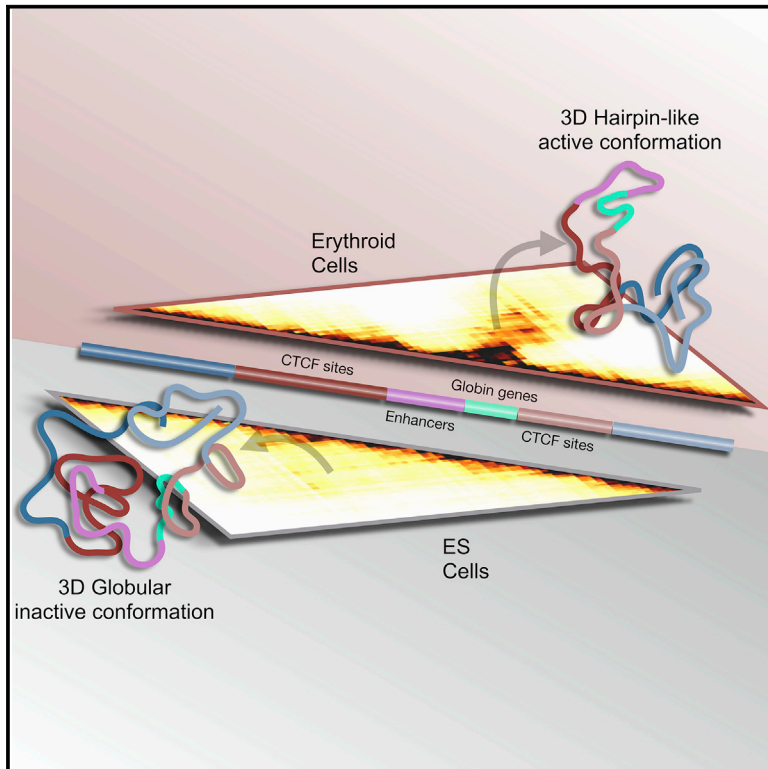


A Dynamic Folded Hairpin Conformation Is Associated with α -Globin Activation in Erythroid Cells

Graphical Abstract



Authors

Andrea M. Chiariello, Simona Bianco, A. Marieke Oudelaar, ..., Douglas R. Higgs, Jim R. Hughes, Mario Nicodemi

Correspondence

chiariello@na.infn.it (A.M.C.),
mario.nicodemi@na.infn.it (M.N.)

In Brief

Chiariello et al. use polymer physics models to infer the 3D conformations of the murine α -globin locus. In the transition from ESCs to erythroid cells, the locus rearranges from an inactive highly intermingled conformation to an active, hairpin-shaped structure, marked by cell-specific three-way contacts accurately predicted by the models.

Highlights

- The structure of the mouse α -globin locus is investigated with polymer physics models
- The conformation is highly intermingled when the globin genes are inactive in ESCs
- The locus is folded in a hairpin-shaped conformation when active in erythroid cells
- The model accurately predicts three-way contacts, as confirmed by TriC experimental data



A Dynamic Folded Hairpin Conformation Is Associated with α -Globin Activation in Erythroid Cells

Andrea M. Chiariello,^{1,7,*} Simona Bianco,^{1,7} A. Marieke Oudelaar,^{4,5} Andrea Esposito,^{1,2} Carlo Annunziatella,¹ Luca Fiorillo,¹ Mattia Conte,¹ Alfonso Corrado,¹ Antonella Prisco,³ Martin S.C. Larke,^{4,5} Jelena M. Telenius,^{4,5} Renato Sciarretta,¹ Francesco Musella,¹ Veronica J. Buckle,⁴ Douglas R. Higgs,⁴ Jim R. Hughes,^{4,5} and Mario Nicodemi^{1,6,8,*}

¹Dipartimento di Fisica, Università degli Studi di Napoli Federico II, and INFN Napoli, Complesso Universitario di Monte Sant'Angelo, 80126 Naples, Italy

²Berlin Institute for Medical Systems Biology at the Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Berlin, Germany

³CNR-IGB, Pietro Castellino 111, Naples, Italy

⁴MRC Molecular Haematology Unit, MRC Weatherall Institute of Molecular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, UK

⁵MRC WIMM Centre for Computational Biology, MRC Weatherall Institute of Molecular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, UK

⁶Berlin Institute of Health (BIH), MDC-Berlin, 13125 Berlin, Germany

⁷These authors contributed equally

⁸Lead Contact

*Correspondence: chiariello@na.infn.it (A.M.C.), mario.nicodemi@na.infn.it (M.N.)

<https://doi.org/10.1016/j.celrep.2020.01.044>

SUMMARY

We investigate the three-dimensional (3D) conformations of the α -globin locus at the single-allele level in murine embryonic stem cells (ESCs) and erythroid cells, combining polymer physics models and high-resolution Capture-C data. Model predictions are validated against independent fluorescence in situ hybridization (FISH) data measuring pairwise distances, and Tri-C data identifying three-way contacts. The architecture is rearranged during the transition from ESCs to erythroid cells, associated with the activation of the globin genes. We find that in ESCs, the spatial organization conforms to a highly intermingled 3D structure involving non-specific contacts, whereas in erythroid cells the α -globin genes and their enhancers form a self-contained domain, arranged in a folded hairpin conformation, separated from intermingling flanking regions by a thermodynamic mechanism of micro-phase separation. The flanking regions are rich in convergent CTCF sites, which only marginally participate in the erythroid-specific gene-enhancer contacts, suggesting that beyond the interaction of CTCF sites, multiple molecular mechanisms cooperate to form an interacting domain.

INTRODUCTION

During development, gene activity is controlled by *cis*-regulatory elements, such as promoters and distal enhancers, that physi-

cally contact their target genes in a dynamic architecture (Dekker and Mirny, 2016; Spielmann et al., 2018). Chromatin has a complex network of interactions at different scales, including TADs (Dixon et al., 2012; Nora et al., 2012) and other structures (Dekker and Mirny, 2016; Fraser et al., 2015; Phillips-Cremins et al., 2013; Rao et al., 2014). Importantly, genomic contacts play a key role in regulating transcription by modulating the associations between genes and regulators. Disruption of such interactions may be associated with abnormal gene expression and development (Franke et al., 2016; Lupiáñez et al., 2015; Valton and Dekker, 2016).

To investigate the physical interactions between genes, promoters, enhancers, and boundary elements during cell differentiation and gene activation, we examined the globin loci, which have been biologically well characterized (see, e.g., Hay et al., 2016; Oudelaar et al., 2018). Here, we used chromatin models from polymer physics (Barbieri et al., 2012; Bianco et al., 2018; Chiariello et al., 2016), informed with recent high-resolution Capture-C data (Oudelaar et al., 2018), to obtain spatial information at the level of single alleles.

Our investigation shows that in embryonic stem cells (ESCs) the silent globin genes are embedded in a large, compact chromatin domain characterized by abundant spurious contacts associated with broad intermingling with its flanking regions. In erythroid cells, upon activation of the genes, the three-dimensional (3D) structure of the locus undergoes a significant change whereby the α -globin genes form a separated domain, with very specific contacts with their enhancers, in a more open conformation having much less intermingling with the flanking regions. In our model, derived contact maps are in good agreement with pairwise contact data from Capture-C and previous models of the locus based on a-priori assumptions about the epigenetic factors involved in the architecture (Brackley et al., 2016). In



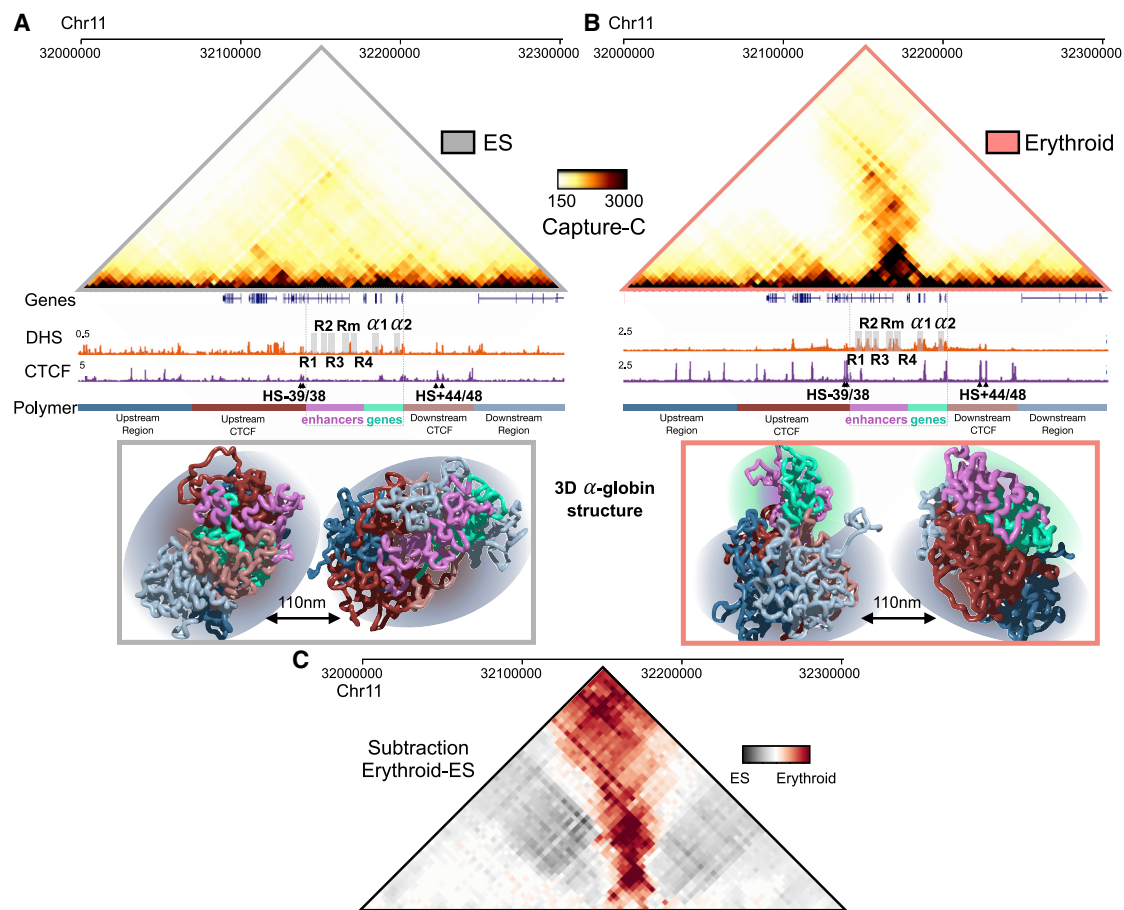


Figure 1. Single-Allele 3D Conformations of the α -Globin Locus Are Highly Tissue Specific

(A) 4-Kb resolution Capture-C data from Oudelaar et al. (2018) of a 300-kb region (chr11: 32,000,000–32,300,000, mm9) around the α -genes in ESCs. Gene annotation, DNaseI hypersensitive sites (DHS), and CTCF are shown below. The gray bars indicate the $\alpha 1$ and $\alpha 2$ gene promoters and the R1, R2, R3, Rm, and R4 enhancers. The dashed box highlights the enhancer-promoter region. The linear color scheme used for the polymer structures is also shown. The SBS polymer model corresponding contact maps are shown in Figure S1. Examples of 3D conformations from the SBS model are reported in the gray box below. The average size of the colored sub-region results is approximately 110 nm.

(B) The same locus in erythroid cells.

(C) The normalized Capture-C subtraction matrix (erythroid-ESC) highlights the structure of the rearrangements occurring in the locus.

addition, we find that the α -globin gene domain forms a novel architectural pattern: a hairpin structure folded on itself. The 3D conformations derived with the model are tested against independent, recently published fluorescence in situ hybridization (FISH) data (Brown et al., 2018) measuring pairwise distances, together with Tri-C data (Oudelaar et al., 2018) measuring triple contacts formed between the genes and their regulatory regions, thus extending the traditional focus on pairwise interactions of computational studies (Baù et al., 2011; Brackley et al., 2016; Gürsoy et al., 2017). The structural rearrangements are meaningfully linked to epigenetic changes in the locus, suggesting that several factors act cooperatively to orchestrate the reorganization necessary for α -globin regulation, beyond the CTCF/cohesin paradigm. In our polymer physics approach, the folding of the α -globin locus is originated by the thermodynamic mechanism of micro-phase separation, resulting from the interaction of chromatin with its binders (Barbieri et al., 2012; Chiariello et al., 2016).

Overall, the 3D conformations derived by our study are compatible with previous experimental observations (Anguita et al., 2004; Davies et al., 2016; Hay et al., 2016; Hughes et al., 2014) and recapitulate them under a unique, simple structural rationale.

RESULTS

We studied the α - and β -globin loci in mouse ESCs and in primary erythroid cells, where the globin genes are respectively silent and active. The mouse α -globin cluster and its five enhancer elements (named R1, R2, R3, Rm, and R4) extend over a ~50-kb region (Figures 1A, 1B, and S1), which is flanked by CTCF-binding sites (Hay et al., 2016; Oudelaar et al., 2018). An analogous organization characterizes the β -globin locus, with the gene cluster and its enhancers (named LCR HS1–6; Figure S1; STAR Methods) flanked by CTCF-binding sites.

The 3D Structure of the Globin Locus Undergoes a Deep Change during Differentiation

To investigate the locus 3D structure in detail, we first considered a genomic region spanning 300 kb around the α -globin genes (chr11: 32000000–32300000, mm9 mouse genome) by using high-quality Capture-C data (Oudelaar et al., 2018) for ESCs (Figure 1A) and erythroid cells (Figure 1B) at 4-kb resolution. Comparing the two cell types, we find that the locus exhibits a highly tissue-specific organization. Specifically, in ESCs the locus is organized into a large uniform domain, where no preferential contact forms between its regulatory elements, whereas in erythroid cells sharper interaction domains form, one of which contains the globin genes and their enhancers. Such architectural reorganization is also highlighted by the erythroid-ESC subtraction matrix (Figure 1C; STAR Methods), where a striking pattern of contact differences is observed, stretched along the anti-diagonal direction of the matrix and involving the regions flanking the α -globin genes at the level of the whole locus.

To quantitatively explore such structural rearrangements and to understand their corresponding 3D structure, we used the Strings & Binders Switch (SBS) polymer model of chromatin, which has been previously shown to describe with good accuracy Hi-C and GAM (genome architecture mapping) contact data genome wide (Barbieri et al., 2012; Chiariello et al., 2016). The SBS model describes the physical scenario in which contacts between distal DNA-binding sites are mediated by bridging molecules, such as transcription factors (TFs). The binding sites of the polymer model specific to the α - (and β -) globin loci in ESCs and erythroid cells are identified using the PRISMR (polymer-based recursive statistical inference method) method (Bianco et al., 2018), a machine learning approach based only on the above Capture-C data, without prior knowledge of TFs involved (STAR Methods; Tables S1 and S2). Next, by massive molecular dynamics (MD) simulations from only polymer physics, we derived with unprecedented detail the thermodynamic ensemble of 3D conformations of the single allele locus in the two cell types (Figures S1A and S1B). From such structures, we observe that in ESC distal regions in the locus have a high degree of intermingling (Figure 1A, gray box), whereas erythroid cells have more discrete interaction domains (Figure 1B, red box), in agreement with the Capture-C data (Figure S1C). As an estimate of the length scales involved, we find that the average size of these regions is approximately 110 nm (Figure S1C). Interestingly, similar rearrangements in the α -globin structure were also found in human GM12878 and K562 cells, although with a different computational approach applied on 5C data (Gürsoy et al., 2017).

