A First-principles Approach to Large-scale Nuclear Architecture

Ankit Agrawal
Computational Biology Group
The Institute of Mathematical Sciences,
C.I.T. Campus, Taramani, Chennai 600 013, India
and
Homi Bhabha National Institute,
Training School Complex, Anushaktinagar,
Mumbai 400 094, India

Nirmalendu Ganai
Department of Physics, Nabadwip Vidyasagar College,
Nabadwip, Nadia 741 302, India
and
Theoretical Soft Matter and Biophysics,
Institute of Complex Systems,
Forschungszentrum Jülich,
D-52425 Jülich, Germany

Surajit Sengupta
TIFR Centre for Interdisciplinary Sciences,
36/P Gopanpally Village,
Serilingampally Mandal, Ranga Reddy District, Hyderabad 500 107, India

Gautam I. Menon*
Computational Biology Group
The Institute of Mathematical Sciences,
C.I.T. Campus, Taramani, Chennai 600 013, India
and
Homi Bhabha National Institute,
Training School Complex, Anushaktinagar,
Mumbai 400 094, India

Abstract

Model approaches to nuclear architecture have traditionally ignored the biophysical consequences of ATP-fueled active processes acting on chromatin. However, transcription-coupled activity is a source of stochastic forces that are substantially larger than the Brownian forces present at physiological temperatures. Here, we describe a first-principles approach to large-scale nuclear architecture in metazoans that incorporates cell-type-specific active processes. The model predicts the statistics of positional distributions, shapes and overlaps of each chromosome. Our simulations reproduce common organising principles underlying large-scale nuclear architecture across human cell nuclei in interphase. These include the differential positioning of euchromatin and heterochromatin, the territorial organisation of chromosomes including both gene-density-based and size-based chromosome radial positioning schemes, the non-random locations of chromosome territories and the shape statistics of individual chromosomes. We propose that the biophysical consequences of the distribution of transcriptional activity across chromosomes should be central to any chromosome positioning code.

Keywords: Chromatin, human genome, nuclear architecture, active processes, polymer model, transcriptomics, molecular dynamics, Hi-C

Correspondence: menon@imsc.res.in

Highlights

- First-principles predictive model for large-scale nuclear architecture incorporating non-equilibrium activity
- Differential activity and looping patterns underly cell-type-specific features of such architecture
- Differential positioning of inactive and active X chromosomes an emergent property
- Simulations of the model recapitulate many known features of nuclear architecture and predict new ones

Introduction

Chromosomes are not distributed at random within the interphase nucleus, an observation that is central to our current understanding of large-scale nuclear architecture in the interphase nuclei of metazoans (Meaburn and Misteli, 2007; Cremer and Cremer, 2010; Bickmore and van Steensel, 2013). Gene rich, more open, early-replicating euchromatin regions are typically distributed more centrally than gene-poor, relatively more compact, late-replicating heterochromatin (Cremer and Cremer, 2010). Chromosomes are organised territorially, with each being segmented into relatively more (A) and less (B) active compartments that are then further subdivided into topologically associated domains (Lieberman-Aiden et al., 2009; Dixon et al., 2012; Fraser et al., 2015). In humans, gene-rich chromosome 19, containing a large number of house-keeping genes, is distributed more centrally across several cell types than the similarly sized but gene-poor chromosome 18 (Croft et al., 1999; Boyle et al., 2001).

This observation generalises to a gene-density-based radial positioning schema for all chromosomes (Takizawa et al., 2008). Gene-rich regions within chromosomes tend to orient towards the nuclear centre, with expressed alleles often found further from the nuclear envelope than ones that are not expressed (Takizawa et al., 2008; Therizols et al., 2014). In some human cell types, chromosomes appear to be positioned by size, with the centres of mass of smaller chromosomes disposed more centrally than those of larger ones (Sun et al., 2000; Bolzer et al., 2005; Kölbl et al., 2012). In female cells, the two X chromosomes are differentially positioned, with the more compact, inactive X-chromosome found somewhat closer to the nuclear envelope than the active one (Dyer et al., 1989; Jégu et al., 2017). Actively transcribed chromosomes tend to have rougher, more elliptical territories than less active ones (Eils et al., 1996; Berezney et al., 2005; Khalil et al., 2007; Sehgal et al., 2014; Jégu et al., 2017). The probability with which two loci along individual chromosomes are found in proximity to each other in ligation assays follows a power-law P(s) $\sim 1/\mathrm{s}^{\alpha}$ with $\alpha \simeq 1$ over an approximately 1 - 8 Mb range, consistent with a fractal globule picture of chromosome structure (Lieberman-Aiden et al., 2009; Mirny, 2011). Currently, experiments suggest that such organization is cell-type dependent and that α ($1 \leq \alpha \leq 1.5$) also varies across chromosomes over a comparable range (Sanborn et al., 2015; Kang et al., 2015).

Most model approaches to nuclear architecture assume *a priori* that chromosomes are structured polymers in thermal equilibrium (Cook and Marenduzzo, 2009; Tark-Dame et al., 2011; Marti-Renom and Mirny, 2011; Heermann et al., 2012; Vasquez and Bloom, 2014; Imakaev et al., 2015). Some models ignore thermal fluctuations altogether in favour of incorporating loop structure as derived from the Hi-C data, while also requiring compatibility with physical restrictions on the overlaps of chromosomes (Imakaev et al., 2015; Amitai and Holcman, 2017; Tjong et al., 2016). Others account for the domain structure of individual chromosomes (Odenheimer et al., 2005; Jost et al., 2014; Jost et al., 2017; Chiariello et al., 2015; Haddad et al., 2017; Ghosh

and Jost, 2017; Zhang and Wolynes, 2017; Tiana et al., 2016; Di Pierro et al., 2016; Di Pierro et al., 2017). As summarized above, large-scale nuclear architecture exhibits generic features that are largely common across cell types. These should severely constrain potential models (Bickmore, 2013). However, set against this stringent requirement, virtually all prior models for such architecture are incomplete: (i) these models fail to predict gene-density based or size-based positioning schemes; (ii) no simulations reproduce the chromosome-specific distribution functions for gene density or chromosome centre-of-mass that FISH-based experiments provide; (iii) the differential positioning of the active and inactive X chromosomes cannot be obtained using any model proposed so far and (iv), the spatial separation of heterochromatin and euchromatin, seen in interphase cell nuclei across multiple cell types, has not been reproduced in model calculations in which this information is not incorporated *a priori*. Understanding these discrepancies remains an outstanding problem.

All molecular machinery associated with chromatin remodelling, transcription and DNA repair is energy-consuming, relying on the hydrolysis of NTP molecules (Flaus and Owen-Hughes, 2011). Recently, we pointed out that this leads to the localised, irreversible consumption of energy at the molecular scale (Ganai et al., 2014). This energy is transduced, through chemo-mechanical "active" processes, into mechanical work (Weber et al., 2012; Zidovska et al., 2013; Chu et al., 2017). Such processes can be modelled *via* recently developed biophysical theories of "active matter" (Menon, 2010; Prost et al., 2015; Marchetti et al., 2013; Needleman and Dogic, 2017). We argued that a description in terms of inhomogeneous, stochastic forces acting on chromatin, equivalent to an effective temperature reflecting local levels of activity, provided the right biophysical setting (Fodor et al., 2015; Hameed et al., 2012). Describing each chromosome as a polymer composed of consecutive monomers, each representing a suitably averaged section of chromatin, different monomers can then be expected to experience different effective temperatures correlating to local active

processes (Ganai et al., 2014; Agrawal et al., 2017; Wang and Wolynes, 2011). Here, extending these ideas, we propose an *ab initio* biophysical approach to predicting both cell-type-specific and cell-type independent features of large-scale nuclear architecture, using data from RNA-seq experiments as a proxy for activity and a Hi-C-derived description of chromosome looping in each cell type. The model provides a unified understanding of a number of common features of large-scale nuclear architecture observed across diverse cell-types.

Results and Discussion

We model human chromosomes in diploid female (XX) cells within interphase, describing each as a polymer made up of monomers linked along a chain. These polymers are confined within a spherical shell that models the nuclear envelope. Each monomer represents a 1Mb section of chromatin (Kölbl et al., 2012; Jackson and Pombo, 1998). Our model chromosomes are dynamic and explore different configurations, based on the forces they experience. Such forces arise from the dense, non-equilibrium and fluctuating environment of the cell nucleoplasm, the interactions of chromosomes and chromosome-nuclear envelope interactions. A number of simulation snapshots of both homologs of chromosomes 18 and 19, against a background of all other chromosomes represented in grey-scale, are shown in Figure 1A. From such snapshots, we compute a variety of statistical properties of chromosomes accessed in experiments (Figure 1B - G).

We work with three models that associate local levels of non-equilibrium transcriptional activity to an effective temperature. In the **gene density model**, proposed in our earlier work, we chose the top 5% of monomers by gene density, assigning them an active temperature in excess of the physiological temperature T_{ph} (Ganai et al., 2014). The gene density model yields fairly accurate representations of the measured distribution function of DNA density S(R) in GM12878 cells, leading to very different distri-

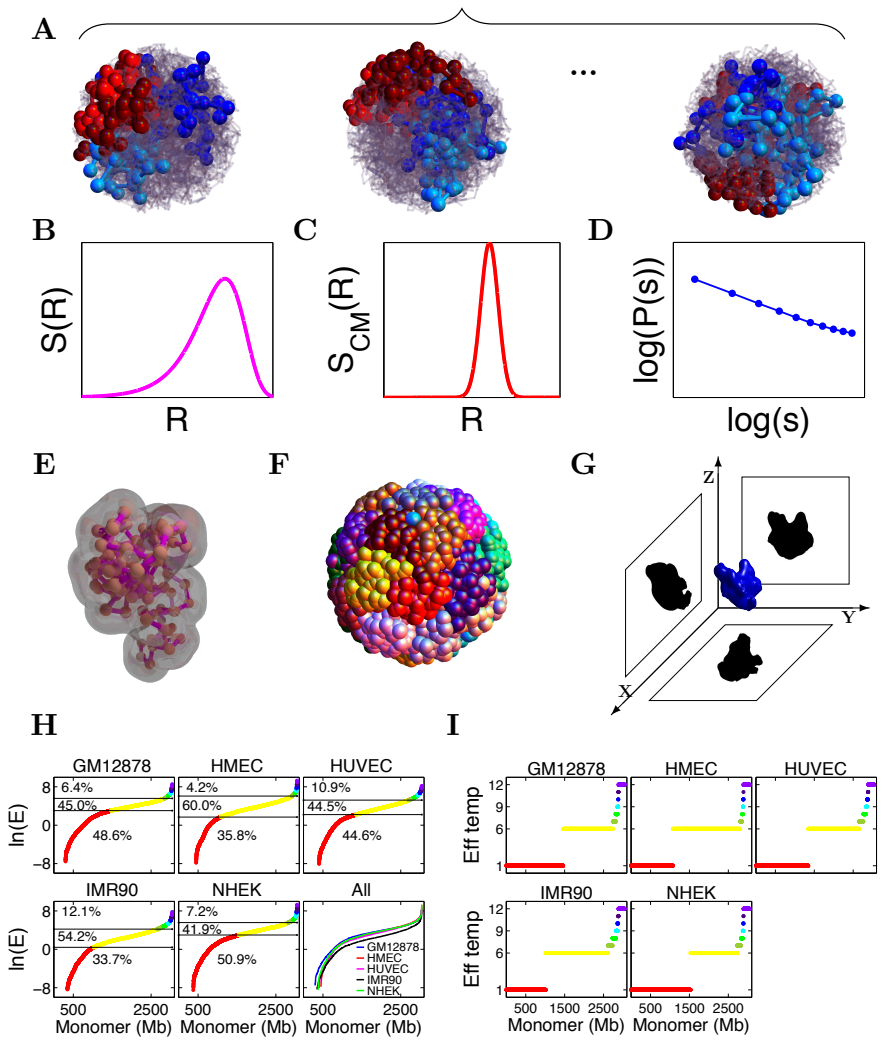

Figure 1: **Model schematics and active temperature assignments**
A. Several simulated configurations of 23 pairs of chromosomes within a spherical nucleus, with pairs of chromosomes 18 and 19 highlighted in the background of other chromosomes, shown in grey scale. Each bead represents a 1 Mb section on each chromosome. We average all calculated quantities, such as distribution functions, over a large number of such configurations in steady state.
B. Schematic of the DNA distribution S(R) of each chromosome, plotted against the radial coordinate R and averaged over many nuclei in our simulations.
C. Schematic of the centre of mass distribution of each chromosome, S_{CM}(R), plotted against the radial coordinate R and computed from an average over many simulated nuclei.

Figure 1 *continued*: **D**. Schematic of the contact probability P(s) between beads of chromosomes, for two monomers separated by an internal (genomic) distance s along the polymer.

E. The shapes of individual chromosome territories extracted from simulation configurations. Such shapes are used to compute a number of geometrical properties of chromosome territories, e.g. their volume, surface area, asphericity and other shape parameters.

F. Typical image of chromosome territories computed in our simulations, with each chromosome colored a different color, illustrating the emergence of territoriality.

G. Schematic illustrating a 2D projection of a three-dimensional chromosome territory, projected along the XY, YZ, and XZ planes. The ellipticity and regularity parameters can be computed from such 2d projections, and compared to 2D FISH data.

H. The logarithm of gene expression values for each 1Mb monomer, plotted in order of increasing gene expression. These are computed from transcriptome data. Data are shown for 5 cell types as indicated in the title to each sub-figure. The horizontal lines drawn motivate our assignment of effective temperatures as discussed in the text, and correspond to our assignment of activity in proportion to gene expression. The last sub-figure plots these data together, illustrating that the shape of the activity profile is largely similar, even though individual monomers in different cell types can be classified differently on the basis of their activity.

I. Assignment of effective temperature to each monomer for the combined model. The red monomers are simulated at $T = 1$, yellow at $T = 6$, yellow-green at $T = 7$, green at $T = 8$, cyan at $T = 9$, blue at $T = 10$, indigo at $T = 11$ and violet at $T = 12$ times the physiological temperature T_{ph}.

butions for the chromosome pairs 18 and 19 *vis a vis.* chromosomes 12 and 20, as seen experimentally (Ganai et al., 2014). However, such a model is insensitive to cell-type-dependent features of nuclear architecture (Bickmore, 2013). Accordingly, in the **gene expression model**, we focused on transcriptomes across a number of model systems, exploring varied ways of associating transcript levels to effective temperatures. Figure 1H shows RNA-seq derived FPKM values summed over 1Mb intervals, indexing transcript levels, across GM12878, HMEC, HUVEC, IMR90 and NHEK cell types. Their distribution follows a Gumbel form (Figure S1). The values in Figure 1H are plotted in increasing order of expression on a logarithmic scale. We chose structured effective temperature assignments that reflect the overall shape of this curve. Surprisingly, the gene expression model did not yield appreciably better results than the much simpler gene density model. To address this, we noted that transcript levels need not directly correlate to activity, since FPKM values are controlled by the rate at which transcripts are both produced and degraded, because non-coding transcription is not fully captured in this version of our model, and because our description averages over the typical time-scales associated with transcriptional "bursts" (Fraser and Bickmore, 2007; Chubb et al., 2006). We felt that a model which included features of both gene density and gene expression models should provide a more accurate representation of inhomogeneous cell-type-dependent activity (Murmann et al., 2005). Accordingly, we decided to assign monomers with a gene density above a present cutoff, the maximum active temperatures, as in the earlier gene-density model. All the results presented in this paper are for this **combined model**. Figure 1I shows temperature assignments, within the combined model, for 5 cell types. Such inhomogeneous (effective) temperature assignments, correlating both to gene density and transcription levels averaged over consecutive 1Mb sections of each chromosome, lie at the core of our work.

Inhomogeneous activity underlies large-scale nuclear architecture

Figure 2A shows an example of simulation-derived chromosome territories for the GM12878 cell type, with each chromosome coloured a different colour. Simulations recover such territorial organization robustly, illustrating how territoriality is an emergent consequence of our model. Figure 2B shows a cut-away profile showing the averaged spatial distribution of active (white, $T/T_{ph} > 1$) and inactive (black, $T/T_{ph} = 1$) monomers in the GM12878 cell type, extracted from snapshots of a typical configuration. Low gene density monomers, shown in black and representing heterochromatin, are enriched towards the boundaries of the nucleus, whereas high gene-density euchromatin regions, shown in white, preferentially occupy the bulk. In Figure 2C, we show a cutaway profile of the time-averaged effective temperature within our simulated nucleus, an indicator of local activity in each spherical shell centred around the origin. In Figure 2D we show the time-averaged gene density across spherical shells, in a similar visualisation. Gene densities, as well as activity, increase towards the nuclear centre.

Our biophysical description of chromosomes and their structuring, given our coarse-graining to the 1Mb scale, reproduces the different spatial distributions of euchromatin and heterochromatin, a feature seen across multiple cell types. A central consequence of our model is that gene densities should correlate to a larger strength of mechanical fluctuations *i.e.* activity, and that the distribution of both these quantities should be attenuated towards the boundaries of the nucleus. This is an emergent property, arising from the combination of differential activity and confinement, that could not have been inferred from how the model was constructed.

