Inferring Chromosome Radial Organization from Hi-C Data

Priyojit Das[1], Tongye Shen[2], and Rachel Patton McCord[2,*]

[1]UT-ORNL Graduate School of Genome Science and Technology, University of Tennessee, Knoxville, TN 37996, USA
[2]Department of Biochemistry & Cellular and Molecular Biology, University of Tennessee, Knoxville, TN 37996, USA
[*]Correspondence to: rmccord@utk.edu

Abstract

Motivation: The nonrandom radial organization of eukaryotic chromosome territories (CTs) inside the nucleus plays an important role in nuclear functional compartmentalization. Microscopy techniques can reveal radial positioning of CTs, but have limited throughput. In comparison, biochemical techniques such as Hi-C has been used extensively to obtain spatial genome organization. To take advantage of the widely available Hi-C data to infer the radial organization of territories, computational techniques are required.
Results: We constructed a computational pipeline that can infer the radial arrangement of CTs using a force-directed network layout algorithm under a regression scheme. The pipeline's computational prediction has a high correlation with the microscopy imaging data for various cell shapes (lymphoblastoid, skin fibroblast, and breast epithelial cells). Furthermore, this computational technique can infer meaningful changes in the arrangement of the CTs in cells having an irregular nuclear shape.
Availability and implementation: The algorithm is implemented using Python. The source code of the algorithm is available at https://github.com/priyojitDas/HIRAC.

1 Introduction

The long linear sequence of eukaryotic DNA is packed tightly inside the nucleus, leading to the formation of a hierarchical genome organization. At the small end of this structural

hierarchy, nucleosomes are formed by wrapping the double-stranded DNA around histone proteins. Moving to larger length scales, different structures ranging from chromatin, loops, topologically associating domains, and compartments emerge (Finn and Misteli, 2019). At the largest scale of chromosome organization, the 3D bodies of individual chromosomes arrange as mostly discrete entities, known as Chromosome Territories (CTs) (Cremer and Cremer, 2010). The arrangement of all of the CTs inside the nucleus with respect to nuclear center and periphery forms the higher-order genome architecture. This higher-order CT organization plays a role in different nuclear mechanisms ranging from DNA replication and gene expression to the processing of RNA (Cremer et al. (1993)).

The organization of CTs inside the nucleus is nonrandom with respect to the nucleus periphery. Based on previous studies, this nonrandom distribution can be largely classified into two categories (Boyle et al., 2001; Sun et al., 2000). In one category, the CTs follow a gene-density based radial positioning, where gene rich chromosomes reside near the nucleus center and gene poor chromosomes are positioned near the nuclear periphery. It has been suggested that gene activity might drive this gene density based CT organization. This idea is supported by the fact that tightly packed heterochromatin regions show association with the nuclear lamina (Van Steensel and Belmont, 2017) and tethering to the lamina often represses transcription (Amendola and van Steensel, 2014). In general, lamin associated domains (LADs) occur more frequently on gene poor chromosomes than on gene rich ones (Guelen et al., 2008). In contrast, the second category shows a chromosome length based CT organization - the smaller chromosomes near center and longer ones near periphery. Cell cycle and nuclear shape have been pointed out as the potential key players for this size-based organization. It has been shown that the proliferating cells follow a gene density based organization compared to the size based distribution in quiescent or senescent cells (Mehta et al., 2007). Apart from this gene density and cell cycle driven radial organization, some studies have found a relationship between nuclear shape and radial distribution (Bolzer et al., 2005). According to these studies, the spherical lymphocyte nucleus follows a gene density based organization, which is conserved across several related species (Tanabe et al., 2002, 2005), while the chromosome size based distribution is more prevalent in ellipsoidal fibroblast nucleus.

The position of each chromosome varies between cells in the population, while following certain tendencies (Kind et al., 2015). Throughout the interphase the CT positions remain stable, but change from one generation to another during mitosis (Lucas and Cervantes, 2002; Walter et al., 2003; Kind et al., 2013; Strickfaden et al., 2010). The probable reason behind this observation is that after the nuclear envelope breakdown during mitosis the chromosomes positions are rearranged. Apart from these cell cycle related changes, rearrangement of the CTs can be seen during cell differentiation (Kuroda et al., 2004) and in different disease conditions. For example, in Hutchinson-Gilford progeria syndrome cells, chr18 appears to shift more toward the nucleus interior compared to its position in normal quiescent fibroblast cells (Mehta et al., 2011). Further, certain gene-rich chromosomes in progeria tend to localize near the periphery in blebs (Bercht Pfleghaar et al., 2015). Recently, researchers have shown

that in breast cancer chrX gains inter-chromosomal interactions and its gene expression changes (Fritz et al., 2014).

The traditional way to study the organization of the CTs inside the nucleus is through imaging techniques. Fluorescence in situ hybridization (FISH) technique is an effective way to locate the position of chromosomes inside the nucleus by painting them with sequence specific fluorescent dyes (Solovei and Cremer, 2010). The throughput of FISH techniques is increasing with recent advances (Shachar et al., 2015), but the approach remains technically difficult, with data available on chromosome positioning from relatively few cell types (Fritz et al., 2016). On the other hand, along with imaging techniques, chromosome conformation capture-based techniques (e.g., 3C, 4C, 5C, Hi-C) have been applied extensively for the study of the spatial genome organization for the last decade (Sati and Cavalli, 2017). These techniques capture the snapshot of chromosomal interactions ranging from specific promoter-enhancer interactions (Schoenfelder et al., 2018) to the large scale inter-chromosomal interactions. Out of the arsenal of different conformation capture techniques, throughout the paper we will only focus on the Hi-C technique, as it captures contacts genome-wide (Golloshi et al., 2018).

Several studies have been performed to reconstruct the 3D folding of individual chromosomes and the genome from Hi-C data at different resolutions using restraint and polymer physics based approaches (Oluwadare et al., 2019). Out of those, only a few focused on the organization of CTs inside the nucleus. In one approach, a coarse-grained polymer based simulation is used to characterize non-random organization of the chromosomes using gene activity and random looping constraints (Ganai et al., 2014). However, this model simulates the genome at very coarse-grain level in a cell-type-independent manner. In another approach, the Chrom3D algorithm uses lamina associated domain (LAD) data along with Hi-C data to capture spatial and radial organization of the chromosomes (Paulsen et al., 2017e). But, the limitation of this method is that, to successfully capture the radial position of CTs, cell-type-specific LAD data is needed which is not widely available for many cell types. An approach that used probabilistic deconvolution of Hi-C maps into a population of single cell structures produced predicted radial positions that matched fairly well with microscopic measurements from a single cell type (Tjong et al., 2016). However, it was not explored whether this approach could predict different CT configurations in different cell types.

