scientific reports

Check for updates

OPEN

Mapping robust multiscale communities in chromosome contact networks

Anton Holmgren , Dolores Bernenko & Ludvig Lizana

To better understand DNA's 3D folding in cell nuclei, researchers developed chromosome capture methods such as Hi-C that measure the contact frequencies between all DNA segment pairs across the genome. As Hi-C data sets often are massive, it is common to use bioinformatics methods to group DNA segments into 3D regions with correlated contact patterns, such as Topologically associated domains and A/B compartments. Recently, another research direction emerged that treats the Hi-C data as a network of 3D contacts. In this representation, one can use community detection algorithms from complex network theory that group nodes into tightly connected mesoscale communities. However, because Hi-C networks are so densely connected, several node partitions may represent feasible solutions to the community detection problem but are indistinguishable unless including other data. Because this limitation is a fundamental property of the network, this problem persists regardless of the community-finding or data-clustering method. To help remedy this problem, we developed a method that charts the solution landscape of network partitions in Hi-C data from human cells. Our approach allows us to scan seamlessly through the scales of the network and determine regimes where we can expect reliable community structures. We find that some scales are more robust than others and that strong clusters may differ significantly. Our work highlights that finding a robust community structure hinges on thoughtful algorithm design or method cross-evaluation.

Mammalian genomes fold into a network of 3D structures that facilitate and regulate genetic processes such as transcription, DNA repair, and epigenetics[1–4]. Most recent discoveries linking genetic processes and genomes' 3D organization derive from chromosome capture methods, such as Hi-C. Hi-C measures the number of contacts between DNA segment pairs and allows researchers to chart chromosome-wide 3D interaction maps[5–7]. These maps depict chromosomes as having 3D structures on a broad range of scales: megabase-scale A/B compartments[5], sub-compartments (A1, A2, B1,…, B4)[8], sub-megabase-scale Topologically Associated Domains (TADs)[9], sub-TADs and short-ranged loops[8]. Some of these structures are associated with epigenetic marks, active genes, and chromatin remodelers, such as CCCTC-binding factors (CTCF), cohesin complexes, and CP190[9–12].

Numerous research groups developed methods rooted in bioinformatics to detect significant 3D structures, foremost TADs and A/B compartments[13–15]. However, recently, there has been an emerging research direction alongside this development that takes advantage of the methods developed in complex network theory. This approach treats the Hi-C data as a weighted network of 3D contacts and groups nodes with above-average connectivity into mesoscale communities[16–19]. While these and many other community detection methods led to several impactful insights, underneath this approach reside an often overlooked fundamental limitation: in most networks, more than one node partition may represent a feasible network community division. Because this limitation is fundamental to the network, this type of degeneracy exists regardless of the community-finding method. Also, the degeneracy becomes increasingly problematic if trying to detect small-scale communities, where there is a significant risk of over-fitting, or in dense networks, where it is hard to determine node-community memberships with significant certainty[20].

This degeneracy problem posits that Hi-C maps' community structure is particularly challenging because Hi-C networks are almost fully connected even if most links are weak. Therefore, we expect that these networks possess several community divisions that cannot be further rated without including new data, e.g., gene expression or epigenetic profiles. Yet more intriguing, this limitation hints that there is a noteworthy probability that community-finding or data-clustering algorithms disagree on the optimal division. This problem likely fueled some debates regarding actual differences between TADs and sub-TADs[1,21].

Integrated Science Lab, Department of Physics, Umeå University, Umeå, Sweden. ✉email: ludvig.lizana@umu.se

This paper explores these limitations by mapping out the landscape of possible network partitions in Hi-C data. To this end, we use the Generalized Louvain Method[22,23] that allows us to detect communities at different network scales. We also developed a method to determine regimes where the solution landscape is degenerate and where we find robust communities.

Results

To study the multiscale 3D organization in chromosomes, we use Hi-C data from the B-lymphoblastoid human cell line (see Section "Assembling chromosome contact data" for references and data handling). As in other approaches[16–18], we convert the Hi-C data into a network, where nodes represent 10^5 base pair long DNA segments (100 kb), and the links stand for segment-segment 3D interactions, where the weights are associated with the Hi-C contact count. In this study, we focus on chromosome 10.

To partition the network and map out multiscale communities, we use the Generalized Louvain method (GenLouvain). GenLouvain separates the network into communities where nodes share more interconnections than some null model (we defer details to Section "Multiscale community detection"). To construct a realistic null model, we assume that the segment-segment contact frequencies decay as a power-law $l^{-\alpha}$, with linear separation l and decay exponent α. This scaling feature appears in established polymer physics models[24] and in Hi-C data[25]. Averaging the Hi-C contacts over many segments gives two regimes: $\alpha \approx 1.08$ for long distances ($\sim$ 500–7000 kb)[5,8], and $\alpha \approx 0.75$ for short distances ($\sim$ 200–1200 kb)[26]. See Eq. (3) in Section "Multiscale community detection" for how we implement this contact scaling in GenLouvain.

Besides the exponent α, GenLouvain has a scale parameter γ. By varying this parameter, users may scan the network hierarchies and find multiscale communities. Using this approach, we sample feasible partitions of the network. We call the collection of these partitions the solution landscape.

Classifying the solution landscape. GenLouvain optimises the modularity quality function Q (Eq. 3) to find mesoscale communities with above-average connectivity. Because the community division problem is NP-hard, it is practically impossible to enumerate all network divisions and determine which one is optimal. Instead, GenLouvain finds feasible divisions using a stochastic search algorithm[27]. But as with most community detection algorithms, GenLouvain sometimes gets trapped in local quality maxima. We illustrate this trapping schematically in Fig. 1 that shows two well-separated local maxima, ■ and ▲, overlaid in a quality contour plot. Depending on starting conditions, GenLouvain will gravitate to ▲ or drift towards ■. To increase the chance of finding the best-quality partition, we run 1,000 independent optimisation passes using different random seeds and compare the Q values.

