Journal of Physics: Complexity

PAPER • OPEN ACCESS

Exploring 3D community inconsistency in human chromosome contact networks

To cite this article: Dolores Bernenko *et al* 2023 *J. Phys. Complex.* **4** 035004

Journal of Physics: Complexity

OPEN ACCESS

RECEIVED
13 March 2023

REVISED
23 July 2023

ACCEPTED FOR PUBLICATION
11 August 2023

PUBLISHED
23 August 2023

PAPER

Exploring 3D community inconsistency in human chromosome contact networks

Dolores Bernenko[1], **Sang Hoon Lee**[2,3] **and Ludvig Lizana**[4,*]

[1] Department of Physics, Integrated Science Lab, Umeå University, SE-901 87 Umeå, Sweden
[2] Department of Physics and Research Institute of Natural Science, Gyeongsang National University, Jinju 52828, Republic of Korea
[3] Future Convergence Technology Research Institute, Gyeongsang National University, Jinju 52849, Republic of Korea
[4] Integrated Science Lab, Department of Physics, Umeå University, SE-901 87 Umeå, Sweden
[*] Author to whom any correspondence should be addressed.

E-mail: ludvig.lizana@umu.se

Keywords: complex networks, chromosome 3D organization, community detection

Supplementary material for this article is available online

Abstract

Researchers have developed chromosome capture methods such as Hi-C to better understand DNA's 3D folding in nuclei. The Hi-C method captures contact frequencies between DNA segment pairs across the genome. When analyzing Hi-C data sets, it is common to group these pairs using standard bioinformatics methods (e.g. PCA). Other approaches handle Hi-C data as weighted networks, where connected node pairs represent DNA segments in 3D proximity. In this representation, one can leverage community detection techniques developed in complex network theory to group nodes into mesoscale communities containing nodes with similar connection patterns. While there are several successful attempts to analyze Hi-C data in this way, it is common to report and study the most typical community structure. But in reality, there are often several valid candidates. Therefore, depending on algorithm design, different community detection methods focusing on slightly different connectivity features may have differing views on the ideal node groupings. In fact, even the same community detection method may yield different results if using a stochastic algorithm. This ambiguity is fundamental to community detection and shared by most complex networks whenever interactions span all scales in the network. This is known as community inconsistency. This paper explores this inconsistency of 3D communities in Hi-C data for all human chromosomes. We base our analysis on two inconsistency metrics, one local and one global, and quantify the network scales where the community separation is most variable. For example, we find that TADs are less reliable than A/B compartments and that nodes with highly variable node-community memberships are associated with open chromatin. Overall, our study provides a helpful framework for data-driven researchers and increases awareness of some inherent challenges when clustering Hi-C data into 3D communities.

1. Introduction

Chromosomes' three-dimensional (3D) folded structure is critical to understanding genetic processes and genome evolution. The discovery of meaningful 3D structures relied on genome-wide chromosome capture data analysis, represented by a pairwise interaction matrix called Hi-C map [1–3]. These maps reveal substructures of different scales, including the dichotomous division into A versus B compartments and smaller-scale topologically associated domains (TADs) scattered across the genome. Molecular biologists find these structures appealing because they define DNA regions with correlated gene expression and epigenetic modifications. In addition, their borders enrich binding sites for architectural proteins such as CCCTC-binding factor (CTCF) (also known as 11-zinc finger protein or the CCCTC-binding factor) in humans and CP190 in *Drosophila melanogaster*[4].

As researchers delved deeper into this topic, they discovered that TADs act as shielded 3D domains with more internal than external contacts, similar to the definition of 'communities' in network science [5–7].

They also noted that some TADs are nested or partially overlapped. Therefore, it is difficult to find a solid mathematical definition for TADs. This problem is best illustrated by a recent publication benchmarking 27 TAD-finding algorithms reporting significant inconsistencies depending on the data set, Hi-C resolution, etc [8]. But despite technical challenges, 'TADs' still represent meaningful genomic regions with above-average internal contact frequencies.

Like TADs, A/B compartments also show cross-scale organization as they typically split into six subclasses (A1, A2, B1, etc) [3, 9]. But unlike TADs, A/B compartments are more consistent across methods. These methods often rest on principal component analysis (PCA) that detects sign shifts of the first principal component. DNA regions with positive signs are assigned to one compartment, A; the rest is denoted B. The A/B division refers to open and closed chromatin, where A often is gene-rich and actively transcribed, whereas B is mostly gene-poor and less active.

Studies of TADs and A/B compartments argue that chromosomes possess a complex multi-scale structure. This complexity is underscored by a suite of publications (by authors of this paper) that use a community-detection method developed for complex networks, modified to obey that the average contact frequency decays as a power-law with the genomic distance separating two DNA segments. This method revealed a spectrum of mesoscale 3D communities in Hi-C data, where A/B compartments and TADs are but two instances [10], and that these communities are semi-nested when overlaying them across the network scales [11]. Also, by sampling many feasible network partitions, we noted that some Hi-C network scales have more partitions than others [12]. If too many of these 'degenerate' partitions exist, it is difficult to find reliable network partitions, which may cause numerical methods to differ. In this work, we complement this idea and take a more node-centric view and study node-community variability.

But these challenges do not differ Hi-C maps from other complex networks. More often than not, complex networks have a blend of overlapping communities at different scales that makes community detection challenging. This complexity complicates finding statistically significant communities, as most community detection methods rely on a mathematical objective function called modularity [6, 7]. This function has a scale or resolution parameter [13], and most community detection algorithms select its value rather arbitrarily without a principled guideline (see [14] for the relationship between the resolution parameter and a parameter from another established community detection framework called the stochastic block model). Recently, some researchers (including one of the authors of this paper) have taken advantage of this fact and tried extracting informative structural properties based on the ensemble of cross-scale 'inconsistently' detected communities [15–17].

This paper builds on some of these methods and findings. In particular, we utilize two representative metrics [16] that quantify local and global inconsistencies. These metrics help highlight the scales offering the most reliable 3D communities in the Hi-C data. We explain these metrics in section 2. Alongside the inconsistency framework developed in [16]. This section also presents the Hi-C data and associated weighted networks. Also, to make the presentation more complete, we briefly explain the main steps in the Hi-C experiment. We also describe how we use protein-binding and epigenetic data to assign chromatin states to DNA regions by calculating folds of enrichment ratios. Finally, we present our findings in section 3. and conclude our paper in section 4.

2. Methods

2.1. Hi-C method and data

The Hi-C method is an experimental technique that explores the 3D structure of genomes [1]. The procedure binds together spatially interacting DNA sequence pairs, which may be several megabases apart on the linear genome. If using a cell population, experimenters obtain a collection of 'binding events' representing the ensemble-averaged contact frequencies.

In practice, Hi-C starts with 'freezing' the chromosomal 3D configuration using formaldehyde. Formaldehyde cross-links spatially close DNA segments and proteins. After removing the cell membrane, a series of steps cut the DNA into numerous fragments, merge the free ends, and tag the cut spot with biotin. These steps are known as restriction digestion, proximity ligation, and biotin filling. The final stages purify DNA from proteins and RNA, remove all cross-links, and pull down biotin-marked fragments that next undergo massive sequencing. After sequencing, the fragments are associated with linear genomic locations and then processed into a contact map ('Hi-C map'), a matrix that illustrates the interaction frequency between all loci.

2.1.1. Transforming Hi-C data into a weighted network

We use the same Hi-C intra-chromosomal contact map as our previous series of studies [10, 11] (human cell line GM12878 (B-lymphoblastoid) [3, 18]). Also, as before [10, 11], we use the MAPQG0 data set at the 100 kilobase-pair (kb) resolution and normalize the interaction map with the Knight–Ruiz (KR) matrix balancing [19]. As a result, we treat each 100 kb chromatin locus as the minimal unit, or 'node', and the normalized interaction weights between nodes i and j as weighted edges, using network science terminology [5].

