Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

Node features of chromosome structure networks and their connections to genome annotation

Yingjie Xu [a], Priyojit Das [b,c], Rachel Patton McCord [d], Tongye Shen [d,*]

[a] Graduate School of Genome Science & Technology, University of Tennessee, Knoxville, TN 37996, USA
[b] Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
[c] Department of Molecular Biology, Massachusetts General Hospital, Boston, MA 02114, USA
[d] Department of Biochemistry & Cellular and Molecular Biology, University of Tennessee, Knoxville, TN 37996, USA

ARTICLE INFO

Keywords:
Graph representation
Network analysis
Chromosome structure
Lamina association
Inverted nucleus
Chromatin state

ABSTRACT

The 3D conformations of chromosomes can encode biological significance, and the implications of such structures have been increasingly appreciated recently. Certain chromosome structural features, such as A/B compartmentalization, are frequently extracted from Hi-C pairwise genome contact information (physical association between different regions of the genome) and compared with linear annotations of the genome, such as histone modifications and lamina association. We investigate how additional properties of chromosome structure can be deduced using an abstract graph representation of the contact heatmap, and describe specific network properties that can have a strong connection with some of these biological annotations. We constructed chromosome structure networks (CSNs) from bulk Hi-C data and calculated a set of site-resolved (node-based) network properties. These properties are useful for characterizing certain aspects of chromosomal structure. We examined the ability of network properties to differentiate several scenarios, such as haploid vs diploid cells, partially inverted nuclei vs conventional architecture, depletion of chromosome architectural proteins, and structural changes during cell development. We also examined the connection between network properties and a series of other linear annotations, such as histone modifications and chromatin states including poised promoter and enhancer labels. We found that semi-local network properties exhibit greater capability in characterizing genome annotations compared to diffusive or ultra-local node features. For example, the local square clustering coefficient can be a strong classifier of lamina-associated domains. We demonstrated that network properties can be useful for highlighting large-scale chromosome structure differences that emerge in different biological situations.

I. Introduction

The chromosomes of eukaryotes are arranged inside the nucleus with a multi-scale architecture of complex folds which can be important for their biological functions [1]. The long-range spatial structures of chromosomes can contribute to the regulation of gene expression by fostering enhancer-promoter contacts locally through loop extrusion or over longer distances (and between chromosomes) through spatial compartmentalization [2,3]. Variations in chromosome structure can both be an important source of functional diversity and a cause of certain pathological traits [4–6]. Chromosome structure capture methods have enabled the study of chromosome organization at the genome scale [7,8]. Hi-C is a high-throughput chromosome 3D structure capture method that can reveal population-based (bulk) chromosome structures expressed by pairwise contact interactions [9,10]. A typical Hi-C analysis uses a two-dimensional contact map that indicates the frequency of contacts between all possible pairs of genomic positions (genomic bins), as illustrated in Fig. 1. However, this two-dimensional matrix may not be intuitive to grasp. It can be difficult to visualize the correlation between the spatial structure and biological annotations, which are often associated with individual genomic regions. Therefore, extracting one-dimensional (1D) indicators from a two-dimensional (2D) contact map can be a useful way to characterize chromosome structural features and compare them to other known linear genomic features, and to discern subtle differences between cell types or upon perturbations.

[*] Correspondence to: Department of Biochemistry & Cellular and Molecular Biology, The University of Tennessee, Knoxville, TN 37996.
E-mail address: tshen@utk.edu (T. Shen).

https://doi.org/10.1016/j.csbj.2024.05.026
Received 13 December 2023; Received in revised form 14 May 2024; Accepted 15 May 2024
Available online 17 May 2024

One commonly adopted 1D indicator is A/B compartment strength (which is often further discretized to a binary A/B classification) for each genomic bin position[9]. A/B compartment strength is obtained from the principal component analysis (PCA) of bulk Hi-C-derived contact matrices. Other methods exist for rendering 2D contact matrices to 1D genomic bin-based information, such as the insulation score and the directionality index to define topologically associating domain (TAD) boundaries [11,12], the gene association domain (GAD) score [13], the distal-to-local ratio (DLR) for inferring the compaction of chromatin region, and interchromosomal fraction (ICF) [14], which have been used to measure dynamic properties such as Loss of Structure (LOS) after perturbation [15]. Here, we explore a set of systematic methods for extracting information using abstract graph representation and network analysis. We aim to assess how these mathematical concepts relate to each other as well as the connection between the underlying genome structure and the biological linear annotation.

Discretizing spatial connection (contact interaction strength) and constructing graph representations of biomolecular structures facilitates the application of a family of node-based network properties (Fig. 2a), such as centralities and clustering coefficients, to understand biological functions. Node-based network properties can measure relationships ranging from two-body interaction (such as edge centrality and assortativity) to higher-order ones including three-body (such as local clustering coefficient LCC3), four-body (such as square clustering coefficient LCC4), and many-body interactions (such as eigenvector centrality). At a relatively small scale, network analysis is applied to study chemical structures [16] and the internal interaction of macromolecules such as proteins. Protein structure networks have been extensively discussed and shown to be useful for studying protein structure and dynamics, [17–24]. Beyond single molecules, abstract graph theories have been used to describe the structure and reaction dynamics of molecular networks of chemicals, from the oxidation of oil paintings to the sol-gel phase transition of polymers [25,26]. It is interesting to examine how such abstract graph representations of structures can be connected to biological function and biophysical aspects of complex and heterogeneous chromosome structures.

Previously, many network analysis for genome biology were molecular interaction networks and co-expression networks [27–29]. There is a variety of investigations using the concept of chromosome network [30], such as chromatin interaction networks [30–33]. In the current study, we emphasize two unique aspects of our approach. First, our construction of network focuses on purely physical and structure aspects of chromosome, with node and edge definitions directly and uniformly derived from Hi-C contact matrices, without including any additional biological annotation, supervision, or inferences regarding the structure or sequence information. In contrast, numerous previous approaches have pre-selected “actively transcribed” nodes or promoter-enhancer focused networks. Additionally, we explore a set of node-based features that reflect various ranges of the interaction between a given chromosome region with its spatial neighbors (and neighbors’ neighbors). Some of the metrics are quite local, such as the degree centrality and local clustering coefficients, while others are quite global, such as closeness and betweenness centralities [22,34]. It is interesting to examine whether specific features of networks reflect strongly a specific biological feature and to what extent network analysis can reveal changes in structure due to cell type variation and environmental influences.

The focus of this work is to construct the structure network of individual chromosomes and compare the network features with selected linear genomic features across different cell types under diverse physiological and pathological conditions. One example of the influence of the chromosome territory environment on chromosome internal structures is found when one compares wild-type (WT) thymocytes, lamin B receptor (LBR) mutant thymocytes, and rod cells. Interphase nuclei of WT thymocytes represent a conventional architecture: heterochromatin regions are primarily located at the nuclear periphery, whereas euchromatin (typically, regions that are gene-rich and largely active) resides in the nuclear interior. FISH experiments [35] have revealed that

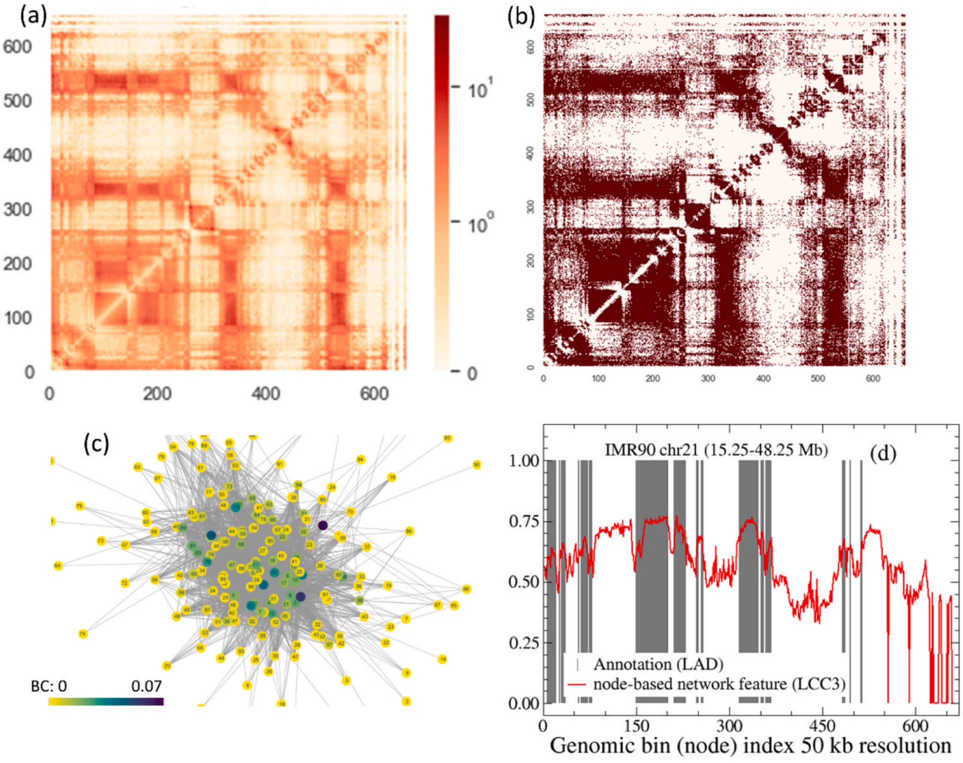

Fig. 1. (a) Hi-C contact map of IMR90 chr21 (region 15.25–48.25 Mb) at 50 kb resolution (b) The corresponding network adjacency matrix representation with a cutoff of 40% coverage (c) The corresponding force-directed network graph representation of panel (b). (d) Network feature (LCC3) output (blue) and LAD annotation (gray) as functions of node (genomic bin) index.

