# Simulating topological domains in human chromosomes with a fitting-free model

# Simulating topological domains in human chromosomes with a fitting-free model

C. A. Brackley,[1] D. Michieletto,[1] F. Mouvet,[1] J. Johnson,[1] S. Kelly,[2] P. R. Cook,[3] and D. Marenduzzo[1]

[1]*SUPA, School of Physics & Astronomy, University of Edinburgh, Peter Guthrie Tait Road, Edinburgh, EH9 3FD, UK*

[2]*Department of Plant Sciences, University of Oxford, South Parks Road, Oxford OX1 3RB, UK*

[3]*Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford, OX1 3RE, UK*

## Abstract

We discuss a polymer model for the 3D organization of human chromosomes. A chromosome is represented by a string of beads, with each bead being "colored" according to 1D bioinformatic data (e.g., chromatin state, histone modification, GC content). Individual spheres (representing bi- and multi-valent transcription factors) can bind reversibly and selectively to beads with the appropriate color. During molecular dynamics simulations, the factors bind, and the string spontaneously folds into loops, rosettes, and topologically-associating domains (TADs). This organization occurs in the absence of any specified interactions between distant DNA segments, or between transcription factors. A comparison with Hi-C data shows that simulations predict the location of most boundaries between TADs correctly. The model is "fitting-free" in the sense that it does not use Hi-C data as an input; consequently, one of its strengths is that it can – in principle – be used to predict the 3D organization of any region of interest, or whole chromosome, in a given organism, or cell line, in the absence of existing Hi-C data. We discuss how this simple model might be refined to include more transcription factors and binding sites, and to correctly predict contacts between convergent CTCF binding sites.

**Hi-C: contact maps, domains and loops**

The conformations adopted by human chromosomes in 3D nuclear space are key contributors to gene activity in health and disease [1], and understanding the principles driving genome folding is one primary goal of biophysicists studying DNA. An important recent experimental breakthrough has been the development of chromosome conformation capture (3C), and its high-throughput derivative – "Hi-C" – which allows contacts between different chromatin segments to be mapped genome-wide [2–4].

Contact maps obtained using Hi-C reflect some underlying chromosomal organization. For example, each chromosome folds into distinct "topologically-associating domains" (TADs) during interphase (but not during mitosis when transcription ceases [5]). Domain size is variable, with higher-resolution studies typically uncovering smaller TADs in the range between 0.1-2 Mbp [3, 4]. TADs are largely specified by the local chromatin environment, as the same 20-Mbp region in a chromosomal fragment or an intact chromosome yield similar contact maps [4]. This organization into TADs is conserved, as they are found in budding yeast [6] and *Caulobacter crescentus*, where they are called "chromosomal interaction domains" or CIDs [7]. CIDs are also separated by strong promoters, and they are eliminated by inhibiting transcription.

Bioinformatic analysis suggests that eukaryotic TADs tend to be epigenetically determined; active and inactive regions typically form separate domains [2–4, 8], with CTCF (the CCCTC-binding transcription factor) and active transcription units (or binding sites for RNA polymerase II) being enriched at inter-domain "boundaries" [3, 4]. These analyses also uncover chromosome loops apparently stabilized by transcription factors bound to promoters and enhancers [4, 9–14], or CTCF bound to its convergent cognate sites (presumably the latter loops are tethered by associated cohesin complexes acting as a molecular "slip-link", or "hand-cuff") [3, 4]. Remarkably, many fewer loops are associated with divergent or parallel CTCF binding sites [4].

While Hi-C data is normally obtained using cell populations of milions of cells, single-cell Hi-C experiments show that no two cells in the same population share exactly the same contacts; nevertheless, the organization is non-random as certain contacts are seen more often than others [15].

These observations point to central roles for transcription orchestrating the 3D organi-

zation of chromosomes, with transcription factors providing molecular ties which stabilize the structure both locally and globally. The results also suggest that CTCF and cohesin are important organizers, with the latter providing an example of a molecular slip-link. Here we discuss results obtained using a simple biophysical model, which is based on the binding of two types of transcription factors to cognate sites on DNA. As we will see, molecular dynamics simulations using this model yield contact maps remarkably similar to those obtained from Hi-C. We further discuss how this model can be extended to incorporate more transcription factors, and molecular slip-links like cohesin.

