Merging 1D and 3D genomic information: Challenges in modelling and validation

Review

# Merging 1D and 3D genomic information: Challenges in modelling and validation☆

Alessandra Merlotti[a, b]

Angelo Rosa[c]

anrosa@sissa.it

Daniel Remondini[a, b, *]

daniel.remondini@unibo.it

[a]Department of Physics and Astronomy (DIFA), University of Bologna, Viale Berti Pichat 6/2, Bologna 40127, Italy

[b]INFN Sez., Bologna, Italy

[c]Scuola Internazionale Superiore di Studi Avanzati (SISSA), Via Bonomea 265, 34136 Trieste, (Italy)

[*]Corresponding author at: Department of Physics and Astronomy (DIFA), University of Bologna, Viale Berti Pichat 6/2, Bologna 40127, Italy.

☆This article is part of a Special Issue entitled: Transcriptional Profiles and Regulatory Gene Networks edited by Dr. Dr. Federico Manuel Giorgi and Dr. Shaun Mahony.

**Abstract**

Genome organization in eukaryotes during interphase stems from the delicate balance between non-random correlations present in the DNA polynucleotide linear sequence and the physico/chemical reactions which shape continuously the form and structure of DNA and chromatin inside the nucleus of the cell. It is now clear that these mechanisms have a key role in important processes like gene regulation, yet the detailed ways they act simultaneously and, eventually, come to influence each other even across very different length-scales remain largely unexplored.

In this paper, we recapitulate some of the main results concerning gene regulatory and physical mechanisms, in relation to the information encoded in the 1D sequence and the 3D folding structure of DNA. In particular, we stress how reciprocal crossfeeding between 1D and 3D models may provide original insight into how these complex processes work and influence each other.

This article is part of a Special Issue entitled: Transcriptional Profiles and Regulatory Gene Networks edited by Dr. Dr. Federico Manuel Giorgi and Dr. Shaun Mahony.

# 1 Introduction

The interplay between the 1D sequence and the 3D folding of DNA in chromosomes and cell nuclei is mediated by the delicate balance between classical physical forces stemming from the DNA nature as a long, tightly packed polymer filament [1,2,3,4,5,6] and complex chemical processes governing DNA and histone methylations, nucleosome positioning and the binding of transcription factors to DNA sequence [7,8] whose actions represent fundamental driving mechanisms in cell-fate decision [9]. For these reasons, understanding how the 1D genome affects its 3D spatial organization (and, viceversa) is a challenging task that requires a deeper understanding of both, the physico/chemical forces governing DNA folding and the mechanisms beyond gene regulation: advancing along this ambitious direction is compelling now more than ever, as it stands as the prerequisite for the comprehension of complex pathologies such as cancer [10], laminopathies and premature aging diseases like Hutchinson-Gilford progeria and Werner syndromes [11,12].

In eukaryotes, every ≈ 200 basepairs of the long DNA filament of each chromosome wrap around the histone complex [14], by creating a necklace-like linear sequence of *nucleosomes*, commonly known as the ~~10nm~~ ~~10~~10 ~~nm~~nm chromatin fiber, see Fig. 1. The present understanding of chromosome organization on spatial scales beyond the ~~10nm-fiber~~ ~~10~~10 ~~nm-fiber~~nm-fiber(in particular with respect to the existence of the “elusive” ~~30nm-fiber~~ ~~30~~30 ~~nm-fiber~~nm-fiber [15,13,16,17,18,19]) appears still remarkably confused.

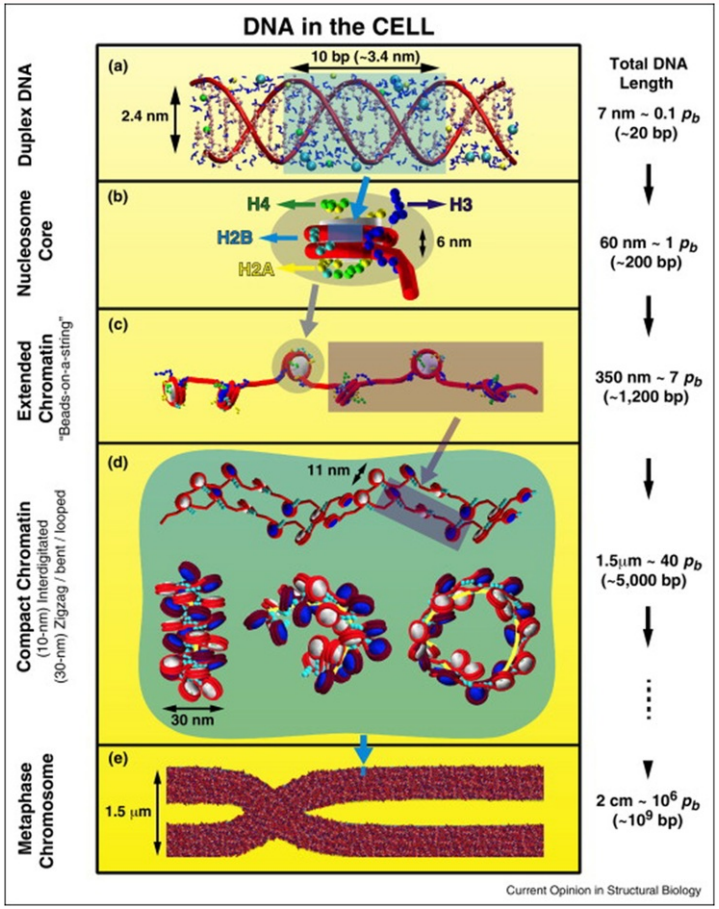



**Fig. 1** Principles of chromosome folding. I. Schemating cartoon of the 10 nm-fiber structure resulting from DNA wrapping around the histone complex. Chromatin folding beyond the 10 nm-fiber up to the scale of the whole chromosome remains controversia

*Source: Reproduced with permission from Ref. [13].*

However ambitious though, merging the information coming from the 1D/3D levels of knowledge promises to be increasingly affordable in the next future especially thanks to the recent, dramatic progress in sequencing techniques, such as the recent ATAC-seq and ChIA-Drop, which helped gaining new insights into the comprehension of 3D DNA organization as a function of 1D epigenetic "marks", in particular by allowing to map chromatin accessibility and nucleosome positioning genome-wide in a faster and more sensitive way than MNase-seq and DNase-seq [20] as for ATAC-seq, and revealing promoter-centred multivalent interactions in the ChIA-Drop case [21].

At the same time, high-precision/high-resolution experimental techniques have now greatly contributed to expand our understanding of the physico/chemical properties of DNA *in vivo*:

- Chemical "painting" of DNA sequences by "fluorescence *in situ* hibridization" (FISH) (Fig. 2) shows that chromosomes fold into compact conformations (chromosome "territories" [22,27]), which have non-random, gene-correlated locations inside the nucleus [23] and are crucial to cell correct behavior [22,23,27]: in particular, territories help keeping some sort of "physical barrier" between close-by chromosomes (see Fig. 2), with minimal amount of tangling [28] at the borders.
- Then, the internal structure inside each territory discloses itself by *chromosome conformation capture* techniques (3C [29]) and HiC [24]), which are based on chromatin-chromatin cross-linking followed by DNA sequencing (Fig. 3, top): this procedure showed that chromosomes display a checkerboard pattern of interactions [24] revealing some compartmentalization into open/closed mega-basepair-sized sub-domains (Fig. 3, top). At higher resolution, chromosomes cluster [25] into "topologically-associated domains" (TADs), regions separated by boundaries enriched for specific protein factors and identified by the unusually high number of contacts recorded in the each TAD's interior which drops suddenly at the

boundaries (see the heat maps in Fig. 3, bottom). Interestingly, chromosome organization into TADs appears “universal”, being both stable across different cell lines and across different species [26].

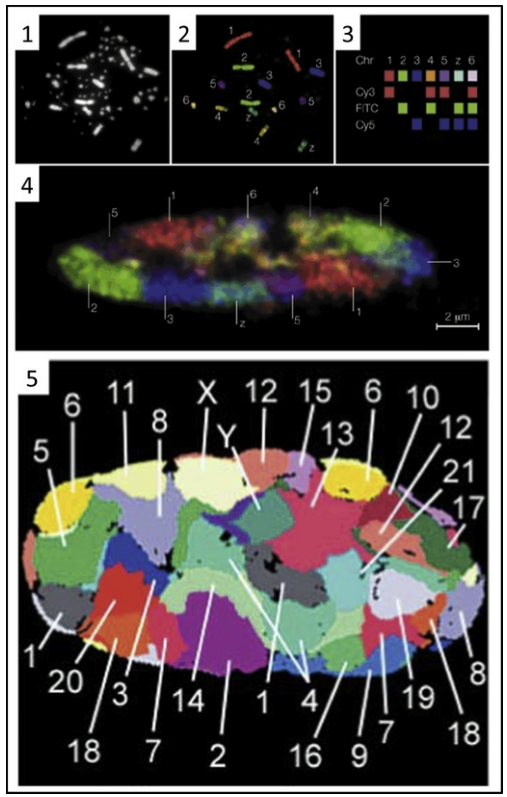
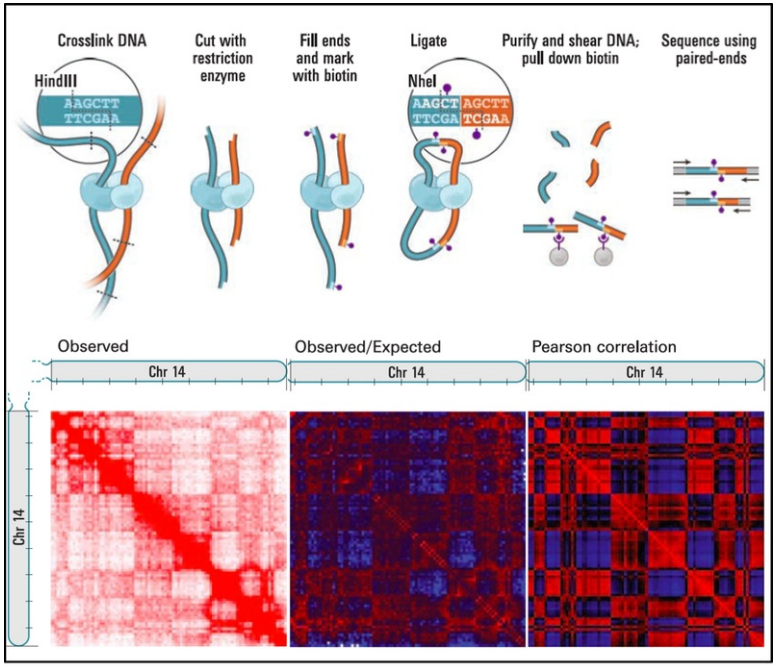



**Fig. 2** Principles of chromosome folding. II. Chromosome “painting” by FISH (panels 1 to 3) reveals that chromosomes occupy distinct territories within the nucleus: panel 4 and panel 5 show examples of chromosome territories in chicken and human fibroblasts, respectively.