To test the model accurately, we compared its contact maps, derived from the predicted 3D conformations, with Capture-C data. To take into account for population variability, the model envisages the locus in either open or compact thermodynamics states (Barbieri et al., 2012; Chiariello et al., 2016). Therefore, we can consider the optimal mixture of maps associated with open and compact conformations, best describing the experimental contact data (STAR Methods). With this approach, we find a Pearson correlation coefficient of $r = 0.96$ in the two cell types and an average distance-corrected Pearson correlation of $r' = 0.89$ (Figure S1A; STAR Methods). As a further check, the HiCRep method (Yang et al., 2017), specific to Hi-C data, returns

an average coefficient SCC (stratum-adjusted correlation coefficient) = 0.84, well above random controls ($p = 0$; STAR Methods). The contact map derived from only compact states also gives high coefficients, yet they are lower than the mixture model, with average values of $r = 88$ and $r' = 0.8$.

Importantly, from both Capture-C data and the 3D structures of erythroid cells, it appears that the globin flanking CTCF sites do not participate in the specific enhancer-promoter interactions but tend to interact and intermingle with other CTCF sites lying at the borders of the enhancer-promoter domain (Videos S1 and S2). Quantitative analysis performed on our polymer ensemble supports this hypothesis, whereas the upstream and downstream CTCF regions tend to intermingle (see below for the formal definition of intermingling) between themselves; they do not intermingle with the enhancer and promoter region (Mann-Whitney U test, $p = 10^{-22}$; Figure S1D). A similar structural reorganization from ESCs to erythroid cells is also observed from the ensemble of 3D structures obtained from high-resolution Capture-C data performed on the β -globin locus (chr7: 110840000–111140000, mm9; Figures S1E and S1F), although with a less complex interaction pattern than the α -globin locus in erythroid cells.

Next, to support our results based on Capture-C data and to investigate the structural conformation of the globin loci at larger genomic scales, we considered Hi-C data at a 20-kb resolution in ESCs (Giorgetti et al., 2016) and primary erythroid cells (Oudelaar et al., 2018) for a 3.3-Mb-long genomic region containing the α -globin genes (chr11: 29900000–32200000). As for the previously discussed case, the extended locus undergoes a similar dramatic rearrangement with a more pronounced TAD-like structure surrounding the α -globin genes in erythroid cells (Oudelaar et al., 2018; Figure S2A). To describe the structural modifications accompanying changes in gene expression, using the same modeling approach we derived the ensemble of single-allele 3D conformations in the two cell types. To test the model-derived contact maps inferred from the predicted 3D conformations, as before we compared them with Hi-C data, finding that they accurately match the structural features in both cell types (Figure S2B), having a Pearson correlation coefficient above $r = 0.94$ and an average SCC = 0.65 from the HiCRep method (STAR Methods). The derived 3D structures show that in ESCs chromatin is characterized by a higher degree of intermingling (see below), with distal regions broadly touching each other (Figure S2C, left), whereas in erythroid cells it exhibits a more compartmentalized organization, in particular regarding the region containing the globin genes (Figure S2C, right). With a completely analogous approach, we modeled a 3.3-Mb region containing the β -globin locus (chr7: 109300000–112600000, mm9) and found very similar results (Figures S2D–S2F).

To quantify all the structural properties discussed above and to validate the model, we first compared our results with recently published FISH data (Brown et al., 2018). Specifically, we considered the distances between the enhancer-promoter region (here named the α -domain) and its two flanking regions F1 and F2 (Figure 2A). In agreement with experiments, we find that the two regions F1 and F2 come into closer spatial proximity in the transition from ESCs to erythroid cells (Kolmogoroff-Smirnov test, $p > 0.01$), with a clear shift on the left of the whole distance distribution (model average distance in ESCs,

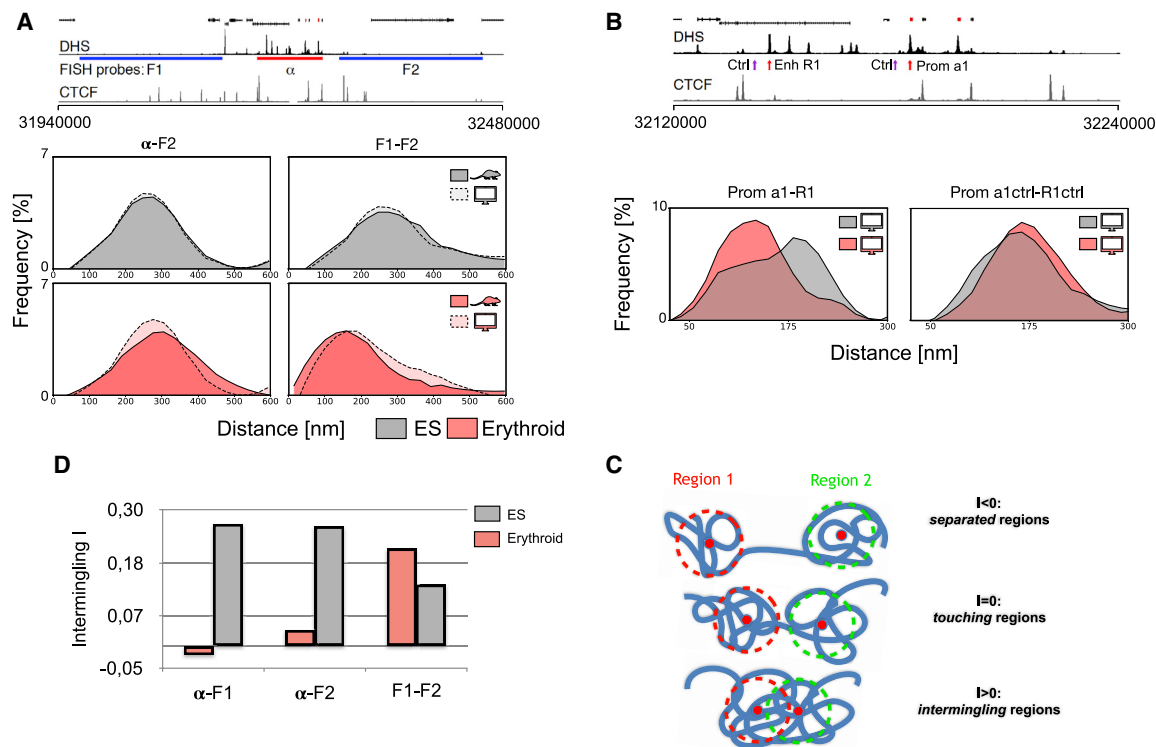


Figure 2. The α -Globin Genes Are Highly Intermingled with Their Flanking Regions in ESC, but Not in Erythroid Cells Where the Genes Are Activated

(A) The distance distribution of the locus from our polymer model well matches independent FISH data from [Brown et al. \(2018\)](#) (Kolmogoroff-Smirnov test, $p > 0.01$). In the transition from ESCs to erythroid cells, the F1 and F2 regions, flanking the α region (containing the α -globin genes and their enhancers), come in closer spatial proximity. In ESCs, the mean distance of F1–F2 is 328 ± 144 nm (FISH, 308 ± 112 nm) and of α -F2 is 276 ± 90 nm (FISH, 275 ± 92 nm); in erythroid cells, the values are 228 ± 99 nm (FISH, 201 ± 119) and 286 ± 85 nm (FISH, 301 ± 98).

(B) Distribution of distances between the promoter and the enhancers in ESCs and erythroid cells. As a control, the distribution between two non-regulatory, genetically equally spaced sites is also given.

(C) An illustration of our measure of intermingling, I , between pairs of genomic regions.

(D) The SBS-model-derived median value of the intermingling, I , shows that the α region is highly intermingled with its flanking regions F1 and F2 in ESCs, whereas it is not in erythroid cells. In all cases, the intermingling distributions (shown in [Figure S3A](#)) are statistically different between ESCs and erythroid cells ($p < 10^{-4}$, Wilcoxon rank-sum test).

328 ± 144 nm; in erythroid cells, 228 ± 99 nm). Furthermore, in erythroid cells, we find that more specific enhancer-promoter contacts occur than in the ESC case, consistent with the change in activity of the genes ([Figure 2B](#)).

Next, to dissect the high-order architectural relationship between the α -domain and the flanking regions, we quantified the degree of intermingling (I) between pairs of genomic regions, by measuring their overall spatial overlap across the model derived single-allele 3D conformations (STAR Methods). When I is negative ($I < 0$), the two regions are well separated compared to their own size; if I is approximately zero ($I \approx 0$), they are roughly touching each other; finally, if $I > 0$, the two regions are intermingling ([Figure 2C](#)). From our 3D polymer conformations, we computed the median degree of I of the polymer stretches overlapping with F1, F2, and α -domain of the FISH probes of [Figure 2A](#). We find that in ESCs the α -domain is highly intermingled with the flanking F1 and F2 regions, and interestingly F1 and F2 are also intermingled, but to a lesser extent, consistent with the genomic proximity constraints ([Figure 2D](#), gray bars). A sharp

change occurs in erythroid cells, where we find that the level of intermingling between the α -domain and both flanking F1 and F2 regions sharply drops to $I \approx 0$, despite their physical separation not changing dramatically between ESCs and erythroid cells (relative variation of the mean $\sim 4\%$; see [Figure 2A](#)). Conversely, the intermingling between F1 and F2 significantly increases with respect to ESCs ([Figure 2D](#), red bars). Such results, thus, provide a quantitative structural interpretation of Capture-C and microscopy experiments ([Brown et al., 2018](#)). The full intermingling distribution is reported in [Figure S3A](#) and reveals a broad level of structural variability of the simulated polymer 3D structures, in line with previous computational studies about the α -globin structure ([Brackley et al., 2016](#); [Gürsoy et al., 2017](#)) and with recent experimental results in single-cell variability ([Bintu et al., 2018](#)). Furthermore, within single dynamics, the intermingling index results only weakly fluctuate around its equilibrium value ([Figure S3B](#)), hinting a global stability of the structure over time-scales, if mapped onto physical units, roughly estimated around tens of minutes or hour (STAR Methods).

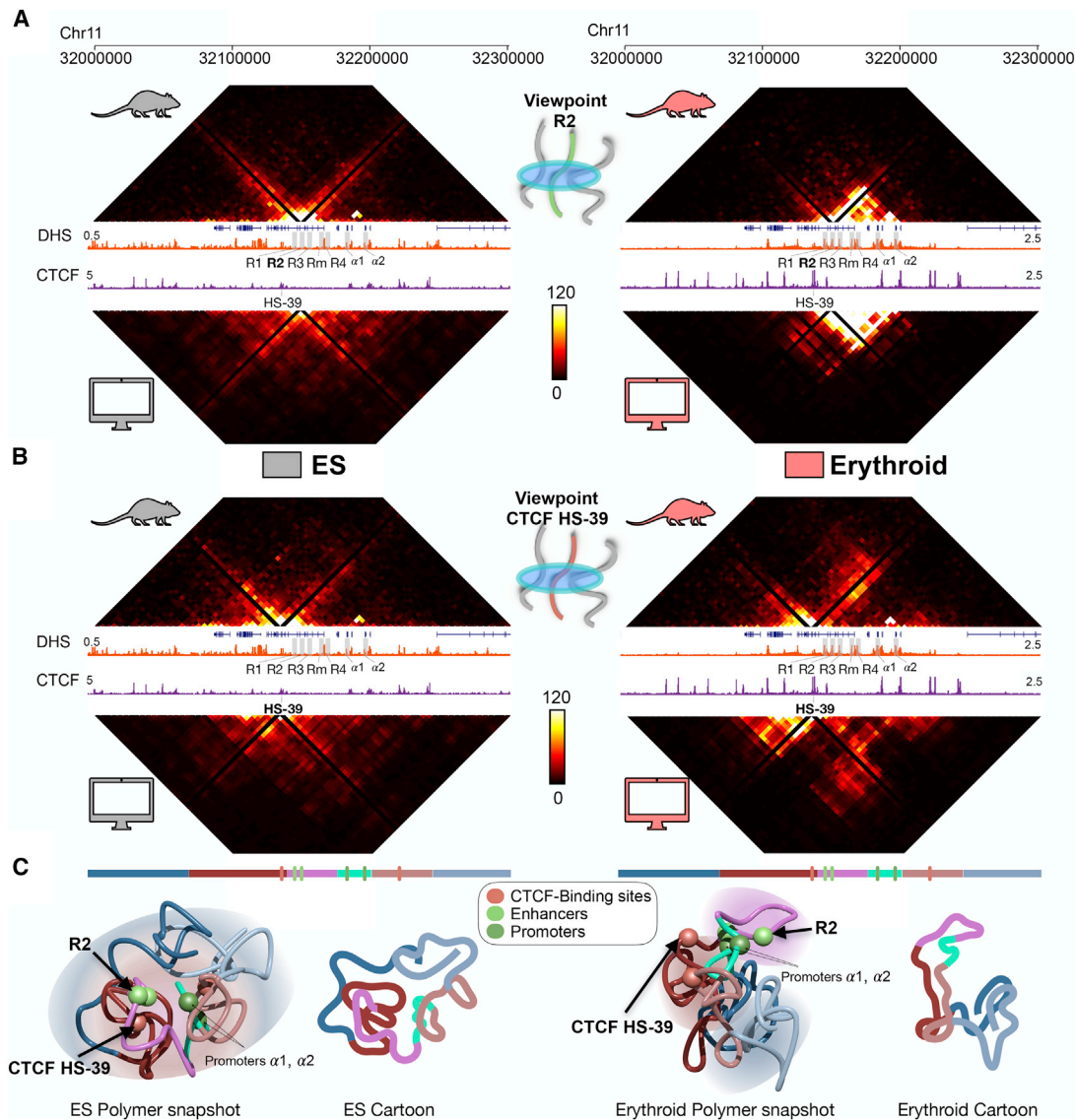


Figure 3. Model Predictions on Triple Contacts in the α -Globin Locus Well Match Independent Tri-C Data

(A) Comparison of triple contacts with the alpha enhancer R2 observed in Tri-C experiments (top matrices, from Oudelaar et al., 2018) and in our polymer model (bottom matrices) in ESC (left) and erythroid cells (right) (chr11: 32,000,000–32,300,000, mm9). The gray bars and labels highlight the position of enhancers and the promoters. In bold the viewpoint is highlighted.

