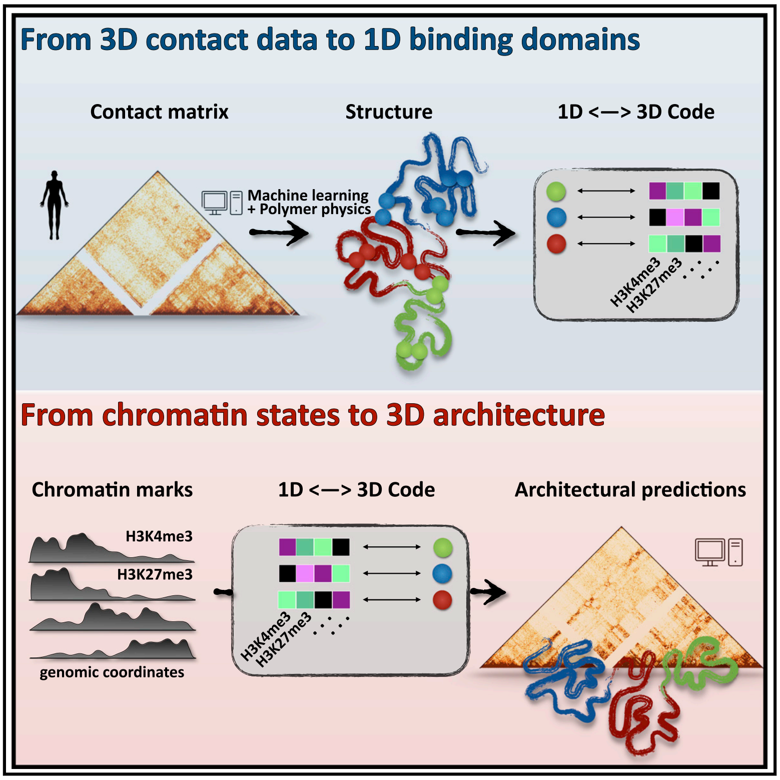
Cell Reports

Article

Polymer physics reveals a combinatorial code linking 3D chromatin architecture to 1D chromatin states

Graphical abstract



Highlights

- Polymer physics identifies a code linking 3D chromatin structure to 1D chromatin states
- The code is shown to be sufficient to explain Hi-C data genome wide
- DNA-DNA contact binding sites have a genomic overlapping, combinatorial organization
- The code is validated by *de novo* predictions of contact maps in wild-types and mutants

Authors

Andrea Esposito, Simona Bianco, Andrea M. Chiariello, ..., Mattia Conte, Raffaele Campanile, Mario Nicodemi

Correspondence

andresposito@na.infn.it (A.E.),
mario.nicodemi@na.infn.it (M.N.)

In brief

Esposito et al. identify a code linking 3D chromatin structure to 1D chromatin states by combining polymer physics and machine learning. The code is sufficient to produce 3D conformations consistent with Hi-C genome wide. That sheds light on how the multitude of specific regulatory DNA contacts is orchestrated by chromatin organizing factors.

Esposito et al., 2022, Cell Reports *38*, 110601
March 29, 2022 © 2022
https://doi.org/10.1016/j.celrep.2022.110601

CellPress

Cell Reports

CellPress
OPEN ACCESS

Article

Polymer physics reveals a combinatorial code linking 3D chromatin architecture to 1D chromatin states

Andrea Esposito,[1,4,*] Simona Bianco,[1,2,4] Andrea M. Chiariello,[1] Alex Abraham,[1] Luca Fiorillo,[1] Mattia Conte,[1] Raffaele Campanile,[1] and Mario Nicodemi[1,2,3,5,*]
[1]Dipartimento di Fisica, Università di Napoli Federico II, and INFN Napoli, Complesso Universitario di Monte Sant'Angelo, 80126 Naples, Italy
[2]Berlin Institute for Medical Systems Biology, Max-Delbrück Centre (MDC) for Molecular Medicine, Berlin, Germany
[3]Berlin Institute of Health (BIH), MDC, Berlin, Germany
[4]These authors contributed equally
[5]Lead contact
*Correspondence: andresposito@na.infn.it (A.E.), mario.nicodemi@na.infn.it (M.N.)
https://doi.org/10.1016/j.celrep.2022.110601

SUMMARY

The mammalian genome has a complex, functional 3D organization. However, it remains largely unknown how DNA contacts are orchestrated by chromatin organizers. Here, we infer from only Hi-C the cell-type-specific arrangement of DNA binding sites sufficient to recapitulate, through polymer physics, contact patterns genome wide. Our model is validated by its predictions in a set of duplications at *Sox9* against available independent data. The binding site types fall in classes that well match chromatin states from segmentation studies, yet they have an overlapping, combinatorial organization along chromosomes necessary to accurately explain contact specificity. The chromatin signatures of the binding site types return a code linking chromatin states to 3D architecture. The code is validated by extensive *de novo* predictions of Hi-C maps in an independent set of chromosomes. Overall, our results shed light on how 3D information is encrypted in 1D chromatin via the specific combinatorial arrangement of binding sites.

INTRODUCTION

The genome of higher organisms has a complex spatial organization within the cell nucleus (Bickmore and Van Steensel, 2013; Dekker and Heard, 2015; Dekker and Mirny, 2016; Dixon et al., 2016; Misteli, 2007; Sexton and Cavalli, 2015), as revealed by recent technologies such as Hi-C (Beagrie et al., 2017; Bintu et al., 2018; Boettiger et al., 2016; Cattoni et al., 2017; Finn et al., 2019; Lieberman-Aiden et al., 2009; Quinodoz et al., 2018). Chromosomes are folded in a sequence of Mb-sized domains enriched in self-contacts, named topologically associating domains (TADs) (Dixon et al., 2012; Nora et al., 2012), some of which show strong loop interactions at their borders, referred to as loop domains (Rao et al., 2017). Further chromatin structures include sub-TADs and larger domains such as A/B compartments (Lieberman-Aiden et al., 2009) and meta-TADs (Fraser et al., 2015). Importantly, such an organization serves vital functional purposes; for instance, distal enhancers control their target genes by establishing physical contacts with them, disruptions being linked to human diseases (Krijger and De Laat, 2016; Oudelaar et al., 2017; Spielmann et al., 2018). However, it remains largely unknown how the multitude of specific DNA contacts, e.g., between transcribed and regulatory regions, is orchestrated by epigenetic signals and chromatin-organizing molecules such as transcription factors (TFs).

To rationalize the complexity of Hi-C data, polymer models from statistical physics (Barbieri et al., 2012; Bohn and Heermann, 2010; Brackley et al., 2013, 2016a, 2017; Chiariello et al., 2016; Fudenberg et al., 2016; Jost et al., 2014; Nicodemi and Pombo, 2014; Nicodemi and Prisco, 2009; Di Pierro et al., 2016; Sanborn et al., 2015; Di Stefano et al., 2016) and a variety of computational methods (Li et al., 2017; Lin et al., 2018; Nir et al., 2018; Serra et al., 2017) have been developed. A class of models, such as the Strings and Binders (SBS) model (Nicodemi and Prisco, 2009), has focused on the classical scenario where loops and contacts between distal DNA sites are established by diffusing molecules such as TFs, or by some effective interaction potential, bridging cognate binding sites by thermodynamic mechanisms of phase separation (Barbieri et al., 2012; Bianco et al., 2019; Bohn and Heermann, 2010; Brackley et al., 2013, 2016a, 2016b; Chiariello et al., 2016, 2020; Conte et al., 2020; Jost et al., 2014; Nicodemi and Prisco, 2009; Di Pierro et al., 2016; Di Stefano et al., 2016). Another interesting classical scenario has been considered by off-equilibrium polymer models where loops are formed by extrusion, e.g., by molecules that bind to DNA and extrude a loop (Brackley et al., 2017; Fudenberg et al., 2016; Sanborn et al., 2015), based on prior knowledge of the involved molecular factors, such as CTCF binding sites. Additionally, computational methods have been introduced for deriving the genome 3D architecture from DNA sequence and epigenomic data, independently of the underlying physical processes (Fudenberg et al., 2020; Qi and Zhang, 2019; Schwessinger et al., 2020; Zhang et al., 2019).

Here, we use a previously developed machine-learning approach (PRISMR [Bianco et al., 2018]) that infers from only Hi-C data of a given genomic region the minimal set of binding sites best explaining its contact patterns through the molecular mechanisms envisaged by the SBS polymer model. PRISMR was originally applied to Mb-wide genomic regions. Here, we extend it to explain high-resolution Hi-C data genome-wide in human and mouse cell types, improving the statistical power of our method by three orders of magnitude. That provides the base to identify the location and combination of the putative binding sites underlying chromatin physical contacts and to derive a first characterization of their molecular features, thus returning a code linking genome-wide architecture, epigenetics (i.e., chromatin states), and function.

To validate our approach, we first show that the SBS polymer model informed with the inferred binding sites recapitulates 5 kb resolution *in situ* Hi-C data in human cells (Rao et al., 2014) and 40 kb resolution Hi-C data in murine cells (Dixon et al., 2012) with high accuracy, illustrating that its minimal ingredients are sufficient to make sense of a substantial fraction of contact patterns genome-wide. For the sake of simplicity, we focus on the SBS model, but the method can be extended to accommodate additional mechanisms, such as loop extrusion (Brackley et al., 2017; Fudenberg et al., 2016; Sanborn et al., 2015).

Next, we validate the model by comparing its predictions about the impact of mutations on chromosome conformation against independent experimental capture Hi-C (cHi-C) data at the *Sox9* locus (Franke et al., 2016), providing insights into how those distinct mutations produce different 3D structures (e.g., neo-TADs) and different enhancer hijackings, resulting in different phenotypes.

Importantly, we find that the model's different binding domains fall in similar classes, well matching functional chromatin states derived in linear epigenetic segmentation studies (Boettiger et al., 2016; Ernst et al., 2011; Gifford et al., 2013; Ho et al., 2014; Javierre et al., 2016). However, we discover that they have an overlapping, combinatorial genomic distribution, lacking in linear segmentations, required to explain Hi-C contacts with high accuracy genome-wide. Finally, we employed the identified code linking architecture and epigenetics to successfully predict *de novo*, from only histone marks, the contact matrices of independent chromosomes, as validated by distinct Hi-C data.

Overall, our results provide insights on how the 1D combinatorial arrangement of a comparatively small number of binding site types, barcoded by distinctive epigenetic signatures, encodes the architectural information guiding chromatin-organizing factors to form specific 3D contacts across chromosomal scales.

RESULTS

Distinct, yet genomically overlapping binding domains explain Hi-C data genome-wide in a cell type-specific manner

To dissect the molecular mechanisms that contribute to chromatin folding, we used the PRISMR machine learning procedure (Bianco et al., 2018) to infer the minimal SBS polymer model best explaining Hi-C contact maps across chromosomes in human and murine cells (Figure 1A; see STAR Methods). In the SBS model (Nicodemi and Prisco, 2009), a chromatin filament is modeled as a self-avoiding string of beads, including specific binding sites for diffusing molecules (here named binders). The motion of beads and binders is subject to polymer thermodynamics. The binders can bridge distal cognate sites along the sequence via classical interaction potentials (Kremer and Grest, 1990), thus producing loops and physical contacts (Figure 1B). In particular, in the SBS model, contact domains of homologous sites are spontaneously established by their cognate binders via a thermodynamic mechanism known as polymer microphase separation (Barbieri et al., 2012; Chiariello et al., 2016; Conte et al., 2020).

The PRISMR procedure learns from experimental Hi-C contact data the minimal number of binding sites, their location along the polymer chain, and their different types (visually represented by different colors, Figures 1A and 1B) that best reproduce the input data through polymer thermodynamics (see STAR Methods). Below, the set of all binding sites of a given type (color) along the polymer chain is named binding domain. Importantly, PRISMR uses just Hi-C data as input, with no prior knowledge of binding factors.