*Source: Panels 1 to 4 are reproduced with permission from Ref. [22], Panel 5 is reproduced from Ref. [23] under Creative Commons License.*



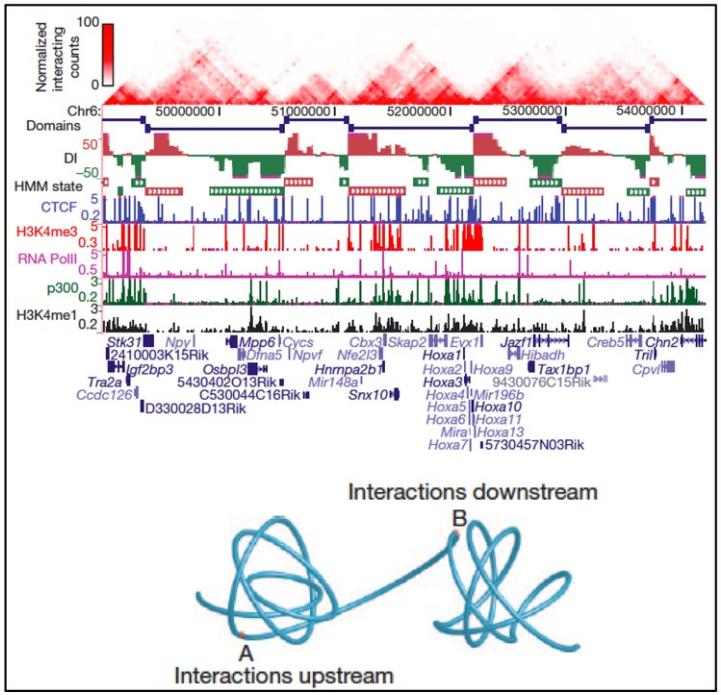




**Fig. 3** Principles of chromosome folding. III. (Top) Chromatin contacts by HiC (top) at 1 Mbp-resolution can be visualized in terms of heat maps. These maps show a “plaid-pattern” structure of intra-chromosome contacts stemming from chromosome

compartmentalization into two (A/B) sub-compartments (bottom). Reproduced with permission from Ref. [24]). (Bottom) At higher resolution ( $\lesssim 100$ kbp), chromosomes appear organized into *topologically-associated domains* (TADs), regions characterized by unusually frequent contacts well separated by narrower regions almost interaction-depleted [25,26]. TADs correlate well with known epigenetic marks. Reproduced with permission from Ref. [25].

With these premises, the hard core of the challenge lies in elaborating appropriate models that take into account or even integrate the 1D/3D levels of information. In this review, we discuss the state-of-the-art of computational approaches which – in our opinion – have best contributed to shed new light on this fascinating and promising research field. To this purpose, we adopt the following outline:

- In Section 2 we discuss 1D models for understanding non-random features in DNA sequences.
- In Section 3 we present a comprehensive catalogue of theoretical models based on polymer physics which describe relevant aspects of chromosome folding. Unless stated differently, we consider models for chromosome conformations during interphase [14], *i.e* chromosomes within nuclear confinement.

In both sections we talk mainly of genome structure in higher eukaryotes (like mammals), occasionally though we generalize to other classes of organisms.

Finally, we conclude the work (Section 4) by highlighting promising directions for future work, in particular with regard to what one can possibly learn by exploiting the connections between 1D and 3D modeling approaches.

# 2 Reading the sequence: 1D models for nucleotide organization

At the 1D level, the DNA sequence can be represented as an ordinary string of text composed by four letters corresponding to the four nucleotides: A, C, G, ~~and~~and T. This simple representation allowed to treat genomes as symbolic sequences and thus to exploit the knowledge developed in the fields of physics and statistics to extract information about their structure. In particular, two approaches have revealed helpful to identify some peculiar structural properties of genomic sequences that are involved in gene expression regulation, such as coding and non-coding regions [30,31], enhancers [32] and CpG islands [33]: DNA random walks and dinucleotide interdistance.

## 2.1 Random walks on DNA sequences

One of the first models of random walk on DNA sequence [30] was defined according to the following rule: the walker steps up ($u(i) = +1$) if a pyrimidine ('C' or 'T' nucleotides) occurs at position $i$ along the sequence, otherwise for the opposite case of a purine ('A' or 'G' nucleotides) the walker steps down ($u(i) = -1$). This simple rule allows to calculate the displacement of a walker after $l$ steps as

$$y(l) = \sum_{i=1}^{l} u(i) \tag{1}$$

and to identify regions with different purine-pyrimidine content by plotting $y(l)$ as a function of nucleotide distance $l$ (see Fig. 4), where positive slopes correspond to high concentration of pyrimidine and negative slopes correspond to high concentration of purines [34]. The power of this simple approach is that different hypotheses on DNA sequence organization can be mapped onto specific "null models" about the characteristics of such random walks, and can thus be tested against the properties of the real sequences. A fundamental statistical quantity characterizing any walk is the root mean square fluctuation $F(l)$ around the average displacement:

$$F^2(l) = \overline{[\Delta y(l)]^2} - \overline{[\Delta y(l)]}^2 \tag{2}$$

where $\Delta y(l) = y(l_0 + l) - y(l_0)$ and the bars indicate an average over all positions $l_0$ on the gene. The calculation of $F(l)$ is a key step in order to identify "anomalous" diffusion. In fact, in pure "random" walks $F(l) \sim l^{1/2}$; otherwise, $F(l) \sim l^{\alpha}$, with $\alpha \neq 1/2$, thus revealing long-range correlations between walk steps, corresponding to correlations in nucleotide positioning process. One of the earliest and most relevant results obtained by applying this method concerns the identification of coding and non-coding sequences inside genes [30]. In particular, long-range correlations were identified as systematic markers of the presence of intron-containing genes and non-transcribed genomic regulatory elements, whereas, the absence of long-range correlations is characteristic of cDNA sequences and genes without introns (Fig. 4).

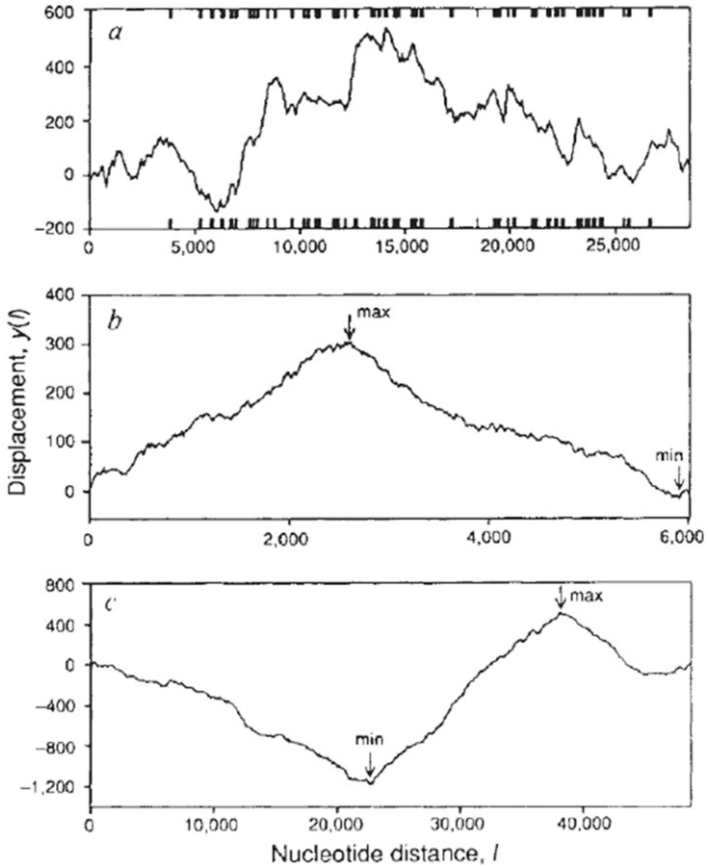



**Fig. 4** The DNA walk representation of intron-rich human $\beta$-cardiac myosin heavy-chain gene sequence (a), its cDNA (b), and the intron-less bacteriophage $\lambda$ DNA sequence (c). Note the more complex fluctuations for the intron-containing gene in (a) compared with the intron-less sequences in (b) and (c). Heavy bars denote coding regions of the gene.

