEDGEMENTS

We thank Olivier Martin and Elissa Lei for helpful comments on the manuscript.

## FUNDING

Intramural Program of the National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health [KIA 15508 to A.D.]. I.J. is supported by a Novartis grant (CRG) and thanks Région Île-de-France and ISC-PIF for financial and logistic support. This study was also supported by the Sixth European Research Framework (project number 034952, GENNETEC project), PRES UniverSud Paris, CNRS and Genopole (to F.K.). Funding for open access charge: Intramural Program, National Institute of Diabetes, and Digestive and Kidney Diseases, National Institutes of Health [KIA 15508 to AD].

*Conflict of interest statement.* None declared.

## REFERENCES

1. Krivega, I. and Dean, A. (2012) Enhancer and promoter interactions—long distance calls. *Curr. Opin. Genet. Dev.*, **22**, 1–7.
2. Ong, C.T. and Corces, V.G. (2011) Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat. Rev. Genet.*, **12**, 283–293.
3. Sexton, T., Bantignies, F. and Cavalli, G. (2009) Genomic interactions: chromatin loops and gene meeting points in transcriptional regulation. *Semin. Cell Dev. Biol.*, **20**, 849–855.
4. Wallace, J.A. and Felsenfeld, G. (2007) We gather together: insulators and genome organization. *Curr. Opin. Genet. Dev.*, **17**, 400–407.
5. Schoenfelder, S., Sexton, T., Chakalova, L., Cope, N.F., Horton, A., Andrews, S., Kurukuti, S., Mitchell, J.A., Umlauf, D., Dimitrova, D.S. *et al.* (2010) Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat. Genet.*, **42**, 53–61.
6. Handoko, L., Xu, H., Li, G., Ngan, C.Y., Chew, E., Schnapp, M., Lee, C.W., Ye, C., Ping, J.L., Mulawadi, F. *et al.* (2011) CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat. Genet.*, **43**, 630–638.
7. Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragozy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J.,