2.2. Network community detection and inconsistency

One of the most popular ways to detect network communities [6, 7]—densely-connected substructures—is to maximize the objective function called modularity[5]

$$\mathcal{M} = \frac{1}{2m}\sum_{i\neq j}\left[\left(A_{ij} - \gamma P_{ij}\right)\delta(g_i, g_j)\right] . \tag{1}$$

Here, A_{ij} denotes the adjacency matrix elements corresponding to the interaction weights between nodes i and j ($A_{ij} = 0$ indicates no edge), and P_{ij} is the expected edge weight based on *a priori* information. The most popular choice is when considering only the overall tendency of node-node interaction $P_{ij} = k_i k_j/(2m)$, where k_i represents node i's strength (the sum of its weights) and m is a normalization constant ensuring that $-1 \leqslant \mathcal{M} \leqslant 1$. Finally, g_i is the community index of node i and δ is the Kronecker delta. A key parameter in our study is the resolution parameter γ, which controls the overall community scale [13].

In principle, maximizing the modularity function with respect to all of the possible community divisions, encoded as $\{g_i\}$ in equation (1), is a mathematically well-defined deterministic concept. However, due to the computational limitation imposed by the problem, it is prohibitively difficult to find the exact solution, e.g. from the comprehensive enumeration of the network divisions. Therefore, most network community detection algorithms rely on various types of approximations or parameter restrictions. Many algorithms take a stochastic approach to sample the community partitions, just as in standard Monte Carlo [20]. One example is the Louvain-type algorithms [21, 22] we use here (detailed in section 2.3).

Although stochastic approaches like Louvain have been successful in terms of speed and accuracy in many community detection applications, their stochastic nature may produce *multiple* results that sometimes include *inconsistent* elements. Researchers tend to work around this inconsistency [23, 24] by choosing the most consistent, or reproducible, network partition. However, one of the authors of this paper has turned this inconsistency into an advantage, using it to probe network structural information [15, 16] at both global and local levels (figures 1 and 2). In particular, by studying inconsistency measures one may pinpoint scale regimes or specific node collections that are the most statistically reliable (at the global level) or flexible (at the local level). For a detailed theoretical framework, we defer to [16]. But below, we remind the reader of the essential parts used in this analysis.

One metric we study is partition inconsistency (PaI). It quantifies the global degree of inconsistency among community partitions in the entire network (figure 1). PaI is based on a recently developed similarity measure called element-entric similarity [25], which overcomes notable biases, for example, skewed cluster sizes in normalized mutual information (and other conventional similarity measures). The similarity measure $S_{\alpha\beta}$ between community configurations α and β is defined as

$$S_{\alpha\beta} = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \frac{1}{2d}\sum_{j=1}^{N}|f_{ij}^{\alpha} - f_{ij}^{\beta}|\right) , \tag{2}$$

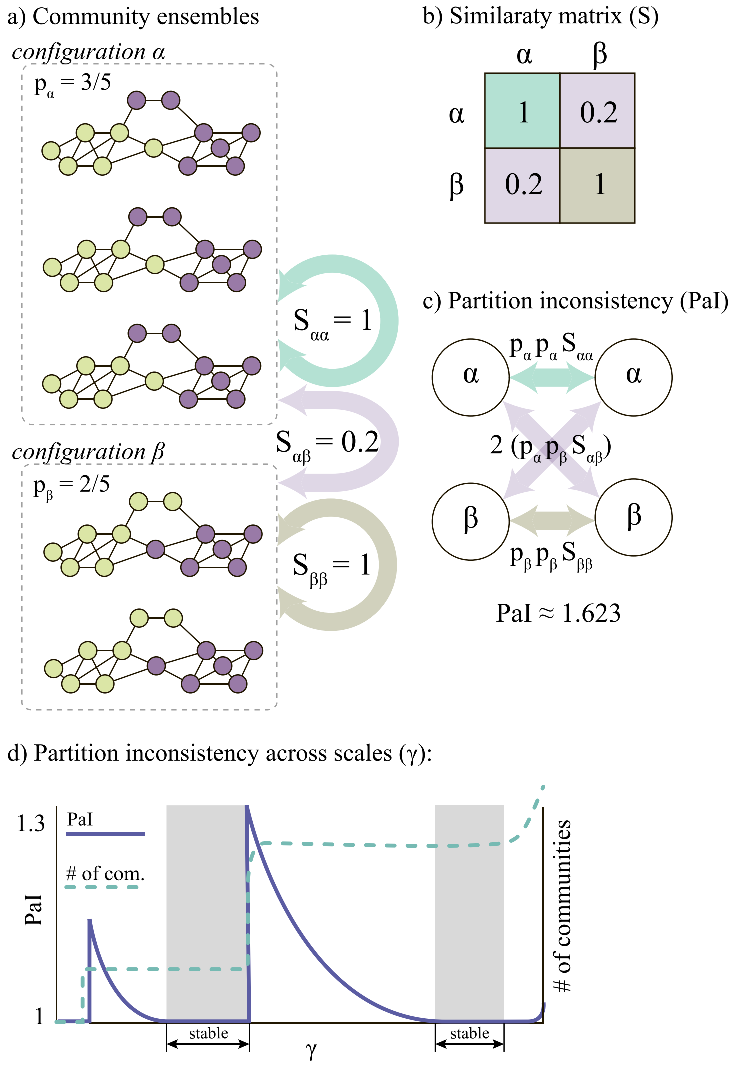
where N is the number of nodes and f_{ij}^{α} is node j's relative importance in the stationary state of the personalized PageRank algorithm (PPR) [26] starting from the node i in configuration α. Also, d is a damping factor we set to 0.9 (default value in [25]).

Based on equation (2), we define PaI value as

$$\mathrm{PaI} = \left(\sum_{\alpha=1}^{\mathcal{C}}\sum_{\beta=1}^{\mathcal{C}} p_\alpha p_\beta S_{\alpha\beta}\right)^{-1} , \tag{3}$$

which corresponds to the average similarity between all configuration pairs. The terms form a weighted sum of $S_{\alpha\beta}$ by the proportion p_α and p_β of each configuration pair α and β, respectively, that appear in an ensemble of community detection results composed of $\mathcal{C}$ configurations. The PaI value ($\geqslant$1) indicates the

[5] We emphasize that we are using the terminology for weighted networks since we use the weighted version of the Hi-C interaction map. For binary networks, they are reduced to the conventional version: A_{ij} is either 0 or 1, representing the absence or presence of the edge, and k_i is the number of neighboring nodes to i (the degree).



	α	β
α	1	0.2
β	0.2	1

Figure 1. Illustration of partition inconsistency (PaI). (a) A community ensemble composed of two configurations α and β. The similarity measure [25] quantifies the degree of similarity, S, between the community configurations. The elements of the similarity matrix **S** are calculated using equation (2). (b) The similarity matrix **S** for configuration pairs visualized in panel (a). (c) Based on the similarity measure, we calculate the PaI value using equation (3) (as in [16]). (d) Cross-scale PaI curve (varying the resolution parameter γ) alongside the average number of communities. Shaded areas show ranges of statistically reliable community scales.

effective number of independent configurations. A small (large) PaI value represents more consistent (inconsistent) regimes, respectively. Using PaI, one extracts the most statistically reliable ranges of community scales by focusing on the (local) minima of PaI, in particular, alongside another meaningful evidence of stable communities: the number of communities stays flat at a specific integer value (illustrated in figure 1(b)).