*Source: Reproduced with permission from Ref. [30].*

Moreover, long sequences (thousands of base pairs) were found inside non-coding regions, which were characterized by long-range correlations, and this led Buldyrev et al. [35] to apply a generalized Lévy-walk model to non-coding sequences, and to hypothesize the existence of DNA loops. In generalized Lévy-walks, the typical walk step $l_j$ can be very long (in fact, walk steps are distributed according to a power-law distribution $P(l_j) \propto 1/l_j^{\mu}$ with $2 < \mu < 3$), implying the existence of correlations between the displacements of nucleotides at very long mutual distances.

The authors provided a molecular basis for the power-law distribution of step lengths by hypothesizing that, in order to be inserted into DNA, a macromolecule should form a loop of length $l_j$, whose ends come close to each other in the space. In fact, Buldyrev et al. et al.et al. pointed out that the long uncorrelated subsequences inside non-coding regions may correspond to repetitive elements, such as LINE-1, or retroviral sequences.

## 2.2 Dinucleotide interdistance

Another approach results very powerful at identifying structural genomic features at the 1D level: the study of dinucleotide interdistance distributions. The idea is inspired to the theory of first-return-time distributions in stochastic and deterministic processes by H. Poincaré, who developed this model to study the trajectories of bounded dynamical systems [36].

Referring to genome sequences, the analysis can be carried out through the following steps: given a dinucleotide $XY$, where $X$ and $Y$ can take any value among $\{A,C,G,T\}$, its interdistance distribution $\hat{p}(\tau)$ can be calculated by

(i) identifying the positions $x_j$ (~~$j = 1,2,..,j = 1,2,...$~~ $j = 1,2,...$ ) of each *XY* along the sequence, (ii) calculating the distance between two consecutive XY as $\tau_j \equiv x_{j+1} - x_j$, (iii) counting the abundance of a given interdistance value $\tau$ and (iv) estimating its relative frequency $\hat{p}(\tau)$ according to the formula:

$$\hat{p}(\tau) = \frac{\#\{j = 1, 2, \ldots | \tau_j = \tau\}}{\#\{j = 1, 2, \ldots | \tau_j\}}, \quad \textbf{(3)}$$

where the numerator counts all values where $\tau_j = \tau$ while the denominator runs over all unrestricted values $\tau_j$.

The first analysis of this quantity on genome sequences [37] showed that dinucleotide interdistance distributions have a pronounced period-3 oscillatory behaviour in protein-coding regions which is absent in the whole-genome distributions and appears to be related to the triplet structure of the protein-coding genetic code. Furthermore, the comparison between real distributions and randomly generated ones revealed that the behaviour of CG dinucleotides is considerably different from all the others. This study opened the avenue to subsequent works that led to methods for the identification of CpG islands [33], and to a more general characterization of CG interdistances in association to DNA methylation functionalities [38,39]. In particular, the work of Paci et al. [38] revealed that CG interdistance distribution in higher-order organisms greatly differs from all other dinucleotides (see the comparison between *Homo sapiens* and *Mus musculus* in Fig. 5), showing the strong exponential decay

$$\hat{p}(\tau) \sim e^{-\tau/b}. \quad \textbf{(4)}$$

This difference seems to be related to the different role that methylation plays in this class of organisms [38].

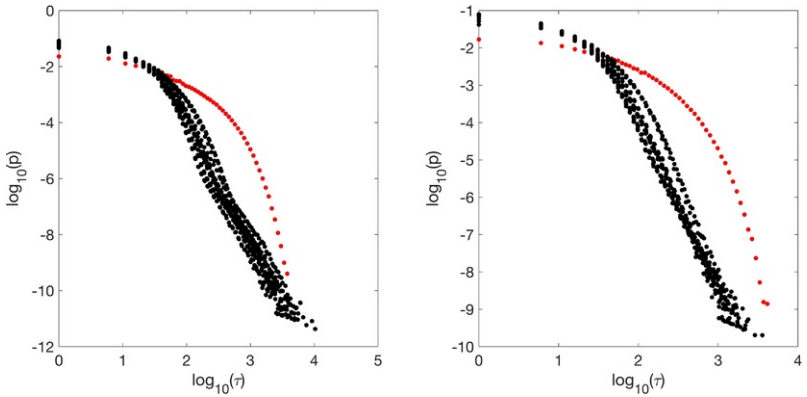



**Fig. 5** Distribution functions of dinucleotide interdistances ($\tau$, measured in units of DNA basepairs) in log-log scale for *Homo sapiens* (left) and *Mus musculus* (right). The distribution for CG dinucleotides is represented in red.

Interestingly, in higher-order organisms the characteristic "length-scale" $b$ of Eq. (4) measuring the average contour length distance between consecutive CGs showed a value ~~$200bp < b < 300$ bp~~ ~~$200$ bp $< b < 300$ bp~~$200$ bp $< b <$ $300$ bp, which is comparable to the typical DNA filament wrapped around the histone complex [14].

An even deeper analysis of CG interdistance distributions was performed in human genome, by identifying the so-called Gamma-distribution

$$\hat{p}(\tau) \sim \tau^{a-1} e^{-\tau/b} \quad \textbf{(5)}$$

as the best fitting model distribution [39]. Furthermore, in this work the authors extended the study to a large variety of organisms spanning all available ranges of biological complexity, finding that the value of parameter $b$ is correlated to the biological complexity of the organism category: in fact, it steadily increases moving from bacteria to vertebrates (see Fig. 6, left) and it is strongly correlated to CG density (CG%), displaying in particular a power-law behavior $b \propto \mathrm{CG}\%^m$. The study showed that all categories, except vertebrates, are characterized by an exponent $m \sim -1$, which is compatible with a simple null model predicting that the average distance between dinucleotides is inversely proportional to the dinucleotide density inside the sequence. For vertebrates instead, the exponent $m$ takes the value $\simeq -0.5$ which is significantly higher in comparison to the other classes of organisms considered (see Fig. 6, right): we speculate that this might be related to a different mechanism for CG positioning along the genome connected to the DNA methylation process that CG dinucleotides undergo in this class of organisms.

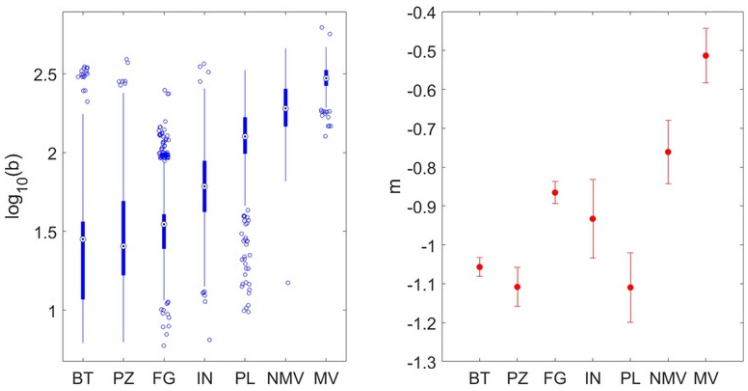



**Fig. 6** (Left) Box-plots for the Gamma-distribution scale parameters *b* (see Eq. (5)) for seven categories of organisms: bacteria (BT), protozoa (PZ), fungi (FG), invertebrates (IN), plants (PL), non-mammal vertebrates (NMV) and mammal vertebrates (MV). (Right) Estimated average values and error bars for *m* exponents relative to the same classes of organisms.

These results show how detailed features of 1D DNA sequence are able to capture key properties of gene regulatory mechanisms that go beyond the 1D environment, such the extension of coding and noncoding regions, or the footprints of DNA methylation. In general, the relative positioning of specific DNA sequences along the genome might reflect their role in a specific 3D context, in particular where complex loop structures can bring close to each other motifs quite far apart along the sequence, similar to what happens during the folding process of peptide chains. The identification of the correct distributions of these distances can help to restrict the type of modelling processes able to generate them, thus helping to clarify the biology, chemistry and/or physics behind these far-from-trivial conformations.

# 3 Folding the sequence: 3D models of chromosome organization in eukaryotes

"Predicting" 3D chromosome structure starting from the 1D DNA sequence – a question reminding in some way of the analogous protein folding problem [40] – is a long-standing problem in cell biology and a very challenging one. Although the two problems (DNA folding and protein folding) may appear similar, a huge difference lies in the fact that DNA structure is not only guided by the chemical properties of its components (as for protein peptides) but relies on the complex interplay with many epigenetic factors (histones, noncoding RNAs, cohesins, lamines, etc.) that can be guided by "signals" set along the native DNA sequence (transcription factor binding sites, enhancer/promoter binding, DNA methylation, etc.) some of which, possibly, might be still unknown. Moreover, chromosome state may depend on other important factors like, to mention a few, the particular cellular type, phase of the cellular cycle, gene activity and the mechanisms beyond DNA repair [14].

Fortunately, the rapid development and increasing availability of structural data on chromosome organization (FISH and HiC *in primis*, see Section 1) alongside with the more and more sophisticate analysis tools (see Section 2) which are now capable of detecting finer and finer correlations in the 1D DNA sequences are rapidly shifting the field towards a more confident description of how chromosomes fold inside the nucleus and how it reverberates on chromosome function.

As for it, in recent years there has been an impressive "explosion" of models trying to fill the missing conceptual gap between the 1D DNA or chromatin sequence and the 3D chromosome packing inside the nucleus. Interestingly, most of (if not all) these models have been proposed by physicists and are based on the (rather obvious) assumption that chromosomes are long polymer chains subject to the same classical [41,42] laws of polymer physics: these laws can then be used to predict the *in vivo* chromosome behavior and, then, make quantitative and testable predictions.

As it has been stressed in the ~~Introduction~~ Introduction (Section 1), chromosome structure inside the nucleus remains highly controversial. It is no surprise then, that there exists a conspicuous literature concerning different polymer models presenting alternative scenarios to illustrate the link between chromosome sequence and folding. In the next sections we will discuss in more detail some of these models and the physical bases of each of them.

