Modelling Genome-wide Topological Associating Domains in Mouse Embryonic Stem Cells

Y. Zhan[1], L. Giorgetti[1*] and G. Tiana[2*]
[1] Friedrich Miescher Institute for Biomedical Research, Basel, CH-4058, Switzerland
[2]Center for Complexity and Biosystems and Department of Physics, Università degli Studi di Milano and INFN, Milano, I-20133, Italy
* luca.giorgetti@fmi.ch, guido.tiana@unimi.it

Abstract

Chromosome conformation capture (3C)-based techniques such as 5C and Hi-C revealed that the folding of mammalian chromosomes is highly hierarchical. A fundamental structural unit in the hierarchy is represented by topologically associating domain (TADs), sub-megabase regions of the genome within which the chromatin fibre preferentially interacts. 3C-based methods provide the mean contact probabilities between chromosomal loci, averaged over a large number of cells, and do not give immediate access to the single-cell conformations of the chromatin fibre. However, coarse-grained polymer models based on 5C data can be used to extract the single-cell conformations of single TADs. Here we extend this approach to analyse around 2500 TADs in murine embryonic stem cells based on high-resolution Hi-C data. This allowed to predict the cell-to-cell variability in single contacts within TADs genome-wide, and correlations between them. Based on these results, we predict that TADs are more similar to ideal chains than to globules in terms of their physical size and three-dimensional shape distribution. Furthermore, we show that their physical size and the degree of structural anisotropy of single TADs are correlated with the level of transcriptional activity of the genes that it harbours. Finally, we show that a large number of multiplets of genomic loci co-localise more often than expected by random, and these loci are particularly enriched in promoters, enhancers and CTCF-bound sites. These results provide the first genome-wide structural reconstruction of TADs using polymeric models obeying the laws of thermodynamics, and reveal important universal trends in the correlation between chromosome structure and transcription.

Keyword: chromatin model, Hi-C data, promoter-enhancer interaction, CTCF, big data

Introduction

The three-dimensional organisation of chromatin is thought to play an important role in the transcriptional regulation of gene expression (Hübner et al. 2013). In recent years, experiments based on chromosome conformation capture (3C, and the derived 4C, 5C and Hi-C, reviewed in (Denker and de Laat 2016)) shed light into the structure of chromatin in the kilobase to megabase scale, where regulatory enhancer-promoter interactions occur, and revealed the existence of a new level of chromatin organisation called topologically associating domains (TADs) (Dixon et al. 2012; Nora et al. 2012). The biological importance of TADs, and their internal sub-structures (Rao et al. 2014), stem from the fact that they seem to act as regulatory microenvironments by controlling the interaction frequencies of enhancers and promoters, and consequently contributing to the establishment of correct transcriptional levels and patterns of gene expression.

3C-based techniques provide maps of the contact probabilities between pairs of genomic loci, averaged over millions of cells. However, this averaged information gives only a partial view of the complicated behaviour of the chromatin fiber, and misses several properties that characterize chromatin folding within TADs, which might be important for transcriptional control. An example of those properties is the degree at which different pairs of loci co-localize. If Hi-C maps provide information on how often a locus A is in contact with locus B, and how often locus B is in contact with locus C, it does not quantify explicitly how often these two contacts occur simultaneously unless more sophisticated approaches are used (Olivares-Chauvet et al. 2016). Moreover, it is not possible to extract any information about the shape and the three-dimensional size of a TAD directly from Hi-C maps. Finally, fluctuations over time of any conformational properties of the fibre cannot come directly from the experimental data.

Polymer models (Baù et al. 2010; Mirny 2011; Barbieri et al. 2012; Brackley et al. 2013; Benedetti et al. 2013; Rosa and Zimmer 2014; Giorgetti et al. 2014; Johnson et al. 2015; Goloborodko et al. 2016) are a useful tool to increase the amount of information that can be derived from 3C-based experiments, and 5C and Hi-C interaction maps in particular. We previously showed that a coarse-grained model of chromatin, in which a chain of beads interact through a contact potential tuned to reproduce the experimental contact properties of a given TAD, is able to extract realistic 3D conformations of the chromatin fibre in single cells (Giorgetti et al. 2014). This model is based on the minimal assumption that chromosomal loci interact through a two-body, short-range interaction, and remain agnostic on the actual molecular mechanisms underlying it (e.g. direct looping vs. loop extrusion between sites bound by CTCF(Merkenschlager and Nora 2016), provided that the resulting interaction can be approximated by this simple model). Nevertheless, the conformations derived were shown to correctly predict independent observables, such as the distribution of distances between pairs of loci measured by DNA fluorescence in situ hybridisation (FISH), and how it is affected by mutations (Giorgetti et al. 2014; Tiana et al. 2016). Moreover, this model is thermodynamically correct, in the sense that the polymer obeys the laws of motion imposed by physics.

Co-localization of multiple loci was clearly observed in the X inactivation centre of mouse ESC cells by DNA-FISH (Nora et al. 2012). The present model proved efficient in identifying multi-body interactions in the same system, specifically those involving the *Linx*, *Tsix* and *Chic1* loci, whose co-localization was then verified by DNA-FISH (Giorgetti et al. 2014). Interestingly, also a different model, based on the correct thermodynamics of proteins as well, was able to highlight many-body contacts in the *Sox9* locus (Chiariello et al. 2016), thus testifying to the robustness of this phenomenon. In a recent work (Olivares-Chauvet et al. 2016) 3C-based techniques were extended to identify experimentally multiple interactions.

