PLOS COMPLEX SYSTEMS

RESEARCH ARTICLE

Identifying stable communities in Hi-C using multifractal network modularity

Lucas Hedström [ID]*, Antón Carcedo, Ludvig Lizana [ID]

Integrated Science Lab, Department of Physics, Umeå University, Umeå, Sweden

* lucas.hedstrom@umu.se

Abstract

Chromosome capture techniques like Hi-C have expanded our understanding of mammalian genome 3D architecture and how it influences gene activity. To analyze Hi-C data sets, researchers increasingly treat them as DNA-contact networks and use standard community detection techniques to identify mesoscale 3D communities. However, there are considerable challenges in finding significant communities because the Hi-C networks have cross-scale interactions and are almost fully connected. This paper presents a pipeline to distill 3D communities that remain intact under experimental noise. To this end, we bootstrap an ensemble of Hi-C datasets representing noisy data and extract 3D communities that we compare with the unperturbed dataset. Notably, we extract the communities by maximizing local modularity (using the Generalized Louvain method), which considers the multifractal spectrum recently discovered in Hi-C maps. Our pipeline finds that stable communities (under noise) typically have above-average internal contac,t frequencies and tend to be enriched in active chromatin marks. We also find they fold into more nested cross-scale hierarchies than less stable ones. Apart from presenting how to systematically extract robust communities in Hi-C data, our paper offers new ways to generate null models that take advantage of the network's multifractal properties. We anticipate this has a broad applicability to several network applications.

Author summary

Understanding the 3D structure of DNA inside cells is crucial for studying gene activity, as the two are often intricately connected. This DNA 3D structure is commonly analyzed using data from a technique called Hi-C, but the complex nature of DNA interactions makes this challenging. Our study introduces a new method to identify stable structures in DNA folding data that are resistant to noise. By generating multiple noisy versions of the Hi-C data and applying a specialized network clustering algorithm, we found that certain DNA structures remain more stable than others. These stable structures have stronger internal connections, are hierarchically organized at different scales, and consist of more active DNA regions. Our method not only improves the analysis of Hi-C data

but also offers new ways to study complex networks to better understand gene regulation and organization.

1. Introduction

Mammalian genomes fold into complex 3D structures that help regulate genetic processes such as transcription, DNA replication and repair, and epigenetics [1–5]. Several of these insights derive from chromosome capture techniques, such as Hi-C [6–8], that quantifies the number of physical contacts between all DNA segment pairs in the genome across a cell population. Data from these experiments allowed the construction of comprehensive chromosome-wide 3D interaction maps, highlighting local 3D structures with high internal frequencies, such as Topologically Associated Domains (TADs) and the large-scale binary division into A/B compartments generally associated with active (A) and inactive (B) chromatin [9,10].

However, these 3D structures represent but two instances on a wide spectrum of structures that fold into each other, forming complex, often nested, hierarchies [11–15]. These hierarchies make it difficult to design robust TAD-finding algorithms ("TAD callers") and for the field to agree on common TAD definitions [16,17]. This ambiguity is likely one reason there exists a large collection of TAD callers [18], each associated with their own parameters and arbitrary choices. In practice, this means that two callers may disagree on the TAD boundaries using the same dataset, let alone two datasets deriving from theoretically identical replicate experiments with experimental and biological noise [19].

To help resolve some of these issues, we recently proposed a method to extract robust 3D communities in Hi-C data [20]. In that paper, we used a network community detection algorithm (Generalized Louvain) to generate an ensemble of feasible network partitions from a public Hi-C data set [21] and quantified the most conserved node-community relationships. This approach distinguishes robust from variable communities and tracks how the robustness changes with the network scale (and chromatin state). We found that robustness is highly scale-dependent and typically worse for small communities, including TADs. When robustness is low, there are many ways to split the network into mathematically feasible community partitions, which is likely one reason TAD callers disagree, as each focuses on slightly different data features [22].

This paper extends our study [20] by addressing how community robustness changes under experimental noise rather than variability in the community detection method. To this end, we first estimate noise levels from actual Hi-C maps (from humans [21]) and construct an ensemble of blurred maps by adding noise to each contact (or pixel). We call these bootstrapped maps [23]. Next, we extract 3D communities by considering the Hi-C data as a network and using the Generalized Louvain algorithm. Finally, we quantify the overlap between the bootstrapped maps relative to the unperturbed map. We find that stable communities typically have above-average internal contact frequencies, tend to be enriched in active chromatin marks, and fold into more nested cross-scale hierarchies than less stable ones.

In addition, we made yet another improvement that reaches beyond Hi-C data. This improvement takes advantage of the network's multifractal spectrum [24], in a way we anticipate is useful for other community detection approaches. When using Generalized Louvain, or other methods resting on maximum modularity [25], users must specify a background network connectivity or null model. Simple null models like the random Newman-Girvan model [26] are common choices in network science due to limited alternatives. However, this model does not apply to Hi-C data since it disregards the spatial constraints associated

with the DNA segments (represented as nodes) belonging to a long linear molecule. One way to deal with this specific problem is to introduce a distant-dependent null model, where the contacts decay as a power law (as discovered in Hi-C) [14,27]. But as the power-law exponent changes with distance, too, (from ≈ 0.75 to ≈ 1.08 [28]) this introduces a new arbitrary parameter as to when to switch exponent. While we explored both exponents in [20], here we remove this arbitrary choice by implementing a one-parameter null model that we fit the Hi-C map's bifractal spectrum using techniques from [24]. This approach is not limited to Hi-C data and should apply to any complex network.

2. Materials and methods

Our computational method has several steps (Fig 1). First, we generate bootstrapped samples representing noisy Hi-C data. Second, we find communities using a method that considers the bifractal structure in Hi-C data. Third, we step through the pipeline to determine the stable communities that are invariant under noise.

2.1. Collecting and preprocessing experimental data

Chromosome contacts. We downloaded Hi-C data for the B-lymphoblastoid human cell line (GM12878) [21] from the GEO database (`MAPQG0` dataset, 100 kb resolution) [29]. We only consider intra-chromosome contacts in our analysis. We interpret the data file as a weighted network ("Hi-C network") with elements A_{ij} in sparse form, where each node represents 100 kb of DNA, and the link weight is the measured contact count.