But for some networks, the solution landscape does not split into two distinct peaks as in Fig. 1. For example, the quality may be nearly identical even in distant parts of this landscape. This means that it is challenging to distinguish the optimal partition since they are degenerate. To detect such degeneracies, we calculate the distance between partitions P and P' using the weighted mean Jaccard distance

$$d_{PP'} = \sum_i \min_j \left(1 - \frac{C_i^P \cap C_j^{P'}}{C_i^P \cup C_j^{P'}} \right) \frac{|C_i^P|}{\sum_k |C_k^P|}, \tag{1}$$

where C_i^P are the nodes in community i in P[28]. Because the distances $d_{PP'}$ are not symmetric ($d_{PP'} \neq d_{P'P}$), we use the average:

$$d = \frac{d_{PP'} + d_{P'P}}{2}. \tag{2}$$

When $d = 0$, the partitions are identical. And if $d = 1$, they are completely dissimilar. We acknowledge that there are other thinkable distance metrics, such as variation of information, but using such metrics will not change the solution landscape's qualitative topology[20].
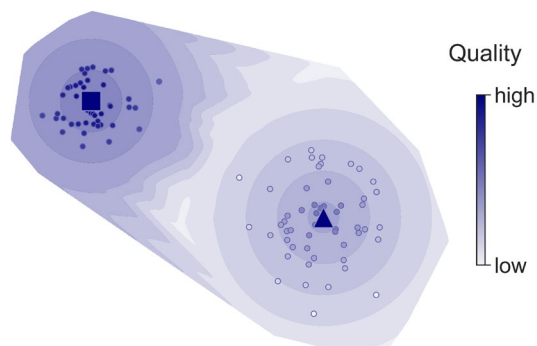

Figure 1. Solution landscape of network partitions (circles) on a quality contour plot. The (locally) best-quality partitions appear on the landscape's peaks (■ and ▲). The square partition has the highest quality.

Next, we classify the solution landscape using the Jaccard distances d and the partition qualities Q. We find three broad landscape categories depending on the variability of d and Q, $\mathrm{Var}(d)$ and $\mathrm{Var}(Q)$. First, if both $\mathrm{Var}(d)$ and $\mathrm{Var}(Q)$ are low, we find structurally similar partitions of almost the same quality. Second, we find dissimilar partitions of different qualities when both are high. For partitions in the third category (arguably the most interesting case), where $\mathrm{Var}(d)$ is high and $\mathrm{Var}(Q)$ is low, we may find dissimilar partitions having similar quality where no partition should be preferred over any other. In our notation, this case represents a degenerate solution landscape. The fourth regime (low $\mathrm{Var}(d)$ and high $\mathrm{Var}(Q)$) is unsound as we find similar partitions with relatively large quality differences. This means that as long as we find similar partitions, there is no need to study the variability in Q to guarantee that GenLouvain found the global quality maximum.

Identifying robust core communities.

We identified three solution landscapes in the previous section using the variabilities among the partitions' quality and pairwise distances. However, this only provides a qualitative assessment of the landscape's overall characteristics. Even when there are distinct peaks, there are always some deviations close to these peaks, where node assignments may differ. To quantify these differences, we tessellate the solution landscape by clustering the partitions and determining robust node-community assignments in each cluster.

We start by grouping similar partitions into clusters and comparing their sizes and qualities. The partition with the locally highest quality represents the cluster centre. To cluster similar partitions relative to the cluster centre ($d < d_{\mathrm{max}}$), we use a clustering algorithm[28], modified to maximise in-cluster similarity. Below, we summarise the main steps:

1. Order all partitions by their quality Q and let the best partition form a cluster centre (Fig. 2a).
2. Create new cluster centres with any partitions that are separated by at least d_{max} from any already present cluster centres (Fig. 2b).
3. Assign the remaining partitions to the closest cluster centre (Fig. 2c).

In this procedure, the critical parameter is the distance threshold d_{max}. This value balances the cluster size and partition similarity with the rest of the cluster. In this analysis, we use $d_{\mathrm{max}} = 0.10$, implying that the best-matching communities' weighted average fraction of shared nodes is at least 90 percent.

Next, after finding the cluster centres, we study if some network communities are more robust than others. We want to know if specific nodes co-appear in the same community in most partitions within a cluster while other nodes tend to change community memberships. To do this, we first select clusters in the solution landscape with at least 100 partitions (Fig. 3a–b). Then, we search for the largest node subset C_i' of each community C_i in P that is clustered together in at least a fraction p of the other co-clustered partitions[30]. We call these subsets core
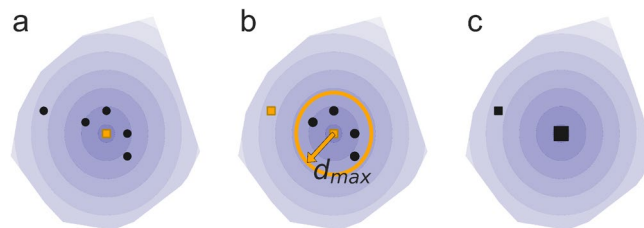

Figure 2. Partition clusters in the solution landscape. (**a**) Partitions with different quality and distance to the best quality partition (■). (**b**) The first partition separated by at least d_{max} from any cluster centre forms a new centre. This process repeats until all clusters are separated by at least d_{max}. (**c**) All partitions are assigned to the nearest cluster.
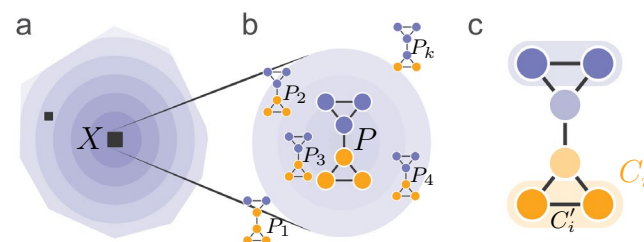

Figure 3. Identifying core communities in a cluster centre. (**a**) The best-quality cluster X (large ■) in the solution landscape. (**b**) The cluster centre P and co-clustered partitions $P_1, P_2, \ldots, P_k$ inside the cluster with possibly different community assignments. (**c**) Core communities of the best partition P are found in a fraction p of the co-clustered partitions.

communities of the cluster centre (Fig. 3c). The parameter p balances core communities' size with how many partitions in the cluster that supports them. We use $p = 0.9$ to compensate for that the partitions in the clusters are allowed to differ by 90 percent on average.