The model predicts positional distributions of individual chromosomes

Chromosome-specific distribution functions S(R) are obtained experimentally using confocal slices of FISH images from an ensemble of fixed nuclei. Our computed S(R) for chromosomes 18 and 19 in the 5 cell types we examine are shown in Figure 3A. All data are averaged over the two autosomal homologs, as their positioning was

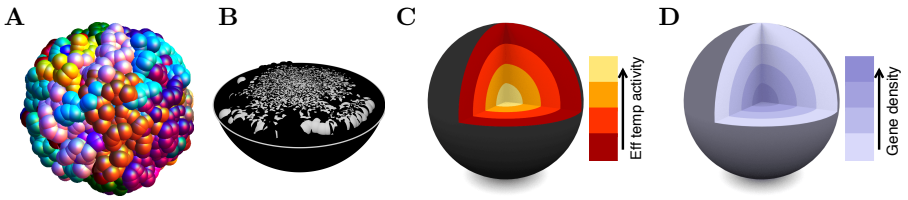

Figure 2: **Model predictions for large-scale features of nuclear architecture**
A. Chromosome territories computed in our simulations, with each chromosome colored a different color. Note the tendency of each chromosome to overlap relatively little, visually representing territoriality.
B. A cut-away sphere representation of the average spatial distribution of euchromatin (or active white) and heterochromatin (or inactive black) monomers as computed for the GM12878 cell type. Here, the active monomers are defined as those having an effective temperature in excess of the physiological one. Heterochromatin is found more peripherally compared to euchromatin which is located towards the nuclear interior.
C. A cutaway sphere representation of average effective temperatures within the simulated nucleus, as computed for the GM12878 cell type. This illustrates the larger effective temperatures, indicating enhanced activity, obtained towards the centre of the nucleus, in comparison to a lower effective temperature in the nuclear periphery.
D. A cutaway sphere representation of the average gene density within the simulated nucleus, computed for the GM12878 cell type. This illustrates the excess in gene density seen towards the centre of the nucleus in comparison to the gene density in the nuclear periphery. This separation of gene-dense and gene-poor 1Mb segments of chromatin correlates to the distinction in the spatial positioning of euchromatin and heterochromatin.

found to be equivalent. For the GM12878 cell type, we compare our results with experimental results extracted from Ref. (Kreth et al., 2004). S(R) for chromosomes 18 and 19 exhibit well separated peaks, a feature that holds across cell types. For comparison, the R^2 rise of S(R) towards the nuclear envelope, expected for uniformly distributed chromosomes, is shown in each subfigure. Figure S2 shows our calculated S(R) for all chromosomes in the 5 cell types. Figure S3 shows S(R) for the GM18278 cell type where we compare the predictions of the gene expression model and the combined model. In Figure S4 we show how S(R) for the GM18278 cell type varies when we include or exclude looping and when we include or exclude activity. The predictions of the different models differ substantially for both the gene-rich chromosomes as well as

the smallest chromosomes. In general, both looping and differential activity are needed to best represent available experimental data.

Figure 3B shows S(R) for chromosomes 12 and 20. For the GM12878 cell type, we compare our results with results from Ref. (Kreth et al., 2004). Note that simulation data for the different cell types all yield similar plots for S(R), with the exception of the GM12878 cell type where, although the simulation and experimental data peak at different locations, the overall shape of the curve is rendered accurately, including the relative shift in peak positions. Figure 3C shows the distribution of centres of mass of specified chromosomes, S_{CM} (R), for chromosomes 18 and 19. For GM12878 cells, we compare our results with experimental data from Ref. (Kalhor et al., 2011). The centre of mass distributions are captured well, especially for chromosome 19. The broader distribution of S_{CM}(R) for chromosome 18 is also in agreement with the left tail of the experimental data, although the experiments show a weaker and more outward shifted peak than the simulation prediction. Broadly, differences in positioning of chromosomes across cell types are more apparent in S_{CM}(R) compared to S(R). Figure S5 shows S_{CM}(R) for all cell types across all chromosomes. We compare the predictions for S_{CM}(R) in the gene expression and combined models in Figure S6. Results for S_{CM}(R) with different combinations of activity and loops are shown in Figure S7. Overall, apart from the gene rich chromosomes, we do not see a substantial difference between the predictions of the gene expression and combined models. The largest variability across cell types is seen in chromosomes 1, 4, 7, 11, 12, 16, 21 and 22. S_{CM}(R) for gene poor chromosomes are sharply peaked while gene rich chromosomes have broader distributions across all cell types.

Figure 3D shows the partial distribution functions S(R) for inactive and active monomers in the GM12878 cell type for chromosomes 1, 13, 19 and 21. The distribution for active monomers is shifted towards the nuclear centre whereas for the inactive monomers, it is seen to be shifted towards the nuclear periphery. These results relate to

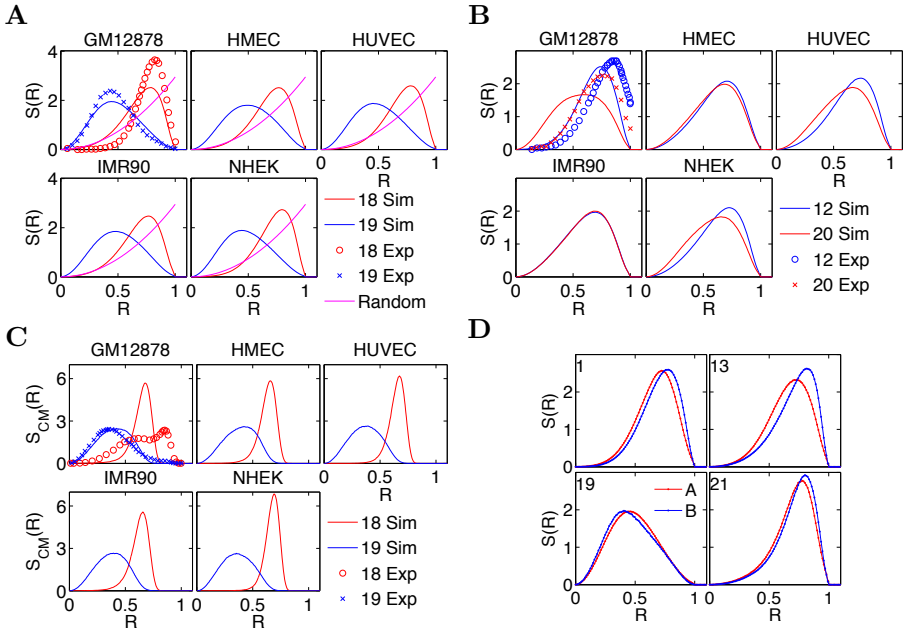

Figure 3: **Predicted radial distribution functions S(R) compared to experimental data**

A. Distribution of monomer density S(R), reflecting the local density of DNA, for chromosomes 18 and 19 (red and blue lines respectively) across 5 cell types as indicated in the titles of each sub-figure. Experimental data obtained from Ref. (Kreth et al., 2004) for the GM12878 cell type is plotted together with the simulation predictions, (red ovals: Chr 18, blue crosses: Chr 19). If chromosomes are distributed randomly across the nucleus, $S(R) \sim R^2$ is expected, as shown with magenta lines.

B. Distribution of the density of monomers, reflecting the local density of DNA, for Chr 12 and 20 (blue and red lines respectively) across 5 cell types as indicated in the titles of each sub-figure. Experimental data obtained from Ref (Kreth et al., 2004) for the GM12878 cell type is plotted (red ovals: Chr 18, blue crosses: Chr 19), together with the simulation prediction.

C. Distribution of chromosome centres of mass for Chr 18 and 19 (red and blue lines respectively) shown for 5 cell types, as indicated in the titles of each sub-figure. Experimental data obtained from Ref. (Kalhor et al., 2011) for the GM12878 cell type are plotted (red ovals: Chr 18, blue crosses: Chr 19) together with the simulation prediction.

D. Density distribution S(R) of active (red) and inactive (blue) monomers for the GM12878 cell type. These are plotted for 4 chromosomes largest Chr1, smallest Chr 21, gene poor Chr 13 and gene rich Chr 19. The distribution of active monomers is more interior with respect to inactive monomers. Here, inactive monomers refer to those monomers assigned a temperature of T = 1; all other monomers are active.

the experimental observation that active alleles are positioned more towards the interior of the nucleus, an effect strong enough to be apparent in our simulations (Fedorova and Zink, 2009; Therizols et al., 2014).

We conclude that there are subtle differences in the gene density distribution function S(R) as well as in the mean centre-of-mass distribution S_{CM} (R) of chromosomes across cell types. These originate both in differences in activity profiles across different cell types as well as variations in their loop content.

Model predictions for size- and gene-density-dependent chromosome positioning

Figures 4A-4D show our computation of the mean centre of mass of each chromosome within the combined model for the cell types GM12878 and IMR90. Also shown are fits to straight lines as a function of chromosome size (Figures 4A and 4C) and as a function of gene density (Figures 4B and 4D). The fit against chromosome size excludes both the two smallest (21 and 22) and the 7 largest (1, 2, 3, 4, 5, 6, 7) chromosomes. We also show, in the same figure, experimental data for the GM12878 cell type extracted from Figure 6 of Ref. (Kalhor et al., 2011). These data are shown as filled circles in (Figures 4A and 4B) which represents our predictions for the same cell type, and as open circles in (Figures 4C and 4D), where their locations are only indicative.

The simulations reproduce most of the experimental systematics. The positions of virtually all chromosomes, with the exception of chromosome 21, lie within the error bars of the experimental data. Note that the experimental and simulation data coincide for some chromosomes. The positions of chromosomes 7, 9, 13, 17, 18, 19 and 20 are very close to the experimental data, reproducing the unusual non-monotonicity in their positions.

If the smallest and largest chromosomes are excluded, an approximate size-dependence of chromosome positions relative to the nuclear centre is predicted. However, the ac-

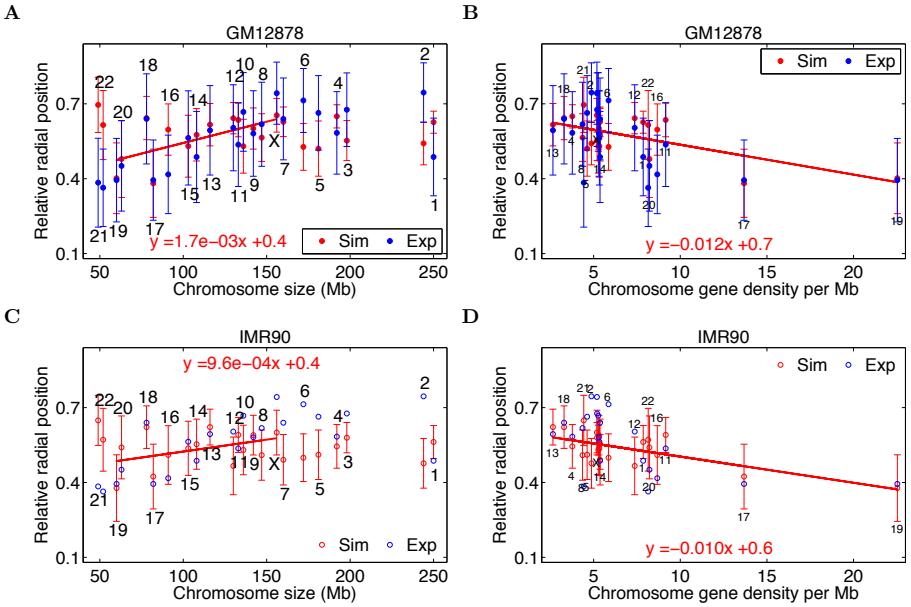

Figure 4: **Predicted chromosome centre-of-mass locations compared to experimental data**

A. Predictions for the mean centre-of-mass location for each chromosome, for the GM12878 cell type and as a function of chromosome size, compared to experimental data of GM12878 extracted from Ref. (Kalhor et al., 2011). Simulation and experimental points are shown using red and blue filled circles respectively, together with error bars, which for the simulations are calculated from the second moment. The relative radial position 0 and 1 represent the centre and periphery of the nucleus. Chromosome numbers are indicated above or below each error-bar. The simulation points, after dropping chromosomes 1,2,3,4,5,6,7, 21 and 22 (see text), are fitted to a straight line whose slope is indicated in the bottom of the sub-figure.

B. Predictions for the mean centre-of-mass location for each chromosome, for the GM12878 cell type, compared to experimental data extracted from Ref. (Kalhor et al., 2011), and plotted against the function of chromosome gene density. The simulation points are fitted to a straight line whose slope and intercept are indicated in the bottom of the sub-figure.

C. Predictions for the mean centre-of-mass location for each chromosome, computed for the IMR90 cell type, as a function of chromosome size, with experimental data for the GM12878 cell type cf. Ref. (Kalhor et al., 2011) shown only for illustrative purposes. Simulation and experimental points are shown together using red and blue ovals respectively, along with simulation error bars only. Simulation points, excluding chromosomes 1,2,3,4,5,6,7, 21 and 22, are fitted to a straight line whose slope and intercept are provided in the top of the sub-figure.

D. Predictions for the mean centre-of-mass location for each chromosome, for the IMR90 cell type, with experimental data for the GM12878 cell type cf. Ref. (Kalhor et al., 2011) shown only for illustrative purposes, plotted against chromosome gene density. All the simulation points are fitted to a straight line whose slope and intercept are provided in the bottom of the sub-figure.

tivity associated with each individual chromosome also plays a role in determining its position. The mean centre-of-mass locations for chromosomes in different cell types are similar but not identical. Chromosomes 18 and 19, although similarly sized, have very different positions relative to the nuclear centre, as also seen in the data of Figures 3A and 3C. Note that chromosomes 21 and 22 in Figure 4A are positioned more towards the exterior of the nucleus in the simulations than in the experimental data. When chromosome centres-of-mass are plotted against gene densities the slope of the straight line is negative in all cell types (Figures 4A and S8). Thus, depending on the region that is fit, one can have reasonable fits to both size dependence and gene density dependence of chromosome centres of mass relative to the nuclear centre. The fact that the smallest chromosomes, Chr 21 and 22, lie outside of the fit to chromosome size may reflect aspects of their activity that our method does not resolve, as well as variations in loop assignments.

Figure S9 shows the mean centre of mass position as computed for the GM12878 cell type, across a variety of simulation conditions, including for the gene density model as well as for the combined model with various choices for the incorporation of loops and activity. Figure S9A shows results for the gene expression model. In Figure S9B we show results for the case in which we allow differential activity but ignore looping. In Figure S9C we show results for the case in which differential activity is absent but looping, as prescribed by the Hi-C data, is retained. All monomers then experience the same effective temperature, which we take to be the thermodynamic temperature. Finally, in Figure S9Ds, we show results for the case where looping is absent as well, so that this case corresponds to the case of chromosomes without loops at thermal equilibrium. From these, we conclude that in the absence of both activity and looping, chromosome positioning is only weakly structured. Our simulations indicate that chromosome positioning is very weakly size-dependent or even independent of size in all conditions where activity is switched off. Allowing for loops induces

some changes in positioning but these results do not match with experiment. Allowing for activity, but ignoring loops, leads to a differential positioning of chromosomes but, if anything, the size dependence of chromosome positions is *opposite* to that seen in the data. Only models which incorporate both activity and looping are successful in reproducing both the approximate size dependence of radial positioning, while also accounting for those specific cases which fall outside this general trend, such as chromosome 19.

The model predicts the centre-of-mass positions of most chromosomes with reasonable accuracy, well within the error bars on the measurements for virtually all chromosomes. Finally, the fact that a number of broad features of the experiments are reproduced in the model suggests that the large-scale structure and positioning of individual chromosomes are principally determined by inhomogeneous activity across chromosomes, the presence of loops and confinement.

Shapes and statistical features of individual chromosome territories compare well to experiments

Figures 5A and 5B show territories for the two autosomes corresponding to chromosomes 12, 20, 18 and 19. In Figure 5C, we show comparisons between 2d FISH data for chromosome regularity and ellipticity on WI38 cells, for which data is available (Sehgal et al., 2014), to predictions from our simulations for the GM12878 and IMR90 cell types. Both IMR90 and WI387 are lung fibroblast cell lines. Chromosomes are indexed, along the x-axis, in order of their gene density. The simulation results and experimental data appear to follow each other, with the simulations finding the same dip and subsequent rise of both ellipticity and regularity around chromosome 22. Both ellipticity and regularity peak for chromosome 11, a feature both of the simulations and of the experiments. The ellipticity and regularity also appear to decrease weakly with increasing gene density, although individual chromosomes may deviate from this

general trend (Figure 5C).

Figures 5D and 5E show the summed volume overlap (SVo), sometimes referred to as the intermingling, and used to understand chromosome-chromosome interactions in trans, of different chromosomes in our model. The ordering of chromosomes according to their gene density per chromosome as shown on the x-axis is the same as the ordering used for the 2d projected data in Figure 5C. The largest overlap is for the most gene rich chromosome. There are perceptible differences in the overlaps of chromosomes in the GM12878 and the IMR90 cell types.

In summary, the simulations reproduce broad features of individual chromosome territories. More active chromosomes appear to deviate more from a spherical shape and tend to have rougher territories (Berezney et al., 2005). The summed volume overlap appears to increase approximately linearly with chromosome gene density, with the Xi being an exception to this trend. Activity and looping tend to have countervailing trends, since activity expands chromosome territories while looping contracts them.

Simulations reproduce the differential positioning of the active and inactive X chromosome

Experiments investigating the positioning of the active and inactive homologs of the X chromosomes within interphase have consistently found that they are differentially positioned. The inactive X chromosome (the heterochromatic Barr body) is located most often towards the periphery of the nucleus (Jégu et al., 2017). This contrasts to the more central disposition of the active X chromosome Xa, which is larger and more extensively transcribed than the more compact Xi. We thus specifically investigated the positioning and other structure properties of the Xa and Xi chromosomes, since we expected that they would provide an example of where our methods, which emphasise the role and importance of activity, would yield predictions that other models could not.