In the present work, we have extended this approach to study the conformational properties of TADs and their internal sub-structures at the genome-wide level in mouse embryonic stem cells, using Hi-C data at 20kb resolution (Giorgetti et al. 2016) as an input for the model. TADs were identified using a custom algorithm [Zhan et al, submitted]. For each domain that is bigger than 240kb (a total of 2473 TADs), a polymer model was derived and its equilibrium properties were investigated and correlated with the activity of the genes contained within them.

The Model

The starting point of the model is a mouse ESC Hi-C map M_{ij} at a resolution of 20 kb, from Ref. (Giorgetti et al. 2016), which was iteratively corrected using the ICE method (Imakaev et al. 2012). All samples were mapped to mouse mm9 using QuasR (Gaidatzis et al. 2015). Hi-C read counts M_{ij} were converted into contact probabilities p_{ij} that are taken as input by the model following a procedure similar to (Giorgetti et al. 2014) (see Supplementary Materials). TADs were identified using the CaTCH algorithm (Zhan et al., manuscript submitted (provided to the reviewers)), which we developed to identify and compare hierarchies of nested interaction domains in Hi-C maps. We identified TADs as the set of domains that is maximally similar to the domains called on the same dataset using the directionality index method (Dixon et al. 2012). TADs simulated in this work have a slightly smaller median size than directionality index TADs (640kb vs. 800 kb), due to the fact that the CaTCH algorithm often splits large, inactive directionality index TADs in two sub-domains (Zhan et al). The complete list of simulated domain coordinates is provided in Supplementary Table S1.

In this work, each TAD is studied independently of the others; see the ‘Size and shape of TADs’ section below for a discussion on the validity of this assumption. Each region is modelled as an inextensible chain of beads. A bead represents effectively a genomic segment of 20 kb *(*which would correspond to ≈134 nm in euchromatic regions, based the estimations discussed in (Giorgetti et al. 2014), corresponding to the resolution of the Hi-C experiments (see Fig. 1a)). We termed a the distance between consecutive beads, which sets the elementary length unit of the model; all lengths in the study will be thus given in units of a. An intrinsic assumption of the model is that such a distance is constant along the genome, and thus that the internal structure of a 20kb-element does not change dramatically from one locus to another. However, importantly, it should be noted that the actual spatial extension of each bead (i.e. the number of nm that it occupies in space) could depend on the particular genomic location and notably on the local chromatin composition (notably eu- vs. heterochromatin).

Pairs of beads *i* and *j* interact with a spherical-well potential

$$u_{ij} = \begin{pmatrix} +\infty & if\ |r_i - r_j| < R_{HC} \\ B_{ij} & if\ R_{HC} < |r_i - r_j| < R \\ 0 & |r_i - r_j| > R \end{pmatrix}$$

where $\left|r_i - r_j\right|$ is the Euclidean distance between the positions of beads i and j, $R_{HC}=0.24a$ is a hard-core radius which sets the diameter of the fibre and $R=0.6a$ is the interaction range (see Fig. 1b).

The matrix B_{ij} contains the interaction energies between each pair of beads, which summarize the effect of all biomolecules (transcription factors, architectural proteins, histone modifications, etc.) associated with the corresponding 20 kb segments, which might mediate such interaction energies. The interaction energy matrix is optimized according to the algorithm described in (Giorgetti et al. 2014) to minimize the χ^2 between the experimental Hi-C map and the one that is back-calculated from the model. The variance between two experimental replicates (Giorgetti et al. 2016), binned at 40kb resolution, was used to weigh the χ^2 measure.

The model was applied to each TADs of the 20 mouse chromosomes larger than 12 model beads (2473) (see Fig. 1c and d) and iterated until the χ^2 between the Hi-C map and the model contact map converges to 1. The final set of conformations associated with $\chi^2\approx1$ represents the most likely equilibrium ensemble of the TAD compatible with the experimental data (Tiana et al. 2016). These ensembles of conformations are then analysed.

Size and shape of TADs

We started by analysing the volume of space occupied by polymer models of the various TADs. Specifically, we quantified the compactness of the polymer (in units of the bead size a) using the root mean square radius of gyration s (see Eq. S4 in the Supplementary Materials), which represents the radius of a hypothetical sphere that contains the typical conformation of the TAD. The value of s as a function of the linear size N of each simulated TAD is displayed in Fig. 2a. The data appear to follow the behaviour of an ideal non-interacting chain, in which s scales as $s{\sim}N^{1/2}$ (Grosberg and Khokhlov 1994), but is also compatible with the scaling of a compact globule ($s{\sim}N^{1/3}$) (solid and dashed line, respectively, in Fig. 2a). However, fluctuations follow a linear dependence on s (Fig. 2b), which is typical of an ideal chain (see Eq. S6 in the Supplementary Materials) and not of a globule, in which σ_s decreases with s. The polymer behaviour of chromatin within single TADs, thus, seems to be largely similar to that of an ideal chain. This in fact is not unexpected, because it refers to small chromosomal segments belonging to a much longer chromosome polymer, confined by the nuclear envelope and by its interactions with other chromosomes into a chromosomal territory of finite volume. The Flory theorem (Doi and Edwards 1986) states that in such a situation, the forces acting on the small segment cancel out and the segment behaves as an ideal chain.

The ideal-chain behaviour is also compatible with our previous observation, based on 5C-based reconstruction of the *Tsix* and *Xist* TADs in ESCs, that chromatin behaves as an ideal chain in those two TADs (Tiana et al. 2016). Our previous calculations showed that the heterogeneous interactions optimized to fit the 5C data set the polymer model of the chromatin fiber close to the theta-point of the system, i.e. the temperature at which all polymers behave ideally. We now provide evidence that this seems to be the case at the genome-wide level.