We applied PRISMR first to *in situ* Hi-C data on human GM12878 B-lymphoblastoid cells at 5 kb resolution (Rao et al., 2014) (Figure 1C). We derived the PRISMR contact matrices at 5 kb resolution and compared them with Hi-C data across chromosomes (Figures 1E and S1A) to check whether the model inferred binding domains (Figure 1D) could explain Hi-C contacts genome-wide. We computed their Pearson correlation coefficient r, their distance-corrected Pearson correlation coefficient r′, and their HiCRep stratum-adjusted correlation coefficient SCC (Yang et al., 2017). The last two measures, in particular, also account for genomic proximity effects (see STAR Methods). Model and experimental data were found to be comparatively similar across the set of entire chromosomes, as r, r′, and SCC range around r = 0.94, r′ = 0.74, and SCC = 0.86, respectively (Figure S1B). Importantly, PRISMR captures Hi-C contact patterns not only at large chromosomal scales (Figures 1C–1E) but also at shorter scales, i.e., at the TAD and sub-TAD levels (Figures 1F, S1C, and S1D). Notably, from the SBS model, the thermodynamics ensemble of chromosomal 3D conformations can also be derived; a snapshot, e.g., of a single-molecule conformation of chromosome 20 is pictured in Figure 1G.

Additionally, to prove the general validity of the method, we tested its performance on a mouse embryonic stem cell (mESC) Hi-C dataset at 40 kb resolution (Dixon et al., 2012), finding that the PRISMR inferred and experimental contact matrices have high correlation values across chromosomes, comparable to those reported above for the 5 kb human data (Figures S2A and S2B). The overall features of the binding domains in mESC are similar to those of human GM12878 cells, but the details of their arrangement along the sequence is cell type specific.

The model binding domains (colors) are the output of PRISMR. The algorithm returns 30 different binding domains per chromosome in GM12878 cells (Figure 1D; see STAR Methods).

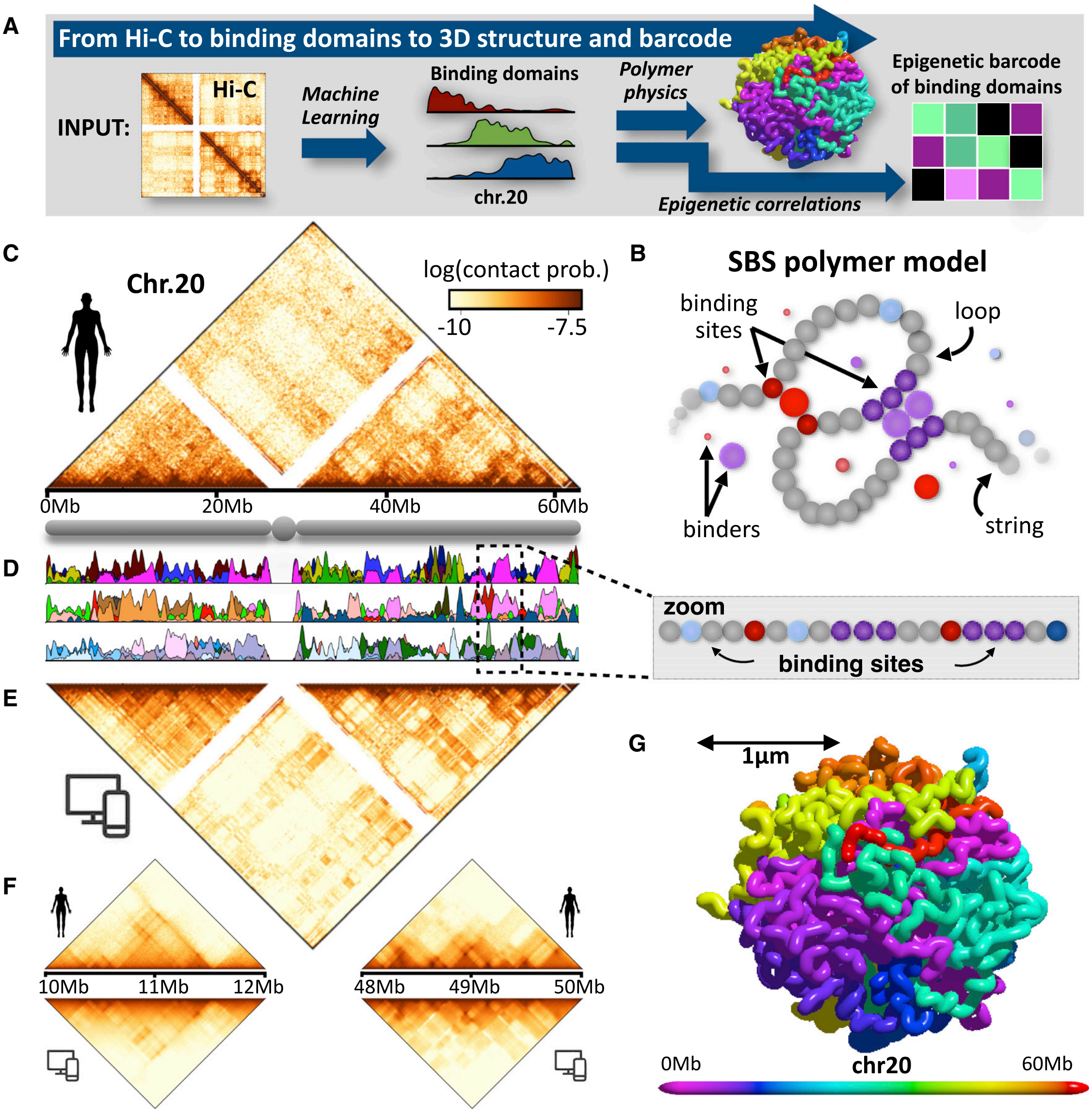


Figure 1. Distinct yet genomically overlapping binding domains explain Hi-C data across chromosomes in a cell-type-specific manner

(A) Our method combines machine learning and polymer physics to infer from only Hi-C data the genomic location of the minimal set of binding sites required to recapitulate chromatin conformations genome-wide by use of the SBS polymer model of chromatin. Additionally, by correlations with epigenetic data, the inferred binding domains can be assigned a molecular barcode.

(B) Scheme of the SBS polymer model of chromatin: it quantifies the scenario where diffusing binders bridge and loop distal cognate binding sites. Each colored bead is a single binding site. The genomic location of the binding sites encodes the 1D information whereby their cognate binders produce the 3D structure via polymer physics.

(C) *In situ* Hi-C data (Rao et al., 2014) of the entire chromosome 20 at 5 kb resolution in human GM12878 cells.

(D) Plots displaying the position and abundance of the different types of binding sites (binding domains) along chromosome 20, as inferred by our method. For visualization purposes, the different domains, each represented by a different color, are drawn in groups of 10 in different rows. Albeit derived from only Hi-C data, the binding domains have specific correlations each with a set of epigenetic marks, and the colors reflect those associations (see Figure 3).

(E) The model-inferred contact matrix of chromosome 20 has a Pearson, distance-corrected Pearson, and stratum-adjusted correlation with Hi-C respectively equal to r = 0.97, r' = 0.85, and SCC = 0.92. Similar results are found across chromosomes (Figure S1A) and in murine cells (Figure S2A).

(F) Comparison of Hi-C (top triangle) and model (bottom triangle) contacts in two 2-Mb-wide genomic regions along chromosome 20.

(G) Time snapshot of the 3D structure of the SBS model of chromosome 20.

Interestingly, we find that the binding domains arrangement along a chromosome is highly non-trivial: the different types of binding sites do not simply occupy separate, contiguous regions but are spread across the whole chromosome and overlap each other (Figure 1D). In particular, although a single binding domain includes on average the equivalent of a genomic length

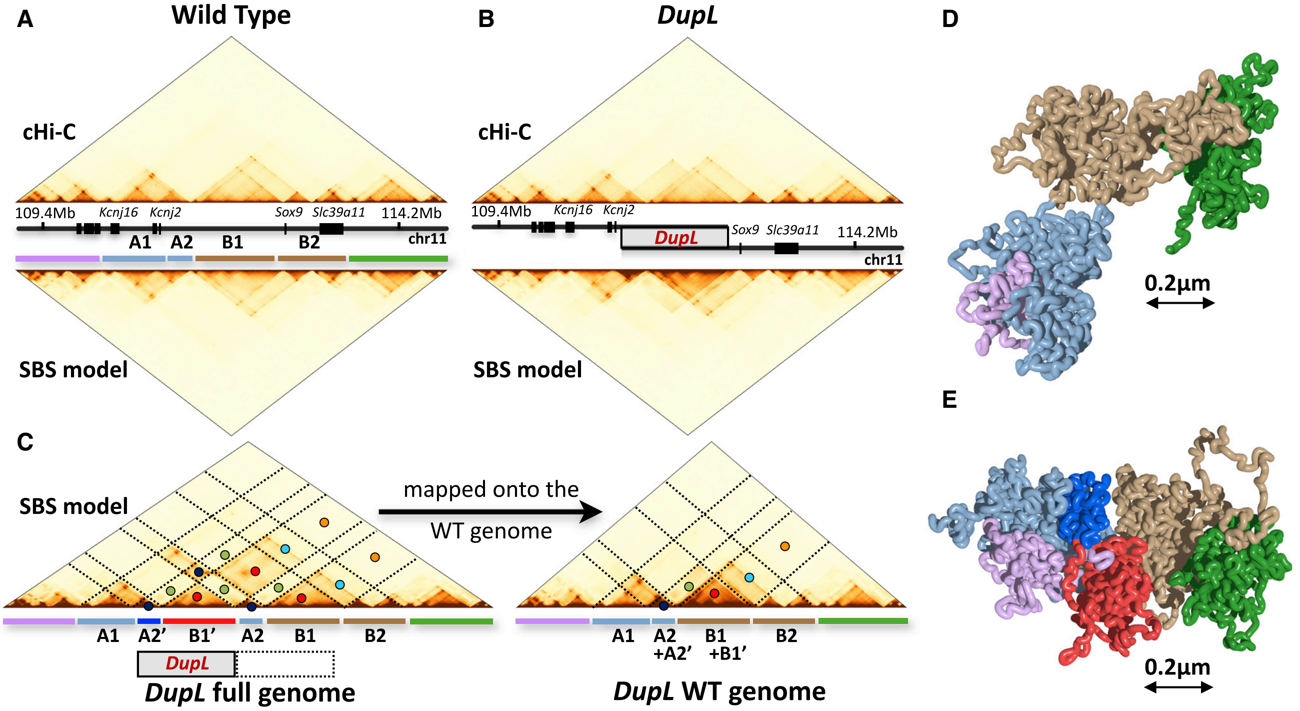


Figure 2. The inferred binding domains are validated against mutations at the *Sox9* locus

(A) Contact data (Franke et al., 2016) of the wild-type *Sox9* locus from cHi-C experiments in E12.5 limb buds (top) and of the corresponding SBS model (bottom) have a correlation r = 0.97 and r′ = 0.87.

(B) Based on the wild-type (WT) model, the contact map of a mutant bearing the *DupL* duplication is predicted from polymer physics (bottom). It has a good correlation (r = 0.92, r′ = 0.63) with independent *DupL* cHi-C data (top [Franke et al., 2016]). Model predictions are also validated across the other available *Sox9* mutations (Figure S3).

(C) Mapping the model contacts on the *DupL* full genome clarifies the origin of the associated neo-TAD (red). The colored circles mark corresponding interaction regions as mapped on the WT and *DupL* full genomes.

(D and E) Snapshots of the model-predicted 3D conformation of, respectively, the WT and *DupL* locus (the color scheme reflects the colored bars in (A) and (C) with its neo-TAD. Different mutations result in different 3D structures, and distinct enhancer-hijackings, explaining their phenotypes (Figure S3).

of 3.1 ± 1.9 Mb, it covers genomic extensions r_{Int}, more than one order of magnitude longer, up to tens of Mbs (Figure S2C; STAR Methods), hence capturing contacts occurring up to chromosomal scales. The distribution of r_{Int}, $P(r_{Int})$, is significantly different from a random control model obtained by bootstrapping the location of binding sites and is asymptotically consistent with a power-law scaling, $P \sim 1/r_{Int}$, typical of hierarchical structures made of domains within domains, as in Cantor sets (Figure S2C; STAR Methods). The broad range of values of r_{Int} shows that chromatin interactions extend above the size of single TADs, with higher-order 3D structures formed at scales below and above the A/B compartment level (Fraser et al., 2015). The derived 3D structures of chromosomes (Figure 1G) shows indeed that, rather than being a linear chain of TADs, they tend to fold on themselves in complex structures, such as meta-TADs (Fraser et al., 2015).