To better accomplish this purpose, it is instructive to classify the models into two categories:

1. In the first category (Section 3.1), we place those models which rely on relatively few, minimal physical assumptions. The idea behind these approaches is that certain features of chromosome organization are common to all species and, in some respect, are *more important* than the details contained in each DNA sequence which make each species so different from any one else. Minimal models of this kind are extremely useful and instructive because they constitute the paradigm to understand the "nuclear" forces which continuously remodel the genomes.

2. In the second category (Section 3.2), we consider those polymer models which are constructed to satisfy a certain number of constraints obtained from experimental results. For this reason, we name these *data-driven models*. These approaches are now becoming especially popular, for one hopes to employ them in the near future to provide accurate predictions on how genomes react when the "native" conditions upon which they were constructed change as the result of some stress on the cell or because of some induced mutation on the DNA sequence.

## 3.1 Chromosome organization by generic, "bottom-up" polymer physics

### *3.1.1 ~~1.~~ The role of topological constraints*

Chromosomes are constituted by long chromatin filaments tightly packed inside the nucleus. By neglecting all details related to the heterogeneity of DNA sequences, at first approximation the entire system of chromosomes contained in the nucleus can be described as a solution of polymer chains [41,42] subject to thermal fluctuations. Under these conditions topological constraints, which are known to force nearby polymer chains to move randomly by *sliding* past each other without *passing* through each [41,42], are expected to play a key role by affecting chromosome structural and dynamical properties.

In fact, it is a non-trivial question to ask how a single centimeter-long chromosome can be efficiently stored inside the nucleus which is typically about thousand times [14] narrower. While the presence of histone complexes and territories point towards the fact that chromosomes maintain a certain level of compactness, they say nothing about how compactness can be practically and efficiently achieved. In this respect, physical theories of polymers may become useful.

A major turning point occurred in the late '80s when Grosberg and colleagues published two influential papers [48,43], suggesting that the DNA or the chromatin fiber of a single chromosome should exist in an unknotted, off-equilibrium state which they termed "the crumpled globule", see Fig. 7 (A). Intuitively, this model can be constructed by assuming that the linear DNA sequence folds by hierarchical compaction from small up to the largest scales: this fractal-like conformation features the two advantages of being maximally packed *and* knot-free.

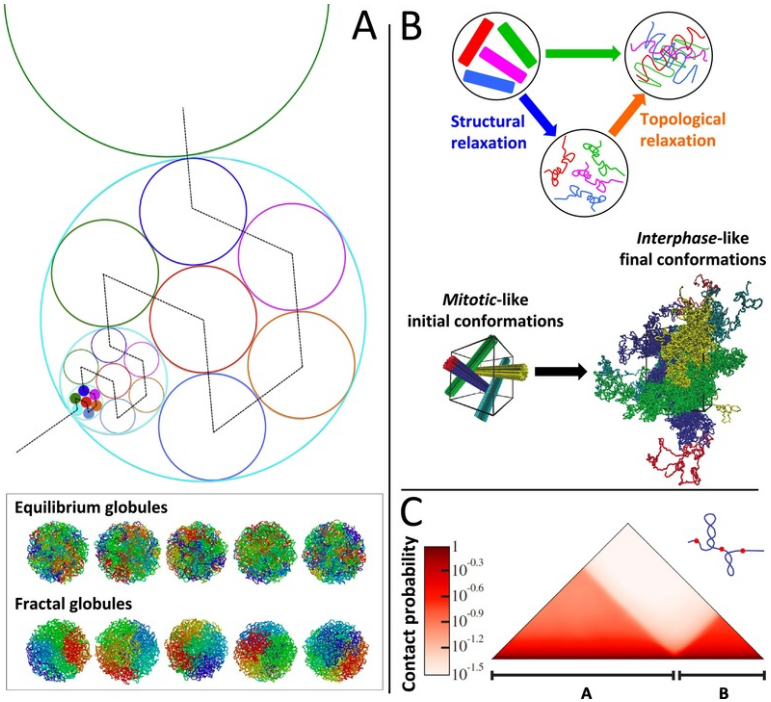



**Fig. 7** The role of topological constraints in chromosome organization. (A, top) Schematic illustration of the "crumpled globule", showing the different layers in the hierarchical folding. The fundamental units (the monomers, filled spheres) fold into globular structures of larger sizes (the smaller empty spheres), acting in turn as "super"-monomers in the following crumpling event. The process proceeds then at the next stage, and so on. The final structure resembles a fractal [43] with maximal compactness. (A, bottom) Examples of polymer conformations obtaine by computer simulations, illustrating the structural differences between equilibrium and crumpled (fractal) globules. Reproduced with permission from Ref. [24]. (B) Because structural and topological relaxations of mitotic-like conformations have markedly different time-scales, chromosomes remain effectively "trapped" into territorial-like conformations [44,45,1,2]. Reproduced from Ref. [44] under Creative Commons License. (C) Chromatin fibers with (negative) levels of supercoiling form TAD-like structures [46,47], reproducing contact patterns observed i

HiC experiments.

*Source: Reproduced from Ref. [46] under Creative Commons CC BY License.*

From the theoretical point of view, two possible mechanisms leading to a crumpled globule were suggested: either by fast switching the solvent conditions of the polymer chain from “good” to “bad” (*i.e.*, polymer self-interactions turn from repulsive to attractive [48]) or by fast confinement of the polymer into a narrow region [24]. Either way, the chain has no time to fully relax from its initial (knot-free) conformation, the final state being crumpled and displaying the presence of domains. Conversely, when the process of crumpling is slow, the final state is akin to an “equilibrium” globule with no domains (see the comparison between the two contrasting sets of model polymer conformations in Fig. 7 (A)). Although interesting from the theoretical point of view, fast crumpling is not expected to take place inside the cell.

In 1999, Langowski and collaborators introduced the so-called *random-loop model* [49]: interphase chromosome structure was described in terms of a self-repulsive random polymer with pairs of monomers permanently bound to form small loops on the scale of ~ 100 kbp, rosette-like sub-compartments on larger scales and territories imposed by artificial confinement of the polymer chain. The model was later [50] applied to describe the 3D structure of the murine immunoglobulin *heavy-chain* locus. The random-loop model appears in qualitative agreement with chromatin organization into TADs and territories, however this is not entirely surprising because these motifs were directly *imposed* on the model and, then, not really *explained*.

Instead, crumpled conformations can be easily obtained through a very simple physical mechanism which looks almost as the “reverse” of the one considered in the construction of a crumpled globule. In two publications [44,45] Rosa and Everaers presented a polymer model for chromosome organization implying that territories emerge “spontaneously” as the result of the slow relaxation of the mitotic-like original chromosome structure (Fig. 7 (B)): in other words, the microscopic topological chromatin state remains quenched *in time* with no chance to relax and chromosomes get trapped into crumpled, territory-like conformations. It was proposed [44] then that the physical mechanism underlying chromosome compaction is the same driving the folding of untangled ring polymers in concentrated solutions [2,51,52,53]. As demonstrated in [44,45], the proposed model is able to capture *quantitatively* generic chromosome features like internal chromatin-chromatin distances and HiC contact frequencies with no fitting parameters, and can be used to model chromosome dynamics on time-scales from seconds to days in real time. Third, it can be also naturally generalized [54] so to take into account the heterogeneity of DNA sequence.

We conclude the section connecting chromosome organization and the topological properties of the chromatin fibers by mentioning some recent work by the Stasiak's group in Lausanne [46,47] which suggests a possible link between the presence of supercoiling in chromatin and TADs (mentioned in Section 1). Chromosomal DNA is expected to be naturally supercoiled due to continuously ongoing processes like replication and transcription. This excess of supercoiling is expected to never relax, once again because of the typically large size of chromosomes. It may thus induce local crumpling of the chromatin fiber, similar to what occurs to a familiar phone cord when excessive twist is applied. By fine-tuning the amount of supercoiling in a numerical polymer model for chromatin fibers, Stasiak and colleagues showed that the phenomenology of TADs, summarized by the excess of intra-domain contacts with respect to inter-domain contacts (see Fig. 7 (C)), can be generically captured.

### 3.1.2 ~~H.~~ Sequence-specific chromatin-chromatin interactions

The polymer models presented in Section 3.1.1 show that notable chromosome features like intra-DNA positions and contacts may be quantitatively understood in terms of the same theoretical mechanisms describing the phenomenology of entangled polymer solutions. On the other hand, there is more to chromosome biology which requires a thorough discussion.

In this respect, it is known that certain species of protein complexes present in the nucleus tend to bind to specific DNA target sites and influence chromosome organization: important examples include the CCCTC binding factor (CTCF) involved in promoter activation or repression and methylation-dependent chromatin insulation [55] and the trascription units which by clustering into transcription “factories” [56] mediate and regulate the production of transcripts. The role of these protein-DNA interactions in chromosome architecture has been addressed in an increasing number of publications.

In the so-called “strings-and-binders-switch” (SBS) polymer model [57], chromatin is described as a *block copolymer* where a certain fraction of monomers (the “binders”) act as binding sites for freely diffusive particles, see Fig. 8 (A). The binding of particles to DNA is dynamic (binders attach and detach intermittently at finite rates), the mechanism being described in terms of two phenomenological parameters: the binder affinity ($E_X$) and the binder concentration ($c_m$). It is then possible to construct a phase diagram in the $E_X$-$c_m$ space where a single line separates swollen from compact polymer conformations, as in the classical $\theta$-collapse [42,57] in polymer physics. The SBS model predicts that as per adaptation to continuously-changing external conditions chromatin is switching between these two states through a suitable combination of the concentration/affity of the binders, thus accounting qualitatively for the observed fluctuations in chromatin loci spatial positions and contacts as measured in FISH and HiC.