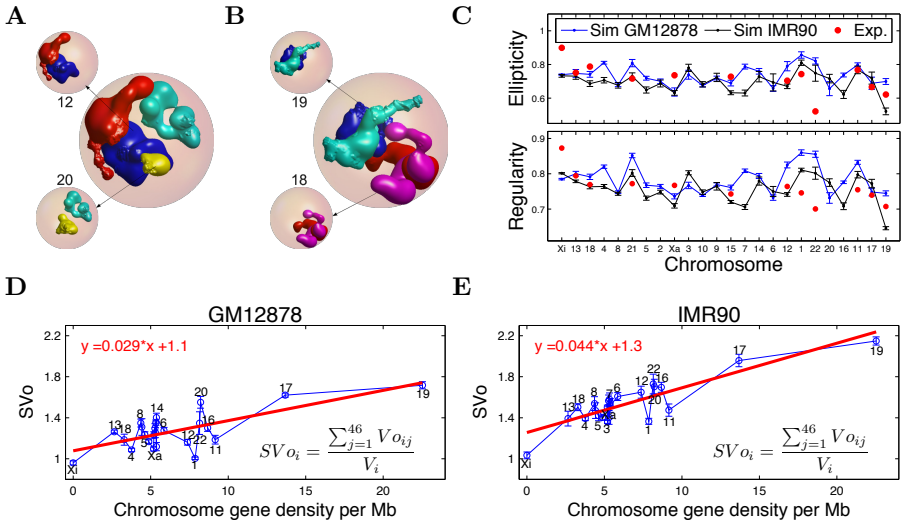

Figure 5: **Structural properties of individual simulated chromosomes in our model**
A. Snapshot of simulated configurations of both homologs of chromosomes 12 and 20. Each chromosome is colored differently so that they can be separately visualized.
B. Snapshot of simulated configurations of both homologs of chromosomes 18 and 19. Each chromosome is colored differently so that they can be separately visualized.
C. Ellipticity and Regularity for each chromosome as predicted by the model and obtained from simulations representing the GM12878 (blue) and IMR90(black) cell types. These are compared to experimental data (red oval symbols) from 2d FISH experiments Ref. (Sehgal et al., 2014) for a cell type closely related to the IMR90 cell type. Ellipticity values of 1 represent a perfect elliptical chromosome and regularity values of 1 refer to a perfectly regular chromosome, without roughness. The x axis is plotted in order of increasing gene density.
D-E. Summed volume overlap (SVo) of chromosomes in GM12878 and IMR90 cell types, with the x-axis plotted in order of increasing gene density per chromosome. There is a weak increase with gene density in both cell types, shown as the solid line, representing the best linear fit to the data. The IMR90 cell shares more volume overlaps with other chromosomes compared to the GM12878 cell type. The (self-) volume overlap (Vo) for the same chromosome is taken to be 0.

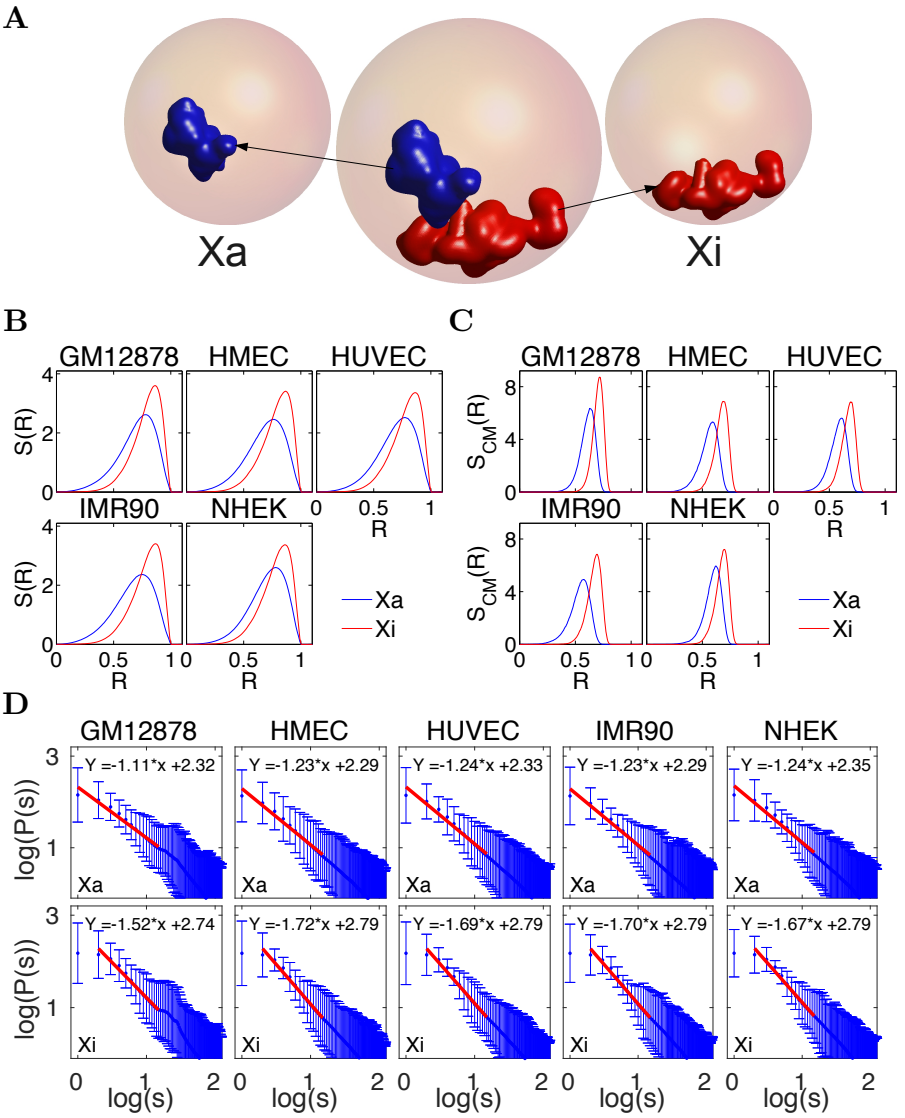

Figure 6: **Predicted distributions and structural features of the inactive and active X chromosomes**

A. Snapshot of a typical configuration of the active Xa and inactive Xi chromosome obtained from simulations.

B Monomer density distribution S(R) vs R, for the Xi and Xa chromosome as obtained from simulations across 5 cell types, named in the header to each subfigure. The inactive X chromosome, Xi, is shown in red and the active X chromosome, Xa, is shown in blue.

C. Distribution of the location of the centre-of-mass of the Xi and Xa chromosome as obtained from simulations across 5 cell types, named in the header to each subfigure. The inactive X chromosome, Xi, is shown in red and the active X chromosome, Xa, is shown in blue.

Figure 6 *continued*: **D**. Contact probability P(s) vs s, for the active (top row) and inactive (bottom row) X chromosomes, computed for 5 cell types within our simulations. The active X chromosome exhibits a reasonable power-law decay of P(s) with an exponent α between 1.1 and 1.25. The inactive X chromosome shows a reduced region of power-law scaling, with an exponent across this reduced range which is between 1.5 and 1.7. Red lines show the power-law fit in both cases, with the fit parameters indicated within each sub-figure.

Figure 6A shows a simulation snapshot of active and inactive X chromosome territories. Figure 6B shows our predictions for how these chromosomes are differentially positioned across all the cell types we study, through S(R). The inactive chromosome has an S(R) which is sharply peaked close to the nuclear periphery. Although the active X chromosome has a peak at a comparable location, its distribution has a long tail towards the nuclear centre. Figure 6C shows the calculated distribution of the centre of mass $S_{CM}(R)$ for these multiple cell types, verifying this essential distinction.

We can compute the contact probabilities P(s) by applying a cutoff to the monomer-monomer distance distributions obtained in our simulation, averaging across a large number of simulation configurations. Figure 6D shows our computation of the contact probability P(s) for both Xa and Xi, across the 5 cell types we study. The cell type is shown at the top of each sub-figure. The active X chromosome shows more prominent power-law scaling of the contact probability than the inactive X chromosome, where any fit to a power law can only be over a far shorter genomic scale. Exponents for the power-law scaling of P(s) range from 1.11 - 1.24, with the smallest values obtained for the GM12878 cell type. For the inactive X chromosome, the power-law exponent obtained over a limited range varies from 1.52 - 1.72, significantly different from the range of α values obtained in the active case. The variation in the scaling of P(s) between Xa and Xi should be accessible experimentally.

Structural features of individual chromosomes are well described in our model

Figure 7A exhibits our results for contact probability P(s) of chromosome 1, across the five different cell types we study here. The data for small s show a power-law P(s) $\sim 1/\mathrm{s}^{\alpha}$ behaviour over approximately a decade. The exponent is smallest for the GM12878 cell type, where our fits yield $\alpha \approx 1.06$. This value is very close to that obtained experimentally across the same region of genomic separation (Lieberman-Aiden et al., 2009; Sanborn et al., 2015). Values of α for all other cell types are consistently larger, with the exception of the IMR90 cell type. Overall, fitting α directly to the data across cell types yields $0.97 \leq \alpha \leq 1.27$. We see $P(s) \sim 1/s^{\alpha}$ with $\alpha \simeq 1$ over a $1 - 10$ Mb range, as predicted by the fractal globule model, even though our model lacks virtually all the requisite ingredients for this model. All we require is that activity is differentially distributed along the chromosome, that we account for looping as drawn from the Hi-C data, and that we account for crowding by other chromosomes, all features that previous work elides. Figures S10 and S11 shows plots of P(s) for all chromosomes computed for the GM12878 and IMR90 cell type. This P(s), for each chromosome, is best described within the 1 - 10 Mb range in terms of a wide range of exponents in the range $0.97 \leq \alpha \leq 1.40$.

Our model specification can be relaxed in several ways, so that we can examine and quantify independent contributions to this behaviour. For the specific case of chromosome 1, we have also investigated the predictions of the gene expression model, as shown in Figure S12A, as well as the effects of varying both activity and looping in the combined model (Figures S12B-S12D). Both the gene expression and combined models exhibit values of α which lie close to the experimental data, which sees values of α close to 1. In Figures S12B-S12D, we show results for the combined model with varying combinations of activity and looping. In the absence of both activity and loops, the exponent is close to the $\alpha = 1.5$ expected for simple polymers. Adding loops or activity reduces this exponent. However, only the combined model, which includes both activity and looping obtains α values closest to those in experiments.

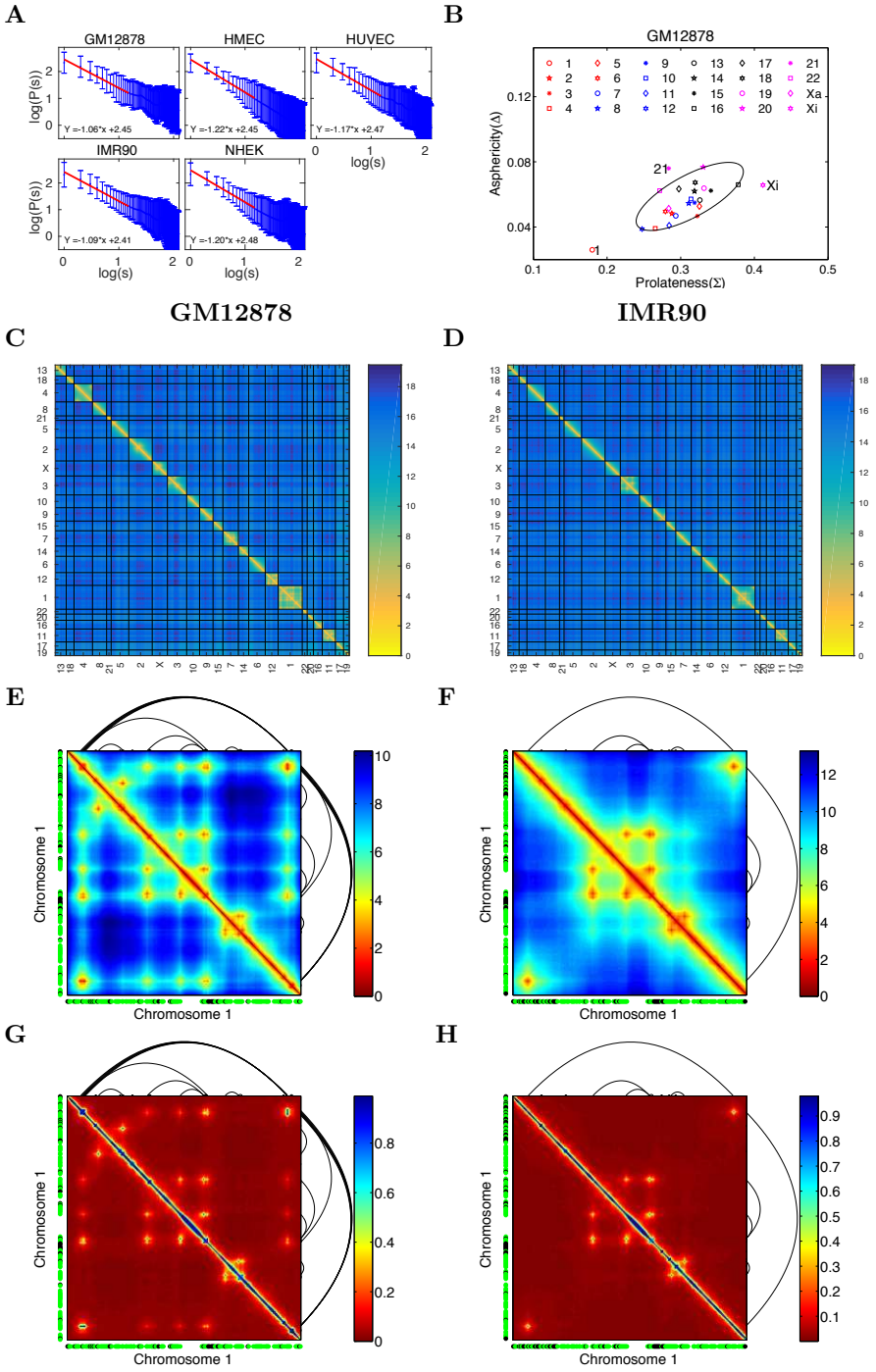

Figure 7: **Predicted chromosome contact probabilities, distance maps and contact maps**

Figure 7 *continued*: **A**. Contact probability P(s) as a function of genomic distance for Chr 1, computed across a range of 1 - 15 MB and plotted for 5 different cell types. Our data is plotted with blue dots displayed with error-bars. Depending on the region that is fit, a power law scaling is obtained with an exponent between roughly 1.17 and 1.22; these fits are shown with red colors.
B. Calculated average values of the prolateness parameter (Σ) and the asphericity parameter (Δ) for the GM12878 cell type. Larger (smaller) chromosomes have smaller(larger) values of Σ and Δ, implying that larger chromosomes are more close to spherical, while smaller chromosomes prefer a more prolate, rod-like shape. The data suggests that values of Σ and Δ for Chr 1, 21 and Xi take more extremal values than for the other chromosomes, as shown by the ellipse drawn together with the data.
C. Heatmap of mean distances between monomers, the distance map, in which chromosomes are ordered by their gene density, shown for the GM12878 cell type.
D. Heatmap of mean distances between monomers, the distance map, in which chromosomes are ordered by their gene density, shown for the IMR90 cell type.
E-F. Heatmap of the distance matrix for chromosome 1, expanded out from Figures 7C and 7D. The locations of the permanent loops inferred from the Hi-C data are plotted in black. Individual monomers at T = 6 and $7 \leq T \leq 12$ are shown in green and black, adjacent to the X and Y axis, respectively.
G-H. Contact map inferred from the distance matrix for chromosome 1, (Figures 7C and 7D). The locations of the permanent loops inferred from the Hi-C data are plotted in black. Individual monomers at T = 6 and $7 \leq T \leq 12$ are shown in green and black, adjacent to the X and Y axis, respectively.

In Figure 7B, we show the spread of the asphericity parameter Δ and the shape parameter Σ, across chromosomes in the GM12878 cell type. The simulations yield a linear relationship between Δ and Σ. Larger chromosomes have smaller values of Δ and Σ. Thus, an outcome of our model is that larger chromosomes will tend to be more spherical. The regularity and ellipticity indices calculated for the 2-d projections are in reasonable agreement with experimental trends (Figure 5C). However, we predict that the asphericity and prolateness of the Xi chromosome should provide an exception to the general trend for other chromosomes. We find that the data appears to fall into two classes, one a more compact set corresponding to all chromosomes with the exception of Chr 1, 21 and Xi, contained within an elliptical domain as shown in Figure 7B. Values of Δ and Σ for these special three chromosomes appear to be somewhat displaced from the locations for the other chromosomes, falling approximately onto the periphery of a larger ellipse. We show similar plots for other cell types in Figure S13. In the absence of activity, both if loops are present or absent, the Δ and Σ values for these chromosomes falls within the inner elliptical region across the 5 cell types we consider (Figure S14).

Figure 7C shows a heat map of monomer distances of chromosomes, indexed in increasing order of gene density for the GM12878 cell type. A similar plot is shown for the IMR90 cell type in Figure 7D. One feature of the data is that the more active chromosomes show smaller values of inter-chromosomal distance, likely reflecting the fact that more active regions are enriched towards the nuclear centre. In Figures 7E and 7F, we show the enlarged distance maps for chromosome 1. Applying a cutoff to such data, we can derive the likelihood of contacts arising from intra-chromosomal interactions, yielding P(s). Solid lines outside the figure body indicate those permanent attachments between different monomers that the Hi-C data provides. Note that regions connected by such loops exhibit a larger overlap. Figures 7G and 7H shows the contact maps inferred after applying a cutoff to the corresponding distance map.

