Polymer coil-globule phase transition is a universal folding principle of *Drosophila* epigenetic domains

Antony Lesage[a,*], **Vincent Dahirel**[b], **Jean-Marc Victor**[a], **and Maria Barbi**[a,*]

[a]Sorbonne Université, CNRS, Laboratoire de Physique Théorique de la Matière Condensée, LPTMC, F-75005 Paris, France. ; [b]Sorbonne Université, CNRS, Phenix, F-75005 Paris, France.

This manuscript was compiled on July 30, 2018

Localized functional domains within chromosomes, known as *topologically associating domains* or TADs, have been recently highlighted. In the case of *Drosophila*, TADs are biochemically defined by epigenetic marks, this suggesting that the 3D arrangement may be the "missing link" between epigenetic coloring and gene activity. Recent observations (Boettiger *et al.*, Nature 2016) on *Drosophila* fly Kc$_{167}$ cell provide access to structural features of these domains with unprecedented resolution thanks to super resolution experiments. In particular, they give access to the *distribution* of the radii of gyration for domains of different linear length and associated with three different transcriptional activity states: active, inactive or repressed. Intriguingly, the observed scaling laws lacked a consistent interpretation in polymer physics. Our methodology is conceived as to extract the best information from such super-resolution data, and to place these experimental results on a theoretical framework. We show that the experimental data are compatible with the behavior of a *finite-sized polymer*. The same generic polymer model leads to quantitative differences between active, inactive and repressed domains. Active domains behave as pure polymer coils, while inactive and repressed domains both lie at the coil-globule cross-over. For the first time, both the "color-specificity" of the persistence length and the mean interaction energy were estimated, leading to important differences between epigenetic states.

Epigenetic domains|polymer|Drosophila|coil-globule|phase transition

The 3D genome organization inside the cell nucleus is one of the most challenging questions of modern cell biology. Increasing evidence suggests that the complex and dynamical spatial arrangement of chromosomes is a keystone of gene regulation hence cell differentiation. Topologically associating domains (TADs) are one of the emerging features in this field. TADs are identified thanks to chromosome conformation capture techniques and may be defined as genomic regions whose DNA sequences physically interact with each other more frequently than with sequences outside the TAD (1). In *Drosophila*, these self-interacting genomic regions appear to be biochemically defined by epigenetic marks specific to various gene activity states (2). These states are called *colors* (3).

Obtaining a physical description of the spatial organization of chromatin inside epigenetic domains is then a crucial issue. Traditional optical imaging techniques, however, cannot be used for this purpose, since their resolution is limited by diffraction to a few hundred nanometers while the typical size of epigenetic domains is in the 0.1 to 1 μm range. This limitation has been overcome by the use of super-resolution imaging, as recently achieved notably by Zhuang's and Nollmann's groups. The former used STORM to image *Drosophila* epigenetic domains at the single-cell level and measured the radius of gyration of each individual snapshot for every imaged domain (4). The latter used SIM to image *Drosophila* epigenetic domains at the single-cell level too and revealed transient, color-specific modulated contacts between and within epigenetic domains (5).

Here we propose a theoretical framework enabling to reproduce and interpret the experimental distributions of gyration radii of Zhuang's group. As more data will follow this pioneering work, we attend here to define the best methodology to extract informations from series of images of equilibrium conformations of polymers. Our methodology includes two ingredients: a theoretical framework from polymer physics and a Bayesian-based parameter inference. Our working hypothesis is that polymer theory is actually relevant for describing the conformations of epigenetic domains in *Drosophila*, provided that a *finite-sized* interacting self-avoiding polymer model is adopted. Finite-size effects, which we think have been underappreciated in the analysis of chromatin structure so far, refer to deviations from the scaling behavior expected in the limit of infinitely long polymer chains. This includes in particular the existence of a size-dependent coil to globule transition temperature (6, 7), and a crossover regime for which the polymer is neither a pure globule, nor a pure coil. To analyze the data, our model is used with the following assumptions:

(i) the *same* general polymer model can describe all the observations whatever the epigenetic color;

(ii) different colors correspond to different sets of model parameters;

(iii) the ensemble of domains of a given color can be fitted with a unique set of parameters, whatever the size of the

Significance Statement

New exciting insight into *Drosophila* chromosome folding has been recently provided by super-resolution microscopy of epigenetic domains. No doubt new results will follow for other organisms. Such experiments provide statistical distributions of geometrical parameters of the domains. We provide here theoretical and data-processing tools to get most of the information out of these data. Our approach, based on polymer physics, allows us to get local parameters such as chromosome flexibility or linear compaction, from the distributions of global features. Such a multiscale analysis can be achieved because *Drosophila* chromosomal domains are found to be close to their coil-globule crossover region, where a particularly rich behavior is observed for finite size polymers. The biological relevance of this coil-globule crossover is discussed.

[*] To whom correspondence should be addressed. E-mail: lesage@lptmc.jussieu.fr, maria.barbi@sorbonne-universite.fr

domain, its genomic context, or other characteristics.

1. Theoretical Model

The standard polymer model at the thermodynamic limit. A chromosomal domain is a linear chain of units and can therefore be modeled as a *polymer*. A central quantity in polymer theory is the ensemble average root-mean-square (rms) value of the radius of gyration R_g, hereafter denoted $\overline{R}_g$, which measures the 3D extent of a chain (see **SI** for a precise definition). The polymer state is characterized by the *scaling behavior* of $\overline{R}_g$ with the polymer length (accounted for, equivalently, by the number of monomers N or by the number of base pairs L_{bp}):

$$\overline{R}_g \propto L_{\mathrm{bp}}^{\nu}$$

where the scaling exponent ν is the so-called *Flory exponent.*

For a polymer at equilibrium and in the thermodynamic limit ($N \to \infty$), two different folding modes have been predicted and measured (9). They depend on the relative strength of the monomer-monomer and solvent-monomer interactions with respect to temperature $\varepsilon / k_B T$. In good solvent (low $\varepsilon / k_B T$), the favorable interaction with the solvent leads to an effective repulsion between monomers. Hence, the polymer expands into a decondensed, disordered state called *coil*, described as a *self-avoiding walk* (SAW) characterized by $\nu = 3/5$. In poor solvent (high $\varepsilon / k_B T$), monomer-monomer attractions become predominant, and the polymer collapses into a state called *globule* with $\nu = 1/3$.