Other properties that can be extracted from TAD structural models are however far from those of an ideal chain. Notably, the average number of contacts between different monomers in a single TAD is displayed in Fig. 2c. In an ideal chain, the number of contacts proportional would be expected to scale as $N^{1/2}$ (see Eq. S8 in the Supplementary Materials); however, in spite of the large variability, this does not seem compatible with the calculated data (solid curve in Fig. 2c) where the number of contacts seems to grow linearly with the physical size of the TAD (dashed curve in Fig. 2c). This is similar to the behaviour of a homopolymer that is collapsed into a globule (Lifshitz et al. 1978).

Another interesting property of polymers is their degree of isotropy, which measures to which extent the conformations of the polymer are spherically symmetric. The isotropy of a conformation can be quantified by a parameter λ that assumes values close to 0 in case of a perfectly spherical symmetry and 1 for a linear conformation of the chain (see Supplementary Materials). The average values of λ for the different TADs are displayed in Fig. 2d as a function of their gyration radius, normalized by $N^{1/2}$, in order to evaluate their compactness independently on the genomic length of the TAD. As a reference, an ideal chain is typically rather anisotropic, displaying $\lambda=0.85\pm0.13$ (cf. solid line in Fig. 2d), while a globular homopolymer displays $\lambda=0.25\pm0.10$. The values of λ for simulated TADs are in between these two reference cases, but overall show increased similarity to an ideal chain rather than a globule. The spatial isotropy shows only a weak dependence on the compactness of the chain, increasing slightly in less compact chains.

The isotropy of TADs was observed experimentally by super-resolution imaging (Boettiger et al. 2016) and as result of simulations of a simple polymer model, different from ours, of the Hox-B locus (Annunziatella et al. 2016).

In summary, although single TADs display properties (such as the number of contacts) that are more similar to those of homopolymer globules, overall their compactness, fluctuations and, (roughly) their anisotropy are more similar to those of an ideal chain. This is likely to reflect the fact that unlike homopolymers, a chromosomal region of the size of a TAD displays complex and heterogeneous interactions (in part mediated by proteins such as CTCF that promote chromosomal interactions), which are distributed unevenly along the chain and can be responsible for properties which depart considerably from those of homopolymers (Zhan et al. 2016).

Relation between structural properties of TADs and gene activity

We then set out to address the question whether any of the polymer properties defined in the previous section is related to level of transcriptional activity within a TAD. We quantified the level of expression of genes in ESC samples matched with the Hi-C data using published RNA-seq datasets (Gendrel et al. 2014). Expression levels were estimated by quantifying the RPKM (Reads Per Kilobase per Million mapped reads) of each gene:

$$RPKM = \frac{M}{N*L} \cdot 10^3 \cdot 10^6,$$

where M is the mapped reads to the gene (exons), L is the length of the region (sum of all exons) and N is the total number of mapped reads. We then defined the overall transcriptional activity of a TAD as the sum of the RPKM of each gene belonging to that TAD. The boxplot of Fig. 3a displays TAD activity as a function of the mean gyration radius s and shows that transcriptional activity appears to decrease with increasing gyration radius (in monomer units a). This effect can also be observed when plotting the gyration radius of the active and of the inactive TADs, respectively, as a function of the genomic size (see Fig. 3b; the curve, that describes the behaviour of active TADs, displays a scaling exponent as a function of genomic size of 0.31, which is slightly smaller the value 0.35 of inactive TADs. The transcriptional activity within TADs also decreases with TAD spatial size, measured in units of a (see Fig. S1 in the Supplementary Materials). This behaviour seems to contradict previous results (le Dily et al. 2014; Fabre et al. 2015; Boettiger et al. 2016), but is difficult to interpret for two reasons: on the one hand, TADs with different genomic size N could have similar gyration radii, depending on their internal interactions and consequent polymer state; on the other hand, in this analysis we implicitly assumed that the density of bases (i.e. the spatial size of a bead) is constant across the genome, whereas (as discussed above) this could not be the case and might in fact critically change in euchromatic vs. heterochromatic regions.

To better understand structural correlates of transcriptional activity, we plotted in Fig. 3c the RPKM as a function of the relative fluctuations of the gyration radius quantified by its normalized standard deviation σ_s/s, which is an adimensional quantity that does not depend on the actual physical size of a monomer. This quantity provides, independently on the genomic size of a TAD, information about the phase of the polymer. If the relative fluctuations are small, the polymer is globular; fluctuations increase as the polymer swells into a coil (Lifshitz et al. 1978). As shown in Fig. 3c the transcriptional activity appears to decrease as the TAD conformation swells.

Finally, Figure 3d shows that the genetic activity within TADs is positively correlated with the mean linearity λ of polymer conformations.

It should be noted that if the chain were described by a simple homopolymeric theory, the correlation of transcriptional activity of TADs with their anisotropy and the anti-correlation with their relative fluctuations would be contradictory, because globules are more spherical than coils (cf. red lines in Fig. 2d). However, the heterogeneity of the interactions allows TADs to display both these features.

The contact probability between any pairs of loci (cf. Fig. 4a) can be estimated, according to its thermodynamic interpretation (Fudenberg and Mirny 2012) from the normalized Hi-C matrix M_{ij} without the need of any polymer model. The average $\bar{M}$ of all pairs of each TAD correlates with the activity of the TAD (see Fig. 4b; r=0.28; p-value=10^{-16}). Notably, transcriptional activity increases with $\bar{M}$ for $\bar{M} \leq 70$ and then reaches a plateau. This effect is mainly due to pairs of loci that are at relatively small genomic distances, since transcriptional activity correlates well with the mean of the Hi-C elements restricted to pairs that are separated by 100 kb (cf. Fig. 4c), but much less with the mean over pairs separated by 600 kb (cf. Fig. 4d).