## A toy model, and some basic principles

We first introduce a toy model which is schematically described in Figure 1A: a chromatin fiber (represented by a flexible bead-and-spring chain) interacts non-specifically with bi- or multi-valent spheres (this toy model is analogous to the "strings-and-binders" model of [16, 17]). The red spheres in Figure 1A represent transcription factors or complexes that can bind to two or more sites on the fiber; consequently they can form "molecular bridges" that stabilize loops. These factors stick to the chromatin fiber via a generic attractive interaction. If the interaction strength is large enough to allow multivalent binding, then the bound proteins spontaneously cluster, a phenomenon first observed and discussed in [18]. This clustering is accompanied by the formation of chromatin "domains", in which intra-domain contacts are enriched over inter-domain ones. The (generic) principle underlying clustering – which occurs in the complete absence of any specified DNA-DNA or protein-protein interaction – has been called the "bridging-induced attraction"– as it does not occur with univalent proteins that cannot stabilize loops [18–21].

The basic mechanism underlying this attraction is a simple thermodynamic positive feedback loop (Fig. 1B). First, proteins bind to chromatin, and – as they are at least bivalent – they can form a molecular bridge between two different DNA segments. This bridging brings distant parts of the chromosome together to increase the local chromatin concentration; this makes it more likely that additional proteins in the soluble pool will bind as they diffuse by. And once they have bound, these proteins will form additional molecular bridges which increase the chromatin concentration further. As this cycle repeats, protein clusters form, and these nucleate TAD-like structures. [We assume that the protein concentration

is sufficiently low that proteins cannot completely cover the fiber even when all bind. If, instead, the protein concentration is very large, then bridging induces macroscopic collapse of the whole fiber [16, 17, 22].]

In this simple case in which the transcription factors only bind non-specifically, the bridging-induced attraction yields clusters that continue to grow in size, ultimately giving one single cluster in steady state [22]. However, most transcription factors also bind specifically, as well as non-specifically. A simple modification of the toy model includes a stronger specific binding (of, e.g., red proteins to pink chromatin beads in Fig. 1C). Clusters still form via the bridging-induced attraction (Fig. 1C), but now they no longer grow indefinitely; instead, they reach a self-limiting size. This is because clustering of specifically-bound beads creates rosettes, or other structures with many chromatin loops, and bringing these together is entropically costly. Crucially, the entropic cost rises super-linearly with loop number, and this arrests cluster growth [18, 19, 23].

Another simple consequence of this generic organizing principle is that multivalent binding naturally creates "specialized" clusters. Imagine that two types of transcription factor (i.e., "red" and "green") bind specifically to different beads on the fiber (i.e., pink and light green; Fig. 1C). Then, the bridging-induced attraction works for the red and green factors separately. For instance, red factors increase the local concentration of pink chromatin binding sites, this recruits more red proteins, etc. Consequently, the clusters that emerge tend to contain either red factors plus pink beads or green factors plus light-green beads. If red and green proteins represent complexes containing RNA polymerase II and III respectively, this naturally explains why distinct foci/"factories" are seen in human cells that contain one or other enzyme, but not both [24]. As discussed in the next Section, a similar mechanism probably underlies the organization of the "A/B" compartments uncovered in Hi-C experiments [2].


**A minimal, fitting-free, polymer model for chromosome folding**

The toy model of Figure 1 was extended in [19] to give a minimal fitting-free predictive model for genome organization. The model is fitting-free because it is based solely on 1D information on the protein binding (or epigenetic) landscape . Thus, unlike other commonly-used approaches, it does not rely on contact information as an input, so its predictive power

4

is enhanced. In the version proposed in [19], the whole of chromosome 19 in GM12878 cells was modeled (Fig. 2A). In this case, each chromatin bead contained 3 kbp, and factors were of two types – "active" (modeling complexes of polymerases and transcription factors) or "inactive" (modeling heterochromatin-associated proteins like HP1$\alpha$, or even a simple linker histone like H1 – as both proteins are known to bind the genome in multiple places [25, 26]). Beads in the chromatin fiber are "colored" according to bioinformatic data to specify whether they bind the active or inactive proteins. Thus, active beads were colored using the "active" Broad ChromHMM tracks [27] [48] on the hg19 assembly (i.e., using states 1,4,5 in the HMM track that signify an "Active Promoter" or "Strong Enhancer" to specify strong binding, and states 9 and 10 that signify "Transcriptional Transition" or "Transcriptional Elongation" to specify weak binding). Inactive beads were colored using either the appropriate HMM tracks or GC content – the latter is illustrated here as a low GC content is such a good predictor of an inactive (heterochromatic) nature.