Taken together, the high correlations found between the SBS model and the Hi-C contact data show that the 1D binding domains inferred by PRISMR contain information sufficient to recapitulate 3D contact patterns genome-wide in human and mouse cells. That sheds light on the molecular mechanisms shaping chromosome architecture, supporting the view that the combinatorial action of a comparatively small number of TFs, mediating the interactions between cognate binding sites, can spontaneously fold chromatin into its 3D structure.

Validation of the inferred binding domains against duplications in the *Sox9* locus

To validate the binding domains inferred by our approach, i.e., the determinants of folding and their envisaged mode of action, we compared our model predictions against previous independently produced cHi-C data in E12.5 limb bud cells from mice carrying homozygous structural variants in the *Sox9* locus (Franke et al., 2016). We considered three mutations (Figures 2, and S3A–S3C): a 0.4-Mb duplication (*DupS*) in the non-coding DNA region within the *Sox9* gene TAD (intra-TAD duplication), associated with female to male sex reversal in humans; a 1.6-Mb duplication (*DupL*) that encompasses the neighboring TAD boundary with no phenotypic effect; and a slightly longer, 1.8-Mb duplication (*DupC*), associated with limb malformation, which also includes *Kcnj2*, the next flanking gene. Specifically, we implemented those mutations in the SBS polymer model of the wild-type region in limb buds inferred by PRISMR and derived the novel contact matrices from polymer physics with no fitting parameters whatsoever. We found, in agreement with the experiments, that a separate chromatin domain (termed a "neo-TAD" [Franke et al., 2016]) arises in the inter-TAD duplications *DupL* and *DupC*, whereas the intra-TAD duplication *DupS* does not affect the overall TAD structure. The Pearson and distance-corrected Pearson correlation coefficients between the model predicted and cHi-C contact maps across the three mutations are as high as $r = 0.95$ and $r' = 0.76$, reflecting their good degree of similarity (Figure 2, and S3A–S3C).

Those results provide a stringent validation to our approach and demonstrate that predictions on the 3D structure of chromatin based on the inferred binding domains can be accurate to the point of anticipating ectopic contacts produced by disease-associated mutations.

Mutation-specific enhancer hijackings within *Sox9* neo-TADs link to different phenotypes

To understand the origin of the ectopic contacts in the mutated systems, within our model we dissected the interactions of the duplicated from the original DNA sequences and the corresponding 3D structures, information inaccessible via Hi-C data (Figures 2, S3D, and, S3E).

DupS is fully included within the TAD encompassing *Sox9* (Figure S3A). Within our model, a TAD and its corresponding enrichment of interactions derive from the presence of a prevailing type of binding sites in that DNA region. Hence, the duplicated and the original sequence in *DupS* (region B2′ and B2 in Figure S3D) share many homologous binding sites, which produce the contacts between such regions visible in the interaction matrix mapped along the full, duplicated genome (Figures S3A–S3D). When those contacts are mapped back onto the wild-type sequence, an excess of interactions appears localized around the mutated region within the corresponding TAD, but no major changes to the overall contact pattern, as experimentally found in cHi-C data (Franke et al., 2016). The model-derived 3D structure of the mutated locus shows, indeed, that the duplicated region remains well embedded into the rest of the locus (Figure S3F).

Conversely, in *DupL*, the duplicated region spans two TADs (Figure 2). In our model, those TADs are produced by different prevailing types of binding sites. Accordingly, the portion of the duplication within the *Sox9* TAD (region B1′ in Figure 2C) has enriched contacts with itself and its corresponding original sequence (B1) but less with the portion of the duplication within the flanking TAD and its original sequence (regions A2′ and A2, respectively). Since B1′ is enriched in self-contacts but has comparatively less interaction with its neighboring genomic regions A2′ and A2, it forms a neo-TAD, remaining partially isolated from the rest, as seen in a snapshot of the 3D structure of the locus (Figure 2E, red region). Since the isolated neo-TAD does not include main genes, *DupL* has no phenotype (Franke et al., 2016).

Finally, *DupC* produces a neo-TAD, much as *DupL*; however, it now includes a copy of the next flanking gene, *Kcnj2* (Figure S3C). As seen in the contact matrix of the full genome, within the neo-TAD the duplicated *Kcnj2* establishes ectopic contacts with the duplicated part of the regulatory region of *Sox9*. So, *Kcnj2* is mis-expressed, leading to the associated phenotype (Franke et al., 2016).

In brief, our findings clarify how mutations impact chromatin architecture and the mode of action of the 3D structure in regulating gene activity. In particular, they explain how the considered structural genomic variations at the *Sox9* locus differently alter 3D conformation and gene regulation by specific enhancer hijackings, resulting in distinct phenotypes.

Histone mark profiles of binding domains provide a code linking architecture to epigenetics

To shed light on the nature of the model inferred binding sites (Figure 1D), we correlated their genomic locations with histone mark tracks available in the ENCODE database (Dunham et al., 2012) for the GM12878 cell line (Figure 3). We employed the binding domains derived from even-numbered chromosomes to compute such correlations, in order to later use the derived barcode linking binding site types and epigenetics to predict independently the architecture of odd-numbered chromosomes. In our analysis, we retained only statistically significant correlation values, i.e., those above a random control model, with sites having bootstrapped genomic positions (STAR Methods). We find, across chromosomes, that each binding domain correlates with a specific combination of different epigenetic factors rather than with a single one (Figure S4A). Next, since the different binding domains tend to fall into groups with similar epigenetic profiles, we performed a hierarchical clustering to identify genome-wide significantly distinct epigenetic classes (Figure S4A STAR Methods). By use of the Akaike information criterion (AIC) (Akaike, 1974), we derived that there are nine statistically different groups (Figures 3A and S4B).

Three classes of binding domains strongly correlate with active chromatin marks (Figure 3A), but they are distinct from an epigenetic point of view. Whereas class 1 is enriched for only active marks, classes 2 and 3 are both enriched also in H3K9me3. Also, class 3 shows a stronger correlation with H3K4me1 compared with class 2, a histone mark associated especially with active enhancer regions (Boettiger et al., 2016; Ernst et al., 2011; Gifford et al., 2013; Ho et al., 2014; Javierre et al., 2016). Interestingly, the genomic positions of the sites of the first three classes (Figure 3B) are partially correlated with each other (Figure S4F; STAR Methods). Their histone signatures are also consistent with DNA accessibility, early replication time, and RNA-seq transcription data (Figure S4C). That supports the view that the binding sites in classes 1, 2, and 3 are responsible and genome-wide, especially for specific contacts between transcribed and regulatory regions, mediated by factors such as active Pol-II, as experimentally demonstrated at a number of loci (Barbieri et al., 2017). Class 4 has the typical signature of bivalent chromatin, with H3K27me3 combined with active marks. Its binding sites could be responsible for interactions between regions including, for instance, poised genes and their regulators, as seen in FISH co-localization experiments (Barbieri et al., 2017). Classes 5 and 6 are significantly correlated with H3K27me3 and, for example, could be responsible for the experimentally observed self-interacting domains of polycomb repressive complex (PRC)-repressed chromatin (Kundu et al., 2017). Interestingly, the first six classes all correlate with CCCTC-binding factor (CTCF) binding sites (Figure S4C), but not the remaining classes. That confirms the significance of CTCF in regulating chromatin architecture and gene activity (see, e.g., Tang et al.,

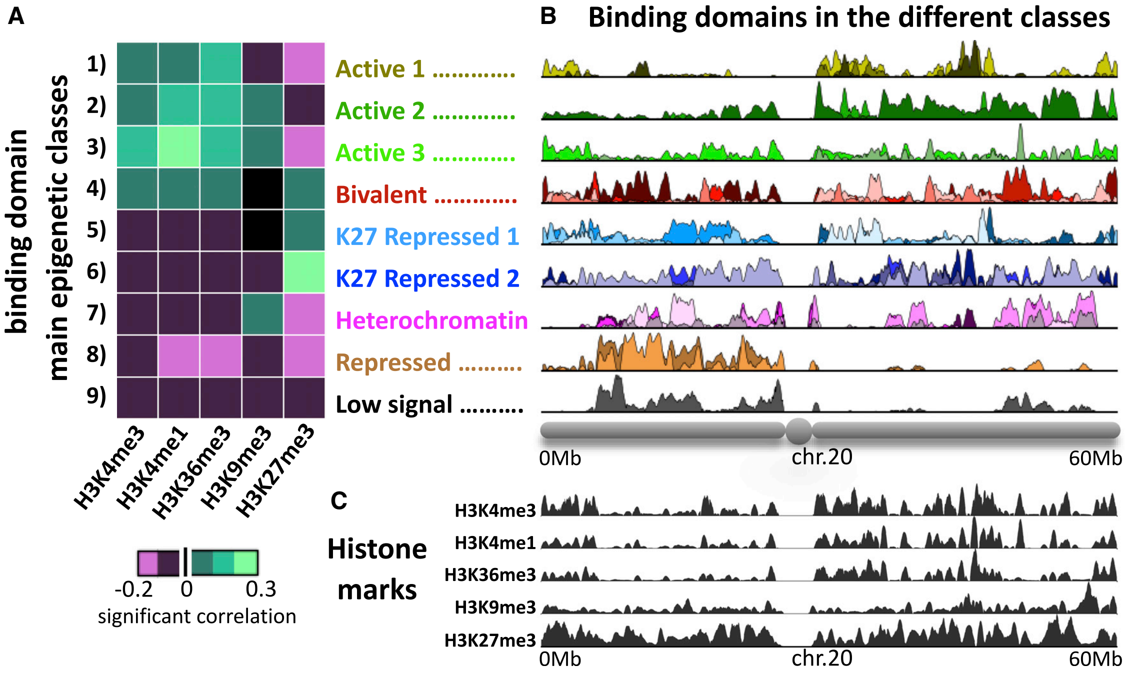


Figure 3. Histone mark profiles of binding domains provide a code linking architecture to epigenetics

(A) The model binding domains, inferred from Hi-C data only, correlate each with a specific set of epigenetic tracks. They cluster in nine main classes genome-wide according to their correlations with the ENCODE key histone marks shown (Figure S4). The epigenetic profile, i.e., the barcode, of the centroid of each class is shown in the heatmap. The nine classes match well the chromatin states derived in epigenetic segmentation studies.

(B) Binding site abundance along chromosomes is not uniform ($p < 0.05$), as shown here for chromosome 20.

(C) Profile of histone marks along chromosome 20.

2015), pointing out that its role can be modulated by different sets of histone marks and molecular factors.

Classes 7 and 8 display a lack of active marks, but whereas class 8 does not correlate with any of the used histone marks, class 7 shows a correlation with H3K9me3, a mark usually associated with constitutive heterochromatin and lack of TF binding, explaining the tendency of heterochromatin regions to cluster in space. Finally, class 9 (named “low signal”) has a very low correlation with available histone marks. However, consistently with previous studies (Boettiger et al., 2016; Ernst et al., 2011; Gifford et al., 2013; Ho et al., 2014; Javierre et al., 2016), it covers almost 15% of the genome, whereas the other classes range from around 2% to 10% in genomic coverage (Figure S4D). Interestingly, the different classes are significantly differently enriched over the different chromosomes and are not consistent with a uniform random genomic distribution (Figure S4E; $p < 0.05$; STAR Methods).