Before constructing the Hi-C network, we normalize the data using the standard Knight-Ruiz matrix balancing algorithm [30] so that the sum of all weights to a node equals unity.
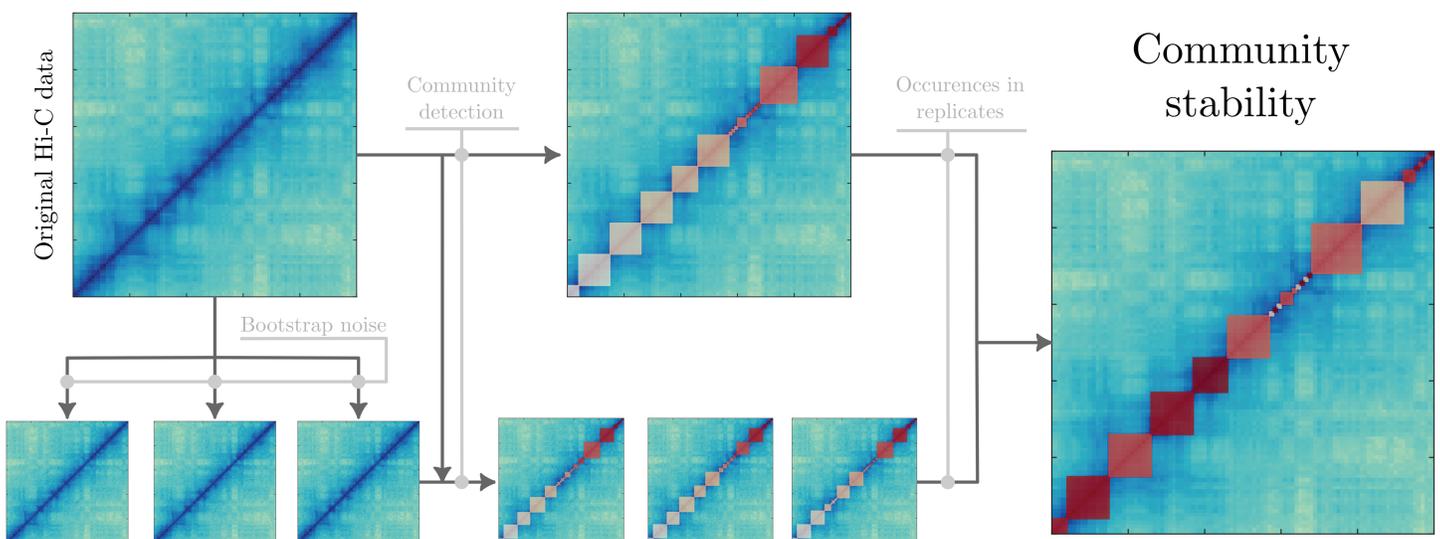
Fig 1. A schematic showcasing the methodology that computes the community stability for a Hi-C network. Using the original Hi-C data, we bootstrap by sampling data from the contact distributions associated with each distance (Fig 2). We then run a community detection algorithm on the original Hi-C data to get the original partition (GenLouvain). To reduce the number of samples required for good statistics, we use the original partition as an initial partition for GenLouvain. Each sample gives a list of partitions, each with unique communities, shown as boxes along the diagonal colored by the community number ID. Based on the overlap (calculated with the Jaccard index), we count the number of times each community in the original Hi-C data appears in the bootstrapped data, which we use as a proxy for stability. (rightmost panel) Original partition, where a darker color indicates communities with higher stability.

Note that the centromere (with no measured contact counts) must be removed before KR-normalization to avoid convergence issues. However, since the indices on the Hi-C matrix represent physical distances, we reintroduce the centromere before performing community detection (see Sect 2.3).

Chromatin states. To relate our results to different functional genomic regions, we downloaded a dataset from ENCODE [31] of 15 different chromatin types created from a multivariate hidden Markov model (denoted HMM states) [32,33]. To reduce the number of states, we grouped the data into five effective categories: Promoters, Enhancers, Transcribed regions, Heterochromatin, and Insulators (see Appendix C in S1 File 4 for complete definitions in terms of the standard HMM states). Moreover, this data set allows us to calculate the chromatin state's folds of enrichment (FE) for each Hi-C bin. We calculate enrichment relative to the chromosome-wide average (see Appendix C in S1 File 4).

2.2. Bootstrapping networks

To estimate community stability across replicates, we bootstrapped data that simulated noise from a typical Hi-C experiment. This noise comes from multiple sources [19], such as randomness of ligations between loci [34] or random contacts between loci due to fluctuations [6]. In Fig 2, we show histograms of the logged Hi-C contacts $\log(A_{ij})$ for four fixed distances in chromosome 10. We note that the distributions at all these distances are log-normal, albeit with their unique mean and standard deviation. We use these empirical distributions as a proxy for experimental noise.

To generate bootstrapped Hi-C maps with noise, we calculate the standard deviation σ_d of the logged contacts for a fixed distance $d = |i - j|$ as

$$\sigma_d = \sqrt{\langle \log^2(A_{ij}) \rangle_d - \langle \log(A_{ij}) \rangle_d^2}, \tag{1}$$
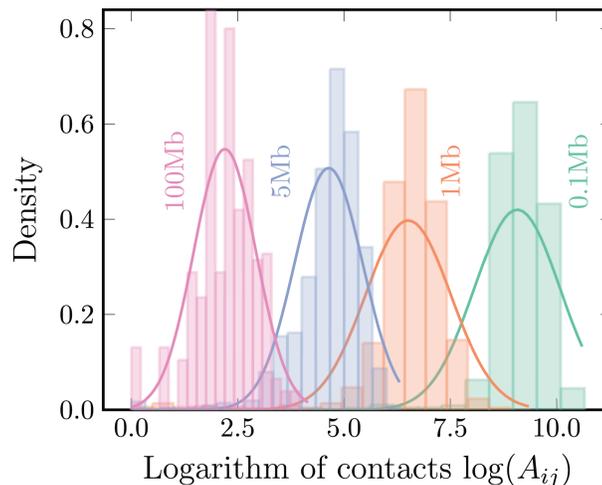
Fig 2. Histogram of the logged Hi-C contacts A_{ij} at four distances for chromosome 10 (bars). Overlaying the histogram as solid lines, we depict log-normal distributions fitted with the mean and standard deviation from the data.

where $\langle...\rangle_d$ denotes the average at the distance d. Next, we bootstrap new noisy contacts A'_{ij} for each bin using

$$A'_{ij} = A_{ij} \times e^{N(0,(\epsilon\sigma_d)^2)} \tag{2}$$

where ϵ is the noise amplitude, and $N(0, (\epsilon\sigma_d)^2)$ denotes the normal distribution with mean 0 and variance $(\epsilon\sigma_d)^2$. As shown below, we tune the amplitude ϵ to not completely disrupt the chromosome's community structure (e.g., the TADs), yet yield maps with realistic noise.

2.3. Detecting communities using a multifractal null model

To find Hi-C network communities, or "3D communities" [14], we use the Generalized Louvain method (GenLouvain) implemented in MATLAB [35]. GenLouvain searches for network partitions that maximize the modularity function Q, capturing local deviations from the expected background connectivity or null model. In its general form, the modularity is expressed as

$$Q = \frac{1}{2m} \sum_{i \neq j} \left(A_{ij} - \gamma P_{ij}^{(0)} \right) \delta_{C_i, C_j} \tag{3}$$

where A_{ij} are entries in the weighted adjacency (Hi-C) matrix, m is the total weight of this matrix, γ is a scale parameter setting the effective community size, C_i is node i's community assignment, and $P_{ij}^{(0)}$ is the null model. The null model is essential as it carries the underlying assumptions of the network structure. We discuss how this is chosen below. The last factor, δ_{ij}, denotes the Kronecker-delta function and indicates that the sum calculates the partial sum of all local modularities within a community.