The phase transition between the two regimes is observed when the effective repulsion between monomers compensates their attraction. This happens nearly exactly when the second virial coefficient of a solution of monomers becomes zero (10). For given polymer and solvent, this condition is satisfied at a specific temperature called Θ (*theta*) temperature or Θ *point*. We note $\varepsilon_\theta = k_B \Theta$ the corresponding interaction energy. The Flory exponent at the Θ *point* is that of an ideal chain, $\nu = 1/2$.

Standard polymer scalings do not explain the scaling exponents inferred from super-resolution microscopy. We used the full ensemble of measurements of Boettiger *et al.* (4) to analyze (supplementary Fig. S1) the *mean* of the radii of gyration for *Drosophila* domains of different lengths and belonging to the three epigenetic states: (i) the active *red* types, covering the expressed regions, (ii) the inactive *black* states and (iii) the repressed *blue* domains, characterized by the presence of Polycomb group (PcG) proteins.

Power low fit exponents						
State:	Active		Inactive		Repressed	
Means	0.34	± 0.02	0.30	± 0.02	0.21	± 0.01
Medians	0.37	± 0.02	0.30	± 0.03	0.24	± 0.02

Table 1. Summary of exponents obtained from the fit of Boettiger's data (4) for active (A), inactive (I) and repressed (R) epigenetic domains through a power low fit of either the mean or median values of the radii of gyration for all different colors and lengths.

The inactive and repressed datasets display scaling exponents ν of 0.30 and 0.21, which are, surprisingly enough, both smaller than the expected value of the globular state $\nu = 1/3$, while the active dataset is well fitted with $\nu = 0.34$ (See Table 1). Plots displaying the power law fits for *mean* and *median* of radii of gyration are presented in the supplementary Fig. S1. Intriguingly, the apparent ν exponents of the median for the three colors are slightly larger ($\nu = 0.37$, 0.30, 0.24 for active, inactive and repressed, respectively, see Table 1). This is a strong indication of a crossover between two scaling regimes, whence the need of a finite-size scaling analysis.

***Finite-sized* corrections to scaling.** Self-attracting homopolymers undergo a coil-globule transition at a critical temperature $\Theta_N < \Theta$ (with a corresponding critical energy $\varepsilon_{\theta_N} > \varepsilon_\theta$) that depends on the polymer length N (6, 7). In order to describe the behavior of a finite-size polymer, we used a refined version of the semi-empirical *finite-size polymer theory* first introduced by one of us (11, 12). Thanks to the comparison with extensive lattice simulations, we were able to express the polymer free energy $\mathcal{F}(R_g^2|N,\varepsilon)$ as a function of its instant radius of gyration and to derive $\overline{R}_g$ as a function of N. Fig. 1**A** compares theoretical curves and simulation results, and shows a few significant simulation snapshots. The theoretical expression also remarkably fits the simulated *distributions* of gyration radii (supplementary Fig. S2). A few more details are given in **Materials and Methods** and **SI**.

The most striking feature of Fig. 1**A** is the non-monotonicity when $\varepsilon > \varepsilon_\theta$. In this regime, $\overline{R}_g(N)$ displays a characteristic knee-point around some value $N = f(\varepsilon/k_B T)$ (13) which defines the crossover region, giving *e.g.* close radii of gyration for the $N = 109$ (coil) and $N = 538$ (globule) blue snapshots. This behavior is quite unusual among critical phenomena and leads to dramatic finite-size effects. Remarkably, the same kind of behavior is observed in the case of a block copolymer, where block conformations are affected by finite-size effects likewise isolated polymers (14). Similar effects can thus be expected in the case of epigenetic domains embedded in larger chromosomal regions.

Mapping experimental data on the adimensional theoretical model. The aforementioned model relates dimensionless quantities, such as the number of monomers N and the gyration radius R_g in Kuhn length units. This model will be used to infer physical parameters from experiments giving the domain size in nanometer, for a known length of the domain in base pairs $\mathrm{L_{bp}}$. Such inference requires the introduction of two scales within the definition of the monomer (or Kuhn segment): the Kuhn length in nanometer, noted K_{nm}, and the Kuhn length in bp K_{bp}. K_{bp} relates the number of monomers N to $\mathrm{L_{bp}}$: $N = L/K = L_{\mathrm{bp}}/K_{\mathrm{bp}}$. K_{nm} yields a physical length scale to the size distribution predicted by the model. The correspondence between K_{nm} and K_{bp} is *a priori* not known. It depends on the local chromatin linear compaction in bp/nm c, as $K_{\mathrm{bp}} = c\,K_{\mathrm{nm}}$, and can thus vary in different domains. c is difficult to estimate, because the nucleosome fiber architecture is not directly observable. In the following, we will then consider K_{nm} and K_{bp} as two independent parameters of our model, in addition to ε, the interaction energy between Kuhn segments. We will, however, rely on the plausible hypothesis that c is homogeneous *within* one epigenetic domain and is the same for all domains of the same color.