While PaI describes the network's global inconsistency, we use another metric, membership inconsistency (MeI), to quantify local (individual-node) inconsistencies (figure 2). The MeI measure for each node *i* is based on the Jaccard index defined as

$$J_{i\alpha\beta} = \frac{|\psi_{i\alpha} \cap \psi_{i\beta}|}{|\psi_{i\alpha} \cup \psi_{i\beta}|}, \tag{4}$$

where $\psi_{i\alpha}$ indicates the set of nodes belonging to the same community as node *i* in configuration α. Then, the MeI measure is defined as

$$\Psi_i = \left(\sum_{\alpha=1}^{C} \sum_{\beta=1}^{C} p(\psi_{i\alpha}) p(\psi_{i\beta}) J_{i\alpha\beta} \right)^{-1}, \tag{5}$$

where $p(\psi_{i\alpha})$ is the relative frequency of appearance $\psi_{i\alpha}$ in the ensemble. MeI represents the effective number of independent communities for a specific node across different community configurations. As

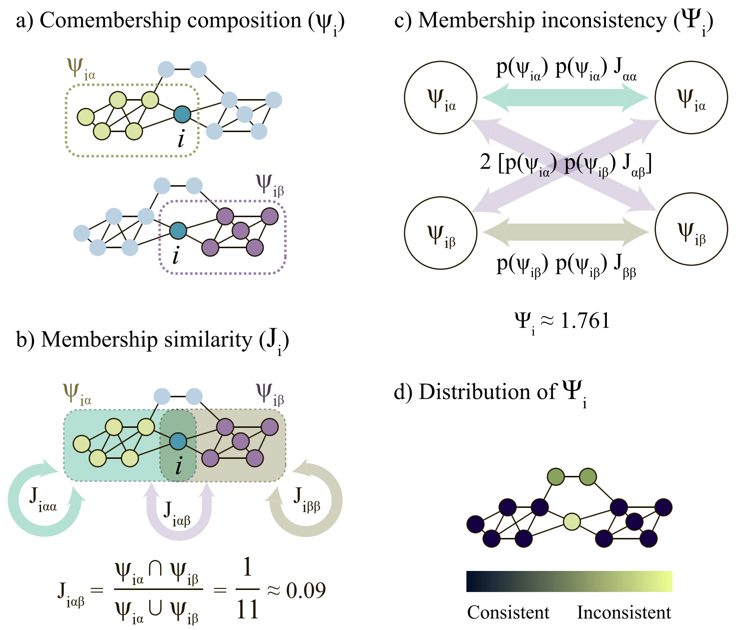

a) Comembership composition (ψ_i)
$\psi_{i\alpha}$
i
$\psi_{i\beta}$
i
c) Membership inconsistency (Ψ_i)
$p(\psi_{i\alpha})\, p(\psi_{i\alpha})\, J_{\alpha\alpha}$
$\Psi_{i\alpha}$
$\Psi_{i\alpha}$
$2\,[p(\psi_{i\alpha})\, p(\psi_{i\beta})\, J_{\alpha\beta}]$
$\Psi_{i\beta}$
$\Psi_{i\beta}$
$p(\psi_{i\beta})\, p(\psi_{i\beta})\, J_{\beta\beta}$
$\Psi_i \approx 1.761$
b) Membership similarity (J_i)
$\psi_{i\alpha}$
$\psi_{i\beta}$
i
$J_{i\alpha\alpha}$
$J_{i\alpha\beta}$
$J_{i\beta\beta}$
$J_{i\alpha\beta} = \frac{\psi_{i\alpha} \cap \psi_{i\beta}}{\psi_{i\alpha} \cup \psi_{i\beta}} = \frac{1}{11} \approx 0.09$
d) Distribution of Ψ_i
Consistent
Inconsistent

Figure 2. Illustration of the membership inconsistency (MeI). (a) The co-membership composition for node i in configurations α and β. In panel (b) we calculate the membership similarity matrix $\mathbf{J}_i$ for each node i using equation (4). (c) From this co-membership structure, we then calculate the MeI value Ψ_i using equation (5) (as in [16]). (d) Example distribution of MeI for a small network.

shown in figure 2(b), the MeI values properly detect the functionally flexible or 'bridge' nodes participating in different modules[6]. In section 3, we use PaI and MeI to study global and local community inconsistencies of Hi-C maps and relate to these metrics to other biological data.

2.3. GenLouvain method

The stochastic community detection method we utilize throughout our work is version 2.1 of GenLouvain [22] (see https://github.com/GenLouvain/GenLouvain for the latest version). GenLouvain is a variant of the celebrated Louvain algorithm [21], which is one of the most widely used algorithms and is popular due to its speed and established packages in various programming languages.

Starting from single-node communities, the algorithm accepts or rejects trial merging processes based on the modularity (equation (1)) change in a greedy fashion. More specifically, each iteration consists of two steps: (i) looping over the nodes, and (ii) treating each community as a 'supernode' [27]. In the first process, each node is put into the neighbors community which results in the largest modularity increase $\Delta\mathcal{M}$ compared with the current partition. The second process transforms the network into a weighted network, where each community from the first step is represented as a supernode. The edges between supernodes are assigned weights indicating the number of edges joining nodes in the corresponding groups (note that self-edges for each supernode are allowed here to represent the internal edges). These steps are repeated until increasing $\mathcal{M}$ is no longer possible by moving a (super)node to a different community.

To determine the community stability across network scales, we run GenLouvain several times, at least 100, for each resolution parameter γ and then calculate the global (PaI) and local (MeI) inconsistency metrics [16].

2.4. Cross-scale node-membership correlations

We use GenLouvain to produce an ensemble of community partitions from Hi-C data for fixed scale parameters γ. However, some of these partitions seem correlated. To better understand these correlations, we use a graphical embedding technique to illustrate high-dimensional data. Specifically, we use t-SNE (t-distributed stochastic neighbor embedding) [28].

t-SNE is an algorithm that maps high-dimensional data into 2D, preserving both local and global structures. This contrasts with PCA, which preserves global structures. While PCA projects the data onto the dimensions with the highest variance, t-SNE focuses on maintaining relative distances between data points.

[6] The MeI measure introduced in [16] is a more principled and improved measure than the original 'companionship inconsistency (CoI)' measure first introduced in [15], by considering the possibility of more than two community memberships.

Notably, unlike PCA, the axes in a t-SNE plot lack specific meanings and represent abstract dimensions used to retain the data structure in the reduced space. This property makes t-SNE ideal for visualizing complex data sets.

The t-SNE algorithm aggregates data points based on some distance metric. While there are several choices, we use the so-called correlation distance D, which is common for random vectors and defined as

$$D = 1 - r(\mathbf{u}, \mathbf{v}), \tag{6}$$

where $r(\mathbf{u}, \mathbf{v})$ is the correlation between the vectors $\mathbf{u}$ and $\mathbf{v}$, conventionally defined as

$$r(\mathbf{u}, \mathbf{v}) = \frac{(\mathbf{u} - \bar{\mathbf{u}}) \cdot (\mathbf{v} - \bar{\mathbf{v}})}{||(\mathbf{u} - \bar{\mathbf{u}})||_2 ||(\mathbf{v} - \bar{\mathbf{v}})||_2}, \tag{7}$$

where $\bar{u}$ and $\bar{v}$ are the mean of the elements (so that $\bar{\mathbf{u}} = \bar{u} \times [1, 1, 1, \ldots]$) and $|| \cdots ||_2$ is the Euclidean norm. According to equation (6), $D = 0$ if they are perfectly correlated ($r = 1$) and $D \approx 1$ if they are uncorrelated ($r \approx 0$).

In our analysis, we create these vectors from 100 GenLouvain runs. Each vector is a binary representation of the node-community membership for one node (each element is 1 or 0) depending on whether the node belongs to a specific community at a particular GenLouvain iteration. For our analysis, we used scikit-learn [29]. To reach the best visualization results, we tuned several parameters: perplexity: 20, early exaggeration: 8, initialization: random, and the number of iterations: 1356.

2.5. Chromatin states and enrichment

In results (section 3), we analyze the inconsistency measures PaI and MeI in terms of chromatin states derived from an established chromatin division [30] that we downloaded from ENCODE [31] (GM12878, Accession: wgEncodeEH000784). This data set constitutes a list of start and stop positions associated with chromatin states called peaks. These peaks result from integrating several biological data sets, e.g. ChIP-seq and RNA-seq, with a multivariate hidden Markov model (HMM). The authors [30] use 15 'HMM states' (S1–S15): active promoter (S1), weak promoter (S2), inactive/poised promoter (S3), strong enhancer (S4 and S5), weak/poised enhancer (S6 and S7), insulator (S8), transcriptional transition (S9), transcriptional elongation (S10), weakly transcribed (S11), Polycomb-repressed (S12), heterochromatin (S13), and repetitive/copy number variation (S14 and S15).

