The scaling features of the 3D organization of chromosomes are highlighted by a transformation á la Kadanoff of Hi-C data

Chiariello A.M., Bianco S., Annunziatella C., Esposito A. and Nicodemi M.

EPL (Europhysics Letters)
2018 FEB 08 ; 120(4): 40004
2018 FEB 08 (first published online)
doi: 10.1209/0295-5075/120/40004

Publisher: Institute of Physics Publishing (IOP)

The scaling features of the 3D organization of chromosomes are highlighted by a transformation *à la* Kadanoff of Hi-C data

Andrea M. Chiariello[1,⋆] , Simona Bianco[1,⋆] , Carlo Annunziatella[1,⋆] , Andrea Esposito[1,2,⋆] and Mario Nicodemi[1,3,†]

[1] *Dipartimento di Fisica, Universitá degli Studi di Napoli Federico II, and INFN Napoli, Complesso Universitario di Monte Sant'Angelo, 80126 Naples, Italy.*

[2] *Berlin Institute for Medical Systems Biology at the Max Delbruck Center for Molecular Medicine in the Helmholtz Association, Berlin, Germany.*

[3] *Berlin Institute of Health (BIH), MDC-Berlin, 13125 Berlin, Germany.*

[⋆] *Equal contribution authors.*

[†] *Corresponding author: mario.nicodemi@na.infn.it*

PACS 05.20.-y – Statistical Mechanics
PACS 87.16.Sr – Chromosomes
PACS 05.10.Cc – Renormalization group methods

Abstract –Technologies such as Hi-C and GAM have revealed that chromosomes are not randomly folded into the nucleus of cells, but are composed by a sequence of contact domains (TADs), each typically 0.5Mb long. However, the larger scale organization of the genome remains still not well understood. To investigate the scaling behaviour of chromosome folding, here we apply an approach *à la* Kadanoff, inspired by the Renormalization Group theory, to Hi-C interaction data, across different cell types and chromosomes. We find that the genome is characterized by complex scaling features, where the average size of contact domains exhibits a power-law behaviour with the rescaling level. That is compatible with the existence of contact domains extending across length scales up to chromosomal sizes. The scaling exponent is statistically indistinguishable among the different murine cell types analysed. These results point toward a scenario of a universal higher-order spatial architecture of the genome, which could reflect fundamental, organizational principles.

Introduction. – In the last decade, new and powerful technologies, such as the Hi-C [1] and the GAM [2] methods, have been developed to quantitatively explore the three dimensional organization of chromosomes in the cell nucleus. They provide information about the frequency of contacts in space between pairs of DNA segments (loci) genome-wide. From these data, it is emerging that the genome has a complex spatial organization in higher organisms [3–5]. Contact data provide vital information because gene activity can be regulated through the interaction between elements, such as promoters and enhancers, remote along DNA sequence. From the analysis of Hi-C data, it is found that chromosomes are folded into a sequence of so-called 'contact domains' (or TADs, Topologically Associating Domains), typically 0.5Mb long, which have strong local interactions [9, 10] and comparatively well conserved boundary locations across different tissues and species [6, 9]. While