We reformulated the free energy in order to use K_{bp} and K_{nm} explicitly as fitting parameters. It yields the following expression for the probability density of R_g^2:

$$\mathcal{P}_L(R_g^2 \big| \, \varepsilon, K_{\mathrm{nm}}, K_{\mathrm{bp}}) \propto \exp\left[-\beta\mathcal{F}\left(\frac{R_g^2}{K_{\mathrm{nm}}^2}\middle|\,\frac{L}{K_{\mathrm{bp}}},\varepsilon\right)\right]$$

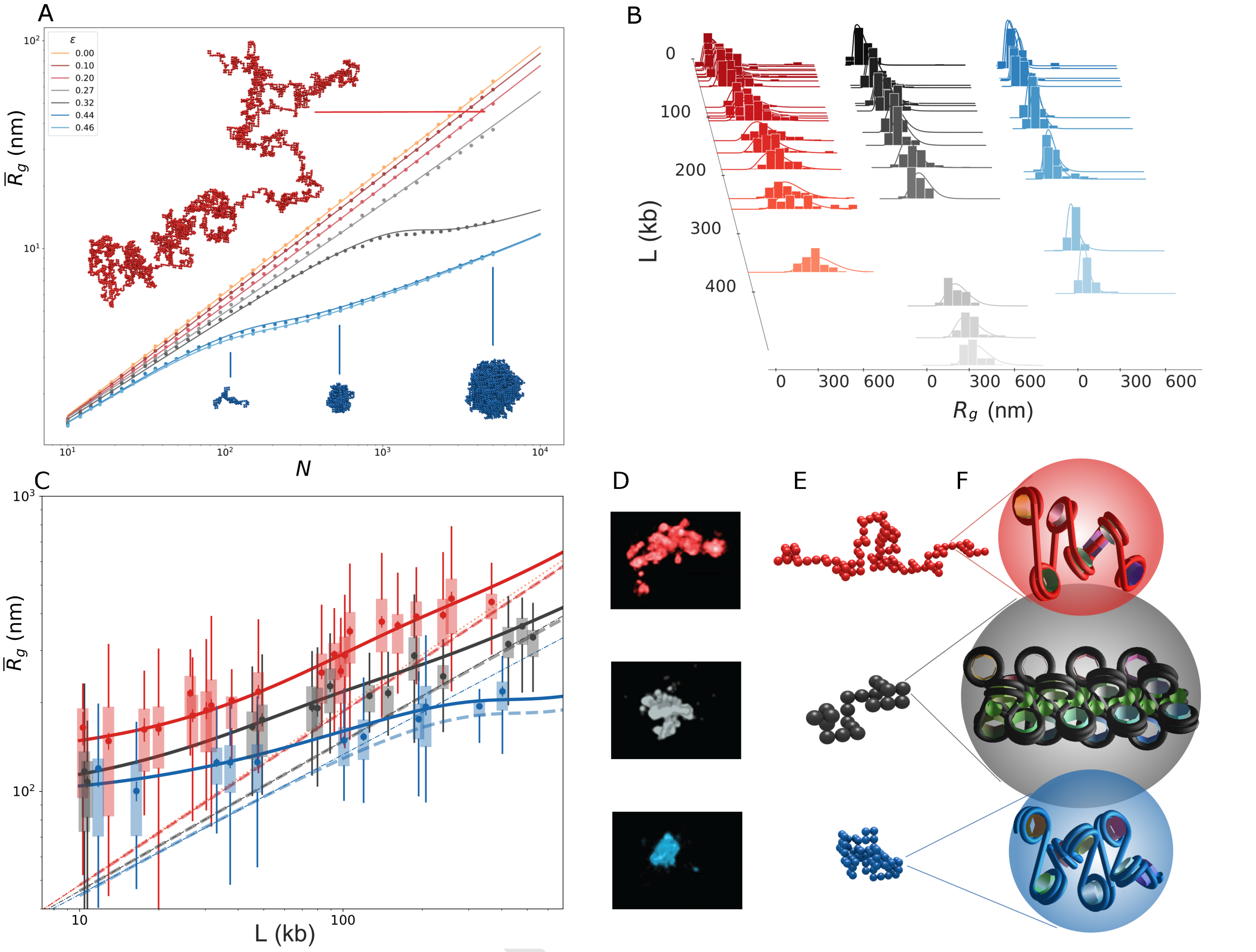

Fig. 1. A. Simulations and theory for a finite size polymer model. Log-log plot of rms radii of gyration $\overline{R}_g$ (in Kuhn length units) against the number of monomers N at different values of ε below and above $\varepsilon_\theta \simeq 0.27$. Lines are the analytical model, dots are obtained by on-lattice simulations (8). Snapshots correspond to $\varepsilon = 0.20\ k_BT$ and $N = 5012$ (red), $\varepsilon = 0.44\ k_BT$ and N =109, 538, 5012 (from left to right, blue). **B. Experimental data fit: distributions.** The three data ensembles from Ref. (4) (histograms) with the corresponding theoretical fitting distributions (lines). Colors refer to epigenetics: red for active, black for inactive and blue for repressed domains. The theoretical histograms have been calculated from the analytical expression of the probability density by using the fitting parameters of Table 2. A more detailed view of the complete set of histograms and fits is given in Fig.s S6, S7 and S8. **C. Experimental data fit: mean gyration radii.** Mean $\overline{R}_g$ as a function of the domain length L calculated from the analytical model with the parameter sets of Table 2: active (red line), inactive (black line), repressed (blue line). Boxplots (same colors) correspond to the experimental data from Ref. (4). Dashed lines are obtained from previous fitting curves by deconvolution, hence correspond to the behavior expected in an haploid system. The orange dotted lines represents the $\nu = 3/5$ typical scaling law. A corresponding fit for *median* values is given in supplementary Fig. S9; **D. Experimental images.** 3D-STORM images adapted from Boettiger *et al.* (4) corresponding to an active, inactive and repressed domain (from top to bottom; 106, 79 and 119 kb respectively). **E. Fitting model snapshots.** Simulation snapshots of the domains shown in (D) obtained with the corresponding fitted parameters K_{nm}, K_{bp} and ε; **F. Corresponding monomers at the fiber scale.** Two-angle models of the nucleosome fibers corresponding to the fitted parameters of the domains shown in (D) and simulated in (E). In the case of black domains, the green spheres are to evoke the presence of H1 histones.

where $\mathcal{F}$ is the system free energy (see Equations 1 to 2).

Correction for tetraploidy. The chromosomes of the tetraploid Drosophila Kc167 line are known to form bundles, sticking together in a regular fashion with a pairing rate of about 80% (15, 16). Up to now, super-resolution imaging techniques do not distinguish paired chromosomes, despite this is in principle possible in STORM experiments for small enough domains, and has been done by SIM (17). Therefore, we chose to describe domains as bundles of four chains.