To understand the relative importance of the different types of binding domains in shaping chromatin architecture, we conducted a set of *in silico* experiments with mutant models where each class, one at the time, was erased. Specifically, from the wild-type chromosome models we removed the binding domains of a given class. Next, we computed the contact maps of the mutated model and measured across chromosomes the variation of the Pearson r and distance-corrected Pearson correlation coefficient r′ between the mutated model and wild-type Hi-C contact map (STAR Methods). The variation was found to be proportional to the genomic coverage of the different classes in both cases (Figures S4G and S4H). That implies that no binding class has a special role in holding the architecture of the genome in place. The linear relation whereby the removal of, say, 10% of binding sites genome-wide results in a roughly 10% reduction of r highlights the structural stability of the system: the removal of a small fraction of binding sites proportionally alters the structure but does not produce a sudden collapse of the architecture.

Finally, as a control of the robustness of the association between binding site types and epigenetics, we applied the same approach to the aforementioned mESCs (Dixon et al., 2012), using the corresponding set of ENCODE histone modifications in mouse, and found an overall analogous classification (Figure S5).

Summarizing, the inferred binding site types have each a specific epigenetic barcode falling in classes that match well those found by previous epigenetic genome segmentation studies (Boettiger et al., 2016; Ernst et al., 2011; Gifford et al., 2013; Ho et al., 2014; Javierre et al., 2016). However, our binding domains are inferred from only Hi-C data without prior knowledge of epigenetics; hence, they bring together independent information on architecture and epigenetics. A crucial feature of the model binding domains to explain contact data is that the different types do overlap with each other along the genome at the resolution of the considered Hi-C data. Therefore, they naturally provide each DNA window with a distinctive set of binding site types. This is an important difference with 1D epigenetic segmentation classes: by definition, those have no genomic overlap; thus, each DNA window is associated with only one of such classes. Epigenetic segmentations have been shown, though, to correlate with Hi-C contacts (Ho et al., 2014; Jost et al., 2014; Di Pierro et al., 2016).

Epigenetic linear segmentation only partially captures chromatin folding

To deepen our comprehension of the interplay of chromosome epigenetics and folding, we investigated the architectural information content retained in 1D epigenetic segmentations of the genome and compared it with the DNA barcoding given by the classes of our binding domains (Figure 4). As done in previous studies (Boettiger et al., 2016; Ernst et al., 2011; Gifford et al., 2013; Ho et al., 2014; Javierre et al., 2016), we segmented chromosomes in nine epigenetic classes based only on ENCODE histone marks (Figure S6A). For simplicity, we opted for nine classes to match the number of the different types of binding domains found above (which is comparable to those in previous segmentation studies). Next, we derived *in silico* the contact maps predicted by a polymer model based only on such a 1D epigenetic segmentation. Specifically, we considered a polymer model where chromatin physical interactions occur only between homologous 1D-segmented epigenetic regions. Interestingly, although the overall contact patterns from such a model visually resemble Hi-C patterns, their distance-corrected Pearson correlation r′ with Hi-C data is low: the mean value of correlations across chromosomes are r = 0.79 and r′ = 0.17 (e.g., for chromosome 20, r = 0.80, r′ = 0.21; Figure S6F). To check that our results were not affected by more complex choices of segmentation, we also considered the established ChromHMM (Ernst and Kellis, 2012; Kundaje et al., 2015) 15-state segmentation of the GM12878 cell line (Figure S6C). Although the number of classes is higher than the one in our 1D segmentation above, their epigenetic profiles are similar, and the correlations between the corresponding model contact maps and Hi-C data are similar, too, with a mean across chromosomes of r = 0.78 and r′ = 0.16 (e.g., for chromosome 20, r = 0.78, r’ = 0.19; Figure S6G). Hence, the patterns derived from a polymer model constructed from 1D epigenetic segmentation is only partially better than one where Hi-C pairwise interactions are replaced by the average value corresponding to that genomic separation. Conversely, a SBS model with nine types of binding domains, based on epigenetics classes, genomically overlapping as discussed before, has higher correlations, with a mean across chromosomes of r = 0.87 and r′ = 0.45 (STAR Methods); and, as stated, the model with the full set of inferred binding domains has a mean of r = 0.94 and r′ = 0.74. To highlight similarities and differences between the experimental and the 1D segmentation model predicted contact patterns, we show a zoom of a 20-Mb-wide region on chromosome 20 in Figures 4A–4C and a 10-Mb zoom in Figures S7A–S7F.

To understand the partial failure of 1D epigenetic segmentation in explaining contact data (Figures 4B and 4C), for each pair of genomic sites we identified the binding domain that mostly contributes to their pairwise interaction within the full SBS model (Figures 4D–4F; STAR Methods). For clarity, we focused on the case-study 20-Mb-wide region on chromosome 20 of Figure 4. Plaid patterns are visible in its Hi-C contact map, as expected from A/B compartments (Figure 4A); they are also visible in the matrix of the most contributing binding domains (Figure 4F), where rich and fine substructures appear as well. Consider, for instance, the TAD associated with region C in Figure 4. The interactions within that TAD are mainly related to binding domains in class 7 (magenta; Figure 4F), which is indeed the most abundant within the genomic region where C is located (Figure 4E). Its interactions with the upstream region A can be simply traced back to homotypic interactions within class 7 itself, which is also the most abundant in A. However, the flanking region B, in which class 6 (dark blue) is the first most abundant, also interacts with C. That occurs because class 7 is the second most abundant in B and because in C class 6 is, in turn, the second most abundant. Such an example illustrates that a linear epigenetic segmentation model with homotypic interactions fails to account for the complexity of the observed contact pattern because a homotypic interaction between B and C would only occur if the two regions belonged to the same class. Analogously, the contacts between regions A and B originate from different overlapping binding domains included in those regions. Similar reasoning can be extended to the plaid pattern of A/B compartments (which is a specific example of a two-classes genome 1D segmentation) capturing the overall interactions between homologous active and repressed regions, respectively (Lieberman-Aiden et al., 2009; Rao et al., 2014). Yet, a much more complex and finer structure of contacts exists (including interactions across A and B compartments). Indeed, it has been shown that polymer models based on a linear epigenetic classification of domains are forced to include heterotypic interactions to accurately explain Hi-C data (Di Pierro et al., 2016). Consistent with such a picture, polymer models informed with tracks of binding sites of multiple proteins/factors, i.e., models where in a given DNA region more than one protein/factor can bind, perform well in recapitulating complex microscopy and Hi-C contact patterns (see, e.g. Barbieri et al., 2017; Brackley et al., 2016b).

Overall, homotypic interactions between the domains of a coarse-grained linear epigenetic segmentation of the genome, such as compartment A/B, are not enough to explain the specificity of Hi-C patterns with high accuracy, since a complexity of relevant heterotypic contacts exists between those regions. The origin of those heterotypic interactions is understood within our analysis showing that multiple binding domains are present in a genomic segment. Their genomic 1D combinatorial overlaps associate a distinctive interaction profile to each DNA segment, containing the information required to produce spontaneously, through physics mechanisms like those discussed above, the complex details of the system 3D conformations (Figure 4). In turn, the specific set of histone marks barcoding each binding domain provides a code linking epigenetic to architecture.

The epigenetic barcode of binding domains predicts *de novo* chromatin architecture

To validate the identified association between linear epigenetic features and chromosome conformations, we considered a reverse approach whereby, starting from only epigenetics data through the mentioned code, we identify the key binding sites of a set of independent chromosomes and, next, predict their contact matrices via polymer physics (Figure 5A). Specifically, we

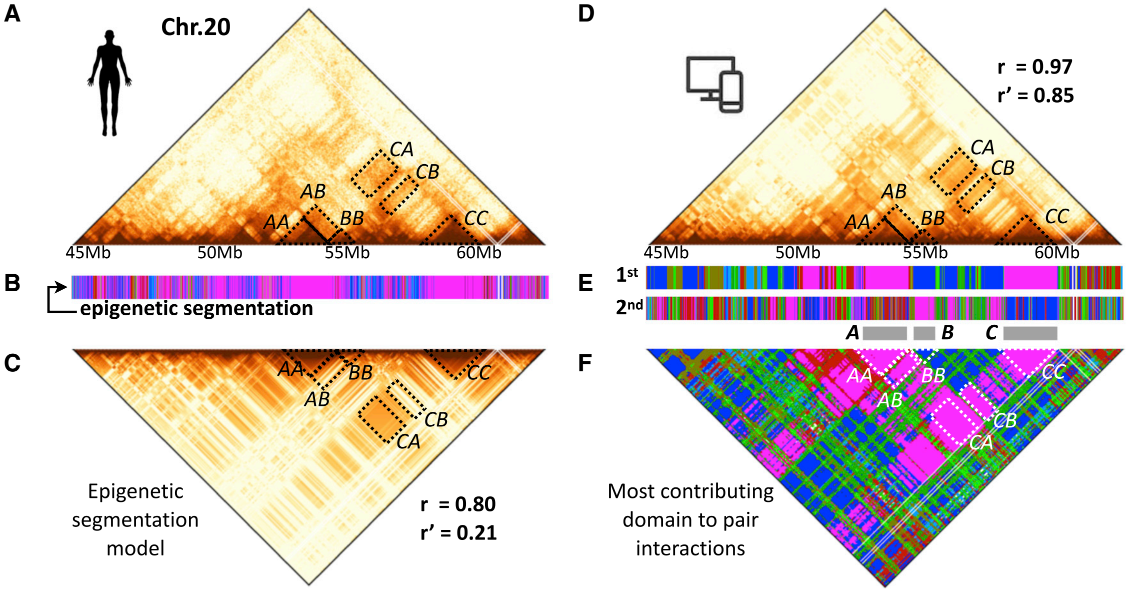


Figure 4. Chromatin architecture patterns are only partially captured by linear epigenetic segmentation

(A and B) (A) *In situ* Hi-C data (Rao et al., 2014; scales as in Figure 1) of a 20-Mb-wide region on chromosome 20 in GM12878 and (B) its linear epigenetic segmentation are shown.

(C) Contact map of the entire chromosome 20 of a model based only on homotypic interactions between linear segmented epigenetic domains has a Pearson correlation r = 0.80 with the Hi-C data, but it has a low distance-corrected correlation r′ = 0.21, showing only a partial improvement over a control model where each interaction is replaced by the average at the corresponding genomic separation. Here, a 20-Mb region is zoomed to highlight the different patterns.

(D) Contact map of the inferred SBS model of chr20 has r = 0.97 and r′ = 0.85 with its Hi-C data.

(E) PRISMR-inferred first and second most abundant binding-site types of the SBS model of the 20-Mb region are shown.

(F) The plot of the SBS most contributing binding domain to each pairwise contact highlights that a combinatorial overlap of different binding-site types along the sequence, missing in linear segmentations is required to capture the complexity and specificity of interaction patterns. For example, interactions (CC) within the TAD in region C are mainly related to binding domains in class 7 (magenta), the most abundant one in C. A and C also interact mainly through class 7, the most abundant in A, too. Yet, region B, where class 6 (dark blue) is the most abundant, interacts with C mainly through class 7, its second most abundant. Analogously, contacts between A and B originate from different overlapping binding domains in those regions.

exploited the epigenetic barcoding provided by the classification of the binding domains of even-numbered chromosomes, as previously described, to identify *de novo* the binding sites of odd-numbered chromosomes. To determine the locations and types of the binding sites, we partitioned each 5-kb genomic window (5 kb is the resolution of Hi-C) of odd-numbered chromosomes in equal-sized, 0.5-kb sub-windows, which we epigenetically profiled by measuring the abundance of the mentioned key set of histone marks (STAR Methods). We then computed the correlations between the epigenetic profile of each sub-window and the centroids of the nine epigenetic classes of the binding domains of even-numbered chromosomes (Figure 3A). We focused on those epigenetics classes because they recapitulate the main functional groups found in segmentation studies; additionally, considering nine types of sites is more stringent than considering all the binding domains found on even chromosomes, as exploiting such a larger set of domains would only improve the results. Finally, each sub-window of odd-numbered chromosomes was assigned a binding site type corresponding to the epigenetic class having the highest correlation (Figure 5A; STAR Methods).