Using the TAD polymer models we also calculated the Euclidean three-dimensional distance between pairs of genomic loci. Consistent with what observed at the level of average contacts, TAD activity positively correlates with the average distance between genomic loci in the small-distance range (100kb, cf. Fig. 4e). This reflects the fact that at small genomic distances, contact probabilities and mean distances are well correlated (see upper panel in Fig. S2 in the Supplementary Materials). However, this is not the case for pairs of loci located at large genomic distance (e.g. 600 kb), whose spatial distances are much less correlated with contact probabilities (see lower panel in Fig. S4). This discordance hints at the fact that comparing Hi-C results and distance measurements (for example by DNA FISH) is not a trivial task (Giorgetti and Heard 2016; Dekker 2016) , and suggests that this might be true especially at relatively large, TAD-scale genomic distances (several hundred kb)(Fudenberg and Imakaev 2016).

In summary, highly active TADs appear to be more polarized along a specific axis (‘elongated’) than inactive ones, to display a smaller amount of structural fluctuations relative to their overall size, and to display a higher amount of short-range contacts (in the 100-kb range). This reflects the higher compartmentalization of active TADs in highly interacting, hierarchical sub-domains of preferential interactions reflecting CTCF-mediated sub-TAD architectures and interactions between active regulatory elements (Rao et al. 2014; Zhang et al.).

Co-localised loci are enriched in active promoters and enhancers

While Hi-C data inform only on the probability that two loci are in contact over the cell population, polymer models can also predict the probability that multiplets of loci are found in contact with each other simultaneously. For each TAD, we calculated the probability p_{ijk} that the triplet of loci i, j, k are in contact (cf. Fig. 4a), and further computed the quantity $p^*_{ijk} = p_{ijk} - \left(p_{ij}p_{jk} + p_{ik}p_{kj} + p_{ki}p_{ij}\right)/3$ expressing the excess probability that the three loci co-localise with respect to the probability that the three of them come in contact independently on each other. Thus, p^*_{ijk} is 0 if the three loci co-localize merely because of nonzero pairwise contact probabilities of each pair in the triplet, whereas it is larger than 0 if there is a specific cooperative mechanism which involves the three loci. We required that i, j and k have a mutual genomic distance of at least 60 kb (i.e., 3 beads in the model). As a negative control, we also calculated the same quantity for each TAD in 40 simulations derived by a randomly reshuffled interaction matrix B_{ij}. For each TAD, we then recorded the triplets of loci whose excess probability p^*_{ijk} was found to be larger than in 95% of the randomized controls and filtered the data by requiring $p^*_{ijk} > 0.02$.

We observed that TADs display a variable number of co-localized triplets, from 0 to ~4000 per TAD, to be compared with the number of possible triplets $\sim N^3/3!\sim 10^5$. Interestingly, the number of co-localized multiplets does not correlate with the activity within single TADs (cf. Fig. S4 in the Supplementary Materials). We then investigated whether co-localized sites contain preferentially active promoters, enhancers or CTCF bound sites. CTCF bound sites were detected using ChIP-seq from ENCODE project (Cheng et al. 2014). In particular, macs2 (Zhang et al. 2008) was applied to ChIP-seq datasets to identify significant CTCF-bound sites (peaks). Active promoters were defined as of those genes where $log_2\ RPKM > 1$. To define a region to be enhancers we used as landmark region enriched in H3K27ac (peaks from macs). Then we analysed the ratio between H3K4me1/H3K4me3 (ChIP-seq normalised signal) and labelled a region as an enhancer if H3K4me1/H3K4me3 is enriched. In Fig. 4b we display the probability P defined as the fraction of co-localized triplets of loci containing a promoter each (marked as "ppp"), a promoter in two of them and an enhancer in the other ("ppe"), a promoter, an enhancer and a CTCF site ("pec"), etc. These probabilities are compared with a negative control obtained reshuffling the colocalized triplets in each TAD and with the theoretical probability obtained from the densities of promoters, enhancers and CTCF in the genome (see Supplementary Materials).

Enrichments displayed in the figure are statistically significant, displaying a p-value <0.05 with respect to the bootstrap. A comparison of the significance of the different combinations was carried out on the basis of their Z-score.

Triplets of CTCF are remarkably significant (Z-score=493), suggesting that the molecular mechanism employed by this protein is not just a dimerization (Pant et al. 2004), but can involve in some way multiple copies of CTCF. A similar significance is displayed by triplets made of an enhance, a promoter and CTCF (Z-score=491), suggesting that CTCF could play a direct role in the approaching of enhancers to promoters. The significance of triplets of promoters (Z-score=451) and of two promoters and one enhancer (Z-score=410) could emphasise the fact that multiple promoters are controlled by few factors. Triplets involving more than one enhancer are less enriched (*eee* has Z-score=262 and *pee* has Z-score=294).

The list of triplets containing promoters and enhancers is reported in the Supplemental Table S2. Among them are, as an example, the triplet *Tsix/Chic1/Linx*, whose co-localisation was already predicted and verified in FISH experiments (Giorgetti et al. 2014).

Conclusions

Hi-C datasets give information about interaction probabilities of pairs of chromosomal loci. Using a polymeric model, these data can be de-convolved into equilibrium ensembles of conformations that allow one to predict the shape and size of single TADs, their structural fluctuations and the correlations in the formation of sub-TAD contacts loci.