(B) Analogous comparison from the CTCF-39 viewpoint.

(C) Examples of single conformations of coarse-grained 3D polymer snapshots highlight the different positions, and multi-way contacts, of genes and regulators. Note how, in the erythroid structure, the enhancer region (colored in purple) tends to interact with the gene region (colored in cyan), and the other similarly colored parts of the polymer, symmetric to the gene-enhancer region, tend to interact between themselves, resulting in a more elongated conformation with respect to the ESC case. The schematic two-dimensional (2D) cartoons highlight such peculiar conformations.

The Architecture Is Marked by Multi-way Contacts in Mouse Erythroid Cells

Next, we investigated the complexity of the high-order structure of the locus, beyond the pairwise interactions. We identified multiple contacts and, in particular, the probability of interactions between triplets of sites across the locus by using the 3D conformations predicted by our polymer model (which is based only on pairwise Capture-C data). First, we focused on triplets formed by the R2 enhancer (the strongest α -globin enhancer; Hay et al.,

2016) in ESCs and erythroid cells, represented as matrices of 3-way contacts from its viewpoint (Figure 3A). To validate our *in silico* findings, we compared them with independent Tri-C experimental data (Oudelaar et al., 2018) (at 4-kb resolution) and found that the model and Tri-C triplets have a Pearson correlation coefficient of $r = 0.8$ for both ESCs and erythroid cells (a random control case has $r \approx 0.15$; see STAR Methods). In agreement with experimental observations, we find that in ESCs there are no specific multi-way interactions of R2 with

other sites in the locus (Figure 3A, left matrices), beyond those linked to proximity effects. That finding further clarifies the nature of the highly intermingled structure of the locus in ESCs, where no major multi-way regulatory contacts occur. Conversely in erythroid cells R2 exhibits enriched three-way contacts within the DNA region spanning the R1, R3, Rm, and R4 enhancers and the $\alpha 1$ and $\alpha 2$ promoters (Figure 3A, right matrices), consistent with the previous finding that such a domain is separated and characterized by a low level of intermingling with its flanking regions (F1 and F2). Indeed, in erythroid cells, we estimated that more than 80% of the most prominent 3-way contacts of R2 (STAR Methods), having an average distance of approximately 100 nm, are triplets made within the enhancer-promoter region, whereas in ESCs, this fraction is roughly halved and originated by proximity effects.

Next, we focused on triplets from the CTCF HS-39 site viewpoint (Figure 3B). As in the R2 case, the pattern of multi-way interactions in ESCs does not show any preferential enrichment (Figure 3B, left matrices). In erythroid cells, a diffuse interaction with the region downstream of the enhancer-promoter α -domain is observed, and yet, no specific strong interactions are observed (Figure 3B, right matrices). Again, the model accurately recapitulates the experimental data (Pearson $r = 0.84$ and $r = 0.77$ for ESCs and erythroid cells, respectively; $r \approx 0.17$ for the control case; STAR Methods). As a further check, we found that HiCRep returns similar results, with an average value of $SCC = 0.68$, again well above the random control ($p = 0.0$; STAR Methods).

The predicted and experimentally observed specificity of the multiple contacts in the murine α -globin correlates with the different activation state of the locus, which was also found in the human α -globin locus (Gürsoy et al., 2017). The coarse-grained 3D structures in the snapshots of Figure 3C, derived from our MD simulations, provide an effective visual summary of the results so far described on the α -globin architecture in ESCs (left) and erythroid cells (right). They represent a typical example from the ensemble of folded structures derived with our simulations.

The 3D Structural Rearrangements during Differentiation Associate with Changes in Chromatin Marks

Our machine-learning (PRISMR)-derived SBS polymer model of the α -globin locus envisages that five main different types of binding sites contribute to the architecture by interactions with their cognate binders, thus determining its spatial structure (Figure 4). Each type of binding site along the polymer accumulates at specific regions and, yet, typically also has a broad distribution along the locus. Therefore, different colors tend to segregate in distinct clusters, according to a mechanisms known as micro-phase separation (Brackley et al., 2013). Those distinct clusters have partial overlaps with each other due to the genomically overlapping binding sites that induce long-range contacts. To characterize the origin of pairwise interactions (Figure 4A), we analyzed the contribution of each putative binding site to the contact pattern, i.e., the fraction of a pair interaction caused by a given type of binding site (STAR Methods). In this way, we can identify the most contributing type to each contact in the

map and produce a matrix highlighting only the most contributing binding type to each pairwise interaction, i.e., a matrix where each pixel represents the most important type to the correspondent contact (Figure 4A, bottom matrices). Notably, in the α -globin locus, we find that in ESCs $\sim 40\%$ of the total contacts are dominated by one single type of binding site, as shown by the most contributing binding site matrix (Figure 4A, left bottom matrix, with a dominating blue triangle in the middle) and the main binding sites tracks (Figure 4B, left), ranked according to their contribution to the contacts of the overall locus (Table S1). Conversely, the erythroid pairwise interactions are associated with a more diverse set of sites and with different types of binding sites, which form the inner active α -domain and the flanking structures (Figure 4A, right bottom matrix). Interestingly, we observe that those sites seem to locate in a mirror symmetric configuration around the α -genes (see next paragraph) (Figure 4B, right). Next, to characterize the molecular nature of the putative model binding sites, we considered our PolII and H3K27me3 datasets and several publicly available epigenomic features at this locus (Figure 4C; STAR Methods) and investigated the correspondence between those epigenetic tracks and the location of the binding sites. To this aim, we performed a cross-correlation analysis by calculating the Pearson coefficient between the genomic track of each binding site type with each analyzed chromatin mark (Bianco et al., 2018; STAR Methods). Our correlation analysis highlights that each different type of binder (i.e., a “color”) envisaged by our model generally does not represent a single distinct factor associated to chromatin, but rather combinations of molecular factors, including CTCF and other factors. Specifically, in ESCs, we find a simple pattern of correlation where the dominating binding site type (the blue track in Figure 4B) correlates with CTCF and cohesin (Figure 4D, left panel), the second most contributing correlates with open chromatin sites and H3K4me3, and the others generally not showing any significant correlation (STAR Methods). Conversely, in erythroid cells, more complex correlations are found, where the binding site types associated with the inner α -globin domain correlate with histone marks of active transcription together with nascent RNA and RNA polymerase II (Pol II) (consistent with the activation of the genes), whereas those in its flanking regions with a combination of CTCF/cohesin and other marks (Figure 4D, right panel) support the scenario where the structural organization emerges from a combinatorial action of different molecular factors (Hnisz et al., 2017; Oudelaar et al., 2018; Pereira et al., 2018), including but not limited to cohesin-mediated interactions between CTCF sites (Buckle et al., 2018; Pereira et al., 2018). Analogous results are obtained from the correlation analysis between binding sites and chromatin marks in the β -globin locus, where a dominant binding site type is identified in ESCs ($\sim 35\%$ of contacts) and more binding sites cooperatively contribute to the more complex structure in erythroid cells (Figure S4; Table S2).

The Globin Locus Forms a Folded Hairpin Structure in Erythroid Cells

Finally, we investigated the structural nature of the specific pattern of interactions observed in the α -globin locus in erythroid cells. Visual inspection of Capture-C data (reported again in

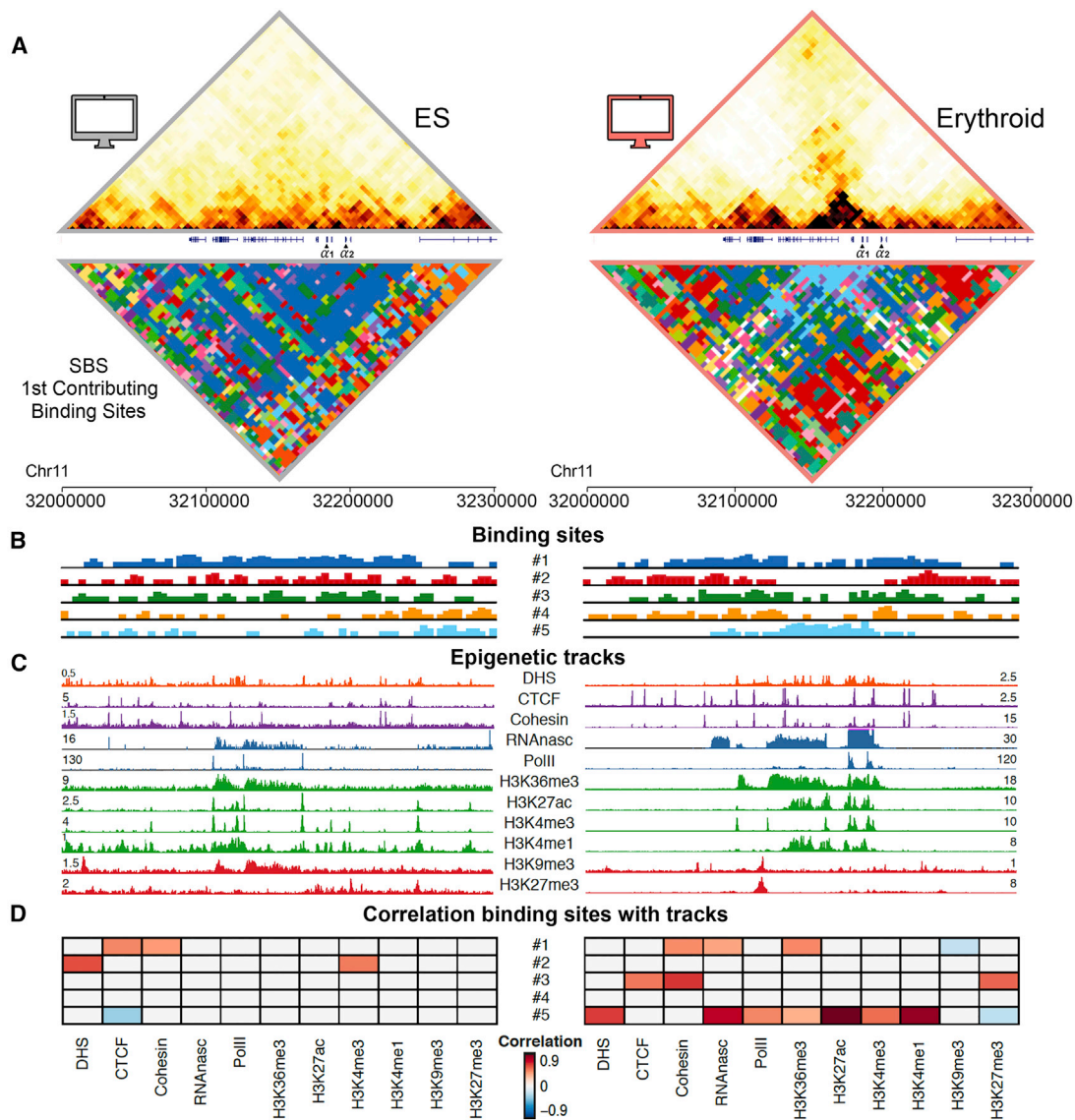


Figure 4. The Binding Domains Underlying the 3D Structure of the α -Globin Locus Correlate with Cell-Type-Specific Epigenetic Features (A) The locus contact map obtained from our 3D polymer model (top matrices) have high correlations with Capture-C (in ESCs, Pearson $r = 0.96$ and distance-corrected Pearson $r' = 0.87$; in erythroid cells, $r = 0.96$ and $r' = 0.91$; the HiCRep coefficient in ESCs is $SCC = 0.75$ and in erythroid is $SCC = 0.92$). On the bottom, the most contributing binding site type to each contact is shown. (B) Top five most contributing binding sites to the locus interactions, sorted by their contribution to the contact map of the overall locus. In ESCs, the first most contributing binding type accounts for about 40% of total contacts, whereas in erythroid cells, 3D contacts are associated with a more complex set of binding site types. In particular, domain #5 localizes in the globin domain and the other ones surround it in a more spread, mirror-like symmetric configuration. (C) Epigenetic features along the α -globin locus. DHS, DNase hypersensitive sites; RNAnasc, nascent RNA expression; Pol II, RNA polymerase II occupancy. (D) Pearson correlations of the binding domains with epigenetic tracks show that the main binding domain in ESCs is correlated with CTCF and cohesin. In erythroid cells, the main binding sites also correlate with chromatin marks of active transcription.

Figure 5A at 4-kb resolution and in Figure S5A in coarse-grained 8-kb version to highlight the long-range contacts) shows that the structure is not compatible with a simple TAD (Dixon et al., 2012) or a loop domain (Rao et al., 2014), which would be associated respectively with a plain “square” or “dot” in the contact map. Rather, the experimental data are much closer to an anti-diagonal structure (Figure 5B), as also confirmed by the subtraction matrix (Figures 1C and S1B) between ESCs and erythroid cells.

Supported by the intermingling analysis (Figures 2D and S3A), we reasoned that such a pattern is architecturally consistent with a hairpin-like organization, where the polymer folds around a specific region in such a way to ensure spatial proximity between symmetric (or roughly symmetric) flanking regions (Figure 5B). The 3D structures derived from the SBS model trained on the Capture-C data support the hypothesis that a hairpin folded on itself is formed, as shown by the position of some

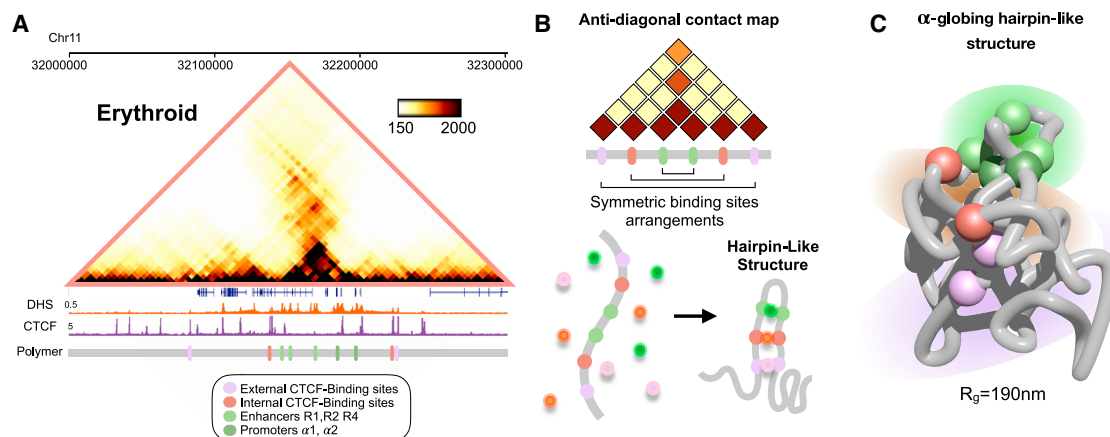


Figure 5. The α -Globin Gene Region in Erythroid Cells Forms a Folded Hairpin-like Structure

(A) α -Globin Capture-C data in erythroid cells (same data as Figure 1B) have an anti-diagonal pattern not compatible with a simple TAD or loop domain. In the linear polymer representation, regulatory elements in a symmetric mirror-like position are highlighted.