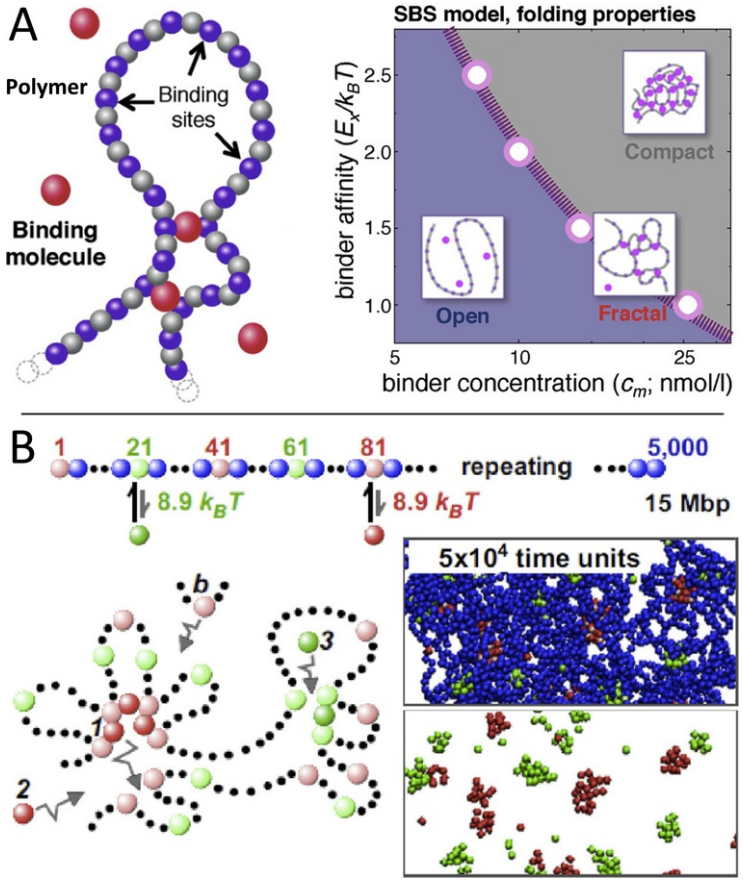



**Fig. 8** The role of sequence-specific chromatin-chromatin interactions in chromosome organization. I. (A) In the "strings-and-binders-switch" (SBS) model, chromatin acts as a block copolymer with site-selective affinity $E_X$ for specific binders at concentration $c_m$. Chromatin folding/unfolding can be represented in terms of the phase diagram in these two parameters. Reproduced with permission from Ref. [57]. (B) Protein-like particles mimicking transcription factors binding to cognate sites on a block copolymer model promote chromosome compaction by forming rosettes and TAD-like domains. The model predicts also the spontaneous self-assembly of proteins into factories.

*Source: Reproduced from Ref. [58] under Creative Commons CC BY License.*

In a variation of the SBS-model, Brackley et al. [58] pointed out that protein-like particles mimicking transcription factors which bind reversibly to cognate sites on a block copolymer model promote chromosome compaction, see Fig. 8 (B). This model outline a picture where a single chromosome is organized into spatial motifs like rosettes and topological domains similar to the ones observed in HiC experiments. Interestingly, as a by-product the model predicts that proteins self-organize into clusters (or, factories [56]) due to a "bridging-induced attraction" which is mediated by polymer folding.

Alternatively (or, in addition) to the action of the binders, the observed chromosome organization may be the consequence of the partitioning into a small [61] set of distinct epigenomic domains which cluster together by *epigenome-dependent* attractive interactions. Jost and collaborators [59,62] have implemented this idea into a copolymer model, where each monomer of a specific epigenomic domain bind exclusively to monomers of the same species. The chromatin fiber associated to each chromosome thus segregate by a physical mechanism known as *microphase separation* (see Fig. 9 (A)) which displays a checkerboard pattern of contacts which may explain chromosome structure into TADs (reported in Section 1). In a related study involving a very similar computational set-up, Shi ~~*et al.*~~ ~~et al.~~et al. have shown [63] that chromatin dynamics is highly heterogeneous, reflecting the observed cell-to-cell variations in the contact maps: folding is a two-step, hierarchical process which involves the formation of TAD-like chromatin domains (or, droplets) followed by their "fusion" inside the entire territory.

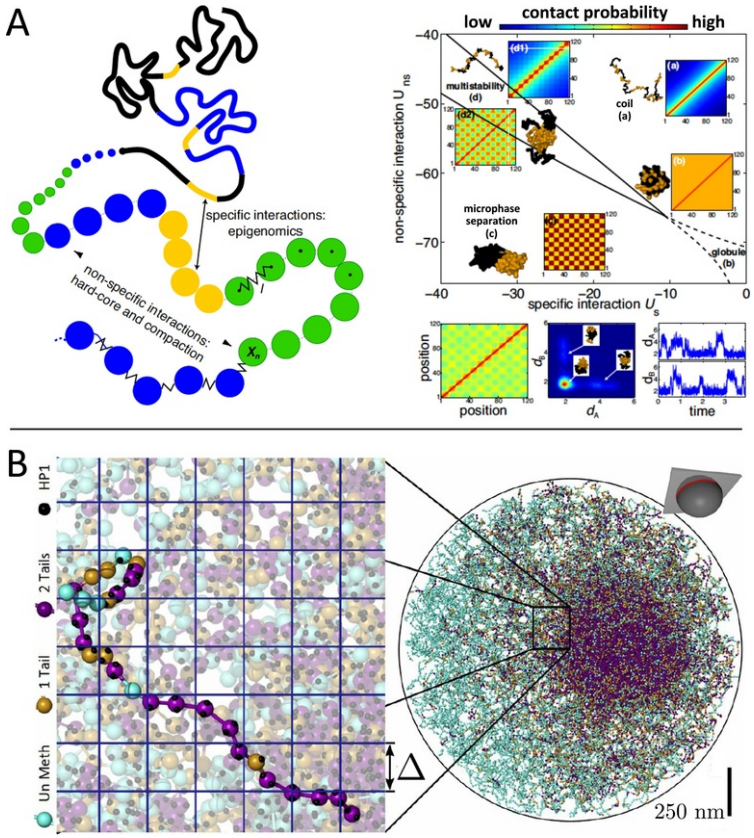



**Fig. 9** The role of sequence-specific chromatin-chromatin interactions in chromosome organization. II. (A) Chromosomes may fold due to epigenome-specific attractive interactions promoting microphase segregation and TAD-like organization. Reproduced from Ref. [59] under Creative Commons CC BY License. (B) Trimethylated chromosomal sites attracting each other by the mediated action of oligomerized HP1 model proteins drive phase segregation into compact, heterochromatin domains *vs.* swollen, euchromatin domains. Reproduced with permissior from Ref. [60].

An interesting hypothesis on the connection between epigenetic marks (specifically, histone methylation) and chromosome folding has been recently formulated by MacPherson et al. [60]. By using Monte Carlo computer simulation of a nucleosome-resolved polymer model complemented by H3K9me3-methylation patterns from ChIP-seq data, the authors suggested that dimerization of HP1 single protein units which bind preferentially to methylated chromatin sites drive chromatin segregation into heterochromatin (dense and H3K9me3-rich) and euchromatin (open and H3K9me3-poor) domains, see Fig. 9 (B). The segregation results in plaid-patterned heat-maps resembling those obtained in HiC experiments.

### *3.1.3 ~~III.~~ Out-of-equilibrium effects: loop extrusion and activity-induced phase separation*

Life is a dynamic process maintained through the continuous contribution of external energy sources: as such, in recent years a conspicuous body of work on the experimental and theoretical aspects of non-equilibrium physics has had a tremendous impact on our understanding of how living matter works [66]. In this respect, chromosomes are no exception. In the following, we summarize a few works which have contributed to highlight the role of non-equilibrium mechanisms with regard to chromosome organization.

Ganai et al. [64] suggested that certain reported correlation between chromosome positioning within the nucleus and gene density (see Section 1) can be understood as the consequence of different “activity” levels: similarly to the approaches described in previous sections chromosomes are modeled as coarse polymers, however – in contrast to the purely passive systems discussed so far – here each monomer is classified according to its level of activity (proportional to gene density) and coupled to a specific, effective temperature. Thus, a higher effective temperature means a larger activity. With the addition of a given amount of permanent loops between chromatin fibers, this models shows that chromosomes tend to be partitioned into clusters of

different temperatures, see Fig. 10 (A). A rigorous physical explanation of this phenomenon was provided in Ref. [67] and later confirmed in Ref. [68] by means of systematic computer simulations: even small temperature gaps induce phase separation in systems of colloids or polymer chains. In spite of the intrinsic *out-of-equilibrium* nature of the system, it can nonetheless be shown that the phenomenon can be captured by the analogy to the classical *equilibrium* theory of binary mixtures which phase separate as the result of distinct chemical affinities [42].