Applying this strategy to all TADs genome-wide in mouse embryonic stem cells allowed us to extract structural properties, and to correlate them with genomic datasets. Model-driven structural reconstruction predicts that the overall structure of TADs displays some features that are distinctive of ideal polymeric chains, as expected by general physical arguments applying to a small polymer chain surrounded by other polymers. However TAD structural properties are strongly affected by the heterogeneity of interactions between loci within the TAD, irrespective of the molecular mechanisms that generate them. The global level of gene activity within single TADs correlates positively with their degree of structural anisotropy; this suggests that high local gene activity is accompanied by a ‘linearization’ of the three-dimensional conformations of the chromatin fibre and a suppression of structural fluctuations, as a consequence of looping interactions between CTCF sites and active regulatory regions and compartmentalization in sub-TAD domains. These two effects require a complex set of interaction to stabilize the conformation of the chromatin fibre within TAD, and could not be possible, for instance, in a simple homopolymeric model (i.e. a genome without locus-specific interactions). Indeed, the number of contacts between pairs of loci at small genomic distance is a good proxy for global transcriptional activity, whereas the number of contacts between genes at large genomic distance (although within the same TAD) does not seem to correlate with activity.

Importantly, our simulations also provide information on the tendency to form simultaneous contacts within single TADs. In agreement with the experimental findings of (Olivares-Chauvet et al. 2016), the amount of co-localising triplets amounts to a small fraction of all possible ones. However, they are enriched in promoters and enhancers, which suggests a potential role of their co-localization in single TADs in setting up their functional interactions.

Acknowledgements

Research in L.G.’s lab is supported by Novartis Research Foundation. We would like to thank NIBR Computing Resources for technical support.

References

Annunziatella C, Chiariello AM, Bianco S, Nicodemi M (2016) Polymer models of the hierarchical folding of the Hox-B chromosomal locus. Phys Rev E 94:042402.

doi: 10.1103/PhysRevE.94.042402

Barbieri M, Chotalia M, Fraser J, et al (2012) Complexity of chromatin folding is captured by the strings and binders switch model. Proc Natl Acad Sci USA 109:16173–16178. doi: 10.1073/pnas.1204799109

Baù D, Sanyal A, Lajoie BR, et al (2011) The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. Nat Struct Mol Biol 118:107-114. doi: 10.1038/nsmb.1936

Benedetti F, Dorier J, Burnier Y, Stasiak A (2013) Models that include supercoiling of topological domains reproduce several known features of interphase chromosomes. Nucl Acid Res 42:2848-2855. doi: 10.1093/nar/gkt1353

Boettiger AN, Bintu B, Moffitt JR, et al (2016) Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. Nature 529:1–15. doi: 10.1038/nature16496

Brackley CA, Taylor S, Papantonis A, et al (2013) Nonspecific bridging-induced attraction drives clustering of DNA-binding proteins and genome organization. Proc Natl Acad Sci USA 110:E3605–11. doi: 10.1073/pnas.1302950110

Cheng Y, Ma Z, Kim B-H, et al (2014) Principles of regulatory information conservation between mouse and human. Nature 515:371–375. doi: 10.1038/nature13985

Chiariello AM, Annunziatella C, Bianco S, et al (2016) Polymer physics of chromosome large-scale 3D organisation. Sci Rep 6:29775. doi: 10.1038/srep29775

Dekker J (2016) Mapping the 3D genome: Aiming for consilience. Nat Rev Mol Cell Biol 17:741–742. doi: 10.1038/nrm.2016.151

Denker A, de Laat W (2016) The second decade of 3C technologies: detailed insights into nuclear organization. Genes Dev 30:1357–1382. doi: 10.1101/gad.281964.116

Dixon JR, Selvaraj S, Yue F, et al (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485:376–380. doi: 10.1038/nature11082

Doi M, Edwards SF (1986) The Theory of Polymer Dynamics. Oxford University Press

Fabre PJ, Benke A, Joye E, et al (2015) Nanoscale spatial organization of the HoxD gene cluster in distinct transcriptional states. Proc Natl Acad Sci USA 112:13964–13969. doi: 10.1073/pnas.1517972112

Fudenberg G, Imakaev M (2016) FISH-ing for captured contacts: towards reconciling FISH and 3C. bioRxiv 081448. doi: 10.1101/081448

Fudenberg G, Mirny LA (2012) Higher-order chromatin structure: bridging physics and biology. Curr Opin Genet Dev 22:115–124. doi: 10.1016/j.gde.2012.01.006

Gaidatzis D, Lerch A, Hahne F, Stadler MB (2015) QuasR: quantification and annotation of short reads in R. Bioinformatics 31:1130–1132. doi: 10.1093/bioinformatics/btu781

Gendrel A-V, Attia M, Chen CJ, et al (2014) Developmental dynamics and disease potential of random monoallelic gene expression. Dev Cell 28:366–380. doi: 10.1016/j.devcel.2014.01.016

Giorgetti L, Galupa R, Nora EP, et al (2014) Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. Cell 157:950–963. doi: 10.1016/j.cell.2014.03.025

Giorgetti L, Heard E (2016) Closing the loop: 3C versus DNA FISH. Genome Biol 17:215. doi: 10.1186/s13059-016-1081-2

Giorgetti L, Lajoie BR, Carter AC, et al (2016) Structural organization of the inactive X chromosome in the mouse. Nature 535:575–579. doi: 10.1038/nature18589

Goloborodko A, Marko JF, Mirny LA (2016) Chromosome Compaction by Active Loop Extrusion. Biophysj 110:2162–2168. doi: 10.1016/j.bpj.2016.02.041