Mapping the solution landscape of human chromosome 10. In this section, we study the degeneracy of the Hi-C network for human chromosome 10, applying the results from the previous section (see Section "Assembling chromosome contact data" for data handling). Particularly, we wish to know how the solution landscape and core communities change with the parameter α associated with chromatin folding and GenLouvain's scale parameter γ that sets the typical community size (see Section "Multiscale community detection"). To make the ensuing discussion less abstract, we express γ as a characteristic community size $\hat{s}$ (number of base pairs). This change simplifies the analysis, particularly when relating our results to established chromatin divisions.

Since the community sizes are relatively heterogeneous for most γ values, we calculate $\hat{s}$ using the perplexity of the community sizes (see Eqs. 4 and 5 in Section "Characteristic community size"). We choose this metric because it is a better representation of characteristic sizes than the median or the arithmetic mean. We depict the explicit $\hat{s}$–γ relationships in Fig. S1 for $\alpha = 0.75$ and $\alpha = 1.08$.

In Fig. 4a–d, we plot the solution landscapes for four pairs of α and $\hat{s}$, each landscape spanning 1,000 GenLouvain runs. Just as in Figs. 2 and 3, we illustrate clusters as markers on top of Q contour plots made using DensMAP[29]. Each marker's diameter is proportional to the size of the cluster, and the colour represents the cluster's quality.

The panels (a–d) illustrate typical landscape behaviours. For example, (a) highlights a case where it is hard to find the optimal partition and distinguish the best community division because all partitions have nearly identical qualities but have dissimilar community structures. This leads to numerous size-one cluster centres scattered across the landscape. We characterise this case as degenerate because there is substantial variability among the cluster centres pairwise distances and low variability in quality (high Var(d) and low Var(Q)). So, in this case, we cannot be sure which cluster centre GenLouvain will gravitate towards from some random initial condition.

For larger community sizes ($\hat{s} \sim 70$ Mb), the solution landscape becomes much easier to analyse because we have only a few large clusters. For example, in (b), GenLouvain recovers the same cluster centre most of the time. Also, around (b), we find the most peaked solution landscapes where all partitions belong to a single cluster.

In panels (a) and (b), we used the looping exponent $\alpha = 1.08$, which is the genome-wide averaged contact decay in human cells for distances $\gtrsim 1$ Mb. However, $\alpha = 0.75$ fits the data better for shorter distances (0.5–1.2 Mb). With this in mind, we made similar analyses as above but for $\alpha = 0.75$ (Fig. 4c–d). This change made a noteworthy difference for the small communities [panel (c)]: the landscape has a clear cluster centre and
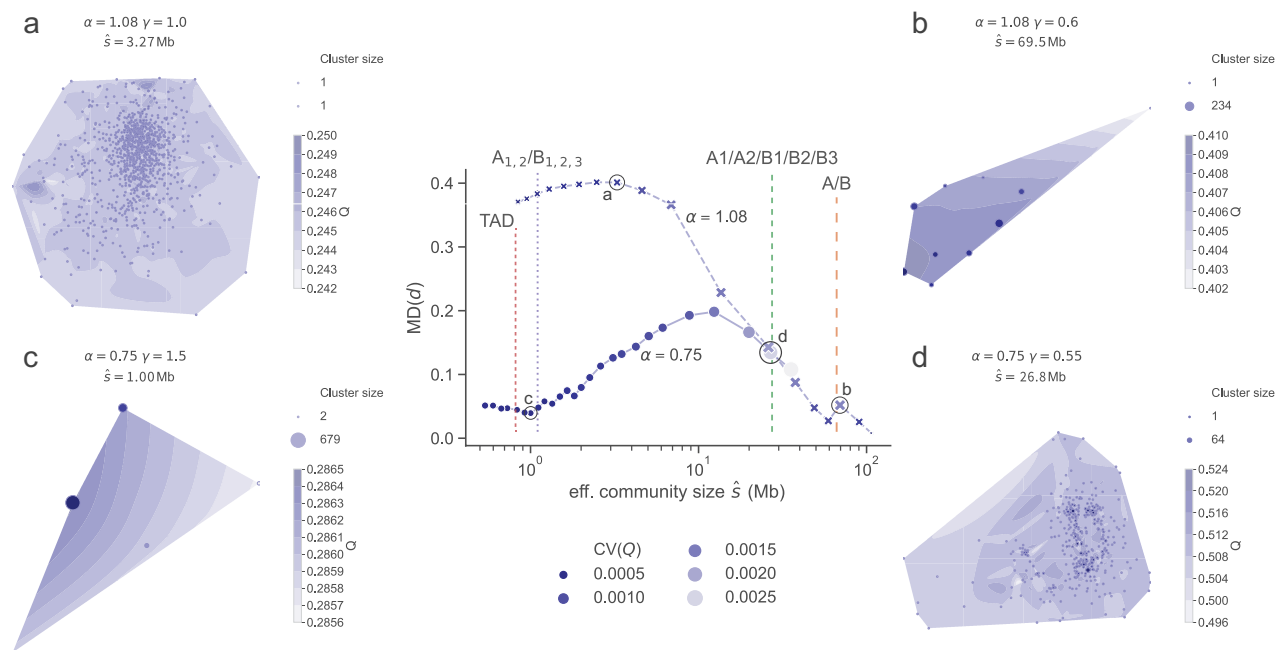

Figure 4. Solution landscapes at different scales. Mean difference (MD) of pairwise partition distances d for different α (main panel), surrounded by selected solution landscapes (**a–d**). The marker radius and colour are proportional to the quality Q's coefficient of variation. As vertical lines, we show the effective sizes of chromatin divisions (summarised in Table 1). TADs' effective size is 0.33 Mb. To fit them in the size axis, we show their effective size when omitting TADs smaller than five Hi-C-bins (~ 0.83 Mb). We visualise the solution landscapes using DensMAP[29] on a contour plot of the quality scores. In panels (**a–d**) the distance between any two points is at least d_{max}.

a reliable, optimal solution. However, forcing GenLouvain to assemble large communities with $\alpha = 0.75$ makes it increasingly degenerate up to a point (d) when the solution landscape has a global maximum alongside many local maxima with slightly lower Q.