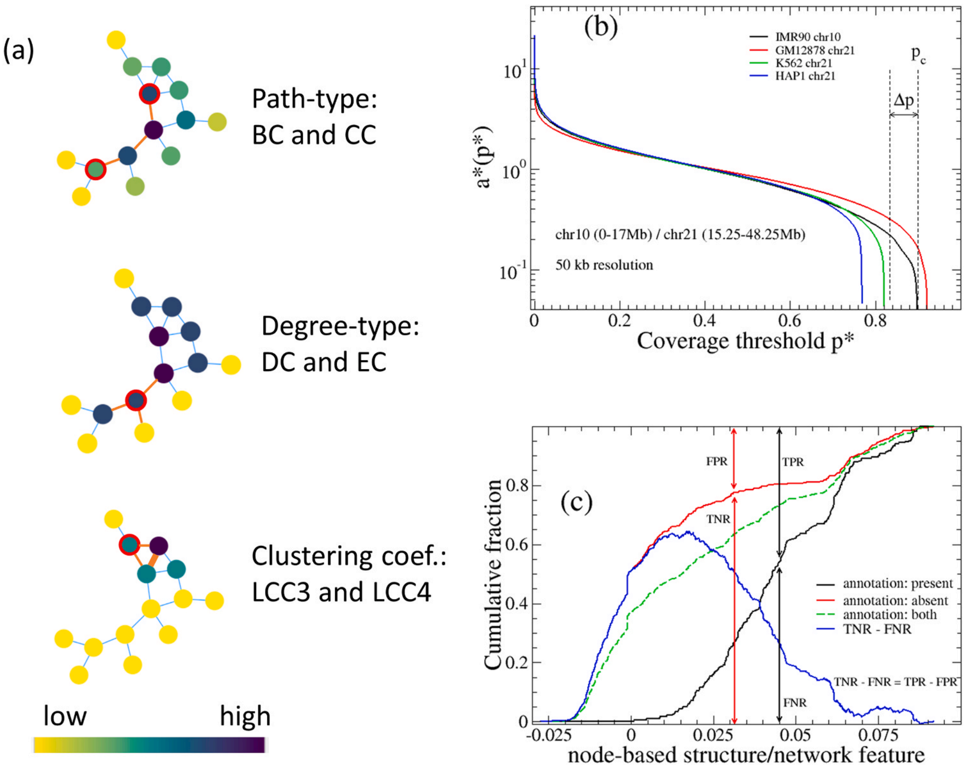

Fig. 2. (a) Pedagogical graphs are shown as illustrations of several network properties. The relative value of network property for each node is color labeled. Both BC and CC are path-oriented measurements. BC counts the paths that go through a target node while CC focuses on the paths terminating at a target node. In contrast, DC and EC focus on the number of neighbors of the target node. LCC indicates the number of closed local paths a node has. (b) The coverage (normalized rank of contact) p^* shown as a function of contact strength a^*. Δp indicates the coverage margin. (c) Cumulative fraction curves for a label being present (black), absent (red), and reference (green). ROC is further defined based on associated concepts TPR, TNR, FPR, and FNR.

the structure is inverted in rod photoreceptor cells of nocturnal mammals, and likewise partially inverted in LBR -/- mutant thymocytes [36]. The previous direct comparison of Hi-C data between such inverted cells and the normal counterpart found straightforward changes between rod cells and WT thymocytes, but the difference between LBR mutant and WT thymocytes is subtle and the A/B compartment representation was unchanged despite the partial inversion. Here, we ask whether distinct features can be extracted by network properties and how a network viewpoint enhances our understanding of this type of biological comparison.

Another application of graph representation is the influence of cell differentiation on chromosome structure. Here, we investigate how chromosome structural features are altered between progenitor cells and differentiated cells using a set of blood cell types [37]. Additional applications include how chromosome structures of haploid cell HAP1 differ from those of ΔCAP-H2 mutant [38] and how HAP1 differs from diploid cell. We also explore the effect of cell cycle on chromosome structures [39], the differences between a cancer cell type and its normal counterpart, and how the nuclear lamina-chromosome interaction is reflected in the chromosome structure network (CSN).

II. Methods and systems

A. *Contact network construction from chromosome structure information*

An abstract graph is a useful way of representing structural components and their relationship. For each graph (network), two types of elements, vertices and edges, are present. The basic structural unit is termed a vertex, which is also called a node in computational sciences or a site in physical sciences. An edge that connects two vertices is termed a link in computational sciences or a bond in physical sciences. We will use these terms interchangeably. In this study, each node is a chromosome region (genomic bin) of size 50 kb unless specified otherwise, whereas each link indicates two such regions are close spatially and their association (contact interaction) can be detected in Hi-C experiments.

The bulk Hi-C data can be represented as a contact matrix A, where a_{jk} indicates the number of contacts recorded between regions j and k, assuming that each region is indexed from 1 to N. There are a total of N nodes in the system. Though self-edges are typically included in a graph representation and are used for downstream analyses, those contacts (diagonal elements) are not considered for the current analysis. Essentially, we would introduce a cutoff value to discretize the values of a_{jk} to either 1 when $a_{jk} \geq a^*$ or 0 otherwise. Here 1 indicates a link (edge) is formed between nodes j and k. Such discretized contact matrix is termed the adjacency matrix of a network. The discretization and construction of a graph representation helps us simplify the relationship between nodes. However, one needs to be careful when choosing a threshold. Multiple methods of defining the threshold exist and each has a different emphasis. For example, the FitHiC method uses a random polymer model as a reference and considers a link formed if the contact value is more than the reference value [40,41]. Such a method might provide different answers depending on the data resolution and it may convert to an overly saturated network in practical cases, especially when the resolution is low and bin size is large.

In this study, we determined the threshold by selecting the median value of non-zero contact strengths, thereby maximizing the Shannon Entropy of edge formation within the reference random network. Essentially, 50% of the $N(N-1)/2$ pair of nodes are connected. In a binary classification, the two extreme cases are an “all-linked” complete graph and a “none-linked” empty graph. Selecting half linked ensures the sharpest contrast with a random network under this constraint. Thus, threshold p = ½ maximizes the entropy of the corresponding random network, S= -[(1-p) ln (1-p) + p ln p]. which enhances the signal-noise ratio. Note that this threshold selection is sensitive to the data quality, i.e., the normalized number of nonzero a_{jk}. A properly selected threshold value a^* sensitively depends on experimental conditions, such as sampling size. Since the meaning of cutoff value a^* is not intuitive for selecting links, we define a concept termed link saturation (coverage) p^*, a quantitative comparison with the complete network, where all edges are formed. With a running cutoff, we can plot the function $a^*(p^*)$ as a curve to connect these two concepts. As shown in Fig. 2b, the j*th* ranked contact with strength a_j contributes to a point (p^*, a^*)= $(\frac{j}{\left\lceil \frac{N(N-1)}{2} \right\rceil}, a_j)$ on the curve $a^*(p^*)$. When a^* is 0, all links are formed and one obtains a complete graph where there are total $N(N-1)/2$ links in the graph. With increasing a^*, some of the links are removed. Parameter p^* can be defined as the ratio between the number of total links of a graph and the link number of a complete graph. We can use a cutoff link saturation (p^*) to specify a^* and further define the contact network, which means only the largest contacts (p^* ×100%) were considered formed. The proper cutoff value ensures that network properties maintain structure information while, at the same time, the cutoff filters out other contacts deemed insignificant.

Note that practically, p^* may not always be set at a high value that is close to 1, as it is affected by how many zero elements the contact matrix has, which in turn is affected by several factors. Certain regions of the chromosome are highly repetitive and cannot be easily resolved. As a result, when there are only total $\mathrm{N_c}$ nonzero values, we can define $p_c = 2\,N_c/[N(N-1)]$. Since $\mathrm{p_c}$ represents the maximum link saturation at a contact threshold of zero, clearly p^* should be less than p_c. Other factors that affect the sampling of a contact map include crosslinking duration and sequence depth. These factors create another layer of uncertainty. Practically, we can define the number of elements a_{jk} that only have a single "hit", N_1, and a safety margin parameter $\Delta_p = 2\,N_1/[N(N-1)]$. Together we have $p^* < p_c - \Delta_p$. Ideally one could choose $p^* = \frac{p_c - \Delta_p}{2}$. Note that the raw (discrete) number of contact counts is only used for selecting proper threshold p^*, whereas a^* requires iterative correction (ICE balancing [42]) to adjust biases in Hi-C data collection before one can render a network representation. Here N_1 and N_c indicate the quality of the sampling. For example, for the whole chr21 of the IMR90 dataset [43] at 250 kb resolution, $p_c = 53.1\%$ and $\Delta_p = 1.2\%$. Thus, one should not choose p^* greater than 50% in such a case. In Fig. 2b, we plot the threshold value a^* vs cutoff link saturation p^* for chr10 (region 0–17 Mb). One can see that a small (tolerant) a^* threshold will include all the nonzero elements of the contact matrix and results in a near maximum coverage p^*. Note that p^* does not necessarily achieve 100% as not all genome bin interactions are detected, especially at higher resolutions. For the case of IMR90 chr10, $p_c = 89.9\%$ and the coverage margin parameter $\Delta_p = 6.6\%$. Ideally one should choose p^* at the flattest region of the $a^*(p^*)$ curve since such cutoff will make the selection least sensitive to the choice of a* . According to the results, it is better to choose a cutoff of 40% under 250 kb resolution. Practically, we use $p^* = 40\%$ unless specified otherwise.