As just mentioned, to calculate Q, one must specify a realistic $P_{ij}^{(0)}$. The most common choice is random connections, known as the Newman-Girvan null model [26], but it does not account for constraints associated with long DNA polymer chain folded in 3D inside the cell nucleus [36,37]. Previous work generalized the Newman-Girvan model to include a power-law with linear node separation d as observed in Hi-C maps and predicted by theoretical polymer models [14]. However, the power-law exponent tends to change with distance [6,28]. For example, within TADs, the contacts decay on average as $d^{-0.75}$, whereas it is closer to $d^{-1.08}$ for longer distances. Therefore, if using a single decay function as the null hypothesis, one must introduce an arbitrary cutoff at which the decay exponent should change. To circumvent this problem, we propose another null model embracing the multifractal, or bi-fractal, properties of Hi-C maps in humans and mice [24]. Admittedly, this model also has a free parameter, but the fitting to Hi-C data is more systematic.

This multifractal model is the so-called Hierarchical Domain Model (HDM) [24]. It originates from turbulence (where it is better known as the β-model [38] and aims to understand the hierarchical energy transfer among various scales of fluid motion and, ultimately, the famed Kolmogorov scaling.) In practice, the HDM is an iterative procedure that starts with a 2×2 matrix with diagonal values a and off-diagonal values b ($a > b$). At each iteration n, the off-diagonal value becomes a new 2×2 block $2^n \times 2^n$ matrix, where diagonal blocks are multiplied with the original matrix. By interpreting the elements in the $2^n \times 2^n$ matrix as pixel values (or contact probabilities) in a Hi-C map, it is possible to vary a and b to fit the HDM to the empirical data. This was the key insight in [24]. Mathematically, we express the HDM

matrix elements $H_a(i,j)$ as (derived in [24])

$$H_a(i,j) =$$
$$\begin{cases} \dfrac{1}{2^n - d} \sum_{k=0}^{n-1} 2^n g(d2^{k-n}) \dfrac{a^k b^{1-\delta_{kn}}}{4^{\max(n-k-1,0)}} & \text{if } d < 2^n \\[2ex] \sum_{k=0}^{n-1} 2^n g(2^k - 2^{k-n}) \dfrac{a^k b^{1-\delta_{kn}}}{4^{\max(n-k-1,0)}} & \text{otherwise} \end{cases} \tag{4}$$

where

$$g(x) = \begin{cases} x & \text{if } 0 \le x \le 0.5 \\ 1 - x & \text{if } 0.5 \le x \le 1 \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

and $n = \lfloor \log_2(N) \rfloor$ (N is the size of the Hi-C matrix). Also, we imposed the condition $b = 0.5 - a$ to ensure proper normalization at each iteration. This means that the HDM only depends on a (hence the single subscript in $H_a(i,j)$). Using least-squares, we fitted to our Hi-C data and found that $a = 0.434$ with minor variation among chromosomes. This value yields an accurate decay exponent in Hi-C contact maps at all scales (see Appendix A and Fig A in S1 File, 4).

Finally, based on the mathematical definition of $H_a(i,j)$, we express the null model $P_{ij}^{(0)}$

$$P_{ij}^{(0)} = 2m \frac{H_a(i,j)}{\sum_{i,j} H_a(i,j)}. \tag{6}$$

We defer to the Appendix A in S1 File, 4 for more details.

2.4. Reducing community detection variance

Since the GenLouvain algorithm tries to optimize the modularity function in a Monte Carlo-like fashion, it often finds different local minima, producing slightly different partitions with similar quality [20] (also discussed in Appendix A in S1 File, 4). To reduce the number of samples required to get good statistics, we use the partition detected in the original data as the starting partition and then generate a collection of bootstrapped samples, see Fig 1. This procedure reduces the number of required samples since we do not have to account for varying initial conditions.

Note that the "stable communities" we study here differ from the "core communities" in ref [20]. While core communities are robust across an ensemble of GenLouvain partitions associated with a fixed γ, the stable communities maintain robustness under artificial experimental noise relative to a select partition.

3. Results

3.1. Fine-tuning noise amplitudes

To find realistic noise levels ϵ when bootstrapping our noisy Hi-C maps, we benchmarked against measured TAD stabilities between replicate experiments [16]. These experiments found the overlap of TAD boundaries (points in the Hi-C map where two TAD communities meet) between replicates to be 62%. We generated communities using GenLouvain, where we set the scale parameter γ to achieve an effective community size $\hat{s} \sim 0.88$ Mb that

best agrees with the TADs considered in [16] (see Appendix B in S1 File, 4). Then we generated sets of communities for n_{samp} bootstrapped samples for different values of ϵ. To calculate the TAD boundary similarity between two sets of community boundaries b_i and b_j from two bootstrapped samples i and j, we use the Jaccard index

$$J_{\text{TAD}}(b_i, b_j) = \frac{|b_i \cap b_j|}{|b_i \cup b_j|} \tag{7}$$

which varies between 1 (identical) and 0 (dissimilar).

However, Eq (7) has a few problems when using it to quantify boundary overlaps. For example, under high noise amplitudes, the communities shrink until each node is its own community, or the communities split into randomly scattered nodes over the network. This situation yields $J_{\text{TAD}}(b_i, b_j) \sim 1$, indicating a large overlap between community borders, but where the partitions no longer represent TAD-sized communities. To solve this problem, we multiply the Jaccard index in Eq (7) by a "penalty factor" containing the fraction of the number of communities in the bootstrapped sample compared to the original. This gives

$$J_{\text{TAD}}^*(b_i, b_j) = J_{\text{TAD}}(b_i, b_j) \times \left(1 - \frac{|b_{\text{min}} - |b_o||}{b_{\text{min}}}\right) \tag{8}$$

where b_o are the TAD boundaries calculated from the original dataset without noise and $b_{\text{min}} = \min_{k=i,j}(|b_k|)$.

To better understand how the penalty factor works, consider two illustrative examples. One is where the bootstrapped partitions contain the same number of communities as the original dataset but shifted along the chromosome. Here, the overlap is low, making the term $|b_i \cap b_j|/|b_i \cup b_j|$ small. Another case is when the bootstrapped samples have many small communities (in the extreme case, where every node represents a separate community). Then, the overlap between bootstrapped samples will be large, indicating a high $J_{\text{TAD}}(b_i, b_j)$, but the number of communities compared to the original partition differs significantly. Therefore, the last term $|b_{\text{min}} - |b_o||/b_{\text{min}}$ approaches 1, leading to small $J_{\text{TAD}}^*(b_i, b_j)$, suggesting poor overlap.

Finally, to provide a global measure of TAD stability in an ensemble of bootstrapped samples, we calculate the mean $J_{\text{TAD}}^*(b_i, b_j)$ across all boundaries

$$J_{\text{TAD}}^* = \frac{2}{n_{\text{samp}}(n_{\text{samp}} - 1)} \sum_{b_i \neq b_j} J_{\text{TAD}}^*(b_i, b_j) \tag{9}$$

where n_{samp} is the number of bootstrapped samples.