Apart from these four examples, we made a parameter sweep of community sizes $\hat{s}$ for $\alpha = 0.75$ and $\alpha = 1.08$. But instead of creating landscape plots for each parameter pair, we calculated the Jaccard distances $d_1, d_2, \ldots, d_i, \ldots$ (Eq. 2) between all partition pairs. Then we calculated the simple average $MD(d) = E[d_i]$ and the coefficient of variation $CV(Q)$ of all partition qualities $Q_1, Q_2, \ldots$. The middle panel shows how $MD(d)$ varies with $\hat{s}$ for $\alpha = 1.08$ (crosses) and $\alpha = 0.75$ (circles) where we colour-coded the markers using $CV(Q)$. This plot allows us to identify scale regimes where $MD(d)$ is large but $CV(Q)$ is small, which is a hallmark of a degenerate solution landscape. For example, the plot demonstrates that $\alpha = 1.08$ is not a suitable folding parameter to find reproducible small-scale communities in the range $\sim$ 1–4 Mb.

In the middle panel, we also indicate $\hat{s}$ of published chromatin divisions, like TADs (> 0.5 Mb) and A/B compartments (see Section "Assembling chromosome contact data"), by vertical dashed and dotted lines. The scales close to (b) (encircled) corresponds to characteristic A/B compartment sizes, $\hat{s} = 66$ Mb. Using $\alpha = 1.08$, this scale is associated with a non-degenerate landscape leading to a reliable partition of the Hi-C network. But interestingly, we note that there seems to be an even better division at a slightly smaller $\hat{s}$. This panel also shows that we must use $\alpha = 0.75$ to find reliable partitions with sizes similar to TADs $\hat{s} = 0.33$ Mb. Finally, sandwiched between A/B compartments and TADs, there is yet another commonly used Hi-C division denoted A1, A2, and B1,…, B3. This regime has less reliable communities because the landscape is flatter (exemplified in d).

Robust communities of chromosome 10. After classifying the solution landscape in Fig. 4, we analyzed how robust the partitions are by identifying the core communities across $\hat{s}$. As illustrated in Fig. 3, we extract robust communities by first clustering similar partitions and then quantifying the internal cluster differences. We quantify these differences by calculating the fraction of identical node-community memberships. We omit clusters with less than ten percent of the total partition ensemble for a given $\hat{s}$–α combination (100 out of a 1,000 partitions). We find robust communities when large clusters have a high fraction of nodes assigned to core communities (note marker sizes in Fig. 5). This finding holds for both folding parameters, $\alpha = 0.75$ and $\alpha = 1.08$. Conversely, we find a fuzzy community structure when small clusters have the same relative quality Q/Q_{max} and a small fraction of core-assigned nodes.

For $\alpha = 0.75$, we observe that the most robust scale is $\hat{s} \sim 10^0$ Mb. Here, one dominating cluster contains more than half of all partitions in which the communities contain nodes interacting primarily over short distances. These communities are mostly unbroken DNA sequences (Fig. S3a) similar to TADs. But there are exceptions. For instance, we find a few large communities that join nodes from linearly separated DNA segments. We illustrate the complete scale-dependent node-community memberships in Fig. S3a. This figure shows how the nodes redistribute between communities when $\hat{s}$ changes. Apart from observing stable communities (e.g, the beginning of the chromosome), we note that the 3D folding is not perfectly hierarchical, in which smaller communities form larger and larger super-structures. Albeit small, there are deviations that make the folding structure semi-nested[18].

For $\alpha = 1.08$, we detect more than 80 percent core nodes when $\hat{s} > 40$ Mb and the most robust scale for $\hat{s} \sim 100$ Mb. But this scale is trivially robust as most nodes are in a giant community (Fig. S3b). A more interesting case is where $\hat{s} \sim 60$ Mb and $\hat{s} \sim 90$ Mb, with the former having a slightly larger fraction of core node assignments. While $\hat{s} \sim 60$ Mb is similar to typical sizes of A/B compartments (Fig. 4), we find multiple clusters when $\hat{s} \sim 70$ Mb that have similar quality but with lower core-node fractions.
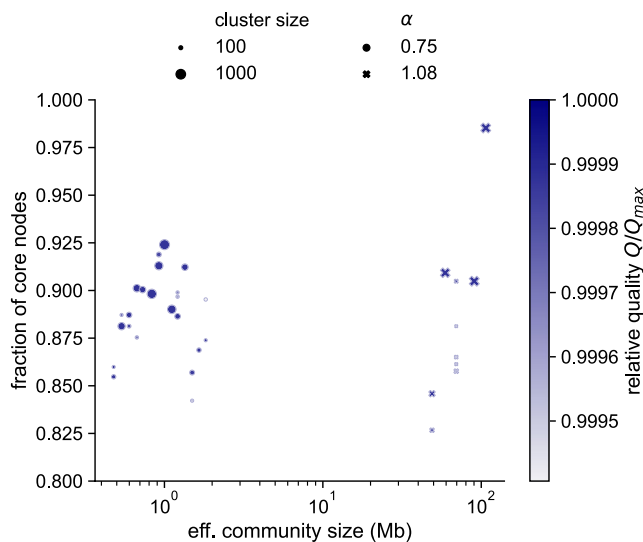

Figure 5. Cluster sizes and the fraction of core nodes. Core nodes are clustered in at least 90 percent of the cluster's partitions. We only show clusters with at least 100 partitions.

Overall, we note that GenLouvain can detect reliable core communities at two distinct network scales ($\hat{s} \sim 1$ Mb and $\hat{s} \sim 60$ Mb) depending on the value of the folding parameter α. To investigate if there are other stable network scales, we made a sweep of α values for each $\hat{s}$ and calculated the mean partition distances MD(d). As shown in the heat map Fig. 6, the most robust regimes are the top-left and bottom-right, where MD(d) is the smallest. In the bottom left corner, where $\alpha \sim 1$ and $\hat{s}$ are small, we find the most degenerate solution landscape.

Finally, we investigated the local robustness of DNA regions after community divisions. This analysis aims to identify nodes with high variability in their community membership. We refer to these nodes as "fringe nodes" as they typically lie at the interface between multiple communities. Positioned in this way, fringe nodes may belong to several communities in a partition ensemble without causing too much difference in the overall modularity. We speculate that DNA regions associated with fringe nodes have multiple functions or are correlated with different DNA-binding proteins.