Though Hi-C experiments can capture intra-chromosomal and inter-chromosomal contacts comprehensively, the contact data is population averaged. Due to population-averaged nature of the Hi-C and also the variability of the inter-chromosomal interactions between cells, inferring the radial organization of CTs from Hi-C data becomes difficult. To overcome this limitation associated with the Hi-C data, we have developed a computational pipeline named - **H**i-C **I**nferred **R**adial **A**rrangement of **C**hromosomes (HIRAC) that generates multiple 3D network configurations using a force-directed graph layout algorithm from Hi-C data. This approach helps us model the heterogeneous variability of CT interactions between cells

in the population. Further, combining this approach with a regression technique, HIRAC predicts the radial organization of the CTs and shows high concordance with microscopy imaging results. The parameter values used in this pipeline can be easily computed and are tuned to work with different cell types. Moreover, this computational pipeline can be used to study the rearrangement of the CTs during various diseased conditions and biochemical perturbations for which Hi-C data exists.

2 Methods

Hi-C experiments capture spatial genome organization by measuring the interaction frequency between different genomic fragments, yielding information about both intra-chromosomal and inter-chromosomal interactions. By binning the data at a specific resolution, the Hi-C contact information can be represented in a square symmetric matrix form. For example, when the whole human genome (22 autosomes, one X and one Y chromosome) is binned at 2.5 Mb resolution, it produces a total of 1253 genomic bins, and the size of Hi-C contact matrix becomes 1253×1253. With the help of the Hi-C contact information and a force-directed graph drawing layout, we construct an ensemble of 3D network graphs where nodes are genomic bins and edge weights are the number of contacts between two bins. Subsequently, by fitting a geometrical object to these network graphs, we obtain the distance of chromosome territories (CT) from the center of the fitted object. In the last step, we use a regression technique to calculate the consensus radial organization of the CTs. The schematic representation of the whole pipeline is given in Fig. 1.

Before we feed Hi-C data into the HIRAC pipeline, two pre-processing steps were performed on the Hi-C matrix. First, to remove the GC content and Hi-C read count biases, the raw contact matrix was normalized using the ICE normalization technique (Imakaev et al., 2012). In the second step, the unmappable (repetitive) genomic regions were removed from the normalized Hi-C matrix. Also, we did not consider chrY in our analysis since this chromosome is not present in both male and female cell types, and we removed those corresponding bins from the Hi-C matrix. This modified Hi-C matrix was used for further downstream processing.

2.1 Construct 3D Network Graphs from a Hi-C Matrix

Let $H_{m \times m}$ represent a Hi-C contact matrix where m is the number of genomic bins of specific size, with elements h_{ij} representing the normalized number of contacts between i^{th} and j^{th} bins. Many inter-chromosomal elements of the contact matrix have contact frequencies similar to the average background level. To exclude contacts that may primarily reflect background and instead extract strong interactions more likely to distinguish 3D chromosome

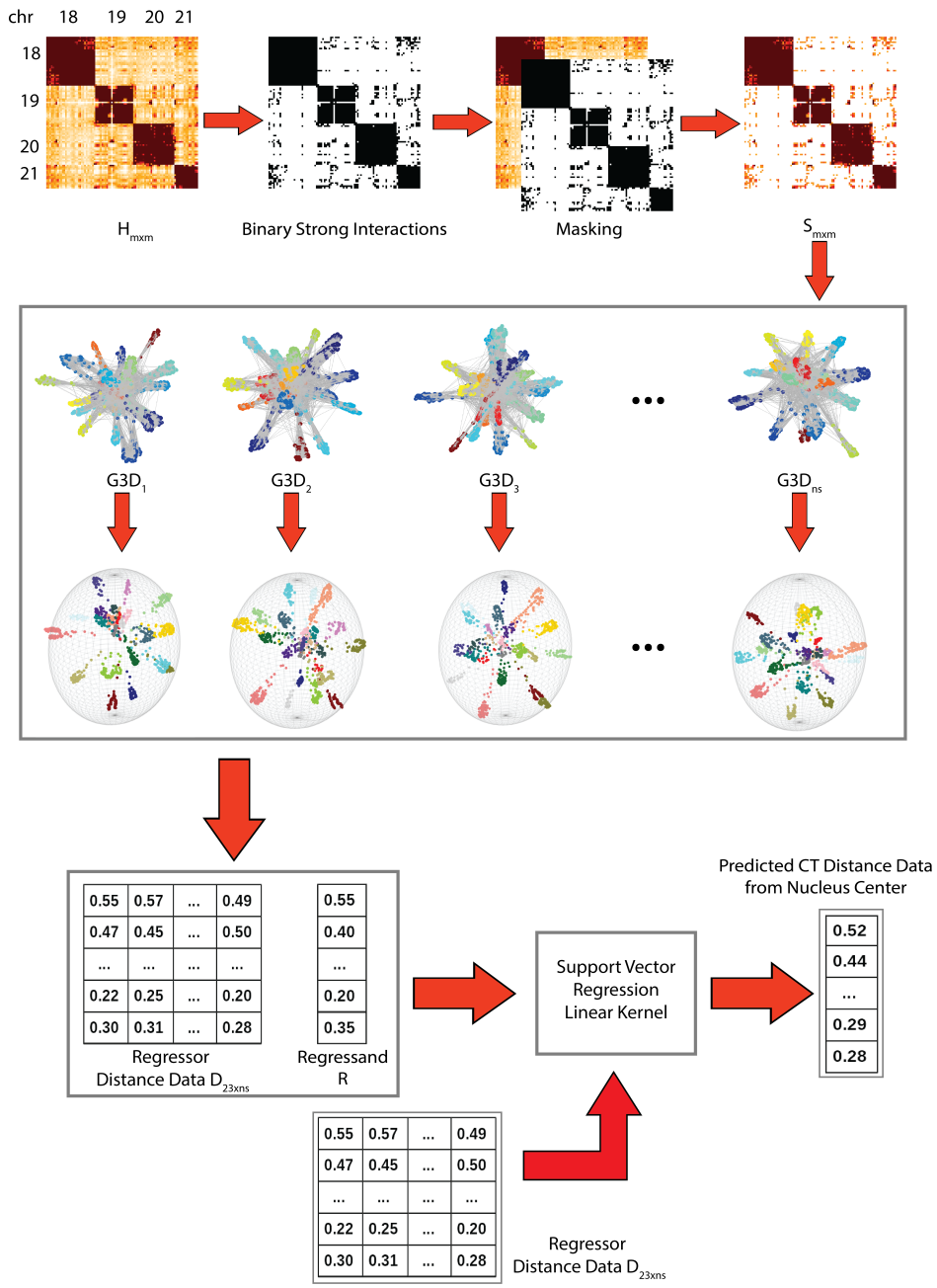

Figure 1: Overview of the HIRAC computational pipeline. After thresholding the whole genome Hi-C map, the resultant strong interaction matrix is transformed into multiple 3D network structures and a minimum volume ellipsoid is fitted to each structure. Consensus CT positions are then obtained from the population of structures using a regression technique

positions between cell types, we set a predefined cutoff limit and applied this to the matrix $H_{m\times m}$, to obtain a matrix of only filtered strong interactions $S_{m\times m}$.