The borders of the axes show, in black and green, the active temperatures associated to specific monomers belonging to those chromosomes. The black colour refers to the most active monomers, with an effective temperature of 12 in units of the physiological temperatures whereas the green colour shows monomers with an effective temperature in the range $6 - 11$. Monomers with a lower effective temperature are not shown. Regions with the same high effective temperature appear to contact each other more, but these are further modulated by the presence of internal loops. Note the presence of a dark banded region towards the centre of chromosome 1, associated to a large inactive region on this chromosome. This is a prominent feature of the experimental data, also seen in other cell types (Rao et al., 2014). We display similar plots for other cell types in Figure S15.

To summarise, our model yields structural information for chromosome structures and shapes that are broadly in agreement with available data. Our simulated distance maps lack the fine detail of distance maps computed in Hi-C experiments, which provide data for contacts at the smaller scales of 10 - 100 kB, but nevertheless are relevant to experiments that probe large-scale structuring. Our computed P(s) contains about a decade or so of power-law decays, with exponents that are comparable to those seen in experiments. Our model-based predictions for trends in the asphericity and prolateness of chromosomes with chromosome size and gene density are testable.

Conclusions

Model descriptions of chromosomes must bridge multiple scales, ranging from microscopic length-scales of a few angstroms to scales of microns, of order the nuclear size. For now, brute-force atomistic simulations of the 23 pairs of chromosomes in human nuclei contained within the densely crowded, fluid and confined environment of the nucleoplasm are impossible. They are likely to remain so at least for the foreseeable future. Understanding how microscopic descriptions connect to macroscopic ones

thus requires intuition for the processes that act to couple these scales, so that model building, which is as much about what to leave out as it is about what to leave in, can proceed.

The model described in this paper stresses a specific biophysical effect, ignored in previous work, of relevance to the modelling of chromosomes in living cells. We began by emphasising the relevance of non-equilibrium effects arising from local transcriptional activity for descriptions of nuclear architecture (Chu et al., 2017; Almassalha et al., 2017). We proposed that the intensity of active processes should increase with increased transcription levels. We mapped a reasonable measure of local transcriptional activity, inferred from combining population-level measures of local RNA-output with estimates of the local gene density, into an effective temperature seen by each monomeric unit in our polymer model of chromosomes. We then performed simulations of these confined polymers, with properties chosen to reflect generic biophysical aspects of chromosomes. The monomers in our simulation represented 1Mb sections of chromosomes, although we could have defined our model at the smaller scales of 0.1 or even 0.01 Mb. However, the averaging inherent in summing transcriptional output over a 1Mb scale renders the model relatively less sensitive to errors and noise in this input. Further, the 1Mb scale is believed to be an appropriate building block for chromosome territories. A more detailed and explicit model for non-equilibrium activity and its consequences for an active temperature description would be useful, but the form such a model ought to take is presently unclear and best left to more extensive investigations. Irrespective of potential quantitative improvements on the model front, the broad trends we describe here should be largely robust.

We made a number of other modelling choices. First, we ignored the important role of lamin proteins in anchoring specific lamin-associated domains (LADS) to the nuclear lamina, as well as the interactions of specific gene loci with nuclear pore complexes (Mattout et al., 2015). While these are important omissions, they can at least

be qualitatively justified by the biophysical intuition that the activity-based physical segregation of chromosomes is a bulk or "volume" effect that should dominate, at the simplest level of description, over "surface" effects arising from interactions with the nuclear envelope. Thus, modelling the effects of interactions of LADS with the nuclear lamina by introducing weak monomer-specific interactions with the inner surface of the confining sphere in our simulations might be expected to modify the results we present here for specific chromosomes, but hopefully in a controllable manner. Second, we ignored nucleoli, formed around nucleolus organizer regions containing multiple copies of rRNA genes, with such regions located on the short arms of the acrocentric chromosomes 13,14,15, 21 and 22 (Németh and Längst, 2011). We can account at least qualitatively for the presence of the nucleolus, a relatively large and dense nuclear landmark, by excluding a pre-decided sub-volume of space within the simulated nucleus from being occupied by other chromosomes and adding a weak attractive attraction that favours association to monomers associated to the p-arms of the acrocentric chromosomes. Third, we simulated the nucleus as a spherical shell containing our model chromosomes, although nuclear shapes exhibit considerable variability and much of the experimental data comes from experiments on the relatively flattened nuclei of fibroblasts (Bolzer et al., 2005). Our model could be generalised to account for the effects of variable nuclear shapes. Fourth, we ignored potential interactions *in trans* between chromosomes. Such interactions could potentially arise from the looping out of loci on different chromosomes to interact at transcription factories (Maharana et al., 2016). We could account for this by making designated monomers on different chromosomes "sticky" with respect to each other, thus coupling regions of different chromosomes that are known to physically localise together when co-transcribed. Fifth, in using RNA-seq data as a proxy for activity, we ignored non-coding transcription, since RNA-seq largely provides steady-state gene expression. Inferring activity from newer methods such as GRO-seq, which also extracts nascent and rapidly degraded transcripts, may

help to provide a more accurate view of transcription-coupled activity. Last, the role of nuclear actin and associated motors remains unclear, although they could potentially contribute additional source of non-equilibrium noise (de Lanerolle, 2012). Indeed, all the possible improvements on our model that we list above could be incorporated, but only at the expense of more model detail and a number of further assumptions. These modifications of our model would have obscured the core argument of this paper, that the primary driver of many features of nuclear architecture is non-equilibrium activity that is inhomogeneous across chromosomes, so we choose to leave these questions to future work.

In first-principles approaches, a small set of initial model assumptions argued for on general grounds must yield consistent explanations and descriptions for all data, not just those the model abstracts in its construction. The advantage of simple models is that they enable us to concentrate on underlying principles that are often obscured by the complexity of real data, including intrinsic heterogeneities across cell populations, varied experimental and analysis procedures and the lack of sufficient statistics in some cases. Prior models for nuclear architecture in mammalian cells fail to reproduce many general attributes of nuclear architecture known from experiment. These properties – certainly their important trends – are emergent in our calculations, since they were not directly encoded in our model specification. This suggests that our methodologies provide hitherto unavailable biophysical insights into the determinants of large-scale nuclear architecture in metazoans.

Methods

Interactions of model chromosomes

Our model chromosomes (diploid, XX) occupy the interior of a spherical shell of radius R_0. The interaction between neighbouring monomers is of the finitely-extensible non-linear elastic (FENE) form. These monomers further interact with (non-neighbouring)

monomers via a Gaussian interaction, the Gaussian core potential used to model polymer brushes (Stillinger, 1976)

Simulation Methodology

We adapt the widely-used LAMMPS code implementing Brownian dynamics (Plimpton et al., 2007) for a polydisperse polymer system with a local monomer-dependent effective temperature. For each monomer, LAMPPS applies a Langevin thermostat, via an over-damped equation of motion, with a different "effective" temperature T_i associated to each monomer, reflecting its local level of activity. In thermal equilibrium, we have $T_i = T_{eq}$ for all monomers.

Units and Normalization

We work in de-dimensionalised units, discussed in Supplementary Information.

Deriving effective temperatures

In the **gene density model**, the gene content of each 1Mb region is obtained from the GENCODE database (Harrow et al., 2012). Monomers containing a number of genes which fall below a preset cutoff are termed as "inactive" or "passive" and are characterized by an effective temperature $T = T_{ph} = 310K$. Monomers possessing a larger number of genes are termed as "active" and assigned an effective temperature $T_a > T_{ph}$. For the **gene expression model**, we infer activities from transcriptome data, using FPKM values from processed RNA-seq output (Consortium et al., 2012). For the **combined model**, we use the same temperature assignments as for the **gene expression model** but, in addition, also take the top 5% of monomers by gene density as inferred from GENCODE, promoting them to a temperature of $T = 12$.

Models for the looping of individual chromosomes

We use Hi-C data on GM12878, NHEK, IMR90, HUVEC and HMEC cells, obtained from data made publicly available by the authors of Ref. (Rao et al., 2014), to represent the effects of long-range looping within a chromosome. We ignore loops smaller than

the 1 Mb scale since these are folded into our description of a single monomer. Across these cell types we have 236 (GM12878), 50 (NHEK), 116 (IMR90), 51 (HUVEC), and 13 (HMEC) loops that are larger than the 1 Mb scale and which our model accounts for. These loops are represented by permanent FENE bonds, with an effective interaction strength that is the same as those of the springs which connect un-looped but connected monomers.

Summary of Analysis

We calculate $S_i(R) = 4\pi R^2 P_i(\vec{R})$, where $P_i(\vec{R})\, dR$ is proportional to the probability of finding a monomer of chromosome i at a radial vector $\vec{R}$ from the origin. For a uniform distribution, $S_i(R) = 4\pi R^2$. We compute $S_i(R)$ for every model chromosome indexed by i. We measure activity in successive radial shells by performing a configurational average over the effective temperature of every monomer in that shell. From these we extract a quantity similar to $S(R)$, but normalize by $4\pi R^2$, so that the quantity plotted in the cut-away sphere representation simply represents the activity at radial distance R. The quantity S(R) measures the DNA density associated with a specific chromosome, across a radial shell at distance R from the nuclear centre, averaged over a large number of nuclei. The quantity S_{CM}(R) measures a similar distribution, but of the chromosome centre of mass. We calculate the distribution of centres of mass of each individual chromosome similarly. To visually examine configurations we color-coded monomers belonging to individual chromosomes.

Calculation of 3D shape of CTs

For each chromosome in our simulation, we draw a 3-d grid across the nucleus with a grid spacing of $0.2 - 0.6$. We represent the monomers as spheres about which the density decays as a gaussian with defined width. Separating these monomers are cylindrical regions. The density about the axis of each cylinder is assume to fall off also as a Gaussian with specified width. The density at any given grid point, associated to a single chromosome, can then be computed by adding up the contributions from all

spherical and cylindrical regions as defined above. Once such a density field is obtained, we can find the surfaces on which it attains a fixed value, the "implicit surface". We adjust the scales governing the decay of the density distribution associated with the monomers and the cylindrical regions separating them as well as the associated constant specifying the implicit surface to optimise geometrical quantities associated with chromosome territories vis a vis experiments. Once fixed, these parameters remain the same for all chromosomes.

Geometric properties of CTs

To calculate the two-dimensional properties of chromosome territories, we use the algorithm of Ref. (Sehgal et al., 2014). To compare our simulation data with data from 2d FISH, we project three-dimensional chromosome territories into the xy plane. We use the ellipticity calculations of (Žunić and Žunić, 2013). In the three-dimensional case, once we associate an implicit surface to a chromosome, that surface can further be triangulated using standard methods, such as the ISOSURFACE command in MATLAB. The total surface area of the chromosomes is obtained by adding the area of these triangles. To calculate the volume of the chromosome we count the number of grid points whose grid density values are more than the given isovalue density c. The asphericity Δ and shape (or prolateness) Σ parameters of a particular chromosome are calculated from the semi-axis lengths a, b, and c of the smallest ellipsoid that encloses all the monomers (Millett et al., 2009; Rawdon et al., 2008). The Khachiyan algorithm is used to find the smallest ellipsoid which encloses all the data points (Todd and Yıldırım, 2007).

Contact probability and contact map

The contact probability is computed using numerical calculations of the contact frequency of monomers of a given chromosome, averaged over a large number of configurational snapshots. When two monomers i and j of the same chromosome are separated in 3-d space by 2.5 units in terms of our scaled unit distance, we assume that

they are in contact. If two monomers are in contact, they are close in distance. However, measures that look at the frequency of contacts will assign a larger frequency to such monomers which are predisposed to be close by.

Author contributions

Conceptualization: GIM; Methodology: GIM, SS and AA; Formal Analysis and Validation: AA, NG and GIM; Investigation: AA and NG; Writing: Original Draft: GIM and AA; Writing - Review and Editing: GIM, AA, SS and NG; Visualization: AA and GIM; Software: AA and NG; Supervision: GIM; Project Administration: GIM

Declaration of Interest

The authors declare no competing interests.

Acknowledgments

We thank Ana Pombo, G.V. Shivashankar, Jean-Francois Joanny, Jacques Prost, Joe Howard, Sriram Ramaswamy, Mehran Kardar, B.J. Rao and Frank Jülicher for useful discussions. We are especially grateful to Kundan Sengupta, Sandhya Koushika and P. Ranjith for detailed input and discussions regarding our manuscript. The work of GIM was supported in its early stages through a DAE-SRC Outstanding Researcher Fellowship. We acknowledge the use of the Annapurna HPC at IMSc.

References

Agrawal, A., Ganai, N., Sengupta, S. and Menon, G. I. (2017). Chromatin as active matter. Journal of Statistical Mechanics: Theory and Experiment *2017*, 014001.

Almassalha, L. M., Bauer, G. M., Wu, W., Cherkezyan, L., Zhang, D., Kendra, A., Gladstein, S., Chandler, J. E., VanDerway, D., Seagle, B.-L. L. et al. (2017). Macrogenomic engineering via modulation of the scaling of chromatin packing density. Nature Biomedical Engineering *1*, 902.

Amitai, A. and Holcman, D. (2017). Polymer physics of nuclear organization and function. Physics Reports *678*, 1–83.

Berezney, R., Malyavantham, K. S., Pliss, A., Bhattacharya, S. and Acharya, R. (2005). Spatio-temporal dynamics of genomic organization and function in the mammalian cell nucleus. Advances in enzyme regulation *45*, 17–26.

Bickmore, W. A. (2013). The Spatial Organization of the Human Genome. Annual Review of Genomics and Human Genetics *14*, 67–84.

Bickmore, W. A. and van Steensel, B. (2013). Genome architecture: domain organization of interphase chromosomes. Cell *152*.

Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., Müller, S., Eils, R., Cremer, C., Speicher, M. R. et al. (2005). Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. PLoS biology *3*, e157.

Boyle, S., Gilchrist, S., Bridger, J. M., Mahy, N. L., Ellis, J. A. and Bickmore, W. A. (2001). The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. Hum Mol Genet *10*.

Chiariello, A. M., Bianco, S., Piccolo, A., Annunziatella, C., Barbieri, M., Pombo, A. and Nicodemi, M. (2015). Polymer models of the organization of chromosomes in the nucleus of cells. Modern Physics Letters B *29*, 1530003.

Chu, F.-Y., Haley, S. C. and Zidovska, A. (2017). On the origin of shape fluctuations of the cell nucleus. Proceedings of the National Academy of Sciences *114*, 10338–10343.

Chubb, J. R., Trcek, T., Shenoy, S. M. and Singer, R. H. (2006). Transcriptional pulsing of a developmental gene. Current biology *16*, 1018–1025.

Consortium, E. P. et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57.

Cook, P. R. and Marenduzzo, D. (2009). Entropic organization of interphase chromosomes. The Journal of cell biology *186*, 825–834.

Cremer, T. and Cremer, M. (2010). Chromosome territories. Cold Spring Harb Perspect Biol *2*.

Croft, J. A., Bridger, J. M., Boyle, S., Perry, P., Teague, P. and Bickmore, W. A. (1999). Differences in the localization and morphology of chromosomes in the human nucleus. J Cell Biol *145*.

de Lanerolle, P. (2012). Nuclear actin and myosins at a glance. Journal of Cell Science *125*, 4945–4949.

Di Pierro, M., Cheng, R. R., Lieberman Aiden, E., Wolynes, P. G. and Onuchic, J. N. (2017). De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture. Proceedings of the National Academy of Sciences *114*, 12126–12131.

Di Pierro, M., Zhang, B., Aiden, E. L., Wolynes, P. G. and Onuchic, J. N. (2016). Transferable model for chromosome architecture. Proceedings of the National Academy of Sciences *113*, 12168–12173.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S. and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*.

Dyer, K., Canfield, T. and Gartler, S. (1989). Molecular cytological differentiation of active from inactive X domains in interphase: implications for X chromosome inactivation. Cytogenetic and Genome Research *50*, 116–120.

Eils, R., Dietzel, S., Bertin, E., Schröck, E., Speicher, M. R., Ried, T., Robert-Nicoud, M., Cremer, C. and Cremer, T. (1996). Three-dimensional reconstruction of painted human interphase chromosomes: active and inactive X chromosome territories have

similar volumes but differ in shape and surface structure. The Journal of cell biology *135*, 1427–1440.

Fedorova, E. and Zink, D. (2009). Nuclear genome organization: common themes and individual patterns. Current opinion in genetics & development *19*, 166–171.

Flaus, A. and Owen-Hughes, T. (2011). Mechanisms for ATP-dependent chromatin remodelling: the means to the end. The FEBS journal *278*, 3579–3595.

Fodor, É., Guo, M., Gov, N., Visco, P., Weitz, D. and van Wijland, F. (2015). Activity-driven fluctuations in living cells. EPL (Europhysics Letters) *110*, 48005.

Fraser, J., Williamson, I., Bickmore, W. A. and Dostie, J. (2015). An overview of genome organization and how we got there: from FISH to Hi-C. Microbiology and Molecular Biology Reviews *79*, 347–372.

Fraser, P. and Bickmore, W. (2007). Nuclear organization of the genome and the potential for gene regulation. Nature *447*.

Ganai, N., Sengupta, S. and Menon, G. I. (2014). Chromosome positioning from activity-based segregation. Nucleic Acids Research *42*, 4145–4159.

Ghosh, S. K. and Jost, D. (2017). How epigenome drives chromatin folding and dynamics, insights from efficient coarse-grained models of chromosomes. bioRxiv *0*, 1.

Haddad, N., Jost, D. and Vaillant, C. (2017). Perspectives: using polymer modeling to understand the formation and function of nuclear compartments. Chromosome Research *25*, 35–50.

Hameed, F. M., Rao, M. and Shivashankar, G. (2012). Dynamics of passive and active particles in the cell nucleus. PLoS One *7*, e45843.

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S. et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome research *22*, 1760–1774.

Heermann, D. W., Jerabek, H., Liu, L. and Li, Y. (2012). A model for the 3D chromatin architecture of pro and eukaryotes. Methods *58*, 307–314.