Given the simplicity of this model, it is striking to see how well it allows correct prediction of the positions of TADs and their boundaries (Fig. 2C,D). For example, 85% boundaries are correctly identified to within 100 kbp; some inter-domain interactions are even correctly captured (see the off-diagonal blocks in the contact maps). While this agreement can certainly be improved by adding biological detail, we stress that it is especially remarkable as it appears in a fitting-free minimal model (the only relevant parameters are interaction strengths and cut-offs, but little difference is found if these are set to ensure multivalent binding). The model can be applied, in principle, to any chromosome for which appropriate bioinformatic data is available (e.g., Broad ChromHMM track or histone modifications [28]); consequently, it can be used genome-wide in different cell lines and organisms. It can also be used to predict the contact map of any region of interest, and – of course – it can be applied at a higher resolution [28].

As in the toy model, active and inactive factors (and their cognate biding sites) cluster separately, and the model naturally yields the A (active) and B (inactive) compartments seen in Hi-C contact maps. Moreover, the proteins cluster to give structures reminiscent of both nuclear "bodies" (e.g., Cajal, polycomb and promyelocytic leukemia bodies), and factories containing RNA polymerases II and III – all structures rich in distinct proteins binding to different DNA sequences [29–32]. The number of protein clusters is significantly smaller than that of chromatin domains: therefore our model predicts that a number of

5

TADs will come together into a single protein cluster (say, a transcription factory), but different TADs might interact in different cells.

As these simulations reproduce the overall Hi-C organization well, it is of interest to ask what is special about beads at, or close to, boundaries between TADs. Figure 3 shows that the boundary beads *in silico* are depleted of inactive beads and enriched in active marks: this is consistent with bioinformatic analyses showing that boundaries are depleted in heterochromatic marks like HK39me3 and K3K27me3, and enriched in active ones like H3K4me3, as well as in transcription start sites and binding sites for RNA polymerase II [3]. An intriguing additional signal is that beads enriched at boundaries in silico are often non-binding beads – which naturally form boundaries as they possess few contacts; this is consistent with  15% Hi-C boundaries lacking any particular mark [3]. Finally, we note that, by using toy models, Refs. [19, 33, 34] showed that (permanent) chromatin loops (e.g., maintained by CTCF) may also act as boundaries, whose strength varies according to the force field used. This finding may be the reason why active beads are enriched at boundaries (they often constitute the base of loops, although these are dynamic ones).

**Beyond the minimal model: adding colors and slip-links**

The minimal model described this far generally yields contact maps like those obtained from Hi-C data [19]; however, exceptions do exist. In general, the percentage of TAD boundaries predicted accurately increases with transcriptional activity (the organization of chromosome 19 is predicted well, perhaps because it is the one containing the most active genes). In less-active regions, boundaries are sometimes predicted less accurately: e.g., Fig. 4A shows a region, in chromosome 14, where the minimal model fails at correctly predicting the location of some TADs (most of which are inactive). This raises the questions whether it is possible to improve the "coloring" of inactive beads, and/or add more colors. Capture Hi-C results provide a way of adding more colors. Thus, Mifsud et al. [9] distinguished contacts between promoters on the basis of their histone marks, and found there chromatin regions bearing the H3K9me3 or the H3K27me3 mark interacted with other regions with the same mark, whereas "mixed" contacts between K27 and K9 trimethylated regions were very rare. H3K9me3 binds HP1 to yield constitutive heterochromatin [35]. H3K27me3 is a classic inactive mark associated with facultative heterochromain and binding of polycomb-

group repressing complexes; it marks "blue chromatin" in Drosophila [36]. Therefore, we improved our model by stipulating that heterochromatic beads were classified according to histone modifications (instead of GC content), with two different colors for beads bearing the H3K9me3 or H3K27me3 mark [49]; we then also included in the simulations two proteins binding to these marks (modeling, e.g., PcG-protein complexes, such as PRC1 binding to H3K27me3 marks [21, 37]).

Figure 4B shows that, once the two different heterochromatin beads are distinguished, the simulation predicts TAD patterns more accurately. We stress that the refined model is still fitting free as it does not rely on Hi-C data for input, but only assumes knowledge of 1D protein binding landscape, or histone modification profiles.

Another (fitting-free) model similar in spirit to the one presented here is the "block-copolymer" model used to study folding of Drosophila chromosomes [38]. [For a non-fitting free version, see [39, 40].] In this case, chromatin beads interact directly, so bridging proteins are implied but not explicitly modeled. This approach is equivalent to the one used in Figures 2-4 if bridging proteins are abundant enough to saturate binding sites; however, the two models differ in the regime where only some binding sites are occupied. The model used in Figures 2-4 also naturally explains the formation of nuclear bodies, and so can be used to study their biogenesis and kinetics (this is not possible with the block-copolymer model where bridging proteins are assumed to be uniformly distributed at all times).