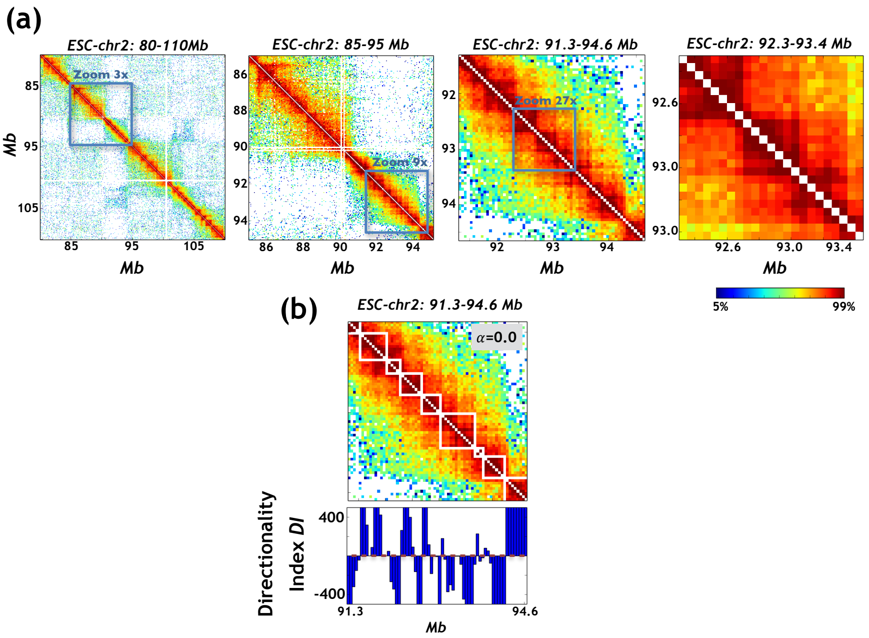

Fig. 1: Example of Hi-C matrices for ESC cell line, from [6]. **(a)** Four nested regions on chromosome 2, ranging from $30Mb$ to $\sim 1Mb$. Each matrix contain a complex pattern of interaction: the TAD structure is visible in the smallest matrix (strong red squares along the diagonal), but bigger interaction domains exist as the genomic window is increased. To highlight long-range contact, data are shown in logarithmic scale. **(b)**Contact domains are identified using the Directionality Index DI ([6,9]): as the DI signal becomes positive, a boundary is annotated. The domains identified in the reported region are highlighted by white squares.

TADs are currently the focus of intense investigations, it has been observed that chromosomal interactions exist also at different scales, within TADs [6–8] and at larger scales. It has been discussed, for instance, that TADs interact in 10Mb wide 'A/B compartments' [1] and form higher order structures, named metaTADs [6]. Yet, the global organization of chromosomes remains not fully understood. Here, we investigate the scaling features of spatial organization of chromosomes by implementing a computational procedure inspired by the Renormalization Group methods of Statistical Mechanics, and in particular Kadanoff transformations [11, 12] applied directly to published murine genome-wide Hi-C data [6]. By rescaling iteratively the interactions between different regions of the genome, we find that the average size of the contact domains identified at each rescaling level exhibits an approximate power-law behaviour over two decades in genomic scales, across different cell types along murine neuronal differentiation [6]. That complex scaling behaviour points to-ward a scenario where chromosome folding is marked by structures across different scales. Interestingly, the scaling exponents in the studied cell types are very similar, suggesting a universal global organization of the genome, which could reflect fundamental organizational principles [13].

Dataset analyzed. – To investigate the chromatin folding in different cell types, we study recently published Hi-C data [6] in a murine neuronal differentiation cell line, from mouse embryonic stem cells (ESC), intermediate neuronal precursor cells (NPC) and post-mitotic neurons (Neurons). We consider published intra-chromosomal Hi-C data at a resolution of 50kb; they can be visualized as symmetric square matrices (see Fig. 1 or [6],

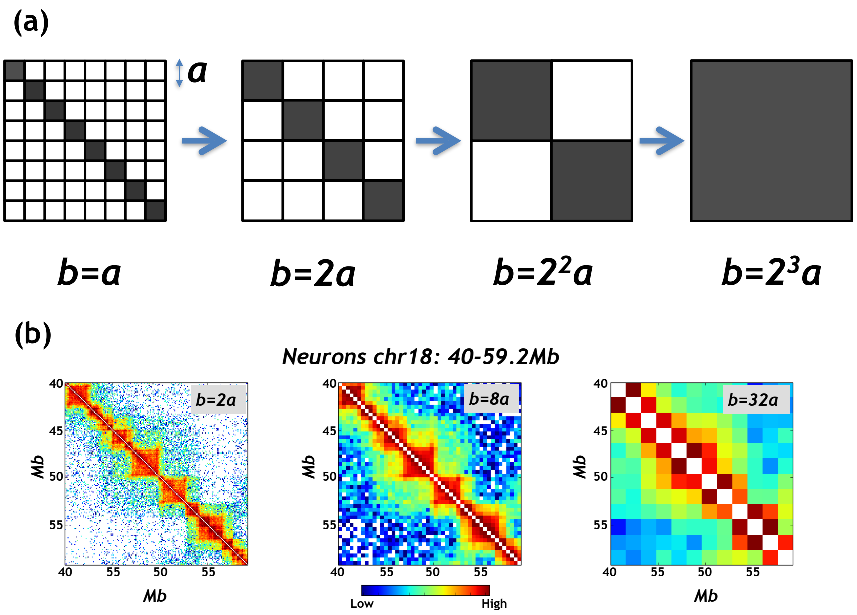

Fig. 2: Coarse-graining approach used to investigate the scaling features of genome architecture.**(a)** Schematic representation of the scaling transformation iteratively applied to each Hi-C matrix. In our notation, a is the length unit of the bins, that is the resolution of the original Hi-C matrices $(50Kb)$**(b)** The coarse-graining procedure is applied on a $\sim 20Mb$ region of chromosome 18, Neurons cell line. Here, we show the results for the rescaling factors $b = 100Kb$ (left matrix), $b = 400Kb$ (central matrix) and $b = 1.6Mb$ (right matrix).