Imakaev, M. V., Fudenberg, G. and Mirny, L. A. (2015). Modeling chromosomes: Beyond pretty pictures. FEBS Letters *589*, 3031 – 3036.

Jackson, D. A. and Pombo, A. (1998). Replicon clusters are stable units of chromosome structure: evidence that nuclear organization contributes to the efficient activation and propagation of S phase in human cells. The Journal of cell biology *140*, 1285–1295.

Jégu, T., Aeby, E. and Lee, J. T. (2017). The X chromosome in space. Nature Reviews Genetics *18*, 377–389.

Jost, D., Carrivain, P., Cavalli, G. and Vaillant, C. (2014). Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains. Nucleic acids research *42*, 9553–9561.

Jost, D., Vaillant, C. and Meister, P. (2017). Coupling 1D modifications and 3D nuclear organization: data, models and function. Current opinion in cell biology *44*, 20–27.

Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F. and Chen, L. (2011). Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat Biotechnol *30*.

Kang, H., Yoon, Y.-G., Thirumalai, D. and Hyeon, C. (2015). Confinement-induced glassy dynamics in a model for chromosome organization. Physical review letters *115*, 198102.

Khalil, A., Grant, J., Caddle, L., Atzema, E., Mills, K. and Arnéodo, A. (2007). Chromosome territories have a highly nonspherical morphology and nonrandom positioning. Chromosome research *15*, 899–916.

Kölbl, A. C., Weigl, D., Mulaw, M., Thormeyer, T., Bohlander, S. K., Cremer, T. and Dietzel, S. (2012). The radial nuclear positioning of genes correlates with features of megabase-sized chromatin domains. Chromosome Research *20*, 735–752.

Kreth, G., Finsterle, J., Von Hase, J., Cremer, M. and Cremer, C. (2004). Radial arrangement of chromosome territories in human cell nuclei: a computer model approach based on gene density indicates a probabilistic global positioning code. Biophysical journal *86*, 2803–2812.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S. and Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*.

Maharana, S., Iyer, K. V., Jain, N., Nagarajan, M., Wang, Y. and Shivashankar, G. (2016). Chromosome interminglingthe physical basis of chromosome organization in differentiated cells. Nucleic acids research *44*, 5148–5160.

Marchetti, M. C., Joanny, J., Ramaswamy, S., Liverpool, T., Prost, J., Rao, M. and Simha, R. A. (2013). Hydrodynamics of soft active matter. Reviews of Modern Physics *85*, 1143.

Marti-Renom, M. A. and Mirny, L. A. (2011). Bridging the resolution gap in structural modeling of 3D genome organization. PLoS Comput Biol *7*.

Mattout, A., Cabianca, D. S. and Gasser, S. M. (2015). Chromatin states and nuclear organization in development a view from the nuclear lamina. Genome Biology *16*, 1–15.

Meaburn, K. J. and Misteli, T. (2007). Cell biology: chromosome territories. Nature *445*.

Menon, G. I. (2010). Active matter. In Rheology of complex Fluids pp. 193–218. Springer.

Millett, K. C., Plunkett, P., Piatek, M., Rawdon, E. J. and Stasiak, A. (2009). Effect of knotting on polymer shapes and their enveloping ellipsoids. The Journal of chemical physics *130*, 04B623.

Mirny, L. A. (2011). The fractal globule as a model of chromatin architecture in the cell. Chromosome Res *19*.

Murmann, A. E., Gao, J., Encinosa, M., Gautier, M., Peter, M. E., Eils, R., Lichter, P. and Rowley, J. D. (2005). Local gene density predicts the spatial position of genetic loci in the interphase nucleus. Experimental cell research *311*, 14–26.

Needleman, D. and Dogic, Z. (2017). Active matter at the interface between materials science and cell biology. Nature Reviews Materials *2*, 17048.

Németh, A. and Längst, G. (2011). Genome organization in and around the nucleolus. Trends in genetics *27*, 149–156.

Odenheimer, J., Kreth, G. and Heermann, D. W. (2005). Dynamic simulation of active/inactive chromatin domains. Journal of biological physics *31*, 351–363.

Plimpton, S., Crozier, P. and Thompson, A. (2007). LAMMPS-large-scale atomic/molecular massively parallel simulator. Sandia National Laboratories *18*.

Prost, J., Jülicher, F. and Joanny, J.-F. (2015). Active gel physics. Nature Physics *11*, 111.

Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D. and Robinson, J. T. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *59*.

Rawdon, E. J., Kern, J. C., Piatek, M., Plunkett, P., Stasiak, A. and Millett, K. C. (2008). Effect of knotting on the shape of polymers. Macromolecules *41*, 8281–8287.

Sanborn, A. L., Rao, S. S., Huang, S.-C., Durand, N. C., Huntley, M. H., Jewett, A. I., Bochkov, I. D., Chinnappan, D., Cutkosky, A., Li, J. 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*, E6456–E6465.

Sehgal, N., Fritz, A. J., Morris, K., Torres, I., Chen, Z., Xu, J. and Berezney, R. (2014). Gene density and chromosome territory shape. Chromosoma *123*, 499–513.

Stillinger, F. H. (1976). Phase transitions in the Gaussian core system. The Journal of Chemical Physics *65*, 3968–3974.

Sun, H. B., Shen, J. and Yokota, H. (2000). Size-dependent positioning of human chromosomes in interphase nuclei. Biophysical journal *79*, 184–190.

Takizawa, T., Meaburn, K. J. and Misteli, T. (2008). The meaning of gene positioning. Cell *135*, 9–13.

Tark-Dame, M., van Driel, R. and Heermann, D. W. (2011). Chromatin folding–from biology to polymer models and back. J Cell Sci *124*, 839–845.

Therizols, P., Illingworth, R. S., Courilleau, C., Boyle, S., Wood, A. J. and Bickmore, W. A. (2014). Chromatin decondensation is sufficient to alter nuclear organization in embryonic stem cells. Science. *346*.

Tiana, G., Amitai, A., Pollex, T., Piolot, T., Holcman, D., Heard, E. and Giorgetti, L. (2016). Structural fluctuations of the chromatin fiber within topologically associating domains. Biophysical journal *110*, 1234–1245.

Tjong, H., Li, W., Kalhor, R., Dai, C., Hao, S., Gong, K., Zhou, Y., Li, H., Zhou, X. J., Le Gros, M. A. et al. (2016). Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. Proceedings of the National Academy of Sciences *113*, E1663–E1672.

Todd, M. J. and Yıldırım, E. A. (2007). On Khachiyan's algorithm for the computation of minimum-volume enclosing ellipsoids. Discrete Applied Mathematics *155*, 1731–1744.

Vasquez, P. A. and Bloom, K. (2014). Polymer models of interphase chromosomes. Nucleus *5*, 376–390.

Wang, S. and Wolynes, P. G. (2011). Communication: Effective temperature and glassy dynamics of active matter.

Weber, S. C., Spakowitz, A. J. and Theriot, J. A. (2012). Nonthermal ATP-dependent fluctuations contribute to the in vivo motion of chromosomal loci. Proceedings of the National Academy of Sciences *109*, 7338–7343.

Zhang, B. and Wolynes, P. G. (2017). Genomic energy landscapes. Biophysical journal *112*, 427–433.

Zidovska, A., Weitz, D. A. and Mitchison, T. J. (2013). Micron-scale coherence in interphase chromatin dynamics. Proceedings of the National Academy of Sciences *110*, 15555–15560.

Žunić, D. and Žunić, J. (2013). Shape ellipticity based on the first Hu moment invariant. Information Processing Letters *113*, 807–810.

Supplementary Information

A First-principles Approach to Large-scale Nuclear Architecture

Ankit Agrawal[1,2], Nirmalendu Ganai[3,4], Surajit Sengupta[5], and Gautam I Menon[1,2,*]

[1]Computational Biology Group, The Institute of Mathematical Sciences, C.I.T. Campus, Taramani, Chennai 600 013, India
[2] Homi Bhabha National Institute, Training School Complex, Anushaktinagar, Mumbai 400 094, India
[3] Department of Physics, Nabadwip Vidyasagar College, Nabadwip, Nadia 741 302, India
[4]Theoretical Soft Matter and Biophysics, Institute of Complex Systems, Forschungszentrum Jülich, D-52425 Jülich, Germany
[5]TIFR Centre for Interdisciplinary Sciences, 36/P Gopanpally Village, Serilingampally Mandal, Ranga Reddy District, Hyderabad 500 107, India
[*]Correspondence should be addressed to G.I.M., (menon@imsc.res.in)

S1 Model Overview

We model individual chromosomes are modelled as polymer chains, composed of spherical monomers representing consecutive 1Mb sections of each chromosome. There are 23 pairs of chromosomes, describing a human diploid female (XX) cell. These model chromosomes are confined within a spherical shell representing the nucleus. The monomers are connected through nonlinear springs and interact via a finitely-extensible nonlinear elastic (FENE) interaction (Kremer and Grest, 1990). Individual monomers are self-repelling and the total number of monomers is 6086 for a diploid XX human cell. Monomers experience forces from other monomers, arising from both bonded and non-bonded interactions. Additionally, each monomer experiences random forces arising from thermal as well as active fluctuations. We treat such active noise as analogous to thermal noise, drawing particular realisations of the noise from a Gaussian distribution with zero mean and a variance set by the effective temperature.

To represent activity we assign an active temperature to monomers in three ways. In the simplest model, the "gene density" model, the temperature assigned to each monomer reflects the gene density associated with the specific region of the chromosome associated to that monomer. The gene content of each such 1Mb region is obtained from the GENCODE database (Harrow et al., 2012). We label each monomer as active or inactive, associating one of two temperatures ($T = 1, 12$) to the monomer depending on its label. Temperatures are scaled in units of physiological temperature $T_{ph} = 310K$.

A second model, the "gene expression" model, assumes that the temperature assigned to each monomer is proportional to the amount of RNA transcript generated across that region of chromosome. Transcript levels then provide a proxy for the intensity of active processes locally. We infer these levels using transcriptome data, as described below. Both the gene expression and gene density models fit specific aspects of the experimental data well. However, a model description which appears to provide the most comprehensive fits to the data combines features of both these models. The model used in this paper, the "combined model", bases itself largely on the gene expression model but also assigns high activity to a fraction of monomers with the highest values of gene density.

An important part of our model is the incorporation of existing prior information regarding the looping of chromosomes. Here, we use data from Hi-C experiments to assign permanent loops to our model chromosomes. We simulate our system using standard Brownian dynamics methods. After verifying that the system has achieved a non-equilibrium steady state, we compute all properties of interest, including the distribution of DNA density and of chromosome centre of mass, territorial organisation, shape statistics and spatial distance maps from which we can infer potential contacts.

S1.1 Interactions of model chromosomes

Our model chromosomes occupy the interior of a spherical shell. The radius of this shell is R_0, which we take to be 17.2 in the reduced units we derive below. The interaction between neighbouring monomers (labeled as $i, i+1$, with position coordinates $\mathbf{r}_i, \mathbf{r}_{i+1}$) is of the FENE form

$$V_{neighbour\ monomers}(\mathbf{r}_i, \mathbf{r}_{i+1}) = -\frac{1}{2} K r_0^2 \ln \left[1 - \left(\frac{r}{r_0} \right)^2 \right], \tag{1}$$

where K is a spring constant and r_0 is the maximum length of the bond; in contrast, our earlier work used the simpler harmonic spring interaction (Ganai et al., 2014). These monomers further interact with (non-neighbouring) monomers via a Gaussian interaction, the Gaussian core potential used earlier to model polymer brushes (Stillinger, 1976)

$$V_{monomer-monomer}(\mathbf{r}_i, \mathbf{r}_j) = V_0 \exp\left(-B|\mathbf{r}_i - \mathbf{r}_j|^2\right), \qquad |\mathbf{r}_i - \mathbf{r}_j| < r_{cut} \tag{2}$$

The effective pair potential at zero separation, V_0, is chosen to be of order $k_B T_{ph}$, with k_B the Boltzmann constant and T_{ph} the physiological temperature: $B = 1.0$,

$V_0 = 1.5$ and $r_{cut} = 3.5$. The interaction between each monomer and the confining sphere vanishes if the monomer centre falls within the sphere. Outside the sphere and within a cutoff r_c, the monomer experiences a Lennard-Jones potential that diverges as the distance to the cutoff is reduced. The parameters are $\epsilon = 250$, $\sigma = 1$ and $r_c = 1$ for this Lennard-Jones potential. We take $r_0 = 10$ and $K = 4.17$, for both bonded neighbors and non-bonded monomers connected by long range loops.

S1.2 Simulation Methodology

Our numerical evolution of the system of monomers adapts the widely-used LAMMPS code implementing Brownian dynamics (Plimpton et al., 2007) for a polydisperse polymer system with a FENE interaction between monomers and a monomer-specific effective temperature. The effective temperature is incorporated as a local monomer-dependent effective temperature. For each monomer, LAMPPS applies a Langevin thermostat, via the following over-damped equation of motion,

$$\zeta \frac{d\mathbf{r}_i}{dt} = \mathbf{F}_i + \eta_i \tag{3}$$

where $\mathbf{r}_i$ represents the location of the ith monomer, ζ is a drag coefficient, $\mathbf{F}_i$ accounts for all non-stochastic forces acting on the monomer and η_i represents stochastic forces (gaussian, with vanishing cross-correlations) arising from both active and thermal fluctuations. The noise is assumed Gaussian distributed, with cross-correlations vanishing at all times irrespective of monomer labels. The diagonal correlators, at equal times and for the same monomer, are non-zero and obtained from $\langle \eta_i^x(t)\eta_j^x(t')\rangle = \langle \eta_i^y(t)\eta_j^y(t') = \langle \eta_i^z(t)\eta_j^z(t')\rangle = 2k_B T_i \zeta \delta_{ij}\delta(t-t')$. Here T_i is an "effective" temperature associated to each monomer, reflecting its local level of activity. We represent each of the components of $\eta_i/\sqrt{\zeta}$ as the product of a Gaussian random number with zero mean and unit variance with the quantity $\sqrt{2k_B T_i/\zeta}$. In thermal equilibrium, we have $T_i = T_{eq}$ for all monomers. The largest number of monomers (249) are found in Chromosome 1 and the smallest (47) in Chromosome 21.

In comparison to the model in Ref. (Ganai et al., 2014), the model described here has 12 fewer monomers, a consequence of differences between the GeneCards and GENCODE databases.

S1.3 Units and Normalization

Following Refs. (Kreth et al., 2004; Ganai et al., 2014), we consider a chromatin volume fraction $0.1 \leq \phi \leq 0.2$. The monomer is assumed to have a diameter $d \simeq 500nm$; the equilibrium domain separation is $\ell_0 \simeq 600nm$. Both these quantities accord with computed Kuhn lengths of $\approx 300nm$ (Kreth et al., 2004; Rosa and Everaers, 2008). Assuming that the radius of the nucleus is $R_0 \simeq 8.6\mu m$ yields a packing fraction of $\phi \simeq 0.15$. We ignore the marginal differences in nuclear volume across cell types; such volumes differ by at most a factor of 1.5 for the cell types we consider. We scale all lengths in units of d and measure energies in units of $k_B T_{eq}$. All chromosomes are fairly tightly confined to R_0. We can choose units of time (τ) such that $\zeta = 1$. We can approximate the value of ζ appropriate for this calculation from the

Stokes relation: $\zeta \simeq 6\pi\eta_s R$ where R is the hydrodynamic radius appropriate to the monomer size. Assuming that the appropriate value of the viscosity at such scales is $\eta_s \sim 10\eta_w$, with η_w the viscosity of water $(8.9\times10^{-4} Pa\cdot s)$, its numerical value is then $\zeta = 8.38\times10^{-8} N\cdot s/m$. With this choice, τ is then $(500.0)^2\zeta/6k_BT \simeq 8.16\times10^{-1}$s. Since $\tau \approx 10^{-1}s$; our choice of time-step of 0.001 thus corresponds to real-time evolution by 10^{-4}s.

S1.4 Deriving effective temperatures

S1.4.1 Gene density model

In the gene density model, the gene content of each 1Mb region is obtained from the GENCODE database (Harrow et al., 2012). Monomers containing a number of genes which fall below a preset cutoff are termed as "inactive" or "passive" and are characterized by an effective temperature $T = T_{ph} = 310K$. Monomers possessing a larger number of genes are termed as "active" and assigned an effective temperature $T_a > T_{ph}$, argued for thus: The hydrolysis of a single ATP molecule yields approximately $20k_BT_{ph}$ of energy, some part of which is transduced into local mechanical work with the remaining part begin dissipated. Thus, the characteristic energy scales associated with such hydrolysis events, modelled in terms of an effective temperature, must be some reasonable fraction of this scale. We have experimented with several different choices of T_a as well as the cutoff, finding that a relatively small spread between physiological and active temperatures is sufficient to generate the activity-dependent biophysical structuring this paper discusses. For concreteness, here we take the maximum value for the active temperature to be $T_a = 12$, where T_a is measured in units of T_{ph}. (Measurements of the diffusion constants of individual gene loci in bacteria and yeast provide evidence for a similar spread in local "effective" temperatures, as inferred from an Einstein relation, and that a related variation in local active forces has been suggested to explain observations from colloidal micro-rheology in the nucleus (Zidovska et al., 2013; Weber et al., 2012; Hameed et al., 2012))

S1.4.2 Gene expression model

For the gene expression model, we infer activities from transcriptome data. We downloaded processed RNA-seq data from the ENCODE project website (Consortium et al., 2012). The data provides FPKM (Fragments Per Kilobase of transcript per Million mapped read, with fragment referring to a pair of reads for paired-end data) values that quantify transcripts generated across the human genome. We used five cell types GM12878, HMEC, HUVEC, IMR90 and NHEK for our analysis. We consider all genes whose FPKM value lies above a specified cutoff to reduce noise in the data for each cell type. We then summed the FPKM value for all these genes whose chromosome position (mid position of start and end coordinate of a gene) lies within our 1 Mb interval to assign an activity value to that monomer. We went on to assign effective temperatures proportional to such activity values using a derivative cutoff method. We examined the expression data discussed above, sorted it in increasing order of expression, and then took a numerical derivative of this data, using the $diff$ function

of MATLAB. We set a cutoff of 0.2 for this slope but our results were largely insensitive to where this cutoff was chosen. Given the generic shape of the plot for the sorted activity, which has two regions in which activity increases sharply separated by a plateau region, this procedure automatically yields two cutoff lines and three demarcated regions. Monomers in the plateau region are assigned a constant value of active temperature, between T_{ph} and the maximum value of the active temperature $12T_{ph}$.