A recent study by Chiariello et al. [41] offers another avenue to improve simulation accuracy by using some information from Hi-C experiments (but then the model is not fitting free). In practice this is done through an iterative procedure which finds the minimal arrangement of binding sites and colors which best explain the Hi-C contact map; for example, simulations involving 16 colors gave contact maps for the Sox6 locus that were indistinguishable from those obtained by Hi-C (correlation coefficient 95%).

An important unaddressed aspect concerns loops (or "loop domains") stabilized by CTCF [4]. As discussed above, CTCF is more likely to bridge two cognate binding sites [4, 42] when sites are in a "convergent" orientation compared to a "divergent" one. Polymer models to explain this have been proposed [43, 44]; they involve loop-extrusion factors and slip-links that are simultaneously bound (linked) to beads on two different chromosomal segments and which can slide (slip) along the segments (in practice, these factors/slip-links are cohesin and/or condensin). These models can account for the observed CTCF orientation bias, as

7

they assume that the loop-extrusion factors can stably stick only to one side of CTCF (which is true of cohesin). However, these models also require some as-yet undiscovered motor protein with a processivity sufficient to generate loops of hundreds of kb. Moreover, CTCF and its convergent sites cannot be the sole organizer of boundaries, as knock-outs of CTCF have only minor effects on domain organization in mammals [45–47], and bacteria possess domains but no equivalent of CTCF. Nevertheless, CTCF directionality and cohesins clearly play an important role in the formation and establishment of several eukaryotic loops, so it will be of interest to incorporate these components into our model.

## Acknowledgements

---

[1] Cavalli, G., and Misteli, T. (2013). Functional implications of genome topology. *Nat. Struct. Mol. Biol.* **20**, 290-299.

[2] 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.

[3] 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.

[4] Rao, S.S.P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S. and Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1-16.

[5] Naumova, N., Imakaev, M., Fudenberg, G., Zhan, Y., Lajoie, B.R., Mirny, L.A., and Dekker, J. (2013). Organization of the mitotic chromosome. *Science* **342**, 948-953.

[6] Hsieh, T.-H., S., Weiner, A., Lajoie, B., Dekker, J., Friedman, N., and Rando, O.J. (2015). Mapping nucleosome resolution chromosome folding in yeast by micro-C. *Cell* **162**, 108-119.

[7] Le, T.B., Imakaev, M.V., Mirny, L.A., and Laub, M.T. (2013). High-resolution mapping of the spatial organization of a bacterial chromosome. *Science* **342**, 731-734.

[8] Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**, 458-472.

[9] Mifsud, B., Tavares-Cadete, F., Young, A. N, Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S. W., Andrews, S., Grey, W., Ewels, P. A. *et al.* (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Gen.* **47**, 598-606.

[10] Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., and de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.* **38**, 1348-1354.

[11] Li, G., Ruan, X., Auerbach, R.K., Sandhu, K.S., Zheng, M., Wang, P., Poh, H.M., Goh, Y., Lim, J., Zhang, J. *et al.* (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84-98.

[12] Jin, F., Li, Y., Dixon, J. R., Selvaraj, S., Ye, Z., Lee, A. Y., Yen, C. A., Schmitt, A. D., Espinoza, C. A., and Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290-294.

[13] Zhang, Y., Wong, C. H., Birnbaum, R. Y., Li, G. L., Favaro, R., Ngan, C .Y., Lim, J., Tai, E., Poh, H. M., Wong, E. *et al.* (2013). Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. *Nature* **503**, 290-294.

[14] Heidari, N., Phanstiel, D. H., He, C., Grubert, F., Jahanbani, F., Kasowski, M., Zhang, M. Q., and Snyder, M. P. (2014). Genome-wide map of regulatory interactions in the human genome. *Genome Res.* **24**, 1905-1917.

[15] Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E. D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59-64.

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

[17] Barbieri, M., Chotalia, M., Fraser, J., Lavitas, L.M., Dostie, J., Pombo, A., and Nicodemi, M. (2012). Complexity of chromatin folding is captured by the strings and binders switch model. *Proc. Natl. Acad. Sci. USA* **109**, 16173-16178.

[18] Brackley, C. A, Taylor, S., Papantonis, A., Cook, P. R., and Marenduzzo, D. (2013). Non-specific bridging-induced attraction drives clustering of DNA-binding proteins and genome organization. *Proc. Natl. Acad. Sci. USA* **110**, E3605-3611.