$$\begin{cases} s_{ij} = h_{ij} & \text{for } h_{ij} > h_{cut} \\ s_{ij} = 0 & \text{otherwise} \end{cases} \tag{1}$$
$$s_{ii} = 0$$

where s_{ij} is the element of matrix $S_{m\times m}$, which represents the normalized number of contacts between i^{th} and j^{th} bins and h_{cut} is the cutoff limit. For 2.5 Mb genomic resolution, the value of cutoff limit was set to: $h_{cut} = 95^{th}$ percentile of all the elements of $H_{m\times m}$. The significance of and reason for choosing this 95th percentile cutoff is discussed in the **Results** section. Due to the application of this cutoff limit h_{cut}, the resultant matrix $S_{m\times m}$ contains mostly the intra-chromosomal interactions and a few strong inter-chromosomal interactions.

This strong interactions matrix $S_{m\times m}$ was treated as a weighted adjacency matrix to generate an undirected weighted graph G, where nodes are genomic bins and the number of interactions between the bins is used as the edge weight. Next, we applied a 3D Fruchterman-Reingold (FR) force based layout (Fruchterman and Reingold, 1991) to draw the undirected Hi-C graph in 3D space. This graph drawing layout adds an attractive force between the connected nodes and creates repulsion between the nodes that are not connected. Along with this, gravitational force is used in this layout to pull the nodes towards the center. The rationale behind using the FR layout is that it uncovers the intrinsic structure of the network, as the strong interactions only matrix $S_{m\times m}$ has a large number of intra-chromosomal interactions and fewer inter-chromosomal interactions. For each chromosome, the nodes will be clustered together due to high intra-chromosomal contact among themselves. On the other hand, the chromosomes that have higher inter-chromosomal contacts among themselves will be placed close together in 3D space compared to the other chromosomes with less inter-chromosomal contacts. Since, from the Hi-C matrix, we could model only one copy of each chromosome, from here we considered individual chromosomal territory locations. In reality, this will represent the average of the chromosome locations of each homolog. The resulting 3D network graph infers the radial organization of the CTs inside the nucleus.

Though the 3D FR layout can infer the average arrangement of the CTs from the Hi-C graph G, it has been shown that the positioning of the CTs in individual cells changes from mother cell to daughter cell during mitosis (Lucas and Cervantes, 2002; Walter et al., 2003). To mimic this variable CT arrangement phenomenon, n_s number of independent runs of the FR algorithm were performed on the undirected Hi-C graph G with different random initial configurations, which led to the formation of n_s different 3D network graphs: $G3D_1, G3D_2, ..., G3D_{n_s}$.

2.2 Fit Geometrical Structure to the 3D Network Graphs

For each 3D network graph, the next objective is to find the distance of each CT from the center of that network graph. This step was performed in two parts.

In the first part, given a 3D network graph, the 3D Cartesian coordinates of the nodes were extracted and then Khachiyan's algorithm used to find a minimum ellipsoid volume enclosing the set of nodes (Todd and Yıldırım, 2007). The center of the fitted object was calculated and assigned as the nucleus center. Along with this, the center of mass of each of the individual CTs was calculated from the coordinates of the nodes of the network graph.

$$CENTER_i = Fit(\text{coordinates of } G3D_i), i \in \{1, 2, ..., n_s\} \tag{2}$$

$$COM_j^i = \frac{\sum \text{coordinates of nodes of CT } j \text{ in } G3D_i}{\text{\# of nodes of CT } j \text{ in } G3D_i} \tag{3}$$

where $CENTER_i$ is the center of the object fitted to the 3D network graph coordinates $G3D_i$, COM_j^i represents the center of mass of a particular CT j in the structure $G3D_i$ and $j \in \{1, 2, ..., 22, X\}$.

After obtaining the center of mass of each CT and nucleus center, the Euclidean distance was calculated between the center of mass of each CT and the nucleus center. Also, to remove the heterogeneity originating due to the different minimum volume ellipsoid fit of different 3D network structures, for each structure the distance values were normalized in the [0,1] range using Min-max normalization. In this way, by iterating over all of the 3D network graphs, for each CT, n_s distances from the center of the nucleus were obtained. This distance information was represented in a matrix $D_{23\times n_s}$ of size $23 \times n_s$, where rows represent 23 CTs (22 autosomes and one X chromosome), columns correspond to n_s 3D network graphs and each matrix entry d_{ji} is distance of the CT j from the nucleus center of network graph $G3D_i$.

2.3 Calculate the Consensus Radial Organization of CTs from the Population of Networks

The last part of the pipeline constitutes the calculation of a consensus radial arrangement of CTs using the distance values obtained from the 3D network graphs ensemble along with chromosome gene density and length information. Taking a simple average of the CTs distance across all of the structures would not work due to non-normal distribution of the CTs and the fact that the simple averaged positions do not correlate well with microscopy data. So, additional chromosomal properties must be added to the Hi-C data to predict the radial positions. As it has already been shown that radial organization of CTs depends on gene density and chromosome length to some extent, here, we modeled this effect using a regression technique to obtain the consensus position of CTs. The Epsilon-Support Vector Regression technique (Drucker et al., 1997) with linear kernel was used here to predict the consensus position from the population. The matrix values $D_{23\times n_s}$ were used here as the regressors and for the regressand or response variable, we designed a variable based on chromosome gene density and chromosome length. Human chromosome gene density (genes/Mb) was

obtained from the published book titled "Short guide to the human genome" (Scherer, 2008) and length information from UCSC Genome Browser hg19 (Haeussler et al., 2018). Also, before using the gene density and length information to calculate response variable, both data were normalized to keep the range within $[0, 1]$.