Once we obtained the genomic locations of the binding sites along odd-numbered chromosomes, we computed their contact matrices via the SBS polymer model and compared them with the corresponding *in situ* Hi-C maps (Figure S7G). Figures 5B and 5C show, for example, the contact data of the entire chromosomes 19 and 21 predicted by use of the above-defined code, which as stated links binding sites, i.e., architecture, to epigenetic marks. In all the considered cases, the predicted matrices well capture the patterns of interactions seen in Hi-C data across genomic distances. The correlation and distance-corrected correlation coefficients (Figure S7H) are much higher than those found by 1D epigenetic segmentation, as seen above (e.g., r = 0.91 and r′ = 0.47 and r = 0.91 and r′ = 0.63 for, respectively, chromosomes 19 and 21).

Taken together, our results show that the barcode linking epigenetics marks to the binding sites inferred by PRISMR from Hi-C data, albeit still incomplete, can predict the genome's 3D architecture to a good level of accuracy. A crucial difference between our and epigenetic segmentation strategies to predict chromatin contacts is the intrinsically overlapping nature of binding domains, lacking in segmentations, which is necessary to recapitulate accurately the complex pattern of chromatin interactions.

DISCUSSION

To infer from Hi-C data the different types of DNA binding sites determining chromosome architecture and their genomic

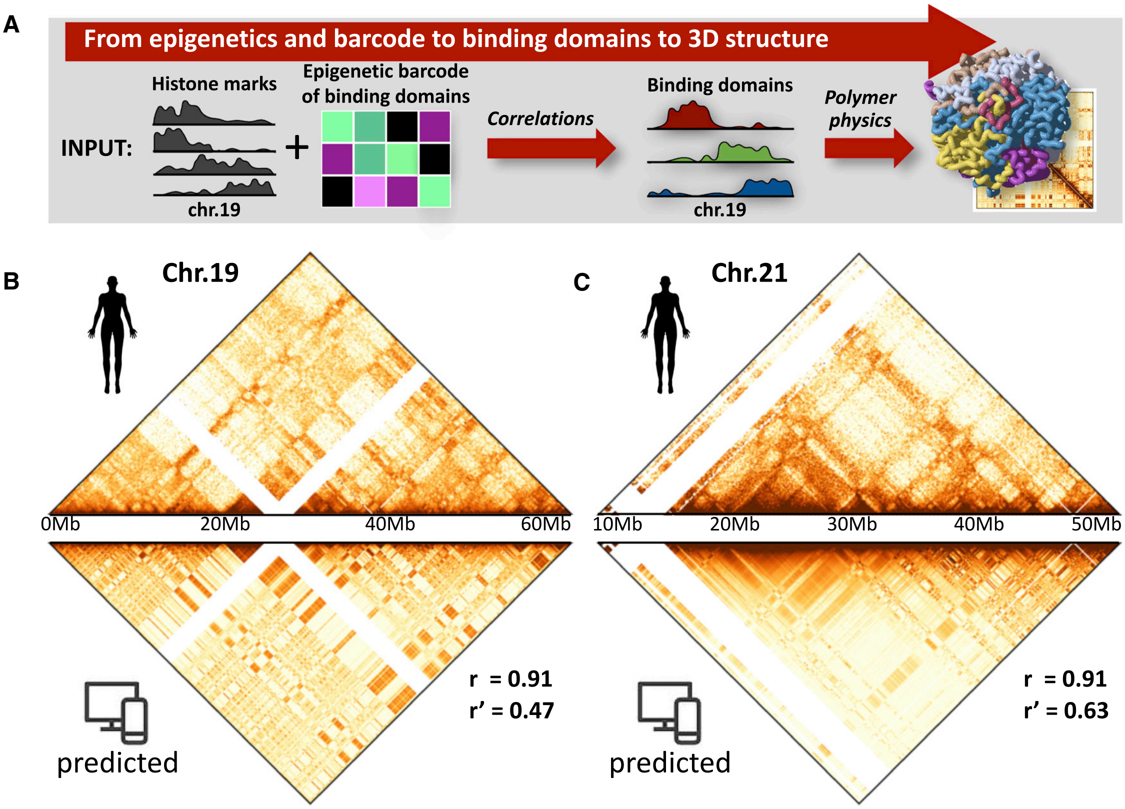


Figure 5. The epigenetic barcode of binding domains predicts *de novo* chromatin contacts

(A) In a reverse approach, we correlate the epigenetic profiles of the binding domains from even chromosomes with epigenetic signals from odd chromosomes to identify the binding sites of the latter. Next, we use the SBS polymer model to predict 3D structures and contact matrices of odd chromosomes to be compared against independent Hi-C data.

(B) Top: *in situ* Hi-C data (Rao et al., 2014; scales as in Figure 1) of the entire chromosome 19 in GM12878. Bottom: the predicted contact matrix has a correlation, a distance-corrected correlation, and a stratum adjusted correlation with Hi-C respectively equal to r = 0.91, r′ = 0.47, and SCC = 0.65.

(C) Top: Hi-C data of the entire chromosome 21. Bottom: the predicted contact matrix has correlations with Hi-C equal to r = 0.91, r′ = 0.63, and SCC = 0.50.

position, we employed a procedure based on machine learning and the physics of the SBS polymer model of chromatin (Bianco et al., 2018). The SBS model quantifies the scenario where molecular factors (such as TFs) establish DNA contacts and loops between distal cognate binding sites (Nicodemi and Prisco, 2009), and our procedure returns the putative binding sites specific to the model of each given chromosome. We found that the 3D structures derived by the model informed with the inferred binding domains explain Hi-C data genome-wide with high accuracy in human GM12878 B-lymphoblastoid cells (Rao et al., 2014) and mESCs (Dixon et al., 2012). That shows that the basic molecular ingredients considered by the model are sufficient to explain contact patterns across genomic scales. As the identified binding domains encode the molecular information required to fold chromatin, they provide an architectural code whereby 3D conformations can be assembled based on the 1D sequence (Figure 6). To explain folding with high accuracy, they have a cell type-specific combinatorial organization along chromosomes, which is needed to control the intricate multitude of genomic interactions captured in Hi-C maps and their functional specificity, via a comparatively smaller number of molecular factors. Additionally, the non-trivial arrangement of binding domains provides structural stability to the 3D conformation of the genome, as experimentally reported (Barutcu et al., 2018; Kubo et al., 2017; Nora et al., 2017; Rao et al., 2017; Rodríguez-Carballo et al., 2017). We found that binding domains produce chromatin interactions extending across chromosomal scales, from below the size of TADs, to A/B compartments, up to entire chromosomes, in a hierarchy of higher-order 3D structures as in the meta-TAD picture (Fraser et al., 2015).

Next, we associated each of the Hi-C inferred binding domains to an epigenetic profile based on their genomic correlation with a few important ENCODE histone marks. The model binding domains turn out to belong to main epigenetic classes, similar in human and mouse cell types, which well match known chromatin states (e.g., active, poised, repressed) derived by linear segmentation studies (Boettiger et al., 2016; Ernst et al., 2011; Gifford et al., 2013; Ho et al., 2014; Javierre et al., 2016). However, as stated, the identified binding domains have broad overlaps along the genome, a feature missing in linear segmentations but required to explain Hi-C accurately. The few coarse-grained epigenetic classes discussed here constitute only a first, simplified description of the epigenetic features of the binding domains that shape chromatin architecture. More generally, their barcode

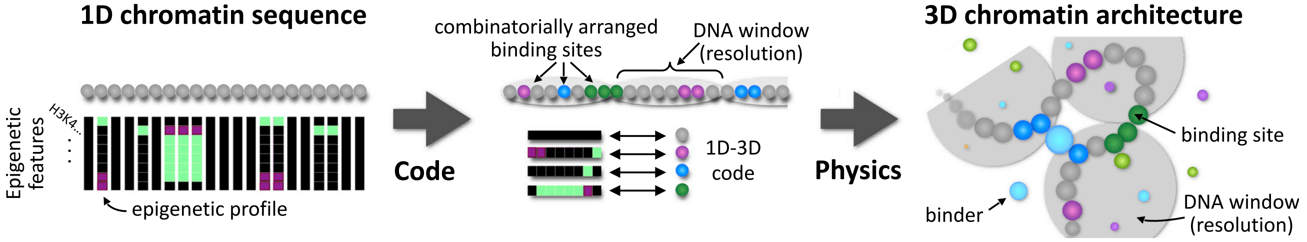


Figure 6. Chromatin 3D architectural information is encrypted in a combinatorial 1D arrangement of epigenetically barcoded binding sites
Our approach infers, from Hi-C data only, the minimal set of binding sites along the 1D genome sequence (middle) required to produce, via interactions with diffusing cognate binding molecules (i.e., via polymer physics) 3D structures (right) consistent with Hi-C data. Next, we find that different combinations of epigenetic factors (left, vertical bars) mark the distinct inferred binding-site types (bead colors), which fall into epigenetic classes well matching functional chromatin states known from linear segmentation studies. However, the binding sites have a genomic overlapping, combinatorial organization, lacking in epigenetic segmentations, necessary to explain Hi-C contacts with high accuracy genome-wide. The resulting code linking specific sets of epigenetic signals to different types of binding sites (middle, bottom) can predict *de novo* chromatin conformations, e.g., after genetic or epigenetic variations, showing that the inferred combinatorial 1D arrangement of binding sites carries accurate 3D architectural information.

is expected to be associated with a broader set of (still partially unknown) molecular factors, including histone marks, CTCF (Rao et al., 2014), active/poised Pol-II (Barbieri et al., 2017), lncRNAs (Quinodoz et al., 2021) and additional factors, such as PRC1 (Kundu et al., 2017), PRC2 (Barbieri et al., 2017), and MLL3/4 (Yan et al., 2018). Furthermore, molecular mechanisms beyond those envisaged by the SBS model, such as DNA loop extrusion (Brackley et al., 2017; Buckle et al., 2018; Fudenberg et al., 2016; Sanborn et al., 2015), appear to play a role in chromosome folding, and the code can be extended to accommodate them.

The inferred binding domains and the associated architectural interaction code were tested by making predictions on the changes of the 3D structure caused by a set of structural variants at the *Sox9* locus linked to human diseases. Notably, the predicted contact maps were confirmed by independent cHi-C data in cells carrying such mutations (Franke et al., 2016) in a stringent validation because there are no available fitting parameters. The model also helps understanding how the mutations differently affect the 3D structure of the locus (e.g., forming neo-TADs) and how that differently impacts gene regulation and, hence, phenotype by enhancer hijackings.

Finally, in a reverse approach, based on the discovered code linking epigenetics to the binding domains and, hence, to the 3D architecture, we identified the binding sites of an independent set of chromosomes from only their epigenetic marks. Those binding sites were sufficient to predict *de novo*, via the SBS model, the contact matrices of those chromosomes with good accuracy, validating the inferred epigenetic-architecture code. The binding domains have a cell type-specific genomic arrangement, yet their overall features, as much as their epigenetic classes, are similar in the human and mouse cells investigated here, hinting toward a general organizational principle.

Overall, the agreement between our results and the independent experimental Hi-C data strengthens the scenario where chromatin 3D architectural information is encoded in a 1D combinatorial arrangement of epigenetically barcoded sites, which can be inferred across chromosomes and cell types by our computational approach. By integration of different genomic data, it provides a quantitative picture of the cause-effect relationship between epigenetics, architecture, and function, which can help the development of tools in biomedicine to infer the link between genotype and phenotype through the features of the genomic landscape.

Limitations of the study

A limitation of our study is the 5 kb resolution of the data employed for computational feasibility, as it limits the accuracy of our model in determining the link between 3D architecture and epigenetic marks such as histone tracks and TF binding sites. Furthermore, the inclusion in the model of additional molecular signals and integration of data from different experimental assays, such as GAM, SPRITE, or microscopy, could improve its predictive power.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- METHOD DETAILS
  - ○ The String & Binders Switch model of chromatin
  - ○ The PRISMR method
  - ○ Details on the application of PRISMR genome-wide
  - ○ Structural variants at the Sox9 locus and validation of PRISMR
  - ○ Matrix similarity evaluation
  - ○ Molecular dynamics simulations
  - ○ Characterization of the binding domains arrangement along chromosomes
  - ○ Epigenetic analysis of the binding domains
  - ○ Characterization of epigenetic classes of binding domains
  - ○ Most abundant and most contributing binding domains to chromatin pairwise contacts

  - ○ Epigenetic linear segmentation models
  - ○ Prediction of *de novo* chromatin structures from epigenetic data by combinatorial barcode
- ● QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2022.110601.