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

[20] Le Treut, G., Kepes, F., and Orland, H. (2016). Phase Behavior of DNA in the Presence of DNA-Binding Proteins *Biophys. J.* **110**, 51-62 (2016).

[21] Michieletto, D., Marenduzzo, D., and Wani, A. H. (2016) Chromosome-wide simulations uncover folding pathway and 3D organization of interphase chromosomes. arXiv:1604.03041.

[22] Johnson, J., Brackley, C.A., Cook, P.R., and Marenduzzo, D. (2015). A simple model for DNA bridging proteins and bacterial or human genomes: bridging-induced attraction and genome compaction. *J. Phys. Condens. Matter* **27**, 064119.

[23] Marenduzzo, D. and Orlandini, E. (2009). Topological and entropic repulsion in b iopolymers. *J. Stat. Mech.*, L09002.

[24] Xu, M., and Cook, P. R. (2008). Similar active genes cluster in specialized transcription factories. *J. Cell. Biol.* **181**, 615-623.

[25] Kilic, S., Bachmann, A.L., Bryan, L.C., and Fierz, B. (2015). Multivalency governs HP1$\alpha$ association dynamics with the silent chromatin state. *Nat. Commun.* **6**, 7313.

[26] Mack, A. H., Schlingman, D. J., R D Salinas, R. D., Regan, L., and Mochrie, S. G. J. (2015) Condensation transition and forced unravelling of DNA-histone H1 toroids: a multi-state free energy landscape. *J. Phys.: Cond. Matt.* **27**, 061406.

[27] 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.

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

[29] Sleeman, J.E., and Trinkle-Mulcahy, L. (2014). Nuclear bodies: new insights into assembly/dynamics and disease relevance. *Curr. Opin. Cell Biol.* **28**, 76-83.

[30] Pombo, A., Jackson, D.A., Hollinshead, M., Wang, Z., Roeder, R.G., and Cook, P.R. (1999).

Regional specialization in human nuclei: visualization of discrete sites of transcription by RNA polymerase III. *EMBO J.* **18**, 2241-2253.

[31] Cook, P.R. (1999). The organization of replication and transcription. *Science* **284**, 1790-1795.

[32] Papantonis, A., and Cook, P.R. (2013). Transcription factories; genome organization and gene regulation. *Chem. Rev.* **113**, 8683-8705.

[33] Benedetti F., Dorier J., Burnier Y., and Stasiak A. (2014). Models that include supercoiling of topological domains reproduce several known features of interphase chromosomes. *Nucleic Acids Res* **42**, 2848-2855.

[34] Hofmann, A., and Heermann, D. W. (2015). The role of loops on the order of eukaryotes and prokaryotes. *FEBS Lett.* **589**, 2958–2965.

[35] Lehnertz, B, Ueda, Y,, Derijck, A. A, Braunschweig, U., Perez-Burgos, L., Kubicek, S., Chen, T., Li, E., Jenuwein, T., and Peters, A. H. (2003). Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. *Curr. Biol.* **13**, 1192-1200.

[36] Filion, G. J., van Bemmel, J. G., Braunschweig, U., Talhout, W., Kind, J., Ward, L. D., Brugman, W., de Castro, I. J., Kerkhoven, R. M., Bussemaker, H. J. and van Steensel, B. (2010). Systematic Protein Location Mapping Reveals Five Principal Chromatin Types in Drosophila Cells. *Cell* **143**, 212-224.

[37] Wani, A. H., Boettiger, A.-N., Schorderet, P., Ergun, A., Munger, C., Sadreyev, R. I., Zhuang, X., Kingston, R. E., and Francis, N. J. (2016). Chromatin topology is coupled to Polycomb group protein subnuclear organization. *Nat. Comm.* **7**, 10291.

[38] 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.

[39] Giorgetti, L., Galupa, R., Nora, E.P., Piolot, T., Lam, F., Dekker, J., Tiana, G., and Heard, E. (2014). Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell* **157**, 950-963.

[40] Tiana, G., Amitai, A., Pollex, T., Piolot, T., Holcman, D., Heard, E., Giorgetti, L. (2016). Structural Fluctuations of the Chromatin Fiber within Topologically Associating Domains. *Biophys. J.* **110**, 1234-1245.

[41] Chiariello, A. M., Annunziatella, C., Bianco, S., Esposito, A., and Nicodemi, M. (2016).