The response variable R has three components to model the relation between chromosome length and gene density. The first component corresponds to the loess (locally estimated scatterplot smoothing) (Cleveland, 1979) fitted values between gene density and chromosome length. Here, the loess technique was used to fit a smoothed curve between these chromosomal properties. The second and third components are related to the normalized gene density and normalized chromosome length respectively. Along with this, three parameters α, β and γ were used to regulate the contribution of each of components to the response variable R. We will discuss the values of the parameters in the following Results sections.

$$R = \alpha * Loess(GD, LN) + \beta * \exp(1 - GD) + \gamma * LN \tag{4}$$

$$\alpha + \beta + \gamma = 1.0 \tag{5}$$

where GD is gene density and LN is chromosome length.

Following the regression model training phase with the $D_{23 \times n_s}$ and R, the model was used to predict the response with the same training input variable $D_{23 \times n_s}$. The output of the model represented the consensus radial arrangements of the CTs inside the nucleus.

3 Results

3.1 Parameter estimation using lymphoblastoid and fibroblast cells

To estimate the parameters associated with different methods of the HIRAC pipeline, we used two different cell types with different nuclear shapes (Supplementary Table 1). Human blood lymphoblastoid cell line GM12878 has a spherical nucleus and follows a gene density based radial CTs organization (Boyle et al., 2001). On the other hand, BJ1-hTERT human skin fibroblast cells have an ellipsoidal nucleus, which shows chromosome length based CTs organization (Bolzer et al., 2005). The whole genome Hi-C count matrices at 2.5 Mb resolution were obtained from Sanders et al., 2019 and processed as described in **Methods**.

After obtaining those Hi-C count matrices, we tested different values of h_{cut} starting from 5 to 99th percentile to define strong interactions. This h_{cut} mainly affects inter-chromosomal interactions, which contain randomly ligated contacts and nearly uniform low frequency contacts along with biologically significant contacts. To determine which cutoff

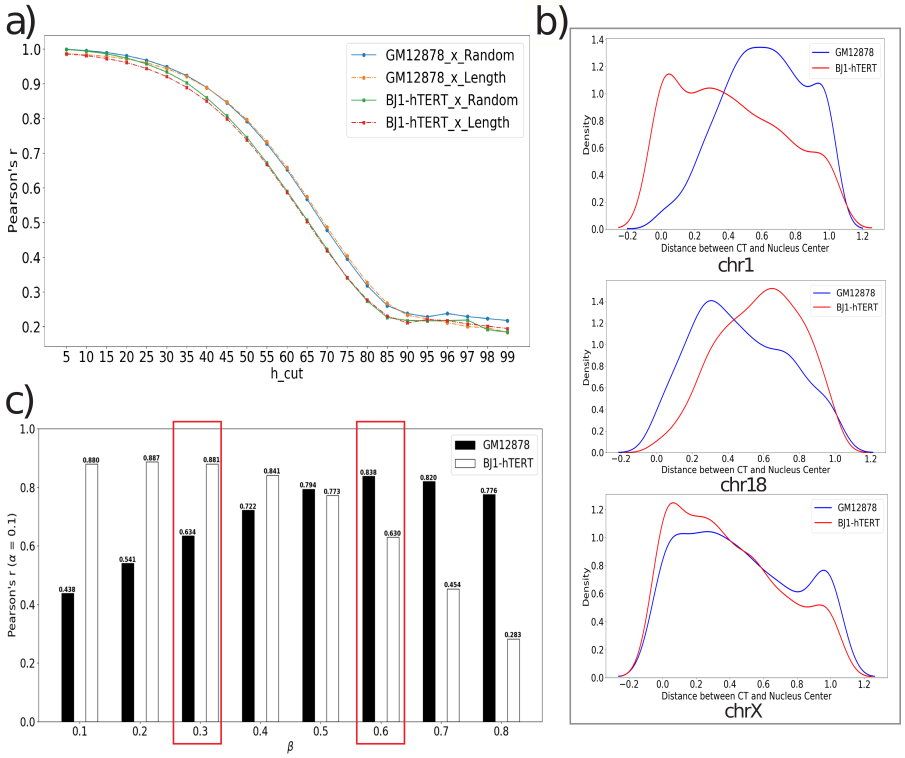

Figure 2: Parameter estimation of the HIRAC pipeline. **a)** Solid lines: correlation of strong inter-chromosomal contacts per chromosome between original Hi-C and randomly ligated Hi-C at different values of h_{cut}. Dotted lines: correlation between original Hi-C strong inter-chromosomal contacts per chromosome and chromosome length at different values of h_{cut}. **b)** The modeled distance distribution of 3 CTs (chr1, chr18, and chrX) from nucleus center for both GM12878 and BJ1-hTERT cells using a n_s value of 1000. **c)** Grid search result for different values of β and γ parameter with a fixed $\alpha = 0.1$ using GM12878 and BJ1-hTERT. For each combination of β and γ parameters, the predicted CTs distance data was compared with corresponding imaging data. As the α, β, and γ parameters are related, only the β parameter is shown here

limits best extract strong interactions without noise across the whole genome, we compared inter-chromosomal contacts of the original count matrix with the inter-chromosomal contacts of a simulated whole genome random ligation contact matrix.

$$\begin{aligned} CHRTRANS_{ij} &= \sum \text{strong interactions between chr } i \text{ and chr } j \\ CHRTRANS_{ii} &= 0 \end{aligned} \tag{6}$$

For each h_{cut} value, the inter-chromosomal contacts passing the threshold were summed between each pair of chromosomes to a single bin (Eqn. 6) and Pearson's correlation was

applied between sums obtained from original and randomly ligated Hi-C data. The random ligation Hi-C matrix was generated using the method described in Yardımcı et al., 2019. From Fig. 2a, it can be seen that with the increase in h_{cut} values from 45 to 85, the inter-chromosomal similarity between randomly ligated and real data decreased rapidly and reached a stable value around the 90th percentile. Similarly, we compared the inter-chromosomal interaction sums with the pairwise product of the chromosome lengths to measure how much the number of interactions was primarily driven by chromosome length (i.e. two large chromosomes will have more interactions at random overall than two small chromosomes). This analysis produced a similar correlation trend as the random ligation effect comparison - strong interaction sums are no longer primarily explained by chromosome length at 90-95 percentile h_{cut} (Fig. 2a). Based on these two comparison results for both GM12878 and BJ1-hTERT, it is evident that the effect of random ligation and length scale on strong interactions becomes less and less as the value of h_{cut} increases. So, we decided to use 95th percentile as h_{cut} value as at this limit there was little random ligation and length scale effect on the strong interactions.