We show J_{TAD}^* over a range of noise amplitudes ϵ across all chromosomes in Fig 3. As expected, we find perfect overlap ($J_{\text{TAD}}^* \to 1$) when the noise level becomes low ($\epsilon \to 0$). Also, as ϵ increases, the overlap reduces below the experimental benchmark (0.62, dashed horizontal line) until reaching a plateau where each node represents separate communities. Using the 0.62-line, we may find numerically the critical noise amplitude ϵ^* for each chromosome. We note that ϵ^* varies between $\sim 0.1 - 0.3$. This parameter presents an analog to stability since a higher ϵ^* for a specific chromosome, compared to the mean, indicates that the chromosome is more resistant to noise, as the original TADs are still extractable with a larger noise factor. From Fig 3, we choose $\epsilon^* = 0.18$ as the critical noise amplitude for all chromosomes (vertical dashed line).
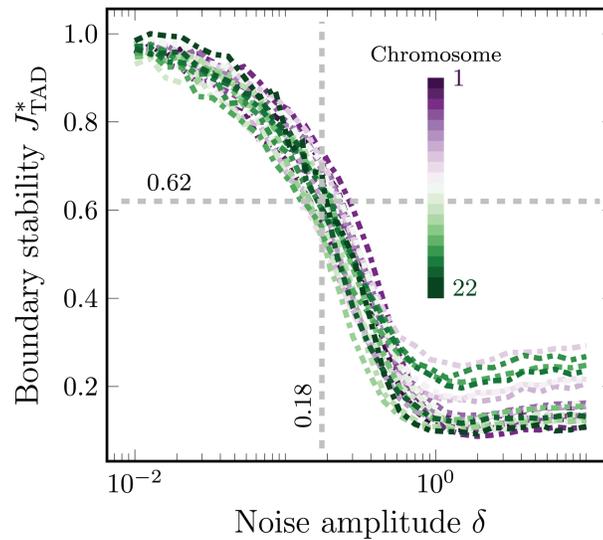
Fig 3. TAD boundary stability for different noise amplitudes for all chromosomes. The dash-dotted lines show the modified Jaccard index in Eq 9 (averaged over 100 samples per chromosome). As the noise ϵ increases, the overlap decreases. We note that every chromosome follows a similar decaying pattern, where the stability drops and reaches a plateau. The horizontal dashed line at $J^*_{\text{TAD}} = 0.62$ indicates the experimental boundary overlap from [16]. This intersects all J^*_{TAD} lines at slightly different critical noise amplitudes. The vertical dashed line shows the mean of all these amplitudes, $\epsilon^* = 0.18$.

https://doi.org/10.1371/journal.pcsy.0000053.g003

3.2. Identifying stable communities

Once we tuned the noise amplitude, we generated several bootstrapped Hi-C maps (see Sect 2.2) and extracted the parts of the community partition that stayed consistent under noise relative to the original dataset. The specific procedure reads as follows (inspired by [23]):

- Get the original community partition P_o.
- Generate n_{samp} bootstrapped samples from the original dataset.
- Get communities from each bootstrapped sample P_n, based on the starting partition P_o.
- Identify the communities from the original partition P_o that are also in the bootstrapped samples. To do this, we calculate the Jaccard indices in each sample considering all nodes between community pairs and pick the ones with the largest Jaccard index.
- Filter all pairs with a Jaccard index smaller than some cutoff J_{cutoff} and calculate the fraction of bootstrapped samples that each community in P_o appears in.

Important with this algorithm is that the Jaccard index no longer considers only the boundary nodes like in Eq 9, but rather the Jaccard index of all nodes between two communities. Based on these steps, we calculate the stability fraction S_{C_i}, for community C_i in the original partition P_o as

$$S_{C_i} = \frac{1}{n_{\text{samp}}} \sum_{n=1}^{n_{\text{samp}}} \theta \left(\max_{C_j \in P_n} J(C_i, C_j) - J_{\text{cutoff}} \right). \tag{10}$$

Here, $\theta(\cdot)$ denotes the Heaviside step function, and $J(C_i, C_j) = |C_i \cap C_j|/|C_i \cup C_j|$ is the Jaccard index considering all nodes in C_i and C_j.

The cutoff J_{cutoff} is a free parameter. We set it to 1.0, meaning that we define a community C_i as stable if all of its nodes remain in the original dataset and the bootstrapped samples (however, this choice is not too important since the stability metric correlates for smaller cutoffs. See Appendix D in S1 File, 4.) After setting this cutoff, we generated several bootstrapped Hi-C maps and calculated the stability fraction S_{C_i} (Eq (10)) for each community C_i associated with the original dataset (P_o). As before, we set γ to get the effective community size $\hat{s} \sim 0.88Mb$ for each chromosome.

We show the distributions of the stabilities S_{C_i} as boxplots in Fig 4, one box per chromosome, sorted by the S_{C_i} medians. By reading the chromosome numbers and the x-axis, we note no strong correlation between chromosome size and stability, as both large and small chromosomes appear on either end. To compare this result to experiments, we note that the same work that calculated the median TAD boundary similarity to be 62% also used an advanced similarity metric TADsim to measure the structural similarity of TAD sets between replicates [16]. They found this overlap to be lower, about 50% between replicates. This is close to our median stability between all chromosomes, which turns out to be $\sim$0.58.

Until this point, we studied each chromosome separately. In the sections that follow, we aggregate all the data and consider genome-wide averages. We also used $n_{\text{samp}} = 1000$ bootstrapped samples per chromosome to calculate the stability, which we bin to a resolution of 1%.

3.3. Understanding stability and community connectedness

The boxplots and the distribution in Fig 4 foreshadow that some of the communities have a more resilient community structure than others under noise. Here, we ask how much this stability depends on the local structure of the underlying network. In particular, we study the communities' internal node strength. This metric differs from general node strength—the sum of all edge weights of a node—by restricting the sum to nodes within a community. Node strength is a good predictor of, for example, node centrality in a fully connected network [39] or search times to a random node [40–43].
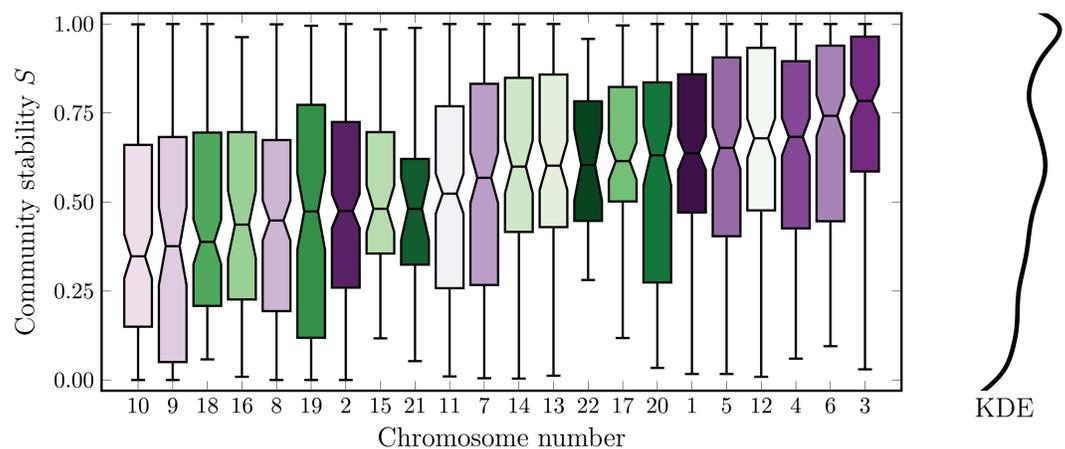
Fig 4. Boxplots showing the distribution of community stability at an effective community size of $\sim$0.88Mb across all chromosomes using $\epsilon^* = 0.18$ and a Jaccard index cutoff of 1.0. The kernel density estimation (KDE) for the distribution of all stabilities over all chromosomes is shown to the right. With these parameters, the stability median over all chromosomes is 0.58. We note that the total distribution has two peaks around $\sim$1.0 and $\sim$0.6.