B. Network properties

Once the discretization of the contact matrix is achieved and an abstract graph is constructed (Fig. 1a-c), we can further calculate site-based network properties. For a given network, there are a range of node-based network properties one can construct, from properties that are quite local and reflect only the connectivity of how a genomic bin with its linked neighbors, to those that are global and collective. Here, we mainly study centrality properties, clustering properties, and a hybrid of centrality and clustering properties. As a comparative overview, the two centralities (closeness centrality CC and betweenness centrality BC) that are defined from paths are quite global, while the two eigensystem-based properties, eigenvector centrality (EC) and A/B strength (MI-PCA), are semi-local. Local clustering coefficients (LCCs) are a step more local, and finally degree centrality (DC) is the most local. The order (from global to local node features) is: CC and BC, EC and MI-PCA, LCC4, LCC3, and DC.

Generally speaking, betweenness centrality (BC) measures how often a node appears on the shortest paths between two nodes, while closeness centrality (CC) focuses on nodes being the ends of a shortest path (Fig. 2a). Both properties are path-based parameters and emphasize the global features of the network. In contrast, degree centrality (DC) and eigenvector centrality (EC) measure local properties. Both DC and EC count the number of neighbors of a node, but the main difference is that EC considers a self-consistent weight assigned to each node, and it is more global than DC. Highly connected nodes weigh more than less connected nodes. The local clustering coefficient (LCC) is another way of associating network features to node-based values. LCC is a way of indicating how dense the links are around a node, and in comparison, it is more local than some of the centralities. Specifically, the LCC value for node i is given by $LCC3_i = 2T_i/[D_i(D_i-1)]$ [34]. The term T_i represents the count of triangles that include vertex i and $2/[D_i(D_i-1)]$, the total possible triangles given the degree of ith node, D_i, which defines the number of neighbors of vertex v. Using adjacency matrix, A, one can generalize LCC3 and describe completed neighboring squares for ith node [44,45], $LCC4_i = \sum_{jk} A_{ij} A_{jk} A_{ki}/[D_i(D_i-1)]$ where $D_i = \sum_j A_{ij}$ and $A_{ij} = 1$ when a link exists between i and j, 0 otherwise. As we will demonstrate below, LCC4 (LCC-even, in general) is largely independent from LCC3 (LCC-odd) and especially useful for studying certain types of networks such as bipartite networks [46]. Both provide distinct perspectives of network features. Besides the loop-based generalization we adopted, there are some other expansions of definitions, such as clique-based LCCs [47] that can provide an additional insight on the architecture of networks in general.

C. Statistical analysis

Although site-based contact PCA is not deduced from the properties of a network, it is a popular way of reducing 2D contact map information to site-based values [48]. For MI-PCA and downstream discretization of A/B compartment analysis, each element x_j of the top eigenvector (one with the largest eigenvalue) is associated with each genomic bin j, and the A/B compartment is assigned to each bin depending on the sign of the eigenvector element at that bin. As one group of elements is often associated with gene rich, while the other gene poor, one can assign the signs so that j belongs to the A compartment when $x_j > 0$ and belong to B when $x_j < 0$. The absolute value $|x_j|$ also is used to symbolize the strength of the compartment.

We would like to quantitatively compare local, genomic bin-associated network properties and linear annotations of linear genome information, which are often expressed as discrete (even binary) states. The conventional scatter plot and correlation coefficients may not work in such cases. Instead, we report the area under the cumulative fraction curves (CFCs), which are essentially an integration of the raw scatter plot function.

For each pair of network properties and binary annotation information, one can construct three cumulative fraction curves, $\mathrm{CF_P}$ (black), $\mathrm{CF_N}$ (red), and $\mathrm{CF_R}$ (green dashes), as illustrated in Fig. 2c. Each genomic bin position has a network value, e.g., LCC3 and a binary annotation, e.g., LAD or non-LAD. For $\mathrm{CF_P}$, the cumulative fraction curve examines the positive response and increases by a fraction of $dy = 1/N_p$ at position $\mathrm{x = LCC3_i}$, when ith genomic bin is a LAD and $dy = 0$ when it is not a LAD. Here N_p is the total number of nodes with LAD status. Conversely, $\mathrm{CF_N}$ examines the negative response and increases when a bin is not a LAD by $dy = 1/N_n$. Here, parameter N_n is the total number of non-LAD nodes. The reference curve $\mathrm{CF_R}$ increases by $dy = 1/(N_n + N_p)$ at $\mathrm{LCC3_i}$ regardless of LAD status. When we make a hypothesis that larger values of LCC3 are associated LAD status, the value of CF_P is the false negative rate (FNR) while 1-FNR is true positive rate (TPR). Similarly, one can obtain the true negative rate (TNR) and false positive rate (FPR) from the CF_N curve. We can further define TNR-FNR (=TRP-FRP) and find an optimal threshold value to classify these two states (LAD or not) from LCC3 values. Note that the positive likelihood ratio LR+ =TNR/FNP is a similar method of discerning these two curves. Further, the widely applied receiver operating characteristic (ROC) curve is plotted (x, y) = (FPR, TPR). The area under the ROC curve (AUC) is a value between 0 and 1, where a larger value implies a stronger performance measurement for the classification of an annotation from a given network property.

D. Systems and data

1. Hi-C data of human cell lines**:** All structural data of chromosomes used in this study came from bulk Hi-C data. In this study, we used publicly available (www.ncbi.nlm.nih.gov/geo) Hi-C data from the following human cell types: lymphoblast GM12878 (GSE63525) [43], fibroblast IMR90 (GSE63525) [43], leukemia cell line K562 (GSE63525) [43], HAP1 (GSE95014) [38] – a haploid cell derived from K562 [49], condensin II deletion in HAP1 cells (GSE163625) [38] and Hi-C data from different blood cell types: erythroblast, neutrophil, and megakaryocyte, with permission from the PCHI-C Consortium [37].

   Hi-C raw data is expressed as a contact frequency between regions of the genome at a specific resolution (genomic bin). With proper procedures, one can describe the information as a contact matrix. Unless otherwise specified, all Hi-C data came through MAPQGE30-filtering and was further normalized by ICE balancing [42]. Furthermore, we apply a sequence distance normalization, i.e., $a_{ij} \rightarrow a_{ij}/b_k$ with $b_k = \sum_{ij} u_{ij} \times \delta_{|i-j|-k} / \sum_{ij} \delta_{|i-j|-k} = \sum_i u_{i,i+k}/(N-k)$, where Kronecker delta selects k = |i −j|, as stated in Ref. [48]. Previously, we showed that unlike ICE balancing which is a necessary site-based correction, both distance-normalized and untreated versions can be useful, where the former is focused on interactions that occur more often than expected in a random polymer, and therefore perhaps are mediated by specific biological mechanisms, while the latter represents how often two regions actually come in contact in a nucleus, which has implications for biochemical reactions [48]. Given our focus on studying structure patterns, we opted for distance normalization for all data presented, except in the case of chicken mitotic chromosome, where we compared both.

   We focus on intra-chromosomal interactions in this study, and human chromosome 21 is examined by default. Since a part of the chr21 is poorly covered by Hi-C experiments due to its repetitive nature, we used the 15–48 Mb region. To generalize our conclusions, we also examined another (larger) chromosome, chr10 (~ 0–17 Mb) for a comparison and obtained consistent results.

   There are many interesting linear annotation data on chromosomes available (some are direct measurements and others are derived): chromosome histone modification marks, lamina-associated domains (LADs), gene expression, and derived information such as chromosome sub-compartments and chromatin states. We focus specifically on examining chromatin state, LADs, and compartment subtype in this work. We utilized these biological properties since they are closely associated with the structural properties of chromosomes and thus may manifest as network properties or show biological significance.

2. Lamina-Associated Domain: Lamin is a nuclear inner membrane protein that is critical for various biological processes within the nucleus, such as chromatin organization, DNA repair, and gene expression. In mammals, both A- and B-type lamins at the nuclear periphery interact with hundreds of large chromatin domains known as lamin-associated domains (LADs) [50]. LADs might be closely tied to geometrical properties of chromosome structure that are therefore reflected in a graph representation. LADs also change their spatial localization in response to cell type-specific gene expression during differentiation and development [51]. For this study, LAD data came from Tig3 fibroblasts [48] for comparison to the similar IMR90 fibroblast Hi-C data.

3. Chromatin state: Chromatin states (CSs) map epigenomic marks such as histone modifications, histone variants, open chromatin regions, and other associated marks to their likely functional roles in gene regulation [52]. Each region of the chromosome is assigned to only one of the 15 CS states according to their combination of epigenetic marks: CS1 =Active Promoter, CS2 =Weak Promoter, CS3 =Poised Promoter, CS4 =Strong Enhancer 1, CS5 =Strong Enhancer 2, CS6 =Weak Enhancer 1, CS7 =Weak Enhancer 2, CS8 =Insulator, CS9 = Transcription Transition, CS10 = Transcription Elongation, CS11 =Weak Transcription, CS12 =Repressed, CS13 =Heterochromatin/low signal, CS14 =Repetitive/ Copy Number Variation, CS15 =Repetitive/ Copy Number Variation. We used CSs of GM12878 in this study.