While identifying these states is automated (using chromHMM [32]), their names result from human interpretations, reflecting typical biological functions. Also, selecting 15 states is somewhat arbitrary, with the typical range spanning 5–20. Another limitation to consider is the resolution. It is approximately one nucleosome (∼200 base pairs) because the data stems from epigenetic histone modifications.

The start and stop regions for these 15 HMM do not match perfectly with the Hi-C bins. To classify every Hi-C bin into one of these HMM states, we calculate the folds of enrichment (FE) relative to a chromosome-wide average according to the following steps.

1. Count the number of peaks k_X per bin, where $X = \mathrm{S}1, \ldots, \mathrm{S}15$. Because some peaks span multiple bins, we only count the peak starts.
2. Calculate the peak frequency's expected value using the hypergeometric test (chromosome-wide sampling without replacement). The expected number of X peaks per bin is calculated as $\bar{k}'_X = K_X \times (n/N)$, where, n is the number of peaks of any state in a bin, N is the total number of peaks per chromosome, and K_X is the total number of peaks for state X.
3. Calculate the folds of enrichment FE_X for each HMM state X per bin by dividing the observed by the expected peak number, $\mathrm{FE}_X = k_X/\bar{k}'_X$.

We note each Hi-C bin can be enriched in several chromatin states. Based on enrichment, we divide Hi-C bins into five groups (A–D) if $\mathrm{FE}_X > 1$:

(A) Promoters: $X =$ S1 and S2.
(B) Enhancers: $X =$ S4, S5, S6, and S7.
(C) Transcribed regions: $X =$ S9, S10, and S11.
(D) Heterochromatin and other repressive states: $X =$ S3, S13, S14, and S15.
(E) Insulators: $X =$ S8

Apart from these five groups, we assign bins that are not enriched in any state to the category 'NA'.

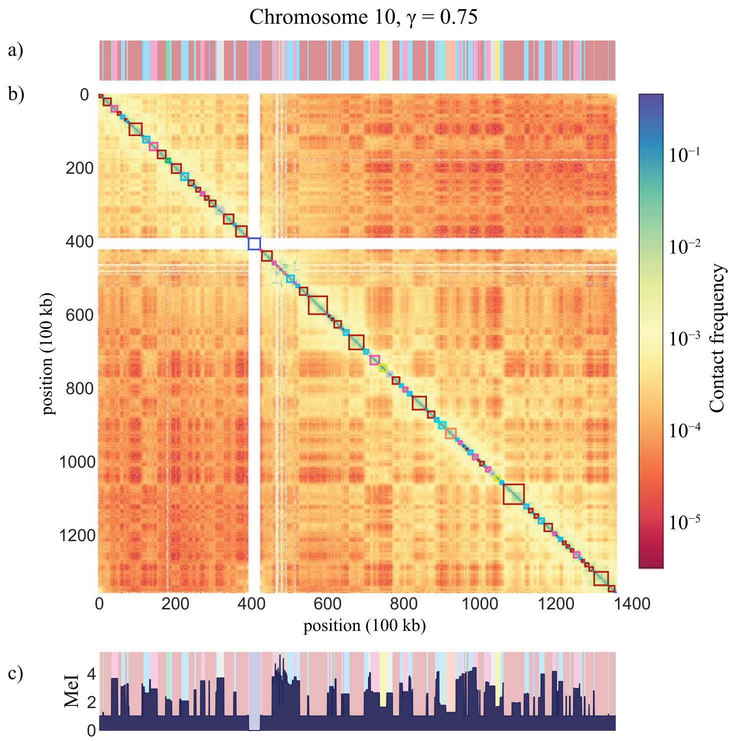


Figure 3. Membership inconsistency analysis for $\gamma = 0.75$. The horizontal axis represents the genomic position along human chromosome 10. (a) Community membership of DNA segments shown as rectangles along the DNA sequence. Rectangles of the same color enclose DNA segments belonging to the same community (a single configuration). (b) Hi-C data is depicted as a heat map of contact frequencies between DNA-segment pairs. Rectangles along the main diagonal represent DNA segments' community membership from panel (a). (c) Nodes' membership inconsistency measured across 100 GenLouvain realizations. The dark-blue plot shows MeI scores of DNA segments in a chromosome (the network nodes). The background of this panel shows the color-coded community membership from panel (a).

2.6. Interpreting boxen plots

In section 3, we visualize the MeI data using so-called boxen or letter-value plots. These contrasts regular box plots with outliers, where boxen plots better represent the actual distribution of the data.

In more detail, boxen plots are based on so-called letter values. These are recursively defined order statistics with specific depths d_i that splits the data set. These depths are calculated as:

$$d_1 = (1+n)/2, \;\; d_i = (1 + \lfloor d_{i-1} \rfloor)/2, \tag{8}$$

where n is the number of data points and $\lfloor \ldots \rfloor$ denotes the floor function. Here, d_1 corresponds to the median that splits the dataset into two halves. Next, d_2 splits these two halves to get the fourths. Splitting yet more times gives the eights, sixteenths, etc.

This recursive division is similar to a regular boxplot. First, we have the median in the middle, which splits the data points into two halves. Then, the remaining data spaces are halved again to get the quantiles. The boxes around the median then indicate the end of these quantiles. But boxen plots do not truncate after the quantiles. The remaining data spaces are divided yet again, adding more and more boxes to the plot that represents an increasingly smaller fraction of the dataset. The splitting stops after reaching some stopping criterion, typically the 95% confidence interval.

3. Results

3.1. Local community inconsistency

We illustrated the local inconsistency associated with a single Hi-C map in figure 3. This map depicts the number of contacts between all 100 kb DNA-segment pairs (KR normalized) in human chromosome 10. Along the diagonal, we highlight the GenLouvain-derived communities [22], at resolution parameter $\gamma = 0.75$. Squares sharing colors have the same community membership. These colors are better illustrated in the stripe above the map, showing how communities appear along the linear DNA sequence. We note that some scattered segments have the same color, which indicates that communities assemble DNA segments in 3D proximity, not only 1D adjacent neighbors. This contrasts the conventional notion of TADs, which