Grosberg AY, Khokhlov A (1994) Statistical Mechanics of Macromolecules. AIP Press

Hübner MR, Eckersley-Maslin MA, Spector DL (2013) Chromatin organization and transcriptional regulation. Curr Opin Genet Dev 23:89–95. doi: 10.1016/j.gde.2012.11.006

Imakaev M, Fudenberg G, McCord RP, et al (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods 9:999–1003. doi: 10.1038/nmeth.2148

Johnson J, Brackley CA, Cook PR, Marenduzzo D (2015) A simple model for DNA bridging proteins and bacterial or human genomes: bridging-induced attraction and genome compaction. J Phys Cond Mat 27:064119. doi: 10.1088/0953-8984/27/6/064119

le Dily F, Baù D, Pohl A, et al (2014) Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. Genes Dev 28:2151–2162. doi: 10.1101/gad.241422.114

Lifshitz I, Grosberg A, Khokhlov A (1978) Some problems of the statistical physics of polymer chains with volume interaction. Rev Mod Phys 50:683-713. doi: 10.1103/RevModPhys.50.683

Merkenschlager M, Nora EP (2016) CTCF and Cohesin in Genome Folding and Transcriptional Gene Regulation. Annu Rev Genomics Hum Genet 17:17–43. doi: 10.1146/annurev-genom-083115-022339

Mirny LA (2011) The fractal globule as a model of chromatin architecture in the cell.

Chromosome Research 19:37–51. doi: 10.1007/s10577-010-9177-0

Nora EP, Lajoie BR, Schulz EG, et al (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature 485:381–385. doi: 10.1038/nature11049

Olivares-Chauvet P, Mukamel Z, Lifshitz A, et al (2016) Capturing pairwise and multi-way chromosomal conformations using chromosomal walks. Nature 540:296-300. doi: 10.1038/nature20158

Pant V, Kurukuti S, Pugacheva E, et al (2004) Mutation of a single CTCF target site within the H19 imprinting control region leads to loss of Igf2 imprinting and complex patterns of de novo methylation upon maternal inheritance. Mol Cell Biol 24:3497–3504. doi: 10.1128/MCB.24.8.3497-3504.2004

Rao SSP, Huntley MH, Durand NC, et al (2014) A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell 159:1–29. doi: 10.1016/j.cell.2014.11.021

Rosa A, Zimmer C (2014) Computational models of large-scale genome architecture. Int Rev Cell Mol Biol 307:275–349. doi: 10.1016/B978-0-12-800046-5.00009-6

Tiana G, Amitai A, Pollex T, et al (2016) Structural Fluctuations of the Chromatin Fiber within Topologically Associating Domains. Biophys J 110:1234–1245. doi: 10.1016/j.bpj.2016.02.003

Zhan Y, Giorgetti L, Tiana G (2016) Looping probability of random heteropolymers helps to understand the scaling properties of biopolymers. Phys Rev E 94:032402. doi: 10.1103/PhysRevE.94.032402

Zhang Y, Liu T, Meyer CA, et al (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol 9:R137. doi: 10.1186/gb-2008-9-9-r137

Zhang Y, Mariani L, Barozzi I, et al Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes (submitted).

Figures

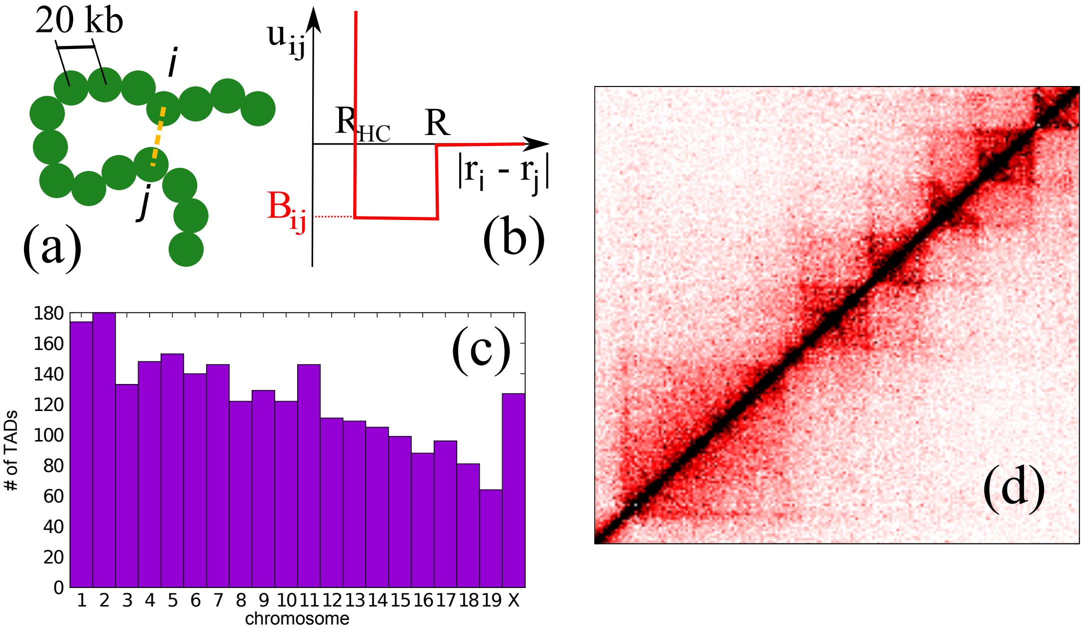

Figure 1: (A) Chromosomes are represented as chains of beads, each representing a unit of 20 kb. Pairs of beads interact with a potential u_{ij} (B) shaped as a spherical well. (C) The number of TADs identified in each chromosome. (D) A zoom into the Hi-C map showing five TADs.

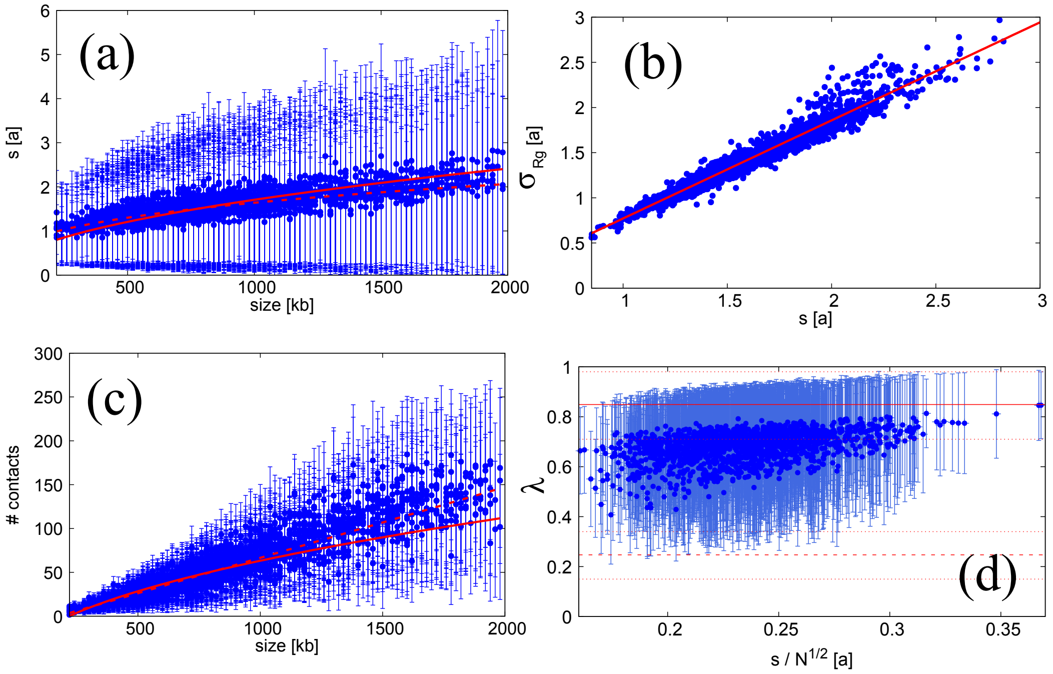

Figure 2: (a) The root square mean gyration radius *s* of each TAD as a function of its size; the error bars report the associated fluctuations, that is the standard deviation σ_{Rg} of the gyration radius. Lengths are in units of the inter-bead distance *a*. The red curves indicate the fits with the curves expected for an ideal-chain behaviour (continuous curve) and for a random globule (dashed curve). (b) The standard deviation plotted as a function of the mean gyration radius; the red line displays the linear behaviour expected for an ideal chain. (c) The average number of contacts associated with each TAD as a function of the linear size of the TAD; in red, the fits (cf. (a)). (d) The average linearity parameter λ of each TAD as a function of the normalized mean gyration radius; the continuous and the dashed line indicate the average value calculated, respectively, for an ideal chain and a random coil; the dotted lines indicate their standard deviations.