https://doi.org/10.1371/journal.pcsy.0000053.g004

However, in our case, the stability of a community is not related to how well the nodes are connected to other parts of the network. Instead, we quantify the internal community connectedness. We define this as the sum of all edge weights within a community (including self-loops) normalized by the total node strength

$$\hat{s}_{C_i}^{\text{int}} = \frac{\sum_{j \in C_i} \sum_{i \in C_i} A_{ij}}{\sum_{j} \sum_{i \in C_i} A_{ij}} \qquad (11)$$

where "int" stands for internal node strength. Plotting $\hat{s}_{C_i}^{\text{int}}$ for all communities across all chromosomes against the community stability S_{C_i} (Eq (10)), we find that high internal node strength suggests high community stability under noise (Fig 5). This means that strong inter-connected communities are less affected by random noise during bootstrapping.
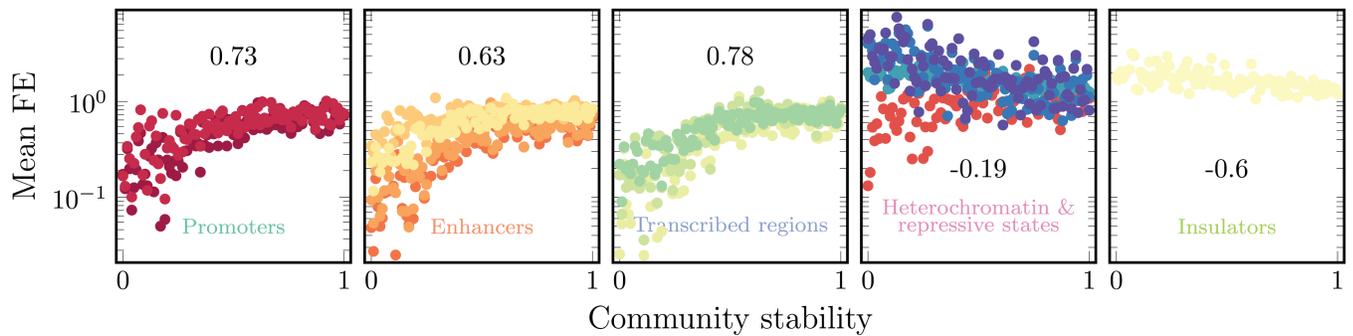
3.4. Relating community stability to chromatin states

The previous section studied how stability is connected to network structure. Next, we perform a similar analysis using biologically relevant markers: chromatin states. In particular, we ask if our stability metric calculated on TAD-sized communities is associated with the enrichment or depletion of active or inactive genomic regions, which could suggest underlying mechanisms that form strongly connected communities. For example, there are several indications of insulation enrichment at TAD borders and facilitated promoter-enhancer interactions within TADs [44,45].

To study this, we used data that classify genomic regions into 15 different functional states (see Appendix C in S1 File, 4), which we grouped into five categories (defined in Appendix C in S1 File, 4). We plotted the folds of enrichment (FE) of these five groups against community stability for TAD-sized communities ($\hat{s} \sim 0.38$ Mb) in Fig 6. For three of the groups associated with active genomic regions (promoters, enhancers, and transcribed regions), there is a
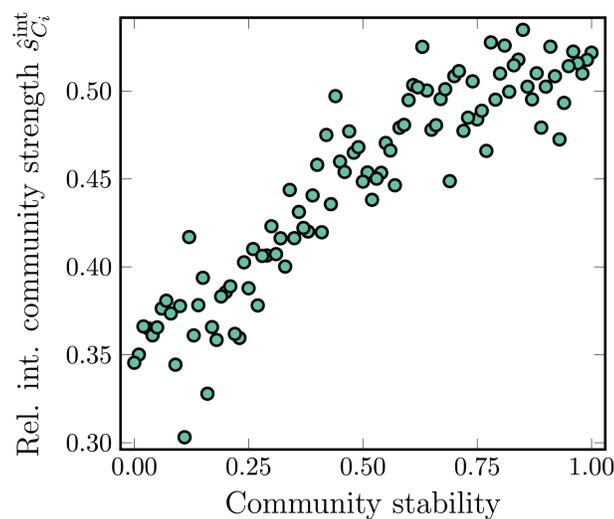
Fig 5. The mean internal node strength $\hat{s}_{C_i}^{\text{int}}$ versus community stability S_{C_i} at an effective community size of ~0.88Mb across all chromosomes. The mean internal node strength is binned by the stability. We observe a nearly linear relationship between the two quantities, indicating that the internal strength within a community is a good predictor of community stability between bootstrapped samples.

Fig 6. Community stability versus folds of enrichment (FE) for several chromatin groups (colored markers, defined in Appendix C in S1 File, 4). The number inside each panel shows the Spearman correlation coefficient. We note a convincing positive correlation with enrichment in the three leftmost panels (promoters, enhancers, and transcribed regions). We note negative correlations (one weak and one less so) for the two rightmost panels (heterochromatin & repressive states, insulators).

clear positive relation between stability and FE. However, repressed regions (heterochromatin & repressive states) show no such correlation, whilst insulator regions show a negative correlation to stability. This indicates that the active genomic regions are associated with strong stability towards noise or communities having strong internal connectivity, and that insulators repress the internal connectivity in communities to lessen their stability.

3.5. Quantifying cross-scale nestedness of stable communities

Thus far, we have focused on the stability of communities of a particular size. However, by tuning the GenLouvain parameter γ we can get communities of varying sizes. This allows us to study how communities at one scale nest with communities at larger scales. Such studies complement previous work finding that active chromatin congregates in communities with a significant hierarchical nestedness across community sizes $\hat{s}$ [15], in particular, hierarchical TADs [46,47]. Inspired by this, we ask: do nodes in stable communities of one size stay within stable communities of a larger size? Or is community nesting related to stability?

To this end, we first calculated the community stability S_{C_i} for different effective community sizes. We choose four commonly studied chromosome scales: TADs ($\hat{s} \sim 0.38$MB), A/B segments ($\hat{s} \sim 1.04$MB), $A_{1,2}$, $B_{1,2,3}$ structures ($\hat{s} \sim 49$MB) and A/B compartments ($\hat{s} \sim 60$MB) (see Appendix C in S1 File, 4). Next, we assign node stability s_i to each node i, which we choose as the community stability S_{C_j} of the community j the node resides in. Thus, as γ varies, we obtain s_i values for each size $\hat{s}$. The change in s_i reflects how stability changes over different network scales.