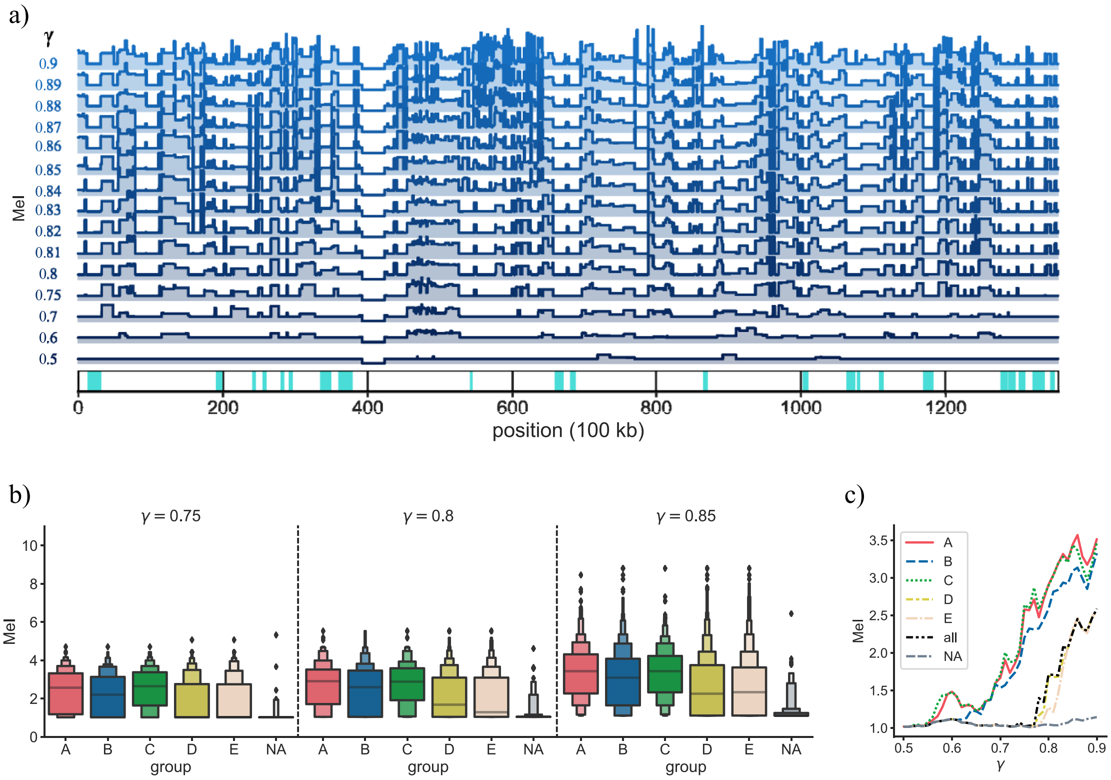


Figure 4. Membership inconsistency (MeI) across network scales ($0.5 \leqslant \gamma \leqslant 0.9$) on human chromosome 10. (a) MeI profile for each resolution parameter γ. Below the waterfall plot, we draw blue squares that enclose DNA segments with low MeI scores across all scales. (b) Distribution of MeI scores shown for three network scales ($\gamma = 0.75, 0.80, 0.85$). We separated the distributions into A,. . ., and E groups (and 'NA'), highlighting the nodes' chromatin state. Each distribution is shown as a boxen plot with a center line marking the median, and the boxes indicate successive quantiles [33]. (c) Median MeI across γ and chromatin groups.

comprise contiguous DNA stretches. To separate notations, we denote unbroken units of DNA stretches in a community as *domains*.

It is essential to realize that the domains and communities in figures 3(a) and (b) represent a single configuration, or partition, of Hi-C network communities at one specific resolution parameter value γ. Since GenLouvain uses a stochastic maximization algorithm, we expect to find other partitions if running it several times on the same data set, some of which may differ substantially. To quantify this variability, we generated 100 independent network partitions and calculated the local inconsistency measure MeI [16] that quantifies how many different community configurations a single node effectively belongs to. We plot the MeI profile along chromosome 10 in figure 3(c). This profile shows that about half of the domains do not change community membership (the median value of MeI = 1.02), whereas the rest show significantly more variability (MeI $\approx$ 4). We also note that the MeI score is relatively uniform within each domain and that sharp MeI transitions occur near domain boundaries.

Based on previous work [16, 24], we anticipate that the MeI profile changes with the network scale. Therefore, we scanned through a wide range of community scales, extracted 3D communities, and calculated the MeI profile. We show the result from such a sweep in figure 4(a), where each MeI profile is associated with one γ value. We note that some DNA regions have low MeI scores ($\gamma > 0.6$), which indicates that nodes in those regions mostly appear in the same communities for most γ values. We indicate this as colored rectangles below the MeI profiles. But other DNA regions show the opposite behavior. These regions contain nodes that often do not appear in the same communities, which results in high and variable MeI values. Overall, the local node inconsistency grows as γ becomes larger.

3.2. Local inconsistency and chromatin states

To appreciate the MeI variations from a biological perspective, we analyzed them relative to local chromatin states. As outlined in the Methods (section 2.5), we use five states and calculate the folds of enrichment for each node. We denote the chromatin states as promoters (A), enhancers (B), transcribed regions (C), heterochromatin and other repressive states (D), and insulators (E).

Below the MeI profile in figure 4(b), we show boxen plots for three γ values. The boxen or letter-value plot offers a more detailed visualization of the data distribution than a traditional box plot with outliers [33].

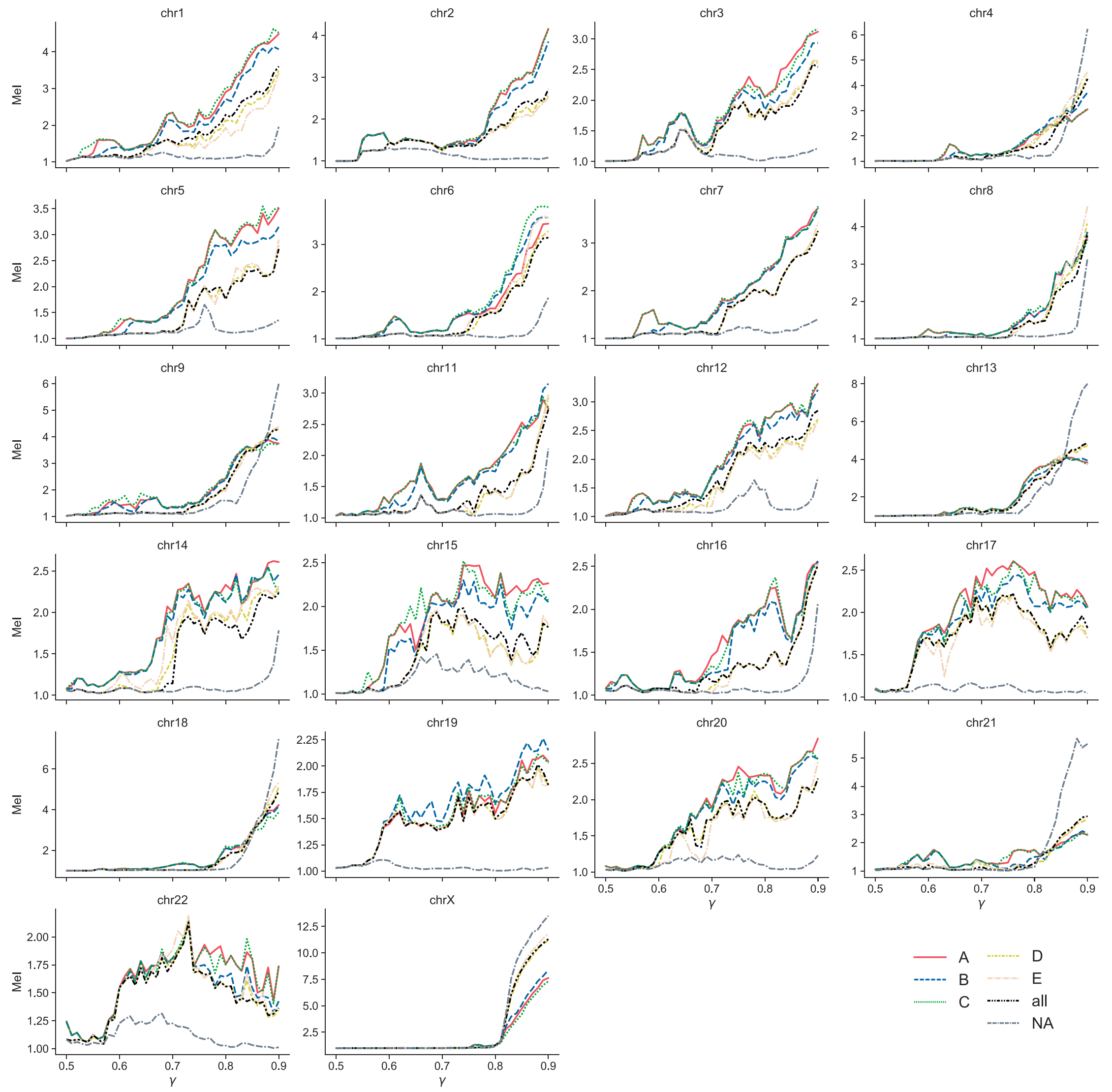


Figure 5. Median MeI for human chromosomes 1–9, 11–22, and X across γ values and chromatin groups.

The boxes are determined by calculating a series of quantiles that successively divide the data set into more segments than the typical quartiles in regular boxplots (halves, fourths, eights, etc). This division provides a more detailed picture of the data distribution, especially revealing the behaviors in the tails. We defer to methods (section 2.6) for more details.