The parameter n_s represents the number of configurations required to mimic the heterogeneity in CT arrangement originating during cell division within the same cell type. To estimate this parameter, for different values of n_s we compared the fitted distance distribution of all CTs between GM12878 and BJ1-hTERT cells using non-parametric hypothesis testing - a Two-sided Mann-Whitney U test. From the statistical test results, it can be observed that with increasing n_s value, more chromosomes show significant differences in their CT distributions between lymphoblastoid and fibroblast cells, reaching a maximum at $n_s = 1000$ (Supplementary Fig. 1 and 2). Hence, we set the n_s parameter to 1000 in the proposed pipeline. Some of the CT distance distributions have been shown in Fig. 2b.

The last part of the pipeline has three parameters that control the contribution of gene density and chromosome length to the arrangement of CTs. The grid search technique was used here to estimate the values of the three parameters. As the range of each of the values of the parameters was within $[0, 1]$, values starting from 0.0 to 1.0 with an interval of 0.1 were selected for each of the parameters for the estimation purpose. Out of all possible combinations of three parameter values ($11 \times 11 \times 11$), only the combinations satisfying Eqn. 5 were tested. From these filtered combinations, for each combination of values, consensus CT distances were obtained by using the regression technique described in the **Methods** section. Next, the predicted results were compared with imaging results. We obtained experimental CT distance data for lymphoblastoid and fibroblast cell from Boyle et al., 2001 and Bolzer et al., 2005 respectively. To maintain uniformity with the predicted data, the experimental data were normalized to the range [0,1]. The grid search procedure was applied to both GM12878 and BJ1 cells. By comparing the results of both gene density and length based distribution, we found that the α parameter is insensitive to the nuclear shape (Supplementary Table 2) while β and γ vary. In the case of GM12878, which follows a gene density based CT organization, higher β and lower γ value produced higher correlation with imaging results, while lower β and higher γ value gave the best results for BJ1-hTERT ellipsoidal

nuclei, where size-based organization is followed. Finally comparing all correlation results, we decided to set α to 0.1 and two related parameters β and γ to $(0.6, 0.3)$ and $(0.3, 0.6)$ for spherical and ellipsoidal nuclei respectively (Fig. 2c). Having set the five parameters, we

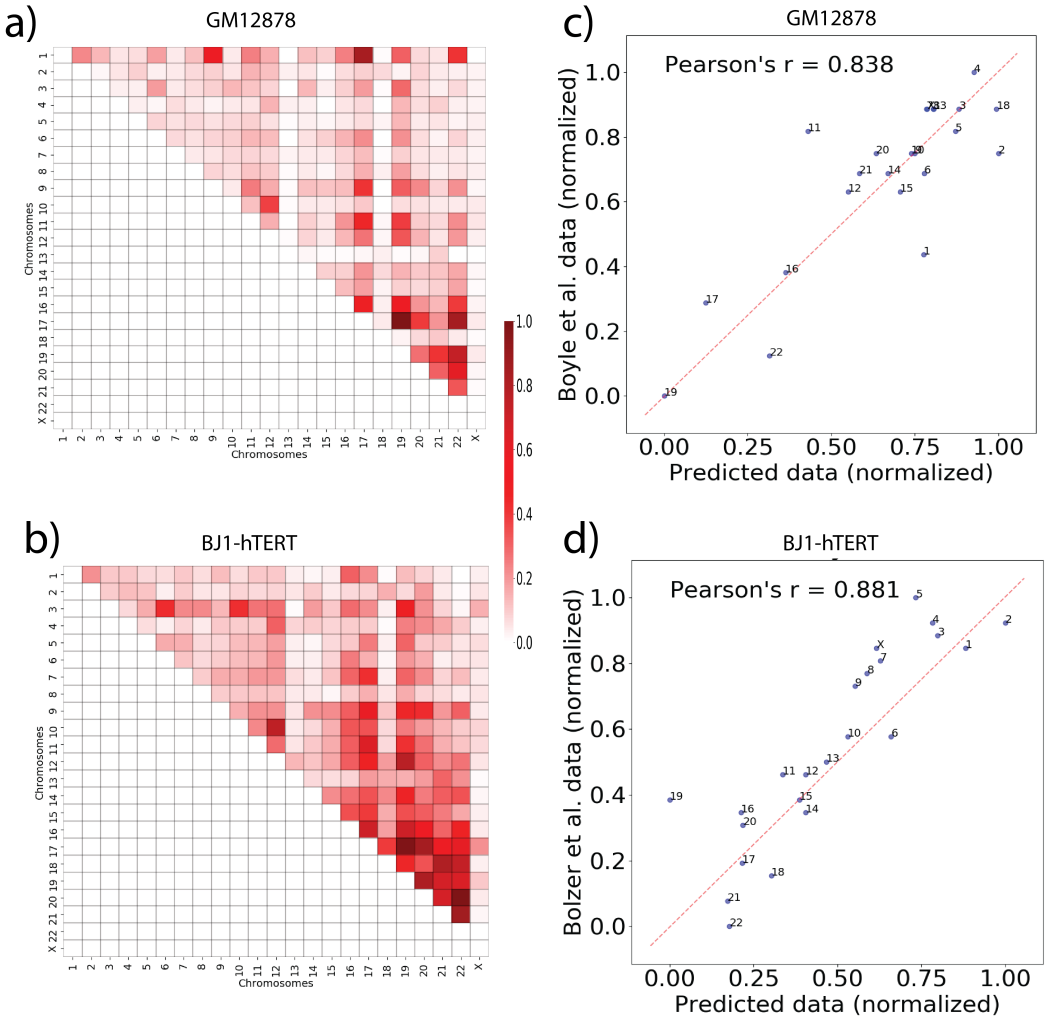

Figure 3: Performance evaluation using GM12878 and BJ1-hTERT cells. **a-b)** The inter-chromosomal strong interaction matrix obtained from GM12878 and BJ1-hTERT binned to a single bin per chromosome for an h_{cut} value of 95th percentile. **c)** Pearson's correlation between predicted CT distance from GM12878 Hi-C data and lymphoblastoid microscopy imaging data. **d)** Pearson's correlation between predicted CT distance from BJ1-hTERT Hi-C data and fibroblast microscopy imaging data

calculated the CT distance predictions for GM12878 and BJ1-hTERT. Fig. 3b and Fig. 3d shows the correlation of the GM12878 and BJ1-hTERT predicted results with the microscopy results. From Fig. 3b, it can be seen that, in GM12878, most of the CT positions correlate well with the experimentally obtained position with slight displacements, apart from the CT11 position, which moves towards center compared to the experimental data. Similarly,

predicted CT distances for the BJ1-hTERT cell shows high similarity with the corresponding imaging data. For this particular cell, out of all the CTs, CT19 showed a higher amount of displacement towards the center in the predicted result. When the technique was tested on simulated random ligation Hi-C data using both the lymphoblastoid and ellipsoidal parameter set, it produced a weaker correlation with the corresponding imaging results in both the cases compared to the original results (Supplementary Fig. 5). This demonstrates that real Hi-C contacts contribute important information to our modeled distances. After comparing each cell type prediction with corresponding imaging data using nuclear shape specific β and γ parameters, we performed a cross-comparison using different β and γ parameters and imaging data. From the comparison result, it is evident that when the parameter set corresponding to the nuclear shape was used, a higher correlation was obtained compared to other non-matching combinations (Table 1). Furthermore, we checked the consistency of the predicted results between two Hi-C replicates of the same cell type and found very little difference in chromosome positioning (Supplementary Fig. 3 and 4).