   Note that the annotation state of a genome sequence, such as chromatin state, exists at a different (and often higher) resolution than that of the network nodes. In the current study, CS data has a much higher resolution of 200 bp. When we examine cumulative fractions and make comparisons between chromatin states and network properties, we operate at the lower resolution (e.g., 50 kb) of the two datasets and directly assign the annotation state of the exact genome position of the node (the lower bound of the corresponding genome bin). Our straightforward definition works well for correlation calculations. However, in our attempt to illustrate the statistical features of network properties for a particular chromatin state using a distribution (e.g. a violin plot), such as in the case of CS3 (poised promoter), we encountered a challenge since this annotation is typically very small and is not always assigned to the genomic bin node. To resolve this issue, we practically operate at the higher resolution (200 bp) mode and interpolate by assigning the same network property value to all bins that fall within each 50 kb range (one structure node). Thus, when constructing the violin plots, multiple chromatin state labels can be assigned to the same node and quantitative weights ensured.

4. Subcompartments**:** In addition to the binary classification of A/B compartment (largely correlated with euchromatin vs heterochromatin) [53], additional studies classify regions into multiple compartment subtypes (sub-compartments) using a range of data, including interchromosomal contacts, strength of PCA eigenvectors, gene expression, and epigenetic information [43,54–58]. Here, we compared our network properties across several cell types using two sub-compartment definitions, which extends the A/B definition to six states, A2, B1, B2, B3 and B4 [43,53].

5. Mouse Interphase nuclei**:** A partially inverted architecture was observed when lamin B receptor is deleted from thymocytes. Compared to WT thymocytes, these LBR mutant cells cannot anchor lamin associated domains onto the nuclear membrane [36]. Though it is relatively simple to discern the differences in Hi-C maps from WT thymocytes (conventional) and rod cells (inverted) because these are quite different cell types, it is difficult to distinguish between LBR -/- (partially inverted) and WT (conventional) architecture using Hi-C contact maps directly. Therefore, we examine whether network properties can distinguish these two nuclear organizations based on their subtle changes in Hi-C data. We used the gene-dense mouse chr11 for this study. The original data resolution is 20 kb, and we use 200 kb for better statistics in the network analysis.

6. Chicken mitotic chromosome: Chromosomes go through dramatic structure transformations during the cell cycle. We study CSNs derived from Hi-C data of mitotic chicken chromosome 21 during the transition from G2 into prophase (time = 0, 2, 5, 7, 10, 30, 60 min)

from Ref. [39]. ICE balance and distance normalization were applied. The resolution of the contact matrices is 40 kb.

III. Results and discussion

A. *Effects of coverage and resolution on chromosome structure network properties*

We first report how resolution and threshold may affect the analysis using chr21 of IMR90 (region 15.25–48.25 Mb) as an example. By varying coverage parameter p^*, we found that a higher cutoff value such as 80% results in constant high values of the node-based network property LCC3, as the nodes of the CSN are almost fully connected. Conversely, if the threshold p^* is as low as 20%, the resulting curve has larger variation with some nodes being zero, as the CSN has fewer links and is broken into disconnected smaller subnetworks (Fig. 3a). These results suggest a p^* threshold of 40%, the same as chosen based on observations from Fig. 2b, is also appropriate to yield network properties that vary meaningfully across the chromosome.

Resolution also has an impact on network properties, as shown in Fig. 3b. For example, there is a potentially spurious high peak for LCC3 values near 37.5 Mb (bin index 89). However, the peak disappears if the resolution is 50 kb (bin index 445), when the genomic bin size is decreased from 250 kb to 50 kb. We observe that the resolution has a relatively small effect on EC, which indicates eigenvector centrality is less sensitive to resolution selection since EC is more of a semi-global property compared to local properties such as LCC3. Generally, a smaller bin size provides a high-resolution description only if there are sufficient samples in each bin.

We next compare the similarities and differences in the quantified network properties, LCC3, EC, BC, CC, and LCC4, using the same parameters (250 kb bins and p^*= 40%) (Fig. 3c). BC exhibits the most frequent variations, while the other four properties show changes more gradually across the genomic region. Overall, BC values are nearly constantly low while CC values are constantly high. EC and LCC3 have a wider range of variation. However, their peaks and valleys are not correlated, which indicates that different network properties reflect distinct features of the CSN. Using several different network properties may provide a more comprehensive understanding of chromosome structure. For example, at positions around 35,000–37,500 kb (arrow, Fig. 3c), EC is almost zero, the LCC4 value is around 0.3, and the LCC3 value is high. The low CC and EC suggests that the region has few neighbors, while the contrast between LCC3 and LCC4 values indicates that there is a short-distance ring structure in the area. Thus, this region is like an isolated island, likely to form links within itself rather than with others.

The network properties provide different measurements of the structure features. One can further define a derived network property as a combination of basic network properties: BC, CC, EC, and LCCs. A previous study pointed out possible correlations between derivative network properties, such as BC vs DC × (1 − LCC3) [59]. We directly compared the correlation between different network properties and between network properties and A/B compartment strength in Supplementary Material (SM) Fig. S1. Among all the basic network properties studied, the majority show little correlations, indicating that they are largely independent metrics. Only a few pairs of properties show correlation in restricted regions. For example, LCC4 and A/B compartment strength show correlation at high value regions whereas EC and A/B correspond at low value regions. We found only one strongly correlated pair: EC and DC × LCC3. Additionally, as shown in SM Fig. S2, a derived network property, the LCC4/LCC3 ratio, is correlated to the A/B compartment classification. This ratio compares the relatively longer connectivity (neighbors of nearest neighbors) to short-range connectivity (nearest neighbors). Nodes with high LCC4/LCC3 ratios are likely a part of the inactive region (B compartment) and vice versa. This distinction further underscores the utility of graph properties in

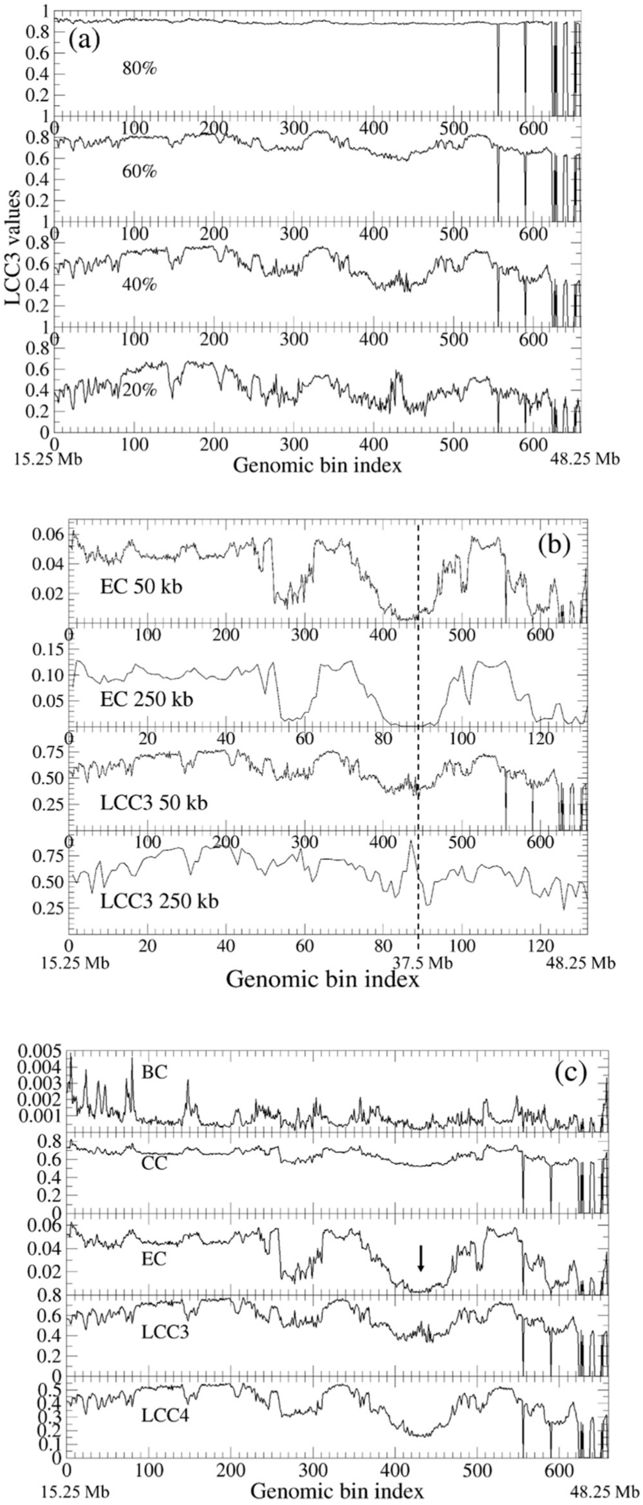

Fig. 3. (a) LCC3 value as a function of node (genomic bin) index, under different coverage thresholds p^* (max, 80%, 60%, 40%, 20%) at 50 kb resolution for IMR90, chr21:15.25–48.25 Mb. (b) IMR90 chromosome 21 (same region as above) 40% p^* cutoff EC and LCC3 values at different resolution (50 kb vs 250 kb bin sizes). (c) Comparison of different network properties including BC, CC, EC, LCC3 and LCC4 across the same IMR90 chr21 region as in a and b. Arrow indicates a notable region of divergent values between network properties at a given genomic location.

characterizing of biologically related structure features.