Each subplot illustrates the distribution of MeI scores associated with each chromatin group (A–E); 'NA' represents nodes not enriched in any chromatin type. Following the MeI medians (horizontal lines), we note that groups A–C have consistently higher values than the rest. In panel (c), we explore this observation more thoroughly and plot the median MeI for several γ values. The lines show that MeI grows with γ, where the A–C groups are more inconsistent than the chromosome-wide average (denoted 'all'). They also have higher MeI scores than the D–E groups that tend to follow the average. These two chromatin groups, A–C and D–E, represent the classical division into open (active) and closed (inactive) chromatin, or so-called eu- and heterochromatin.

We find similar patterns when calculating the MeI scores for all chromosomes (figure 5). We note that open chromatin nodes (A–C) are systematically more inconsistent than nodes categorized as closed chromatin and the chromosome-wide averages. However, there are significant differences in how the MeI scores change with γ. Some chromosomes, such as 1–12, show steady but wiggly growth. Others, e.g. 13–15, 20, and 22, exhibit a steep growth for small γ values that turns into a plateau. In some cases, such as chromosome 22, this plateau starts to decline, indicating that node-community memberships become more consistent.

These MeI plots pose several intriguing biological questions that go beyond the scope of this paper. In supplementary material (figures S1–S22), we provide the complete MeI analysis like the one in figure 4.

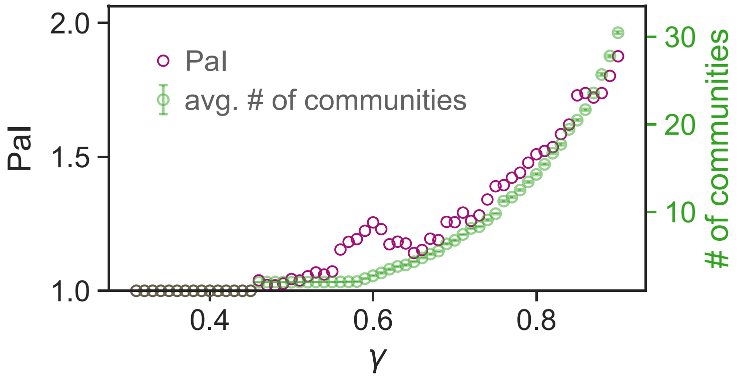


Figure 6. Global inconsistency measured by PaI (the violet circles) for human chromosome 10, along with the average number of communities (the green circles) across a range of γ values with the small (smaller than the symbols themselves for all cases) error bars representing the standard deviation.

3.3. Global inconsistency

The previous subsection analyzed the cross-scale local inconsistency measure (MeI). Here, we extend the analysis to all human chromosomes using the global inconsistency PaI instead of MeI (PaI yields one number per γ value instead of a chromosome-wide profile). The PaI score measures the effective number of independent network partitions (see section 2). By the mathematical construction of PaI [16], if there is no special scale of communities, the null-model behavior of PaI as a function of γ would be as follows. As γ gets larger from $\gamma = 0$, the PaI value starts to increase as the average number of communities increases enough to form a certain level of inconsistency (for $\gamma = 0$, PaI trivially vanishes because there cannot be any inconsistency for the single community composed of all of the nodes). On the other extreme case of $\gamma \to \infty$, each node tends to form its own singleton community, so again, there is no inconsistency, or PaI becomes zero. Therefore, if there is no particular characteristic scale of communities, the PaI curves against γ would be single-peaked ones without any nontrivial behavior such as local minima. In reality, there are characteristic scales of communities, where PaI reach its local minima [16], which indicate the most meaningful community scale. As our results show, the Hi-C communities also exhibit such characteristic scales. To better understand this metric, we revisit chromosome 10 before analyzing all human chromosomes.

We plot the PaI values for chromosome 10 in figure 6 as violet circles. When γ is small ($<$0.5), we note that PaI has a plateau extending over several γ values. Such a plateau is ideal for stable community partitions (see figure 1). However, this case is trivial because the community comprises the entire network (PaI $= 1$, thus one effective community). Next, if γ increases above 0.5, PaI starts to fluctuate, which indicates that partitions become more variable. Notably, the growing trend stops at $\gamma \approx 0.6$, and PaI decays to eventually reach a local minimum at $\gamma \approx 0.65$. This local minimum represents relatively stable communities, hinted by the small effective number of independent partitions at that scale. As we increase γ above 0.65, the community structure becomes less and less stable, along with the rapidly growing number of communities (the green circles). However, asymptotically, the number of independent community ensembles grows slightly less than the number of communities per ensemble, indicating that they have different γ behaviors. For example, at $\gamma = 0.9$, there are 1.75 independent ensembles (effective), each of which is composed of 25 communities. As a final remark, it is essential to realize that the growing trend of PaI with γ does not necessarily imply the lack of intrinsic organizational scales. Instead, it indicates fuzzy scale transitions where we observe a short range of stable communities at the local PaI minimum.

When examining the PaI curve for chromosome 10, we noticed one significant local minimum with a relatively stable community partition. Next, we ask if similar inconsistency patterns appear across all human chromosomes. To this end, we plotted PaI against γ for chromosomes 1–22 and X (except chromosome 10) in figure 7.

We found several commonalities. First, similar to chromosome 10, most PaI curves have at least one local minimum and maximum. Some chromosomes even have two minima (e.g. chromosomes 1, 3, 9, 14, etc), which indicate multiple stable scales of communities. Also, when γ becomes large enough, the network enters the multi-community regime. Second, as γ grows, so does PaI. This growth indicates that community structures have become increasingly inconsistent. Although it is natural to observe higher inconsistency for larger numbers of communities as there are more possible combinations. As future work, it would be informative to check its scaling behavior across different chromosomes and classify chromosomes based on the functional shapes of PaI and the number of communities.

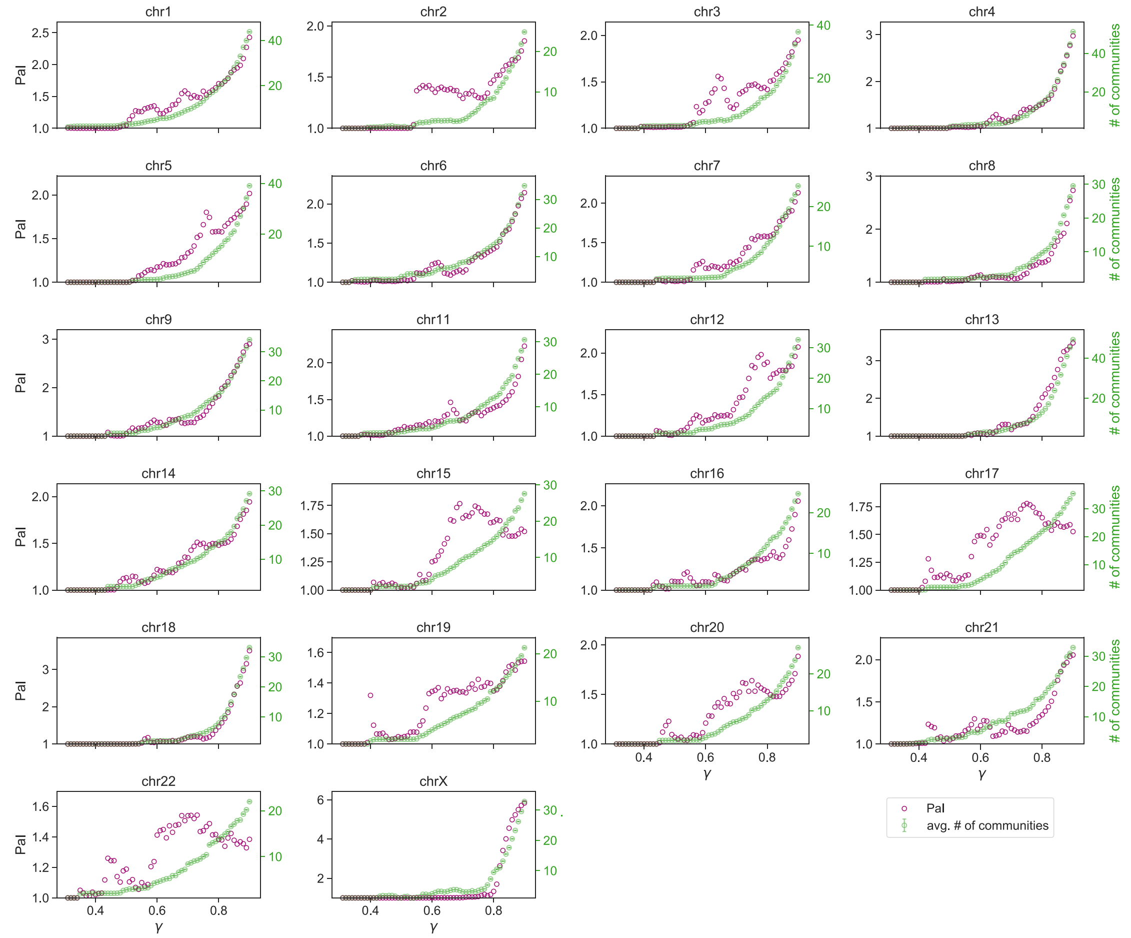


Figure 7. Chromosome-wide global community inconsistency measured by PaI. Each panel shows the values of the PaI metric (the violet circles) and the average number of communities (the green circles) across a range of γ values with the small (smaller than the symbols themselves for all cases) error bars representing the standard deviation.