logarithmic scale). Each pixel of the matrix $x_{i,j}$ contains the interaction frequency between the DNA regions (loci) i and j. In Fig. 1a, Hi-C data are shown for four nested regions on chromosome 2, in ESC, spanning from 30Mb to 1Mb in size. A complex pattern of interaction is visible for each genomic window represented. In particular, the typical TAD structure, with strong squares along the diagonal, is seen in the higher-resolution matrices (Fig. 1a right panels), but as the size of the considered genomic window is increased it is also clear that much bigger interaction domains exist, encompassing multiple TADs. These considerations prompted us to investigate the scaling features of genomic interactions.

Identification of contact domains. – To identify the basic contact domains, or TADs, in the system, i.e., the regions with enriched intra-domain contacts along the diagonal of Hi-C matrices, many algorithms have been proposed [9, 14]. Here, we use the pipeline described in [6]. Briefly, we calculated, for each 50kb window, the Directionality Index (DI, [9]), which measures the difference of interaction that a locus has with its neighbouring upstream or downstream loci. An example of DI signal is reported in Fig. 1b. Briefly, the boundaries of the domains (superimposed on the matrix as white squares) are identified when the signal changes sign from negative to positive, i.e., the tendency to interact shifts from left to right, by crossing a given amplitude threshold (all details in [6]). An example of TADs is given in Fig. 1b.

Coarse-graining of Hi-C matrices. – To investigate the scaling features of genomic interactions, we considered an iterative computational procedure inspired by the Renormalization Group. We applied a scale transformation to the chromosomes where pairs of consecutive $50Kb$ bins are fused in a new, twice as large bin; that corresponds to a coarse-graining of the original Hi-C matrix in 2x2 blocks (Fig. 2a). The overall interaction between the new bins is then the average of the values contained in the corresponding block of the original matrix. Specifically, the renormalized interaction $x'_{i',j'}$ between the rescaled bin i' and j' is:

$$x'_{i',j'} = 1/4 \sum_{i,j} x_{i,j} \tag{1}$$

where the sum runs over the original bins i and j included respectively in the new blocks i' and j', and $x_{i,j}$ is their Hi-C interaction. The result is a coarse-grained Hi-C matrix, $x'_{i',j'}$, having a linear size scaled by a factor 2 relative to $x_{i,j}$, as schematically represented in Fig. 2a. The described procedure is next applied iteratively to the coarse-grained Hi-C matrices. For instance, in Fig. 2a, this is repeated 4 times, and the initial 16x16 matrix eventually becomes just one single bin. In our notation, we name a the length unit of the bins of the original matrix ($50Kb$ in our case), and b is the rescaling factor. So, if we apply the transformation 4 times, the final rescaling factor b is $2^4a = 16a = 800Kb$. In other words, b is the genomic length of the single bin at the considered coarse-graining level. Fig. 2b shows the effect of the transformation applied to real Hi-C data of a region on chromosome 18 ($\sim 20Mb$ long), in Neurons. The results are plotted for the transformation at three levels of the rescaling procedure ($b = 2a = 100Kb$, $b = 8a = 400Kb$ and $b = 32a = 1.6Mb$). For each cell type, we applied the coarse-graining protocol to the 20 chromosomes of the mouse genome, starting from the original 50Kb resolution of Hi-C data. The sizes in bins of the original $50Kb$ resolution matrices range from 3944x3944 (chromosome 1, the largest one) to 1227x1227 (chromosome 19, the smallest one). They set the maximal number of iterations of the procedure for each chromosome.