To identify fringe nodes, we first select the best cluster from each α–γ pair. Next, using the core communities derived before, we count how many of the cluster's partitions a designated node co-clusters with the core community. We call this quantity c. Based on c, we calculate complement $1 - c$, which signifies the difficulty in assigning a stable community to a designated node.

We depict the $1 - c$ genomic profile in Fig. 7, where each row corresponds to varying γ values. Notably, we observe extended stretches where $1 - c \approx 0$, particularly for small γ values, indicating stable community assignments. Between these regions, we note clusters associated with high fringe assignments. These DNA regions, or nodes, represent DNA loci with variable community memberships. For example, there is a cluster of fringe nodes close to the centromere (position ≈ 45 Mb), consistently appearing as a band across all α–γ pairs. Additionally, we note that the number of fringe nodes does not significantly increase with higher γ values, despite more communities. Instead, new DNA loci appear with high $1 - c$ values.

Established chromatin divisions differ from optimal network communities. In Fig. 4, we indicated typical sizes of a few established chromatin divisions, like large TADs and A/B compartments, by vertical lines. These chromatin divisions have size distributions that differ from typical network communities. To make a better comparison, we varied γ to find the network partition that is most similar to the chromatin divisions, disregarding that the effective size $\hat{s}$ may differ from $\hat{s}_{TAD}$ or $\hat{s}_{A/B}$. Then we quantified the similarity by calculating the adjusted mutual information (AMI), commonly used to compare partitions. The AMI is 1 when the two partitions are identical and 0 when inseparable from chance. We summarise the results of our AMI analysis in Table 1.

For TADs (Table 1, top row), we find the best correspondence when $\hat{s} = 0.77$ Mb, which is larger than TADs' effective size $\hat{s}_{TAD} = 0.33$ Mb. Here, the AMI score is 0.53, indicating that the community structures show significant deviations. This deviation is likely because median TAD sizes are close to the data resolution we use (0.1 Mb). The AMI score is similar for A/B compartments (AMI = 0.47), but the scales match better ($\hat{s} = 66$ Mb vs $\hat{s} = 59$ Mb). We find the best overlap with the small-scale $A_{1,2}/B_{1,2,3}$ segments (denoted "A/B segments" in Table 1) with $\hat{s} = 1.8$ Mb and AMI = 0.72. We do not compare our results with A1/A2/B1/B2/B3 sub-compartments because we cannot detect robust communities in this regime.

Finally, in Fig. 8, we visualise how the node-community membership differs between the A/B compartments and the optimal network partition at $\hat{s} = 59$ Mb. We observe that most sub-compartments are isolated into a single network community. But the A2 sub-compartment includes Hi-C bins assigned to the two largest communities.

Discussion

Hi-C networks are densely connected. Therefore, finding reliable community structures across various scales is challenging. To better understand this problem, we have mapped out the solution landscape of feasible partitions in a chromosome contact network at different organization scales. We sampled 1,000 partitions using different scale- and DNA-looping parameters to detect regimes associated with robust or degenerate solution landscapes. We classified these regimes in terms of the variabilities of the partition's qualities and pairwise distances. Then
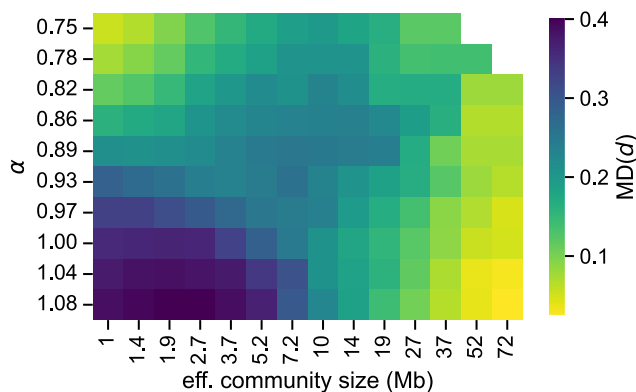

Figure 6. Mean absolute difference (MD) of pairwise partition distances d for chromosome 10 for different α.
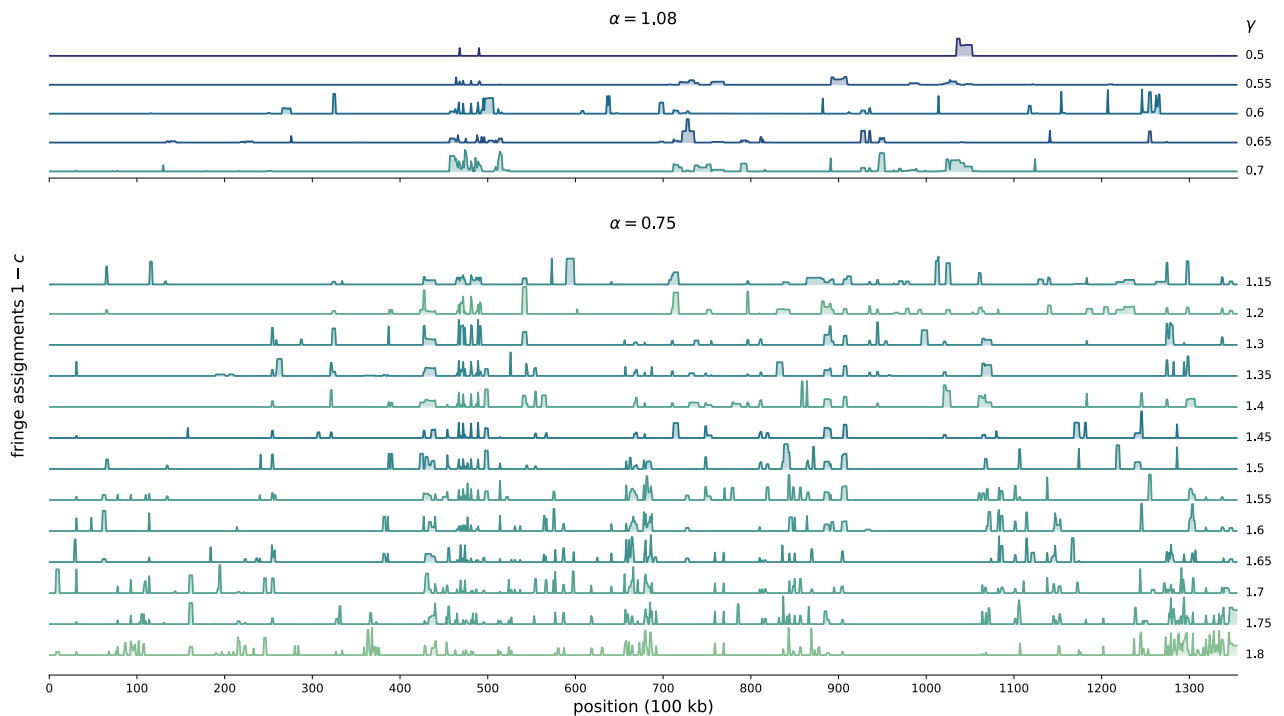
Figure 7. Fringe assignments for different decay parameters α and scale parameters γ along chromosome 10. The fringe assignment is measured as $1 - c$, where c is the fraction of partitions where a node share the same community assignments as the cluster centre. Increasing the scale parameter γ decreases the effective community size $\hat{s}$ (Fig. S1). Colours indicate the total fraction of fringe nodes, with darker blue colours having the least fringe nodes. We only include γ-values supported by clusters with at least 100 partitions.