			Predicted Data	
		Parameters	GM12878	BJ1-hTERT
Microscopy Imaging Data	Lymphoblastoid	$\alpha = 0.1$, $\beta = 0.6$, $\gamma = 0.3$	**0.8379**	0.7773
		$\alpha = 0.1$, $\beta = 0.3$, $\gamma = 0.6$	0.6391	0.5607
	Fibroblast	$\alpha = 0.1$, $\beta = 0.6$, $\gamma = 0.3$	0.6001	0.6298
		$\alpha = 0.1$, $\beta = 0.3$, $\gamma = 0.6$	0.8271	**0.8805**

Table 1: Cross comparison between different parameter sets and microscopy imaging data

3.2 Performance evaluation using a breast epithelial cell line

After calibrating and testing the performance of the HIRAC technique on the GM12878 and BJ1 cells, we applied the pipeline to the MCF10A cell. MCFA10A is a non-tumorigenic breast epithelial cell which, when grown in 2D culture, has an ellipsoidal nucleus and follows a size based territorial organization (Fritz et al., 2014). The Hi-C contact data of this cell at 2.5 Mb resolution was downloaded from Barutcu et al., 2015. We obtained corresponding imaging data for 8 CTs (CT1, CT4, CT11, CT12, CT15, CT16, CT18 and CT21) of the MCF10A from Fritz et al., 2014. In the cultured cells used for Hi-C data, the cell nucleus would adopt an ellipsoidal shape, so we used the fibroblast parameter set here (even though this is an epithelial cell type). For the 8 CTs, when the predicted data was compared with the experimental data, a high correlation was obtained (Fig. 4a).

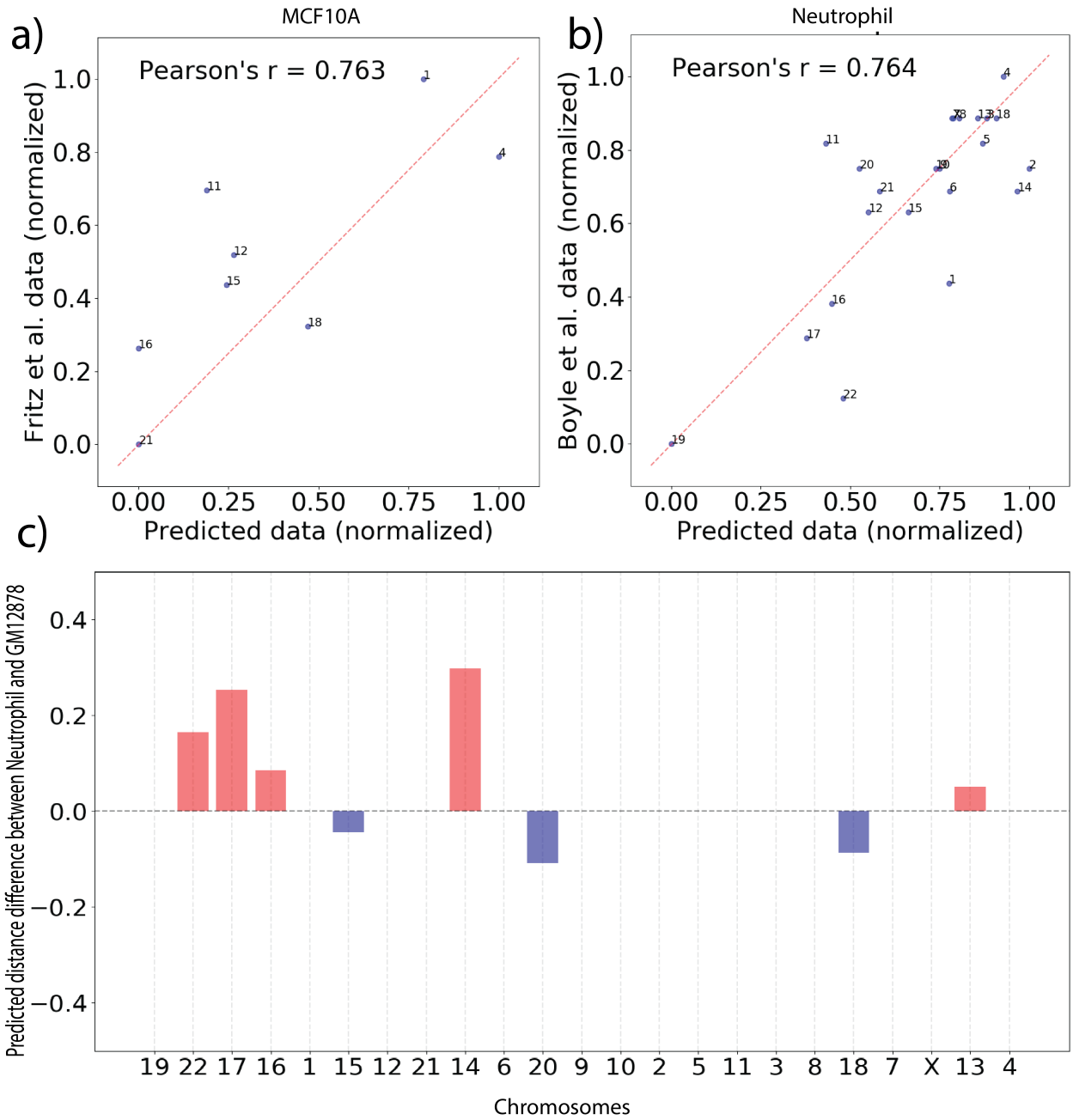

Figure 4: The application of the HIREC technique on MCF10A and neutrophil Hi-C data. **a)** Pearson's correlation between predicted CT distance from Hi-C data and microscopy imaging data for MCF10A cells. **b)** Pearson's correlation between predicted CT distance from neutrophil Hi-C data and lymphoblastoid microscopy imaging data. **c)** Displacement of the neutrophil CTs relative to GM12878 radial position of CTs