B. Network properties display distinct structural features of chromosomes for different cell types

There are several epigenetic differences between chromosomes of different cell types. We evaluated whether network properties and CSN can discern the structural differences that result from epigenetic differences. We first compared the CSN of chr21 between a normal cell type (GM12878, non-cancerous lymphoblast) and a cancer counterpart (K562; leukemia). As shown in Fig. 4a, the overall fluctuation of LCC3 and LCC4 values of K562 is much more subdued than the corresponding fluctuation of normal cells (GM12878). This plateaued pattern, which does not vary significantly between sequence neighbors, suggests that many nodes on a chromosome have a similar (and even) degree of connectivity with other nodes in the network. This loss of network features in bulk cancer cell data may reflect the loss of functional organization of the chromosome. Therefore, unlike normal cells which have highly organized chromosome structure, the degree of chromosome organization in cancer cell K562 is weakened. Although most of the remaining peak and valley positions of K562 are similar to that of GM12878 (especially for EC), there are exceptions. For example, K562 is missing the peak at about 37 Mb for LCC3, potentially a site of more dramatic chromosome structural difference.

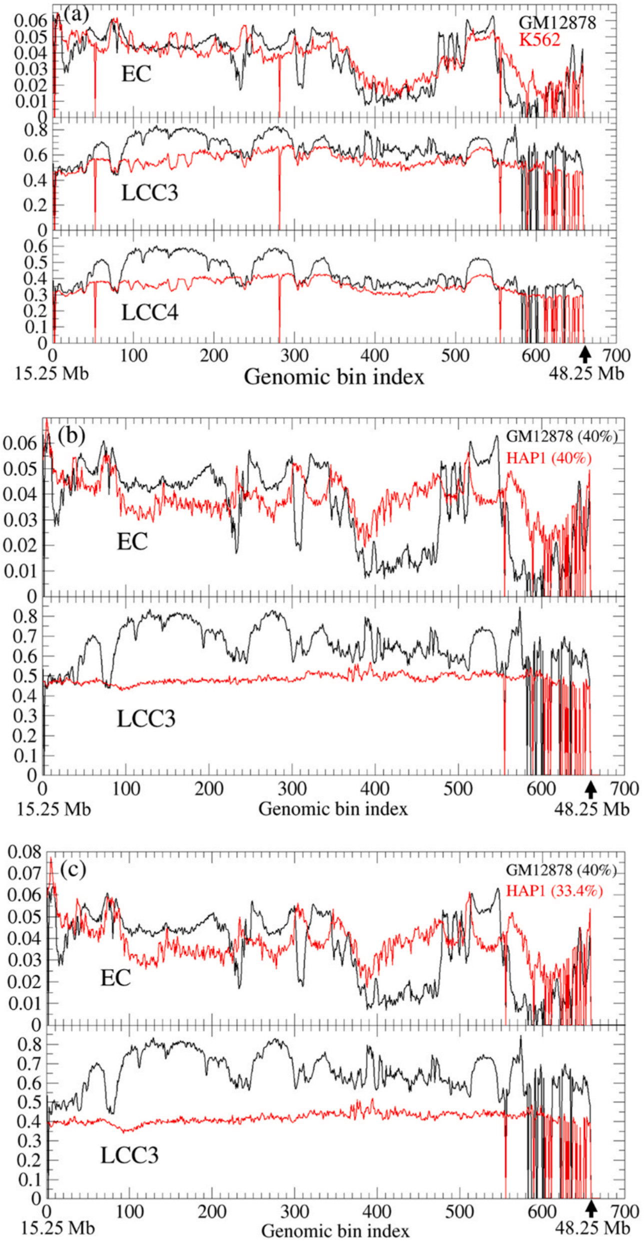

Fig. 4. Network properties (EC and LCC3/4) as functions of network node index for cancer cell type K562 are shown in (a) and for haploid cell HAP1 in (b). The corresponding results for lymphoblast GM12878 are shown as a comparison. Data of region 15.25–48.25 Mb for chr21 (50 kb resolution) is used. Constant coverage of 40% is used. (c) Same comparison as (b) but HAP1 coverage is set to 33.4%, so that p^*/p_c is kept constant.

In addition to changes in chromosomal structure in cancer, we also examined the impact of interactions between homologous chromosomes on genome structure. In this case, haploid cells HAP1 vs. diploid cells GM12878 were compared. HAP1 has only a single copy of each chromosome except chromosome 15 (the second copy of chr15 is only an incomplete fragment). Therefore, chromosomes of haploid cells may lose some of the interchromosomal interactions of the corresponding diploid cells. We constructed contact networks for HAP1 using the same cutoff ($p^* = 40\%$) and displayed the EC and LCC3 results in Fig. 4b. Note that there is a disparity in the coverage of Hi-C sequencing data, where GM12878 has a higher p_c than HAP1 ($p_c = 92.5\%$ vs 77.2%). Another aspect to consider when interpreting HAP1 and K562 results is the translocations that occur in these cell lines [60]. For example, chr21 and chr12 exhibit translocations in the region near chr21:27 Mb and chr12:23 Mb in K562. These genomic rearrangements can affect contacts near the edges of the translocation region.

When using the same cutoff p^*, the system with lower p_c has fewer dispensable edges to choose and thus leads to a structure network that demonstrates an ensemble average of configurations, an evenly connected global network. Since p_c (data coverage) influences the effect of p^*, we also used an alternative definition, a scaled p^* to compare different systems. Specifically, instead of using a constant p^*, we use different p^* for different systems while keep $\frac{p^*}{p_c} =$ constant. For this case, we make sure that $\frac{p^*_{HAP}}{p_{cHAP}} = \frac{p^*_{GM}}{p_{cGM}}$. As shown in Fig. 4c, we set $p^* = 33.4\%$ to construct the network of HAP1, to ensure 33.4%/77.2% = 40%/92.5%. When we compare Figs. 4b and 4c, the results are not sensitive to the selection of p^*, i.e., lowering the cutoff value of HAP1 cells, did not significantly affect LCC3 and EC values. Similar to the cancer cell K562, from which HAP1 is derived, the LCC3 values are almost constant, and each position has a similar number of random contacts with its surrounding network. Subsequent tests showed that even if p^* is lowered to 20%, the LCC3 value of HAP1 would still be largely constant with slightly more random fluctuation (SM Fig. S3). Therefore, this phenomenon is robust and not an effect of p^*, which suggests the lack of well-defined chromosome structure features in HAP1. Since HAP1 is both haploid cell and originates from cancer cell line, it is unclear whether one or both factors contribute to the phenomenon. However, upon examining the EC values, we observe that HAP1 exhibits further attenuation of the organizational structure (for example, in the 19.5–24 Mb region, SM Fig. S3) compared to that of K562. This structural changes might be caused by the haploid nature of the HAP1.

We also studied a HAP1 mutation (ΔCAP-H2) [38], as shown in SM Fig. S4. CAP-H2 is a subunit of condensin II, which is involved in chromosome looping and condensation. Previous research reported large scale chromosome structure changes and notable changes in interchromosomal contacts with CAP-H2 deletion [46]. However, local changes are difficult to discern from the Hi-C contact maps alone (SM Fig. S4). Although condensin II primarily affects regions near centromeres, our selected region (q-arm of chr21) still exhibits clear effects of this mutation using network analysis. We observed that notable changes occur in LCC3 and LCC4 while EC is largely preserved. Their values decreased across the entire chromosome, indicating a general loss of intermediate-scale connectivity along the chromosome, consistent with the loss of loops formed by condensin. Additionally, the CAP-H2 mutant cells exhibited greater variations in LCC3 and LCC4 compared to the WT, indicating that certain genomic regions are more strongly influenced by the loss of condensin than others. Therefore, network properties can reveal differences in decondensation after condensin deletion, which may not be readily apparent from the 2D contact map alone.

Rod cells in mice have a unique inverted nuclear structure. It is difficult to visualize the difference between the "inverted" and "non-inverted" structures from the Hi-C contact map directly. While distance-based methods, such as image-based chromatin tracing [61], may be better at discerning this structure difference than Hi-C, we demonstrate that CSN metrics can still assist in revealing differences in the Hi-C information,. We analyzed the differences between WT type and LBR mutation using the chromosome structure networks. As shown in Fig. 5, BC, CC and EC show a similar distribution pattern of high values at both ends and low in the middle. Conversely, LCCs are higher in the middle than at both ends. This suggests that the nodes in the central region of mouse chr11 have fewer neighbors but are more locally connected, whereas the nodes at the ends are more urban. A scenario of long-distance enhancer-promoter interaction at the ends and highly correlated regulation complexes in the middle can be a realization of such CSNs. Compared with the wild type, the value of network properties fluctuated less in the LBR mutant. Similarly, the LCC values of LBR mutant are lower than that of wild type, while BC and CC do not change much, which suggests that the LBR mutation has weaker short-distance contacts and a loss of some local organization. Thus, chromosomes of LBR mutant cells have weakened short range contacts compared to those of WT cells.

As a further comparison of related cell types, we compared three cell types from different branches of the same hematopoietic lineage: erythroblasts, neutrophils and megakaryocytes. As indicated in the SM Fig. S5, the network properties of erythroblast and megakaryocyte cells are relatively close, whereas neutrophils are far from both erythroblasts and megakaryocytes. For example, the BC and LCC values for neutrophil structure have extremely high peaks around chr10:12.5 Mb (50 bin index). Such a relationship is consistent with the differentiation path of the cell types. Both erythroblast and megakaryocyte cells can further give rise to additional cell types (such as erythrocytes and platelets) in processes that involve the loss of the nucleus, while neutrophils are at a terminal stage [38]. This difference may be reflected in the increased structure specialization of the neutrophil chromosomes. These selected three cell types exhibit significant overall differences from each other, raising the question of how effectively network properties can distinguish between more closely related cell types. To explore this question further, we also compared a pair of more closely related cell types, B cell [59] and GM12878 in SM Fig. S5 (d) and (e). We observe that node features are much closer to each other than those in (a-c), yet certain notable differences persist.