	Characteristic size (Mb)		Most similar partition		
	Median	$\hat{s}$	α	$\hat{s}$	AMI
TADs	0.18	0.33	0.75	0.77	0.53
A/B segments	0.30	1.1	0.75	1.8	0.72
A1,...,B3	31	27	–	–	–
A/B compartments	64	66	1.08	59	0.47

Table 1. Comparing optimal network partitions with established chromatin divisions. We derived the sizes for A1,A2,B1,…,B3 by aggregating $A_{1,2}/B_{1,2,3}$ segments ("A/B segments") and the A/B sizes from merging A1/.../B3 sub-compartments (see Fig. S2). Notation: effective size $\hat{s}$, and adjusted mutual information (AMI), chromatin looping exponent α. We found no similar partition for A1,…,B3.

we used a partition clustering approach and compared cluster sizes and qualities. Also, studying the proximity of the best-quality partition, we find robust core communities supported by at least 90 percent of the proximate partitions. Finally, varying the looping parameter α We find robust small-scale communities for $\alpha = 0.75$ and larger-scale communities for $\alpha = 1.08$, roughly corresponding to TADs and A/B compartments. Between these extremes, we find a regime opaque to community detection methods.

Our results derive from 100 kb Hi-C data. However, our approach is not restricted to any specific resolution or interaction matrix. It can efficiently analyze various chromatin interaction matrices such as single-cell Hi-C (scHi-C)[31], HiCap[32], HiChIP[33], and distance matrices[34]. Nevertheless, modifications to the GenLouvain null model may be necessary for some of these scenarios.

We mapped out the multiscale solution landscape in Fig. 4 and discovered regimes where the landscape is degenerate, as illustrated in panel (a). It is critical to note this degeneracy problem is not easily resolved using another community detection method because strong communities might not exist in the data at that scale. Therefore, different methods will provide different answers. We circumvented some degenerate scales by modifying the null model's folding parameter. For example, at $\hat{s} \sim 1$ Mb, changing α from 1.08 to 0.75, GenLouvain recovers the same optimal partition most of the time. However, this approach is not straightforward to generalise.

Furthermore, we found two distinct regimes in the α–$\hat{s}$ parameter space where community detection is easy (in Fig. 6). But this finding does not exclude other robust network scales. In GenLouvain's modularity function, we assumed that node-node contacts decay as a power law with some exponent α. While this is consistent with the average contact decay in Hi-C maps and established polymer physics models (e.g., the Gaussian chain or the
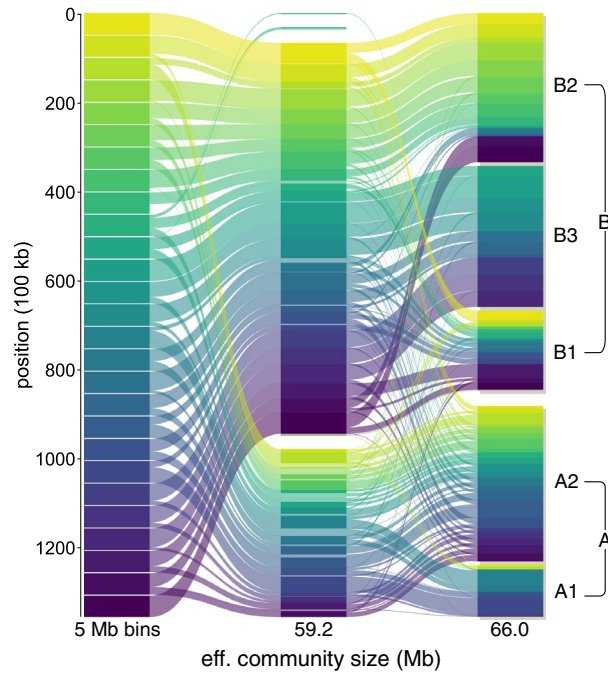
Figure 8. Core communities and A/B compartments of chromosome 10. The leftmost column represents 5 Mb bins coloured by position. The middle column represents the partition most similar to the A/B compartments ($\hat{s} = 66$ Mb), with communities ordered by average position. Transparent segments do not belong to the core. The right-most column represents A/B sub-compartments.

fractal globule), there could be other functional forms that better describe the actual folding mechanism or a blend of several competing mechanisms (e.g., short-ranged loop-extrusion and long-ranged phase separation)[35]. This amounts to improving the null model, which we leave as an avenue for future research.

We found that established chromatin divisions differ from the optimal GenLouvain partition associated with identical characteristic sizes (Table 1). Even if sweeping through a range of characteristic sizes, we still find significant differences with the most similar GenLouvain partition. We achieved the best match for $A_{1,2}/B_{1,2,3}$ segments, and the matching communities are robust. While we cannot reach perfect overlap using one single characteristic size, we point out that it is conceivable to increase the overlap if considering partitions from several $\hat{s}$. This indicates that our approach might find most chromatin divisions but not at a single $\hat{s}$. This finding helps benchmark our results to other published TAD-finding methods and offers a systematic approach to highlight deviations from expected network partitions under the null model (power law decaying contacts).