Scale invariance and power law behaviour. – To explore the features of genomic interactions at different length scales, we employed our coarse-graining scheme. In particular, at each coarse-graining level, we calculated the 'contact domains' in the correspondingly rescaled Hi-C matrix, using the procedure described above. In this way we access the emerging 'contact domains' at different scales. The domain size, d, distribution is reported in Fig. 3a, in ESC cells, for the first three levels of coarse-graining. The average size $< d >$, highlighted as vertical dashed line, increases as the transformation is repeated. Fig. 3b, shows the dimensionless variable $< d > /b$, i.e., the rescaled average domain size in units of a, as a function of the scaling factor b (in dimensionless units). The values are averaged over the 20 chromosomes. The quantity $< d > /b$ represents the average number of bins that form a 'contact domain' at the given coarse-graining level. In case the contact domains of the system are characterized by only one length scale, b_0, e.g., the $0.5Mb$ average size of TADs, the expectation is that, as b grows larger than b_0, $< d > /b$ flattens out to an asymptotic constant value d_∞, $< d > /b \sim d_\infty$. While such a behaviour is indeed observed in a control case (see below), interestingly, we found that the rescaled domain size decays as a power-law, very well described by the function:

$$< d > /b = \bar{d}/b^\gamma + d_\infty \tag{2}$$

where γ is the scaling exponent, $\bar{d}$ a constant. In Fig. 3b, the fits for the three cell lines analysed are shown as dashed lines. Interestingly, the asymptotic value d_∞ is not approached until chromosomal scales. The fits are robust (a chi-squared test has a p-val=1), and return a value for the exponent $\gamma = 0.52 \pm 0.08$ (averaged over the three cell types), which is not compatible with the control model having 'contact domains' of a single characteristic scale (see next paragraph). The scaling behaviour of interaction data is visible in the example of Fig. 3c, where four Hi-C matrices are shown at increasing rescaling levels, from $b = 2a$ to $b = 16a$. Accordingly, the plotted genomic region is increased by a factor 2 each time,

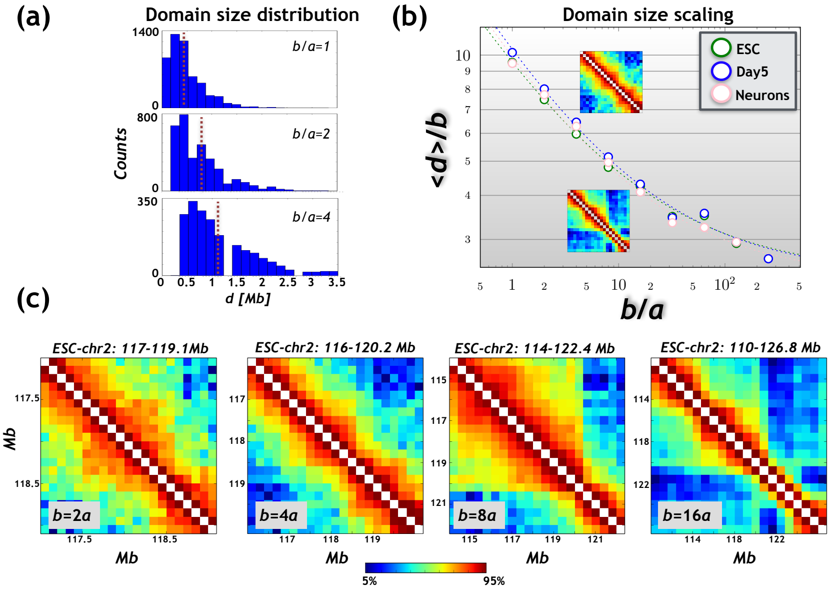

Fig. 3: Domain size exhibits a power law behaviour with the rescaling factor. **(a)** Domain size distribution for the first three level of coarse-graining of the ESC cell line. The NPC and Neurons cell lines have similar distributions. The vertical dashed line represents the average value. **(b)** Scaling of the average domain size $< d > /b$ with the dimensionless rescaling factor b/a. The dashed lines represent the best fit curve described by eq. 2. All the analyzed lines have very similar scaling behaviour. The inset matrices refer to different coarse-graining levels. The plot is in log-log scale. **(c)** Four Hi-C matrices coarse-grained by an increasing rescaling factor, reported in the grey box. The genomic window is doubled each time so to have a constant matrix size. The block structure is visible in each matrix, and no privileged length scale is observed.

in order to keep constant the size of the matrices. From a visual inspection, they exhibit an overall similar structure with blocks of interactions. This suggests that it is not possible identify a privileged length scale, rather the structure has a complex scaling behaviour. The power law behaviour we found points toward a scenario where the organization of genome interactions is characterized by different, increasing length scales up to the sizes of entire chromosomes. Our results also suggest that organization into TADs discovered at small length scales [10, 14] is replicated at higher length scales, with complex, different sizes of domains. This is consistent with the existence of metaTADs [6].