ACKNOWLEDGMENTS

M.N. acknowledges support from NIH grants ID 1U54DK107977-01 and 1UM1HG011585-01, the EU H2020 Marie Curie ITN n.813282, CINECA ISCRA ID HP10CYFPS5 and HP10CRTY8P, Einstein BIH Fellowship Award (EVF-BIH-2016 and 2019), Regione Campania SATIN Project 2018-2020. A.M.C. acknowledges support from "Programma per il Finanziamento della Ricerca di Ateneo Linea B" (FRA) 2020, University of Naples Federico II. S.B. and A.M.C. acknowledge support from the CINECA ISCRA grant ID HP10CCZ4KN. We acknowledge computer resources from INFN, CINECA, ENEA CRESCO/ENEAGRID (Iannone et al., 2019) and *Scope/ReCAS/Ibisco* at the University of Naples.

AUTHOR CONTRIBUTIONS

M.N. designed the project. A.E. and S.B. developed the modeling part. A.E., S.B., A.M.C, A.A., L.F., M.C., and R.C. ran the computer simulations and performed the data analyses. M.N., A.E., S.B., and A.M.C. wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 18, 2021
Revised: January 21, 2022
Accepted: March 9, 2022
Published: March 29, 2022

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. IEEE Trans. Automat. Contr. *19*, 716–723.

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. U S A. *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.

Barutcu, A.R., Maass, P.G., Lewandowski, J.P., Weiner, C.L., and Rinn, J.L. (2018). A TAD boundary is preserved upon deletion of the CTCF-rich Firre locus. Nat. Commun. *9*, 1444.

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.

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.

Bianco, S., Annunziatella, C., Andrey, G., Chiariello, A.M., Esposito, A., Fiorillo, L., Prisco, A., Conte, M., Campanile, R., and Nicodemi, M. (2019). Modeling single-molecule conformations of the HoxD region in mouse embryonic stem and cortical neuronal cells. Cell Rep. *28*, 1574–1583.e4.

Bickmore, W.A., and Van Steensel, B. (2013). Genome architecture: domain organization of interphase chromosomes. Cell *152*, 1270–1284.

Bintu, B., Mateo, L.J., Su, J.-H.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.

Boettiger, A.N., Bintu, B., Moffitt, J.R., Wang, S., Beliveau, B.J., Fudenberg, G., Imakaev, M., Mirny, L.A., Wu, C.T., and Zhuang, X. (2016). Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. Nature *529*, 418–422.

Bohn, M., and Heermann, D.W. (2010). Diffusion-driven looping provides a consistent framework for chromatin organization. PLoS One *5*, e12218.

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. U S A. *110*, E3605–E3611.

Brackley, C.A., Brown, J.M., Waithe, D., Babbs, C., Davies, J., Hughes, J.R., Buckle, V.J., and Marenduzzo, D. (2016a). Predicting the three-dimensional folding of cis-regulatory regions in mammalian genomes using bioinformatic data and polymer models. Genome Biol. *17*, 59.

Brackley, C.A., Johnson, J., Kelly, S., Cook, P.R., and Marenduzzo, D. (2016b). Simulated binding of transcription factors to active and inactive regions folds human chromosomes into loops, rosettes and topological domains. Nucleic Acids Res. *44*, 3503–3512.

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.

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.

Cattoni, D.I., Gizzi, A.M.C., Georgieva, M., Di Stefano, M., Valeri, A., Chamousset, D., Houbron, C., Déjardin, S., Fiche, J.B., González, I., et al. (2017). Single-cell absolute contact probability detection reveals chromosomes are organized by multiple low-frequency yet specific interactions. Nat. Commun. *8*, 1753.

Chiariello, A.M., Bianco, S., Oudelaar, A.M.M., Esposito, A., Annunziatella, C., Fiorillo, L., Conte, M., Corrado, A., Prisco, A., Larke, M.S.C., et al. (2020). A dynamic folded hairpin conformation is associated with α-globin activation in erythroid cells. Cell Rep. *30*, 2125–2135.e5.

Chiariello, A.M.A.M., Annunziatella, C., Bianco, S., Esposito, A., and Nicodemi, M. (2016). Polymer physics of chromosome large-scale 3D organisation. Sci. Rep. *6*, 29775.

Conte, M., Fiorillo, L., Bianco, S., Chiariello, A.M., Esposito, A., and Nicodemi, M. (2020). Polymer physics indicates chromatin folding variability across single-cells results from state degeneracy in phase separation. Nat. Commun. *11*, 3289.

Dekker, J., and Heard, E. (2015). Structural and functional diversity of topologically associating domains. FEBS Lett. *589*, 2877–2884.

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.

Dixon, J.R., Gorkin, D.U., and Ren, B. (2016). Chromatin domains: the unit of chromosome organization. Mol. Cell *62*, 668–680.

Di Pierro, M., Zhang, B., Aiden, E.L., Wolynes, P.G., and Onuchic, J.N. (2016). Transferable model for chromosome architecture. Proc. Natl. Acad. Sci. U S A *113*, 12168–12173.

Dryden, N.H., Broome, L.R., Dudbridge, F., Johnson, N., Orr, N., Schoenfelder, S., Nagano, T., Andrews, S., Wingett, S., Kozarewa, I., et al. (2014). Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. Genome Res. *24*, 1854–1868.

Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nat. Methods *9*, 215–216.

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.

Finn, E.H., Pegoraro, G., Brandão, H.B., Valton, A.L., Oomen, M.E., Dekker, J., Mirny, L., and Misteli, T. (2019). Extensive heterogeneity and intrinsic variation in spatial genome organization. Cell *176*, 1502–1515.e10.

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. (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.

Fudenberg, G., Kelley, D.R., and Pollard, K.S. (2020). Predicting 3D genome folding from DNA sequence with Akita. Nat. Methods. *17*, 1111–1117.

Gifford, C.A., Ziller, M.J., Gu, H., Trapnell, C., Donaghey, J., Tsankov, A., Shalek, A.K., Kelley, D.R., Shishkin, A.A., Issner, R., et al. (2013). Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. Cell *153*, 1149–1163.

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.

Iannone, F., Ambrosino, F., Bracco, G., De Rosa, M., Funel, A., Guarnieri, G., Migliori, S., Palombi, F., Ponti, G., Santomauro, G., et al. (2019). CRESCO ENEA HPC clusters: a working example of a multifabric GPFS spectrum scale layout. In 2019 International Conference on High Performance Computing & Simulation (HPCS) (IEEE), pp. 1051–1052.

Javierre, B.M., Sewitz, S., Cairns, J., Wingett, S.W., Várnai, C., Thiecke, M.J., Freire-Pritchett, P., Spivakov, M., Fraser, P., Burren, O.S., et al. (2016). Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. Cell *167*, 1369–1384.e19.

Jost, D., Carrivain, P., Cavalli, G., and Vaillant, C. (2014). Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains. Nucleic Acids Res. *42*, 9553–9561.

Kirkpatrick, S., Gelatt, C.D., and Vecchi, M.P. (1983). Optimization by simulated annealing. Science *220*, 671–680.

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.

Krijger, P.H.L., and De Laat, W. (2016). Regulation of disease-associated gene expression in the 3D genome. Nat. Rev. Mol. Cell Biol. *17*, 771–782.

Kubo, N., Ishii, H., Gorkin, D., Meitinger, F., Xiong, X., Fang, R., Liu, T., Ye, Z., Li, B., Dixon, J., et al. (2017). Preservation of chromatin organization after acute loss of CTCF in mouse embryonic stem cells. Preprint at bioRxiv. https://doi.org/10.1101/118737.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

Kundu, S., Ji, F., Sunwoo, H., Jain, G., Lee, J.T., Sadreyev, R.I., Dekker, J., and Kingston, R.E. (2017). Polycomb repressive complex 1 generates discrete compacted domains that change during differentiation. Mol. Cell *65*, 432–446.e5.

Li, Q., Tjong, H., Li, X., Gong, K., Zhou, X.J., Chiolo, I., and Alber, F. (2017). The three-dimensional genome organization of Drosophila melanogaster through data integration. Genome Biol. *18*, 145.

Lieberman-Aiden, E., Van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

Lin, D., Bonora, G., Yardimci, G.G., and Noble, W.S. (2018). Computational methods for analyzing and modeling genome structure and organization. Wiley Interdiscip. Rev. Syst. Biol. Med. *11*, e1435.

Misteli, T. (2007). Beyond the sequence: cellular organization of genome function. Cell *128*, 787–800.

Nicodemi, M., and Pombo, A. (2014). Models of chromosome structure. Curr. Opin. Cell Biol. *28*, 90–95.

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

Nir, G., Farabella, I., Pérez Estrada, C., Ebeling, C.G., Beliveau, B.J., Sasaki, H.M., Lee, S.H., Nguyen, S.C., McCole, R.B., Chattoraj, S., et al. (2018). Walking along chromosomes with super-resolution imaging, contact maps, and integrative modeling. PLoS Genet. *14*, e1007872.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, 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.

Nora, E.P., Goloborodko, A., Valton, A.L., Gibcus, J.H., Uebersohn, A., Abdennur, N., Dekker, J., Mirny, L.A., and Bruneau, B.G. (2017). Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. Cell *169*, 930–944.e22.

Oudelaar, A.M., Hanssen, L.L.P., Hardison, R.C., Kassouf, M.T., Hughes, J.R., and Higgs, D.R. (2017). Between form and function: the complexity of genome folding. Hum. Mol. Genet. *26*, R208–R215.

Plimpton, S. (1995). Fast parallel algorithms for short-range molecular dynamics. J. Comput. Phys. *117*, 1–19.

Qi, Y., and Zhang, B. (2019). Predicting three-dimensional genome organization with chromatin states. PLoS Comput. Biol. *15*, e1007024.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Quinodoz, S.A., Ollikainen, N., Tabak, B., Palla, A., Schmidt, J.M., Detmar, E., Lai, M.M., Shishkin, A.A., Bhat, P., Takei, Y., et al. (2018). Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. Cell *174*, 744–757.e24.

Quinodoz, S.A., Jachowicz, J.W., Bhat, P., Ollikainen, N., Banerjee, A.K., Goronzy, I.N., Blanco, M.R., Chovanec, P., Chow, A., Markaki, Y., et al. (2021). RNA promotes the formation of spatial compartments in the nucleus. Cell *184*, 5775–5790.e30.

Rao, S.S.P., Huang, S.C., Glenn St Hilaire, B., Engreitz, J.M., Perez, E.M., Kieffer-Kwon, K.R., Sanborn, A.L., Johnstone, S.E., Bascom, G.D., Bochkov, I.D., et al. (2017). Cohesin loss eliminates all loop domains. Cell *171*, 305–320.e24.

Rao, S.S.P.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., et al. (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., Nguyen Huynh, T.H., 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.

Salamon, P., Sibani, P., and Frost, R. (2002). Facts, Conjectures, and Improvements for Simulated Annealing (SIAM).

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. *112*, E6456–E6465.

Schwessinger, R., Gosden, M., Downes, D., Brown, R.C., Oudelaar, A.M., Telenius, J., Teh, Y.W., Lunter, G., and Hughes, J.R. (2020). DeepC: predicting 3D genome folding using megabase-scale transfer learning. Nat. Methods. *17*, 1118–1124.

Serra, F., Baù, D., Goodstadt, M., Castillo, D., Filion, G.J., and Marti-Renom, M.A. (2017). Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. PLOS Comput. Biol. *13*, e1005665.

Sexton, T., and Cavalli, G. (2015). The role of chromosome domains in shaping the functional genome. Cell *160*, 1049–1059.