(B) Schematically, an anti-diagonal pattern of interactions can be explained with a mirror symmetric arrangement of binding sites, as found in the locus model (Figure 4B). The resulting 3D structure is a folded hairpin-like conformation.

(C) Example of single 3D conformation derived from the model trained on the Capture-C data exhibits a hairpin-like shape, as shown by the position of some key regulatory elements. Promoters and enhancers are represented in green, and CTCF binding sites are represented in red and thistle. To give a sense of the length scales involved in the hairpin structure, we report the estimated gyration radius R_g .

enhancers, promoters, and CTCF binding elements (Figure 5C). The ensemble of 3D conformations is characterized by an intrinsic degree of heterogeneity, and the details of the hairpin-like conformation do vary from structure to structure, as highlighted by the broad distributions of the gyration radii and intermingling index shown in Figures S1C and S3A. Such heterogeneity derives from the intricate folding of the polymer in the 3D space. Consequently, the basic hairpin structure, schematically represented in Figure 5B, can be more complicated in real polymer conformations, as it can fold and bend on itself. Importantly, we observe that in our simulations the hairpin conformation is dynamically stable because the intermingling index I typically exhibits in time small thermal fluctuations around its equilibrium value (Figure S3B). To highlight the correspondence with Capture-C data in the α -globin locus, we also built a simple, basic “toy” model based on the full SBS model, where we conveniently selected only the binding sites corresponding to genomic positions of enhancers, promoters, and CTCF sites (Figure S5; STAR Methods). Interestingly, such a toy model recapitulates the key pattern observed in the data (Figure S5A). The simpler 3D structure of the toy model clearly displays the hairpin-like organization (Figure S5B). To support this scenario, we compared the toy model with the more complex binding site arrangements previously found based only on the information contained in the Capture-C map (STAR Methods). Consistently, we find that the most contributing binding sites in the full model inferred from Capture-C data correspond to the three key binding sites of the toy model (Figure 5C).

DISCUSSION

In this work, we investigated the 3D organization of the α - and β -globin loci by combining high-resolution Capture-C data and

polymer physics modeling, in mouse ESCs where the globin genes are silent, and in erythroid cells where they are active. From experimental evidence, it is known that dramatic structural rearrangements occur in this transition, involving the genes and their enhancers. From polymer models constructed starting only from Capture-C data, we derived a thermodynamic ensemble of 3D conformations reflecting the structural properties of both cell types. Independent Hi-C data provided similar results. Our findings were tested against independent single-cell FISH data, showing that our thermodynamics ensemble of conformations also reproduces with good accuracy the distance distributions for the available experimental site pairs. From our polymer 3D structures, we also predicted the existence of high-order contacts, such as triplets formed by the enhancer R2, which were confirmed by independent Tri-C data.

The physical processes underlying the self-assembly of the locus envisaged by our polymer model are thermodynamic mechanisms of micro-phase separation, i.e., phase transitions deriving from the interactions between chromatin and its binders (Barbieri et al., 2012; Chiariello et al., 2016). Additional off-equilibrium mechanisms, such as the loop extrusion (Fudenberg et al., 2016; Sanborn et al., 2015) or the slip-link (Brackley et al., 2017) mechanism, could contribute to chromatin folding. Yet, our results based on the SBS model show that contact data at the α - and β -globin loci can be explained without invoking loop extrusion. Anyway, the SBS as well as loop extrusion are simplified models, as in real systems a variety of additional complex phenomena can occur, ranging from spontaneous segregation (Caglioti et al., 1998; Nicodemi et al., 2002, 2008), to jamming (Ciamarra et al., 2010, 2011; Grebenkov et al., 2008; Nicodemi and Coniglio, 1998), and stress anomalies (Coniglio et al., 1999; Nicodemi, 1998) to avalanche effects (Hammon et al., 2002). By introducing a quantitative measure of

intermingling, *I*, from our single-allele 3D structures, we found that in ESCs the 3D structures representing the 300 kb around the α -globin genes are characterized by a high level of intermingling, where the two globin gene promoters ($\alpha 1$ and $\alpha 2$) and their five enhancers (R1, R2, R3, Rm, and R4) weakly contact each other in a random and non-specific manner. Conversely, in erythroid cells, a much more compartmentalized structure is observed, with promoters and enhancers organized in a self-interacting domain, not intermingled with its flanking regions. Such a domain is flanked by regions rich in CTCF sites; yet, they do not participate in the specific gene-enhancer contacts (Oudelaar et al., 2018). In particular, we find that such a domain is arranged in a folded hairpin-like structure. Its flanking regions, instead, interact in a high-order structure, maintaining a high level of intermingling and diffuse contacts, which is in agreement with experimental data.

Although our findings clarify the 3D structure of the globin loci, they depict a scenario where multiple molecular mechanisms contribute to formation of the locus conformation, including but not limited to CTCF/cohesin-associated interactions (Schwarzer et al., 2017).

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
 - Primary erythroid cells
 - ES cells
- METHOD DETAILS
 - Strings and Binders Switch (SBS) polymer model
 - Fit of Capture-C and HiC data
 - Molecular Dynamics simulations details
 - Polymer models based on Hi-C data
 - Polymer models based on Capture-C data
 - Toy polymer model based on regulatory elements position
 - Contact pairwise and tri-wise matrices
 - Polymer graphics
 - Polymer intermingling
 - Structural analysis of the binding domains
 - Epigenetics features – Analysis
 - Epigenetics features – Correlation with predicted binding domains
 - Polymer distance distributions – Correlation with FISH data
 - Polymer distance distributions – Analyses of the dynamic globin loci
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND CODE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at <https://doi.org/10.1016/j.celrep.2020.01.044>.

ACKNOWLEDGMENTS

M.N. acknowledges grants from EU H2020 Marie Curie (813282), National Institutes of Health (NIH) (1U54DK107977-01), CINECA ISCRA (HP10CPTY8P), the Einstein BIH Fellowship Award (EVF-BIH-2016-282), Regione Campania SATIN Project 2018-2020, and computer resources from the Istituto Nazionale di Fisica Nucleare, CINECA, ENEA CRESCO/ENEAGRID (Ponti et al., 2014) and SCoPE/ReCaS at the University of Naples. D.R.H., J.H., and V.J.B. are supported by grants from the Medical Research Council (MC_UU_00016 and MR/N00969X/1).

AUTHOR CONTRIBUTIONS

M.N. and A.M.C. designed the project. A.M.C., S.B., C.A., and M.N. developed the modeling part; A.M.C., S.B., A.E., M.C., L.F., A.C., R.S., and F.M. ran the computer simulations and performed their analyses; J.M.T. performed bioinformatic analysis; M.S.C.L. generated the new epigenetics datasets; J.H., V.J.B., and D.R.H. provided conceptual advice; and A.M.C., M.N., S.B., A.M.O., D.R.H., and A.P. wrote the manuscript.

Received: March 16, 2019

Revised: August 13, 2019

Accepted: January 14, 2020

Published: February 18, 2020

REFERENCES

- Allen, M.P., and Tildesley, D.J. (1989). *Computer Simulation of Liquids* (Oxford Science Publications).
- Anguita, E., Hughes, J., Heyworth, C., Blobel, G.A., Wood, W.G., and Higgs, D.R. (2004). Globin gene activation during haemopoiesis is driven by protein complexes nucleated by GATA-1 and GATA-2. *EMBO J.* 23, 2841–2852.
- Annunziatella, C., Chiariello, A.M., Bianco, S., and Nicodemi, M. (2016). Polymer models of the hierarchical folding of the Hox-B chromosomal locus. *Phys. Rev. E* 94, 042402.
- Annunziatella, C., Chiariello, A.M., Esposito, A., Bianco, S., Fiorillo, L., and Nicodemi, M. (2018). Molecular Dynamics simulations of the Strings and Binders Switch model of chromatin. *Methods* 142, 81–88.
- Barbieri, M., Chotalia, M., Fraser, J., Lavitas, L.M., Dostie, J., Pombo, A., and Nicodemi, M. (2012). Complexity of chromatin folding is captured by the strings and binders switch model. *Proc. Natl. Acad. Sci. USA* 109, 16173–16178.
- Barbieri, M., Xie, S.Q., Torlai Triglia, E., Chiariello, A.M., Bianco, S., de Santiago, I., Branco, M.R., Rueda, D., Nicodemi, M., and Pombo, A. (2017). Active and poised promoter states drive folding of the extended HoxB locus in mouse embryonic stem cells. *Nat. Struct. Mol. Biol.* 24, 515–524.
- Baù, 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 α -globin gene domain reveals formation of chromatin globules. *Nat. Struct. Mol. Biol.* 18, 107–114.
- Bianco, S., Lupiáñez, D.G., Chiariello, A.M., Annunziatella, C., Kraft, K., Schöpflin, R., Wittler, L., Andrey, G., Vingron, M., Pombo, A., et al. (2018). Polymer physics predicts the effects of structural variants on chromatin architecture. *Nat. Genet.* 50, 662–667.
- Bintu, B., Mateo, L.J., Su, J.H., Sinnott-Armstrong, N.A., Parker, M., Kinrot, S., Yamaya, K., Boettiger, A.N., and Zhuang, X. (2018). Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science* 362, eaau1783.
- Brackley, C.A., Taylor, S., Papantonis, A., Cook, P.R., and Marenduzzo, D. (2013). Nonspecific bridging-induced attraction drives clustering of DNA-binding proteins and genome organization. *Proc. Natl. Acad. Sci. USA* 110, E3605–E3611.
- Brackley, C.A., Brown, J.M., Waithe, D., Babbs, C., Davies, J., Hughes, J.R., Buckle, V.J., and Marenduzzo, D. (2016). Predicting the three-dimensional

- p>
folding of cis-regulatory regions in mammalian genomes using bioinformatic data and polymer models.
- Genome Biol.*
- 17, 59.
- Brackley, C.A., Johnson, J., Michieletto, D., Morozov, A.N., Nicodemi, M., Cook, P.R., and Marenduzzo, D. (2017). Nonequilibrium Chromosome Looping via Molecular Slip Links. *Phys. Rev. Lett.* 119, 138101.
- Brown, J.M., Roberts, N.A., Graham, B., Waithe, D., Lagerholm, C., Telenius, J.M., De Ornellas, S., Oudelaar, A.M., Scott, C., Szczerbal, I., et al. (2018). A tissue-specific self-interacting chromatin domain forms independently of enhancer-promoter interactions. *Nat. Commun.* 9, 3849.
- Buckle, A., Brackley, C.A., Boyle, S., Marenduzzo, D., and Gilbert, N. (2018). Polymer Simulations of Heteromorphic Chromatin Predict the 3D Folding of Complex Genomic Loci. *Mol. Cell* 72, 786–797.e11.
- Bulut-Karslioglu, A., De La Rosa-Velázquez, I.A., Ramirez, F., Barenboim, M., Onishi-Seebacher, M., Arand, J., Galán, C., Winter, G.E., Engist, B., Gerle, B., et al. (2014). Suv39h-dependent H3K9me3 marks intact retrotransposons and silences LINE elements in mouse embryonic stem cells. *Mol. Cell* 55, 277–290.
- Caglioti, E., Coniglio, A., Herrmann, H.J., Loreto, V., and Nicodemi, M. (1998). Segregation of granular mixtures in the presence of compaction. *Europhys. Lett.* 43, 591–597.
- Chiariello, A.M., Annunziatella, C., Bianco, S., Esposito, A., and Nicodemi, M. (2016). Polymer physics of chromosome large-scale 3D organisation. *Sci. Rep.* 6, 29775.
- Ciamarra, M.P., Nicodemi, M., and Coniglio, A. (2010). Recent results on the jamming phase diagram. *Soft Matter* 6, 2871–2874.
- Ciamarra, M.P., Pastore, R., Nicodemi, M., and Coniglio, A. (2011). Jamming phase diagram for frictional particles. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 84, 041308.
- Coniglio, A., de Candia, A., Fierro, A., and Nicodemi, M. (1999). Universality in glassy systems. *J. Phys. Condens. Matter* 11, A167–A174.
- Davies, J.O.J., Telenius, J.M., McGowan, S.J., Roberts, N.A., Taylor, S., Higgs, D.R., and Hughes, J.R. (2016). Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nat. Methods* 13, 74–80.
- De Gennes, P.G. (1979). *Scaling concepts in polymer physics* (Cornell University Press).
- Dekker, J., and Mirny, L. (2016). The 3D Genome as Moderator of Chromosomal Communication. *Cell* 164, 1110–1121.
- Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485, 376–380.
- Domcke, S., Bardet, A.F., Adrian Ginno, P., Hartl, D., Burger, L., and Schübeler, D. (2015). Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature* 528, 575–579.
- Franke, M., Ibrahim, D.M., Andrey, G., Schwarzer, W., Heinrich, V., Schöpf, R., Kraft, K., Kempfer, R., Jerković, I., Chan, W.L., et al. (2016). Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* 538, 265–269.
- Fraser, J., Ferrai, C., Chiariello, A.M., Schueler, M., Rito, T., Laudanno, G., Barbieri, M., Moore, B.L., Kraemer, D.C., Aitken, S., et al.; FANTOM Consortium (2015). Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Mol. Syst. Biol.* 11, 852.
- Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., and Mirny, L.A. (2016). Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep.* 15, 2038–2049.
- Giorgetti, L., Lajoie, B.R., Carter, A.C., Attia, M., Zhan, Y., Xu, J., Chen, C.J., Kaplan, N., Chang, H.Y., Heard, E., and Dekker, J. (2016). Structural organization of the inactive X chromosome in the mouse. *Nature* 535, 575–579.
- Grebenkov, D.S., Ciamarra, M.P., Nicodemi, M., and Coniglio, A. (2008). Flow, ordering, and jamming of sheared granular suspensions. *Phys. Rev. Lett.* 100, 078001.
- Gürsoy, G., Xu, Y., Kenter, A.L., and Liang, J. (2017). Computational construction of 3D chromatin ensembles and prediction of functional interactions of alpha-globin locus from 5C data. *Nucleic Acids Res.* 45, 11547–11558.
- Hamon, D., Nicodemi, M., and Jensen, H.J. (2002). Continuously driven OFC: A simple model of solar flare statistics. *Astron. Astrophys.* 387, 326–334.
- Hanssen, L.L.P., Kassouf, M.T., Oudelaar, A.M., Biggs, D., Preece, C., Downes, D.J., Gosden, M., Sharpe, J.A., Sloane-Stanley, J.A., Hughes, J.R., et al. (2017). Tissue-specific CTCF-cohesin-mediated chromatin architecture delimits enhancer interactions and function *in vivo*. *Nat. Cell Biol.* 19, 952–961.
- Hay, D., Hughes, J.R., Babbs, C., Davies, J.O.J., Graham, B.J., Hanssen, L., Kassouf, M.T., Marieke Oudelaar, A.M., Sharpe, J.A., Suci, M.C., et al. (2016). Genetic dissection of the α -globin super-enhancer *in vivo*. *Nat. Genet.* 48, 895–903.
- Hnisz, D., Shrinivas, K., Young, R.A., Chakraborty, A.K., and Sharp, P.A. (2017). A Phase Separation Model for Transcriptional Control. *Cell* 169, 13–23.
- Hughes, J.R., Roberts, N., McGowan, S., Hay, D., Giannoulou, E., Lynch, M., De Gobbi, M., Taylor, S., Gibbons, R., and Higgs, D.R. (2014). Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat. Genet.* 46, 205–212.
- Kowalczyk, M.S., Hughes, J.R., Garrick, D., Lynch, M.D., Sharpe, J.A., Sloane-Stanley, J.A., McGowan, S.J., De Gobbi, M., Hosseini, M., Vernimmen, D., et al. (2012). Intragenic enhancers act as alternative promoters. *Mol. Cell* 45, 447–458.
- Kremer, K., and Grest, G.S. (1990). Dynamics of entangled linear polymer melts: A molecular-dynamics simulation. *J. Chem. Phys.* 92, 5057–5086.
- Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* 161, 1012–1025.
- Nicodemi, M. (1998). Force correlations and arch formation in granular assemblies. *Phys. Rev. Lett.* 80, 1340–1343.
- Nicodemi, M., and Coniglio, A. (1998). Macroscopic glassy relaxations and microscopic motions in a frustrated lattice gas. *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics* 57, R39.
- Nicodemi, M., and Prisco, A. (2009). Thermodynamic pathways to genome spatial organization in the cell nucleus. *Biophys. J.* 96, 2168–2177.
- Nicodemi, M., Fierro, A., and Coniglio, A. (2002). Segregation in hard-sphere mixtures under gravity. An extension of Edwards approach with two thermodynamical parameters. *Europhys. Lett.* 60, 684–690.
- Nicodemi, M., Panning, B., and Prisco, A. (2008). A thermodynamic switch for chromosome colocalization. *Genetics* 179, 717–721.
- Nitzsche, A., Paszkowski-Rogacz, M., Matarese, F., Janssen-Megens, E.M., Hubner, N.C., Schulz, H., de Vries, I., Ding, L., Huebner, N., Mann, M., et al. (2011). RAD21 cooperates with pluripotency transcription factors in the maintenance of embryonic stem cell identity. *PLoS One* 6, e19470.
- Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Pilot, T., van Berkum, N.L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* 485, 381–385.
- Oudelaar, A.M., Davies, J.O.J., Hanssen, L.L.P., Telenius, J.M., Schwesinger, R., Liu, Y., Brown, J.M., Downes, D.J., Chiariello, A.M., Bianco, S., et al. (2018). Single-allele chromatin interactions identify regulatory hubs in dynamic compartmentalized domains. *Nat. Genet.* 50, 1744–1751.
- Pereira, M.C.F., Brackley, C.A., Michieletto, D., Annunziatella, C., Bianco, S., Chiariello, A.M., Nicodemi, M., and Marenduzzo, D. (2018). Complementary chromosome folding by transcription factors and cohesin. *bioRxiv*. <https://doi.org/10.1101/305359>.
- Phillips-Cremins, J.E., Sauria, M.E.G., Sanyal, A., Gerasimova, T.I., Lajoie, B.R., Bell, J.S.K., Ong, C.T., Hookway, T.A., Guo, C., Sun, Y., et al. (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* 153, 1281–1295.
- Plimpton, S. (1995). Fast parallel algorithms for short-range molecular dynamics. *J. Comp. Physiol.* 117, 1–19.
- Ponti, G., Palombi, F., Abate, D., Ambrosino, F., Aprea, G., Bastianelli, T., Beone, F., Bertini, R., Bracco, G., Caporicci, M., et al. (2014). The role of medium size facilities in the HPC ecosystem: the case of the new CRESO4

cluster integrated in the ENEAGRID infrastructure. In Proceedings of the 2014 International Conference on High Performance Computing & Simulation (IEEE), pp. 1030–1033.

Rahl, P.B., Lin, C.Y., Seila, A.C., Flynn, R.A., McCuine, S., Burge, C.B., Sharp, P.A., and Young, R.A. (2010). c-Myc regulates transcriptional pause release. *Cell* **141**, 432–445.

Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., and Aiden, E.L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680.

Rosa, A., and Everaers, R. (2008). Structure and dynamics of interphase chromosomes. *PLoS Comput. Biol.* **4**, e1000153.

Sanborn, A.L., Rao, S.S.P., Huang, S.C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J., et al. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci. USA* **112**, E6456–E6465.

Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N.A., Huber, W., Haering, C.H., Mirny, L., and Spitz, F. (2017). Two independent modes of chromatin organization revealed by cohesin removal. *Nature* **551**, 51–56.

Spielmann, M., Lupiáñez, D.G., and Mundlos, S. (2018). Structural variation in the 3D genome. *Nat. Rev. Genet.* **19**, 453–467.

Strikoudis, A., Lazaris, C., Trimarchi, T., Galvao Neto, A.L., Yang, Y., Ntziachristos, P., Rothbart, S., Buckley, S., Dolgalev, I., Stadtfeld, M., et al. (2016). Regulation of transcriptional elongation in pluripotency and cell differentiation by the PHD-finger protein Phf5a. *Nat. Cell Biol.* **18**, 1127–1138.

Valton, A.L., and Dekker, J. (2016). TAD disruption as oncogenic driver. *Curr. Opin. Genet. Dev.* **36**, 34–40.

Wamstad, J.A., Alexander, J.M., Truty, R.M., Shrikumar, A., Li, F., Eilertson, K.E., Ding, H., Wylie, J.N., Pico, A.R., Capra, J.A., et al. (2012). Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. *Cell* **151**, 206–220.

Yang, T., Zhang, F., Yardımcı, G.G., Song, F., Hardison, R.C., Noble, W.S., Yue, F., and Li, Q. (2017). HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Res.* **27**, 1939–1949.

Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B.D., et al.; Mouse ENCODE Consortium (2014). A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355–364.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Capture-C data in mouse ES and Erythroid	Oudelaar et al., 2018	GEO: GSE107940
Tri-C data in mouse ES and Erythroid	Oudelaar et al., 2018	GEO: GSE107940
HiC data in mouse ES	Giorgetti et al., 2016	GEO: GSE72697
HiC data in mouse erythroid	Oudelaar et al., 2018	GEO: GSE107940
DNase-seq in Erythroid	Kowalczyk et al., 2012	GEO: GSE27921
DNase-seq in ES	Domcke et al., 2015	GEO: GSE67867
CTCF Chip-Seq in Erythroid	Hanssen et al., 2017	GEO: GSE97871
CTCF Chip-Seq in ES	Nitzsche et al., 2011	GEO: GSE24030
Rad21 ChIP-seq in Erythroid	Hanssen et al., 2017	GEO: GSE97871
Rad21 ChIP-seq in ES	Nitzsche et al., 2011	GEO: GSE24030
Nascent transcription in Erythroid	Hay et al., 2016	GEO: GSE78835
Nascent transcription in ES	Strikoudis et al., 2016	GEO: GSE63974
Polymerase II Chip-seq in Erythroid	This paper	GEO: GSE107938
Polymerase II Chip-seq in ES	Rahl et al., 2010	GEO: GSE20485
H3K36me3 ChIP-seq in Erythroid	Yue et al., 2014	GEO: GSM946543
H3K36me3 ChIP-seq in ES	Yue et al., 2014	GEO: GSM1000109
H3K27ac, H3K4me3 and H3K4me1 ChIP-seq in Erythroid	Kowalczyk et al., 2012	GEO: GSE27921
H3K27ac ChIP-seq in ES	Wamstad et al., 2012	GEO: GSM1163096/7
H3K4me3 ChIP-seq in ES	Wamstad et al., 2012	GEO: GSM1163084/5/6
H3K4me1 ChIP-seq in ES	Wamstad et al., 2012	GEO: GSM1163115/6
H3K9me3 ChIP-seq in Erythroid	Yue et al., 2014	GEO: GSM946549
H3K9me3 ChIP-seq in ES	Bulut-Karslioglu et al., 2014	GEO: GSE57092
H3K27me3 ChIP-seq in Erythroid	Kowalczyk et al., 2012	GEO: GSE27921
H3K27me3 ChIP-seq in ES	This paper	GEO: GSE107937
FISH data in mouse ES and Erythroid	Brown et al., 2018	
Software and Algorithms		
LAMMPS	Plimpton, 1995	https://lammps.sandia.gov
Antibodies		
Pol II (N20 clone)	Santa Cruz Biotechnology	sc899c, lot #H3115
H3K27me3	Millipore (previously Upstate)	Cat# 07-449; RRID: AB_310624

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Mario Nicodemi (nicodem@na.infn.it).

This study did not generate new unique reagents.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Primary erythroid cells

Primary Ter119+ erythroid cells were obtained from spleens of female C57BL/6 mice (3-9 month old) treated with phenylhydrazine as previously described ([Davies et al., 2016](#)).

ES cells

Mouse ES cells (strain: 129/Ola) were cultured and harvested as previously described ([Davies et al., 2016](#)).

METHOD DETAILS

Strings and Binders Switch (SBS) polymer model

To investigate the spatial structures of the α - and β -globin loci, we used the Strings and Binders Switch (SBS) polymer model (Barbieri et al., 2012; Chiariello et al., 2016; Nicodemi and Prisco, 2009), in which a chromatin filament is modeled by a string made of N beads with binding sites that can interact with different, specific binding factors present in the surrounding environment. Each type of factor can interact only with the corresponding type of binding site on the polymer, with an interaction energy E_{int} and a total concentration c . As predicted by standard polymer thermodynamics (De Gennes, 1979), if these parameters are above threshold, the polymer undergoes the coil-globule phase transition which allow the polymer to collapse in a compact globular conformation (Barbieri et al., 2012).

Fit of Capture-C and HiC data

In order to find the minimal number of different types of binding sites and the positions of the binding sites along the polymer chain modeling the globin loci, we employed a previously described computational procedure (PRISMR method (Bianco et al., 2018)). PRISMR is a Machine Learning algorithm that takes as input only the corresponding experimental contact matrix (Hi-C or Capture-C) and returns the best SBS model describing the data. Briefly, it is based on the minimization of a cost function H , which includes a term H_0 (the distance between the input matrix and the model matrix) and a term H_λ that penalizes overfitting of binding site types (full details can be found in Bianco et al., 2018). At the end of the procedure, the fitted parameters are therefore the number of binding sites n and the distribution of the binding sites along the chain. From each matrix, a specific sequence of binding sites is found and describes the locus under consideration. To create an ensemble of thermodynamics equilibrium single molecule conformations of the loci, we then performed Molecular Dynamics (MD) simulations as previously described (Annunziatella et al., 2018; Chiariello et al., 2016).

Molecular Dynamics simulations details

Beads and binders are subject to Brownian motion at room temperature T and their positions evolve according to the Langevin equation (Allen and Tildesley, 1989). We set the dimensionless diameter σ and mass m of the beads and binders equal to 1 (Kremer and Grest, 1990). To account for excluded volume effects due to the hard-core nature of the particles, we employed a purely repulsive Lennard-Jones potential between any two particles with length scale σ and energy scale $\epsilon = kBT$ (Kremer and Grest, 1990). Additionally, between two consecutive beads of the polymer we imposed a finitely extensible non-linear elastic spring (FENE; Kremer and Grest, 1990), with length constant $R_0 = 1.6\sigma$ and spring constant $K = 30kBT/\sigma^2$ (Annunziatella et al., 2016; Barbieri et al., 2017; Brackley et al., 2013; Chiariello et al., 2016). The interaction between beads and binders of the same type was modeled by an attractive LJ potential, truncated at $R_{int} = 1.5\sigma$ (Chiariello et al., 2016). The interaction energy E_{int} is given by the minimum of this potential and results $8.16kBT$. These parameters are summarized in Table S3. To integrate the Langevin equation, we used the LAMMPS package (Plipton, 1995). The dynamic dimensionless parameters were set to standard values (Annunziatella et al., 2018), with friction coefficient $\zeta = 0.5$, temperature $T = 1$ and integration time step $\delta t = 0.012$ (Kremer and Grest, 1990; Rosa and Everaers, 2008). The system was confined in a cubic simulation box with periodic boundary conditions, with edge size approximately as twice as the gyration radius of a Self-Avoiding Walk (SAW) polymer with N beads, in order to minimize finite size effects. Each simulation started from a standard SAW polymer configuration (Kremer and Grest, 1990). The binders of each color were initially uniformly distributed in the simulation box with a concentration above the coil-globule transition threshold, before binding to the polymer and driving its folding in a globular structure. In that respect, the precise value of the binders does not matter. Anyway, in particular, we set their number to be roughly 20% of the corresponding number of binding sites, as reported in Tables S1 and S2. For each model, we performed up to 10^2 independent simulations, which were equilibrated up to 2×10^8 timesteps, in both the coil and globular phases. Starting from the initial state, the configurations were logarithmically sampled, and taken every 10^6 timesteps when the polymer was completely folded. A filter on the polymer size in the compact state (about 50% gyration radius cut-off) was applied for the following analyses.

Polymer models based on Hi-C data

To model the structure of the extended globin loci, we used Hi-C data at 20 kb resolution, spanning 3.3 Mb around the α - and β -globin genes (chr11:29,902,000-33,202,000 and chr7:109,307,000-112,607,000, respectively). Using the PRISMR method mentioned above, our procedure returned an optimum number of $n = 16$ different binding types, used for both mouse ES and erythroid cells, hence the corresponding polymer chains are made of $N = (3.3\text{Mb}/20\text{kb}) \times 16 = 2640$ beads. The factor 16, which coincides with the number of binding sites, as described in Bianco et al. (2018), is necessary to model the presence of more binding types in one 20kb bin. Here we used a total binder concentration per volume unit of $c \sim 0.05\%$, which is above the coil-globule transition threshold (Barbieri et al., 2012, 2017). This is related to the molar concentration c_m through the relation $c = c_m \cdot (\sigma^3 N_A)$, where σ is the length unit (see below) and N_A is the Avogadro number.

Polymer models based on Capture-C data

To model the structure of the globin loci in deeper detail, we used recent high-resolution Capture-C data (Oudelaar et al., 2018) from viewpoints closely spaced across 300 kb around the α - and β -globin genes (chr11:32,001,350-32,301,350 and chr7:110,841,000-111,141,000, respectively). We modeled these data at 4 kb resolution and derived missing entries in the contact matrices by linearly interpolating the experimental data. Then, as described in the previous section, we found the number of types and position of the binding sites and derived the polymer chains in mouse ES and erythroid cells. Here, the optimum number of binding sites resulted $n = 10$. Anyway, for sake of simplicity, we choose as safe option for all polymers the maximum number of binding sites ($n = 16$) obtained among the datasets analyzed. Therefore, the polymers are made of $N = (300\text{kb}/4\text{kb}) \times 16 = 1200$ beads. The total binder concentration per volume unit used here is $c \sim 0.03\%$, and again ensures the coil-globule transition. In Figures 4B and S4B we showed the first top 5 most important types contributing to the polymer architecture (see next sections). The Capture-C data, normalized as described in Oudelaar et al. (2018), allow a direct comparison between ES and erythroid cell types. Analogously, the total number of contacts obtained from the simulated conformations reflects the structural information contained in the normalized data and thus can be fairly compared. The binding sites number and the corresponding number of cognate binders are reported in Tables S1 and S2, for the α - and β -globin respectively.

Toy polymer model based on regulatory elements position

The toy polymer model for the restricted α -globin locus in erythroid cells, reported in Figure S5, is a SAW with $N = 300\text{kb}/4\text{kb} = 75$ beads. Here, the binding sites are selected from the full Capture-C based model at the position of specific chosen regulatory element (Figure 5C, toy polymer linear scheme at the bottom), according to the 4kb binning used for the experimental Capture-C data and, importantly, correspond to three different most contributing binding site types (Figure 4B, bottom right matrix) located in a symmetric configuration with respect to the globin gene domain. Specifically, we used 3 enhancers (R1, R2 and R4), the $\alpha 1$ and $\alpha 2$ gene promoters and four CTCF sites, with only three different binding site types, distributed therefore on nine different binding sites. Here, the binders are taken in the same number of the binding sites. The contact matrix obtained from MD simulation of the toy model is then coarse-grained by a factor 2 to facilitate the comparison of the simple toy model with the experimental data.

Contact pairwise and tri-wise matrices

To calculate the contact matrices from the polymer configurations, we used a previously described method (Barbieri et al., 2012; Chiariello et al., 2016). We fixed a distance threshold A and measured the distance r_{ij} between any two beads i and j of the same type. If the distance was lower than the threshold, the beads were considered in contact. For each bead pair, the contact score was averaged over the different independent simulations to create an ensemble average, from the configurations in the coil SAW state or from the configurations in the equilibrated compact globular state. The contact matrices shown in the figures are a weighted average, in order to maximize the Pearson correlation coefficients r with the corresponding experimental data (Bianco et al., 2018; Chiariello et al., 2016). This is simply done by scanning the space of weights and selecting the value that gives the highest r . For instance, the average coefficient for the pure compact ensemble of the α -globin from Capture-C is 0.88 while it results 0.96 in the mixture with 0.64/0.36 open/closed fraction in erythroid and 0.67/0.33 in ES (similar numbers are found for the β -globin model). We report also as other measure of similarity with the experimental data the distance-corrected Pearson coefficient r' , which takes into account for the monotonic decay of the contact probability with the genomic distance (Bianco et al., 2018) (again, it is slightly affected by the weighting procedure since in the α -globin the average coefficient results 0.8 for the pure ensemble against $r' = 0.89$ for the mixture). As further check, we employed the stratum-adjusted correlation coefficient SCC from the HiCRep method (Yang et al., 2017), calculated using standard parameters, and find even with this more restrictive method high correlations. In the α -globin (β -globin) model from Capture-C data, we obtain 0.75 (0.52) and 0.92 (0.77) for ES and erythroid respectively. In all cases, the values are higher than random control model ($p\text{-val} = 0.0$, see below). Analogously, in the α -globin (β -globin) model from HiC data, we obtain 0.63 (0.58) and 0.67 (0.45) for ES and erythroid respectively (here, we considered genomic distances up to 750kb to take into account the long-range contacts).

Average distances for a pairwise contact can be estimated from our equilibrium ensemble and result in the Capture-C models approximately 150 ± 90 nm. The subtraction maps were produced by subtracting erythroid-ES matrices, and then normalizing the difference by the average ES score at a fixed genomic distance (Bianco et al., 2018). The model tri-wise matrices that were compared to the Tri-C data were calculated using a similar approach. We fixed the viewpoint on the polymer, that is the bin at 4kb containing the regulatory element of interest R2 or CTCF HS-39, calculated all potential triplets, and considered a triplet formed when all the three possible pairs were in contact, according to the criterion above described for the pairwise contacts. As for the pairwise contacts, we estimated a global, average distance among the three bodies (that is 4kb polymer segments) forming the triplet and found 140 ± 70 nm. Among all the possible triplets with a fixed viewpoint, we consider most prominent those below 20th percentile of their average distance distribution. Indeed, in erythroid the viewpoint R2, that has the most biologically relevant triplets, forms 81% of such strong triplets with loci in the enhancer-promoter region. Their three-body average distance is approximately 100nm, which is likely the distance of a biological three-body contact. Importantly, such estimates are fully compatible with recent experimental findings (Bintu et al., 2018), where an upper contact threshold of 200 nm is used. In analogy with the pairwise matrices, we tested the accuracy of the predicted triplets by calculating Pearson correlation coefficient r and also the SCC from HiCRep between model and Tri-C data and found high values (in ES, R2 and CTCF HS-39, SCC coefficient is 0.66 and 0.77, in erythroid, 0.48 and 0.79) above random

control model ($p\text{-val} = 0.0$). The random control model is an ensemble of matrices generated from the original data with bootstrapped diagonals.

Polymer graphics

The 3D conformations derived from the polymer models shown in the figures are fully equilibrated configurations taken from MD simulations. They are represented as smooth curves obtained from a third order polynomial spline of the spatial coordinates of the polymer beads. Additionally, to better visualize individual *cis*-regulatory elements, we produced coarse-grained structures. Where necessary, finite size effect were corrected by cutting the polymer tails, for visualization purposes. All conformations were produced using the POVray software.

Polymer intermingling

To quantify the degree of intermingling between two genomic regions, we considered the measure $I = (rg1 + rg2 - r12) / (rg1 + rg2)$, in which $rg1$ and $rg2$ are the gyration radii of the two regions in the polymer, and $r12$ is the average distance between their centers of mass (Figure 2C). Three cases can occur: if $I > 0$, the distance is lower than the sum of the gyration radii, and the two regions intermingle; if $I < 0$, the distance is larger than the sum of the gyration radii, and the regions are separated; if $I = 0$, the sum of the gyration radii equals the distance, and the regions are approximately “touching” just on their surfaces. From our simulations, it results that the intermingling I is stable during long dynamics, therefore the hairpin structure does not change dramatically in time, as reported in Figure S3B where an example of single dynamics is shown.

Structural analysis of the binding domains

To visualize the contribution of the different binding domains composing the polymer models of the globin loci at 4kb resolution, we calculated the binding domains with the strongest contribution to each contact and represented these in a matrix (Figure 4A, bottom matrices; and Figure S4A, bottom matrices). The contribution of each binding domain to a given contact was defined as the number of binding site pairs of that type common to the two considered loci, as interactions can only take place between binding sites of the same type. The order of the overall contribution of the domains to the contact pattern was derived from their total number of contacts in the matrices representing the strongest binding site contributions.

Epigenetics features – Analysis

We compared the predicted binding domains to epigenetic features of the globin loci, in erythroid and ES cells, using the following datasets:

- DNase-seq: Erythroid (Kowalczyk et al., 2012); ES (Domcke et al., 2015);
- CTCF ChIP-seq: Erythroid (Hanssen et al., 2017); ES (Nitzsche et al., 2011);
- Rad21 (Cohesin) ChIP-seq: Erythroid (Hanssen et al., 2017); ES (Nitzsche et al., 2011);
- Nascent transcription: Erythroid (Hay et al., 2016); ES (Strikoudis et al., 2016);
- Polymerase II ChIP-seq: Erythroid (new data); ES (Rahl et al., 2010);
- H3K36me3 ChIP-seq: Erythroid and ES (Yue et al., 2014);
- H3K27ac ChIP-seq: Erythroid (Kowalczyk et al., 2012); ES (Wamstad et al., 2012);
- H3K4me3: Erythroid (Kowalczyk et al., 2012); ES (Wamstad et al., 2012);
- H3K4me1: Erythroid (Kowalczyk et al., 2012); ES (Wamstad et al., 2012);
- H3K9me3: Erythroid (Yue et al., 2014); ES (Bulut-Karslioglu et al., 2014);
- H3K27me3: Erythroid (Kowalczyk et al., 2012); ES (new data);

The new ChIP-seq data were generated using previously described procedures (Hay et al., 2016) using antibodies for Polymerase II (Santa Cruz, catalog number: sc899c, lot: H3115) and H3K27me3 (Millipore [previously Upstate], catalog number: 07-449, lot: DAM1641103). To determine the association of these features with the predicted binding domains of the SBS model at 4 kb resolution, we (re-)analyzed these datasets to express the counts at the same resolution. For the marks that are distributed in peaks (DNase, CTCF and Cohesin) we considered both the number and area of the peaks; for the widely distributed marks, we considered all counts within 4 kb bins. In brief, DNase-seq and ChIP-seq data were aligned to the mm9 reference genome using the NGseqBasic mapping pipeline (<http://userweb.molbiol.ox.ac.uk/public/telenius/NGseqBasicManual/external>; V20.0). Duplicated reads and reads mapping to chrM, chrX and chrY were excluded from analysis. In the DNase-seq data and the CTCF and Cohesin ChIP-seq data, peak regions were identified using Macs2 (default parameters, without input-correction). Read coverage over ± 250 bp from the summits of these peaks was counted and annotated to the peak summit base. These counts were binned in 4 kb bins. In all other ChIP-seq datasets, read coverage was counted directly over 4 kb bins. Nascent RNA-seq data were aligned to the mm9 reference genome using TopHat (v.1.1.4b; default parameters) and filtered for ribosomal RNA, reads mapping more than twice and reads mapping to chrM, chrX, chrY. Read coverage was counted over 4 kb bins. All counts were input-corrected, normalized to 10 million reads and floored to integers. Where applicable, counts of replicates were averaged, and negative counts were set to zero.

Epigenetics features – Correlation with predicted binding domains

We compared the epigenetic features described above to the predicted binding domains of the globin loci, in ES and erythroid cells. As these binding domains are *de novo* predictions that could reflect any feature or combinations of features (Bianco et al., 2018), our ability to find good correlations with the epigenetic marks that were considered was limited. Nevertheless, we were able to find distinct correlation patterns for most of the predicted binding domains. These patterns were derived from the Pearson correlation coefficients between the number of binding sites of each type and the counts of each epigenetic feature in corresponding bins. Statistical analyses were performed using a random control model to bootstrap our binding domains and re-calculate their correlations with the considered epigenetic marks. The significant correlations shown in the figures are those above the 95th percentile or below the 5th percentile of the distribution of random correlations.

Polymer distance distributions – Correlation with FISH data

To validate that the dynamic conformations of our polymer models represent real variation in individual cells, we compared derived distance distributions to single-cell fluorescent *in situ* hybridization (FISH) data of the α -globin locus (Brown et al., 2018). To this aim, we first extracted the physical value of the bead diameter σ (Brackley et al., 2016). For each considered probe pair (F1-F2 and α -F2), we calculated the average distance (experimental and simulated) and calculated the conversion factor by equalizing the two values (Brackley et al., 2016). Importantly, we averaged the conversion factors from mouse ES and erythroid cells and obtained $\sigma = 42.6\text{nm}$. Even in this restrictive case, the simulated distributions well described the experimental data (p values > 1%, Kolmogorov-Smirnov test). The distances in the polymer models were calculated between the centers of mass of the polymer regions corresponding to the experimental probes using the model of the extended (3.3 Mb) α -globin locus at 20 kb resolution. In the model distributions, the last ten frames of the simulation run, for each replicate, were considered.

Polymer distance distributions – Analyses of the dynamic globin loci

The distance distributions between individual *cis*-regulatory and control elements in the globin loci were extracted from the models at 4 kb resolution that span 300 kb around the globin genes. In this case, a bead diameter of $\sigma = 25\text{nm}$ was used, obtained by comparing the simulated distance distributions with the FISH data, as described in the previous section. Since the F1 and F2 probes fall outside the modeled region, we considered the distribution of the distances between the beads corresponding to the middle coordinates of the probes. As expected, we find a lower value for σ , since the genomic content of the elementary bead is lower than in the model at 20 kb resolution. From this length scale value follows the timescale, τ , fixed by the standard MD relation $\tau = \eta(6\pi\sigma^3/\epsilon)$ and gives, for a viscosity $\eta = 1\text{cP}$, previously employed in similar models (Brackley et al., 2013; Chiariello et al., 2016), $\tau \sim 0.1\text{ms}$. The distances in the models were calculated between the centers of mass of the polymer regions corresponding to the position of the *cis*-regulatory and control elements in the model distributions, considering the last ten frames of each simulation run.

QUANTIFICATION AND STATISTICAL ANALYSIS

All the statistical tests employed are specified in the text and in the Method Details section. Pearson correlation coefficients and stratum-adjusted correlation coefficients (SCC) from the HicRep method (Yang et al., 2017) were used to compare experimental and simulated pairwise and tri-wise contact matrices. Pearson coefficients were also used to compare model binding sites with epigenetic features. Kolmogorov-Smirnov (KS) tests were used to compare the distributions of physical distances between FISH experiments and models while Mann-Whitneyu were used to compare intermingling distributions.

DATA AND CODE AVAILABILITY

ChIP-seq sequencing data have been submitted to the NCBI Gene Expression Omnibus (GEO; <http://www.ncbi.nlm.nih.gov/geo/>) under accession numbers GSE107937 (H3K27me3) and GSE107938 (Pol II). Custom scripts used are available from the corresponding author on request.

Supplemental Information

A Dynamic Folded Hairpin Conformation

Is Associated with α -Globin Activation

in Erythroid Cells

Andrea M. Chiariello, Simona Bianco, A. Marieke Oudelaar, Andrea Esposito, Carlo Annunziatella, Luca Fiorillo, Mattia Conte, Alfonso Corrado, Antonella Prisco, Martin S.C. Larke, Jelena M. Telenius, Renato Sciarretta, Francesco Musella, Veronica J. Buckle, Douglas R. Higgs, Jim R. Hughes, and Mario Nicodemi

Supplemental Information

for

***A dynamic folded hairpin conformation is associated with
 α -globin activation in erythroid cells***

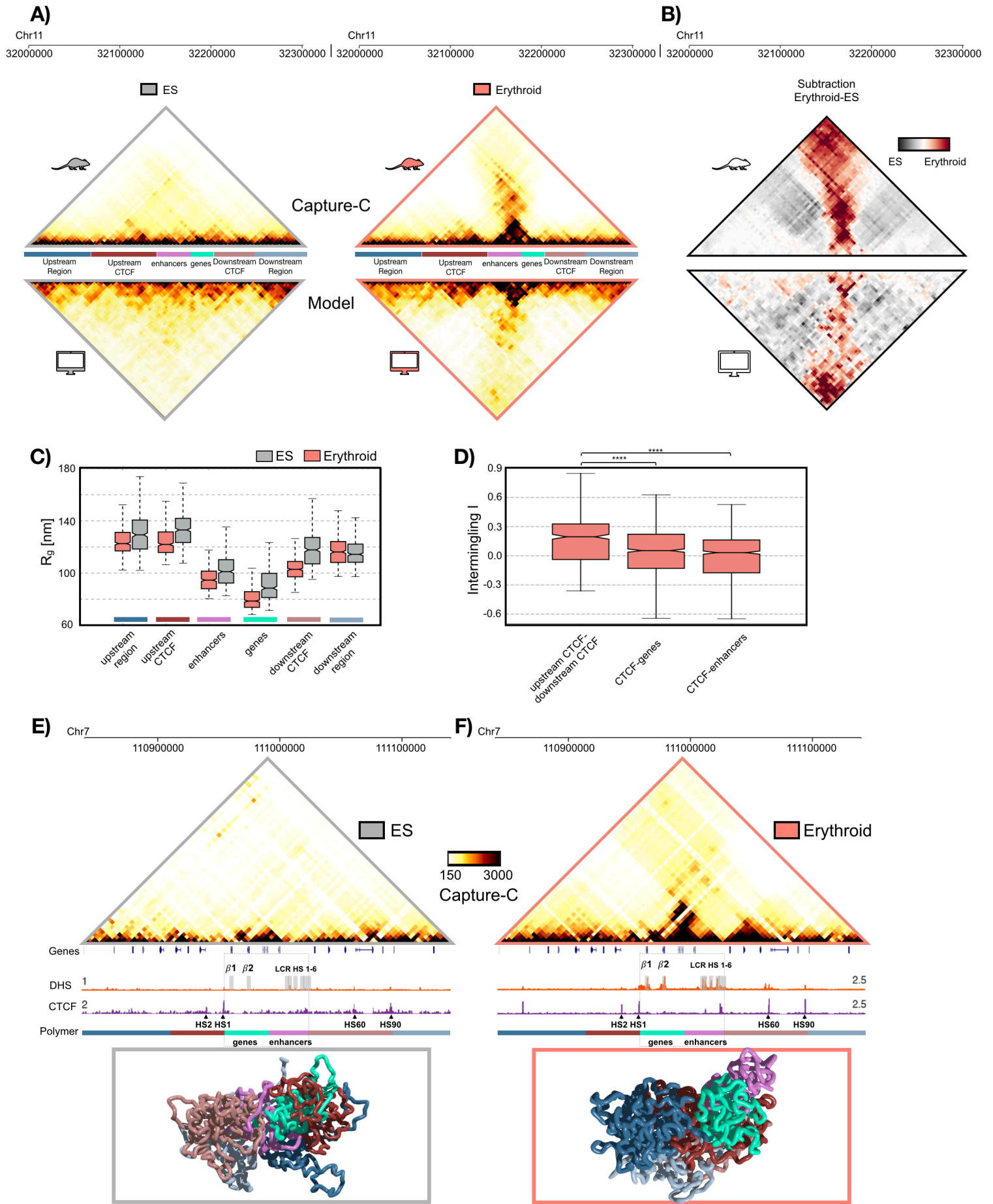


Figure S1 (related to Figure 1). Comparison between the α -globin Capture-C and SBS model inferred contact maps and model derived single-allele conformations of the β -globin locus

A) Capture-C data (top matrices, see Fig.1) and SBS polymer model derived (bottom matrices) contact maps, in ES (left) and erythroid (right) cells. Direct comparison shows that the model accurately recapitulates the experimental data (Pearson correlation $r=0.96$ and distance-corrected correlation $r'=0.87$ and $r'=0.91$ for ES and erythroid respectively). **B)** Subtraction map erythroid-ES of Capture-C (top) and simulated (bottom) contact matrices, normalized by the average contact at a fixed genomic distance in ES. The peculiar elongated shape of the red pixels along the anti-diagonal direction is also well captured by the simulated matrices. **C)** Distribution of the gyration radius in the different polymer sub-regions shown in Panel A and Figure 1A, according to the color scheme used there. Note how those regions are more localized in erythroid than in ES cells. The average gyration radius of the sub-regions considered is approximately 110nm and the maximum distances between the beads of each region is roughly as twice as the average gyration radius. **D)** Distribution of the intermingling in erythroid among the sub-regions containing CTCF, promoters and enhancers. Their coordinates are approximately overlapping with the color scheme of Panel A. Upstream-downstream CTCF regions tend to intermingle between them rather than with enhancer or gene regions, since their average intermingling is approximately zero. The stars indicate Mann-Whitney test $p\text{-val}<10^{-4}$. **E)** High resolution Capture-C data (4kb resolution) for a 300kb window around the β -genes for ES (left) and erythroid (right) cells. Gene annotation, DNaseI Hypersensitive Sites (DHS) and CTCF are shown below. The grey bars indicate the $\beta 1$ and $\beta 2$ gene promoters and the LCS HS 1-6 enhancers. The dashed box highlights the enhancer-promoter region. The colored bar represents the linear color scheme used for the polymer structures. The corresponding 3D conformations obtained from MD are reported below. Their corresponding contact maps are shown in Figure S4 (Pearson correlation with Capture-C is above $r=0.95$). Genomic coordinates: chr7:110,840,000-111,400,000 mm9. **F)** High resolution Capture-C data (4kb resolution) for a 300kb window around the β -globin genes for erythroid cells. Data from (Oudelaar et al., 2018).

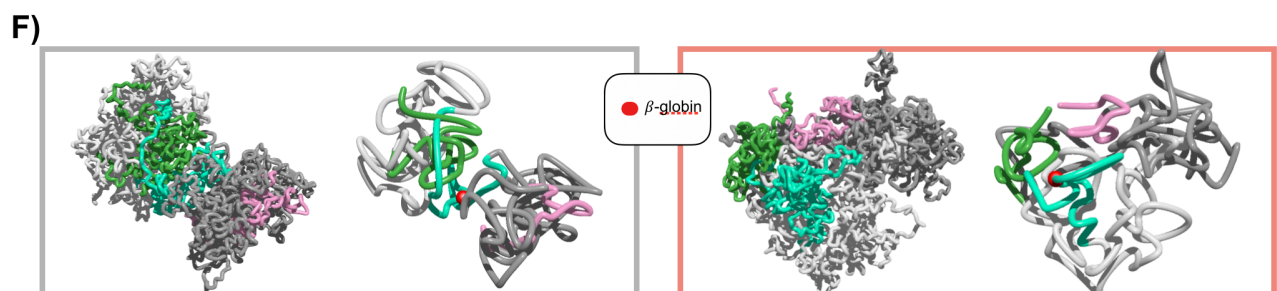
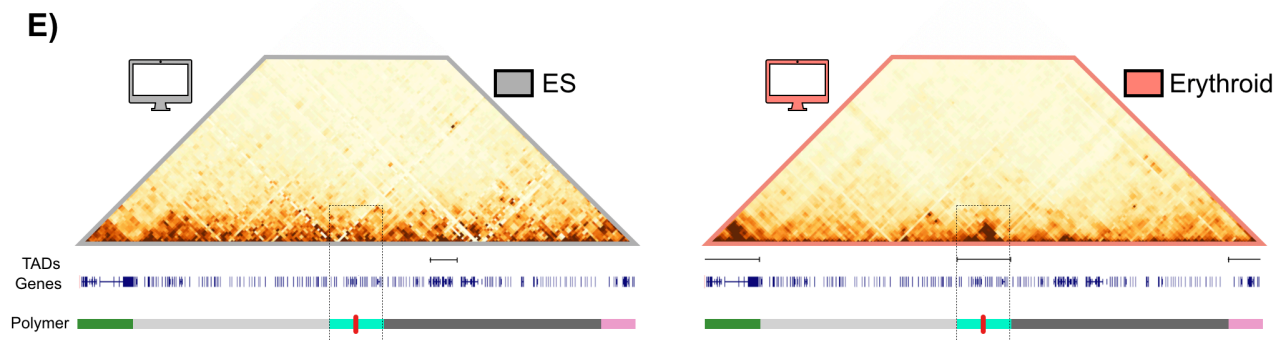
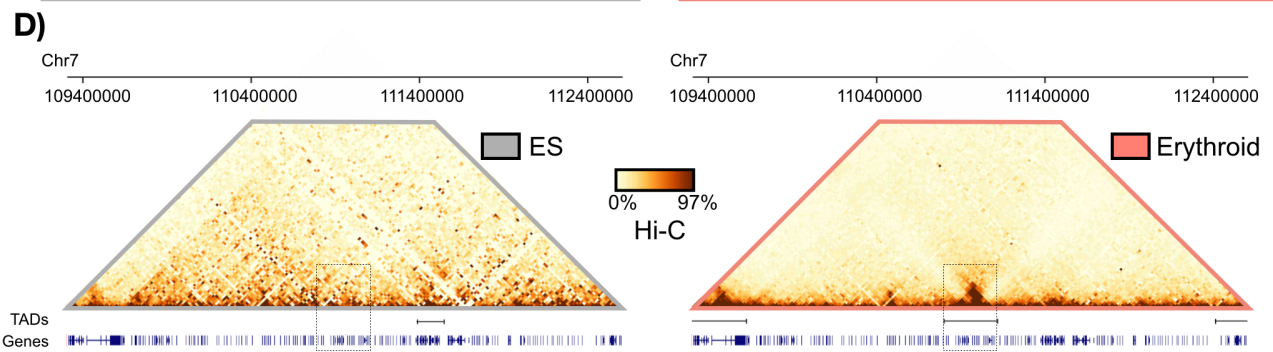
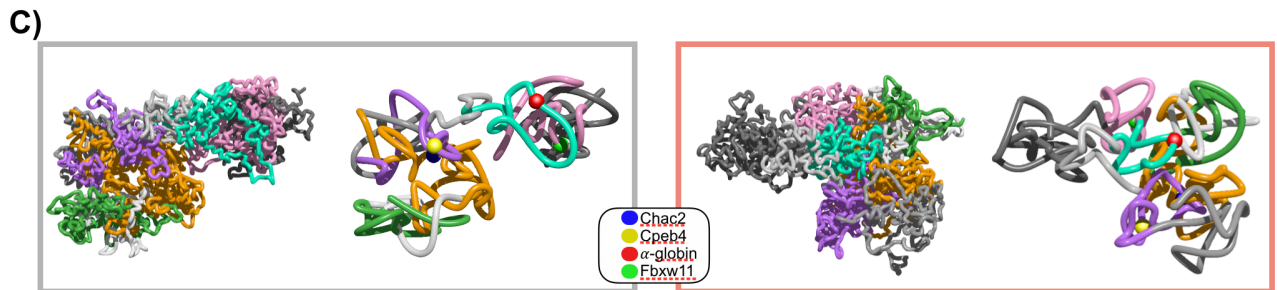
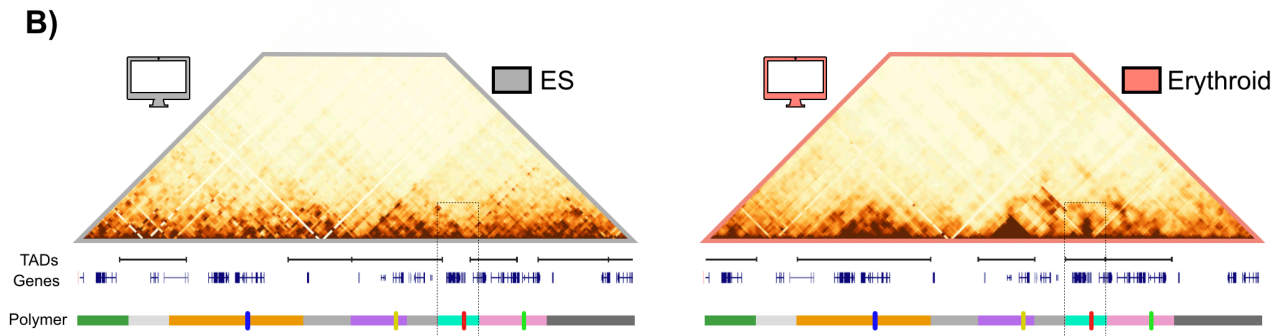
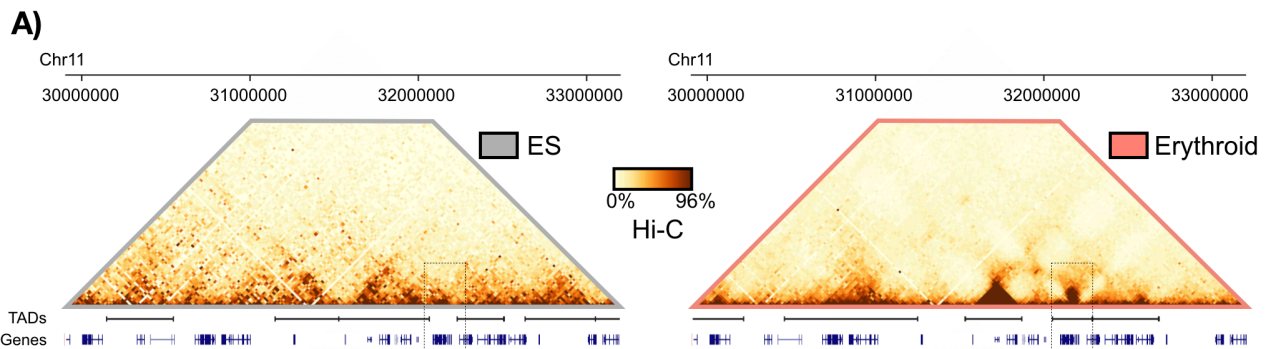


Figure S2 (Related to Figure 1 and STAR Methods). Model derived single-allele conformations of the extended α - and β -globin loci.

A) Hi-C contact matrices (20 kb resolution) of the α -globin locus in ES (left) and erythroid (right) cells. TADs and gene annotation are shown below. Dashed boxes highlight the erythroid TAD containing the α -globin genes. Genomic coordinates: chr11:29,900,000-33,200,000, mm9. Data from (Oudelaar et al., 2018) (erythroid) and from (Giorgetti et al., 2016) (ES). **B)** Contact matrices (20 kb resolution) of the α -globin locus in ES (left) and erythroid (right) cells, obtained from the polymer model. The polymer color scheme shown below is chosen according to the erythroid TAD boundaries (Oudelaar et al., 2018). Pearson correlation coefficients with Hi-C data are $r=0.94$ for ES and $r=0.97$ for erythroid. The HiCRep SCC coefficient in ES is $SCC=0.63$ and in erythroid $SCC=0.67$. **C)** 3D conformations (full and coarse-grained) of the α -globin locus (20 kb resolution) derived from the SBS models of ES (left) and erythroid (right) cells. In ES, the polymer appears much more intermingled than in erythroid case. Some important key genes are also shown. **D)** Hi-C contact matrices (20 kb resolution) of the β -globin locus in ES (left) and erythroid (right) cells. TADs and gene annotation are shown below. Dashed boxes highlight the erythroid TAD containing the β -globin genes. Genomic coordinates: chr7:109,307,000-112,607,000, mm9. Data from (Oudelaar et al., 2018) (erythroid) and from (Giorgetti et al., 2016) (ES). **E)** Contact matrices (20 kb resolution) of the β -globin locus in ES (left) and erythroid (right) cells, obtained from the polymer model. The polymer color scheme shown below is chosen according to the erythroid TAD boundaries. Pearson correlation coefficients with Hi-C data are $r=0.91$ for ES and $r=0.96$ for erythroid. The HiCRep coefficient in ES is $SCC=0.58$ and in erythroid $SCC=0.45$. **F)** 3D conformations (full and coarse-grained) of the β -globin locus (20 kb resolution) derived from the SBS models of ES (left) and erythroid (right) cells. In ES, the β -globin gene is located in a region poorly defined (colored in cyan) and highly intermingled with the rest of the polymer. In erythroid, such region is highly defined.

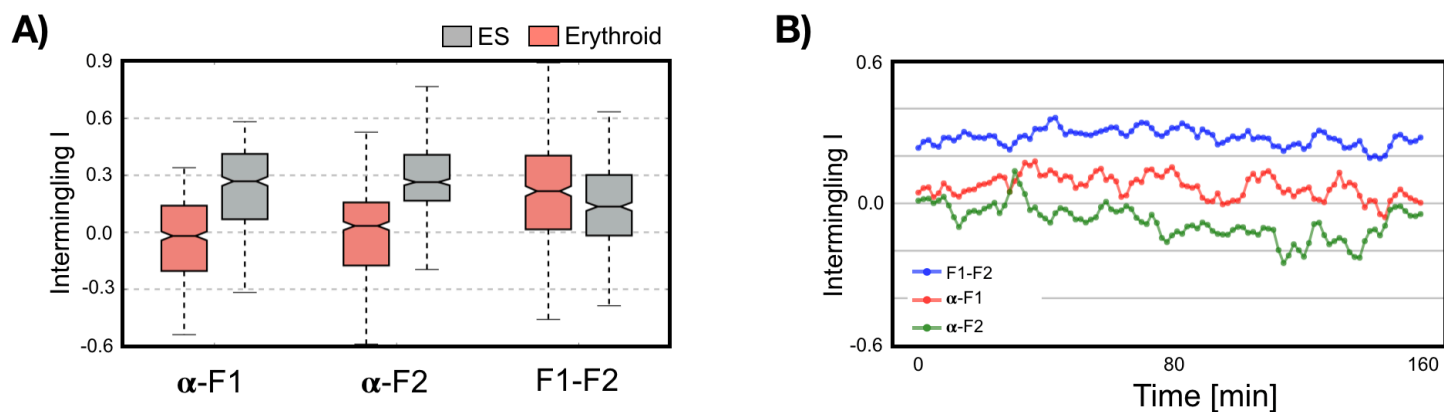


Figure S3 (Related to Figure 2). Analysis of the structural properties reveals heterogeneity and stability in the α -globin higher-order architecture.

A) The full distribution of intermingling among the regions shown in Figure 2A highlights the level of structural heterogeneity in the ensemble of polymer 3D structures. **B)** The dynamics of the Intermingling index I between the regions α -domain, F1 and F2 during a typical Molecular Dynamics run for an erythroid polymer. It exhibits small, thermal fluctuation around its equilibrium value for physical time scales of roughly tens of minutes or hour.

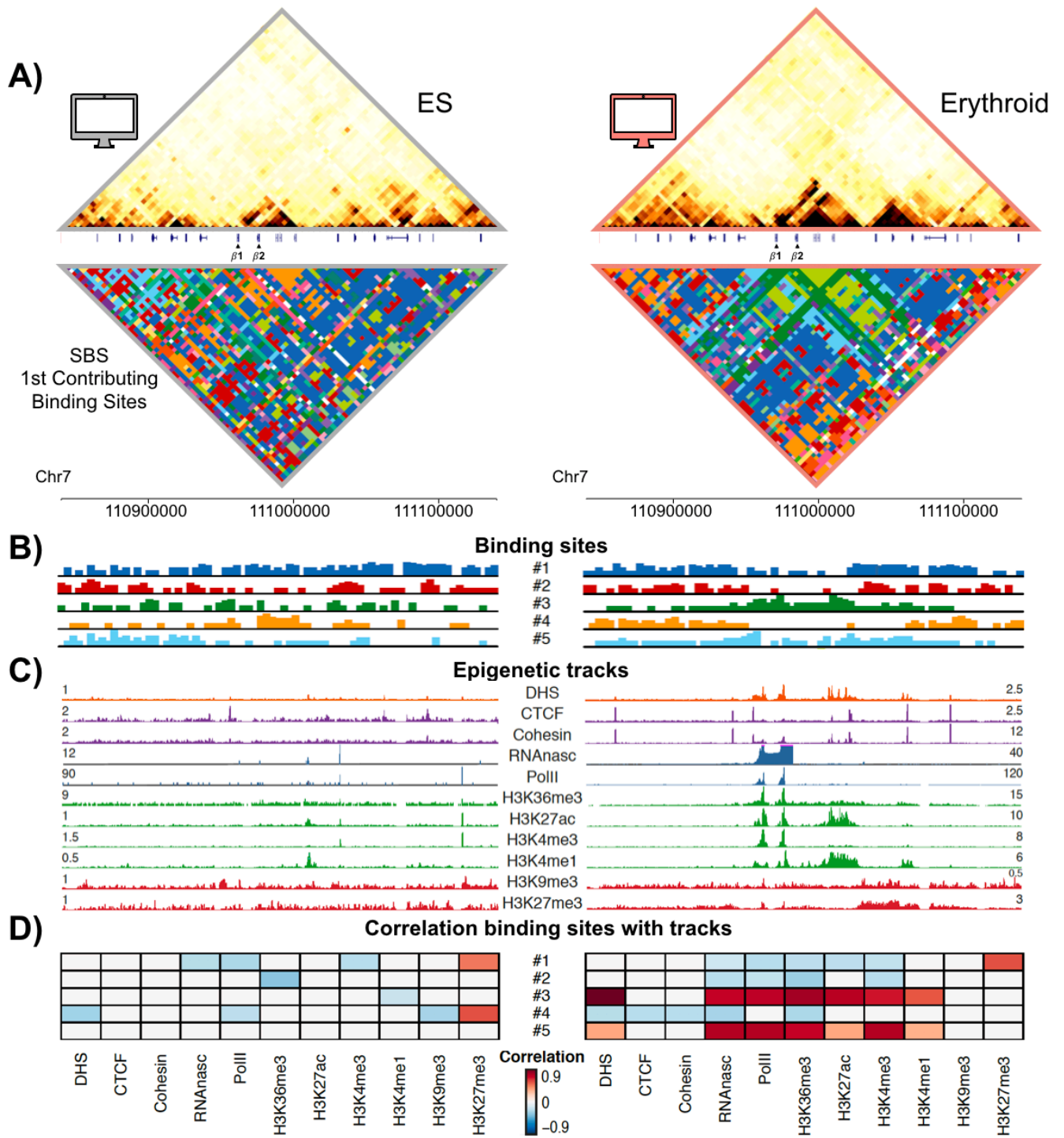


Figure S4 (Related to Figure 4). Binding domains contributing to the 3D structure of the β -globin locus correlate with distinct epigenetic features.

A) Contact map obtained from the 3D polymer model (top matrices). The agreement with experimental Capture-C is high (in ES, Pearson $r=0.95$ and distance corrected Pearson $r'=0.90$, in erythroid, $r=0.96$ and $r'=0.92$). The HiCRep coefficient in ES is $SCC=0.52$ and in erythroid $SCC=0.77$). On the bottom, the most contributing binding sites for each contact is shown. Gene annotations is shown in the middle. **B)** Top five most contributing binding sites to the locus architecture, sorted by their contribution to the overall contact map of the whole locus. In ES, the first most contributing accounts for about 35% of total contacts, while in

erythroid the contacts are associated with more complex set of binding site types. **C)** Epigenetic features of the β -globin locus. DHS = DNaseI Hypersensitive Sites; RNAnasc = nascent RNA expression; PolII = RNA Polymerase II occupancy. **D)** Comparisons of the predicted binding domains to chromatin data show that the dominating binding domain in ES cells is correlated with repressive chromatin marks, while the binding sites associated with the erythroid structure also correlate with chromatin marks of active transcription. Values represent significant Pearson correlation coefficients.

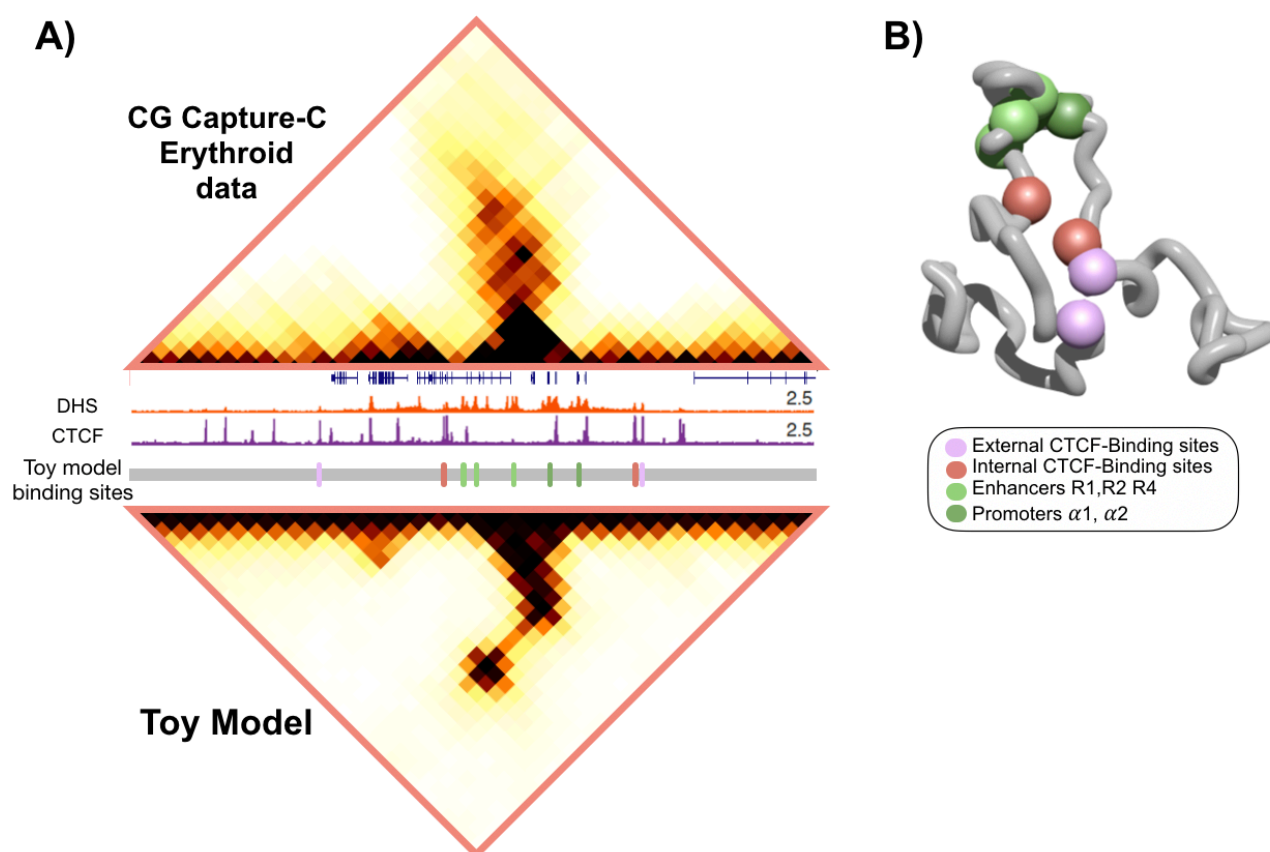


Figure S5 (Related to Figure 5). A simple toy, hairpin-shaped model can help visualising the folded hairpin structure of the α -globin in erythroid data.

A) Coarse-grained (CG) Capture-C data in erythroid cells (top matrix) compared with contact map resulting from the toy model (bottom matrix) with a symmetric arrangement of binding sites corresponding to promoters, enhancers and CTCF sites. **B)** A hairpin-like folded structure obtained from the MD simulation of the toy model.

Tables

Locus	α -globin			
Cell Type	Erythroid		ES	
	# Binding Sites	# Cognate Binders	# Binding Sites	# Cognate Binders
Type 1	104	20	135	27
Type 2	91	18	79	15
Type 3	91	18	78	15
Type 4	73	14	51	10
Type 5	86	17	52	10
Type 6	60	12	68	13
Type 7	63	12	60	12
Type 8	44	8	71	14
Type 9	54	10	38	7
Type 10	52	10	59	11
Type 11	62	12	55	11
Type 12	60	12	49	9
Type 13	54	10	43	8
Type 14	46	9	49	9
Type 15	51	10	42	8
Type 16	47	9	36	7
Inert sites	162	-	235	-
Total number of beads	1200		1200	
Total number of binders		201		186

Table S1 (Related to Figure 4). Details of the simulated polymer model from Capture-C data, for the α -globin in ES and erythroid. For each binding site, the number of sites along the chain and the number of cognate binders is reported. The distributions along the chain of the first 5 types are shown in Figure 4B.

Locus	β -globin			
Cell Type	Erythroid		ES	
	# Binding Sites	# Cognate Binders	# Binding Sites	# Cognate Binders
Type 1	138	27	149	29
Type 2	76	15	88	17

Type 3	106	21	58	11
Type 4	74	14	62	12
Type 5	92	18	73	14
Type 6	74	14	62	12
Type 7	61	12	54	10
Type 8	65	13	69	13
Type 9	45	9	51	10
Type 10	40	8	50	10
Type 11	57	11	67	13
Type 12	50	10	51	10
Type 13	50	10	41	8
Type 14	50	10	41	8
Type 15	34	6	33	6
Type 16	31	6	39	7
Inert sites	157	-	212	-
Total number of beads	1200		1200	
Total number of binders		204		190

Table S2 (Related to Figure S4). Details of the simulated of the polymer model from Capture-C data, for the β -globin in ES and erythroid. For each binding site, the number of sites along the chain and the number of cognate binders is reported. The distributions along the chain of the first 5 are shown in Figure S4B.

SBS Polymer simulation details			
Parameter	Value	Parameter	Value
Bead diameter	$\sigma=1$	Binder diameter	$\sigma=1$
LJ repulsive ($K_B T$ units)	$\epsilon=1$	LJ attractive ($K_B T$ units)	$\epsilon=12$
FENE length constant (σ)	$R_0=1.6$	Bead-Binder interaction range (σ units)	$R_{int}=1.5$
Stiffness (kb units)	4 (Capture-C model)	Bead-Binder Interaction intensity ($K_B T$ units)	$E_{int}=8.16$
	20 (HiC model)		
FENE Spring constant ($K_B T/\sigma^2$ units)	K=30		

Table S3 (Related to STAR Methods). Summary of the Molecular Dynamics parameters employed to simulate the SBS model.