3.3 Variation between radial arrangement of lymphoblastoid and neutrophils

To test the performance of the computational technique on a cell having an irregular nuclear shape, we used neutrophil cell Hi-C data. Neutrophils have a multi-lobed nucleus where

lobes are connected by thin filaments and a toroid shaped genome (Zhu et al., 2017). The Hi-C contact data for netrophils was obtained from Javierre et al., 2016. Though the nucleus has an irregular shape, we used the lymphoblastoid parameter set for this case because the neutrophil is a differentiated product of promyelocyte which has a round nucleus. Our predicted CT positions for neutrophils showed a weaker correlation with lymphoblastoid imaging data than the GM12878 predictions did (Fig. 4b), which is consistent with the known differences between these cell types. This is another indication that the Hi-C contact data provides more information about chromosome positioning than just the underlying length based or gene density based distribution. The divergence of neutrophil from GM12878 CT positioning is expected. This might happen due to organization of pericentrometric heterochromatin and ribosomal DNA to the nuclear lamina in neutrophil. Our model results can also predict which chromosomes would move toward the center or periphery in one cell type vs. another as represented in Fig. 4c. In the case of neutrophils compared to GM12878, gene rich chromosomes are displaced from the nuclear center toward the periphery.

4 Discussion

Here, we have proposed a computational pipeline to the infer radial arrangement of the CTs from the Hi-C contact data using a graph layout algorithm and regression technique. The novelty of the proposed computational method is that it models and predicts the cell-type-specific radial organization of CTs from the Hi-C data. To do that, it combines Hi-C contact information with the chromosome gene density and length information as a guidance. Apart from this CT arrangement prediction, the HIRAC has several other noteworthy features. First, it models the heterogeneity present in the dynamic CT interactions between cells in a population using different configurations of 3D network structures. Second, the parameters used in this pipeline have been estimated using two different cell types having different organization of CTs and can therefore be applied to a variety of other cell types. Also, the parameters, such as h_{cut}, can be tuned easily depending on the resolution of the Hi-C data. Finally, the last part of the computational pipeline is independent of the other two parts and can be optimized to work with some new kind of radial positioning by tuning the α, β and γ parameters.

The HIRAC computational pipeline is highly flexible due to its modular nature and can be integrated with different kinds of genome analysis applications. For example, the inferred positions can be used to characterize the changes in the radial CT organization in a perturbed (e.g. relocation of CTs during DNA damage response (Fatakia et al., 2017)) or diseased cell (e.g. mislocalization of CT18 and CT19 in lamin B2 depleted colorectal cancer cells (Ranade et al., 2017)) compared to a control cell, which in turn allows us to study how these changes affect higher-order chromatin structure. Furthermore, the result obtained from this method can be combined with the polymer simulation of the chromosomes to study conformational structure and dynamics at different resolutions. In addition, the spatial

CT organization produced by this method can serve as the initial structure of the genome simulation for further downstream analysis. The method also generates a population of 3D network structures, which can further be used to characterize interchromosomal dynamics and compared to single cell Hi-C results.

Acknowledgements

We thank Jacob Sanders for pre-publication access to his BJ1-hTERT and GM12878 Hi-C datasets and Dr. Rebeca San Martin and Rosela Golloshi for their valuable suggestions and feedback in the development of the project.

Funding

This research was supported in part by NIH NIGMS grant R35GM133557 to R.P.M.

References

Amendola, M. and van Steensel, B. (2014). Mechanisms and dynamics of nuclear lamina–genome interactions. *Current opinion in cell biology*, 28:61–68.

Barutcu, A. R., Lajoie, B. R., McCord, R. P., Tye, C. E., Hong, D., Messier, T. L., Browne, G., van Wijnen, A. J., Lian, J. B., Stein, J. L., et al. (2015). Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells. *Genome biology*, 16(1):214.

Bercht Pfleghaar, K., Taimen, P., Butin-Israeli, V., Shimi, T., Langer-Freitag, S., Markaki, Y., Goldman, A. E., Wehnert, M., and Goldman, R. D. (2015). Gene-rich chromosomal regions are preferentially localized in the lamin b deficient nuclear blebs of atypical progeria cells. *Nucleus*, 6(1):66–76.

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(5):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. *Human molecular genetics*, 10(3):211–220.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836.

Cremer, T. and Cremer, M. (2010). Chromosome territories. *Cold Spring Harbor perspectives in biology*, 2(3):a003889.

Cremer, T., Kurz, A., Zirbel, R., Dietzel, S., Rinke, B., Schröck, E., Speicher, M. R., Mathieu, U., Jauch, A., Emmerich, P., et al. (1993). Role of chromosome territories in the functional compartmentalization of the cell nucleus. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 58, pages 777–792. Cold Spring Harbor Laboratory Press.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A. J., and Vapnik, V. (1997). Support vector regression machines. In *Advances in neural information processing systems*, pages 155–161.

Fatakia, S. N., Kulashreshtha, M., Mehta, I. S., and Rao, B. J. (2017). Chromosome territory relocation paradigm during dna damage response: Some insights from molecular biology to physics. *Nucleus*, 8(5):449–460.

Finn, E. H. and Misteli, T. (2019). Molecular basis and biological function of variability in spatial genome organization. *Science*, 365(6457):eaaw9498.

Fritz, A. J., Barutcu, A. R., Martin-Buley, L., van Wijnen, A. J., Zaidi, S. K., Imbalzano, A. N., Lian, J. B., Stein, J. L., and Stein, G. S. (2016). Chromosomes at work: organization of chromosome territories in the interphase nucleus. *Journal of cellular biochemistry*, 117(1):9–19.

Fritz, A. J., Stojkovic, B., Ding, H., Xu, J., Bhattacharya, S., Gaile, D., and Berezney, R. (2014). Wide-scale alterations in interchromosomal organization in breast cancer cells: defining a network of interacting chromosomes. *Human molecular genetics*, 23(19):5133–5146.

Fruchterman, T. M. and Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and experience*, 21(11):1129–1164.

Ganai, N., Sengupta, S., and Menon, G. I. (2014). Chromosome positioning from activity-based segregation. *Nucleic acids research*, 42(7):4145–4159.

Golloshi, R., Sanders, J. T., and McCord, R. P. (2018). Iteratively improving hi-c experiments one step at a time. *Methods*, 142:47–58.

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M. B., Talhout, W., Eussen, B. H., de Klein, A., Wessels, L., de Laat, W., et al. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, 453(7197):948.