Polymer physics of chromosome large-scale 3D organisation. *Sci. Rep.* **6**, 29775.

[42] Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J. H., Gorkin, D. U., Jung, I., Wu, H. Y., Zhai, Y. N., and Tang, Y. X. *et al.* (2015). CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell.* **162**, 900-910.

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

[44] Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., and Mirny, L. A. (2016) Formation of Chromosomal Domains by Loop Extrusion. *Cell Reports* **15**, 2038-2049.

[45] Zuin, J., Dixon, J. R., can der Reijden, M. I. J. A., Ye, Z., Kolovos, P., Brouwer, R. W. W., van de Corput, M. P. C., van de Werken, H. J. G., Knoch, T. A., van IJcken, W. F. J. *et al.* Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc. Natl. Acad. Sci. USA* **111**, 996-1001 (2014).

[46] Hou, C., Dale, R., and Dean, A. (2010) Cell type specificity of chromatin organization mediated by CTCF and cohesin. *Proc. Natl. Acad. Sci. USA* **107**, 3651-3656.

[47] Seitan, V. C., Faure, A. J., Zhan, Y., McCord, R. P., Lajoie, B. R., Ing-Simmons, E., Lenhard, B., Giorgetti, L., Heard, E., Fisher, A. G., Flicek, P., Dekker, J., and Merkenschlager, M. (2013). Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments. *Gen. Res.* **23**, 2066-2077.

[48] The Broad ChromHMM track is available on the UCSC Genome Browser. To build it, several data sets for histone modification and protein binding have been analyzed using a hidden Markov model to classify chromatin regions as being in one of several chromatin "states".