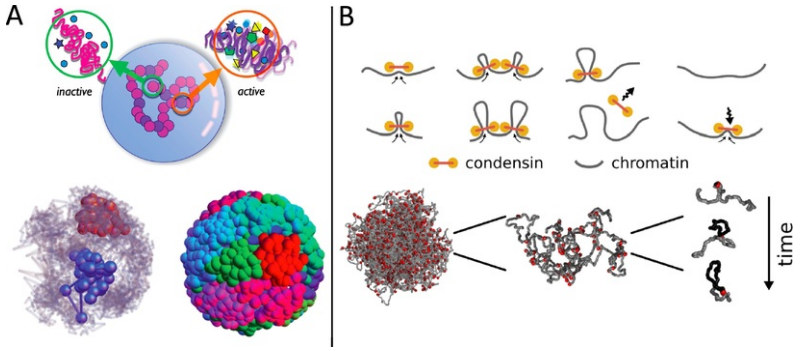



**Fig. 10** The role of active processes in chromosome organization. (A) Chromatin is classified as "inactive" and "active" depending on its gene content (top). Gene-poor and gene-rich chromosomes phase separate and form territories whose spatial positions with respect to the nucleus correlate with experimental observations (bottom). Reproduced from Ref. [64] under Creative Commons CC BY License. (B) The condensin complex (in yellow) bind to the chromatin fiber (in black) and, by moving into opposite directions, effectively produces chromatin loop extrusion. Extrusion stops when two (or more) complexes bump unto each other (top). An apparently disordered tangled mass of chromatin can then self-organize into a regular array of extruded loops (bottom). Reproduced from Ref. [65] under Creative Commons Attribution License.

Recently, it has been pointed out that *active loop extrusion* may be universally responsible for chromosome segregation during mitosis [65,69] and for chromosome compartmentalization into TADs [70]. Specific proteins called "condensins" assemble into complexes and bond together spatially close loci on the chromatin fiber, see Fig. 10 (B). Then, the chromatin filament fixed by the condensins starts to be effectively *extruded* when the complex moves into opposite directions along the fiber. When two condensins collide into each other the translocation process stops. Moreover, with the addition of topoisomerase-II the loop extrusion mechanism is able to simplify chromosome topology by removing knots and links [71,72] between chromatin fibers within the crowded environment of the nucleus.

## 3.2 Building chromosomes by data-driven, "top-down" polymer models

The polymer models illustrated in Section 3.1 employ minimal physical assumptions in trying to capture various aspects of chromosomes phenomenology and, for this reason, they have been generically termed "bottom-up". The most fascinating side of these approaches is that they often make testable predictions which are amenable to experimental validation.

Recently, a number of studies have attacked the problem of chromosome organization from a radically different perspective: instead of explaining experimental observation by employing minimal physics why not using the information contained in the experiments to deduce the *most probable* chromosome conformations compatible with the observations?

In two related studies, Di Stefano and coworkers showed that by just enforcing co*localization* of co*expressed* genes in a polymer model for human chromosome 19 first [73] and then for the entire human genome [77] without major additional constraints, the resulting conformations (see the example shown in Fig. 11 (A)) appear compatible with chromatin classification in A/B sub-domains and with the non-random locations of chromosome territories correlated to gene content (see Section 1).

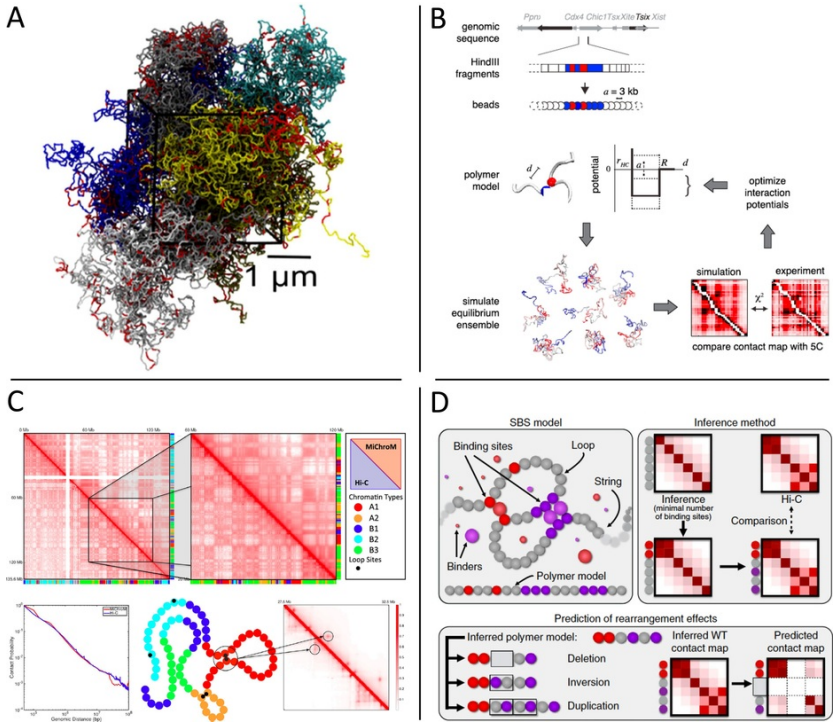



**Fig. 11** Data-driven polymer models. (A) A polymer model promoting colocalization of coexpressed genes in human chromosome 19 produces conformations organized in spatial macrodomains which correlate with HiC [24] predictions. Reproduced from Ref. [73] under Creative Commons License. (B) A single TAD is modeled as a polymer chain whose beads interact *via* a square-well potential with an attractive wall. The energy parameters are optimized by iteration of a Monte Carlo sampling scheme so to maximize the agreement between the *predicted* and *observed* chromosome conformation contacts maps. Reproduced with permission from Ref. [74]. (C) In the Minimal Chromatin Model (MiChroM), chromatin loci are classified into different types (colors) and certain pairs of genomic loci ("anchors") tend to form loops. The interaction potential for the polymer chain is *trained* based on the HiC [24] contact matrix for human chromosome 10, and used then to construct and study the spatial features of the other chromosomes. Reproduced with permission from Ref. [75]. (D) The Polymer-based Recursive Statistical Inference Method (PRISMR) refines the SBS polymer model [57] by "filtering" the simulated chromosome conformations so to derive the minimal set of binding sites and binding molecules which best reproduces the input HiC contact matrix. Instructing the model on wild-type (WT) chromosome data, the effects of genomic mutations (deletions/inversions/duplications) on abnormal chromosome conformations can be then predicted without further additional parameters. Reproduced with permission from Ref. [76].

In order to exploit the nature of TADs and of chromatin-chromatin interactions measured within a single TAD, Giorgetti et al. [74] introduced a computational polymer model (see. Fig. 11 (B)) where sequence-dependent monomer-monomer interactions were obtained upon maximizing the agreement between contact frequencies predicted by the model and the ones measured by ordinary conformation capture techniques. The model, targeted onto a specific region of mouse chromosome X, reveals that the structure of a single TAD measured by HiC reflects a full ensemble of fluctuating conformations across the cell population with no stable loops. Interestingly, the model was later tested by inducing a deletion at a specific locus and measuring the altered spatial distances.

A similar approach, the Minimal Chromatin Model (MiChroM), was introduced recently by Di Pierro et al. [75] with the intent of expanding the analysis to an entire chromosome and trying to export the derived force-field to describe the whole diploid nucleus. Specifically, polymer loci were classified into chromatin types (as in some of the models considered in Section 3.1.2) and the energy parameters describing the interactions between them were trained by using HiC data for human chromosome 10 from a specific cell line, see Fig. 11 (C). The model was then used *to predict* an ensemble of possible structures for the other chromosomes not used for the training of the energy function: interestingly, the obtained maps match well the ones obtained by HiC and the simulated chromosome structures recapitulate other notable features of interphase chromatin, like microphase separation of chromatin types (Section 3.1.2) and the tendency of open chromatin to remain at the periphery of its territory.

Finally, Bianco ~~*et al.*~~ ~~et al.~~et al. refined the SBS model discussed in Section 3.1.2, by introducing the Polymer-based Recursive Statistical Inference Method (PRISMR) [76]: PRISMR works by minimizing a cost function which –

again – takes into account the predicted *vs.* the measured HiC contact frequencies, see Fig. 11 (D). The "optimal" polymer model is then exported to construct chromosome conformations for a number of so-called structural variants of chromosomes which are known to produce anomalous chromatin folding and diseases. The protocol is then shown to be very efficient in detecting mutated chromatin-chromatin interactions which are involved in anomalous phenotypes: the work reports in particular the example of the *EPHA4* locus where specific deletions are associated to anomalous polydactyly.

# 4 Discussion

In this article, we have described some of the most popular modelling approaches to 1D and 3D features of genomic DNA sequences.

With regard to 1D features, we have shown (Section 2) evidence of nontrivial displacement of nucleotides along the sequence: (1) at the single-nucleotide level, since pyrimidines and purines are not randomly distributed but show long-range correlations up to kb scale (Section 2.1) and (2) at a dinucleotide level (Section 2.2), in particular CG-dinucleotides associated to DNA methylation, for which the distribution of mutual interdistances along the genome shows a different behaviour from the other dinucleotides and seems correlated to specific regulation mechanisms (CpG islands) or to organism complexity. Thus, the analysis of 1D sequences in these specific cases reveals important properties that go *beyond* the 1D environment itself, and likely have an impact on (or are influenced by) the surrounding 3D context.

In order to understand how genomes fold in 3D, we have presented recent work about molecular modeling (Section 3) of chromosomes. In this respect, the state-of-the-art is remarkably complex: topological effects (Section 3.1.1), specific DNA-DNA interactions (Section 3.1.2), energy-driven, active (opposed to entropy-driven, passive) mechanisms (Section 3.1.3) are all likely to act concurrently. Future work has to dissect one by one each of these mechanisms with the goal to understand their relative importance with respect to the full picture.

Inspired by the phenomenology of the "protein folding" problem [40] where the aminoacid sequence contains the essential information to drive the protein towards its unambiguous, "native" structure, it is natural to ask to which extent the 1D sequence influences the 3D chromatin architecture, provided that epigenetic factors are a key player to be associated to DNA sequence. Two recent complementary approaches suggested that a significant amount of spatial contacts detected by chromosome conformation capture techniques can be predicted based on the spatial colocalization of transcription-factor binding sites measured by ChIA-PET [78] or from 1D maps of histone modifications and other epigenetic marks [79]. However, in spite of some evidence pointing to some non-trivial interplay between 1D sequence and 3D folding, the full picture remains poorly understood.