There are numerous TAD-finding methods which can be broadly categorized into feature-based algorithms, clustering methods, and graph-partitioning tools[36,37]. In our study, we employ a technique from the graph-partitioning category, which encompasses popular community-detection algorithms based on modularity maximization. For instance, one study[38] identified TADs using the Louvain method but assumed that background connectivity follows a random network under given node degrees (the Newman-Girvan model). This assumption was partly remedied in Ref.[39] that combines maximum modularity and Hi-C-like distance decay to extract communities for different γ values. However, their method considers TADs as continuous DNA stretches, unlike our approach, which treats them as delocalized entities. Our approach uses power-law decaying models where the exponents α closely align with the observed distance decay in contact probability within human Hi-C maps ($\alpha = 1.08$ and $\alpha = 0.75$). This type of decay typically identifies more spatially dispersed and delocalized 3D communities, whereas the Newman-Girvan model tends to group contiguous DNA segments into local communities, similar to TADs. Through polymer simulations[17], we have demonstrated that our generalization effectively partitions spatially proximal monomers into meaningful 3D communities. And by adding the complete scope outlined in this paper, researchers can rate the stability and robustness of these communities by identifying core regions, distinguishing ambiguous nodes, and investigating hierarchical community relationships[18,40].

Finally, while our work focuses on Hi-C contact maps, GenLouvain is commonly used to detect communities in a wide range of networks. Therefore, our work is helpful to other researchers searching for robust communities when facing the degeneracy problem.

Materials and methods

Assembling chromosome contact data.
We downloaded Hi-C data for the B-lymphoblastoid human cell line (GM12878)[8] from the GEO database (`MAPQG0` dataset, 100 kb resolution)[41]. The data file contains measured contact frequencies between DNA segment pairs in a cell population. We only consider intra-chromosome contacts in our analysis, allowing us to study each chromosome by itself. We interpret the Hi-C data as a weighted network in sparse form, where each node represents a 100 kb DNA segment, and the link weight

is the measured contact count. Before constructing the network, we normalise the data using the Knight-Ruiz matrix balancing algorithm.

In addition to Hi-C data, we use datasets associated with existing 3D divisions[8]: A/B sub-compartments and topologically associating domains (TAD) (downloaded from the GEO database[41]). The sub-compartments divide chromosomes into regions called A1, A2, B1, B2, B3, and B4. While A1 and A2 exhibit high gene expression, B1–B3 are associated with repressed and inactive DNA regions (B4 is found only in chromosome 19 and does not participate in our study as we focus on chromosome 10). Also, functionally similar sub-compartments tend to have correlated contact patterns and are generally referred to as A- and B-compartments. Alongside the sub-compartment, we study TADs. Defined by the Arrowhead algorithm[8], TADs are genomic regions with above-average contact frequencies, serving as microenvironments for co-regulated genes. TADs appear as squares along the main diagonal in Hi-C maps.

Multiscale community detection.

To find network communities, we use the Generalized Louvain method (GenLouvain)[23]. GenLouvain searches for network partitions that maximise the modularity function Q, capturing local deviations from the expected background connectivity. While the most common choice is random connections, better known as the Newman-Girvan null model[42], we rescale the expected link weights to mimic that nodes are interconnected DNA segments forming a long polymer chain that is folded in 3D inside the cell nucleus[17]. Empirical data shows that the average link weight ($\propto$ number of contacts) decays as a power-law with linear node separation. After this modification, the parametric modularity (or quality) function is[43]

$$
Q = \frac{1}{2m} \sum_{i \neq j} \left(A_{ij} - \gamma P_{ij}^{(0)} \right) \delta(C^i, C^j),
$$

$$
P_{ij}^{(0)} = \frac{2m\, k_i\, k_j\, |i-j|^{-\alpha}}{\sum_{i' \neq j'} k_{i'}\, k_{j'}\, |i'-j'|^{-\alpha}},
$$

(3)

where A_{ij} are entries in the weighted adjacency (Hi-C) matrix, m is the total weight, γ is the scale parameter, k_i is the strength of node i, and C^i is node i's community assignment. By tuning the scale parameter γ, we get a spectrum of communities of different sizes. With increasing γ, we penalise any links with weights close to the random expectation.

The decay parameter α reflects DNA's 3D folding. This parameter also changes how GenLouvain treats weak (or long-ranged) connections when assembling communities. Particularly, while decreasing α tend to disfavour weak links, working as a threshold for long-range links, increasing α favour weak links. When $\alpha = 0$, we recover the Newman-Girvan null model. Based on empirical data, we study $\alpha = 1.08$ to find large, long-range ($\sim$ 500–7000 kb) communities[5], and $\alpha = 0.75$ to find smaller, short-range ($\sim$ 200–1200 kb) communities[26]. These values reflect two DNA-folding mechanisms: the loop extrusion that forms small-scale 3D structures, and the phase separation that governs the self-aggregation of distant regions.

Finally, we set GenLouvain to randomly regroup nodes to communities proportional to the resulting quality increase. This achieves better solution landscape sampling.

Characteristic community size.

We explore the solution landscapes over varying scale and decay parameters. To compare the partitions' characteristic community sizes, we use a metric that is weakly dependent on spurious singleton communities, unlike the mean and median. Instead, we use the effective community size

$$
\hat{s} = \frac{\text{number of nodes}}{\text{effective number of communities}},
$$

(4)

where we calculate the effective number of communities using the perplexity $2^{H(P)}$ of partition P's community size distribution, with Shannon entropy

$$
H(P) = -\sum_i \frac{|C_i|}{\sum_j |C_j|} \log_2 \frac{|C_i|}{\sum_j |C_j|}.
$$

(5)

Data availability

The `MAPQG0` dataset, sub-compartment, and topologically associating domain (TAD) data[8] was downloaded from The GEO Database with accession number `GSE63525` at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63525. The source code to GenLouvain was downloaded from https://github.com/GenLouvain/GenLouvain. All other source codes are available at https://github.com/lizanalab/mapping2023holmgren.