We account for this effect by convolving the single-chromosome response, described by the theoretical model, with a bundle response. The apparent radii of gyration of the smallest domains are virtually equal to the radius of gyration σ of the bundle section. Hence the bundle response can be fitted on the smallest domains. To account for the boundary constraints on the epigenetic domains, we also let σ vary in a sigmoidal way from a minimum value a_0 to a maximum a_∞, reached within a characteristic length scale N_0 (see **Materials and Methods**).

2. Results

Finite-size polymer theory explains the observed scaling. Our fitting procedure is based on Bayesian inference methods. This allowed us to gain access to the probability distribution

	State	Active		Inactive	Repressed
		From Bayesian fit	From estimate	From Bayesian fit	From Bayesian fit
Fitted parameters:	ε [k_BT]	0.1 ± 0.05		0.32 ± 0.03	0.44 ± 0.04
	$K_{\rm bp}$ [bp]	$K_{\rm nm} \propto K_{\rm bp}^{0.56}$	$\sim$ 1100 $\div$ 1500	3900 ± 1300	1500 ± 550
	$K_{\rm nm}$ [nm]		$\sim$ 32 $\div$ 37	60 ± 9	37 ± 6
	a_0 [nm]	130 ± 7		93 ± 10	94 ± 4
	a_∞ [nm]	290 ± 15		170 ± 10	n/a
	n_0	630 ± 370		10 ± 6	n/a
Derived parameters:	$c = K_{\rm bp}/K_{\rm nm}$ [bp/nm]		$\sim$ 35 $\div$ 40	66 ± 24	40 ± 16
	c_{10} [nucl./10 nm]		$\sim$ 1.9 $\div$ 2.2	3.5 ± 1.5	2 ± 1
	C [nucl./$K_{\rm nm}$]		$\sim$ 6 $\div$ 8	20 ± 7	8 ± 3

Table 2. Summary of physical parameters obtained from the fit of Boettiger's data (4) for active (A), inactive (I) and repressed (R) epigenetic domains through the Bayesian procedure (mean values, see Fig.s S3 to S5). Errors are calculated from marginalized parameter distribution standard deviations. At the bottom, some derived geometrical parameters as the compaction in bp/nm, in nucleosomes per 10 nm, the number of nucleosomes per Kuhn segment C. Derived parameters are calculated by assuming a nucleosome repeat length of 182 bp for active domains, 192 bp for inactive and repressed domains (18) (The numerical results obtained with 182 or 192 bp are very close,in the margin of error). For active domains, we only report physical meaning parameters resulting from the direct fit: the right column estimates are obtained by including architectural features, see Discussion.

in the parameter space: we have then been able to obtain their best values and confidence intervals (and to check for correlations between them; see **Materials and Methods**). Datasets for each of the three epigenetic colors are analyzed as a whole. We performed Markov Chain Monte Carlo (MCMC) Bayesian inference to maximize the log-likelihood of the observation data given the parameters of our theoretical finite-sized self-avoiding polymer ($\varepsilon, K_{\rm bp}, K_{\rm nm}$), corrected for tetraploidy with the bundle parameters (a_0, a_∞, N_0). Probability distributions for the six parameters and for the case of active, inactive and repressed domains are displayed in supplementary Fig.s S3, S4 and S5, respectively. Note the absence of correlation between the bundle parameters and the energy parameter ε in these distributions.

Once obtained the marginal distributions, we define the optimal value for each parameter as its mean value. Confidence intervals have been deduced as standard deviations. Table 2 summarizes the results obtained for the three epigenetic states, together with the resulting linear compaction in different units. In Fig. 1**B**, all the fitted histograms are plotted along with the theoretical curves obtained with the optimal parameters. Separate histograms are given in Fig.s S6, S7 and S8 for the three colors respectively. The comparison shows a remarkably good agreement between the distribution of data and the predicted behavior.

As an *a posteriori* check of the results, we calculated the mean radius of gyration $\overline{R}_g$ as a function of the domain length L from the analytical model, for each color, and compare it to the experimental averages in Fig. 1**C**. A secondary estimate of the goodness of fit is obtained by performing a Pearson's chi-squared test on $\overline{R}_g$. We calculated the reduced $\tilde{\chi}^2$ for active, inactive and repressed domains respectively and find values lower than 0.3, close to what obtained for the power law fits. Interestingly, we could eventually get rid of the bundle on the fitting curves by applying a deconvolution procedure, thus predicting what would be observed with a haploid genome. The resulting curves are shown in Fig. 1**C** as dashed lines. The fitted parameters of these curves are given in Table 2.

Subdomains. Boettiger *et al.* observe a plateau in the plot of $\overline{R}_g$ as a function of the genomic size for the two largest subdomains of the largest repressed domains. They describe this behavior as intermixing. This is also characteristic of globular conformations of polymers (19). In all other cases (all active domains, all inactive domains and all other repressed domains), the plots of $\overline{R}_g$ as a function of the subdomain length are the same as the plots of $\overline{R}_g$ as a function of the domain length (see **SI**). These observations strongly support the existence of a coil-globule transition: only the largest repressed domains are globular enough to exhibit a plateau in the subdomains plot. In all other cases, the conformations are coils either because the domains are above the Θ point (active domains) or because they are too small to be globular (all black domains and all other blue domains).

Positioning the three colors with respect to the coil-globule crossover. Fig. 1**C** clearly shows that active (red) domains have an exponent very close to 3/5 and stay thus very close to the coil regime for all the observed lengths, in agreement with the fitted $\varepsilon = 0.10\ k_BT$ parameter, well below the theoretical transition value of $\varepsilon_\theta \simeq 0.27\ k_BT$. On the contrary, the repressed (blue) domains are well above this limit, with $\varepsilon = 0.43\ k_BT$. In Fig. 1**C**, a plateau is indeed visible around lengths of ~400 kb, with a net deviation from the coil behavior starting from the smallest observed domains.