- Dorschner, M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
8. Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F. and de Laat, W. (2002) Looping and interaction between hypersensitive sites in the active  $\beta$ -globin locus. *Mol. Cell*, **10**, 1453–1465.
  9. Stamatoyannopoulos, G. (2005) Control of globin gene expression during development and erythroid differentiation. *Exp. Hematol.*, **33**, 259–271.
  10. Palstra, R.J., Tolhuis, B., Splinter, E., Nijmeijer, R., Grosveld, F. and de Laat, W. (2003) The  $\beta$ -globin nuclear compartment in development and erythroid differentiation. *Nat. Genet.*, **35**, 190–194.
  11. Vakoc, C.R., Letting, D.L., Gheldof, N., Sawado, T., Bender, M.A., Groudine, M., Weiss, M.J., Dekker, J. and Blobel, G.A. (2005) Proximity among distant regulatory elements at the  $\beta$ -globin locus requires GATA-1 and FOG-1. *Mol. Cell*, **17**, 453–462.
  12. Drissen, R., Palstra, R.J., Gillemans, N., Splinter, E., Grosveld, F., Philipsen, S. and de Laat, W. (2004) The active spatial organization of the  $\beta$ -globin locus requires the transcription factor EKLK. *Genes Dev.*, **18**, 2485–2490.
  13. Song, S.-H., Hou, C. and Dean, A. (2007) A positive role for NLI/Ldb1 in long-range  $\beta$ -globin locus control region function. *Mol. Cell*, **28**, 810–822.
  14. Kim, S.I., Bultman, S.J., Kiefer, C.M., Dean, A. and Bresnick, E.H. (2009) BRG1 requirement for long-range interaction of a locus control region with a downstream promoter. *Proc. Natl Acad. Sci. USA*, **106**, 2259–2264.
  15. Botta, M., Haider, S., Leung, I.X., Lio, P. and Mozziconacci, J. (2010) Intra- and inter-chromosomal interactions correlate with CTCF binding genome wide. *Mol. Syst. Biol.*, **6**, 426.
  16. Phillips, J.E. and Corces, V.G. (2009) CTCF: master weaver of the genome. *Cell*, **137**, 1194–1211.
  17. Murrell, A., Heeson, S. and Reik, W. (2004) Interaction between differentially methylated regions partitions the imprinted genes Igf2 and H19 into parent-specific chromatin loops. *Nat. Genet.*, **36**, 889–893.
  18. Kurukuti, S., Tiwari, V.K., Tavoosidana, G., Pugacheva, E., Murrell, A., Zhao, Z., Lobanekov, V., Reik, W. and Ohlsson, R. (2006) CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2. *Proc. Natl Acad. Sci. USA*, **103**, 10684–10689.
  19. Farrell, C.M., Grinberg, A., Huang, S.P., Chen, D., Pichel, J.G., Westphal, H. and Felsenfeld, G. (2000) A large upstream region is not necessary for gene expression or hypersensitive site formation at the mouse  $\beta$ -globin locus. *Proc. Natl Acad. Sci. USA*, **97**, 14554–14559.
  20. Bender, M.A., Reik, A., Close, J., Telling, A., Epner, E., Fiering, S., Hardison, R. and Groudine, M. (1998) Description and targeted deletion of 5' hypersensitive site 5 and 6 of the mouse  $\beta$ -globin locus control region. *Blood*, **92**, 4394–4403.
  21. Bender, M.A., Byron, R., Ragoczy, T., Telling, A., Bulger, M. and Groudine, M. (2006) Flanking HS-62.5 and 3' HS1, and regions upstream of the LCR, are not required for beta-globin transcription. *Blood*, **108**, 1395–1401.
  22. Splinter, E., Heath, H., Kooren, J., Palstra, R.J., Klous, P., Grosveld, F., Galjart, N. and de Laat, W. (2006) CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes Dev.*, **20**, 2349–2354.
  23. Duan, Z., Andronescu, M., Schutz, K., McIlwain, S., Kim, Y.J., Lee, C., Shendure, J., Fields, S., Blau, C.A. and Noble, W.S. (2010) A three-dimensional model of the yeast genome. *Nature*, **465**, 363–367.
  24. Bau, D., Sanyal, A., Lajoie, B.R., Capriotti, E., Byron, M., Lawrence, J.B., Dekker, J. and Marti-Renom, M.A. (2011) The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules. *Nat. Struct. Mol. Biol.*, **1**, 107–114.
  25. Fraser, J., Rousseau, M., Shenker, S., Ferraiuolo, M.A., Hayashizaki, Y., Blanchette, M. and Dostie, J. (2009) Chromatin conformation signatures of cellular differentiation. *Genome Biol.*, **10**, R37.
  26. Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F. and Chen, L. (2012) Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.*, **30**, 90–98.
  27. Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A. and Cavalli, G. (2012) Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell*, **148**, 458–472.
  28. Rousseau, M., Fraser, J., Ferraiuolo, M.A., Dostie, J. and Blanchette, M. (2011) Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. *BMC Bioinformatics*, **12**, 414.