3.4. Node membership correlations

When analyzing the PaI and MeI scores across γ, we noted that the Hi-C networks exhibit relatively few independent partitions (e.g. $\text{PaI}_{\text{chr10}} < 3$), and that each node belongs to just a few communities (median MeI < 4). This suggests that the community partitions are correlated. To better understand these correlations, we use a stochastic embedding technique called t-SNE that projects high-dimensional data clusters on a 2D plane (see section 2.4). In our case, the data set is the community membership per node over 100 GenLouvain runs.

We show the t-SNE analysis in figure 8 for three γ values, where each filled circle represents a network node (again, using chromosome 10). The closer two circles appear in the plot, the more correlated their node-community memberships are. As we increase γ, we note that node clusters split and that some circles become isolated. We interpret this as the ensemble of network partitions grows with γ and becomes increasingly dissimilar.

While the clustering is identical in panels (a) and (b), we color-coded them differently to highlight specific features. In (a), the colors represent the local inconsistency score (MeI). We note that nodes with high MeI tend to separate from those with low MeI. We also see that the low MeI nodes have relatively stronger correlations, thereby forming more distinct clusters. Panel (a) also indicates that nodes with low MeI have similar node-community memberships.

In panel (b), the color-coding illustrates the chromatin type. To simplify the analysis, we consider two large chromatin groups—Euchromatin and Heterochromatin—instead of the five we used before (section 2.5). These groups reflect the traditional division into open and closed chromatin associated with active transcription and repression, respectively. In terms of our previous definitions, the two groups are:

Euchromatin: A, B, and C,
Heterochromatin: D.

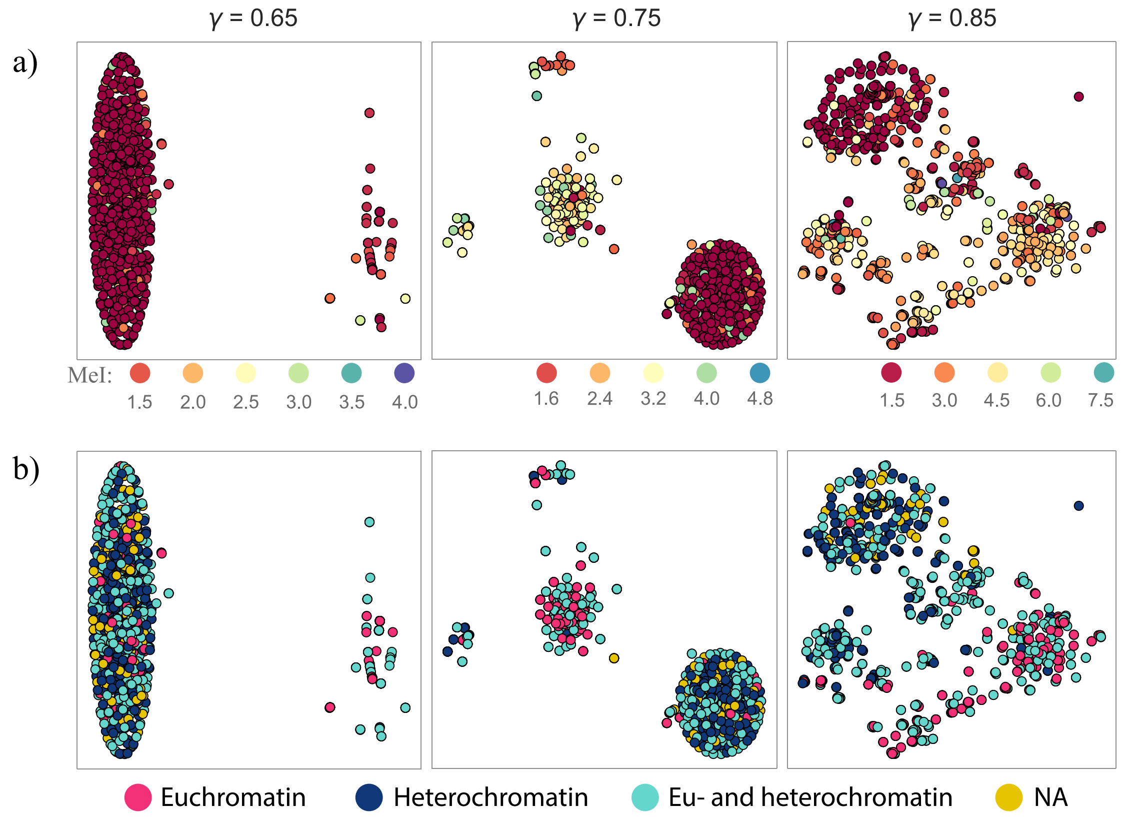


Figure 8. Node-community membership correlations across three resolution parameters ($\gamma = 0.65, 0.75, 0.85$) visualized with t-SNE dimension reduction. Nodes are represented as colored circles. Those having correlated community memberships over several GenLouvain realizations tend to cluster. In contrast, nodes that repel each other belong less likely to the same community. In (a), the colors indicate MeI scores (see the legend under each plot). In (b) chromatin types (Eu- or Heterochromatin).

Note that we disregarded group E ('Insulators') as it is associated with boundaries rather than long chromatin stretches, such as Eu- and Heterochromatin. Also, while most nodes belong to either Eu- or Heterochromatin, some are enriched in both types[7]. We call this group 'mixed'. Finally, there is yet another node group that does not enrich any of the two chromatin types ('NA').

In panel (b), we observe that community membership correlations are associated with chromatin type. For example, when $\gamma = 0.75$, Euchromatin and Mixed nodes separate from the large cluster and form new sub-groups. The nodes in these subgroups generally have high MeI scores indicating a larger variability in their community memberships.

Overall, panels (a) and (b) show a scale-dependent separation between communities associated with active and inactive DNA regions (red and blue nodes repel each other). This separation resembles the A/B compartmentalization but for small-scale 3D structures. Furthermore, the MeI score suggests that these structures have multiple independent ways to assemble if formed from Euchromatin nodes. This observation hints at higher structural variability of the accessible genome, which may reflect the dynamic nature of gene expression processes.

4. Conclusions

There is a growing awareness that Hi-C networks have a complex scale-dependent community structure. While some communities have a hierarchical or nested organization, others form a patchwork of partially overlapping communities. In this regard, Hi-C networks do not represent exceptions. Instead, they belong to the norm: most complex networks show convoluted multi-scale behaviors whenever competing organization principles shape the network structure. These principles force some nodes into ambiguous community memberships, making network partitioning challenging.

[7] Imagine a Venn diagram with two large circles portraying the chromatin enrichment for each node. While most nodes separate into either Eu- or Heterochromatin, the diagram shows a significant overlap where some nodes are enriched in both chromatin types. These nodes belong to the mixed group.