Somehow expectedly, inactive (black) domains display an intermediate regime: with $\varepsilon = 0.32\ k_BT$, they are already above the coil-globule transition but finite-size effects remain strong at the observed lengths. Hence, the crossover plateau is still unreached at these lengths, but a net discrepancy with respect to the short-range coil behavior is observed.

In any case, in the left part of the curves of this figure, all domains are closer to coil conformations due to finite-size effects, that emerge then as a crucial feature in the interpretation of domain super-resolution imaging. A fit of the deconvolved curves slopes in the small domain region ($<$ 60 kb) gives $\nu = 0.51$ and 0.47 for inactive and repressed domains, respectively (we obtain $\nu = 59$ for the active deconvolved fit within the same range).

Getting structural parameters of chromatin. In the case of the inactive and repressed domains, both values of $K_{\rm nm}$ and $K_{\rm bp}$ have been obtained simultaneously by our approach. We obtain for repressed domains $K_{\rm nm} \sim 35$ nm and $K_{\rm bp} \sim 1500$ bp (Table 2). The corresponding compaction $c \sim 40$ corresponds to a compaction $c_{10} \sim 2$ nucl./10 nm. Inactive domains give instead $K_{\rm nm} \simeq 60$ nm and $K_{\rm bp} \simeq 4000$ bp, with a corresponding $c_{10} = 3.5$ nucl./10 nm, hence a nucleosome fiber almost twice as stiff and twice as compact as for repressed domains.

The possibility to determine these structural parameters is a remarkable consequence of the coil-globule crossover. This comes from the existence of different asymptotic scaling laws when N increases, before, during and after the crossover.

The mean linear compaction of the nucleosome fiber (parameter c) is in principle determined by the underlying architecture of the nucleosome fiber, which in turn depends on a few local parameters, namely the nucleosome repeat length (NRL) and the degree of DNA wrapping around the nucleosome. A simple estimation of the elastic properties and of the compaction of this assembly can be obtained by the two-angle model (20) (see SI Fig. S10). The mechanical and structural features estimated here for repressed chromatin features fit easily with what is analytically obtained in the framework of the two-angle model with standard NRL (192 bp) and wrapping angle (negatively crossed nucleosomes).

Inactive chromatin can instead be obtained with an abnormally short NRL only, whatever the wrapping. This may suggest a possible role for the H1 histone, whose presence is characteristic of inactive chromatin (3). By cross-linking the entering and exiting DNAs of each nucleosome, H1 may indeed result in an effective shortening of linker DNAs (21), hence explain the stiffening and compaction of the nucleosome fiber.

At variance with inactive and repressed domains, active domains are in the scale invariant regime ($\varepsilon = 0.1\ k_BT$) where $K_{\rm nm}$ and $K_{\rm bp}$ cannot be computed independently but satisfy instead such a relation as $K_{\rm nm} = \kappa K_{\rm bp}^{\nu}$ with κ some constant and ν the Flory exponent. Hence, one would expect the log-likelihood in the $(K_{\rm nm}, K_{\rm bp})$ plane to be nearly constant along the curve of equation $K_{\rm nm} = \kappa K_{\rm bp}^{\nu}$. We found indeed a power law fit of the marginalized $(K_{\rm nm}, K_{\rm bp})$ distribution of the form $K_{\rm nm} = 0.62\ K_{\rm bp}^{0.56}$ with an exponent very close to the expected Flory exponent. And the log-likelihood in the $(K_{\rm nm}, K_{\rm bp})$ plane is indeed nearly constant along this power-law curve. As a consequence, both averages $K_{\rm nm}$ and $K_{\rm bp}$ obtained from the marginalized distributions (Fig. S3) for active domains are ill-defined. And they are indeed unrealistically small.

In order to identify reasonable ranges for both $K_{\rm nm}$ and $K_{\rm bp}$ in active domains, we combined the observed power law with a second relationship arising from the underlying structure of the nucleosome fiber by means of the two-angle model. To this aim, we calculated the geometry-based $K_{\rm nm}^{\rm geom}$ as a function of c for any given NRL and wrapping angles from the analytical model. We then deduced $K_{\rm nm}(c)$ by replacing $K_{\rm bp} = cK_{\rm nm}$ in $K_{\rm nm} = \kappa K_{\rm bp}^{\nu}$. We thus found an intercept between $K_{\rm nm}^{\rm geom}(c)$ and $K_{\rm nm}(c)$ for relatively open wrapping angles, typical of *open* nucleosomes, and the expected NRL of 182 bp (18). The intercept gives $K_{\rm nm} \sim 35$ nm and $K_{\rm bp} \sim 1300$ bp (Table 2). The corresponding compaction $c \sim 35$ corresponds to a compaction $c_{10} \sim 2$ nucl./10 nm. Interestingly and rather surprisingly, we found by this procedure that the geometrical parameters of active domains are essentially indistinguishable from what previously derived for repressed domains (see supplementary Fig. S10). If confirmed, this finding seems to indicate that active and repressed domains are in fact very close from a structural point of view, and differ essentially only with respect to the interaction energy ε.

To sum up these findings, Figures 1**D** and **E** compare typical STORM images (4) with corresponding simulation snapshots obtained with our model and the corresponding parameters, *i.e.* by using the parameters of Table 2 and a number N of monomers corresponding to the length of the images domain. Figure **F** reproduces the corresponding monomer stretch as obtained with the two-angle model, showing at a glance its physical size and linear density. The precise nucleosome orientation is of course only indicative, since it depends finely on the precise architectural parameters and is expected to display fluctuations *in vivo*.

bundle geometry. As shown in Table 2, we obtain minimal bundle section extents of the order of 100 nm for the three colors (with a slightly larger value for active domains) compatibly with the similar radii of gyration observed for small domains (Fig. 1**C**).

Interestingly, the variation of the bundle section as a function of domain lengths significantly differs for different epigenetic colors. Active domains appear to allow for the largest bundle section spreading, up to $a_\infty \sim 300nm$ provided that the polymer is long enough (N_0 being of the order of 500 monomers, i.e. approximately 15000 nm or 600 kb). At variance, only minor variations in the bundle section extent are obtained for inactive domains, and in the case of repressed domain this effect is so strong that we could indifferently fit the data with a simplified model with a unique $\sigma = a_0$. These findings corroborate the scenario of rather decondensed active domains, for which the looseness of chromosome pairing may result in an faster spreading with the domain length, while this effect is strongly reduced by the strongest packing of inactive and repressed domain, due to their globular configuration.