[49] In practice, we used a threshold in histone modification tracks to color beads, but the exact value of the threshold played a minor role in the results.
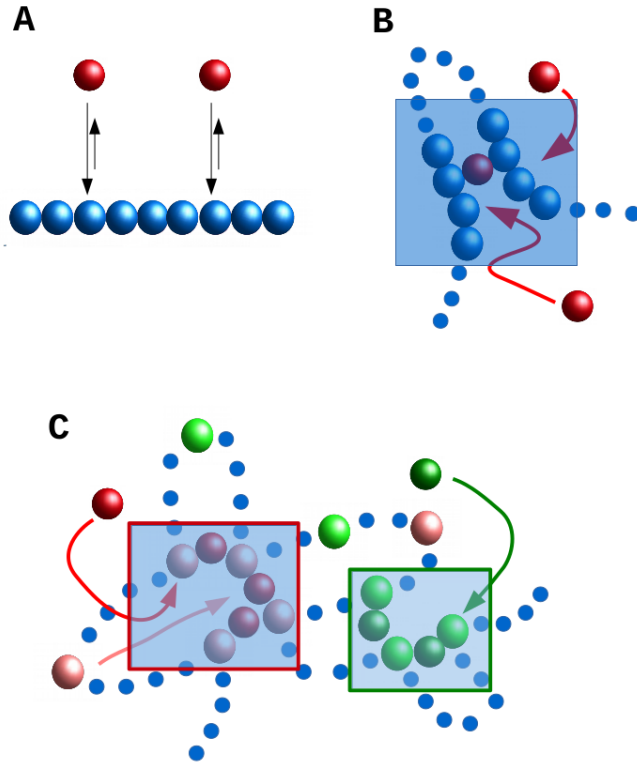
FIG. 1: Schematic representation of the toy model discussed in the text. (A) A chromatin fiber is coarse-grained into a bead-and-spring polymer, where monomers are spherical (blue beads). Proteins (red beads) bind to the chromatin fiber non-specifically (arrows). (B) As proteins are multivalent, upon binding they can create molecular bridges: here the bound red protein contacts two blue chromatin beads, and this increases the local chromatin density (shaded area): therefore, other proteins in the soluble pool are more likely to bind chromatin in this area. This will, in turn, further increase chromatin density creating a (thermodynamic) positive feedback loop which eventually leads to the formation of protein clusters (concomitantly with TAD-like chromatin domains). (C) Schematic of a toy model with specific binding. Now red proteins bind specifically to pink chromatin beads, and green proteins to light-green chromatin beads. As proteins are multivalent, and because pink and light-green beads lie at different places along the fiber, a similar positive feedback as in (B) separately drives the increase of local concentration of pink and light-green chromatin beads (in the two shaded ares), which eventually leads to the formation of specialized clusters of red proteins and pink chromatin binding beads, and of green proteins and light-green binding beads.
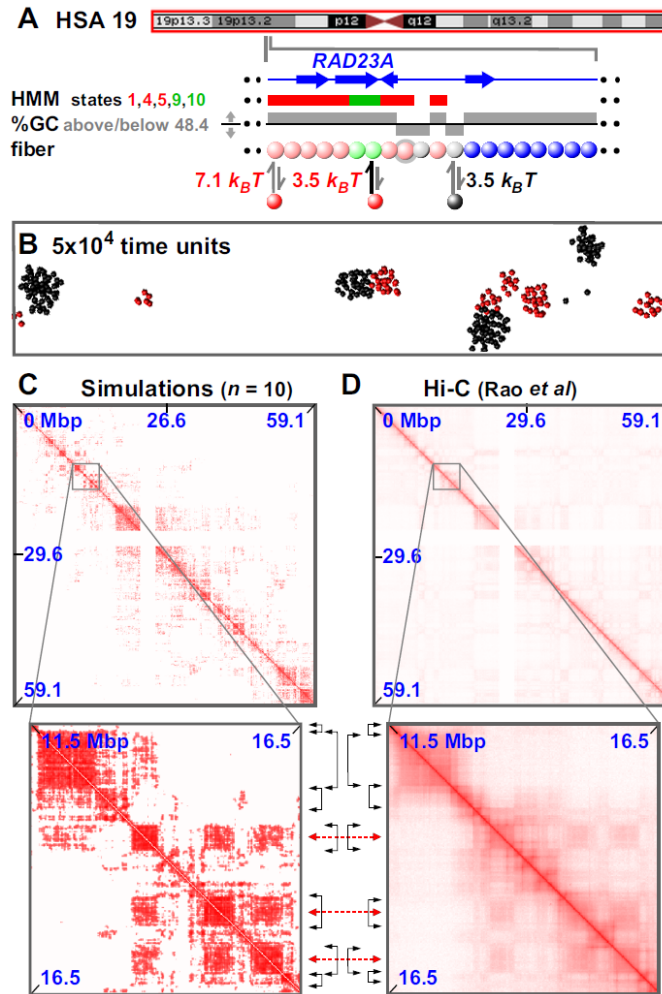
FIG. 2: Fitting-free simulations of chromosome 19 in GM12878 cells. (A) Overview. The ideogram (red box indicates the whole chromosome that was simulated) and Broad HMM track (colored regions reflect chromatin states) are from the UCSC browser; the zoom illustrates an arbitrary region, around *RAD23A*, to show the details of the "coloring". Beads (3 kbp) are colored according to HMM state and GC content: blue beads are non-binding; pink beads correspond to states 1,4,5 in the ChromHMM track; light-green to states 9,10. Grey beads correspond to beads which have <48.4% GC. Pink and light-green beads bind (respectively, strongly and weakly) active factors (red in the figure); grey beads bind to inactive factors, linked to heterochromatization (black in the figure). Note that the coloring rule is such that beads can have multiple colors: for instance, in the zoom two pink beads are also grey (represented by grey halos), so that such beads can bind both red and black