  29. Mateos-Langerak, J., Bohn, M., de Leeuw, W., Giromus, O., Manders, E.M., Verschure, P.J., Indemans, M.H., Gierman, H.J., Heermann, D.W., van Driel, R. *et al.* (2009) Spatially confined folding of chromatin in the interphase nucleus. *Proc. Natl Acad. Sci. USA*, **106**, 3812–3817.
  30. Junier, I., Martin, O. and Kepes, F. (2010) Spatial and topological organization of DNA chains induced by gene co-localization. *PLoS Comput. Biol.*, **6**, e1000678.
  31. Marenduzzo, D., Faro-Trindade, I. and Cook, P.R. (2007) What are the molecular ties that maintain genomic loops? *Trends Genet.*, **23**, 126–133.
  32. Dorier, J. and Stasiak, A. (2010) The role of transcription factories-mediated interchromosomal contacts in the organization of nuclear architecture. *Nucleic Acids Res.*, **38**, 7410–7421.
  33. Cook, P.R. and Marenduzzo, D. (2009) Entropic organization of interphase chromosomes. *J. Cell Biol.*, **186**, 825–834.
  34. Rosa, A. and Everaers, R. (2008) Structure and dynamics of interphase chromosomes. *PLoS Comput. Biol.*, **4**, e1000153.
  35. Mukhopadhyay, S., Schedl, P., Studitsky, V.M. and Sengupta, A.M. (2011) Theoretical analysis of the role of chromatin interactions in long-range action of enhancers and insulators. *Proc. Natl Acad. Sci. USA*, **108**, 19919–19924.
  36. Hou, C., Dale, R. and Dean, A. (2010) Cell type specificity of chromatin organization mediated by CTCF and cohesin. *Proc. Natl Acad. Sci. USA*, **107**, 3651–3656.
  37. Cui, Y. and Bustamante, C. (2000) Pulling a single chromatin fiber reveals the forces that maintain its higher-order structure. *Proc. Natl Acad. Sci. USA*, **97**, 127–132.
  38. Langowski, J. and Heermann, D.W. (2007) Computational modeling of the chromatin fiber. *Semin. Cell Dev. Biol.*, **18**, 659–667.
  39. Strick, T.R., Dessinges, M.-N., Charvin, G., Dekker, N.H., Allemand, J.-H., Bensimon, D. and Coquette, V. (2003) Stretching of macromolecules and proteins. *Rep. Prog. Phys.*, **66**, 1–45.
  40. Reuillon, R., Chuffart, F., Leclaire, M., Faure, T., Dumoulin, N. and Hill, D. (2010) Declarative task delegation in OpenMOLE. In: International Conference on High Performance Computing and Simulation (HPCS 2010), IEEE, Caen, June 28–July 2, pp. 55–62.
  41. Kiefer, C.M., Lee, J., Hou, C., Dale, R.K., Lee, Y.T., Meier, E.R., Miller, J.L. and Dean, A. (2011) Distinct Ldb1/NLI complexes orchestrate gamma-globin repression and reactivation through ETO2 in human adult erythroid cells. *Blood*, **118**, 6200–6208.
  42. Dekker, J. (2006) The three 'C's of chromosome conformation capture: controls, controls, controls. *Nat. Methods*, **3**, 17–21.
  43. de Gennes, P.G. (1988) *Scaling Concepts in Polymer Physics*. Cornell University Press, Ithaca, NY.
  44. Yaffe, E. and Tanay, A. (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.*, **43**, 1059–1065.
  45. Hadjur, S., Williams, L.M., Ryan, N.K., Cobb, B.S., Sexton, T., Fraser, P., Fisher, A.G. and Merkenschlager, M. (2009) Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. *Nature*, **460**, 410–413.
  46. Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C. *et al.* (2006) Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.*, **16**, 1299–1309.
  47. Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B. and de Laat, W. (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.*, **38**, 1348–1354.

48. Bau,D., Sanyal,A., Lajoie,B.R., Capriotti,E., Byron,M., Lawrence,J.B., Dekker,J. and Marti-Renom,M.A. (2011) The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules. *Nat. Struct. Mol. Biol.*, **18**, 107–114.
49. Ragoczy,T., Bender,M.A., Telling,A., Byron,R. and Groudine,M. (2006) The locus control region is required for association of the murine beta-globin locus with engaged transcription factories during erythroid maturation. *Genes Dev.*, **20**, 1447–1457.
50. Song,S.H., Kim,A., Ragoczy,T., Bender,M.A., Groudine,M. and Dean,A. (2010) Multiple functions of Ldb1 required for beta-globin activation during erythroid differentiation. *Blood*, **116**, 2356–2364.
51. Lee,H.Y., Johnson,K.D., Boyer,M.E. and Bresnick,E.H. (2011) Relocalizing genetic loci into specific subnuclear neighborhoods. *J. Biol. Chem.*, **286**, 18834–18844.
52. Hou,C., Zhao,H., Tanimoto,K. and Dean,A. (2008) CTCF-dependent enhancer-blocking by alternative chromatin loop formation. *Proc. Natl Acad. Sci. USA*, **105**, 20398–20403.