To evaluate how CSN metrics perform with dramatic shape changes in entire chromosomes, we examined the transformation of chromosome structure from interphase through prophase to metaphase. We used Hi-C data from chicken chromosome 21 at different stages of chromosome condensation (time = 0, 2, 5, 7, 10, 30, and 60 min) [39] to illustrate the large scale chromosome dynamics. As shown in SM Fig. S6, we compared both distance decay normalized (panel 1) and nonnormalized (panel 2). We found that the node features of CSN do not simply disappear over time, even for the normalized cases. Instead, several types of structure rearrangements occur over time. For example, EC and LCCs indicate a type of structural reorganization occurs at 5 min and another one at 30 min. However, creating a distinct network from uniformly condensed mitotic chromosomes (at later timepoints) becomes more difficult and the discretized contact maps at these stages show few distinctive features. The corresponding results without normalization show that the EC metric can capture certain global features of the structural changes as chromosomes progress to condensed metaphase stages. Overall, we found that LCCs did not reveal hidden features overlooked by a direct inspection of the 2D contact maps. This might be a limitation of the network metrics which focus on the connectivity of only a few nodes and not sensitive to drastic global changes.

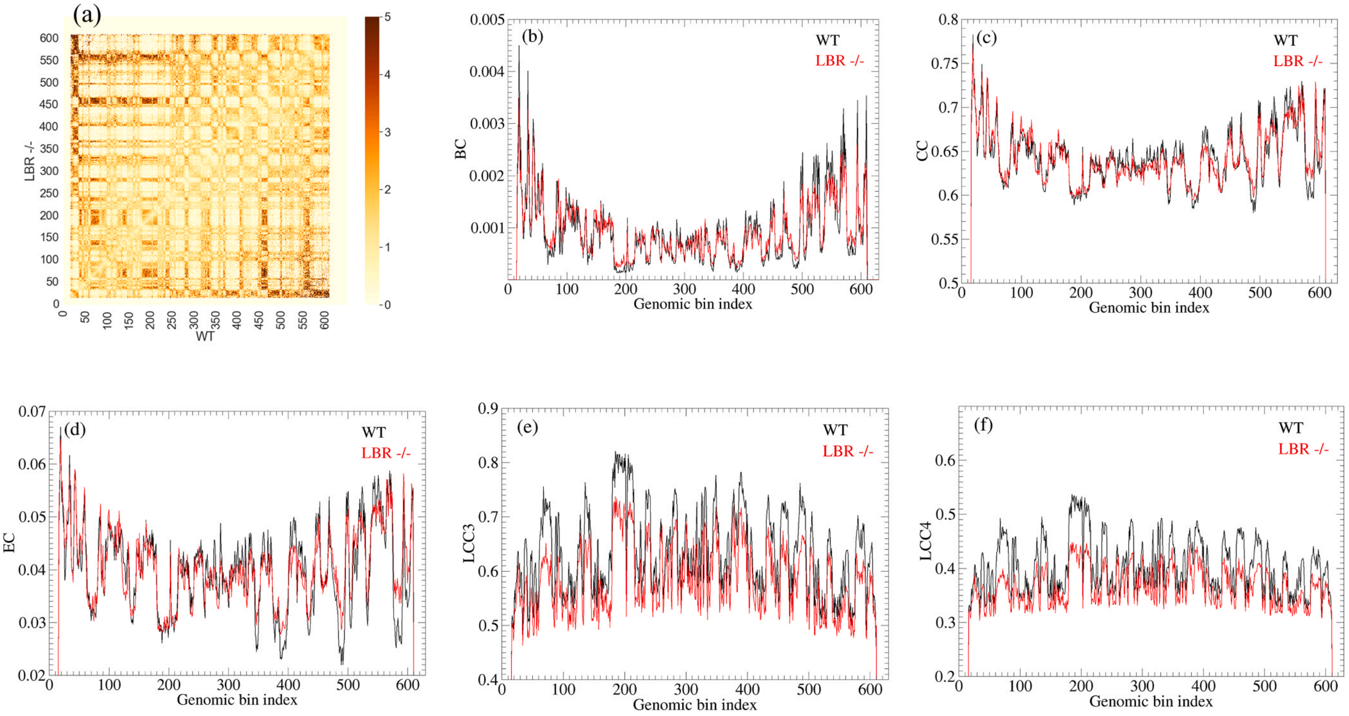

Fig. 5. The structures of chr11 of mouse rod photoreceptor cell wildtype (WT) and LBR mutant are characterized using network properties at a resolution of 200 kb. (a) Heatmap comparison. Above the diagonal is LBR -/- cell. Below the diagonal line is WT cell. (b) BC value comparison between two the above cells. (c) CC value comparison. (d) EC value comparison. (e) LCC3 value comparison. (f) LCC4 value comparison.

C. The connection between network properties and genomic annotations

Having observed that quantitative CSN properties differ between cell types and conditions, we also evaluated whether these properties of the structural network relate to other biologically significant linear annotations along the genome. A major advantage of CSN is the potential for data reduction. Various node-based network properties render the complex structure information (two-dimensional contact matrix originated from Hi-C) into different one-dimensional functions of positions (genomic bins). In this way, it is convenient to compare 1D node properties with known genomic linear annotations. Indeed, previous work has observed that Hi-C network properties such as node degree can vary between promoter and enhancer regions [62].

In our study, we used LAD status and chromatin states as examples to study the relationship between genomic linear annotations and network properties of the structure (LCC3, LCC4, EC, BC, and CC), focusing on IMR90 cells, as shown in Fig. 6. We also compared LAD annotation with A/B compartment strength (top eigenvector MI-PCA of Hi-C contact [48], up to a sign, Fig. 6c), where we observe that LAD labels are more likely associated with compartment B. Note that we did not fix the sign, and compartment B is positive (and A is negative) for this top eigenvector of MI-PCA. Though the A/B compartment value is not a node-based network property, it nevertheless provides a value associated with each genomic position. Researchers have routinely used it to classify compartments, while more advanced methods of classification have been developed in recent years [57]. As described in the Methods, we used cumulative fraction curves (CFCs) to discern the separation of LAD states (LAD versus non-LAD) by different node network properties. As shown in Fig. 6a, EC can clearly separate LADs from non-LADs. The genomic bins associated with LAD labels are more likely to have high EC values. Similar results can also be observed for the CFC of LCC4 in Fig. 6b. In contrast, although MI-PCA can also distinguish LAD and non-LAD regions at low values, the three CFC curves representing LAD, non-LAD and reference overlap at high values (A compartment) (Fig. 6c).

The comparison of how well each network property can distinguish LADs from non-LADs can be seen from the ROC graph in Fig. 6d. When an ROC is high (thus giving a larger AUC value), there is a strong separation of LAD vs. non-LAD by this metric. LCC4 values (AUC = 0.852) are the strongest LAD predictor, followed by EC and MI-PCA, while LAD regions show no distinction based on BC values. Except BC, all five other network properties are associated with differences in LAD classification. Both LCC3 and LCC4 outperform others, including the A-B compartment signal, in resolving LAD classification especially at the small false positive rate region. This suggests that LADs regions have characteristic modes of local clustering within the chromosome structure that are even more distinctive to LADs than B compartment association in general. This observation confirms the importance of quantifying different network properties to study chromosome structure and biological annotation. By examining various network properties that range from global to local, we find that LADs seem to have more distinctive local rather than global network properties.

Besides comparing LAD and MI-PCA (A/B compartment strengths), we also examined the correlation between A/B compartment subtypes and network properties. The difference between A1 and A2 subcompartments is that A1 has higher CG content and shorter genes. B1, B2, and B3 differ in their association with histone modifications. According to the available data, chr21 of IMR90, GM12878, and K562 cells associates with up to four sub-compartments A1, A2, B1, and B2. By comparing LCC3 and LCC4 with A1, A2, B1 and B2 classifications, we found that network property LCCs can distinguish different compartment subtypes in IMR90 (SM Fig. S6). For both subcompartment definition approaches for IMR90 [43,53], LCC3 and LCC4 values show significant distinctions between A1 and B2 regions, with A2 and B1 regions showing intermediate LCC3 and LCC4 values. This pattern is also evident for LCC4 values in GM12878 and K562 cells, although for GM12878 cells LCC3 values are much more similar across all subcompartment types.