Control model. – As a test, we compared our results with a control case model where interactions are confined within a particular scale. Precisely, the control case is made as follows: we consider the coordinates of the TADs identified in the original data ($50Kb$ resolution), and then produce an artificial matrix where an entry is 1 if the bins are within the same TAD, and zero otherwise. In Fig. 4a, an example is shown of the experimental data (and the corresponding domains, in white) and the resulting control matrix model, with red squares. Since we use the actual domain coordinates in each chromosome, we have a set of control matrices that are equal in number and size to the original experimental matrices, but are marked only by the scale of the fundamental TADs in the system. TAD sizes have an exponential distribution with an average of 0.5Mb [6]. Hence, by construction our control matrices do not contain interactions at larger scales. We applied our coarse-graining procedure to the control matrices, derived their corresponding coarse-grained versions and

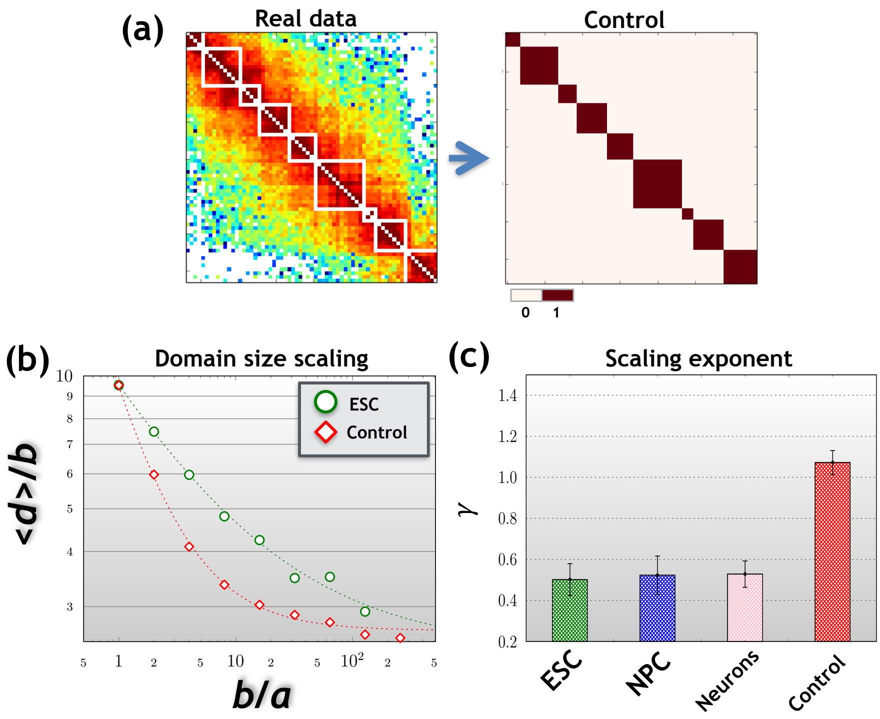

Fig. 4: Control model exhibits a scaling behaviour not compatible with the experimental data.**(a)** The control matrices are made by blocks with 1 if the bin is in a TAD and 0 otherwise. **(b)** Comparison between the scaling of the normalized average size for the ESC cell line (green curve) and the control model (red curve). **(c)** The values of the scaling exponent γ obtained from the best fit. The control model is not compatible with the real data.