In this respect, some recent attempts (Section 3.2) based on "data-driven" polymer physics with input from epigenetic patterns seem to describe well the spatial structure of chromosomes *in vivo* and, in some specific cases, are able to identify critical hot-spots along the sequence associated to mutations in the phenotype. At the same time, the 3D chromosome conformation participates actively in the occurrence of epigenetic phenomena along the 1D sequence, such as the formation of loops between specific chromatin loci having distant locations along the sequence. Therefore, it appears plausible that the 3D chromosome organization is "echoed" in the positioning along the DNA sequence of 1D motifs associated to promoters and enhancers regulating gene expression [80], and that it is a major "driving force" in fixing and stabilizing the complex architectures [81,82] of gene regulatory networks.

Providing answers to these questions represents an exciting challenge which requires concerted experimental and theoretical efforts: the hope of the future is to find a systematic way for addressing unsolved biological and medical challenges linking DNA sequences anomalies, chromosome misfolding and aberrant phenotypic behavior. A combination of 1D and 3D genome information can improve the understanding of pathologies with a "structural" basis, such as the Hutchinson-Gilford progeria syndrome in which a protein associated to nuclear membrane scaffolding and DNA arrangement is mutated [83], or of pathologies such as cancer [10], characterized by significant expression *de*regulation due to epigenetic phenomena and in which specific 1D mutational events can be associated to DNA 3D structure [84].

# Transparency document

The Transparency document associated with this article can be found, in online version.

# Acknowledgments

AR and DR would like to acknowledge networking support by the ~~COST Action CA18127~~ COST ~~Action~~ActionCA18127. DR and AM would like to acknowledge support by the ~~HARMONY~~ HARMONY IMI-2 n. ~~116026.~~ 116026.

# References

**[1]** A.Y. Grosberg, How two meters of DNA fit into a cell nucleus: polymer models with topological constraints and experimental data, *Polym. Sci. Ser. C* **54**, 2012, 1–10.

**[2]** J.D. Halverson, J. Smrek, K. Kremer and A.Y. Grosberg, From a melt of rings to chromosome territories: the role of topological constraints in genome folding, *Rep. Prog. Phys.* **77**, 2014, 022601.

**[3]** A. Rosa and C. Zimmer, Computational models of large-scale genome architecture, *Int. Rev. Cell Mol. Biol.* **307**, 2014, 275–349.

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

**[5]** D. Jost, C. Vaillant and P. Meister, Coupling 1D modifications and 3D nuclear organization: data, models and function, *Curr. Opin. Cell Biol.* **44**, 2017, 20–27.

**[6]** D. Jost, A. Rosa, C. Vaillant and R. Everaers, A polymer physics view on universal and sequence-specific aspects of chromosome folding, In: C. Lavelle and J.-M. Victor, (Eds.), *Nuclear Architecture and Dynamics,* **vol. 2**, 2017, Academic Press, 149–169.

**[7]** A. Arneodo, C. Vaillant, B. Audit, F. Argoul, Y. d'Aubenton-Carafa and C. Thermes, Multi-scale coding of genomic information: from DNA sequence to genome structure and function, *Phys. Rep.* **498**, 2011, 45–188.

**[8]** R. Cortini, M. Barbi, B.R. Caré, C. Lavelle, A. Lesne, J. Mozziconacci, et al., The physics of epigenetics, *Rev. Mod. Phys.* **88**, 2016, 025002.

**[9]** T. Stadhouders, G.J. Filion and T. Graf, Transcription factors and 3D genome conformation in cell-fate decisions, *Nature* **569**, 2019, 345–354.

**[10]** J.-P. Mallm, M. Iskar, N. Ishaque, L.C. Klett, S.J. Kugler, J.M. Muino, et al., Linking aberrant chromatin features in chronic lymphocytic leukemia to transcription factor networks, *Mol. Syst. Biol.* **15**, 2019.

**[11]** H. Heyn, S. Moran and M. Esteller, Aberrant DNA methylation profiles in the premature aging disorders Hutchinson-Gilford Progeria and Werner syndrome, *Epigenetics* **8**, 2013, 28–33.

**[12]** K.N. Dahl, P. Scaffidi, M.F. Islam, A.G. Yodh, K.L. Wilson and T. Misteli, Distinct structural and mechanical properties of the nuclear lamina in Hutchinson–Gilford progeria syndrome, *Proc. Natl. Acad. Sci. U. S. A.* **103**, 2006, 10271–10276.

**[13]** G. Ozer, A. Luque and T. Schlick, The chromatin fiber: multiscale problems and approaches, *Curr. Opin. Struct. Biol.* **31**, 2015, 124–139.

**[14]** B. Alberts, et al., Molecular Biology of the Cell, $6^{th}$ ed., 2014, Garland Science; New York.

**[15]** D.J. Tremethick, Higher-order structures of chromatin: the elusive 30 nm fiber, *Cell* **128** (4), 2007, 651–654.

**[16]** Y. Nishino, M. Eltsov, Y. Joti, K. Ito, H. Takata, Y. Takahashi, et al., Human mitotic chromosomes consist predominantly of irregularly folded nucleosome fibres without a 30-nm chromatin structure, *Embo J.* **31**, 2012, 1644.

**[17]** K. Maeshima, R. Rogge, S. Tamura, Y. Joti, T. Hikima, H. Szerlong, et al., Nucleosomal arrays self-assemble into supramolecular globular structures lacking 30-nm fibers, *Embo J.* **35**, 2016, 1115–1132.

**[18]** K. Maeshima, S. Ide, K. Hibino and M. Sasai, Liquid-like behavior of chromatin, *Curr. Opin. Genet. Dev.* **37**, 2016, 36–45.

**[19]** H.D. Ou, S. Phan, T.J. Deerinck, A. Thor, M.H. Ellisman and C.C. O'Shea, ChromEMT: visualizing 3D chromatin structure and compaction in interphase and mitotic cells, *Science* **357**, 2017, eaag0025.

**[20]** J.D. Buenrostro, B. Wu, H.Y. Chang and W.J. Greenleaf, ATAC-seq: a method for assaying chromatin accessibility genome-wide, *Curr. Protoc. Mol. Biol.* **109**, 2015, 21.29.1.

**[21]** M. Zheng, S.Z. Tian, D. Capurso, M. Kim, R. Maurya, B. Lee, et al., Multiplex chromatin interactions with single-molecule precision, *Nature* **566**, 2019, 558–562.

**[22]** T. Cremer and C. Cremer, Chromosome territories, nuclear architecture and gene regulation in mammalian cells, *Nat. Rev. Genet.* **2**, 2001, 292–301.

**[23]** A. Bolzer, G. Kreth, I. Solovei, D. Koehler, K. Saracoglu, C. Fauth, et al., Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes, *Plos Biol.* **3**, 2005, e157.

**[24]** E. Lieberman-Aiden, N.L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, et al., Comprehensive mapping of long-range interactions reveal folding principles of the human genome, *Science* **326**, 2009, 289–293.

**[25]** J.R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, et al., Topological domains in mammalian genomes identified by analysis of chromatin interactions, *Nature* **485**, 2012, 376–380.

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

**[27]** T. Cremer and M. Cremer, Chromosome territories, *Cold Spring Harbor Perspect. Biol.* **2**, 2010, a003889.

**[28]** M.R. Branco and A. Pombo, Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations, *Plos Biol.* **4** (5), 2006, e138.

**[29]** J. Dekker, K. Rippe, M. Dekker and N. Kleckner, Capturing chromosome conformation, *Science* **295**, 2002, 1306.

**[30]** C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, et al., Long-range correlations in nucleotide sequences, *Nature* **356**, 1992, 168–170.

**[31]** C.-K. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H.E. Stanley and A.L. Goldberger, Mosaic organization of DNA nucleotides, *Phys. Rev. E* **49**, 1994, 1685–1689.

**[32]** A.P. Singh, S. Mishra and S. Jabin, Sequence based prediction of enhancer regions from DNA random walk, *Sci. Rep.* **8**, 2018.

**[33]** V. Afreixo, C.A.C. Bastos, J.M.O.S. Rodrigues and R.M. Silva, Identification of DNA CpG islands using inter-dinucleotide distances, In: *Optimization in the Natural Sciences,* 2015, Springer International Publishing, 162–172.

**[34]** P.M. Iannaccone and M. Khokha, Fractal Geometry in Biological Systems — An Analytical Approach, 1996, CRC Press.

**[35]** S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, M. Simons and H.E. Stanley, Generalized Lévy-walk model for DNA nucleotide sequences, *Phys. Rev. E* **47**, 1993, 4514–4523.

**[36]** H. Poincaré, On the three-body problem and the equations of dynamics, *Acta Math.* **13**, 1890.

**[37]** C. Bastos, V. Afreixo, A. Pinho, S.P. Garcia, J.M.O.S. Rodrigues and P.J.S.G. Ferreira, Inter-dinucleotide distances in the human genome: an analysis of the whole-genome and protein-coding distribution, *J. Integr. Bioinform.* **8**, 2011, 31–42.

**[38]** G. Paci, G. Cristadoro, B. Monti, M. Lenci, M. Degli Esposti, G.C. Castellani, et al., Characterization of DNA methylation as a function of biological complexity via dinucleotide inter-distances, *Philos. Trans. R. Soc. A* **374**, 2016.

**[39]** A. Merlotti, I. Faria do Valle, G. Castellani and D. Remondini, Statistical modelling of CG interdistance across multiple organisms, *BMC Bioinforma.* **19**, 2018, 355.

**[40]** V.S. Pande, A.Y. Grosberg and T. Tanaka, Heteropolymer freezing and design: towards physical models of protein folding, *Rev. Mod. Phys.* **72**, 2000, 259–314.