3. Discussion

The values of the Kuhn lengths $K_{\rm bp}$ (in base pairs) that come out from our analysis are in close agreement with recent estimations from Hi-C data (Giacomo Cavalli, personal communication). We also provide here values of the Kuhn lengths in nanometers $K_{\rm nm}$ which have never been measured in vivo so far. Their relatively small values, as compared with naked DNA in particular, confirm the most recent dynamic measurements of the high flexibility of chromatin in vivo (22).

In previous studies, notably in simulations of 3D genome organization, it has generally been assumed that the size of the monomer ($K_{\rm bp}$ or $K_{\rm nm}$) does not depend on the epigenetic state. We find here that active (red) and repressed (blue) domains have indeed, though surprisingly, the same monomer size ($K_{\rm nm} \sim 35$ nm and $K_{\rm bp} \sim 1500$ bp), whereas inactive (black) chromatin has a monomer size ($K_{\rm bp}$ or $K_{\rm nm}$) about twice as large.

As blue chromatin domains are dispersed among the volume of the so-called active *compartment* (5), the nucleosome fiber structural similarity of active and repressed *domains* may facilitate transitions between active and repressed epigenetic states in the course of cell differentiation.

In addition to measuring their size, we also made the first color-specific determination of the interaction energy ε between chromatin Kuhn segments *in vivo*. Other estimations of interaction parameters have been deduced, in particular from the fit of Hi-C data (23, 24). In a recent study, Falk, Mirny and coworkers have determined the value of the interaction energy parameters in a copolymer model (A and B chromatin compartments) (25). In order to recover the experimental phase separation between chromatin A and B, they found an interaction between B monomers of 0.55 k_BT and a much weaker interaction between A monomers. This is compatible with our results, assimilating the A compartment with active chromatin, and the B compartment with repressed ones.

It is tempting to try to relate the different values of ε obtained for the three epigenetic states to different molecular interaction mechanisms. Caution is needed, since ε is an effective parameter accounting for the overall, mean interaction energy between two Kuhn segments. Simulations of nucleosome fibers with a fine-graining of 10 bp for DNA indicate that, on average, one should expect only one nucleosome-nucleosome contact *in trans* per Kuhn segment (Pascal Carrivain, personal communication). Assuming this, ε appears as a reasonable estimate for single *in trans* interaction, so that a direct comparison between the fitted values becomes possible. In the case of repressed domains, such interaction is known to be mediated by Polycomb proteins which are considered to stabilize condensed chromatin configurations by means of bridges, and we find, coherently, the largest interaction energy $\varepsilon \simeq 0.4\ k_BT$. No condensing protein is known, instead, for active and inactive domains. So, what are the interactions responsible for the coil-globule transition of these domains?

We recall that polymers are at their coil-globule transition when the second virial coefficient of a solution of their monomer becomes zero (10). Now, in Ref. (26, 27), Livolant and coworkers experimentally characterized the interaction between isolated nucleosome core particles at different monovalent salt concentrations. Interestingly, the second virial coefficient steeply decreases to zero and presents a cusp in the salt range 75–210 mM, i.e. around physiological concentrations. Hence, physiological conditions seem to have been selected so that repulsion and attraction between monomers counterbalance.

It is therefore tempting to attribute the coil-globule transition of chromosomes to the vanishing of the second virial coefficient of nucleosome-nucleosome interaction. This is also in line with quite recent measurements of chromosome dynamics in yeast, which has been modeled as a Rouse dynamics slowed down by nucleosome-nucleosome transient interactions (22). Inactive (black) chromatin is very close to the Θ point, indicating that nucleosome-nucleosome interactions might be preponderant within inactive domains. For the active (red) chromatin, we speculate that the lower interaction is linked to a lower interaction between nucleosomes, which is consistent with acetylation of histone tails in transcribing chromatin (28), thus reducing their charge, hence their ability to bridge other nucleosomes (29). For repressed (blue) chromatin, a larger value of ε points toward a stronger interaction, certainly mediated by proteins from the Polycomb family, in agreement with Ph-knockdown experiments of Ref. (4). The detailed modeling of the mechanistic effects involved remains elusive and clearly points to the need for molecular modeling of the Polycomb gene silencing complexes. Interestingly, active and repressed domains have very similar structural parameters, this suggesting for polycomb an action *in trans* rather than *in cis*.

4. Conclusions

Our analysis of the distribution of radii of gyrations of 48 epigenetic domains of Drosophila have enabled to estimate many previously unavailable physical parameters describing chromatin structure and intra-chromatin interactions. In particular, we could get:

First, color-specific measures of the Kuhn length (in base pairs *and* in nanometers) of active, inactive and repressed domains respectively. Strikingly, these measures are similar to Hi-C data in mammals (30) as well as to most recent dynamic measurements in yeast (22). The knowledge of both Kuhn lengths leads to the value of the compaction of the chromatin, *i.e.* the number of nucleosomes per 10 nm. This is a precious indication of the *conformational* state of the nucleosome fiber.

Second, we get the first measure of the interaction energy ε between Kuhn segments. It is very striking that, in all but two cases studied here (95%), the length of epigenetic domains remain small enough so that the domains are still in the coil region of the phase diagram. This suggests that one essential role of the coil-globule transition is to create *dense coils* which at the same time allow to "tidy up" a whole genome in the reduced volume of a cell nucleus while giving access in a reversible way to the transcription machinery. Importantly the high density of chromatin inside cell nuclei is not imposed by nuclear membrane confinement but by transient interactions.