identified with the same pipeline above the emerging 'contact domains' at each iteration. In Fig. 4b, it is reported the $< d > /b$ curve of the control case as a function of the rescaling factor b, and the same curve for the real data (ESC cells) as a comparison (already shown in Fig. 3b). We find that the control case scales with an exponent $\gamma = 1.07 \pm 0.06$, and $< d > /b$ rapidly approaches the asymptotic value. As before, the fit is very robust (chi-squared test p-val=1). The exponent, consistent with 1, is expected because the domains detected at each coarse-graining level are always the same and the number of bins is halved each time the matrix is rescaled. The values of the exponents obtained from the fit are shown in Fig. 4c. The error bars are extracted from the covariance matrix given by the fitting algorithm (Python routine *curve fit* from the *scipy* package). Importantly, the exponents obtained from the real data, in all the analysed cell types, are statistically equal to each other and different from the control case ($\gamma = 0.5 \pm 0.08$,0.52 ± 0.09 and 0.53 ± 0.06 for ESC, NPC and Neurons respectively). In brief, we conclude that Hi-C contact data return a scenario of genomic interactions extending across scales, well beyond the size of fundamental TADs. This appears to be a general feature of genome organization as it is observed across all the investigated cell types.

Robustness of the procedure. – Finally, we tested the robustness of our approach and results against using different definitions of TADs. So, we repeated the above analysis calculating the contact domains, at each coarse-graining level, by using a different threshold values for the DI index; in particular, rather than considering $\alpha = 0.0$ as above, we employed the value of α that returns TADs having average size similar to the original TADs defined in [9] ($\alpha = 0.1$, see [6] for details) that is roughly twice as large as in the $\alpha = 0.0$ case [6].

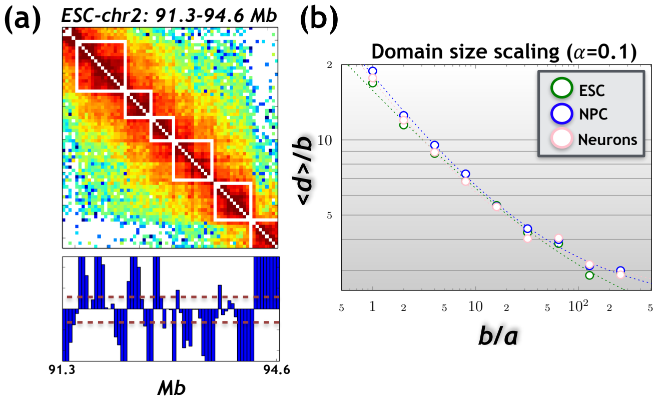

Fig. 5: The scaling behaviour is robust to parameters changes. **(a)** Contact domains are detected with a different threshold, tuned by the α parameter. In this case, we set $\alpha = 0.1$. **(b)** Scaling behaviour is not affected by changing the contact domains definition. Dashed lines are the best fit curves of eq. 2.

Nevertheless, the $< d > /b$ curve is again very well described by the power-law in equation 2. Fig. 4d, shows the curve for the corresponding domain size. The scaling exponent is $\gamma = 0.54 \pm 0.06$ (averaged over the three cell lines), which is completely consistent with our previous result and confirm once again the complex scaling behaviour.

Conclusions. – Overall, this work represents a novel application of classical concepts of Statistical Physics (Kadanoff transformation) to gain a deeper insight into the 3D struc-ture of chromosomes in the nucleus of cells of higher mammals. New technologies such as Hi-C [1] and GAM [2] are revealing novel important details on genome folding in the cell nucleus. It has emerged, in particular, that chromosomes are formed by a sequence of do-mains marked by high frequencies contacts, named TADs, having a typical size of roughly $0.5Mb$. Yet, more recent studies are highlighting the existence of higher order interactions between TADs and chromosomes [6]. To shed light on the matter, in the present work we investigated the scaling behaviour of real Hi-C data, in a murine neuronal differentiation, by applying to experimental Hi-C data [6] a computational procedure inspired by the methods of the Renormalization Group of Statistical Physics. Our results return a view of chromo-somes architecture characterized by complex scaling features, extending from short ($50Kb$) up to chromosomal scales, not compatible with the existence of interactions at only one length scale, say the one of TADs or A/B compartments [1]. In particular, upon rescaling of the interactions, we find that the average contact domain size exhibits a power-law behaviour with a non trivial exponent pointing towards a self similar organization across scales. These results are in agreement with recent findings [2, 6], where the importance of long-range chromosome interactions has been stressed. The discovered complex scaling fea-tures of chromosomal structures could be biologically convenient as they would permit high level of compaction and, at the same time, selective contacts between specific regions [6,15]. One limitation of the present study is that it is entirely based on the empirical definition of TADs or contact domains, in turn based on heuristic analyses of experimental Hi-C data. To understand more deeply the relationship between data and spatial structure, principled theories are needed. To this aim polymer physics models have been developed. Importantly, polymer physics confirms that scaling concepts are very important to understand the genome organization (see, e.g., [16–18] and reviews in [26–29]). Such models represent also a pow-erful tool to investigate the architecture of real loci [19–21, 33–35] and whole chromosomes [30, 31], and to investigate the specific mechanism driving folding [22–25, 32].