References

1. Dixon, J. R., Gorkin, D. U. & Ren, B. Chromatin domains: The unit of chromosome organization. *Mol. Cell* **62**(5), 668–680 (2016).
2. Schwartz, Y. B. & Cavalli, G. Three-dimensional genome organization and function in drosophila. *Genetics* **205**(1), 5–24 (2017).
3. Bonev, B. & Cavalli, G. Organization and function of the 3d genome. *Nat. Rev. Genet.* **17**(11), 661–678 (2016).
4. Denker, A. & De Laat, W. The second decade of 3c technologies: Detailed insights into nuclear organization. *Genes Dev.* **30**(12), 1357–1382 (2016).

5. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**(5950), 289–293 (2009).
6. Sexton, T. *et al.* Three-dimensional folding and functional organization principles of the drosophila genome. *Cell* **148**(3), 458–472 (2012).
7. Dekker, J., Marti-Renom, M. A. & Mirny, L. A. Exploring the three-dimensional organization of genomes: Interpreting chromatin interaction data. *Nat. Rev. Genet.* **14**(6), 390–403 (2013).
8. Rao, S. S. *et al.* A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**(7), 1665–1680 (2014).
9. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**(7398), 376–380 (2012).
10. Kaushal, A. *et al.* Ctcf loss has limited effects on global genome architecture in drosophila despite critical regulatory functions. *Nat. Commun.* **12**(1), 1–16 (2021).
11. Remeseiro, S., Hörnblad, A. & Spitz, F. Gene regulation during development in the light of topologically associating domains. *Wiley Interdiscip. Rev. Dev. Biol.* **5**(2), 169–185 (2016).
12. Szabo, Q., Bantignies, F. & Cavalli, G. Principles of genome folding into topologically associating domains. *Sci. Adv.* **5**(4), eaaw1668 (2019).
13. MacKay, K. & Kusalik, A. Computational methods for predicting 3d genomic organization from high-resolution chromosome conformation capture data. *Brief. Funct. Genomics* **19**(4), 292–308 (2020).
14. Fraser, J. *et al.* Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Mol. Syst. Biol.* **11**(12), 852 (2015).
15. Liu, Y. *et al.* Systematic inference and comparison of multi-scale chromatin sub-compartments connects spatial organization to cell phenotypes. *Nat. Commun.* **12**(1), 1–11 (2021).
16. Sarnataro, S., Chiariello, A. M., Esposito, A., Prisco, A. & Nicodemi, M. Structure of the human chromosome interaction network. *PLoS One* **12**(11), e0188201 (2017).
17. Lee, S. H. *et al.* Mapping the spectrum of 3d communities in human chromosome conformation capture data. *Sci. Rep.* **9**(1), 1–7 (2019).
18. Bernenko, D., Lee, S. H., Stenberg, P. & Lizana, L. Mapping the semi-nested community structure of 3d chromosome contact networks. *bioRxiv* (2022).
19. Boulos, R. E., Arneodo, A., Jensen, P. & Audit, B. Revealing long-range interconnected hubs in human chromatin interaction data using graph theory. *Phys. Rev. Lett.* **111**(11), 118102 (2013).
20. Good, B. H., De Montjoye, Y. A. & Clauset, A. Performance of modularity maximization in practical contexts. *Phys. Rev. E* **81**(4), 046106 (2010).
21. Eres, I. E. & Gilad, Y. A tad skeptic: Is 3d genome topology conserved?. *Trends Genet.* **37**(3), 216–223 (2021).
22. Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**(10), P10008 (2008).
23. Jeub, L. G.S., Bazzi, M., Jutla, I. S. & Mucha, P. J. A generalized louvain method for community detection implemented in matlab. *https://github.com/GenLouvain/GenLouvain* (2011–2019).
24. Mirny, L. A. The fractal globule as a model of chromatin architecture in the cell. *Chromosome Res.* **19**(1), 37–51 (2011).
25. Pigolotti, S., Jensen, M. H., Zhan, Y. & Tiana, G. Bifractal nature of chromosome contact maps. *Phys. Rev. Res.* **2**(4), 043078 (2020).
26. Sanborn, A. L. *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci.* **112**(47), E6456–E6465 (2015).
27. De Meo, P., Ferrara, E., Fiumara, G. & Provetti, A. Generalized louvain method for community detection in large networks. In *2011 11th International Conference on Intelligent Systems Design and Applications* 88–93 (IEEE, 2011).
28. Calatayud, J., Bernardo-Madrid, R., Neuman, M., Rojas, A. & Rosvall, M. Exploring the solution landscape enables more reliable network community detection. *Phys. Rev. E* **100**(5), 052308 (2019).
29. Narayan, A,, Berger, B. & Cho, H. Density-preserving data visualization unveils dynamic patterns of single-cell transcriptomic variability. *bioRxiv* (2020).
30. Rosvall, M. & Bergstrom, C. T. Mapping change in large networks. *PLoS One* **5**(1), e8694 (2010).
31. Nagano, T. *et al.* Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**(7469), 59–64 (2013).
32. Zhigulev, A. & Sahlén, P. Targeted chromosome conformation capture (hicap). In *Spatial Genome Organization: Methods and Protocols* 75–94 (Springer, Berlin, 2022).
33. Mumbach, M. R. *et al.* Hichip: Efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* **13**(11), 919–922 (2016).
34. Bintu, B. *et al.* Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science* **362**(6413), eaau1783 (2018).
35. Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N. & Mirny, L. Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proc. Natl. Acad. Sci.* **115**(29), E6697–E6706 (2018).
36. Zufferey, M., Tavernari, D., Oricchio, E. & Ciriello, G. Comparison of computational methods for the identification of topologically associating domains. *Genome Biol.* **19**(1), 1–18 (2018).
37. Sefer, E. A comparison of topologically associating domain callers over mammals at high resolution. *BMC Bioinform.* **23**(1), 127 (2022).
38. Norton, H. K. *et al.* Detecting hierarchical genome folding with network modularity. *Nat. Methods* **15**(2), 119–122 (2018).
39. Yan, K.-K., Lou, S. & Gerstein, M. Mrtadfinder: A network modularity based approach to identify topologically associating domains in multiple resolutions. *PLoS Comput. Biol.* **13**(7), e1005647 (2017).
40. Bernenko, D., Lee, S. H., Stenberg, P. & Lizana, L.. Exploring 3d community inconsistency in human chromosome contact networks. arXiv preprint arXiv:2302.14684 (2023).
41. Edgar, R., Domrachev, M. & