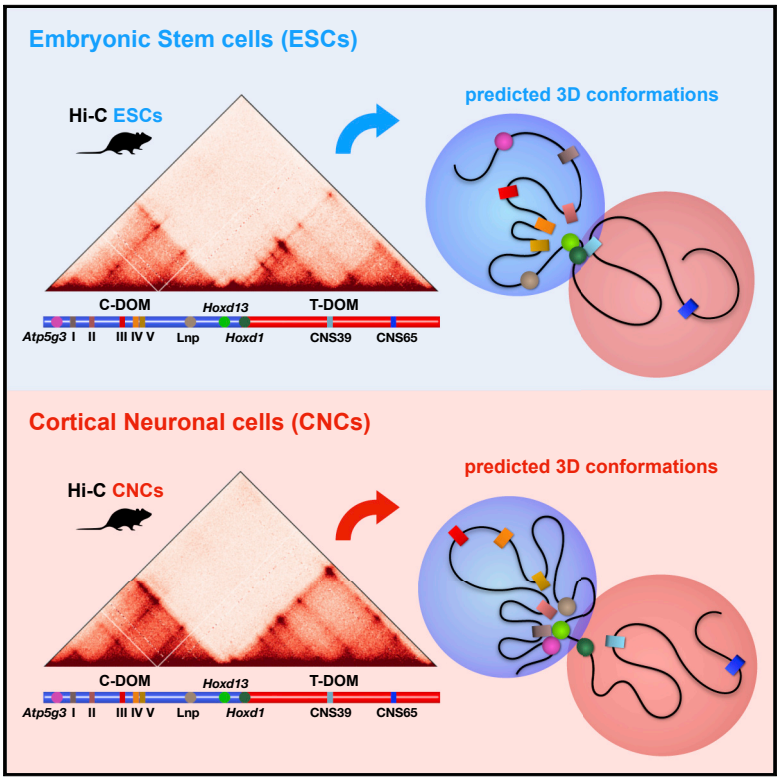
Article

Cell Reports

Modeling Single-Molecule Conformations of the *HoxD* Region in Mouse Embryonic Stem and Cortical Neuronal Cells

Graphical Abstract



Authors

Simona Bianco, Carlo Annunziatella, Guillaume Andrey, ..., Mattia Conte, Raffaele Campanile, Mario Nicodemi

Correspondence

mario.nicodemi@na.infn.it

In Brief

Bianco et al. reconstruct the 3D structure of the murine *HoxD* locus by using polymer models at the single DNA molecule level. The locus architecture rearranges upon differentiation from embryonic stem cells to cortical neurons in connection to epigenetic changes, as cell-type- and gene-specific multi-way contacts are established with regulatory elements.

Highlights

- Single-molecule 3D conformations of the *HoxD* locus are inferred by polymer models
- Model predictions are validated against independent data in wild-type and mutants
- Cell-type- and gene-specific multi-way contacts occur across *HoxD* genes and regulators
- The locus 3D architecture exhibits a cell-type-specific, high cell-to-cell variability

Bianco et al., 2019, Cell Reports *28*, 1574–1583
August 6, 2019
https://doi.org/10.1016/j.celrep.2019.07.013

CellPress

OPEN ACCESS
CellPress

Cell Reports
Article

Modeling Single-Molecule Conformations of the *HoxD* Region in Mouse Embryonic Stem and Cortical Neuronal Cells

Simona Bianco,[1,6] Carlo Annunziatella,[1,6] Guillaume Andrey,[2] Andrea M. Chiariello,[1] Andrea Esposito,[1,3] Luca Fiorillo,[1] Antonella Prisco,[5] Mattia Conte,[1] Raffaele Campanile,[1] and Mario Nicodemi[1,3,4,7,*]

[1]Dipartimento di Fisica, Università di Napoli Federico II, and INFN Napoli, Complesso Universitario di Monte Sant'Angelo, 80126 Naples, Italy
[2]Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland
[3]Berlin Institute for Medical Systems Biology, Max-Delbrück Centre (MDC) for Molecular Medicine, Robert-Rössle Straße, Berlin-Buch 13125, Germany
[4]Berlin Institute of Health (BIH), MDC-Berlin, Berlin, Germany
[5]CNR-IGB, Pietro Castellino 111, Naples, Italy
[6]These authors contributed equally
[7]Lead Contact
*Correspondence: mario.nicodemi@na.infn.it
https://doi.org/10.1016/j.celrep.2019.07.013

SUMMARY

Complex architectural rearrangements are associated to the control of the *HoxD* genes in different cell types; yet, how they are implemented in single cells remains unknown. By use of polymer models, we dissect the locus 3D structure at the single DNA molecule level in mouse embryonic stem and cortical neuronal cells, as the *HoxD* cluster changes from a poised to a silent state. Our model describes published Hi-C, 3-way 4C, and FISH data with high accuracy and is validated against independent 4C data on the Nsi-SB 0.5-Mb duplication and on triple contacts. It reveals the mode of action of compartmentalization on the regulation of the *HoxD* genes that have gene- and cell-type-specific multi-way interactions with their regulatory elements and high cell-to-cell variability. It shows that TADs and higher-order 3D structures, such as metaTADs, associate with distinct combinations of epigenetic factors, including but not limited to CCCTC-binding factor (CTCF) and histone marks.

INTRODUCTION

The transcriptional program of the *HoxD* genes is orchestrated in spatial and temporal correspondence with complex architectural transformations of the locus, as revealed by chromatin contact patterns provided by technologies such as proximity ligation methods (see, e.g., Andrey et al., 2013; Noordermeer et al., 2014). In mouse embryonic stem cells, the *Hox* genes are marked by bivalent chromatin states, with both repressive (H3K27me3) and activating (H3K4me3) signatures (Bernstein et al., 2006; Noordermeer and Duboule, 2013; Schuettengruber et al., 2017; Soshnikova and Duboule, 2009). During mouse embryo development, a collinear activation occurs, as the genes are sequentially turned on according to their genomic position (Deschamps et al., 1999; Kmita and Duboule, 2003). Correspondingly, the locus exhibits a transcriptional-state-dependent 3D compartmentalization, with active genes forming a cluster physically separated from the inactive ones (Noordermeer et al., 2011). Similar complex architectural patterns are also found during limb bud development and in other tissues (Andrey et al., 2013; Noordermeer et al., 2014). The hypothesis has been raised that such 3D compartmentalization has a general functional role, which may help, for instance, the maintenance of the transcriptional states by avoiding contacts between the active and inactive genes, and by restricting the use of enhancer repertoires during development (Andrey et al., 2013; Noordermeer et al., 2011). However, it is unknown how such a regulatory program is implemented at the single-cell level and, in particular, the corresponding folding mechanisms that control contact specificity between genes and regulators at different transcriptional stages.

To investigate those topics, we use models of polymer physics (Chiariello et al., 2016; Nicodemi and Prisco, 2009), as they can provide the 3D structure of the *HoxD* locus at the single DNA molecule level and help dissecting the specific mechanisms of folding. To explore the role of different chromatin states and their impact on the 3D architecture, we compare embryonic stem cells (ESCs) to mouse cortical neuronal cells (CNCs), as the *HoxD* genes change from a poised to a silent state. We also investigate the differences between *in-vitro*-differentiated CNCs and *in vivo* cortex tissue (Cortex). We explore a 7-Mb-wide region around the *HoxD* cluster, encompassing its flanking topologically associating domains (TADs). To investigate the nano-scale details of the 3D structure of the *HoxD* genes and their repositioning upon differentiation, we derive an ensemble of high-resolution, single-molecule 3D conformations. To validate our polymer models, first we show that they describe with high accuracy available Hi-C average interaction data (Bonev et al., 2017; Dixon et al., 2012). Next, we test our results against independent fluorescence *in situ* hybridization (FISH) data (Fabre et al., 2015) and 3-way 4C data (Olivares-Chauvet et al., 2016). Finally, we investigate the effects of a duplication on chromatin

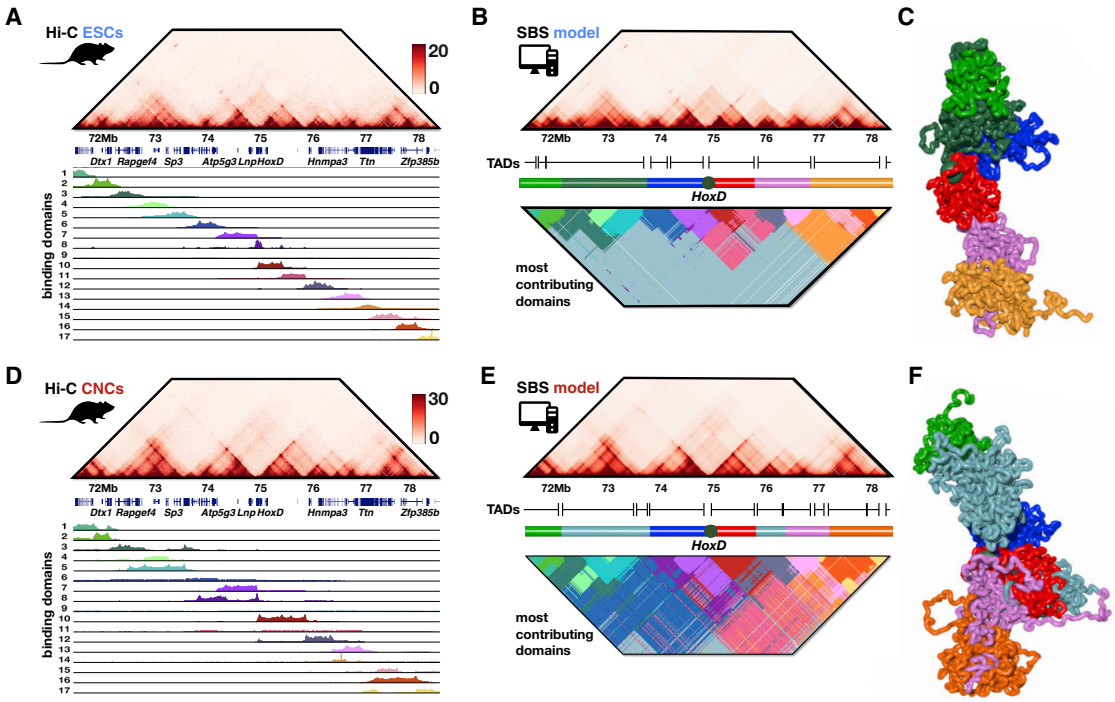


Figure 1. The SBS Model Describes with Good Accuracy Hi-C Patterns in the Extended *HoxD* Region in ESCs and CNCs

(A and B) High-resolution (5 kb) Hi-C data from Bonev et al., (2017) (A, top), and SBS-model-derived average contact matrix (B, top) of the *HoxD* region in mouse ESCs have a Pearson correlation r = 0.92 and a distance corrected correlation r' = 0.48. The model envisaged main binding domains of the locus are shown in (A, bottom), while the main contributing binding domains to the contacts are shown in (B, bottom). The TADs of the locus (black segments, Bonev et al., 2017) correspond to regions enriched for contacts between a specific type of binding sites.

(C–F) As binding domains overlap, TADs have internal structures and interactions with each other. A single-molecule time snapshot visualizes the 3D conformations corresponding to the TADs (C, color scheme in B). The green sphere in the structure highlights the position of the *HoxD* cluster. High-resolution (5 kb) Hi-C data from Bonev et al., (2017) (D, top) and the SBS-model-derived average contact matrix (E, top) in CNCs have a correlation r = 0.93 and r' = 0.59. The model main binding domains (D, bottom) have broader overlaps in CNCs than in ESCs, producing interactions across TADs (E, bottom), as seen in the contact maps. Correspondingly, higher-order structures (metaTADs) are formed, which are visible in the single-cell 3D time snapshot of the locus (F, color scheme in E).

folding (Montavon et al., 2012) and predict the corresponding changes in interaction frequencies, which are tested against independent 4C data (Montavon et al., 2012). Polymer physics provides a principled explanation of TADs and higher-order structures, deriving from the complex interplay of contacts between specific binding sites and cognate bridging molecules, associated with combinations of epigenetic factors, including but not limited to CCCTC-binding factor (CTCF) and histone marks. In all the considered cell types, the TADs of the locus form a sequence of spatially distinct structures having non-trivial contacts across them, forming higher-order chromatin structures, i.e., metaTADs (Fraser et al., 2015). The 3D architecture of the locus has a high cell-to-cell variability in all the studies cases. In ESCs, the *HoxD* genes establish strong many-body contacts with each other. *Hoxd1* and *Hoxd9* also form triplet contacts with their telomeric TAD (T-DOM) and *Hoxd13* with both the centromeric TAD (C-DOM) and T-DOM. In CNCs, triplets between the *HoxD* genes are less frequent and weaker many-body contacts are established from the *Hoxd9* and *Hoxd1* genes with both their centromeric and T-DOM, whereas *Hoxd13* strongly interacts in triple contacts with its C-DOM. That returns a picture of the mode of action of compartmentalization on gene regulation in different cell types, based on specific, simultaneous interactions with multiple genes and regulatory elements.

Overall, our results provide insights about the 3D structure of the murine *HoxD* region, which is unavailable from current experimental Hi-C and microscopy data. In particular, our model provides conformational details at the single DNA molecule level, showing the high cell-to-cell variability of the 3D structure of the *HoxD* region and returning at the same time multi-way contacts and physical distances between genes and regulatory elements. They also illustrate how the 3D architecture changes upon differentiation in connection to corresponding epigenetic changes.

RESULTS

Polymer Models of the *HoxD* Region in ESCs and CNCs

To investigate the 3D structure of the *HoxD* locus in its broader genomic context, we focus first on a 7-Mb region around the murine *HoxD* cluster, in mouse ESCs and CNCs (Figure 1), at a 5-kb resolution as in published Hi-C data (Bonev et al., 2017). To model the region, we use the String & Binders Switch (SBS) polymer model (Nicodemi and Prisco, 2009) that quantifies a

classical scenario where molecules, such as transcription factors, loop DNA by bridging distal cognate binding sites. The SBS was already shown to recapitulate with high accuracy Hi-C, genome architecture mapping (GAM), and FISH data across loci and cell types (Annunziatella et al., 2016; Barbieri et al., 2012, 2017; Bianco et al., 2018; Chiariello et al., 2016; Fraser et al., 2015; Beagrie et al., 2017). In the SBS, a chromatin filament is described as a self-avoiding chain of beads, each of linear size σ. The chain includes specific beads that act as binding sites for diffusing molecules that can bridge cognate sites. The different sets of homologous binding sites are the binding domains of the polymer model, each represented by a different color in our visualization (Figure 1). They are derived by a machine learning procedure, named the polymer-based recursive statistical inference method (PRISMR) (Bianco et al., 2018; Chiariello et al., 2016), that finds the minimal polymer model that best describes the given Hi-C data of the region based only on polymer physics, without requiring any prior biological knowledge (e.g., of DNA binding proteins). Briefly, PRISMR finds the optimal arrangement of binding sites along a considered genomic region by minimizing a cost function equal to the distance between the input experimental contact matrix of the region and the contact matrix, which the SBS polymer gives at thermodynamics equilibrium, plus a Bayesian term to reduce overfitting (Bianco et al., 2018). To consider population effects, the procedure also considers whether the locus in different cells is in either the open (coil) or in the closed (globular) chromatin state. Specifically, it returns the optimal mixture of single-molecule structures, in the coil and in the globular thermodynamics state, best describing the population-averaged Hi-C contact data. The only free parameter of the model is the scale of distances, σ. To estimate σ, we compared our 3D models with available FISH data in ESCs from Eskeland et al., (2010) by imposing that the average distance between *Hoxd1–Hoxd13* is the same, i.e., about 350nm, and found σ = 40nm (STAR Methods).

Pairwise Contact Frequencies and Model Binding Domains

To test the accuracy of our models, we compared the cell-type-specific patterns of 5-kb resolution Hi-C data (Bonev et al., 2017) in ESCs and in CNCs against the model pairwise contact matrices derived by our SBS models (Figures 1A–1D; STAR Methods). The Pearson's correlation between model and Hi-C data is r = 0.92 and r = 0.93 in ESCs and CNCs, respectively. Additionally, to consider the average decay of interactions with genomic distance, we also computed the distance corrected Pearson's correlation, i.e., the correlation between the contact matrices where the average decay is subtracted, which results to be r' = 0.48 and r' = 0.59, respectively (STAR Methods).

To test the robustness of the procedure, we also derived polymer models from lower resolution, namely 40 kb, Hi-C data (Dixon et al., 2012) of the *HoxD* region in mouse ESCs and Cortex tissue. Although finer, smaller structures are visible from the 5-kb-resolution dataset, a comparison of the two datasets binned at the same resolution shows that there are overall similar (Figures S1A and S1B). Accordingly, we obtained a similar agreement between experimental and model pairwise matrices (Figures S1C and S1D).

To dissect the origin of the contact patterns of the locus and to provide a principled definition of the otherwise heuristic notion of TAD, we investigated how such patterns arise from polymer physics by the interactions of the model binding sites. In ESCs, the model identifies 17 main binding domains (20 in total, STAR Methods), each of which tends to have a high overlap with a single TAD or sub-TAD (Bonev et al., 2017) of the locus (Figures 1A, 1B, and S1E; STAR Methods). Figure 1B visualizes the most contributing domain to each pairwise contact, visually illustrating that the TADs in the Hi-C data, identified by Bonev et al., (2017), roughly correspond to DNA regions particularly enriched by contacts linked to one of the binding domains of the model (Figure S1E; STAR Methods). The binding domains tend to overlap with each other along the DNA linear sequence; hence, interactions within a TAD are sometimes associated with more than a single binding domain, reflecting TAD internal structures. The model also identifies binding domains not directly associated to a single TAD, which are more spread over the locus (Figure S1E) and contribute, in particular, to the weaker, yet non-negligible longer range interactions across the locus, producing the visible, complex contact patterns. Similar results are found in CNCs (Figures 1D, 1E, and S1F; STAR Methods) where, for visualization purposes, the color given to the binding domains is chosen based on the highest genomic overlap with the corresponding domain in ESCs (STAR Methods). Interestingly, in CNCs the binding domains have a stronger genomic overlaps with each other with respect to ESCs, originating higher level of inter-TAD interactions (metaTADs; Fraser et al., 2015) seen in Hi-C data (Figures 1B and 1E).

To guess the nature of the molecular factors associated to the different binding domains, we correlated their genomic position with available histone and other epigenetics marks (Feingold et al., 2004; Bonev et al., 2017) and found that each type of binding site corresponds to a different combination of markers, rather than a single factor (Figures S1E and S1F; STAR Methods). The 3D reorganization of the locus from ESCs to CNCs is associated to specific epigenetic changes of the binding domains. For example, domain 8 (purple) in Figures 1A and 1D is the most overlapping one with the *HoxD* genes and so the one mainly involved in the conformational changes of the *HoxD* cluster. In ESCs it is associated to CTCF peaks and to bivalent signatures, i.e., to both active and repressive features, whereas in CNCs only to CTCF and the repressive H3K27me3.

In summary, the high correlation between model and experimental contact data supports a principled interpretation of the interaction patterns based on polymer physics. Structures such as TADs and metaTADs can be explained as regions enriched for contacts between specific types of binding sites, which correlate with different combinations of epigenetic factors, including but not limited to CTCF. The 3D reorganization of the locus from ESCs to CNCs is linked to a broadening of the main binding domains along the linear sequence, correlated to specific epigenetic changes.

The Impact of a Duplication on the Locus 3D Structure and Model Validation

To validate our model, we considered a previously studied 0.5-Mb-long duplication (Nsi-SB) flanking the *HoxD* cluster where

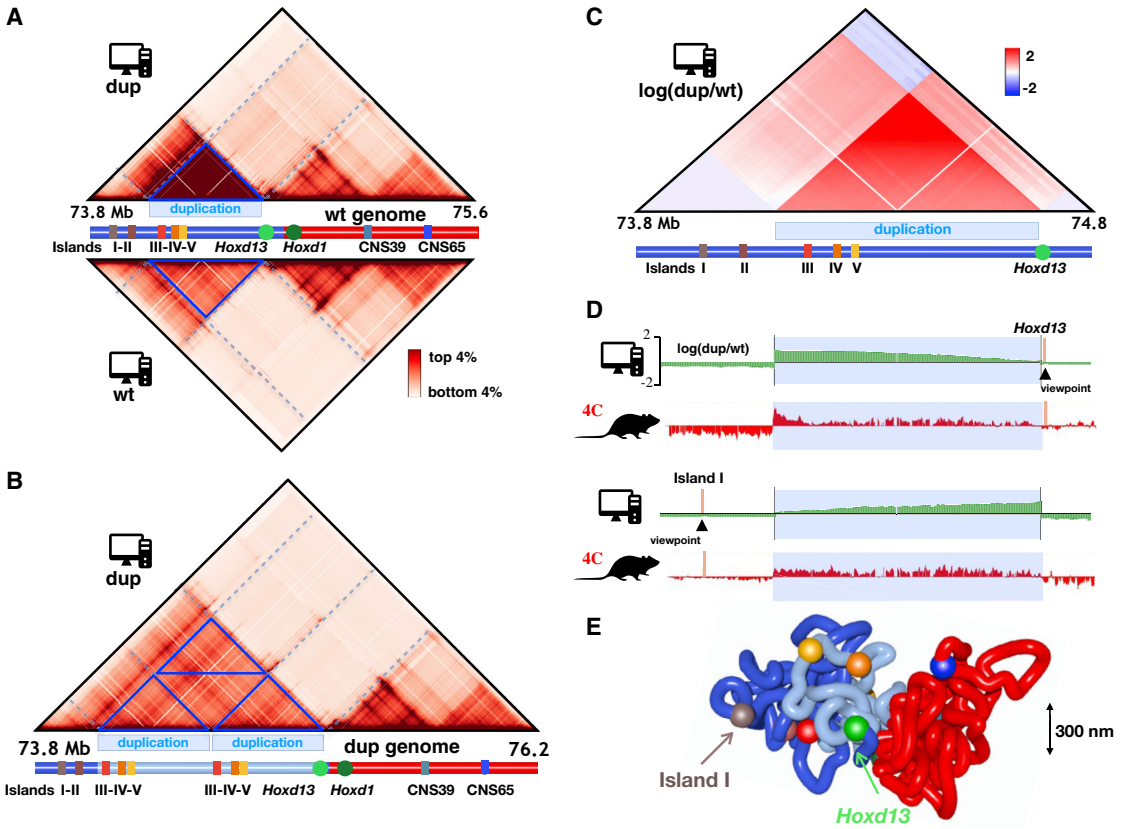


Figure 2. The Model-Predicted Effect of the Nsi-SB Duplication on Folding Is Tested against Published 4C Data

(A and B) As the Nsi-SB duplication is implemented in the high-resolution SBS polymer model of the WT locus in ESCs (Figure 1A, bottom), a pattern of increased contacts is predicted in the contact map (A) of the mutated system (blue triangles). Mapping the contacts on the mutated genome (B) highlights the origin of such pattern of interaction from four different contributions (blue triangles), i.e., interaction of the two identical genomic segments with themselves and with each other. (C–E) The log2 ratio of the mutated and wild-type signals (C) better highlights the interaction changes. Published, higher resolution 4C data (D; Montavon et al., 2012) from the *Hoxd13* and *Island* I viewpoints in murine limb cells (adapted from Montavon et al., 2012) are used to test the model predictions on location and amplitude of interaction changes. A snapshot of the 3D conformations of the locus in the duplicated system is shown in (E) (color scheme in B).

independent experimental 4C data are available in murine limb tissue cells (Montavon et al., 2012). We implemented the duplication in our 5k resolution wild-type (WT) polymer model in ESCs, derived how the locus refolds under only the laws of physics, and compared the predicted contact profiles against 4C data from the two available viewpoints. This is a stringent test of the model because there are no tuning parameters available in the comparison.

The model predicts that at the location of the duplication, an increase in interaction frequency appears with respect to the WT case (Figure 2A), similarly to previous observations in a different cell system (Franke et al., 2016). The origin of such interactions can be dissected in our model by looking at the contacts mapped on the correspondingly mutated genome (Figure 2B). The changes in interaction frequencies are further highlighted using the $\log_2$ ratio of the predicted interaction frequency between the WT and the dup genomes (Figure 2C).

The impact of the duplication on the locus 3D structure can be easily rationalized within our model. In the mutated system, the total self-interaction of the genomic sequence included in the region of duplication is the sum of four different contributions (blue triangles in Figure 2B), i.e., the interaction of the WT original region and of its flanking duplicated region with themselves and with each other (Figure 2B). Hence, by mapping interactions back onto the linear WT genome (Figure 2A), an increase of the signal is expected precisely at the site of the duplication deriving by the addition of the different contributions. That information is unavailable in usual Hi-C data but can be easily extracted with our models. The formation of strips of enhanced interaction between the region of duplication and the rest of the locus (blue dashed lines in Figures 2A and 2B) is explained in the same way as the sum of the two different contributions from the duplicated and original sequences. For example, we found that the interaction frequency of *Hoxd13*, located downstream the duplicated region, with the *Islands III*, *IV*, and *V* is 40% due to its contacts with the centromeric *Island* elements and 60% to contacts with their duplicated telomeric copies. Analogously, the reduction of contacts of *Hoxd13* with the region upstream of the duplication, including in particular its regulatory *Islands I and II*, is related both to the increased genomic distance to it because of the intervening duplicated region and to the increased sequence-specific interactions with

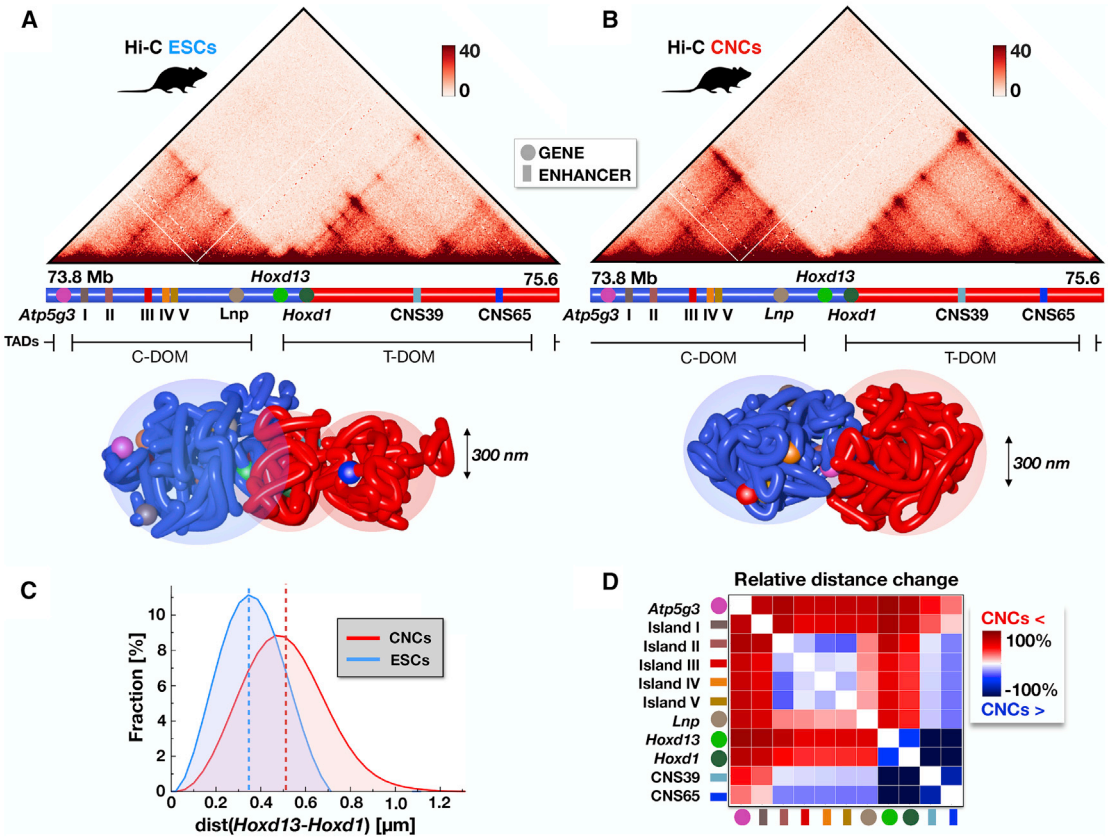


Figure 3. The *HoxD* Region Has a High Cell-to-Cell Variability in ESCs and CNCs

(A and B) Snapshots of the high-resolution model derived 3D single-molecule structure of the restricted *HoxD* region in ESCs (A) and in CNCs (B) help visual its two flanking TADs and their inner structure.

(C and D) The SD to average ratio of the *Hoxd1–Hoxd13* distance distribution is around 30%, highlighting a strong cell-to-cell variability in both ESCs and CNCs (C). The relative average distance change ((ESCs − CNCs)/ESCs) across genes and regulatory elements of the locus (D) shows that the *Hoxd13* and *Hoxd1* genes are around 50% further in CNCs.

the duplicated portion. Importantly, the model predicted scale of the interaction changes produced by the mutation compares well with 4C data (Figures 2C and 2D). The overall agreement gives a validation of the model in a different system, against independent data. Interestingly, that is also a hint that the main binding domains envisaged by the model must be similar in ESCs and limb tissue and roughly maintained in the duplicated region.

In particular, the accuracy of our predictions was tested by using available 4C data from the viewpoints of *Hoxd13* and *Island I* (Figure 2D). The comparison involves different cell types, and Hi-C and 4C data have different resolutions. However, the model predicted and the experimentally tested contact patterns overall match to a good extent. For example, from the *Hoxd13* viewpoint, approximately a 1.5-fold increase of contacts is seen within the region of the duplication, whereas a decrease is found with the region upstream to it, in both model and 4C data. A similar change is also measured from the *Island I* viewpoint. In particular, *Hoxd13* shows reduced contacts with *Islands I* and *II* after duplication in both 4C experiment and model prediction. A 3D snapshot of the conformation of the locus carrying the duplication (Figure 2E, to be compared with a 3D snapshot in WT of Figure 3A) helps visualizing the architectural change. Notably, the Nsi-SB duplication is associated in limbs to a missing phalange in digit II, a phenotype closely resembling the case of a deletion including *Islands I* and *II* (Montavon et al., 2012), suggesting the role of such specific loss of contacts in determining the gene expression level and associated phenotype. Finally, we also checked that very similar results are obtained by implementing the Nsi-SB duplication in our WT ESCs model derived from 40-kb-resolution Hi-C data (Dixon et al., 2012) (Figure S2).

Combined with experiments (Montavon et al., 2012), our results show how the specific 3D organization of the locus is fundamental for the correct regulation of the *HoxD* genes. They are also in line with recent experimental outcomes from different genomic rearrangements at the *HoxD* locus, showing that changes in contact profiles derive from a combination of genomic distance effect, sequence specificity, and TAD boundary relocation (Fabre et al., 2017; Rodríguez-Carballo et al., 2017). Altogether, these results clarify the impact of the studied mutation on the architecture of the *HoxD* region.

The 3D Structure of the Locus and Its Cell-to-Cell Variability

To investigate the variability of the 3D structure at the extended *HoxD* region across different single cells, we explored our ensemble of high-resolution, single-molecule 3D conformations produced by polymer physics at thermodynamic equilibrium (Figures 1C and 1F). To measure the size and variability of the 3D structure around the *HoxD* locus, we recorded the average gyration radius, R_g, of the 2-Mb region encompassing the *HoxD* cluster and its two flanking centromeric (C-DOM) and telomeric (T-DOM) TADs (Figures 3A and 3B), i.e., the radius of its average enclosing sphere. In both cell types, R_g is about 1 μm, with a similar standard deviation of 0.4 μm and 0.5 μm in ESCs and CNCs, respectively.

To characterize the level of cell-to-cell variability of the 3D structure of the *HoxD* cluster, we also measured the distribution of the distances between some key genes and enhancers in the restricted 2-Mb region around the locus (Figures 3A and 3B). As mentioned above, we fixed the average distance between *Hoxd1* and *Hoxd13* equal to 350 nm in ESCs, as found in a previous, independent, measure by FISH (Eskeland et al., 2010). We find, for example, that such distance slightly increases in CNCs to 520 nm (Figure 3C). On the other hand, the average distances between the *HoxD* genes and their regulators in the C-DOM, such as *Islands I–V*, diminishes (Figure 3D). However, the distance distribution between *Hoxd1–Hoxd13* has a high variability, with a standard deviation to average ratio around 30% (SD = 100 nm in ESCs and 170 nm in CNCs), highlighting a strong cell-to-cell variability in both ESCs and CNCs. We also derived cell-to-cell variability of the *Hoxd1–Hoxd13* distance in our 3D models from lower resolution Dixon 2012 data in ESCs and Cortex cells. In ESCs, variability is slightly higher (around 40%), probably because of the lower resolution but also because they are highly dividing cells, whereas ESC data from Bonev et al.,(2017) derive from cell-cycle-staged cells. Interestingly, an even higher population variability is found in Cortex tissue, where the SD to average ratio of *Hoxd1–Hoxd13* is about 60%. The higher population variability of the architecture in Cortex is possibly due to the presence of various cell types, including glia and neurons. We finally compared the distribution of *Hoxd1–d13* distances in ESCs with available FISH data in limb tissue (Fabre et al., 2015) and found that they have a statistically similar shape (Figure S3).

Taken together. our results show that although the poised *HoxD* cluster in ESCs has a 3D structure slightly more compact than in CNCs, the cell-to-cell variability is high in both cell types.

Regulatory Multi-way Contacts

Next, we investigated the combinatorial nature of regulatory interactions in the *HoxD* region, searching for high-multiplicity, many-body contacts. That information is straightforwardly derived within our polymer models, but experimentally it can be currently obtained only at much lower resolution by, say, multi-way 4C or GAM experiments (Allahyar et al., 2018; Oudelaar et al., 2018; Beagrie et al., 2017; Olivares-Chauvet et al., 2016). We find that many-body contacts are abundant in the system and statistically significant with respect to the expected random background (Wilcoxon test p value < 0.01; STAR Methods, Chiariello et al., 2016).

For brevity, we focus on the triplets formed by the promoters and their known regulators in the region, which we find to be strongly gene- and cell-type-specific. Notably, we find that the triplets formed by the *Hoxd13*, *Hoxd9*, and *Hoxd1* genes are almost exclusively restricted to their flanking TADs, showing that such multiple contacts are highly selective, even more than pairwise contacts (Figure 4A). Triplets are also compartmentalized. For instance, in ESCs, the triplets formed by *Hoxd1* and *Hoxd9* are confined mostly to sites in their T-DOM, whereas *Hoxd13* forms non-trivial triplets especially with its C-DOM. In CNCs, instead, weaker many-body contacts are established from the *Hoxd9* and *Hoxd1* genes with both C-DOM and T-DOM, whereas *Hoxd13* strongly interact in triple contacts with its regulatory elements in the C-DOM.

In particular, our model shows that the genes of the locus form specific multiple contacts with their associated enhancer elements and other genes (Figure 4B). In ESCs, we find that the *HoxD* genes display simultaneous triple interactions with each other, which is in agreement with recent single-cell 3-way 4C data (Olivares-Chauvet et al., 2016). In ESCs, we also find triplet interactions between the *HoxD* genes and the *CNS39* regulator downstream and between *Hoxd13* and its regulatory *Islands II–V*. Interestingly, in CNCs we find that triplets between the *HoxD* genes are less frequent. Moreover, in CNCs, *Hoxd13* frequently interacts in triplets with *Lnp* and *Atp5g3* genes and, specifically, with its centromeric regulatory *Island* elements, whereas *Hoxd9* and *Hoxd1* form triplets also with the downstream regulator *CNS39*. To a different extent, in CNCs all the *HoxD* genes gain some interactions with the more distal part of the C-DOM, including *Islands I* and *II* and the *Atp5g3* gene.

Triplet interactions from our ESCs and Cortex tissue low-resolution models have also been derived (Figure S4). As an effect of the lower resolution, in both cell types, triplet contacts appear more spread, but we can still see an effect of compartmentalization in ESCs. In Cortex, conversely, we find that the *HoxD* genes all share broader multiple contacts within a larger metaTAD formed by C-DOM and T-DOM (see Figures S1D and S4), that could be partially due to the presence in Cortex tissue of different cell types, as mentioned before.

Finally, we tested our predicted triplet interactions against available 3-way 4C data (Olivares-Chauvet et al., 2016) from five different viewpoints in ESCs (Figures 5 and S5). Interestingly, model and experiment return similar patterns of triple interactions: for instance, about 70% of the experimental triplets is also detected by the model (Figure 5; STAR Methods).

Our results return a picture where the *HoxD* locus is marked by a complex, cell-type-specific network of high-multiplicity regulatory contacts, where *HoxD* genes interact selectively and combinatorially within their flanking TADs. That could be the mode of action of compartmentalization to fine tune specific gene activity.

DISCUSSION

To shed light on the regulatory interactions occurring at the *HoxD* region and its underlying molecular mechanisms, we used a

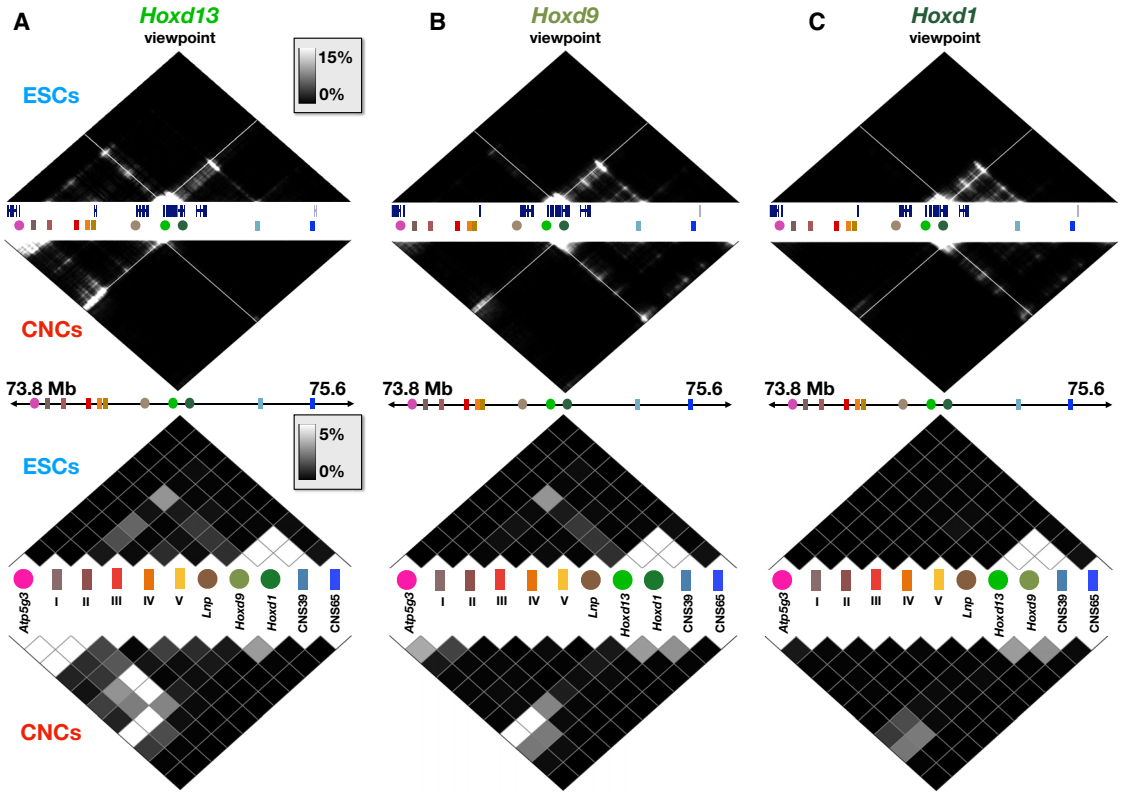


Figure 4. Triple Contact Probabilities of Genes and Regulators at the *HoxD* Region Are Gene and Cell-Type Specific

(A–C) High-resolution-model-derived single-cell triple contact probability from the viewpoint of *Hoxd1*, *Hoxd9*, and *Hoxd13* have a gene- and cell-type-specific compartmentalized structure. *Hoxd13* (A) forms triplets especially with the centromeric TAD, in both ESCs and CNCs. *Hoxd9* (B) and *Hoxd1* (C) form triplets mainly with the telomeric TAD in both ESCs and CNCs. Promoter-specific subset of triplets are formed by genes and regulators within the *HoxD* region (A, B, and C, bottom panels). Such combinatorial interactions could be the mode of action whereby the 3D conformation differentially regulates the genes.

polymer physics approach to reconstruct the 3D structure of the region at a single-molecule level. In particular, we used the SBS model (Nicodemi and Prisco, 2009) that quantifies a well-known scenario of molecular biology where molecules, such as transcription factors, act as chromatin 3D organizers by bridging distal cognate DNA sites in loops. We showed that such a physics model explains with high accuracy available Hi-C data (Dixon et al., 2012; Bonev et al., 2017) and well describes gene distances measured by FISH (Eskeland et al., 2010; Fabre et al., 2015). Its prediction on the 3D organization in cells bearing the 0.5-Mb Nsi-SB duplication was also validated against independent 4C data (Montavon et al., 2012). Among many limitations, one of the advantages of polymer physics is to provide a 3D representation and a principled interpretation of the patterns visible in Hi-C, derived only by the model envisaged basic molecular mechanisms and the laws of physics.

In the emerging scenario, structures ranging from sub- to metaTADs, correspond to regions enriched for contacts between specific types of binding sites mediated by cognate molecules, by a thermodynamic process known in polymer physics as microphase separation (Barbieri et al., 2012; Chiariello et al., 2016; Nicodemi and Prisco, 2009). Interestingly, recent experiments have given evidence that microphase separation is a chromosome organizing mechanism (Hnisz et al., 2017; Larson et al., 2017; Strom et al., 2017). The different binding factors envisaged by the model correlate each with a specific combination of histone marks and molecules such as CTCF, known to be implicated in chromatin organization (see, e.g., Barbieri et al., 2017; Ernst et al., 2011; Ho et al., 2014; Rao et al., 2014; Sanborn et al., 2015), as if an epigenomic combinatorial code underlies chromatin folding. Additional mechanisms can contribute to the folding of the locus, such as those envisaged by the Loop Extrusion model (Brackley et al., 2017; Fudenberg et al., 2016; Sanborn et al., 2015).

We find, in particular, that the architecture of the *HoxD* locus is characterized by a network of specific many-body regulatory contacts, undergoing profound reorganizations in different cell types according to the state of activity of the genes and epigenetics changes. As previously reported (see, e.g., Andrey et al., 2013; Noordermeer et al., 2014), in ESCs the interactions of the poised *HoxD* genes are prevalently located to either the C-DOM or T-DOM. We find that such interactions also involve multiple simultaneous contacts with enhancer elements. For instance, in ESCs *Hoxd13* forms triplets with *Lnp* and its centromeric regulatory *Islands* but also with *Hoxd1*, which does not interact with the *Islands*. In CNCs, the locus architecture is

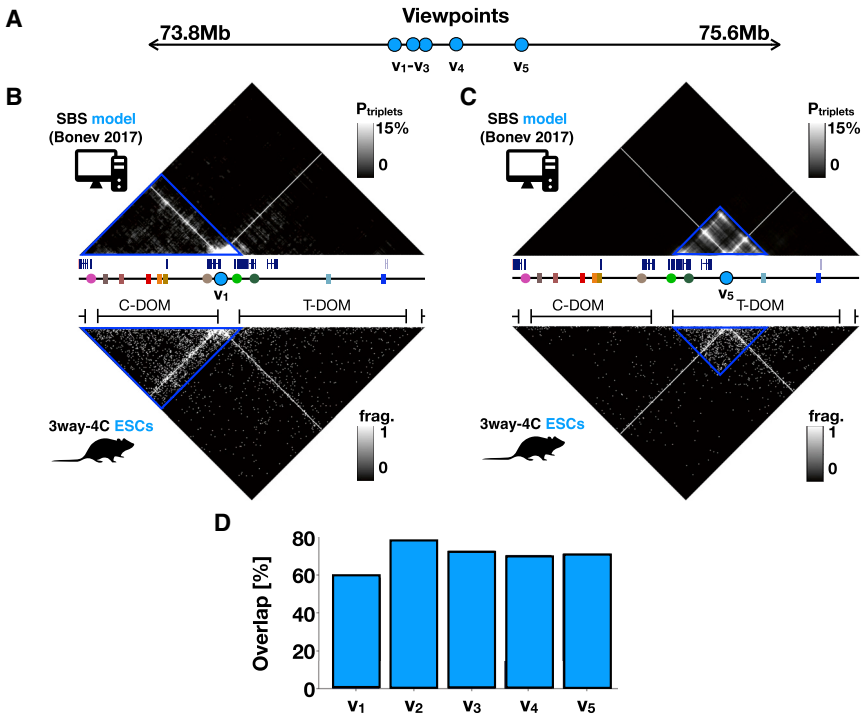


Figure 5. Comparison of Model-Predicted and Experimental Triple Contact Probabilities at the *HoxD* Region in ESCs

(A) Scheme with the 3-way 4C data viewpoints at the *HoxD* locus in ESCs considered in Olivares-Chauvet et al., (2016).

(B–D) The SBS model predicted triplet contact probabilities at 5-kb resolution (B, top) and experimental 3-way 4C data (Olivares-Chauvet et al., 2016, bottom) in ESCs, from viewpoint v_1, internal to the centromeric TAD, and (C) from viewpoint v_5, in the telomeric TAD (results from the other viewpoints are shown in Figure S5). The two datasets have a very high overlap (D) (see STAR Methods).

largely reorganized, as the *HoxD* genes interact less in triplets with each other and all of them gain some triplet interactions with the more centromeric part of the C-DOM, including the *Atp5g3* gene. As a further test of our model, we compared its predictions on triplet interactions against independent 3-way 4C data (Olivares-Chauvet et al., 2016), finding a 70% overlap. The combinatorial nature of high-multiplicity contacts and their gene and cell-type specificity hint toward a key role in the regulation of the transcriptional program of the *HoxD* locus and may explain how architectural compartmentalization finely controls the expression of single genes.

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
- METHOD DETAILS
  - String&Binders Switch model of the HoxD locus
  - Molecular Dynamics Simulations details
  - Pairwise contact matrices and correlations
  - Predicted binding domains
  - Significance of the polymers binding domains
  - Comparison between binding domains and TADs
  - Structural role of the polymers binding domains
  - Correlations of the binding domains with chromatin marks
  - Modeling of the Nsi-SB duplication
  - Cell-to-cell variability
  - Triplets frequencies
  - Polymer 3D representation
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND CODE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.celrep.2019.07.013.

ACKNOWLEDGMENTS

M.N. acknowledges support from NIH grant 1U54DK107977-01; the EU H2020 Marie Curie ITN n.813282; CINECA ISCRA HP10CYFPS5 and HP10CRTY8P; an Einstein BIH Fellowship Award (EVF-BIH-2016-282); Regione Campania SATIN Project 2018–2020; and computer resources from INFN, CINECA, ENEA CRESCO-ENEAGRID (Ponti et al., 2014), and Scope-ReCAS at the University of Naples.

AUTHOR CONTRIBUTIONS

M.N., S.B., and C.A., designed the research project. S.B. and C.A. developed the modeling part; S.B., C.A., A.M.C., A.E., L.F., M.C., and R.C. ran the computer simulations and performed data analyses. G.A. provided conceptual advice. M.N., S.B., C.A., G.A., and A.P. wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 27, 2018
Revised: May 24, 2019
Accepted: July 2, 2019
Published: August 6, 2019

REFERENCES

Allahyar, A., Vermeulen, C., Bouwman, B.A.M., Krijger, P.H.L., Verstegen, M.J.A.M., Geeven, G., van Kranenburg, M., Pieterse, M., Straver, R., Haarhuis, J.H.I., et al. (2018). Enhancer hubs and loop collisions identified from single-allele topologies. Nat. Genet. *50*, 1151–1160.

Allen. Michael, P., and Tildesley, D.J. (1987). Computer Simulation of Liquids (Oxford University Press).

Andrey, G., Montavon, T., Mascrez, B., Gonzalez, F., Noordermeer, D., Leleu, M., Trono, D., Spitz, F., and Duboule, D. (2013). A switch between topological domains underlies HoxD genes collinearity in mouse limbs. Science *340*, 1234167.

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.

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.

Beagrie, R.A., Scialdone, A., Schueler, M., Kraemer, D.C.A., Chotalia, M., Xie, S.Q., Barbieri, M., de Santiago, I., Lavitas, L.M., Branco, M.R., et al. (2017). Complex multi-enhancer contacts captured by genome architecture mapping. Nature *543*, 519–524.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315–326.

Bianco, S., Chiariello, A.M., Annunziatella, C., Esposito, A., and Nicodemi, M. (2017). Predicting chromatin architecture from models of polymer physics. Chromosome Res. *25*, 25–34.

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.

Bonev, B., Mendelson Cohen, N., Szabo, Q., Fritsch, L., Papadopoulos, G.L., Lubling, Y., Xu, X., Lv, X., Hugnot, J.P., Tanay, A., and Cavalli, G. (2017). Multiscale 3D Genome Rewiring during Mouse Neural Development. Cell *171*, 557–572.e24.

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.

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.

Deschamps, J., Akker, E., Forlani, S., De Graaff, W., Oosterveen, T., Roelen, B., and Roelfsema, J. (1999). Initiation, establishment and maintenance of Hox gene expression patterns in the mouse. Int. J. Dev. Biol. *43*, 635–650.

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.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.

Eskeland, R., Leeb, M., Grimes, G.R., Kress, C., Boyle, S., Sproul, D., Gilbert, N., Fan, Y., Skoultchi, A.I., Wutz, A., and Bickmore, W.A. (2010). Ring1B compacts chromatin structure and represses gene expression independent of histone ubiquitination. Mol. Cell *38*, 452–464.

Fabre, P.J., Benke, A., Joye, E., Nguyen Huynh, T.H., Manley, S., and Duboule, D. (2015). Nanoscale spatial organization of the *HoxD* gene cluster in distinct transcriptional states. Proc. Natl. Acad. Sci. USA *112*, 13964–13969.

Fabre, P.J., Leleu, M., Mormann, B.H., Lopez-Delisle, L., Noordermeer, D., Beccari, L., and Duboule, D. (2017). Large scale genomic reorganization of topological domains at the HoxD locus. Genome Biol. *18*, 149.

Feingold, E.A., Good, P.J., Guyer, M.S., Kamholz, S., Liefer, L., Wetterstrand, K., Collins, F.S., Gingeras, T.R., Kampa, D., Sekinger, E.A., et al. (2004). The ENCODE (ENCyclopedia of DNA Elements) Project. Science *306*, 636–640.

Franke, M., Ibrahim, D.M., Andrey, G., Schwarzer, W., Heinrich, V., Schöpflin, 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.

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.

Ho, J.W.K., Jung, Y.L., Liu, T., Alver, B.H., Lee, S., Ikegami, K., Sohn, K.A., Minoda, A., Tolstorukov, M.Y., Appert, A., et al. (2014). Comparative analysis of metazoan chromatin organization. Nature *512*, 449–452.

Kmita, M., and Duboule, D. (2003). Organizing axes in time and space; 25 years of colinear tinkering. Science *301*, 331–333.

Knight, P.A., and Ruiz, D. (2013). A fast algorithm for matrix balancing. IMA J. Numer. Anal. *33*, 1029–1047.

Kremer, K., and Grest, G.S. (1990). Dynamics of entangled linear polymer melts: A molecular-dynamics simulation. J. Chem. Phys. *92*, 5057–5086.

Larson, A.G., Elnatan, D., Keenen, M.M., Trnka, M.J., Johnston, J.B., Burlingame, A.L., Agard, D.A., Redding, S., and Narlikar, G.J. (2017). Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin. Nature *547*, 236–240.

Montavon, T., Thevenet, L., and Duboule, D. (2012). Impact of copy number variations (CNVs) on long-range gene regulation at the HoxD locus. Proc. Natl. Acad. Sci. USA *109*, 20204–20211.

Nicodemi, M., and Prisco, A. (2009). Thermodynamic pathways to genome spatial organization in the cell nucleus. Biophys. J. *96*, 2168–2177.

Noordermeer, D., and Duboule, D. (2013). Chromatin architectures and Hox gene collinearity. Curr. Top. Dev. Biol. *104*, 113–148.

Noordermeer, D., Leleu, M., Splinter, E., Rougemont, J., De Laat, W., and Duboule, D. (2011). The dynamic architecture of Hox gene clusters. Science *334*, 222–225.

Noordermeer, D., Leleu, M., Schorderet, P., Joye, E., Chabaud, F., and Duboule, D. (2014). Temporal dynamics and developmental memory of 3D chromatin architecture at Hox gene loci. eLife *3*, e02557.

Olivares-Chauvet, P., Mukamel, Z., Lifshitz, A., Schwartzman, O., Elkayam, N.O., Lubling, Y., Deikus, G., Sebra, R.P., and Tanay, A. (2016). Capturing pairwise and multi-way chromosomal conformations using chromosomal walks. Nature *540*, 296–300.

Oudelaar, A.M., Davies, J.O.J., Hanssen, L.L.P., Telenius, J.M., Schwessinger, 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.

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 CRESCO4 cluster integrated in the ENEAGRID infrastructure. In Proceedings of the

2014 International Conference on High Performance Computing and Simulation, pp. 1030–1033, 6903807.

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.

Rodríguez-Carballo, E., Lopez-Delisle, L., Zhan, Y., Fabre, P.J., Beccari, L., El-Idrissi, I., Huynh, T.H.N., Ozadam, H., Dekker, J., and Duboule, D. (2017). The *HoxD* cluster is a dynamic and resilient TAD boundary controlling the segregation of antagonistic regulatory landscapes. Genes Dev. *31*, 2264–2281.

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.

Schuettengruber, B., Bourbon, H.M., Di Croce, L., and Cavalli, G. (2017). Genome Regulation by Polycomb and Trithorax: 70 Years and Counting. Cell *171*, 34–57.

Soshnikova, N., and Duboule, D. (2009). Epigenetic temporal control of mouse hox genes in vivo. Science *324*, 1321–1323.

Strom, A.R., Emelyanov, A.V., Mir, M., Fyodorov, D.V., Darzacq, X., and Karpen, G.H. (2017). Phase separation drives heterochromatin domain formation. Nature *547*, 241–245.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Hi-C data for mouse ESCs and CNCs	Bonev et al., 2017	GEO: GSE96107
Hi-C data for mouse ESCs and Cortex	Dixon et al., 2012	GEO: GSE35156
H3K27me3 in ESCs (E14)	ENCODE	ENCFF945LRL
H3K4me1 in ESCs (E14)	ENCODE	ENCFF817CZF
H3K4me3 in ESCs (E14)	ENCODE	ENCFF796LDS
H3K27ac in ESCs (C57BL/6)	ENCODE	ENCFF001XWR
CTCF in ESCs (E14)	ENCODE	ENCFF854IVF
H3K27me3 in C57BL/6 cerebellum	ENCODE	ENCFF001XWB
H3K4me1 in C57BL/6 cortical plate	ENCODE	ENCFF001XWM
H3K4me3 in C57BL/6 cortical plate	ENCODE	ENCFF001XWN
H3K27ac in C57BL/6 cortical plate	ENCODE	ENCFF001XWL
CTCF in C57BL/6 cortical plate	ENCODE	ENCFF001YAA
H3K27me3 in CNCs	Bonev et al., 2017	GEO: GSE96107
H3K4me1 in CNCs	Bonev et al., 2017	GEO: GSE96107
H3K4me3 in CNCs	Bonev et al., 2017	GEO: GSE96107
H3K27ac in CNCs	Bonev et al., 2017	GEO: GSE96107
CTCF in CNCs	Bonev et al., 2017	GEO: GSE96107
Software and Algorithms		
LAMMPS	Plimpton, 1995	https://lammps.sandia.gov
POV-Ray	Persistence of Vision Pty. Ltd. (2004)	http://www.povray.org/

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 (mario.nicodemi@na.infn.it).

METHOD DETAILS

String&Binders Switch model of the HoxD locus

To reconstruct the 3D structure of the *HoxD* locus, we employed the String & Binders Switch (SBS) model (Barbieri et al., 2012; Chiariello et al., 2016; Nicodemi and Prisco, 2009). In the SBS, a chromatin locus is modeled as a Self-Avoiding Walk (SAW) polymer chain of beads, along which a number of binding sites of different types are present, each type interacting with specific diffusing molecular binders. The SBS polymer models of the *HoxD* locus in mouse embryonic stem and differentiated cells, i.e., the optimal arrangement of different types of binding sites along the polymers, have been determined using PRISMR, a previously described machine learning procedure (Bianco et al., 2018; Chiariello et al., 2016), which minimize the distance between experimental Hi-C and model derived contact data of the locus. We employed published Hi-C data from mouse embryonic stem (ESCs) and Cortical Neuronal (CNCs) cells, in a 7Mb wide region around the *HoxD* cluster (chr2:71160000-78160000, mm10), at 5kb resolution, with KR normalization (Knight and Ruiz, 2013). We also employed published 40kb resolution Hi-C data from (Dixon et al., 2012), in mouse ESCs and Cortex tissue. We used Dixon et al. (2012) Hi-C data as released by the authors in their normalized version with combined replicates. The application of our inference procedure resulted in a polymer model including up to n = 20 different types of binding sites in all the studied cell types (Figures 1A and 1D; Figure S1). Finally, in order to derive an ensemble of single cell 3D conformations at equilibrium, Molecular Dynamics (MD) simulations have been performed as previously described (Chiariello et al., 2016; Kremer and Grest, 1990).

Molecular Dynamics Simulations details

In our Molecular Dynamics (MD) simulations of the SBS model, the system of beads and binders evolve according to the Langevin equation (Allen. Michael and Tildesley, 1987), that is numerically solved using the Verlet algorithm implemented within the LAMMPS

software (Plimpton, 1995). The interaction potentials of the system are the standard potentials used in classical polymer physics studies (Kremer and Grest, 1990). The initial states of the polymers are SAW configurations, where binders are randomly distributed within the simulation box (Chiariello et al., 2016). The linear size of the simulation box has been chosen at least as large as two times the gyration radius in the SAW polymer state and periodic boundary conditions have been implemented to minimize finite size effects. Starting from an open conformation, the system evolves up to $5x10^8$ MD timesteps to approach stationary, as verified by plateauing of the gyration radius as a function of time and confirmed by scaling polymer exponents (Chiariello et al., 2016). We derived by MD an ensemble of, at least, 10^2 different configurations for each of the two considered cases.

We sampled the total concentration c of binders from zero to 116nmol/l and the scale of interaction energy between beads and binders equal to $E_{int} \simeq 1k_BT$ and $E_{int} \simeq 8.1k_BT$, which correspond to the coil and globule conformational states respectively, predicted by polymer physics (Chiariello et al., 2016). The size σ of each bead making up the polymer chain has been estimated by imposing that the average distance between *Hoxd13-Hoxd1* genes in ESCs is equal to 350nm, as found from available FISH data (Eskeland et al., 2010), obtaining σ = 40nm.

Pairwise contact matrices and correlations

To compare our 3D modeling results against the experimental Hi-C data (Dixon et al., 2012; Bonev et al., 2017), we computed the average contact matrix from the ensemble of configurations derived by MD approach. To compute the frequency of contact for all the pairs (i, j) of beads, for each 3D conformation we counted how often i and j are in contact. We considered a pair in contact if their physical distance r_{ij} is less than (or equal to) a fixed threshold distance $\lambda\sigma$ (where λ is a dimensionless constant we set equal to $\lambda = 9$) and they are of the same type (Chiariello et al., 2016). We computed separately the average contact matrix for coil and globule states, and next, to take into account heterogeneity effect for cell population, we find the mixture which best describe the locus, minimizing the Pearson correlation *r* with the Hi-C contact matrix (Chiariello et al., 2016). Through that approach, we find a 100% globular state for ESCs and CNCs cells, and a mixture of coil/globule states equal to 66%–34% for ESCs at 40kb resolution and 61%–39% for Cortex cells.

To measure the agreement between the experimental and model data, we calculated the Pearson correlation coefficient, *r,* between Hi-C and the averaged contact matrix obtained from the 3D conformations. Furthermore, as a finer measure to compare experimental and model matrices, we computed a distance-corrected Pearson correlation coefficient, *r'*, where the trivial contribute given by the descending trend of the contact frequency with the genomic distance is subtracted. That is made by correlating the matrices where, from each diagonal, average contact frequency at that corresponding genomic distance has been subtracted (Bianco et al., 2018).