We show the change in s_i over three different scale transitions as a boxplot in Fig 7(a) (green boxes). For relatively small communities (around the size of TADs), we note that the stability difference is smaller in smaller communities. To make a fair comparison, we calculated the same difference between two randomly chosen nodes (orange boxes). These boxes show that the stability difference is smaller in the actual data (green) than in the random case, at least for smaller communities. This indicates that stable communities at one size are generally supported by stable communities at a larger size, indicating the cross-scale nestedness of stable communities.

However, this relationship holds best as communities become smaller. At large sizes, the communities occupy a considerable part of the chromosome, implying that randomization does not change much since the probability of selecting two random nodes from the same
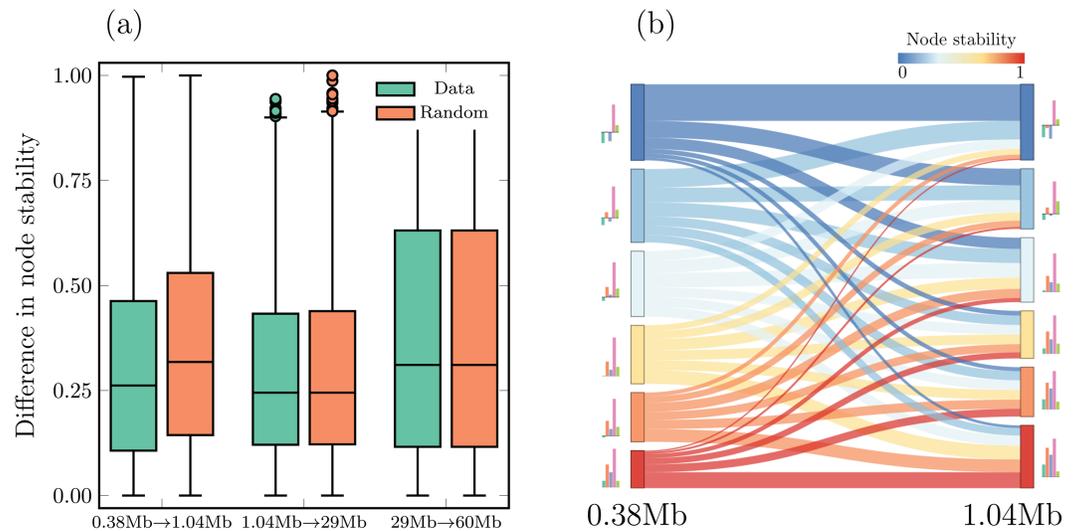
Fig 7. Connection between stability, nestedness, and active genomic regions. (a) Difference in node stability between different network scales (community sizes). Each group in the boxplots shows the distribution of absolute differences in node stability s_i between two different scales. The left boxes (green) show the difference from the actual data, and the right (orange) boxes show s_i after randomization. The x-labels indicate the effective community sizes, and the vertical lines show the median. For smaller communities, we note smaller s_i in the actual data compared to the random case. This indicates that the stability is hierarchical, where a stable node in a smaller community belongs to a similarly stable community with a larger size (or scale). However, this trend does not seem to hold for larger communities. **(b)** Cross-scale stability and chromatin state. The alluvial diagram shows the individual node stabilities between community sizes of 0.38 Mb and 1.04 Mb, rounded with a precision of 0.2. We see that most of the flows from left to right go from between similar stabilities (e.g., most of the dark red at 0.38 Mb remains dark red at 1.04 Mb or slightly lighter). The bars on the sides show the five coarse-grained chromatin states (same colors as in Fig 6, indicated by colors). We see that correlated HMM states stay correlated between different community sizes, where groups 4 and 5 are notable exceptions.

community is high. This does not necessarily imply that the stability is not nested at larger sizes but rather that a much larger partition sample is required to get good statistics.

To better illustrate how nested stability changes between two select chromosome scales ($\hat{s} = 0.38$ Mb $\rightarrow \hat{s} = 1.04$ Mb), we plotted s_i in an alluvial diagram (Fig 7(b)). To increase readability, we rounded the s_i values to a precision of 0.2. We note that most flows from $\hat{s} = 0.38$Mb to $\hat{s} = 1.04$Mb are between nodes with similar stabilities, or between stabilities one step above/below. This gives additional support to our conclusion that stability is nested across network scales. We also depict the chromatin state enrichment FE as bars to the left and right of the alluvial diagram (same data as in Fig 6). We note that the FE levels for all of the chromatin types remain conserved across both scales for all five HMM groupings. This indicates that similar correlations as in Fig 6 are observed for different structural sizes.

4. Discussion

Hi-C represents a promising method to uncover essential relationships between gene activity and chromosome 3D organization. However, extracting biologically meaningful structures from Hi-C data sets poses several challenges. One challenge concerns the 3D contacts forming a nearly fully connected complex network with cross-scale interactions. This leads to inconsistent network partitions when comparing data clustering or TAD-finding methods [13,18,20]. Another challenge comes from experimental noise, where the same clustering method yields slightly different network partitions between theoretically identical replicate

Hi-C experiments. Here, we propose a new method that systematically helps deal with the noise problem. We also propose a new null model for the distance-dependent average contact frequency that builds on Hi-C maps' multifractal properties. Using our method, we find that stable 3D communities typically have a high internal connectedness, tend to be enriched in active chromatin marks, and have a more nested cross-scale hierarchy than less stable ones.

As mentioned previously and discussed before [15], genome organization does not likely follow a perfect hierarchical tree structure. This is partly due to scale-dependent folding mechanisms, such as loop extrusion and chromatin-chromatin interactions, which are critical in forming TADs and A/B compartments [48]. Our work focuses on extracting stable communities under noise and suggests new folding mechanisms that are worthy of further experimental inquiry. We base this conclusion on Fig 7(b), where highly stable communities maintain chromatin state enrichment across chromosome scales. This foreshadows the same mechanisms being responsible for hierarchical folding in select parts of the 3D hierarchy.

We anticipate that our pipeline can handle Hi-C datasets across a range of species. However, it relies on three steps that must all be satisfied: (1) the HDM null-model must fit the data, (2) the distribution of contact counts at a fixed distance must follow a log-normal distribution, and (3) the TADs must be easily extractable. While these conditions are fulfilled for the human Hi-C data studied here, they may not hold for datasets associated with different species, like in [49]. However, we emphasize that all these steps are modifiable to fit the data in question, such as using a different null-model or additional data to fine-tune the noise parameter.

We acknowledge that our results stem from a specific Hi-C data resolution (100 kb). However, our method is not limited to this choice and can handle any resolution and Hi-C data set. In fact, we do not anticipate drastic changes to our findings because previous research finds persistent correlations between TAD boundaries and structure in the 40–100 kb range [16]. When applying the approach in this paper to a different Hi-C resolution, it is essential to double-check the noise amplitude ϵ^*, as it is not a universal parameter.