Spielmann, M., Lupiáñez, D.G., and Mundlos, S. (2018). Structural variation in the 3D genome. Nat. Rev. Genet., 1–15.

Di Stefano, M., Paulsen, J., Lien, T.G., Hovig, E., and Micheletti, C. (2016). Hi-C-constrained physical models of human chromosomes recover functionally-related properties of genome organization. Sci. Rep. *6*, 35985.

Tang, Z., Luo, O.J., Li, X., Zheng, M., Zhu, J.J., Szalaj, P., Trzaskoma, P., Magalska, A., Wlodarczyk, J., Ruszczycki, B., et al. (2015). CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. Cell *163*, 1611–1627.

Yan, J., Chen, S.A.A., Local, A., Liu, T., Qiu, Y., Dorighi, K.M., Preissl, S., Rivera, C.M., Wang, C., Ye, Z., et al. (2018). Histone H3 lysine 4 monomethylation modulates long-range chromatin interactions at enhancers. Cell Res. *28*, 204–220.

Yang, T., Zhang, F., Yardımci, 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.

Zhang, S., Chasman, D., Knaack, S., and Roy, S. (2019). In silico prediction of high-resolution Hi-C interaction matrices. Nat. Commun. *10*, 5449.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
H3K4me1 ChIP-seq in GM12878	ENCODE	ENCFF682WPF
H3K36me3 ChIP-seq in GM12878	ENCODE	ENCFF662QFK
H3K9me3 ChIP-seq in GM12878	ENCODE	ENCFF776OVW
H3K27me3 ChIP-seq in GM12878	ENCODE	ENCFF167NBF
H3K4me2 ChIP-seq in GM12878	ENCODE	ENCFF828CQV
H3K9ac ChIP-seq in GM12878	ENCODE	ENCFF465KNK
H3K27ac ChIP-seq in GM12878	ENCODE	ENCFF180LKW
H3K79me2 ChIP-seq in GM12878	ENCODE	ENCFF396JIR
H3K20me1 ChIP-seq in GM12878	ENCODE	ENCFF831WYD
H2AFZ ChIP-seq in GM12878	ENCODE	ENCFF885XEM
CTCF ChIP-seq in GM12878	ENCODE	ENCFF312KXX
RAD21 ChIP-seq in GM12878	ENCODE	ENCFF567EGK
SMC3 ChIP-seq in GM12878	ENCODE	ENCFF235BXX
POLR2A ChIP-seq in GM12878	ENCODE	ENCFF368HBX
DNase-seq in GM12878	ENCODE	ENCFF264NMW
Repli-seq G1 phase in GM12878	ENCODE	ENCFF001GNK
Repli-seq S1 phase in GM12878	ENCODE	ENCFF001GNR
Repli-seq S2 phase in GM12878	ENCODE	ENCFF001GNT
Repli-seq S3 phase in GM12878	ENCODE	ENCFF001GNX
Repli-seq S4 phase in GM12878	ENCODE	ENCFF001GOA
Repli-seq G2 phase in GM12878	ENCODE	ENCFF001GNN
Total RNA-seq in GM12878	ENCODE	ENCFF273YJY
H3K4me3 in mouse ESCs	ENCODE	ENCFF796LDS
H3K4me1 in mouse ESCs	ENCODE	ENCFF817CZF
H3K36me3 in mouse ESCs	ENCODE	ENCFF001XWZ
H3K9me3 in mouse ESCs	ENCODE	ENCFF001YHE
H3K27me3 in mouse ESCs	ENCODE	ENCFF945LRL
Software and algorithms		
LAMMPS	Plimpton (1995)	https://lammps.sandia.gov
POV-Ray	Persistence of Vision Pty. Ltd	http://www.povray.org/
BEDTools	Quinlan and Hall (2010)	https://bedtools.readthedocs.io/en/latest/
HiCRep	Yang et al. (2017)	https://github.com/qunhualilab/hicrep

RESOURCE AVAILABILITY

Lead contact

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).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyzes existing, publicly available data. Accession numbers for the datasets are listed in the key resources table.

- All data and codes for the genome-wide analysis of binding domains have been publicly released on GitHub (https://github.com/AndreaEsp/ChromBarCode).
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

The String & Binders Switch model of chromatin

To investigate the 3D structure of the genome, we employed the String & Binders Switch (SBS) model (Barbieri et al., 2012; Chiariello et al., 2016; Nicodemi and Prisco, 2009). According to the SBS, a chromatin filament (from small loci to entire chromosomes) is modeled as a self-avoiding walk polymer chain of beads, a fraction of which, named binding sites, interacts with diffusing molecular binders. The interaction between binding sites and binders allows for the formations of loops along the polymer and, therefore, permits its spontaneous folding (Figure 1B). Each bead can be bound only by its specific, cognate type of binders and, to fully describe the complexity of the system, different types of interactions are allowed together with inert sites along the chain that do not interact with any binder (apart from steric effects). We represent these different interactions as different ''colors'' of the system, ''gray'' beads being the non-interacting particles (Figure 1B). Key parameters of the model are the concentration, c, and the binding energy, E_{int}, of each different type of binder. As a function of c and E_{int}, the system of corresponding, cognate binding sites exhibits a coil-globule phase transition from an open conformation (at low concentration or energy) to a globule, compact phase (at high concentration or energy) as extensively discussed in previous studies (Barbieri et al., 2012; Chiariello et al., 2016; Conte et al., 2020). The presence of different sets of binding sites (here named ''binding domains'' and represented with different colors) interacting with different, cognate molecular factors allows the formation of complex 3D structures by microphase separation.

The PRISMR method

To determine the distribution of the different binding sites along the SBS polymer chain, here we used PRISMR, a previously illustrated machine learning procedure (Bianco et al., 2018). The PRISMR algorithm is a polymer physics-based method that, trained on an experimental contact matrix (e.g., Hi-C or GAM), learns the minimal polymer model that, at equilibrium, best describes the input. The learned model is then used to reconstruct the chromosome 3D structure and to make predictions of unseen data, e.g., on the effect of genomic mutations on chromatin organization. Although we focus on the SBS polymer model to describe a chromatin filament, the PRISMR algorithm can be easily generalized to different models.

A detailed description of the PRISMR method can be found in ref (Bianco et al., 2018). Here we just summarize the key points of the algorithm. An SBS polymer model of a genomic region is composed of L beads, depending on the resolution of the input contact matrix of the region. For instance, a 10Mb locus at 10 kb resolution is partitioned in L = 1000 bins. Furthermore, we split each of the L bins into *r* different sub-units, considering that a single DNA bin could include many binding sites and interact with different factors. The SBS polymer is then completely characterized by the arrangement of the binding sites along the chain. Given the number n of different types of binding sites, PRISMR finds the color arrangement along the polymer chain by the minimization, via an iterative Simulated Annealing (SA) Monte Carlo optimization procedure (Kirkpatrick et al., 1983; Salamon et al., 2002), of a specific cost function made of two terms. The first term representing the distance between the experimental and the model predicted contact matrices; the second one is a Bayesian term proportional to the total number of colored sites of the polymer through a parameter λ and penalizes the addition of new colored beads. In this way we account for the necessity to fit well the input data and, at the same time, we attempt to avoid overfitting. After initializing the SBS polymer in a random configuration, by assigning a random color to each bead, a standard iterative SA procedure is performed, as available in public software repositories (see e.g. https://github.com/perrygeo/simanneal), to optimize the model (Kirkpatrick et al., 1983; Salamon et al., 2002). Schematically, each SA step consists in randomly changing the color of a polymer bead, compute the average contact matrix of the new polymer, evaluate the new cost function, compare it with the cost function in the previous step and, based on it, accept or reject the color change. SA steps are iteratively repeated until convergence (Bianco et al., 2018). The entire procedure is repeated many times by varying the polymer initial configurations and the model parameters n, *r*, and λ, to find their optimal values.

Details on the application of PRISMR genome-wide

In this study, we present the first genome-wide application of the algorithm. Precisely, here we applied PRISMR over the somatic chromosomes of the human genome, obtaining, for each chromosome independently, the SBS polymer that best describes its corresponding Hi-C matrix. We employed published *in situ* Hi-C data (Rao et al., 2014) relative to the human GM12878 cell line at 5 kb of resolution and normalized according to the method described in ref (Knight and Ruiz, 2013). To reduce the local noise in the input Hi-C data, we applied a Gaussian filter with a standard deviation equal to 1 along both *x* and *y* directions. The optimal value of the parameters of the algorithm has been estimated as already described in ref (Bianco et al., 2018), that is, we repeated the SA procedure many times starting from different initial conditions and different values of n, *r*, and λ to set these parameters at the values that explain the input data within a given accuracy. As input data for the optimal parameter evaluation, we used the contact matrix of chromosome 12, a medium-sized chromosome, obtaining n = 30 different types of binding sites, *r* = 30 and λ = 3×10-5. The same values for the parameters n, *r*, and λ have been used to obtain the best SBS polymer for all the other chromosomes. However, we checked the robustness of the optimal parameters by evaluating them on different chromosomes and we found a variation of around 15%. For example, we found n = 28 for the smaller chr19, while n = 34 for the larger chr8. Analogously, r and λ change of around 15% across chromosomes. Figure S1A shows the comparison between the chromosome-wide contact matrices inferred by PRISMR (lower triangular maps) and the *in situ* Hi-C matrices (upper triangular maps) at 5 kb resolution. The global pattern obtained by PRISMR is highly correlated with the experimental one as also quantified by the comparatively high values of the Pearson's (r), distance-corrected Pearson's (r') (Bianco et al., 2018) and stratum-adjusted (SCC) (Yang et al., 2017) correlation coefficients (Figure S1B, see below). In the calculation of r and r', to correct for outliers, we did not consider genomic distances below 25 kb. The PRISMR method is highly generalizable across different experiments and data resolution. To test that, we also applied our method to genome-wide Hi-C data in mouse embryonic stem (mES) cells (Dixon et al., 2012) at 40 kb resolution (Figure S2A). The correlations between experimental and model matrices obtained in mouse are as high as the values obtained in human, as shown in Figure S2B.

Structural variants at the Sox9 locus and validation of PRISMR

As a validation of the PRISMR inference method and the SBS model, we implemented in-silico a set of three previously studied structural variants in E12.5 limb bud cells from mice (Franke et al., 2016). Specifically, we first derived an SBS polymer model of the region chr11:109010000-114880000 (mm9), including the *Sox9* gene, from wild-type capture Hi-C (cHi-C) data in E12.5 limb buds (Franke et al., 2016). The cHi-C protocol incorporates a sequence capture step into a Hi-C procedure, so allowing high-resolution analysis of targeted regions of the genome (Dryden et al., 2014). Next, we implemented on the wild-type model, independently, the following duplications: *DupS*, an intra-TAD duplication of the region chr11:111760000-112160000; *DupL*, an inter-TAD duplication of the region chr11:110960000-112520000; *DupC*, another inter-TAD duplication of the region chr11:110760000-112520000. We then computed the PRISMR predicted contact maps for each duplication, under no adjustable parameters, obtaining the following values of correlations r and r', between model and experimental matrices (excluding the effect of strong outliers <5th and >95th percentile): r = 0.97 and r' = 0.87 in the WT, r = 0.95 and r' = 0.76 in *DupS*; r = 0.92 and r' = 0.63 in *DupL*; r = 0.90 and r' = 0.59 in *DupC* (Figure S3).

Matrix similarity evaluation

The agreement between experiment and model matrices has been quantified using Pearson's correlation coefficient, r. We also used two additional measures: 1) the distance corrected Pearson correlation coefficient, denoted by r', that is the Pearson's correlation coefficient between the two matrices where we subtracted from each diagonal (corresponding to a given genomic distance) their average contact frequency; 2) the stratum-adjusted correlation coefficient, denoted by SCC, from the HiCRep (Yang et al., 2017) method with a smoothing parameter h = 10 and an upper bound of interaction distance equal to 5 Mb. These two measures have been used to put aside the expected decreasing trend of the pairwise contact frequency with genomic distance, which tends to dominate in the simple Pearson correlation value.