**[41]** M. Doi and S.F. Edwards, The Theory of Polymer Dynamics, 1986, Oxford University Press; New York.

**[42]** M. Rubinstein and R.H. Colby, Polymer Physics, 2003, Oxford University Press; New York.

**[43]** A. Grosberg, Y. Rabin, S. Havlin and A. Neer, Crumpled globule model of the three-dimensional structure of DNA, *EPL (Europhys. Lett.)* **23**, 1993, 373–378.

**[44]** A. Rosa and R. Everaers, Structure and dynamics of interphase chromosomes, *PLoS Comput. Biol.* **4**, 2008, e1000153.

**[45]** A. Rosa, N.B. Becker and R. Everaers, Looping probabilities in model interphase chromosomes, *Biophys. J.* **98**, 2010, 2410–2419.

**[46]** F. Benedetti, J. Dorier, Y. Burnier and A. Stasiak, Models that include supercoiling of topological domains reproduce several known features of interphase chromosomes, *Nucleic Acids Res.* **42**, 2013, 2848–2855.

**[47]** F. Benedetti, D. Racko, J. Dorier, Y. Burnier and A. Stasiak, Transcription-induced supercoiling explains formation of self-interacting chromatin domains in *S. pombe*, *Nucleic Acids Res.* **45**, 2017, 9850–9859.

**[48]** A.Y. Grosberg, S.K. Nechaev and E.I. Shakhnovich, The role of topological constraints in the kinetics of collapse of macromolecules, *J. Phys. Fr.* **49** (12), 1988, 2095–2100.

**[49]** C. Münkel, R. Eils, S. Dietzel, D. Zink, C. Mehring, G. Wedemann, et al., Compartmentalization of interphase chromosomes observed in simulation and experiment, *J. Mol. Biol.* **285**, 1999, 1053–1065.

**[50]** S. Jhunjhunwala, M.C. van Zelm, M.M. Peak, S. Cutchin, R. Riblet, J.J.M. van Dongen, et al., The 3D structure of the immunoglobulin heavy-chain locus: implications for long-range genomic interactions, *Cell* **133**, 2008, 265–279.

**[51]** A. Rosa and R. Everaers, Ring polymers in the melt state: the physics of crumpling, *Phys. Rev. Lett.* **112**, 2014, 118302.

**[52]** J. Smrek, K. Kremer and A. Rosa, Threading of unconcatenated ring polymers at high concentrations: double-folded vs time-equilibrated structures, *ACS Macro Lett.* **8**, 2019, 155–160.

**[53]** R.D. Schram, A. Rosa and R. Everaers, Local loop opening in untangled ring polymer melts: a detailed “Feynman test” of models for the large scale structure, *Soft Matter* **15**, 2019, 2418–2429.

**[54]** A.-M. Florescu, P. Therizols and A. Rosa, Large scale chromosome folding is stable against local changes in chromatin structure, *Plos Comput. Biol.* **12**, 2016, e1004987.

**[55]** M. Renda, I. Baglivo, B. Burgess-Beusse, S. Esposito, R. Fattorusso, G. Felsenfeld, et al., Critical DNA binding interactions of the insulator protein CTCF, *J. Biol. Chem.* **282**, 2007, 33336–33345.

**[56]** P.R. Cook, A model for all genomes: the role of transcription factories, *J. Mol. Biol.* **395**, 2010, 1–10.

**[57]** M. Barbieri, M. Chotalia, J. Fraser, L.-M. Lavitas, J. Dostie, A. Pombo, et al., Complexity of chromatin folding is captured by the strings and binders switch model, *Proc. Natl. Acad. Sci. U. S. A.* **109**, 2012, 16173–16178.

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

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

**[60]** Q. MacPherson, B. Beltran and A.J. Spakowitz, Bottom–up modeling of chromatin segregation due to epigenetic modifications, *Proc. Natl. Acad. Sci. U. S. A.* **115**, 2018, 12739–12744.

**[61]** T. Sexton, E. Yaffe, E. Kenigsberg, F. Bantignies, B. Leblanc, M. Hoichman, et al., Three-dimensional folding and functional organization principles of the *Drosophila* genome, *Cell* **148**, 2012, 458–472.

**[62]** S.K. Ghosh and D. Jost, How epigenome drives chromatin folding and dynamics, insights from efficient coarse-grained models of chromosomes, *PloS Comput. Biol.* **14**, 2018, e1006159.

**[63]** G. Shi, L. Liu, C. Hyeon and D. Thirumalai, Interphase human chromosome exhibits out of equilibrium glassy dynamics, *Nat. Commun.* **9**, 2018, 3161.

**[64]** N. Ganai, S. Sengupta and G.I. Menon, Chromosome positioning from activity-based segregation, *Nucleic Acids Res.* **42**, 2014, 4145–4159.

**[65]** A. Goloborodko, M.V. Imakaev, J.F. Marko and L. Mirny, Compaction and segregation of sister chromatids via active loop extrusion, *eLife* **5**, 2016, e14864.

**[66]** S. Ramaswamy, The mechanics and statistics of active matter, *Annu. Rev. Condens. Matter Phys.* **1**, 2010, 323–345.

**[67]** A.Y. Grosberg and J.-F. Joanny, Nonequilibrium statistical mechanics of mixtures of particles in contact with different thermostats, *Phys. Rev. E* **92**, 2015, 032118.

**[68]** J. Smrek and K. Kremer, Small activity differences drive phase separation in active-passive polymer mixtures, *Phys. Rev. Lett.* **118**, 2017, 098002.

**[69]** A. Goloborodko, J.F. Marko and L.A. Mirny, Chromosome compaction by active loop extrusion, *Biophys. J.* **110**, 2016, 2162–2168.

**[70]** G. Fudenberg, M. Imakaev, C. Lu, A. Goloborodko, N. Abdennur and L.A. Mirny, Formation of chromosomal domains by loop extrusion, *Cell Rep.* **15**, 2016, 2038–2049.

**[71]** D. Racko, F. Benedetti, D. Goundaroulis and A. Stasiak, Chromatin loop extrusion and chromatin unknotting, *Polymers* **10**, 2018, 1126.

**[72]** E. Orlandini, D. Marenduzzo and D. Michieletto, Synergy of topoisomerase and structural-maintenance-of-chromosomes proteins creates a universal pathway to simplify genome topology, *Proc. Natl. Acad. Sci. U. S. A.* **116**, 2019, 8149.

**[73]** M. Di Stefano, A. Rosa, V. Belcastro, D. di Bernardo and C. Micheletti, Colocalization of coregulated genes: a steered molecular dynamics study of human chromosome 19, *PloS Comput. Biol.* **9**, 2013, e1003019.

**[74]** L. Giorgetti, R. Galupa, E.P. Nora, T. Piolot, F. Lam, J. Dekker, et al., Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription, *Cell* **157**, 2014, 950–963.

**[75]** M. Di Pierro, B. Zhang, E. Lieberman-Aiden, P.G. Wolynes and J.N. Onuchic, Transferable model for chromosome architecture, *Proc. Natl. Acad. Sci. U. S. A.* **113** (43), 2016, 12168–12173.

**[76]** S. Bianco, D.G. Lupiáñez, A.M. Chiariello, C. Annunziatella, K. Kraft and R. Schöpflin, Polymer physics predicts the effects of structural variants on chromatin architecture, *Nat. Genet.* **50**, 2018, 662–667.

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

**[78]** P. Szałaj, Z. Tang, P. Michalski, M.J. Pietal, O.J. Luo, M. Sadowski, et al., An integrated 3-dimensional genome modeling engine for data-driven simulation of spatial genome organization, *Genome Res.* **26**, 2016, 1–13.

**[79]** Y. Zhu, Z. Chen, K. Zhang, M. Wang, D. Medovoy, J.W. Whitaker, et al., Constructing 3D interaction maps from 1D epigenomes, *Nat. Commun.* **7**, 2016, 10812.

**[80]** A. Pombo and N. Dillon, Three-dimensional genome architecture: players and mechanisms, *Nat. Rev. Mol. Cell Biol.* **16**, 2015, 245–257.

**[81]** H. de Jong, Modeling and simulation of genetic regulatory systems: a literature review, *J. Comput. Biol.* **9**, 2002, 67–103.

**[82]** M. Cosentino-Lagomarsino, B. Bassetti, G. Castellani and D. Remondini, Functional models for large-scale gene regulation networks: realism and fiction, *Mol. Biosyst.* **5**, 2009, 335–344.

**[83]** R. McCord, A. Nazario-Toole, H. Zhang, P. Chines, Y. Zhan, M.R. Erdos, et al., Correlated alterations in genome organization, histone methylation, and DNA–lamin A/C interactions in Hutchinson-Gilford progeria syndrome, *Genome Res.* **23**, 2013, 260–269.

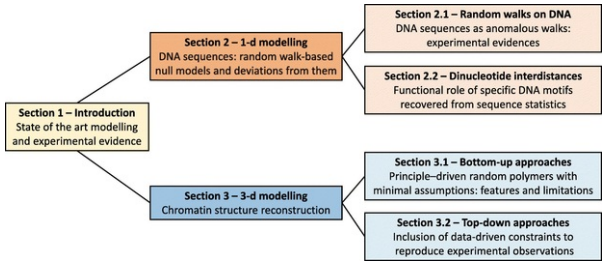
**[84]** G.I. Dellino, F. Palluzzi, A.M. Chiariello, R. Piccioni, S. Bianco, L. Furia, et al., Release of paused RNA polymerase II at specific loci favors DNA double-strand-break formation and promotes cancer translocations, *Nat. Genet.* **51**, 2019, 1011–1023.

# Transparency document

Multimedia Component 1

Transparency document

### Graphical Abstract




alt-text: Unlabelled Image