Also, our comparison to chromatin states originates from a Hidden Markov Model approach (refs [32,33]), which is just one of many chromatin annotations [50]. We selected this dataset because of its extensive use in other studies. However, since our correlations align with current understanding of how correlated noise drives structural conformations [51], we believe our results are not overly reliant on a specific dataset but instead reflect the folding of various chromatin types.

We choose the GenLouvin algorithm to find communities. While it has known shortcomings [20,52], it is one of the most widely used methods in network science because it is relatively simple and allows specifying a theoretical null model. Here, we use a null model that takes advantage of the multifractal spectrum of the Hi-C map. We believe this approach is useful for community detection in any network setting representing a complex system because it provides a straightforward way of parameterizing an explicit null model that otherwise, in lack of better options, often becomes the random Newman-Girvan model. While there exist methods to construct multifractal networks (e.g., [53]), our approach addresses the inverse problem.

Furthermore, ϵ^* also depends on the specific choice of community detection method as we calibrate against the communities associated with the unperturbed Hi-C data set, which has to be fine-tuned again if we chose another TAD-finding method, like in [18]. Our pipeline also allows experimenters to assess how stable 3D communities are under noise from varying sampling depths [19]. This case likely requires a revised model of the noise (Sect 2.2) and subsequent recalibration of ϵ^*. Otherwise, the pipeline is the same.

Conclusion

Our work expands the techniques employed for community detection in Hi-C data and introduces additional methods and metrics to extract communities that survive realistic noise levels. We anticipate having better access to robust 3D communities will help future research uncover causal connections between chromosome contact networks and genetic processes, such as gene expression and repression.

Supporting information

S1 File. Appendix covering details of data extraction and processing. The appendix contains four sections that covers (A) evaluating the HDM null model, (B) matching structure sizes to γ values, (C) preprocessing chromatin states to Hi-C bins and (D) evaluating the results with different J_{cutoff} and overlap metrics.
(PDF)

Acknowledgments

Author contributions

Conceptualization: Lucas Hedström, Ludvig Lizana.

Data curation: Lucas Hedström, Antón Carcedo, Ludvig Lizana.

Formal analysis: Lucas Hedström, Antón Carcedo, Ludvig Lizana.

Funding acquisition: Ludvig Lizana.

Investigation: Lucas Hedström, Antón Carcedo, Ludvig Lizana.

Methodology: Lucas Hedström, Antón Carcedo, Ludvig Lizana.

Project administration: Lucas Hedström, Ludvig Lizana.

Resources: Ludvig Lizana.

Software: Lucas Hedström, Antón Carcedo.

Supervision: Ludvig Lizana.

Validation: Lucas Hedström, Antón Carcedo, Ludvig Lizana.

Visualization: Lucas Hedström, Ludvig Lizana.

Writing – original draft: Lucas Hedström.

Writing – review & editing: Lucas Hedström, Antón Carcedo, Ludvig Lizana.

References

1. Dixon JR, Gorkin DU, Ren B. Chromatin domains: the unit of chromosome organization. Mol Cell. 2016;62(5):668–80. https://doi.org/10.1016/j.molcel.2016.05.018 PMID: 27259200

2. Schwartz YB, Cavalli G. Three-dimensional genome organization and function in Drosophila. Genetics. 2017;205(1):5–24. https://doi.org/10.1534/genetics.115.185132 PMID: 28049701

3. Bonev B, Cavalli G. Organization and function of the 3D genome. Nat Rev Genet. 2016;17(11):661–78. https://doi.org/10.1038/nrg.2016.112 PMID: 27739532

4. Denker A, de Laat W. The second decade of 3C technologies: detailed insights into nuclear organization. Genes Dev. 2016;30(12):1357–82. https://doi.org/10.1101/gad.281964.116 PMID: 27340173

5. Marchal C, Sima J, Gilbert DM. Control of DNA replication timing in the 3D genome. Nat Rev Mol Cell Biol. 2019;20(12):721–37. https://doi.org/10.1038/s41580-019-0162-y PMID: 31477886

6. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326(5950):289–93. https://doi.org/10.1126/science.1181369 PMID: 19815776

7. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, et al. Three-dimensional folding and functional organization principles of the Drosophila genome. Cell. 2012;148(3):458–72. https://doi.org/10.1016/j.cell.2012.01.010 PMID: 22265598

8. Dekker J, Marti-Renom MA, Mirny LA. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat Rev Genet. 2013;14(6):390–403. https://doi.org/10.1038/nrg3454 PMID: 23657480

9. MacKay K, Kusalik A. Computational methods for predicting 3D genomic organization from high-resolution chromosome conformation capture data. Brief Funct Genomics. 2020;19(4):292–308. https://doi.org/10.1093/bfgp/elaa004 PMID: 32353112

10. Liu Y, Nanni L, Sungalee S, Zufferey M, Tavernari D, Mina M, et al. Systematic inference and comparison of multi-scale chromatin sub-compartments connects spatial organization to cell phenotypes. Nat Commun. 2021;12(1):1–11.

11. Weinreb C, Raphael BJ. Identification of hierarchical chromatin domains. Bioinformatics. 2016;32(11):1601–9. https://doi.org/10.1093/bioinformatics/btv485 PMID: 26315910

12. Fraser J, Ferrai C, Chiariello AM, Schueler M, Rito T, Laudanno G, et al. Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. Mol Syst Biol. 2015;11(12):852. https://doi.org/10.15252/msb.20156492 PMID: 26700852

13. Cabreros I, Abbe E, Tsirigos A. Detecting community structures in hi-c genomic data. In: 2016 Annual Conference on Information Science and Systems (CISS). 2016. p. 584–9.

14. Lee SH, Kim Y, Lee S, Durang X, Stenberg P, Jeon J-H, et al. Mapping the spectrum of 3D communities in human chromosome conformation capture data. Sci Rep. 2019;9(1):6859. https://doi.org/10.1038/s41598-019-42212-y PMID: 31048738

15. Bernenko D, Lee SH, Stenberg P, Lizana L. Mapping the semi-nested community structure of 3D chromosome contact networks. bioRxiv. 2022.

16. Sauerwald N, Singhal A, Kingsford C. Analysis of the structural variability of topologically associated domains as revealed by Hi-C. NAR Genom Bioinform. 2020;2(1):lqz008. https://doi.org/10.1093/nargab/lqz008 PMID: 31687663

17. de Wit E. TADs as the caller calls them. J Mol Biol. 2020;432(3):638–42. https://doi.org/10.1016/j.jmb.2019.09.026 PMID: 31654669

18. Sefer E. A comparison of topologically associating domain callers over mammals at high resolution. BMC Bioinformatics. 2022;23(1):127. https://doi.org/10.1186/s12859-022-04674-2 PMID: 35413815

19. Yardımcı GG, Ozadam H, Sauria MEG, Ursu O, Yan K-K, Yang T, et al. Measuring the reproducibility and quality of Hi-C data. Genome Biol. 2019;20(1):57. https://doi.org/10.1186/s13059-019-1658-7 PMID: 30890172

20. Holmgren A, Bernenko D, Lizana L. Mapping robust multiscale communities in chromosome contact networks. Sci Rep. 2023;13(1):12979. https://doi.org/10.1038/s41598-023-39522-7 PMID: 37563218

21. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159(7):1665–80. https://doi.org/10.1016/j.cell.2014.11.021 PMID: 25497547

22. Eres IE, Gilad Y. A TAD Skeptic: is 3D genome topology conserved? Trends Genet. 2021;37(3):216–23. https://doi.org/10.1016/j.tig.2020.10.009 PMID: 33203573

23. Rosvall M, Bergstrom CT. Mapping change in large networks. PLoS One. 2010;5(1):e8694. https://doi.org/10.1371/journal.pone.0008694 PMID: 20111700

24. Pigolotti S, Jensen MH, Zhan Y, Tiana G. Bifractal nature of chromosome contact maps. Phys Rev Res. 2020;2(4). https://doi.org/10.1103/physrevresearch.2.043078

25. Sarnataro S, Chiariello AM, Esposito A, Prisco A, Nicodemi M. Structure of the human chromosome interaction network. PLoS One. 2017;12(11):e0188201. https://doi.org/10.1371/journal.pone.0188201 PMID: 29141034

26. Newman ME, Girvan M. Finding and evaluating community structure in networks. Phys Rev E. 2004;69(2):026113.

27. Yan K-K, Lou S, Gerstein M. MrTADFinder: a network modularity based approach to identify topologically associating domains in multiple resolutions. PLoS Comput Biol. 2017;13(7):e1005647. https://doi.org/10.1371/journal.pcbi.1005647 PMID: 28742097

28. Sanborn AL, Rao SSP, Huang S-C, Durand NC, Huntley MH, Jewett AI, et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc Natl Acad Sci U S A. 2015;112(47):E6456-65. https://doi.org/10.1073/pnas.1518552112 PMID: 26499245

29. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30(1):207–10. https://doi.org/10.1093/nar/30.1.207 PMID: 11752295

30. Knight PA, Ruiz D. A fast algorithm for matrix balancing. IMA J Numer Anal. 2012;33(3):1029–47. https://doi.org/10.1093/imanum/drs019

31. Raney BJ, Barber GP, Benet-Pagès A, Casper J, Clawson H, Cline MS, et al. The UCSC Genome Browser database: 2024 update. Nucleic Acids Res. 2024;52(D1):D1082–8.

32. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473(7345):43–9. https://doi.org/10.1038/nature09906 PMID: 21441907

33. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat Biotechnol. 2010;28(8):817–25. https://doi.org/10.1038/nbt.1662 PMID: 20657582

34. Lajoie BR, Dekker J, Kaplan N. The Hitchhiker's guide to Hi-C analysis: practical guidelines. Methods. 2015;72:65–75. https://doi.org/10.1016/j.ymeth.2014.10.031 PMID: 25448293

35. Jeub L, Bazzi M, Jutla I, Mucha P. A generalized Louvain method for community detection implemented in MATLAB. 2011. https://githubcom/GenLouvain/GenLouvain

36. Grosberg AY, Nechaev SK, Shakhnovich EI. The role of topological constraints in the kinetics of collapse of macromolecules. J Physiq. 1988;49(12):2095–100.

37. Ghosh SK, Jost D. How epigenome drives chromatin folding and dynamics, insights from efficient coarse-grained models of chromosomes. PLoS Comput Biol. 2018;14(5):e1006159. https://doi.org/10.1371/journal.pcbi.1006159 PMID: 29813054

38. Benzi R, Paladin G, Parisi G, Vulpiani A. On the multifractal nature of fully developed turbulence and chaotic systems. J Phys A: Math Gen. 1984;17(18):3521.

39. Opsahl T, Agneessens F, Skvoretz J. Node centrality in weighted networks: generalizing degree and shortest paths. Social Networks. 2010;32(3):245–51. https://doi.org/10.1016/j.socnet.2010.03.006

40. Nyberg M, Ambjörnsson T, Stenberg P, Lizana L. Modeling protein target search in human chromosomes. Phys Rev Res. 2021;3(1):013055.

41. Hedström L, Lizana L. Modelling chromosome-wide target search. New J Phys. 2023;25(3):033024.

42. Noh JD, Rieger H. Random walks on complex networks. Phys Rev Lett. 2004;92(11):118701. https://doi.org/10.1103/PhysRevLett.92.118701 PMID: 15089179

43. Tejedor V, Bénichou O, Voituriez R. Global mean first-passage times of random walks on complex networks. Phys Rev E Stat Nonlin Soft Matter Phys. 2009;80(6 Pt 2):065104. https://doi.org/10.1103/PhysRevE.80.065104 PMID: 20365216

44. Gong Y, Lazaris C, Sakellaropoulos T, Lozano A, Kambadur P, Ntziachristos P, et al. Stratification of TAD boundaries reveals preferential insulation of super-enhancers by strong boundaries. Nat Commun. 2018;9(1):542. https://doi.org/10.1038/s41467-018-03017-1 PMID: 29416042

45. Qu J, Yi G, Zhou H. p63 cooperates with CTCF to modulate chromatin architecture in skin keratinocytes. Epigenetics Chromatin. 2019;12(1):31. https://doi.org/10.1186/s13072-019-0280-y PMID: 31164150

46. Liu E, Lyu H, Peng Q, Liu Y, Wang T, Han J. TADfit is a multivariate linear regression model for profiling hierarchical chromatin domains on replicate Hi-C data. Commun Biol. 2022;5(1):608. https://doi.org/10.1038/s42003-022-03546-y PMID: 35725901

47. Cresswell KG, Stansfield JC, Dozmorov MG. SpectralTAD: an R package for defining a hierarchy of topologically associated domains using spectral clustering. BMC Bioinformatics. 2020;21(1):319. https://doi.org/10.1186/s12859-020-03652-w PMID: 32689928

48. Nuebler J, Fudenberg G, Imakaev M, Abdennur N, Mirny LA. Chromatin organization by an interplay of loop extrusion and compartmental segregation. Proc Natl Acad Sci U S A. 2018;115(29):E6697–706. https://doi.org/10.1073/pnas.1717730115 PMID: 29967174

49. Hoencamp C, Dudchenko O, Elbatsh AMO, Brahmachari S, Raaijmakers JA, van Schaik T, et al. 3D genomics across the tree of life reveals condensin II as a determinant of architecture type. Science. 2021;372(6545):984–9. https://doi.org/10.1126/science.abe2218 PMID: 34045355

50. Baker M. Making sense of chromatin states. Nat Methods. 2011;8(9):717–22. https://doi.org/10.1038/nmeth.1673 PMID: 21878916

51. Goychuk A, Kannan D, Chakraborty AK, Kardar M. Polymer folding through active processes recreates features of genome organization. Proc Natl Acad Sci U S A. 2023;120(20):e2221726120. https://doi.org/10.1073/pnas.2221726120 PMID: 37155885

52. Fortunato S, Hric D. Community detection in networks: a user guide. Phys Rep. 2016;659:1–44.

53. Palla G, Lovász L, Vicsek T. Multifractal network generator. Proc Natl Acad Sci. 2010;107(17):7640–5.