* * *

We acknowledge grants from the NIH ID 1U54DK107977-01, CINECA ISCRA ID HP10CRTY8P, the Einstein BIH Fellowship Award, and computer resources from INFN, CINECA, and Scope at the University of Naples.

REFERENCES

[1] Lieberman-Aiden E. *et al.*, *Science*, **326** (2009) 289-293.
[2] Beagrie R. A., *et al.*, *Nature*, **543** (2017) 519?524.
[3] Bickmore W. A. and van Steensel B., *Cell*, **152** (2013) 1270-1284.
[4] Tanay A. and Cavalli G., *Current Opinion in Genetics & Development*, **23** (2013) 197-203.
[5] Dekker J. and Mirny L., *Cell*, **164** (2016) 1110-21.
[6] Fraser J. *et al.*, *Mol. Syst. Biol.*, **11** (2015) 852.
[7] Sexton T. *et al.*, *Cell*, **148** (2012) 458-72.
[8] Phillips-Cremins J. E. *et al.*, *Cell*, **153** (2013) 1281-1295.
[9] Dixon J. R. *et al.*, *Nature*, **485** (2012) 376-380.
[10] Nora E. P. *et al.*, *Nature*, **485** (2012) 381-385.
[11] Kadanoff L. P., *Physics*, **2** (2966) 263.
[12] Yeomans J. M., *Statistical Mechanics of phase transitions* (Oxford University Press Inc., New York) 1992.
[13] Gilaranz L. J. *et al.*, *Science*, **357** (2017) 199-201.
[14] Rao S. S. P. *et al.*, *Cell*, **159** (2014) 1665-1680.
[15] Sarnataro S. *et al.*, *PLoS ONE*, **12** (2017) e0188201.
[16] Barbieri M. *et al.*, *Proc. Natl. Acad. U.S.A.*, **109** (2012) 16173-16178.
[17] Chiariello A. M. *et al.*, *Scientific Reports*, **6** (2016) 29775.
[18] Nicodemi M. and Prisco A., *Biophys. J.*, **96** (2009) 2168-2177.
[19] Annunziatella C. *et al.*, *Phys. Rev. E*, **94** (2016) 042402.
[20] Barbieri M. *et al.*, *Nat. Struct. Mol. Biol.*, **24** (2017) 515-524.
[21] Giorgetti L. *et al.*, *Cell*, **157** (2014) 950-963.
[22] Brackley C. A. *et al.*, *Proc. Natl. Acad. Sci. U.S.A.*, **110** (2013) E3605-11.
[23] Jost D. *et al.*, *Nucleic Acids Res.* **42**, (2014) 9553-61.
[24] Sanborn A. L. *et al.*, *Proc. Natl. Acad. Sci. U.S.A.*, **112** (2015) E6456-65.
[25] Fudenberg G. *et al.*, *Cell Reports*, **15** (2016) 1-12.
[26] Nicodemi M. and Pombo A., *Curr. Opin. Cell. Biol.*, **28C** (2014) 90-95.
[27] Bianco S. *et al.*, *Chromosome Res.*, **25** (2017) 25-34.
[28] Barbieri M. *et al.*, *Front. Genet.*, (2013) doi: 10.3389/fgene.2013.00113.
[29] Chiariello A. M. *et al.*, *Mod. Phys. Lett. B*, **29** (2015) 1530003.
[30] Rosa A. and Everaers R., *PLoS Comput. Biol.*, **4** (2008) e1000153.
[31] Di Stefano M. *et al.*, *Scientific Reports*, **6** (2016) 35985.
[32] Bohn M. and Heerman D.W., *PLoS ONE*, **5** (2010) e12218.
[33] Scialdone A. *et al.*, *PLoS Comput. Biol.*, **7** (2011) e1002229.
[34] Nicodemi M. and Prisco A., *Phys. Rev. Lett.*, **98** (2007) 108104.
[35] Chiariello A. M. *et al.*, *Front. Neurosci.*, **11** (2017) 559.