Predicted binding domains

For each SBS polymer model, we define its *binding domains* as the different sets of binding sites of the same type along the polymer. In Figures S1E and S1F, the full set of binding domains identified for the *HoxD* locus, in ES and CN cells, are shown, each represented by a different color. We assigned the same color to pairs of ESCs/CNCs binding domains with a similarity criterion, based on their genomic overlap, *q* (Bianco et al., 2018). For any pair of binding domains, *q* is a positive number given by the sum of products of binding sites abundancies of the two colors in each 5kb genomic window, within the 7 Mb *HoxD* locus. *q* is then normalized so that its maximum value, corresponding to the case of a pair of identical binding domains, is *q* = 1. Hence, we linked in an exclusive manner each of the binding domains in ESCs with the most overlapping domain in CNCs. Analogously, the set of binding domains found from 40kb resolution Hi-C data (Dixon et al., 2102) in ESCs and Cortex cells are shown in Figures S1G and S1H, with the most overlapping pairs in the two cell types are also represented with the same color. It is important to notice, however, that binding sites marked with the same color in two cell types are only linked by their similar arrangement along the linear sequence of the locus and we do not make any assumption about their molecular nature, which can in general be different (see Epigenetics correlations subsection).

Significance of the polymers binding domains

To test the statistical significance of the identified binding domains, in each cell type, we compared them with a control random model obtained by bootstrapping the positions of their binding sites. Specifically, we compared the distributions of the genomic overlaps between pairs of binding domains of our polymer models, with the analogous distribution derived from the random model (made of 1000 different realizations of the randomized polymer). We found for example that the distribution of the overlaps between domains is significantly different from the random overlap distribution in both ESCs (p value = 1.6e-26, Kolmogorov-Smirnov test) and Cortex (p value = 1.1e-101, Kolmogorov-Smirnov test) cases. In particular, the average overlap between ESCs domains is *q* = 21.0% (with standard deviation σ = 15.5%), much smaller than the random control average, q_{rand} = 39.18% (σ_{rand} = 0.03%). Analogously the average overlap between Cortex domains is *q* = 19.6% (σ = 11.9%), against a random control average q_{rand} = 41.17% (σ_{rand} = 0.02%). Similar results are obtained in the other cases.

Comparison between binding domains and TADs

To compare the predicted binding domains with the TAD organization of the *HoxD* locus, we computed their genomic overlaps. We used published TAD coordinates for each cell type (Dixon et al., 2012; Bonev et al., 2017). The overlap between a TAD and a binding

domain is defined as above, where the analogous of binding sites abundancy for a TAD is a signal that is equal to 1 if the TAD covers the considered bin and equal to 0 otherwise (Bianco et al., 2018). The resulting overlaps are showed as heat-maps in Figures S1E–S1H.

Structural role of the polymers binding domains

The arrangement of the different binding domains along the *HoxD* locus shape its pattern of contacts. To visualize that, we showed in a matrix, for each pairwise contact, the color of the most contributing binding domain (Figures 1B and 1E). The contribution of a binding domain to a fixed pairwise contact is defined as the product of counts of its binding sites in the two considered bins.

Correlations of the binding domains with chromatin marks

To investigate the molecular nature of our envisaged binding domains, we compared them with published chromatin features available for the studied cell types. Specifically, Chip-seq data with peak-called for mouse embryonic stem and Cortex cells were downloaded from the ENCODE database (Feingold et al., 2004) for H3K27me3, H3K4me1, H3K4me3, H3K27ac, and CTCF (accession numbers: ENCFF945LRL, ENCFF001XWN, ENCFF817CZF, ENCFF001XWM, ENCFF796LDS, ENCFF001XWB, ENCFF001XWR, ENCFF001XWL, ENCFF854IVF, ENCFF001YAA). Chip-seq data for the same factors in CNCs have been taken from Bonev et al., (2017). To compare the considered features with the polymers binding domains, we calculated the Pearson correlation coefficient between the counts of binding sites of a domain and the counts of called peaks in the corresponding bins (Bianco et al., 2017, 2018). To check the statistical significance of the resulting correlations, we considered the distribution of correlations with chromatin marks for the random control model described above, made of 1000 independent polymer arrangements obtained by bootstrapping the original binding sites (Bianco et al., 2018). Precisely, we considered as significant the correlation values falling within the 1^{nd} or the 4^{rd} quartile of the random control distribution. The resulting epigenetic signatures (Figures S1E–S1H) of the model binding domains well match known functional chromatin states, such as active, poised and repressed states (see e.g., Ho et al., 2014). Interestingly, however, the binding domains have a genomic overlapping, combinatorial organization, lacking in epigenetic segmentation studies, necessary to explain Hi-C.

Modeling of the Nsi-SB duplication

To validate our model, we tested its capability to predict the effects on folding of genomic rearrangements. Specifically, we considered the Dup(*Nsi-SB*) duplication, a previously reported 0.5Mb long duplication upstream *Hoxd13*, that has been shown to cause shortening of digit II and for which 4C-seq data are available in mouse limb tissue (Montavon et al., 2012). We implemented the mutation in our WT polymer model of the *HoxD* locus in ESCs by duplicating the portion of the polymer corresponding to the experimental duplication and deriving the model contact frequency matrix (Bianco et al., 2018; Figure 2B). To compare such contact matrix against the WT matrix, we also represented the contact frequencies mapped on the WT genome ("Dup" matrix, Figure 2A). As in Montavon et al., 2012, to make a fair comparison of WT and Dup contact matrices, we scaled them to equal median intensities and subsequently computed the Dup/WT log2 ratio (Figure 2C). Finally, we compared the predicted contact frequencies with the contact profiles from the two available viewpoints in 4C data, *Hoxd13* and Island-I, by plotting the rows corresponding to the two viewpoints in our predicted Dup/WT log2 ratio matrix (Figure 2D). Similar results are obtained by implementing the Dup(*Nsi-SB*) mutation in our lower resolution WT polymer model in ESCs cells (Figure S2).

Cell-to-cell variability

To investigate the differences between CNCs and ESCs conformations, first we studied the average polymer size for a 2Mb long region including the *HoxD* cluster and its flanking C-DOM and T-DOM (Figures 3A and 3B). As estimation of polymer size we considered the gyration radius, defined as $R_g^2 = \Sigma_{i=1}^{N}(\mathbf{r}_i - \mathbf{r}_{CM})^2/N$, where $\mathbf{r}_i$ and $\mathbf{r}_{CM}$ are the positions of the i-th bead and of the center of mass respectively. The mean values and the standard deviations of R_g are calculated over our ensemble of polymer configurations, with the coil-globule mixture previously estimated from contact matrices comparison (see Pair-wise contact matrices and correlations subsection).

Next, to characterize the 3D variability of the *HoxD* locus, we studied the distribution of distances between *Hoxd13* and *Hoxd1* genes (Figure 3C). Interestingly, we found that the *Hoxd13* - *Hoxd1* distance distribution in ESCs has a similar shape compared to available FISH data in limb tissue (Fabre et al., 2015) (Figure S3).

Additionally, to better describe the differences between the ESCs and CNCs types, we computed the relative changes in physical distance for some genes and regulatory regions (Andrey et al., 2013; Montavon et al., 2012). The relative distances shown in Figure 3D are calculated as the ratio $(d_{ES} - d_{CN})/d_{ES}$.

Triplets frequencies

In order to investigate the 3D structures for CNCs and ESCs and capture additional aspects of their spatial organization, we computed the frequencies of triplet contacts. In our analysis we fixed as point of view *Hoxd13, Hoxd9* and *Hoxd1* genes. For each point of view, labeled with index k, we count a triplet contact if the pairs (i, k), (j, k) and (i, j) are simultaneously in contact, i.e., if their distances r_{ik}, r_{jk} and r_{ij} are all less than (or equal to) a fixed threshold distance $\lambda\sigma$ (here, $\lambda = 9$) and they are all of the same type. Then, we normalized over the total number of possible triplets i-j-k. We did such analysis for coil/globule states separately

and then we averaged over these states as discussed above. To test he statistical significance of the triplets, we compared them with corresponding triplets distributions in SAW state (Wilcoxon test: pvalue < 0.001). Finally, we tested our model predictions about triplet interactions against available 3-way 4C data (Olivares-Chauvet et al., 2016) from five different viewpoints in ESCs (Figures 5 and S5). Precisely, we computed the fraction of triplets detected from the experiment that are correctly captured in our model (overlap, Figure 5D).

Polymer 3D representation

In Figures 1C, 1F, 2E, 3A, and 3B, are shown single typical globule state configurations of the *HoxD* region, where, to better visualize the relative position of the interesting regions we pictured ''coarse grained'' versions of the polymers. We interpolated the coordinates of each bead with a smooth spline curve described mathematically by a third-order polynomial. All the figures are produced with the POV-RAY software (Persistence of Vision Pty. Ltd., 2004).

QUANTIFICATION AND STATISTICAL ANALYSIS

All the statistical tests employed are specified in the text and details provided in the Method Details section. Pearson correlations were used to compare experimental and simulated contact matrices and to compare model binding sites with epigenetic features. One-tailed Wilcoxon's rank-sum tests were applied to check the significance of three-way contacts, while Kolmogorov-Smirnov tests were used to compare the distributions of physical distances between experiments and models.

DATA AND CODE AVAILABILITY

Custom scripts used in the current study are available from the corresponding author on request.