For each cell type, given similar curves of sorted expression value, such cutoffs on the data represent the effects of activity on each monomer. We can translate this into an active temperature. In (Figure 1H), for each cell type, we show these cutoff lines. We assign the lowest temperatures to monomers whose activity falls below the value it takes in the plateau region. Monomers associated with the plateau are assigned a common temperature of 6. Finally, monomers with the highest expression values and thus the highest activity, are assigned a temperature which interpolates, in units of 1, between 7 and 12. The effects of variation of activity are strongest for these monomers, as is reasonable since activity increases steeply in this region.

S1.4.3 Combined model

For the combined model, we use the temperature assignments indicated above. However, in addition, we also take the top 5% of monomers by gene density as inferred from GENCODE, promoting them to a temperature of $T = 12$.

S1.5 Summary of Analysis

Our simulations are run for 10^7 time steps, with around 4×10^6 steps discarded to ensure adequate equilibration. All data are averaged over at least 25 independently initialised configurations, with each initial configuration contributing 6000 independent measurements as the simulation proceeds. We verified that the same steady state properties were achieved irrespective of initial (random) configuration. Since the probability of finding a chromosome at a radial separation $\mathbf{r}$ from the origin depends only on the modulus of $\mathbf{r}$, *i.e.* $|\mathbf{r}| \equiv r$, we calculate the probability of finding a monomer belonging to a specific chromosome at a radial distance from the origin, for each chromosome. The main text describes the different distribution functions we calculate.

S1.6 Calculation of 3D shape of CTs

We begin by drawing a 3-d grid across the nucleus with a coarse grid spacing of 0.5 in our scaled units. We calculate the DNA density associated with our polymer model for chromosomes on this grid as described below. Chromosome densities are then defined at the grid points for each simulation configuration. We represent chromosome configurations in the following ways. First, individual monomers are considered as spheres about which the density decays as a normalized gaussian in the radial variable. The characteristic scale of the Gaussian is set by the length scale R_s. We experimented with various choices of R_s to see what might best represent the data. Between these monomers, we assume that the DNA configuration can be described as a cylinder with fixed radius, with the DNA density about the axis of each cylinder assumed to fall off

also as a Gaussian with specified width R_c. We have varied R_c to obtain optimal fits as well, but finally fixed on $R_c = R_s = 2\sigma$ as an optimal choice for all chromosomes across all configurations and all cell types.

The density at any given grid point can then be computed by adding up the contributions from all spherical and cylindrical regions associated to a chromosome. We then interpolate these density values to a smaller grid, typically of spacing 0.17 using MATLAB's INTERP3 function. Once such a density field is obtained, we can find the surfaces on which it attains a fixed value, the "implicit surface".

Our calculation of the density at a given grid point proceeds as follows. Consider the location of the grid point O_j, where j indexes the specific grid point. Let n be the length of chromosome C^k and denote the ith monomer of the kth chromosome as C_i^k. The visual computational of such calculation is shown in (Figure S16). The two consecutive monomer C_i^k and C_{i+1}^k of k chromosome is represented by small circles in, and lines representing the centreline of the associated cylindrical regions. O_j is the grid point where we have to calculate the spherical density D_{sphere} which is contributed by mentioned two monomers shown in (Figure S16A) and similarly contribution by their centerline in form of cylindrical density $D_{cylinder}$ is shown in (Figure S16B). The density at the grid point O_j arising from chromosome k is then computed from the following formula and also shown in (Figure S16C).

$$G(O_j) = \max\left(\sum_{i=1}^{n-1} \exp\left[-\frac{d^{\,2}}{2R_c^2}\right], \sum_{i=1}^{n} \exp\left[-\frac{||\overrightarrow{O_jC_i^k}||^2}{2R_s^2}\right]\right) \begin{cases} \text{if } (\theta > \pi/2)|(\psi > \pi/2) \\ \quad d = 1000 \\ \text{otherwise} \\ \quad d = ||\overrightarrow{O_jC_i^k}|| \sin\theta \end{cases} \tag{4}$$

Here $\cos\theta = \frac{\overrightarrow{O_jC_i^k} \cdot \overrightarrow{C_{i+1}^kC_i^k}}{||\overrightarrow{O_jC_i^k}||||\overrightarrow{C_{i+1}^kC_i^k}||}$ and $\cos\psi = \frac{\overrightarrow{C_{i+1}^kO_j} \cdot \overrightarrow{C_{i+1}^kC_i^k}}{||\overrightarrow{C_{i+1}^kO_j}||||\overrightarrow{C_{i+1}^kC_i^k}||}$. These conditions ensure that (a) the contribution of the associated cylinder to the density at any grid point is cylindrically symmetric about the line joining neighbouring monomers and that (b) the contributions from the spherical regions is also accounted for.

After the density values at all grid points is computed, the isosurface command from MATLAB is used to draw the implicit surface $(F(x, y, z) = c)$ for the chromosome given a density isovalue c. A smaller value of c yields a loose cloud-like surface around chromosomes while larger values of c gives tighter, more well-defined surfaces around them. We chose a value of c such that the chromosome territories it yields are visually equivalent to those obtained in experiments based on similar isosurface representations of experimental FISH data. The 3d surface representation of such chromosome is shown in (Figure 1E). This computation is used in the fits to the experimental data shown in (Figure 5C).

S1.7 Calculation of 2D shape of CTs

To calculate the two-dimensional properties associated to projected chromosome territories. To compare our simulation data with data from 2d FISH as obtained in those experiments and other similar ones, we project three-dimensional chromosome territories

into the xy plane for specificity, noting that averaging over configurations which are rotationally symmetric implies that all projections should be equivalent. A schematic illustrating this is shown in (Figure 1G). The calculations of Ref. (Sehgal et al., 2014; Žunić and Žunić, 2013) rely on the computation of the geometric moments M of the image I. For calculating the ellipticity and regularity of 2d CTs image, we first normalize, scale and rotate the complementary binary image I of our projected chromosome territories such that

1. The area of shape is 1
2. The centroid of shape coincides with the origin and
3. The orientation of shape is 0 (implying that the long axis is parallel to the x-axis).

The scaled moments then are:

$$M_{pq} = \sum_x \sum_y x^p y^q I(x, y) \tag{5a}$$

Where I(x,y) is the pixel intensitites of the grayscale image and $\{p, q\}$ can each be drawn from $0 \ldots \infty$, and refer to the order of moments. We choose each of $\{p, q\}$ from 0,1,2 for specificity.

Given the moments, we compute the ellipticity ϵ from

$$\epsilon = \frac{1}{2} \times \frac{1}{a^2 \times M_{20} + b^2 \times M_{02}} \tag{5b}$$

where a and b are given by

$$a = \sqrt{2\pi^2 L - \pi \times \sqrt{4\pi^2 L^2 - 1}}, b = \sqrt{2\pi^2 L + \pi \times \sqrt{4\pi^2 L^2 - 1}} \tag{5c}$$

and $L = M_{20} + M_{02}$

As ϵ tends to 1, projected chromosome territories resemble ellipses. To calculate the regularity of these projected territories, the area ratio of each CT over its convex hull is determined, using the MATLAB functions BWAREA and BWCONVHULL. If a chromosome is regular, this ratio should be close to 1. As the irregularity in projected CTs increases, this ratio will decrease.

S1.8 Methods for calculating 3d shape statistics for chromosome territories

We use two methods for finding the shape statistics of chromosomes. In the ellipsoidal fit method, the center, rotation, and principal radii of the smallest ellipsoid which encloses all the data points of a polymer are calculated using Khachiyan algorithm (Todd and Yıldırım, 2007). An alternative grid-based method is described in the subsection **Calculation of 2D shape of CTs**. The ellipsoid fit method is an easy and fast method for finding the approximate shapes of the polymer, but is less accurate than the grid-based method. On the other hand, the grid-based method is computationally

expensive. In our work, we calculate the shapes of 2d and 3d territories using the grid-based method but use the simpler ellipsoidal fit method for the shape parameters Δ and Σ described below.

The difference in the computation of chromosome territory properties between the grid-based method and the ellipsoid fit method is illustrated in (Figure S17).

S1.9 Asphericity Δ and shape Σ parameter calculation of chromosomes

The asphericity Δ and the shape parameter Σ of individual chromosomes can be calculated from the 3 semi-principal radii (a,b,c) of the ellipsoid obtained for each ellipsoidal fit to a chromosome territory. The asphericity Δ parameter given in Eq. 6a characterizes the average deviation of the chain conformation from spherical symmetry (Millett et al., 2009; Rawdon et al., 2008). The shape Σ parameter measures the prolateness or oblateness of chromosomes and is defined in Eq. 6b.

$$\Delta(a,b,c) = \frac{(a-b)^2 + (b-c)^2 + (c-a)^2}{2(a+b+c)^2} \tag{6a}$$

$$\Sigma(a,b,c) = \frac{(2a-b-c)(2b-a-c)(2c-a-b)}{2(a^2+b^2+c^2-ab-ac-bc)^{3/2}} \tag{6b}$$

The parameter Δ is bounded in the regime $0 \leq \Delta \leq 1$. The Δ value is 0 for the perfect sphere when $(a = b = c)$ and 1 for rod shape when $(b = c = 0)$. The shape or prolateness parameter Σ is bounded by $-1 \leq \Sigma \leq 1$. The Σ is -1 for perfect oblate shapes, e.g. when $(a = b > c)$ and 1 for perfectly prolate shapes, e.g. when $(a > b = c)$, providing a useful index for chromosome shapes.

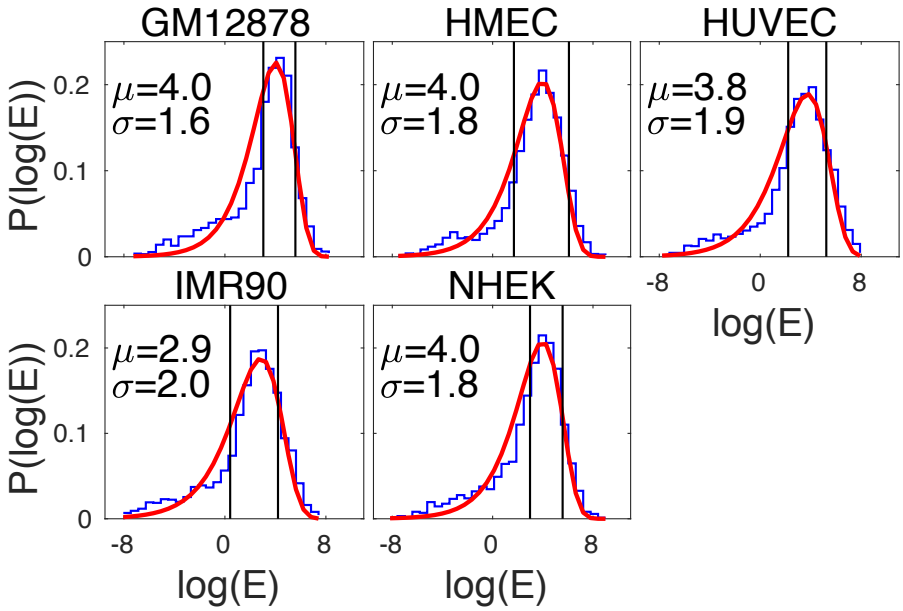

Figure S1: **Distribution of expression values derived from transcriptome data** The histogram of the log of gene expression values as obtained from transcriptome data for 5 cell types is shown in blue. Each cell type name is provided at the top of each subfigure. The sub-figures illustrate a fit of histogram values to an extreme value distribution, the Gumbel distribution, shown in red. The Gumbel distribution for a single variable x takes the form $f(x;\mu,\sigma) = \frac{1}{\sigma}\exp(\frac{x-\mu}{\sigma}).\exp(-\exp(\frac{x-\mu}{\sigma}))$. The fitting parameters μ and σ for the Gumbel distribution in each case are provided in each subfigure. Two black vertical lines, derived from the analysis that led to (Figure1H), demarcate the histogram into 3 regions. The distribution has a long tail towards low expression values and a more sharply decaying right tail towards high expression values.

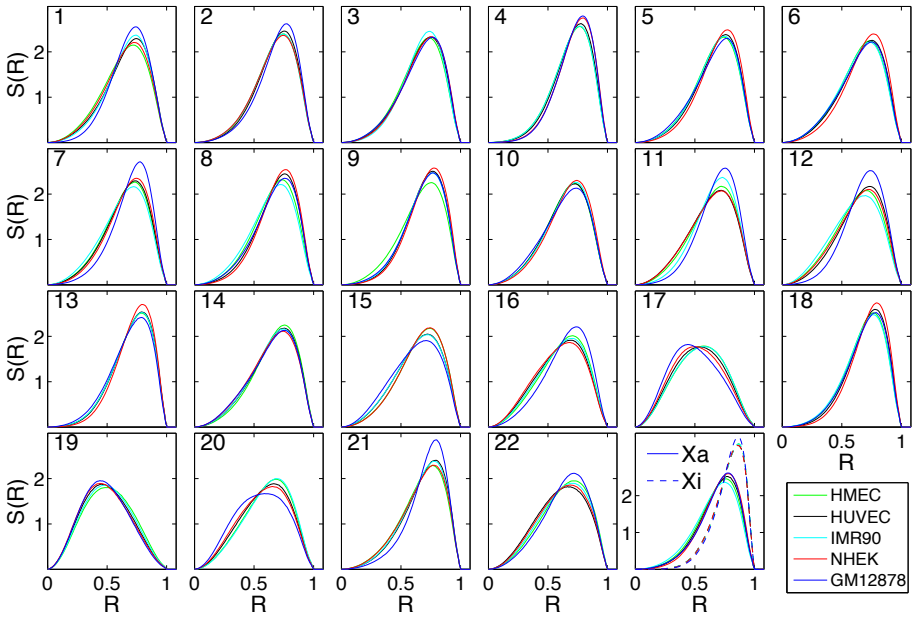

Figure S2: **Computed Distribution functions S(R) for all simulated chromosomes across 5 cell types.** Distribution functions S(R) for each simulated chromosomes for GM12878 (blue), HMEC(green), HUVEC(black), IMR90(cyan) and NHEK(red) is shown. The active chromosome Xa and inactive chromosome Xi are shown with solid and dashed lines. Chromosome numbers are mentioned in the left upper corner of each subfigure.

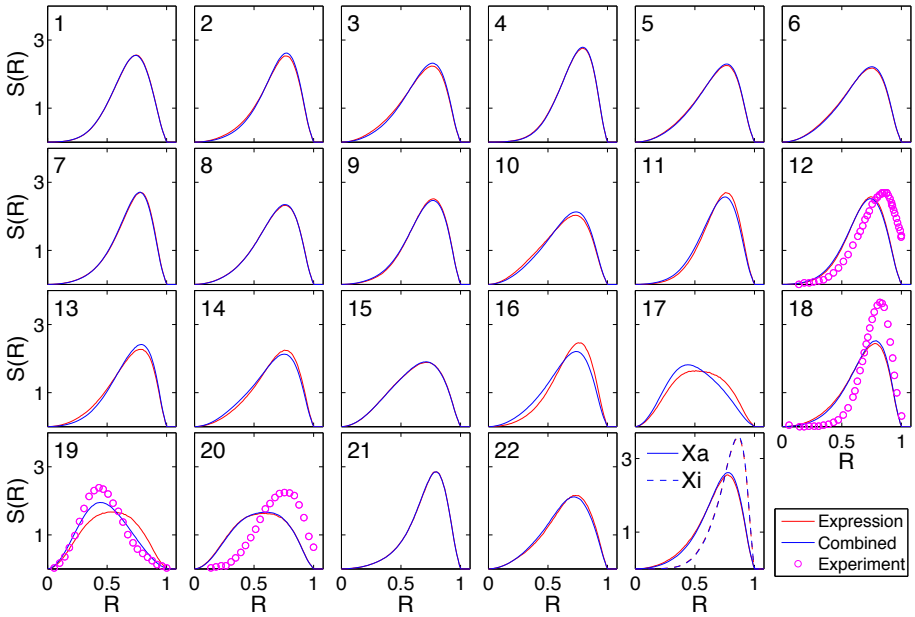

Figure S3: **Distribution functions S(R) for all chromosomes, within the "gene expression" and "combined" models.** S(R) for each chromosome within the gene expression (red) and the combined model (blue) are shown. Experimental data from Ref. (Kreth et al., 2004) for Chromosomes 12, 18, 19 and 20 are shown as magenta-coloured circles.