Haeussler, M., Zweig, A. S., Tyner, C., Speir, M. L., Rosenbloom, K. R., Raney, B. J., Lee, C. M., Lee, B. T., Hinrichs, A. S., Gonzalez, J. N., et al. (2018). The ucsc genome browser database: 2019 update. *Nucleic acids research*, 47(D1):D853–D858.

Imakaev, M., Fudenberg, G., McCord, R. P., Naumova, N., Goloborodko, A., Lajoie, B. R., Dekker, J., and Mirny, L. A. (2012). Iterative correction of hi-c data reveals hallmarks of chromosome organization. *Nature methods*, 9(10):999.

Javierre, B. M., Burren, O. S., Wilder, S. P., Kreuzhuber, R., Hill, S. M., Sewitz, S., Cairns, J., Wingett, S. W., Várnai, C., Thiecke, M. J., et al. (2016). Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell*, 167(5):1369–1384.

Kind, J., Pagie, L., de Vries, S. S., Nahidiazar, L., Dey, S. S., Bienko, M., Zhan, Y., Lajoie, B., de Graaf, C. A., Amendola, M., et al. (2015). Genome-wide maps of nuclear lamina interactions in single human cells. *Cell*, 163(1):134–147.

Kind, J., Pagie, L., Ortabozkoyun, H., Boyle, S., de Vries, S. S., Janssen, H., Amendola, M., Nolen, L. D., Bickmore, W. A., and van Steensel, B. (2013). Single-cell dynamics of genome-nuclear lamina interactions. *Cell*, 153(1):178–192.

Kuroda, M., Tanabe, H., Yoshida, K., Oikawa, K., Saito, A., Kiyuna, T., Mizusawa, H., and Mukai, K. (2004). Alteration of chromosome positioning during adipocyte differentiation. *Journal of cell science*, 117(24):5897–5903.

Lucas, J. and Cervantes, E. (2002). Significant large-scale chromosome territory movement occurs as a result of mitosis, but not during interphase. *International journal of radiation biology*, 78(6):449–455.

Mehta, I. S., Eskiw, C. H., Arican, H. D., Kill, I. R., and Bridger, J. M. (2011). Farnesyltransferase inhibitor treatment restores chromosome territory positions and active chromosome dynamics in hutchinson-gilford progeria syndrome cells. *Genome biology*, 12(8):R74.

Mehta, I. S., Figgitt, M., Clements, C., Kill, I. R., and Bridger, J. M. (2007). Alterations to nuclear architecture and genome behavior in senescent cells. *Annals of the New York Academy of Sciences*, 1100:250–63.

Oluwadare, O., Highsmith, M., and Cheng, J. (2019). An overview of methods for reconstructing 3-d chromosome and genome structures from hi-c data. *Biological procedures online*, 21(1):7.

Paulsen, J., Sekelja, M., Oldenburg, A. R., Barateau, A., Briand, N., Delbarre, E., Shah, A., Sørensen, A. L., Vigouroux, C., Buendia, B., et al. (2017). Chrom3d: three-dimensional genome modeling from hi-c and nuclear lamin-genome contacts. *Genome biology*, 18(1):21.

Ranade, D., Koul, S., Thompson, J., Prasad, K. B., and Sengupta, K. (2017). Chromosomal aneuploidies induced upon lamin b2 depletion are mislocalized in the interphase nucleus. *Chromosoma*, 126(2):223–244.

Sanders, J. T., Freeman, T. F., Xu, Y., Golloshi, R., Stallard, M. A., Martin, R. S., Balajee, A. S., and McCord, R. P. (2019). Radiation-induced dna damage and repair effects on 3d genome organization. *bioRxiv*.

Sati, S. and Cavalli, G. (2017). Chromosome conformation capture technologies and their impact in understanding genome function. *Chromosoma*, 126(1):33–44.

Scherer, S. (2008). *short guide to the human genome*. Cold Spring Harbor Laboratory Press.

Schoenfelder, S., Javierre, B.-M., Furlan-Magaril, M., Wingett, S. W., and Fraser, P. (2018). Promoter capture hi-c: high-resolution, genome-wide profiling of promoter interactions. *JoVE (Journal of Visualized Experiments)*, (136):e57320.

Shachar, S., Voss, T. C., Pegoraro, G., Sciascia, N., and Misteli, T. (2015). A high-throughput imaging-based mapping platform for the systematic identification of gene positioning factors. *Cell*, 162(4):911.

Solovei, I. and Cremer, M. (2010). 3d-fish on cultured cells combined with immunostaining. In *Fluorescence in situ Hybridization (FISH)*, pages 117–126. Springer.

Strickfaden, H., Zunhammer, A., van Koningsbruggen, S., Köhler, D., and Cremer, T. (2010). 4d chromatin dynamics in cycling cells: Theodor boveri's hypotheses revisited. *Nucleus*, 1(3):284–297.

Sun, H. B., Shen, J., and Yokota, H. (2000). Size-dependent positioning of human chromosomes in interphase nuclei. *Biophysical journal*, 79(1):184–190.

Tanabe, H., Habermann, F. A., Solovei, I., Cremer, M., and Cremer, T. (2002). Non-random radial arrangements of interphase chromosome territories: evolutionary considerations and functional implications. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 504(1-2):37–45.

Tanabe, H., Küpper, K., Ishida, T., Neusser, M., and Mizusawa, H. (2005). Inter-and intraspecific gene-density-correlated radial chromosome territory arrangements are conserved in old world monkeys. *Cytogenetic and genome research*, 108(1-3):255–261.

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(12):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(13):1731–1744.

Van Steensel, B. and Belmont, A. S. (2017). Lamina-associated domains: links with chromosome architecture, heterochromatin, and gene repression. *Cell*, 169(5):780–791.

Walter, J., Schermelleh, L., Cremer, M., Tashiro, S., and Cremer, T. (2003). Chromosome order in hela cells changes during mitosis and early g1, but is stably maintained during subsequent interphase stages. *The Journal of cell biology*, 160(5):685–697.

Yardımcı, G. G., Ozadam, H., Sauria, M. E., Ursu, O., Yan, K.-K., Yang, T., Chakraborty, A., Kaul, A., Lajoie, B. R., Song, F., et al. (2019). Measuring the reproducibility and quality of hi-c data. *Genome biology*, 20(1):57.

Zhu, Y., Gong, K., Denholtz, M., Chandra, V., Kamps, M. P., Alber, F., and Murre, C. (2017). Comprehensive characterization of neutrophil genome topology. *Genes & development*, 31(2):141–153.