Additionally, we compared chromatin states with network properties. Chromatin states classify chromatin regions by properties such as their epigenetic marks and transcription state to identify their specific functions in transcriptional regulation [63]. The chromatin state data of GM12878 chr21 (GEO:GSM936082) was used in this comparison. As shown in Figs. 7a and 7b, we show cumulative curves with respect to

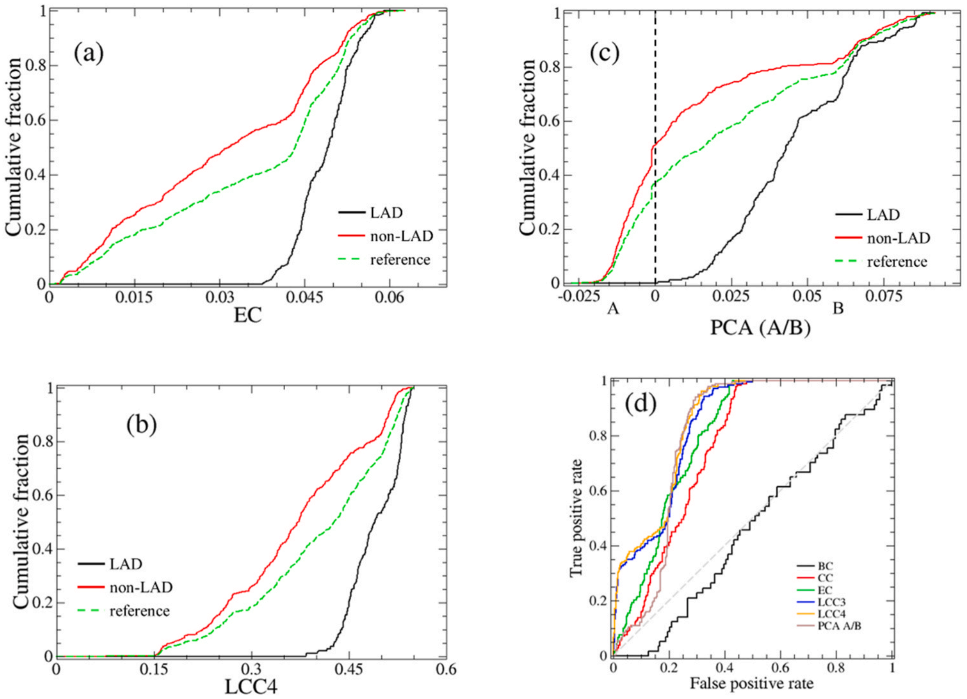

Fig. 6. LAD cumulative fraction (CF) plots of network properties and MI-PCA for a region of IMR90 chr21. Here classifier performance concepts using CF_P (black), CF_N (red), CF_R (green dash) are shown for EC in (a), LCC4 in (b) and PCA in (c). (d) ROC curves of six different network properties for the LAD classification.

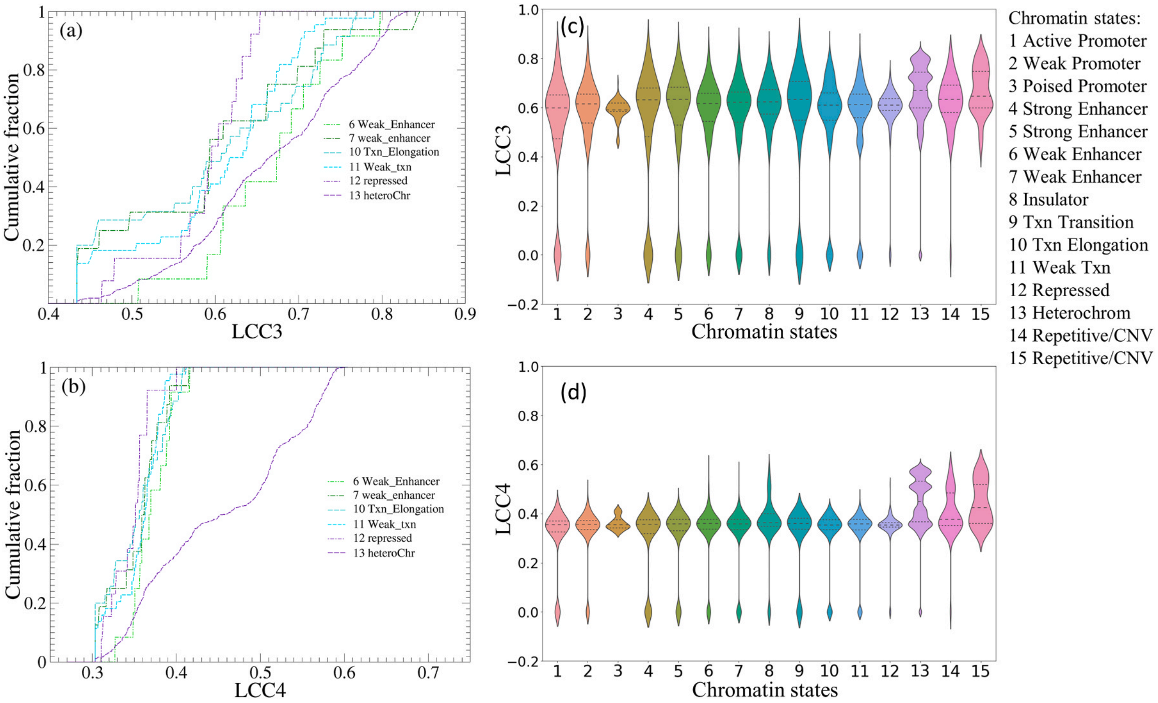

Fig. 7. Using network properties discern selected six chromatin state labels of chr21 GM12878 for LCC3 (a) and LCC4 (b). The corresponding distribution of network properties of all 15 different chromatin states are shown in (c) and (d). The resolution is 50 kb.

LCC values for the six most data-rich chromatin states. Despite the similar nature of the definitions, the accumulation curves of chromatin states in LCC3 and LCC4 are distinct in discerning chromatin state classifications. LCC4 can better discriminate heterochromatin (CS13) versus other states. The distribution of either LCC metric for poised promoter state classification (CS3) is sharply peaked and there are no zero values (Figs. 7c and 7d). This result suggests that the chromatin state label CS3 may have a highly specialized structure motif. This observation may be connected to previous findings suggesting that the promoter-based subnetwork, as opposed to the rest, has a signature of positive assortativity [33]. In addition, we found that some chromatin states, such as heterochromatin and repetitive/CNVs, had more than one aggregation peak at LCC4, which may indicate that there are subtypes of chromosomal structures associated with these CS labels. As previous literature on heterochromatin structure suggests, there is a possible phase separation of heterochromatin [64]. It is an interesting hypothesis that the two sub-states observed for CS15 (heterochromatin) could represent one relatively condensed phase (positive assortativity) and one relatively open structure (negative assortativity) as depicted in Fig. 3c of Ref. [64].

IV. Concluding remarks

We constructed the structural network of chromosomes using Hi-C sequencing data. We found that some node-based network properties, such as LCC3 and LCC4, can be used to characterize important features of chromosome structure. It can render complex two-dimensional interactions into one-dimensional quantitative network properties. Node features of CSN can be used to reveal meaningful structural differences existing between cell types. For example, compared to other network properties and A/B compartment strength (contact PCA), the square local clustering coefficient (LCC4) is a particularly strong classifier for predicting LAD and one specific chromatin state (CS13, heterochromatin). In general, network analysis of CSN can be a useful addition to the growing toolkit for studying bulk Hi-C three-dimensional genome structure information. Similar analyses can also be used in the future to study interchromosomal interactions and chromosome dynamics and probe other types of data such as C-walk [65] and multi-contact capture approaches [66].

CRediT authorship contribution statement

Yingjie Xu: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Rachel P McCord:** Writing – review & editing, Investigation, Data curation, Conceptualization. **Priyojit Das:** Writing – review & editing, Software, Investigation, Data curation. **Tongye Shen:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Conceptualization.

Declaration of Competing Interest

none.

Acknowledgements

We acknowledge insightful discussions with D. Foutch on the network analysis of biomolecular structures. This research was supported in part by National Institute of General Medical Science grant R35GM133557 to RPM. Partial funding for open access to this research was provided by the University of Tennessee's Open Publishing Support Fund.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.05.026.

References

[1] Rowley MJ, Corces VG. Organizational principles of 3D genome architecture. Nat Rev Genet 2018;19:789–800.

[2] Bonev B, Cavalli G. Organization and function of the 3D genome. Nat Rev Genet 2016;17:772.

[3] Robson MI, Ringel AR, Mundlos S. Regulatory landscaping: how enhancer-promoter communication is sculpted in 3D. Mol Cell 2019;74:1110–22.

[4] Liu H, Tsai H, Yang M, Li G, Bian Q, Ding G, Wu D, Dai J. Three-dimensional genome structure and function. MedComm (2020) 2023;4:e326.

[5] da Costa-Nunes JA, Noordermeer D. TADs: Dynamic structures to create stable regulatory functions. Curr Opin Struct Biol 2023;81:102622.

[6] Bruckner DB, Chen H, Barinov L, Zoller B, Gregor T. Stochastic motion and transcriptional dynamics of pairs of distal DNA loci on a compacted chromosome. Science 2023;380:1357–62.

[7] McCord RP, Kaplan N, Giorgetti L. Chromosome Conformation Capture and Beyond: Toward an Integrative View of Chromosome Structure and Function. Mol Cell 2020;77:688–708.

[8] Corsi F, Rusch E, Goloborodko A. Loop extrusion rules: the next generation. Curr Opin Genet Dev 2023;81:102061.

[9] Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science 2009;326:289–93.

[10] Zhang B, Wolynes PG. Topology, structures, and energy landscapes of human chromosomes. Proc Natl Acad Sci 2015;112:6062–7.

[11] Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 2012;485:376–80.

[12] Crane E, Bian Q, McCord RP, Lajoie BR, Wheeler BS, Ralston EJ, Uzawa S, Dekker J, Meyer BJ. Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature 2015;523:240–4.

[13] Zhang C, Xu Z, Yang S, Sun G, Jia L, Zheng Z, Gu Q, Tao W, Cheng T, Li C, et al. tagHi-C Reveals 3D chromatin architecture dynamics during mouse hematopoiesis. Cell Rep 2020;32:108206.

[14] Heinz S, Texari L, Hayes MGB, Urbanowski M, Chang MW, Givarkes N, Rialdi A, White KM, Albrecht RA, Pache L, et al. Transcription Elongation Can Affect Genome 3D Structure. Cell 2018;174:1522 (-+).