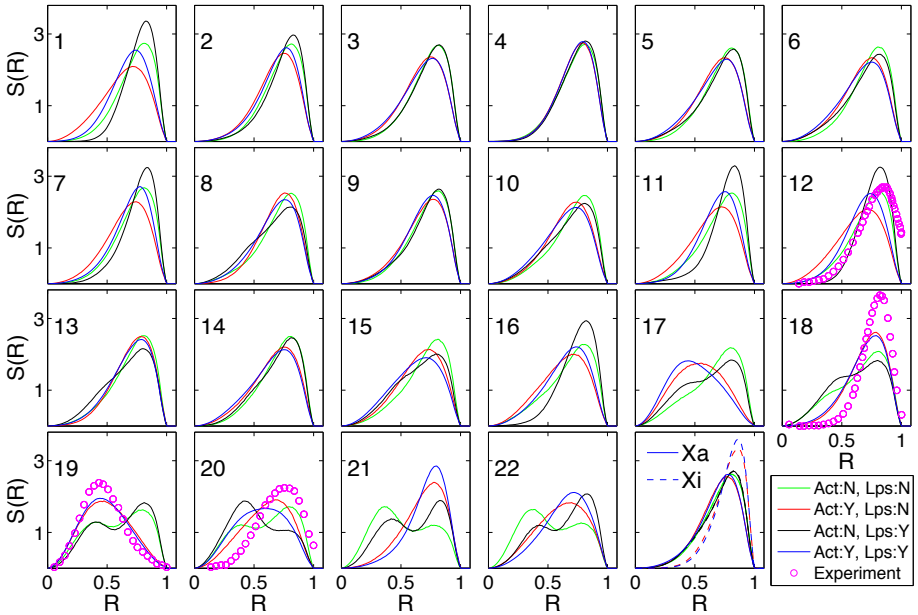

Figure S4: **Distribution functions S(R) for all chromosomes with different combinations of activity and loops.** The monomer density distribution for each chromosome is shown for the models mentioned below. All cases involving activity are shown for the combined model. **Act:N, Lps:N** Both activity and loops are switched off, with all monomers at the same effective temperature of T = 1, shown in green; **Act:Y, Lps:N** Activity is present, implying an inhomogeneous distribution of temperatures, but loops are switched off, shown in red; **Act:N, Lps:Y** Activity is absent but loops are present, shown in black; **Act:Y, Lps:Y** Both activity and loops are present, shown in blue color. This is the original "combined model", also shown in (Figures 3A and 3B) and (Figure S3); The experimental data for chromosomes 12, 18, 19 and 20 are shown as magenta-colored circles, using data from Ref. (Kreth et al., 2004).

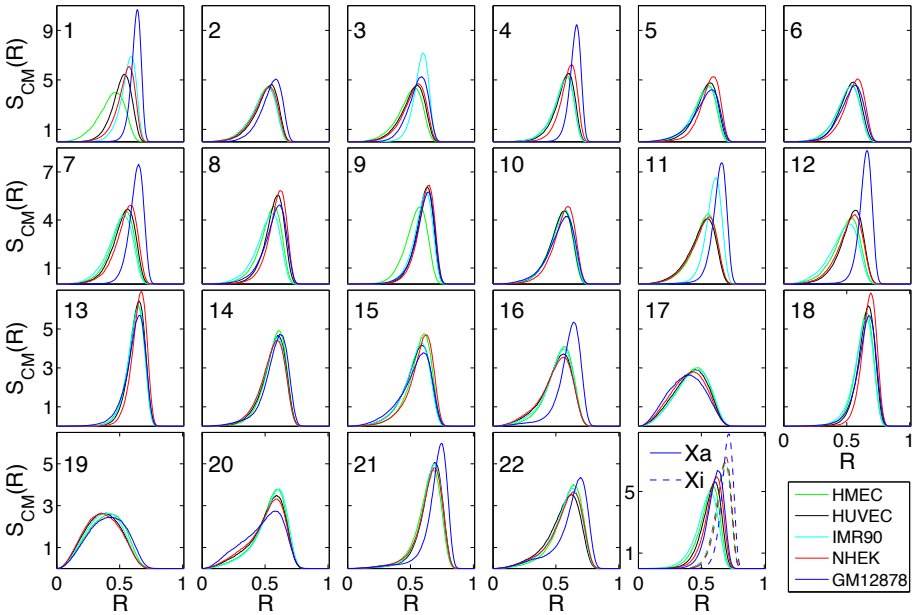

Figure S5: **Centre of mass distributions S_{CM}(R) for all simulated chromosome in 5 cell types.** The centre of mass distribution S_{CM}(R) for each simulated chromosome for GM12878(blue), HMEC(green), HUVEC(black), IMR90(cyan) and NHEK(red) is shown.

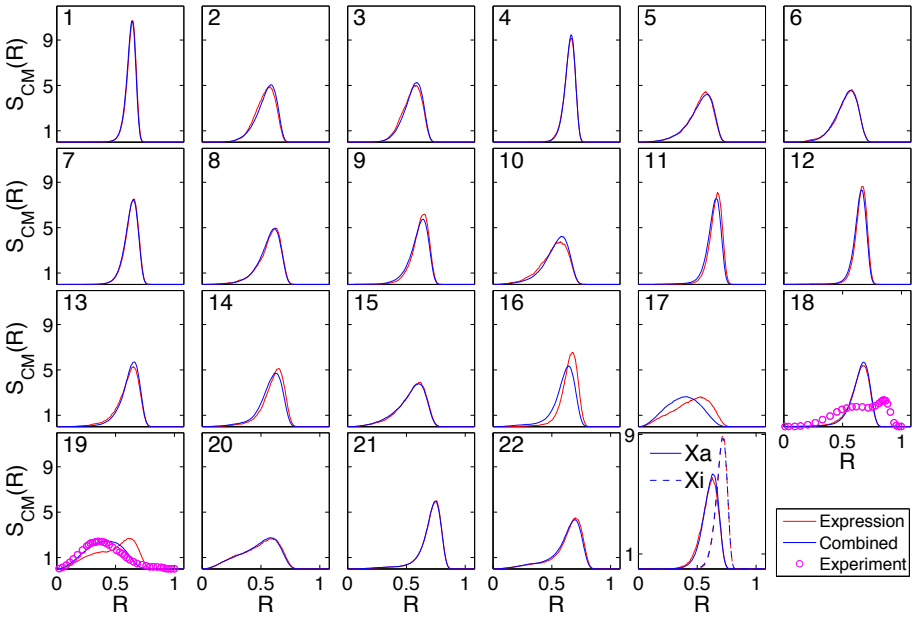

Figure S6: **The centre of mass distribution S_{CM}(R) for all chromosomes, within the "gene expression" and "combined" models.** S_{CM}(R), the centre of mass distribution of each chromosome for gene expression (red) and combined model (blue). Experimental data from Ref. (Kalhor et al., 2011) for Chromosomes 18 and 19 are shown in magenta.

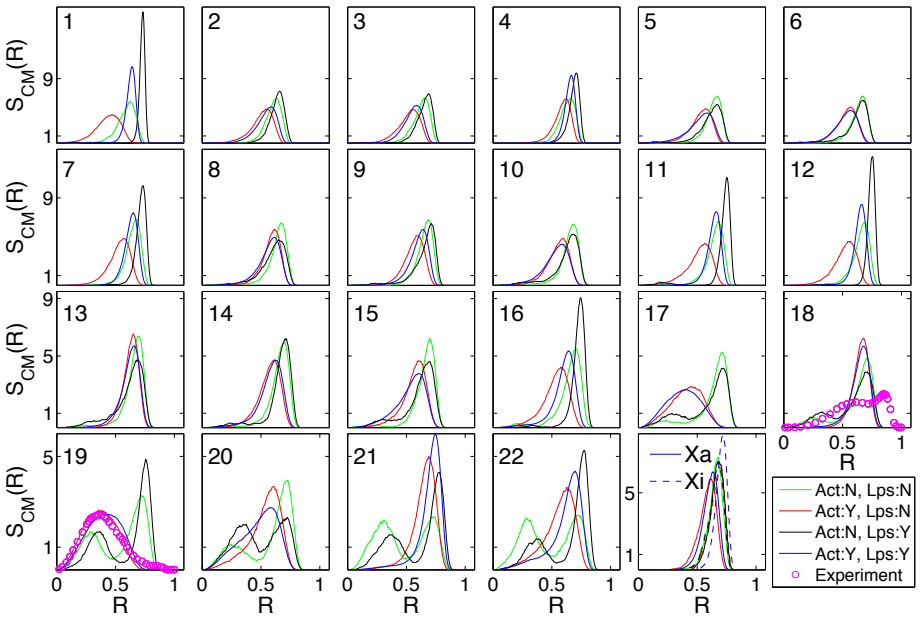

Figure S7: **Centre of mass distribution S_{CM}(R) for all chromosomes with different combinations of activity and loops.** S_{CM}(R) center-of-mass distribution for each chromosome is shown for the models described below. **Act:N, Lps:N** Both activity and loops are switched off, with all monomers at the same effective temperature of T = 1, shown in green; **Act:Y, Lps:N** Activity is present but loops are switched off, shown in red; **Act:N, Lps:Y** Activity is absent but loops are present, shown in black; **Act:Y, Lps:Y** Both activity and loops are present, shown in blue color. This is the original "combined model", also shown in (Figures 3C and S6); The experimental data for chromosomes 18 and 19 are shown using magenta circles using data from Ref. (Kalhor et al., 2011).

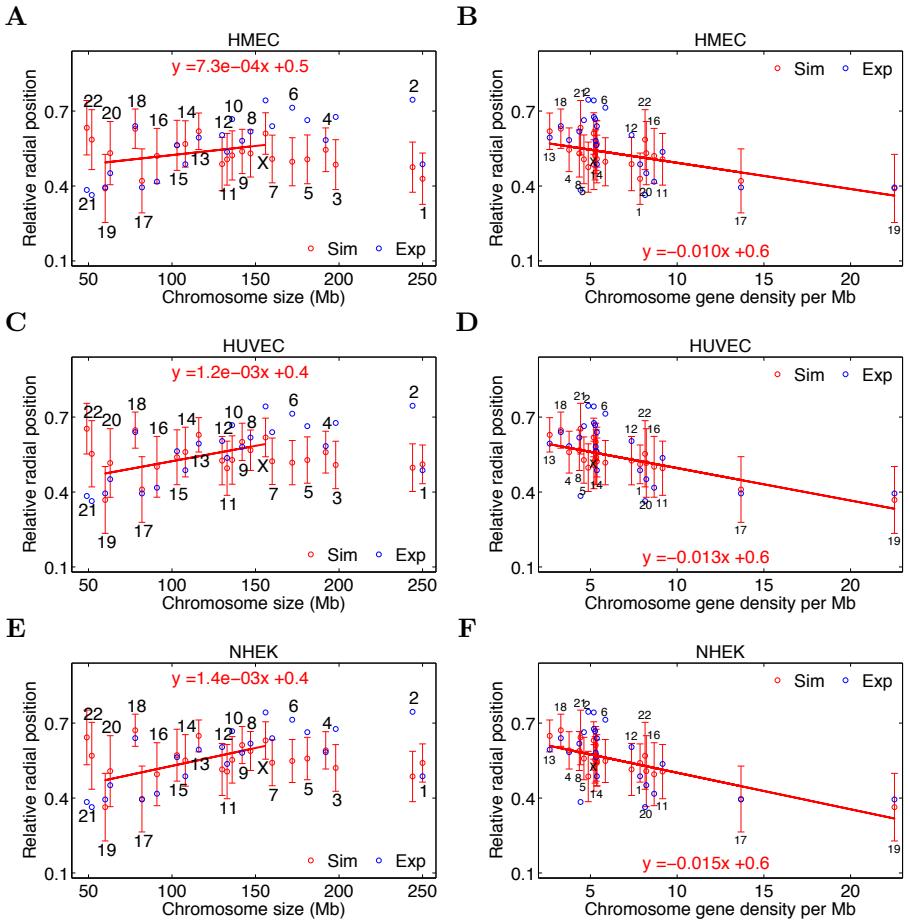

Figure S8: **Mean centre-of-mass location for each chromosome against chromosome sizes and chromosome gene density.** (**Figures A,C,E**). Predictions for the mean center of mass location for each chromosome, for the HMEC, HUVEC, NHEK cell type, as a function of chromosome size, compared to experimental data of GM12878 Ref. (Kalhor et al., 2011) is shown in Figures A, C and E respectively. Simulation and experimental points are shown together using red and blue ovals respectively, along with simulation error bars. The relative radial position 0 and 1 stands for the centre and periphery of nucleus respectively. Chromosome numbers are indicated above or below each error-bar. Simulation points except chromosome 1,2,3,4,5,6,7,21 and 22 are fitted to a straight line whose parameters (slope, intercept) are provided at the top of each sub-figure. (**Figures B,D,F**). Predictions for the mean centre of mass location for each chromosome as in Figures A,C and E, but plotted against the function of chromosome gene density in Figures B,D and F respectively. All the simulation points are fitted to a straight line whose parameters (slope, intercept) are mentioned below each sub-figure.

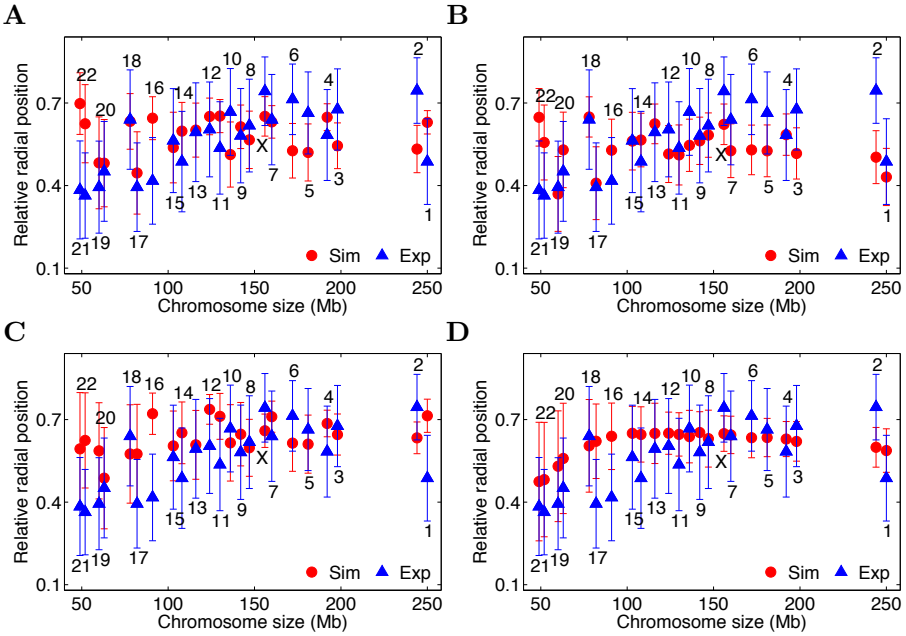

Figure S9: **The relative centre-of-mass position of each chromosome for different combinations of activity and loops.** The relative centre-of-mass position of each chromosome, in increasing order of size, shown for different models; **(A) Gene expression** The effective temperature assignment of monomers is derived from gene expression model of GM12878 cell type and loops are extracted from Hi-C data; **(B) Act:Y, Lps:N** The effective temperature assignment of monomers is taken from the combined model of GM12878 cell type but no loops are present; **(C) Act:N, Lps:Y** All monomers are at the same temperature (no activity), but loops inferred from Hi-C are present; **(D) Act:N, Lps:N** All monomers are at the same temperature (no activity) and loops are also absent; Simulation data points (red oval) are shown together with the experimental data (blue triangle) (Kalhor et al., 2011) along with respective error-bars. Chromosome numbers are mentioned above or below each error-bar.

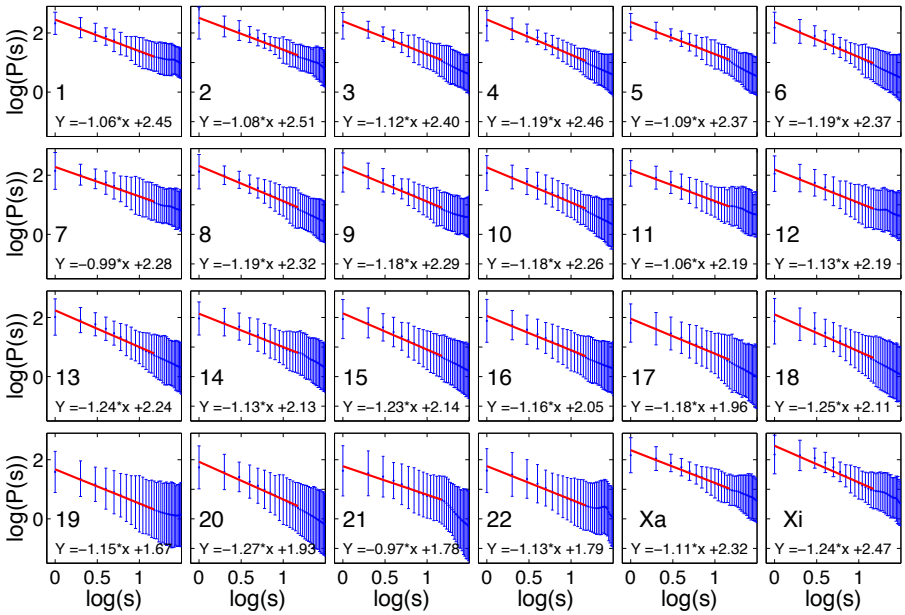

Figure S10: **Contact probability $P(s)$ for each chromosome in the GM12878 cell type.** The scaling of monomer contacts within the initial 10 Mb region for each chromosome is shown with blue dots, accompanied by error-bars. The best power law fit to the data is provided along with each subfigure, in red. The chromosome number is mentioned in the bottom corner of each subfigure. The range of exponents provided by the fit is $0.97 - 1.27$.

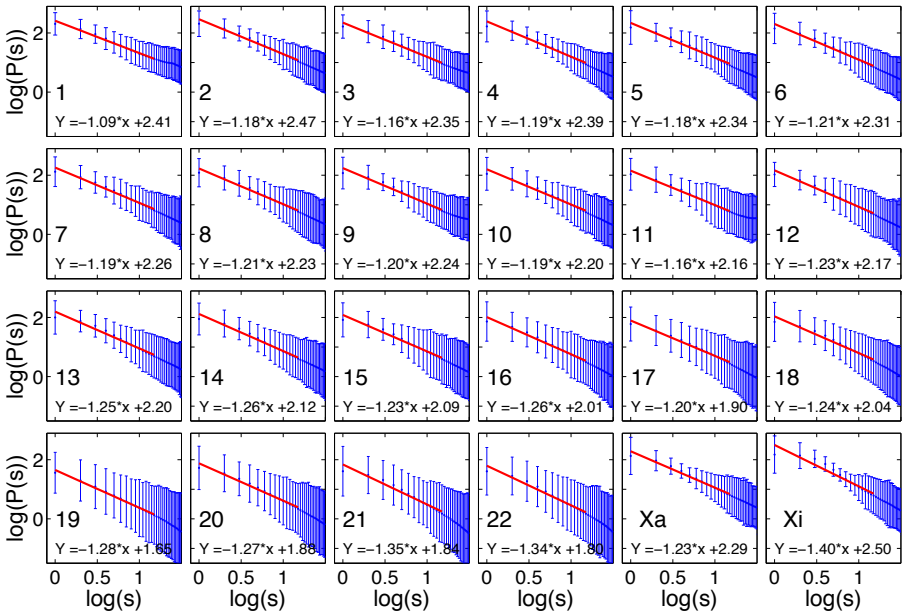

Figure S11: **Contact probability $P(s)$ for each chromosome in the IMR90 cell type.** The scaling of monomer contacts in the initial 10 Mb region for each chromosome is shown with blue dots, accompanied by error-bars. The best power law fit to the data is provided along with each subfigure, in red. The chromosome number is mentioned in the bottom corner of each subfigure. The range of exponents provided by the fit is $1.09 - 1.40$.

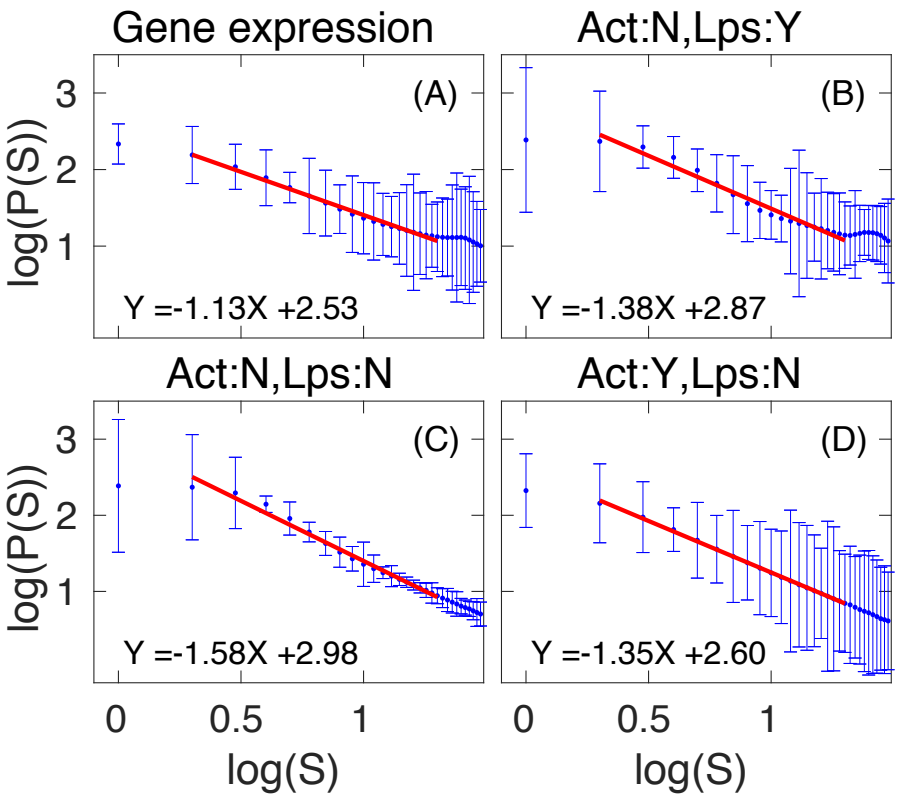

Figure S12: **Contact probability P(s) of chromosome 1 for different combinations of activity and loops.** Contact probability for chromosome 1, fit to a power law for the gene density, gene expression and combined models, plotted in the range [1-20 Mb]. **(A) Gene Expression** $P(s)$ for the gene expression model with effective temperature assignments using RNA-seq of GM12878 cell (Consortium et al., 2012). Loops are obtained from Hi-C experiments (Rao et al., 2014); **(B) Act:N, Lps:Y** $P(s)$ for the case where there is no activity but looping is present; **(C) Act:N, Lps:N** $P(s)$ for the case where both activity and looping are absent ; **(D) Act:Y, Lps:N** $P(s)$ in the case where the effective temperature assignment corresponds to the combined model for the GM12878 cell type but loops are absent.

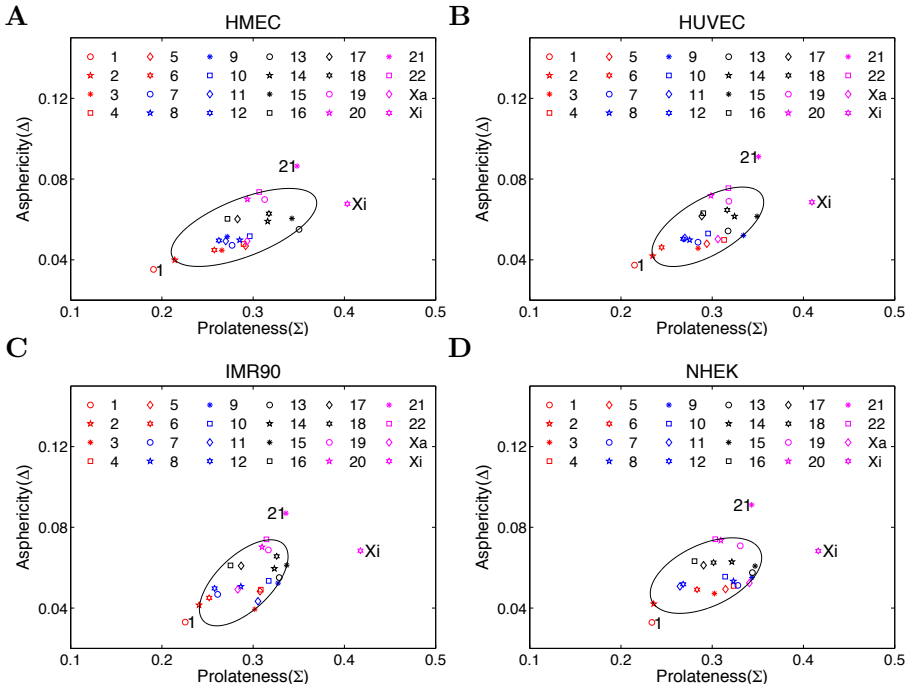

Figure S13: **Asphericity Δ and Prolateness Σ parameter of chromosomes for IMR90, HMEC, HUVEC, and NHEK cell type.** We capture broad trends in the data by plotting an ellipse that includes closely correlated points. Chromosomes 1, 21 and Xi, are clearly excluded while chromosomes 22, 2, 20, 11, 15, 9 and 13 lie within the ellipse.

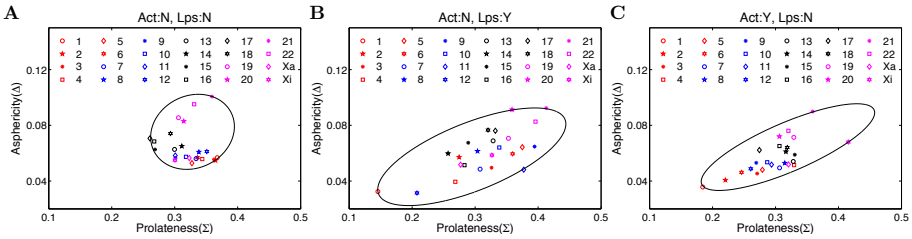

Figure S14: **Asphericity Δ and Prolateness Σ parameter of chromosomes for different combinations of activity and loops.** Δ and Σ of each chromosome is plotted for 3 different models: **Act:N, Lps:N** in which activity is absent and all permanent loops are absent; **Act:N, Lps:Y** in which activity is absent, but permanent loops are present; **Act:Y, Lps:N** in which effective activity is taken from the combined model appropriate to GM12878 cell type but permanent loops are absent; From these figures, we observe that chromosomes achieve a prolate ellipsoidal shape in the absence of activity. Larger chromosomes are more spherical and smaller chromosomes are rougher and more rod-like.

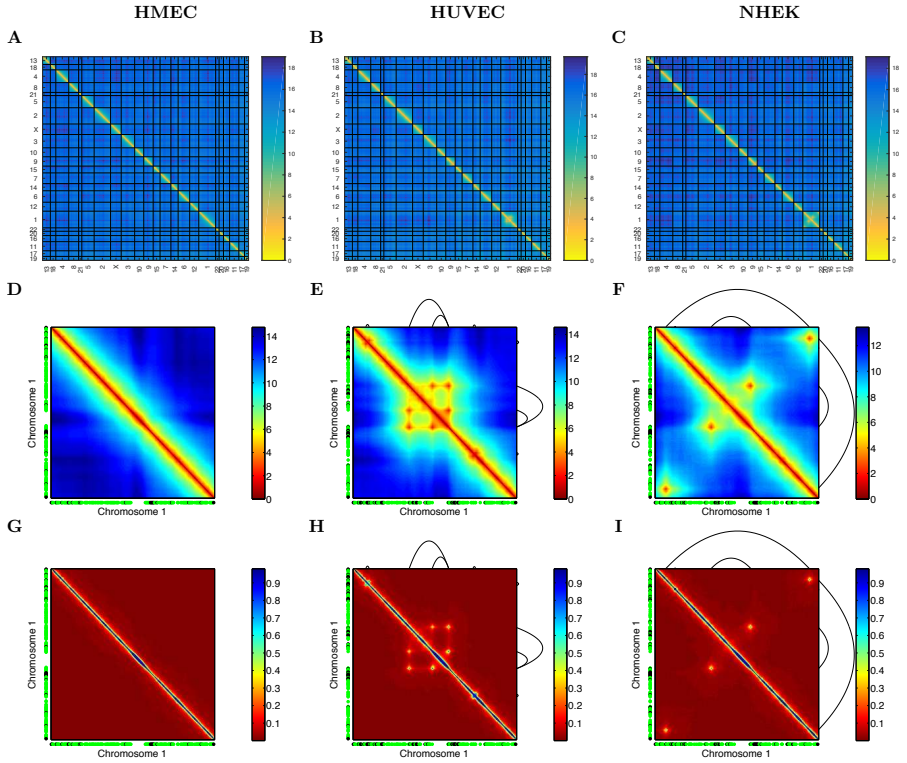

Figure S15: **Simulation-derived distance and contact maps for the HMEC, HUVEC and NHEK cell types. A-C** Distance maps for all chromosomes, ordered according to their gene density; **D-F** Distance map of chromosome 1 displayed along with locations of the permanent loops (black curve) as inferred from the Hi-C data for this cell type. Individual monomers at T = 6 and $7 \leq T \leq 12$ are shown in green and black beside the X and Y axis, respectively; **G-I** Contact map of chromosome 1 displayed along with locations of the permanent loops with black curve as inferred from the Hi-C data. Individual monomers at T=6 and $7 \leq T \leq 12$ are shown in green and black beside the X and Y axis, respectively; Related color-bars for distance and contact maps are shown beside the respective sub-figures.

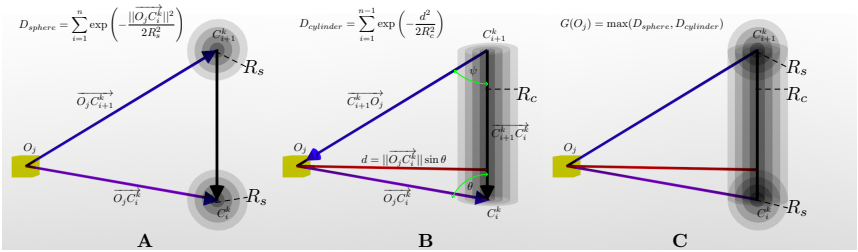

Figure S16: **Schematic showing how the chromosome territory is computed using the grid method. A** Density at yellow grid point (D_{sphere}) is computed due to sphere effect; **B** Density at yellow grid point ($D_{cylinder}$) is computed due to cylinder effect; **C** The actual density at yellow grid point $G(O_j)$ is the maximum of either D_{sphere} or $D_{cylinder}$.

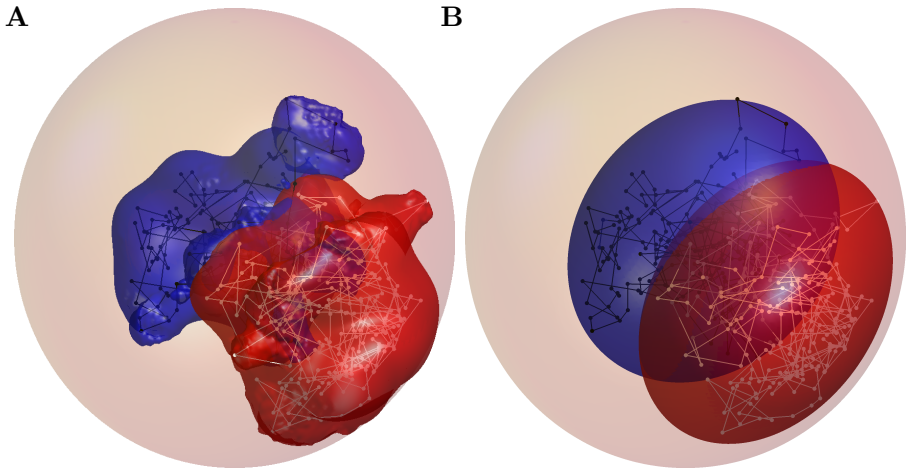

Figure S17: **Comparison of output of grid method and 3d ellipsoid fit method. (A)** Schematic representation of volume overlap of chromosome 1 and 3 by an implicit surface method. The polymer form of chromosome 1 and 3 is represented with white and black color respectively. The 3d surface of chromosome 1 and 3 computed from grid method is shown with red and blue color respectively. **(B)** Schematic representation of volume overlap of chromosome 1 and 3 by an ellipsoidal fit method. The polymer form of chromosome 1 and 3 is represented with white and black color respectively similar to Fig. A. The 3d surface of chromosome 1 and 3 computed from ellipsoid algorithm is shown with red and blue color respectively.

References

Consortium, E. P. et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57.

Ganai, N., Sengupta, S. and Menon, G. I. (2014). Chromosome positioning from activity-based segregation. Nucleic Acids Research *42*, 4145–4159.

Hameed, F. M., Rao, M. and Shivashankar, G. (2012). Dynamics of passive and active particles in the cell nucleus. PLoS One *7*, e45843.

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S. et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. Genome research *22*, 1760–1774.

Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F. and Chen, L. (2011). Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat Biotechnol *30*.

Kremer, K. and Grest, G. S. (1990). Dynamics of entangled linear polymer melts: A molecular-dynamics simulation. The Journal of Chemical Physics *92*, 5057–5086.

Kreth, G., Finsterle, J., Von Hase, J., Cremer, M. and Cremer, C. (2004). Radial arrangement of chromosome territories in human cell nuclei: a computer model approach based on gene density indicates a probabilistic global positioning code. Biophysical journal *86*, 2803–2812.

Millett, K. C., Plunkett, P., Piatek, M., Rawdon, E. J. and Stasiak, A. (2009). Effect of knotting on polymer shapes and their enveloping ellipsoids. The Journal of chemical physics *130*, 04B623.

Plimpton, S., Crozier, P. and Thompson, A. (2007). LAMMPS-large-scale atomic/molecular massively parallel simulator. Sandia National Laboratories *18*.

Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D. and Robinson, J. T. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *59*.

Rawdon, E. J., Kern, J. C., Piatek, M., Plunkett, P., Stasiak, A. and Millett, K. C. (2008). Effect of knotting on the shape of polymers. Macromolecules *41*, 8281–8287.

Rosa, A. and Everaers, R. (2008). Structure and dynamics of interphase chromosomes. PLoS computational biology *4*, e1000153.

Sehgal, N., Fritz, A. J., Morris, K., Torres, I., Chen, Z., Xu, J. and Berezney, R. (2014). Gene density and chromosome territory shape. Chromosoma *123*, 499–513.

Stillinger, F. H. (1976). Phase transitions in the Gaussian core system. The Journal of Chemical Physics *65*, 3968–3974.

Todd, M. J. and Yıldırım, E. A. (2007). On Khachiyan's algorithm for the computation of minimum-volume enclosing ellipsoids. Discrete Applied Mathematics *155*, 1731–1744.

Weber, S. C., Spakowitz, A. J. and Theriot, J. A. (2012). Nonthermal ATP-dependent fluctuations contribute to the in vivo motion of chromosomal loci. Proceedings of the National Academy of Sciences *109*, 7338–7343.

Zidovska, A., Weitz, D. A. and Mitchison, T. J. (2013). Micron-scale coherence in interphase chromatin dynamics. Proceedings of the National Academy of Sciences *110*, 15555–15560.

Žunić, D. and Žunić, J. (2013). Shape ellipticity based on the first Hu moment invariant. Information Processing Letters *113*, 807–810.