It is often stressed that nucleosomes enable to reduce the length of a chromosome by a factor of ten. A new role for the nucleosome emerges from our study. We find here that the specific value of the interaction energy between nucleosomes may allow by itself the existence of a coil-globule transition in the neighborhood of typical physiological conditions, in particular for inactive domains where no known protein-mediated interactions are reported. The key role of histone tail flexibility on chromatin compaction has been shown by computational studies (29). Super-resolution microscopy (4, 5, 31) combined with the methodology presented in this paper now allows to design new experiments to investigate the effect of histone modifications, or histone variants, on the 3D organization of chromatin sub-compartments.

Materials and Methods

Theoretical model. A recall of the main lines of standard polymer theory is given in SI. The case of finite-size polymer of N identical monomers is described by a theoretical model following the main idea of Ref. (11): attractive interactions are directly added to a SAW model. We developed a revised version of such a model by a careful comparison with on-lattice simulations (8) and finally expressed the system free energy as

$$\begin{aligned}\beta\mathcal{F}_N(t|\varepsilon) \quad &= \quad a_1(\varepsilon)Nt + a_2(\varepsilon)Nt^2 + a_3(\varepsilon)(Nt)^{-2/3} \\ &\quad +a_4(\varepsilon)(Nt^2)^{2/3} + 1.13\ln Nt\,, \end{aligned} \qquad [1]$$

(see **SI** for further details). In the case of chromosome domains, the number of monomers is unknown and the accessible physical parameter is the polymer length in base-pairs, L. The two parameters K_{nm}, K_{bp} play the role of rescaling parameters to map the dimensional model on the adimensional case. This results in the following expression of probability density for the squared radii of gyration in nm, as measured in experiments:

$$\mathcal{P}_L\left(R_g^2\middle|\,\varepsilon, K_{\mathrm{nm}}, K_{\mathrm{bp}}\right) = \frac{1}{K_{\mathrm{nm}}^2}p_{L/K_{\mathrm{bp}}}\left(R_g^2/K_{\mathrm{nm}}^2\middle|\,\varepsilon\right)\,. \qquad [2]$$

Bundle correction. Domains are described as a bundle of four Kuhn chains of N segments. The resulting radius of gyration, with $n = 4$, reads

$$R^2 = \frac{1}{n}\sum_{k=1}^{n} R_k^2 + \frac{1}{n}\sum_{k=1}^{n}(\mathbf{G}_k - \mathbf{G})^2\,, \qquad [3]$$

where $\mathbf{G}_k$ is the center of mass of the k-th polymer of the bundle and R_k^2 its radius of gyration, still described by the previous theory $\mathcal{P}_N(R_k^2|\varepsilon)$. The second sum in Equation 3 is the bundle contribution to the total radius of gyration, $B^2 = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{G}_k - \mathbf{G})^2$. We inferred B^2 from the experimental distribution obtained for the smallest epigenetic domains, which displays a (shifted) exponential form $f_\lambda(r^2) = \lambda\, \exp\left(-\lambda(r^2-\mu^2)\right)$ for $r \geq \mu$, $= 0$ otherwise. This is compatible with a random arrangement of the four polymers, with a dispersion $1/\lambda = \sigma^2$ characteristic of the bundle section spreading and where μ^2 accounts for the steric hindrance of each single monomer. Overall, we thus write the R^2 distribution as the convolution $\mathcal{P}_{R^2} = \mathcal{P}_N * f_\lambda$. The bundle section spreading σ^2 depends in general on the polymer length N. We chose σ varying from a minimum value a_0 to a maximum value a_∞, reached within a characteristic length scale N_0:

$$\sigma_N = \frac{a_\infty}{1+(\frac{a_\infty}{a_0}-1)e^{-\frac{N}{N_0}}}. \qquad [4]$$

See **SI** for further details.

Dataset and statistical analysis. Boettinger and co-workers provided us with the ensemble of their radius of gyration measurements. Authors identified candidate domains of a specific length L by applying a moving average filter with a window of same size L on the marker enrichment trace for the marker of the desired epigenetic state. The whole dataset consist in three sets of data for the three different epigenetic colors: active (red), inactive (black) and repressed (blue). These three data sets contain 23, 14 and 11 domains of different lengths, respectively. For each of the 48 domains, the radius of gyration is measured over a set of 20-100 cells. Hence we disposed, for each color and each length L, of a set of n measurements, allowing us to plot an histogram or *distribution*. For each color, we considered the dataset corresponding to the ensemble of measurements as a whole, hence assuming that a unique set of parameters is needed.

For each given color, we denote the set of n different domain lengths explored as $\{L_\ell\}$ with $\ell = 1 \ldots n$. For each length L_ℓ, we denote the set of n_ℓ different measurements of the radius of gyration, corresponding to different cells, as $\{R_g\} = \left\{{R_g}_i^\ell\right\}$ with $i = 1 \ldots n_\ell$ and $\ell = 1 \ldots n$.

We then detected outliers following the procedure described in Ref. (32), which is adapted to the case where the data distribution is skewed. An adjusted outlyingness (AO) is defined by using data skewness as in Equation 3 of Ref. (32). Altogether, 159 points have been eliminated over a total of 2326, i.e. a percentage of 7%.

We maximized the *total* log-likelihood $\mathscr{L} = \sum_\ell^n \mathscr{L}_\ell$, where $\mathscr{L}_\ell$ is defined at *given* length as the logarithm of the product of the probabilities of the dataset $\left\{{R_g}_i^\ell\right\}$ given the set of parameters $\theta = (K_{\rm nm}, K_{\rm bp}, \varepsilon, a_0, a_\infty, N_0)$:

$$\mathscr{L}_\ell = \ln \prod_i^{n_\ell} P\left(\left\{{R_g}_i^\ell\right\} \mid \theta\right) . \qquad [5]$$

We maximized the total log-likelihood by using Bayesian inference to sample the probability distribution of the model parameters. We used a uniform *prior* probability distribution (*naive* Bayesian inference), by only fixing a few constraints on the fitting parameters, namely their positiveness. We then performed Markov Chain Monte Carlo algorithms (MCMC) inference using the Python *emcee* (33).