factors. (B) Snapshot (without chromatin) of central region after $5\times10^4$ units; most clusters contain factors (or proteins) of one color. In other words, active and inactive proteins cluster separately. As discussed in the text, the formation of specialised clusters may underlie both the formation of A/B compartments (when looking at the chromatin interactions) and that of some nuclear bodies (when looking at the protein cluster patterns). (C,D) Comparison between contact maps from simulations and experiments (see Ref. [19] for more details). Between zooms, black double-headed arrows mark boundaries of prominent domains (on the diagonal), and red double-headed ones the centers of off-diagonal blocks making many inter-domain contacts. Reproduced from Ref. [19], with permission.
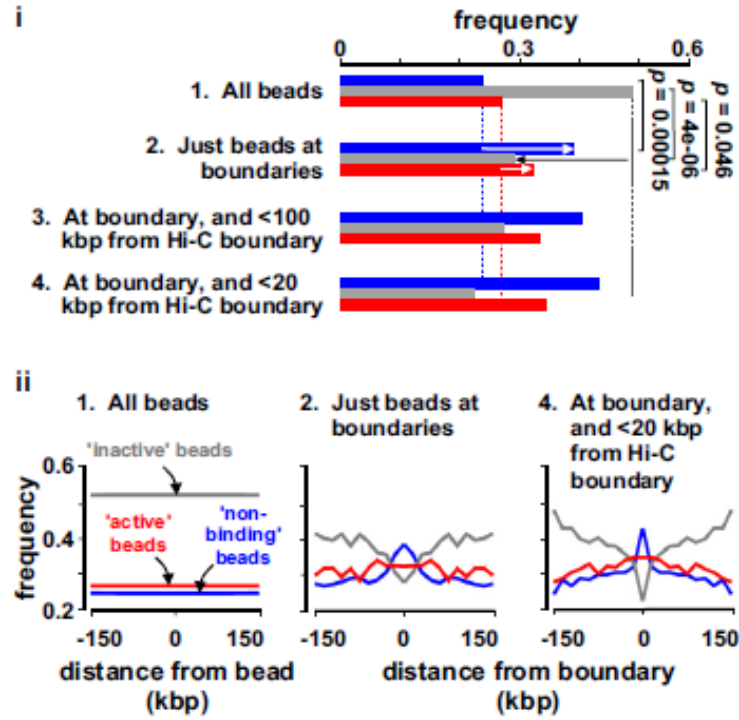
FIG. 3: Characterization of TAD boundaries found *in silico*. These plots are obtained by analyzing the TAD boundaries found in simulations (through combination of an automated method and visual inspection [19]), and by computing the frequencies of non-binding (blue bars), inactive (grey bars) and active (red bars) beads in different sets. Set 1: all beads. Set 2: Beads lying within 100 kbp of a boundary. Sets 3 and 4: The sub-sets of set 2 that also lie within 100 and 20 kbp of a boundary identified in Hi-C data. (i) Beads at boundaries are rich in active and non-binding beads, and depleted of inactive beads (arrows; p values assessed assuming Poisson distributions). (ii) The frequencies of different beads (in sets 1, 2 and 4) in the 150 kbp on each side of either each bead in set 1, or of boundaries in sets 2 and 4. Adapted from Ref. [19].
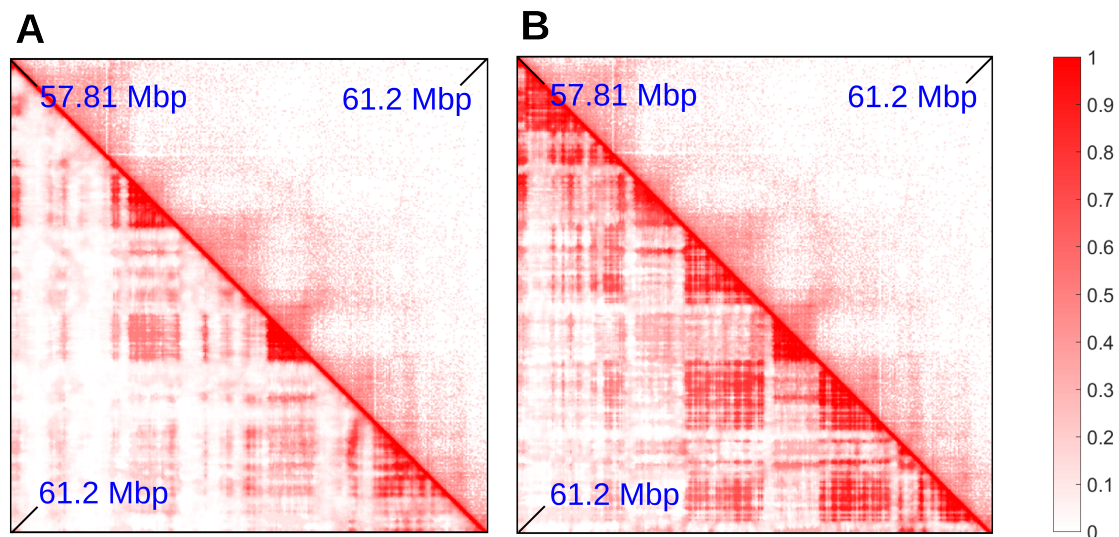
FIG. 4: Adding "colors" to the minimal model. (A,B) Comparison between Hi-C (top triangle) and simulated (bottom triangle) contact maps, for the region between 57.81 and 61.2 Mbp in chromosome 14 in HUVEC cells (coordinates from hg19). Simulations were done similarly to those in Figure 2, and involved 15.5 Mbp of chromatin at 3 kbp resolution, so the region shown is a subset of the whole simulated fragment, chosen to highlight the effect of adding a new species of protein and an additional binding site color to the model. In (A), heterochromatin was colored according to GC content (threshold $\sim$ 40.69%). It can be seen that several TADs are missing in the simulations. In (B), heterochromatin beads are colored according to H3K9me3 and H3K27me3 tracks (so there are now two possible heterochromatic colors). The latter procedure gives better agreement with the Hi-C data.