Molecular dynamics simulations

To obtain 3D conformations of the PRISMR derived SBS models, shown in Figures 1G, 2D, 2E, S3F, and S3G, we performed Molecular Dynamics (MD) simulations. To this aim, we proceeded as described in ref (Chiariello et al., 2016). Briefly, the polymer chain and the binders move in the system according to the Langevin equation, integrated with the LAMMPS software (Plimpton, 1995), using standard dimensionless parameters (Kremer and Grest, 1990). The SBS parameters used are the same reported in ref (Chiariello et al., 2016), i.e., the beads and binders interact with an interaction energy E_{int} = 8.1KbT and the binder concentration is high enough to allow the coil-globule transition (c = 194 nmol/L for the *Sox9* WT and similar values for the duplications). To make MD computation times feasible for the entire chromosome 20, we considered a coarse-grained version of its SBS polymer, having a 50-fold reduced number of beads. All the conformations are taken in the equilibrium globular phase. In all the snapshots, beads coordinates have been interpolated with a smooth third-order polynomial splice curve by using the POV-RAY (Persistence of Vision Pty. Ltd) software.

Characterization of the binding domains arrangement along chromosomes

To study how the different binding domains (colors) span along the genome, we employed two main measures. The first one, that measures the domain size, is the genomic coverage, i.e., the fraction of beads of a given color multiplied by the length of the chromosome it belongs to. Averaging over all the sizes of the domains identified by PRIMSR across chromosomes, we find that the genomic length covered by each domain is on average 3.1 Mb, with a standard deviation of 1.9 Mb, a value close to the mean-size of a TAD. To measure, instead, the range of the interactions due to a single binding domain, we defined r_{int} as two times the standard deviation of the center of mass of that domain. The distribution of r_{int}, P(r_{int}), extends far beyond the size of the single domain, ranging from a few mega-bases to more than 100 Mb (Figure S2C). To check the statistical significance of the domains identified by PRISMR, we compared P(r_{int}) with a control model obtained by randomly bootstrapping the location of our binding sites along the genome, and we found that the two distributions are significantly different (p value<0.001, Wilcoxon's rank sum test). We also found that P(r_{int}) is asymptotically consistent with a power-law scaling, as shown in Figure S2C where the right-hand side of the distribution is well described by a power-law fit (dotted red curve in the graph).

Another way to test the significance of the binding domains identified by PRISMR is to measure their mutual overlap (Bianco et al., 2018), to be compared with the expected level of overlap in the random model of bootstrapped domains mentioned before. To this aim, given a pair of different domains on a chromosome, we defined their overlap q as the sum of products of binding sites occurrences of the two colors in each genomic window, normalized to have $q = 100\%$ in the case of identical domains (the cartoon in Figure S2D gives a visual impression of what q is measuring). We found that the distribution P(q) of the overlap of the binding domains predicted by PRISMR is significantly different (p value<0.001, Wilcoxon's rank sum test) from the one expected in the random control model (red and blue distributions in Figure S2D, respectively).

Epigenetic analysis of the binding domains

To obtain insight into their molecular nature, we analyzed the PRISMR inferred binding domains in the light of epigenetics data. To this aim, we downloaded from the ENCODE database (Dunham et al., 2012) a set of 5 key histone modifications (H3K4me3, H3K4me1, H3K36me3, H3K9me3 and H3K27me3) in the human GM12878 cell line. ChIP-Seq signals were binned at 5 kb resolution by summing the signal contained within each 5 kb window (using the bedtools map tool from the bedtools (Quinlan and Hall, 2010) software). After that, to measure the similarity between our binding domains and the histone marks, we computed Pearson's correlation coefficient between the number of binding sites of each domain and each histone mark profile. Next, we employed a control model to retain only statistically significant correlations. To this aim, first, we computed the Pearson correlations between chromatin mark signals and randomized binding domains signals obtained by bootstrapping their actual genomic locations; then, we retained as significant only the correlation values above the 95th or below the 5th percentile of the distribution of the random correlations. We then collected data in a rectangular matrix X, whose element X_{ij} is either the significant correlation between the i-th binding domains and the j-th histone mark or zero if the correlation was not significant. Since each row of X represents a binding domain's correlation profile with the considered histone modifications, we refer to them as the epigenomic signature of the binding domain. To find binding domains with similar epigenomic signatures, we performed a hierarchical clustering analysis on X using the *Python SciPy* clustering package with 'Euclidean' distance metric and 'Ward' linkage method. To assess the number of clusters in the hierarchical clustering output, we cut the dendrogram at different values (ranging from one to the number of binding domains) and evaluated the Akaike Information Criterion (Akaike, 1974) (AIC) as the number of clusters k is varied. As shown in Figure S4B, while no sharp transitions are present, the curve has a global minimum at $k = 9$. We therefore grouped all the different rows of X in 9 different classes according to their affinity to each cluster (Figure S4A). Each of the 9 classes can be characterized by the epigenetic signature of its centroid, which is the average histone signature of the domains belonging to the given class (Figure 3A). To assign biologically meaningful labels to the obtained classification, we looked at the enrichment of several types of functional annotations. Precisely, we first binned each annotation track at 5 kb resolution, then, for each pair of annotation mark and epigenetic class, we computed the average of the Pearson correlation values between that mark and the binding domains of that class (see Figure S4C). The set of functional annotations in GM12878 cell line considered in this study is taken from ENCODE and include: (1) all remaining available histone modifications; (2) transcription factors binding sites; (3) DNase hypersensitive sites; (4) replication timing data from the Repli-seq assay; (5) transcription data from RNA-seq assay (Figure S4C).

To further test the association between binding domains and epigenetics, we repeated the above analysis for the mouse case. Specifically, we computed correlations among the genome-wide binding domains obtained from Hi-C data in mES cells and a corresponding set of ENCODE histone modifications in that cell line. As shown in Figures S5A–S5C, we found an overall similar epigenetic classification of the binding domains in human and mouse.

Characterization of epigenetic classes of binding domains

The genomic coverage of a given epigenetic class has been computed as the fraction of sites of the binding domains belonging to that class (Figure S4D). To study, instead, how the domains of a given class are distributed along the chromosomes, we counted, for each class, the number of domains falling in each chromosome (Figure S4E, dotted lines are the average values). We found that their distribution is significantly different over the different chromosomes, as measured by the comparison with a uniform distribution obtained by randomly bootstrapping the domains of a given class over the chromosomes (p value<0.05 for each epigenetic class, Kolmogorov-Smirnov test). We also asked whether the genomic positions of the sites of the different classes (Figure 3B) were correlated with each other. To figure out that, we computed the Pearson correlation between the genomic location of the sites of all the possible pairs of epigenetic classes, averaged over the different chromosomes (Figure S4F). We found that classes with similar histone signature correlate with each other and anti-correlate with classes showing a very different histone pattern.

We investigated the impact of the different epigenetic classes on genome architecture by measuring the effect on contact matrices of the withdrawal of the binding domains belonging to each class. Precisely, given the list of the binding domains of a class, we replaced their interacting binding sites with gray, non-interacting elements along each chromosome. We then computed the PRISMR contact matrices of the modified SBS polymer and measured their correlations r and r' with Hi-C. Finally, we evaluated the variation of the correlation, Δr and Δr', with respect to the wild-type model (r = 0.94 and r' = 0.76), averaged over all chromosomes. The variations of r and r' obtained are shown as a function of the genomic coverage of each epigenetic class in Figure S4G.

Most abundant and most contributing binding domains to chromatin pairwise contacts

As the different binding domains can overlap with each other, to better visualize their locations along the genome, we show in Figure 4E (upper bar) the 1st most abundant binding domain, i.e. the one with the largest number of binding sites, per bin. Analogously, Figure 4E (lower bar) shows the 2nd most abundant binding domain per bin. In both cases, to help the visualization, the domains are colored with their epigenetic class color.

The contribution of the different binding domains in forming the interactions between bin pairs is then highlighted in Figure 4F, where the colors of the most contributing binding domains are shown. Specifically, for a given pairwise contact, we defined the contribution of a binding domain to that contact as the number of pairs of its binding site type between the two considered bins. The binding domain having the highest number of binding site pairs is the most contributing one and is colored with the color corresponding to its epigenetic class.

Epigenetic linear segmentation models

To obtain a model based exclusively on the interaction among segments with a similar epigenetic profile, we considered the dataset of five histone modifications discussed in section "epigenetic analysis of the binding domains". We marked each 5 kb genomic window with the z-score value of the signal of each histone mark in that window. Then, we performed a hierarchical clustering analysis to gather the genomic windows with similar histone profiles in 9 different groups, in order to match them with the 9 different types of binding domains found above (Figure S6A). The obtained linear segmentation has been employed to define polymer models with 9 different colors corresponding to the different linear epigenetic classes (Figure 4B), where interactions can only occur between same-colored windows. Finally, we derived in-silico the contact map of such a model and compared it with the corresponding experimental matrix (Figures 4A–4C, S6F, S7A–S7F). We repeated the same analysis by using a 15 state epigenetic segmentation of the human genome, previously obtained (Kundaje et al., 2015) with the ChromHMM software (Ernst and Kellis, 2012). Precisely, we downloaded the ChromHMM states specific for the GM12878 cell line (https://egg2.wustl.edu/roadmap/web_portal/), mapped them on the 5 kb wide genomic windows considered in this study, derived in-silico the corresponding contact maps as above and finally compared them with Hi-C (Figures S6G and S6B–S6D). We have also checked that by finer mapping the ChromHMM states on 200 bp genomic windows, we obtain similar results. Finally, we have considered an additional model by assigning each of the different binding sites the color of the epigenetic class it belongs to. We found that these 9 color SBS models, that in contrast to the linear segmentation model has overlapping binding domains, have higher correlations with Hi-C.

Prediction of *de novo* chromatin structures from epigenetic data by combinatorial barcode

The derived combinatorial code linking 3D conformation to 1D epigenetic signature can be used to predict de novo binding domains in independent chromosomes from epigenetics data only. Specifically, we used the code derived from the set of even-numbered chromosomes in GM12878 to predict the location of the binding sites along the odd-numbered chromosomes in the same cell line. To this aim, we partitioned each of their 5 kb windows (which is the *in situ* Hi-C data resolution) in ten 500-bp sub-windows and binned the signal of the five key histone marks (H3K4me3, H3K4me1, H3K36me3, H3K9me3 and H3K27me3) in those sub-windows. In this way, we obtained a state vector for each sub-window, whose components are the histone marks' abundances in that window. We checked that different sub-windows partitions, ranging from 5 to 20 sub-windows per bin, led to only marginally different results. We then used the obtained state vectors to assign at each 500-bp window a color corresponding to one of the epigenetic classes, so that 9 different types of binding site are allowed. Precisely, we computed the Pearson correlation coefficient between the state vector and each row of the centroid matrix, then assigned to that sub-window a binding site type corresponding to the epigenetic class with the highest correlation. Besides, twenty non-interacting 'gray' beads were added in each sub-window, so to match the number of beads per 5kb-bin of the PRISMR inferred polymer models. The described procedure results in an SBS polymer with 9 different binding domains, each of them interacting in a homotypic fashion. Afterward, we used the SBS model to calculate the predicted polymers' contact matrices and compared them with the independent Hi-C data (Figures S7G and S7H). As reflected by the Pearson and distance corrected Pearson correlations, in all cases, the contact pattern is well described (see for instance chromosomes 19 and 21 in Figure 5).

QUANTIFICATION AND STATISTICAL ANALYSIS

All the statistical tests employed are specified in the text and details provided in the method details section. Pearson, distance-corrected Pearson and stratum-adjusted correlation coefficients (SCC) from the HiCRep method (Yang et al., 2017) were used to compare experimental and simulated contact matrices. Pearson coefficients were also used to compare model binding sites with epigenetic features. Wilcoxon's rank-sum tests were applied to check the significance of binding domain overlaps and range of interaction, while Kolmogorov-Smirnov tests were used to compare the distributions of binding domains over the different chromosomes.