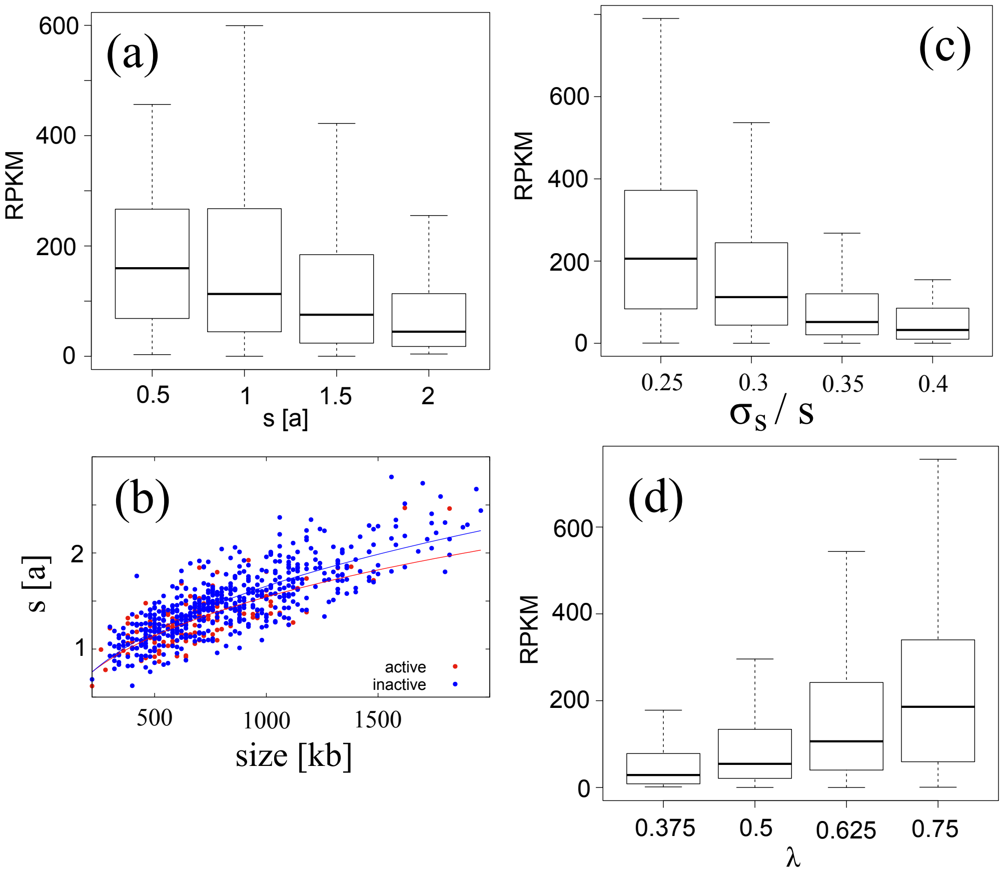

Figure 3: (a) Box plot of the sum of the activity of the genes in each TAD (RPKM) as a function of the mean gyration radius s. (b) The mead gyration radius s of a subset of active TADs, with RPKM larger than 500 (in blue) and of inactive TADs, with RPKM lower than 50 (in red); the lines indicate the power-law fit, which give exponents 0.31 and 0.35, respectively. (c) The activity as a function of the relative fluctuations of the gyration radius. (d). The activity of genes as a function of the linearity coefficient λ.

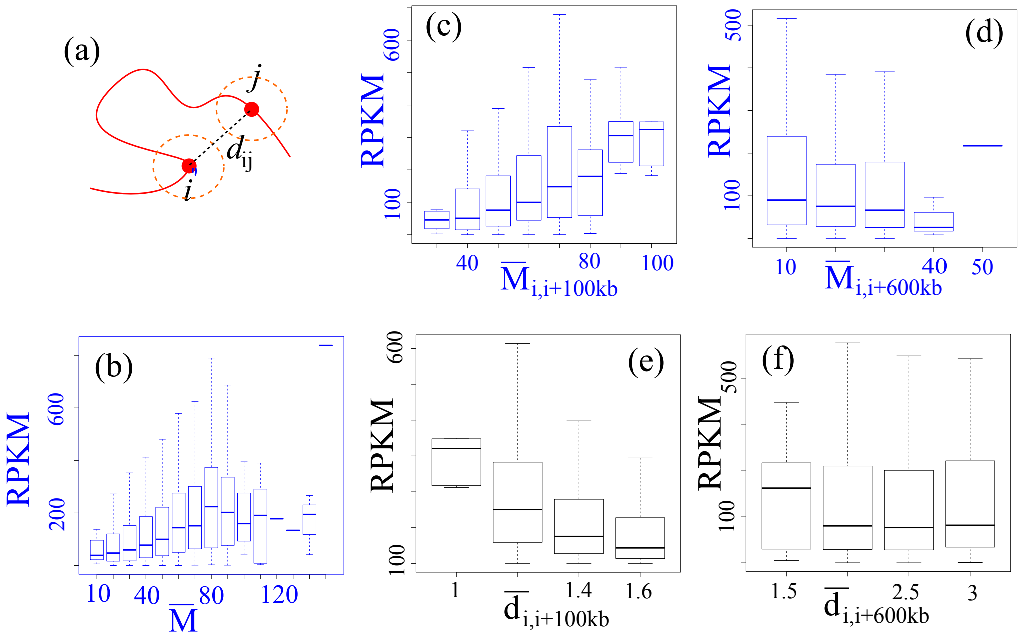

Figure 4: For each pair of loci *i* and *j* the Hi-C map gives the probability M_{ij} that the two are within their interaction distance (orange circles in (a)), and the model permits the calculation of their mutual mean distance $\overline{d_{ij}}$. The activity RPKM of each TAD is plotted versus the mean counts (b), the mean counts restricted to pairs of the TAD at small genomic distances (100kb) (c), that between pairs at large genomic distances (600kb) (d), the average distances between pairs of the tad at small genomic distances (e), that between pairs at large genomic distances (f).

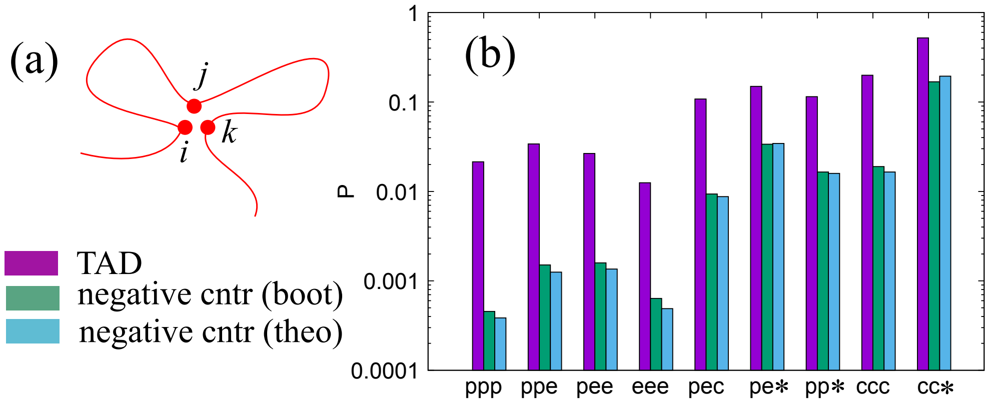

Figure 5: (a) Co-localisation occurs in TADs when three 20-kb loci interact with each other with a probability p^*_{ijk} which is statistically significant (p<0.05) with respect to bootstrapped, random-interacting polymers. (b) The probability P that triplets of co-localized sites contain promoters ("p"), enhancers ("e") and CTCF ("c"). For example, "ppe" means that two of them contain a promoter and the other an enhancer, "pe*" means that a site contains a promoter, another an enhancer and the third can contain anything.