To better understand the scales where this might cause problems when clustering Hi-C data, we have analyzed the node-community variability over an ensemble of network partitions and estimated the ensemble's size. We have found that it typically grows as we zoom in to the network. However, this trend has significant breaks where the ensemble size drops at some specific network scales. This drop narrows the distribution of possible network partitions. We hypothesize that these minima represent the most common partitions of the average 3D chromosome organization (over a cell population).

Moreover, we have found nodes that belong to several communities when calculating the node-community membership variability. These ambiguous nodes act as bridges and are associated with specific chromatin types. For example, we have found the highest variability for nodes classified as enriched in active chromatin. This finding contrasts inactive (or repressed) chromatin nodes that typically exhibit a relatively more consistent community organization. One explanation is Euchromatin's somewhat higher physical flexibility when exploring the nuclear 3D proximity in search of other DNA regions to form functional contacts (see a recent review [34]). An alternative explanation is that fuzzy node-community memberships reflect significant cell-to-cell variations. While some physical interactions might be stable in one cell, they may be absent in another. Therefore, as Hi-C maps portray the average contact frequency over many cells, this variability may manifest in ambiguous nodes.

Finally, a recent study investigated the challenges in finding reliable communities in Hi-C data [12]. This work aims to map out the landscape of feasible network partitions in Hi-C networks and found that the width of the landscape is scale-dependent. Our study takes a more node-centric view, where we calculated the local inconsistency of individual nodes and discovered that some nodes have fuzzy node-community memberships. Both studies highlight that finding reliable communities in Hi-C data is challenging, especially on some scales. One root cause is that Hi-C networks are almost entirely connected (with many weak links). Under these circumstances, we expect that Hi-C networks have several community divisions. Divisions that cannot be distinguished without additional data, such as gene expression or epigenetic profiles. This fundamental problem suggests that there is a significant likelihood of disagreement on the ideal network division between any community-finding or data-clustering methods. This challenge has likely contributed to debates on the actual differences between TADs and sub-TADs [35, 36].

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://github.com/lizanalab/bernenko2023exploring.git.

Acknowledgments

The authors thank Daekyung Lee for providing the Python codes to calculate the inconsistency measures [16]. S H L was supported by the National Research Foundation (NRF) of Korea Grant Nos. NRF-2021R1C1C1004132 and NRF-2022R1A4A1030660. L L acknowledges financial support from the Swedish Research Council (Grant No. 2021-04080).

ORCID iDs

Dolores Bernenko https://orcid.org/0000-0002-6618-8232
Sang Hoon Lee https://orcid.org/0000-0003-3079-5679
Ludvig Lizana https://orcid.org/0000-0003-3174-8145

References

[1] Lieberman-Aiden E *et al* 2009 Comprehensive mapping of long-range interactions reveals folding principles of the human genome *Science* **326** 289–93

[2] Dixon J R, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu J S and Ren B 2012 Topological domains in mammalian genomes identified by analysis of chromatin interactions *Nature* **485** 376–80

[3] Rao S S P *et al* 2014 A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping *Cell* **159** 1665–80

[4] Magaña-Acosta M and Valadez-Graham V 2020 Chromatin remodelers in the 3D nuclear compartment *Front. Genet.* **11** 600615

[5] Newman M 2018 *Networks* 2nd edn (Oxford University Press)

[6] Porter M A, Onnela J-P and Mucha P J 2009 Communities in networks *Not. Am. Math. Soc.* **56** 1082–97 (available at: www.ams.org/notices/200909/rtx090901082p.pdf)

[7] Fortunato S 2010 Community detection in graphs *Phys. Rep.* **486** 75–174

[8] Sefer E 2022 A comparison of topologically associating domain callers over mammals at high resolution *BMC Bioinform.* **23** 127

[9] Sarnataro S, Chiariello A M, Esposito A, Prisco A, Nicodemi M and Oliva B 2017 Structure of the human chromosome interaction network *PLoS One* **12** 1–15

[10] Lee S H, Kim Y, Lee S, Durang X, Stenberg P, Jeon J-H and Lizana L 2019 Mapping the spectrum of 3D communities in human chromosome conformation capture data *Sci. Rep.* **9** 1–7

[11] Bernenko D, Lee S H, Stenberg P and Lizana L 2022 Mapping the semi-nested community structure of 3D chromosome contact networks *PLoS Comput. Biol.* **19** e1011185

[12] Holmgren A, Bernenko D and Lizana L 2022 Mapping robust multiscale communities in chromosome contact networks (arXiv:2212.08456)

[13] Arenas A, Fernández A and Gómez S 2008 Analysis of the structure of complex networks at different resolution levels *New J. Phys.* **10** 053039

[14] Newman M E J 2016 Equivalence between modularity optimization and maximum likelihood methods for community detection *Phys. Rev.* E **94** 052315

[15] Kim H and Lee S H 2019 Relational flexibility of network elements based on inconsistent community detection *Phys. Rev.* E **100** 022311

[16] Lee D, Lee S H, Kim B J and Kim H 2021 Consistency landscape of network communities *Phys. Rev.* E **103** 052306

[17] Riolo M A and Newman M E J 2020 Consistency of community structure in complex networks *Phys. Rev.* E **101** 052306

[18] Edgar R, Domrachev M and Lash A E 2002 Gene expression omnibus: NCBI gene expression and hybridization array data repository *Nucleic Acids Res.* **30** 207–10

[19] Knight P A and Ruiz D 2013 A fast algorithm for matrix balancing *IMA J. Numer. Anal.* **33** 1029–47

[20] Newman M E J and Barkema G T 1998 *Monte Carlo Methods in Statistical Physics* (Clarendon)

[21] Blondel V D, Guillaume J-L, Lambiotte R and Lefebvre E 2008 Fast unfolding of communities in large networks *J. Stat. Mech.* **2008** P10008

[22] Jeub L G S, Bazzi M, Jutla I S and Mucha P J 2011–2019 A generalized Louvain method for community detection implemented in MATLAB (available at: https://github.com/GenLouvain/GenLouvain)

[23] Kwak H, Moon S, Eom Y-H, Choi Y and Jeong H 2011 Consistent community identification in complex networks *J. Korean Phys. Soc.* **59** 3128–32

[24] Lancichinetti A and Fortunato S 2012 Consensus clustering in complex networks *Sci. Rep.* **2** 336

[25] Gates A J, Wood I B, Hetrick W P and Ahn Y-Y 2019 Element-centric clustering comparison unifies overlaps and hierarchy *Sci. Rep.* **9** 8574

[26] Jeh G and Widom J 2003 Scaling personalized web search *WWW '03: Proc. 12th International Conference on World Wide Web* (Association for Computing Machinery) p 271

[27] Menczer F, Fortunato S and Davis C A 2020 *A First Course in Network Science* (Cambridge University Press)

[28] van der Maaten L and Hinton G 2008 Visualizing data using t-SNE *J. Mach. Learn. Res.* **9** 2579–605

[29] Pedregosa F *et al* 2011 Scikit-learn: machine learning in Python *J. Mach. Learn. Res.* **12** 2825–30

[30] Ernst J *et al* 2011 Systematic analysis of chromatin state dynamics in nine human cell types *Nature* **473** 43–49

[31] Encode (available at: http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeBroadHmm)

[32] Ernst J and Kellis M 2010 Discovery and characterization of chromatin states for systematic annotation of the human genome *Nat. Biotechnol.* **28** 817–25

Ernst J Kheradpour P, Mikkelsen TS, Shoresh N, Ward L D, Epstein C B, Zhang X, Wang L, Issner R, Coyne M and Ku M 2011 Mapping and analysis of chromatin state dynamics in nine human cell types *Nature* **473** 43–49

[33] Hofmann H, Wickham H and Kafadar K 2017 Value plots: boxplots for large data *J. Comput. Graph. Stat.* **26** 469–77

[34] Misteli T 2020 The self-organizing genome: principles of genome architecture and function *Cell* **183** 28–45

[35] Dixon J R, Gorkin D U and Ren B 2016 Chromatin domains: the unit of chromosome organization *Mol. Cell* **62** 668–80

[36] Eres I E and Gilad Y 2021 A tad skeptic: is 3D genome topology conserved? *Trends Genet.* **37** 216–23