Once obtained the marginal distributions of all the fitting parameters (see Fig.s S3, S4 and S5) we identified the optimal value for each parameter with its mean value over the distribution. Confidence intervals have been deduced in a similar way by evaluating the standard deviations.

ACKNOWLEDGMENTS. We are very grateful to Alistar Boettiger, Xiaowei Zhuang and co-workers for kindly providing us with their experimental data and for interesting discussions. We thank Giacomo Cavalli and Marcelo Nollmann for helping us to understand some aspects of the experimental approach.

1. Sexton T, et al. (2012) Three-dimensional folding and functional organization principles of the <em>drosophila</em> genome. *Cell* 148(3):458–472.
2. Haddad N, Jost D, Vaillant C (2017) Perspectives: using polymer modeling to understand the formation and function of nuclear compartments. *Chromosome Research* 25(1):35–50.
3. Filion GJ, et al. (2010) Systematic protein location mapping reveals five principal chromatin types in drosophila cells. *Cell* 143(2):212–224.
4. Boettiger AN, et al. (2016) Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. *Nature* 529(7586):418–422.
5. Cattoni Diego I., et al. (2017) Single-cell absolute contact probability detection reveals chromosomes are organized by multiple low-frequency yet specific interactions. *Nature Communications* 8(1):1753.
6. Grassberger P, Hegger R (1995) Simulations of three-dimensional θ polymers. *The Journal of Chemical Physics* 102(17):6881–6899.
7. Caré BR, Carrivain P, Victor TFJM, Lesne A (2014) Finite-size conformational transitions: A unifying concept underlying chromosome dynamics. *Communications in Theoretical Physics* 62(4):607.
8. Lesage A, Dahirel V, Barbi M, Victor JM (in prep.) Finite-size polymer simulations and theory.
9. Nishio I, Sun ST, Swislow G, Tanaka T (1979) First observation of the coil-globule transition in a single polymer chain. *Nature* 281(5728):208–209.
10. Grosberg AI, Khokhlov AR (1994) *Statistical physics of macromolecules.* (New York : American Institute of Physics). Includes bibliographical references (p. 345-346) and index.
11. Victor JM, Lhuillier D (1990) The gyration radius distribution of two-dimensional polymer chains in a good solvent. *The Journal of Chemical Physics* 92(2):1362–1364.
12. Victor J, Imbert J, Lhuillier D (1994) The number of contacts in a self-avoiding walk of variable radius of gyration in two and three dimensions. *The Journal of Chemical Physics* 100(7):5372–5377.
13. F. R, K. B, W. P (2006) The phase diagram of a single polymer chain: New insights from a new simulation method. *Journal of Polymer Science Part B: Polymer Physics* 44(18):2542–2555.
14. Caré BR, Emeriau PE, Cortini R, Victor JM (2015) Chromatin epigenomic domain folding: size matters. *AIMS Biophysics* 2(201504517):517.
15. Williams BR, Bateman JR, Novikov ND, Wu CT (2007) Disruption of topoisomerase ii perturbs pairing in drosophila cell culture. *Genetics* 177(1):31–46.
16. Senaratne TN, Joyce EF, Nguyen SC, C-Ting W (2016) Investigating the interplay between sister chromatid cohesion and homolog pairing in drosophila nuclei. *PLoS Genet* 12(8):e1006169.
17. Szabo Q, et al. (2018) Tads are 3d structural units of higher-order chromosome organization in drosophila. *Science Advances* 4(2).
18. Scacchetti A, et al. (2018) Chrac/acf contribute to the repressive ground state of chromatin. *Life Science Alliance* 1(1).
19. de Gennes PG (1979) *Scaling concepts in polymer physics.* (Cornel University Press).
20. Ben-Haïm E, Lesne A, Victor JM (2001) Chromatin: A tunable spring at work inside chromosomes. *Phys. Rev. E* 64(5):051921.
21. Schiessel H (2002) How short-ranged electrostatics controls the chromatin structure on much larger scales. *EPL (Europhysics Letters)* 58(1):140.
22. Socol M, et al. (2017) In vivo, chromatin is a fluctuating polymer chain at equilibrium constrained by internal friction. *bioRxiv*.
23. Haddad N, Vaillant C, Jost D (2017) Ic-finder: inferring robustly the hierarchical organization of chromatin folding. *Nucleic Acids Research* 45(10):e81.
24. Ghosh SK, Jost D (2018) How epigenome drives chromatin folding and dynamics, insights from efficient coarse-grained models of chromosomes. *PLOS Computational Biology* 14(5):1–26.
25. Falk M, et al. (2018) Heterochromatin drives organization of conventional and inverted nuclei. *bioRxiv*.
26. Mangenot S, Raspaud E, Tribet C, Belloni L, Livolant F (2002) Interactions between isolated nucleosome core particles: A tail-bridging effect? 7:221–231.
27. Mangenot S, Leforestier A, Vachette P, Durand D, Livolant F (2002) Salt-induced conformation and interaction changes of nucleosome core particles. 82:345–56.
28. Lavelle C (2007) Transcription elongation through a chromatin template. *Biochimie* 89(4):516 – 527. DNA Topology.
29. Collepardo-Guevara R, et al. (2015) Chromatin unfolding by epigenetic modifications explained by dramatic impairment of internucleosome interactions: A multiscale computational study. *Journal of the American Chemical Society* 137(32):10205–10215. PMID: 26192632.
30. Sanborn AL, et al. (2015) Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences* 112(47):E6456–E6465.
31. Xu J, et al. (2018) Super-resolution imaging of higher-order chromatin structures at different epigenomic states in single mammalian cells. *Cell Reports* 24(4):873 – 882.
32. Hubert M, Van der Veeken S (2008) Outlier detection for skewed data. *Journal of Chemometrics* 22(3-4):235–246.
33. Foreman-Mackey D, Hogg DW, Lang D, Goodman J (2013) emcee: The mcmc hammer. *PASP* 125:306–312.

***.** Supporting Information (SI)

Supplementary text and figures are available at https://www.overleaf.com/17424831dsychmjdxxvb#/66238769/

Classification : Physical sciences; Biophysics and Computational Biology