[15] Belaghzal H, Borrman T, Stephens AD, Lafontaine DL, Venev SV, Weng Z, Marko JF, Dekker J. Liquid chromatin Hi-C characterizes compartment-dependent chromatin interaction dynamics. Nat Genet 2021;53:367–78.

[16] García-Domenech R, Gálvez J, de Julián-Ortiz JV, Pogliani L. Some New Trends in Chemical Graph Theory. Chem Rev 2008;108:1127–69.

[17] Atilgan, C., Okan, O.B. and Atilgan, A.R. (2012) In Rees, D.C. (ed.), *Annual Review of Biophysics, Vol 41* , Vol. 41, pp. 205–225.

[18] Amitai G, Shemesh A, Sitbon E, Shklar M, Netanely D, Venger I, Pietrokovski S. Network analysis of protein structures identifies functional residues. J Mol Biol 2004;344:1135–46.

[19] Di Paola L, Giuliani A. Protein contact network topology: a natural language for allostery. Curr Opin Struct Biol 2015;31:43–8.

[20] Doshi U, Holliday MJ, Eisenmesser EZ, Hamelberg D. Dynamical network of residue–residue contacts reveals coupled allosteric effects in recognition, catalysis, and mutation. Proc Natl Acad Sci 2016;113:4735–40.

[21] Daily MD, Upadhyaya TJ, Gray JJ. Contact rearrangements form coupled networks from local motions in allosteric proteins. Proteins-Struct Funct Bioinforma 2008;71:455–66.

[22] Albert R, Barabási A-L. Statistical mechanics of complex networks. Rev Mod Phys 2002;74:47–97.

[23] Yao XQ, Momin M, Hamelberg D. Elucidating allosteric communications in proteins with difference contact network analysis. J Chem Inf Model 2018;58:1325–30.

[24] Sethi A, Tian J, Vu, Dung M, Gnanakaran S. Identification of minimally interacting modules in an intrinsically disordered protein. Biophys J 2012;103:748–57.

[25] Kryven I, Duivenvoorden J, Hermans J, Iedema PD. Random graph approach to multifunctional molecular networks. Macromol Theory Simul 2016;25:449–65.

[26] Stauffer D. Introduction to Percolation Theory. Taylor & Francis,; 1985.

[27] Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 2005;4. Article17.

[28] Hao D, Ren C, Li C. Revisiting the variation of clustering coefficient of biological networks suggests new modular structure. BMC Syst Biol 2012;6:34.

[29] Panditrao G, Bhowmick R, Meena C, Sarkar RR. Emerging landscape of molecular interaction networks: Opportunities, challenges and prospects. J Biosci 2022;47:24.

[30] Pancaldi V. Network models of chromatin structure. Curr Opin Genet Dev 2023;80:102051.

[31] Botta M, Haider S, Leung IXY, Lio P, Mozziconacci J. Intra- and inter-chromosomal interactions correlate with CTCF binding genome wide. Mol Syst Biol 2010;6:426.

[32] Sandhu KS, Li G, Poh HM, Quek YL, Sia YY, Peh SQ, Mulawadi FH, Lim J, Sikic M, Menghi F, et al. Large-scale functional organization of long-range chromatin interaction networks. Cell Rep 2012;2:1207–19.

[33] Pancaldi V, Carrillo-de-Santa-Pau E, Javierre BM, Juan D, Fraser P, Spivakov M, Valencia A, Rico D. Integrating epigenomic data and 3D genomic structure with a new measure of chromatin assortativity. Genome Biol 2016;17:152.

[34] Newman M. Networks. Oxford: OUP,; 2018.

[35] Solovei I, Kreysing M, Lanctot C, Kosem S, Peichl L, Cremer T, Guck J, Joffe B. Nuclear architecture of rod photoreceptor cells adapts to vision in mammalian evolution. Cell 2009;137:356–68.

[36] Falk M, Feodorova Y, Naumova N, Imakaev M, Lajoie BR, Leonhardt H, Joffe B, Dekker J, Fudenberg G, Solovei I, et al. Heterochromatin drives compartmentalization of inverted and conventional nuclei. Nature 2019;570:395–9.

[37] Javierre BM, Burren OS, Wilder SP, Kreuzhuber R, Hill SM, Sewitz S, Cairns J, Wingett SW, Varnai C, Thiecke MJ, et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell 2016;167:1369–84. e1319.

[38] Hoencamp C, Dudchenko O, Elbatsh AMO, Brahmachari S, Raaijmakers JA, van Schaik T, Sedeño Cacciatore Á, Contessoto VG, van Heesbeen R, van den Broek B, et al. 3D genomics across the tree of life reveals condensin II as a determinant of architecture type. Science 2021;372:984–9.

[39] Gibcus JH, Samejima K, Goloborodko A, Samejima I, Naumova N, Nuebler J, Kanemaki MT, Xie L, Paulson JR, Earnshaw WC, et al. A pathway for mitotic chromosome formation. Science 2018;359:eaao6135.

[40] Kaul A, Bhattacharyya S, Ay F. Identifying statistically significant chromatin contacts from Hi-C data with FitHiC2. Nat Protoc 2020;15:991–1012.

[41] Ay F, Bailey TL, Noble WS. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Res 2014;24:999–1011.

[42] Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods 2012;9:999–1003.

[43] Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 2014;159:1665–80.

[44] Caldarelli G, Pastor-Satorras R, Vespignani A. Structure of cycles and local ordering in complex networks. Eur Phys J B 2004;38:183–6.

[45] Fronczak A, Hołyst JA, Jedynak M, Sienkiewicz J. Higher order clustering coefficients in Barabási–Albert networks. Phys A: Stat Mech its Appl 2002;316:688–94.

[46] Lind PG, González MC, Herrmann HJ. Cycles and clustering in bipartite networks. Phys Rev E 2005;72:056127.

[47] Yin H, Benson AR, Leskovec J. Higher-order clustering in networks. Phys Rev E 2018;97:052306.

[48] Das P, Golloshi R, McCord RP, Shen T. Using contact statistics to characterize structure transformation of biopolymer ensembles. Phys Rev E 2020;101:012419.

[49] Haarhuis JHI, van der Weide RH, Blomen VA, Yanez-Cuna JO, Amendola M, van Ruiten MS, Krijger PHL, Teunissen H, Medema RH, van Steensel B, et al. The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. Cell 2017;169:693–707. e614.

[50] Kind J, Pagie L, de Vries SS, Nahidiazar L, Dey SS, Bienko M, Zhan Y, Lajoie B, de Graaf CA, Amendola M, et al. Genome-wide Maps of Nuclear Lamina Interactions in Single Human. Cells *Cell* 2015;163:134–47.

[51] Das P, San Martin R, McCord RP. Differential contributions of nuclear lamina association and genome compartmentalization to gene regulation. Nucleus 2023;14:2197693.

[52] Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 2011;473:43–9.

[53] Xiong K, Ma J. Revealing Hi-C subcompartments by imputing inter-chromosomal chromatin interactions. Nat Commun 2019;10:5069.

[54] Wen Z, Zhang W, Zhong Q, Xu J, Hou C, Qin ZS, Li L. Extensive chromatin structure-function associations revealed by accurate 3D compartmentalization characterization. Front Cell Dev Biol 2022;10:845118.

[55] Nichols MH, Corces VG. Principles of 3D compartmentalization of the human genome. Cell Rep 2021;35:109330.

[56] Yaffe E, Tanay A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat Genet 2011;43:1059–65.

[57] Kariti H, Feld T, Kaplan N. Hypothesis-driven probabilistic modelling enables a principled perspective of genomic compartments. Nucleic Acids Res 2023;51:1103–19.

[58] Dodero-Rojas E, Mello MF, Brahmachari S, Oliveira Junior AB, Contessoto VG, Onuchic JN. PyMEGABASE: predicting cell-type-specific structural annotations of chromosomes using the epigenome. J Mol Biol 2023;435:168180.

[59] Fatima U, Hina S, Wasif M. A novel global clustering coefficient-dependent degree centrality (GCCDC) metric for large network analysis using real-world datasets. J Comput Sci 2023;70:102008.

[60] Malod-Dognin N, Pancaldi V, Valencia A, Pržulj N. Chromatin network markers of leukemia. Bioinformatics 2020;36:i455–63.

[61] Su JH, Zheng P, Kinrot SS, Bintu B, Zhuang X. Genome-Scale Imaging of the 3D organization and transcriptional activity of chromatin. Cell 2020;182:1641–59. e1626.

[62] Thibodeau A, Márquez EJ, Shin D-G, Vera-Licona P, Ucar D. Chromatin interaction networks revealed unique connectivity patterns of broad H3K4me3 domains and super enhancers in 3D chromatin. Sci Rep 2017;7:14466.

[63] Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. Nat Protoc 2017;12:2478–92.

[64] Abdulla AZ, Salari H, Tortora MMC, Vaillant C, Jost D. 4D epigenomics: deciphering the coupling between genome folding and epigenomic regulation with biophysical modeling. Curr Opin Genet Dev 2023;79:102033.

[65] Olivares-Chauvet P, Mukamel Z, Lifshitz A, Schwartzman O, Elkayam NO, Lubling Y, Deikus G, Sebra RP, Tanay A. Capturing pairwise and multi-way chromosomal conformations using chromosomal walks. Nature 2016;540:296–300.

[66] Dotson GA, Chen C, Lindsly S, Cicalo A, Dilworth S, Ryan C, Jeyarajan S, Meixner W, Stansbury C, Pickard J, et al. Deciphering multi-way interactions in the human genome. Nat Commun 2022;13:5498.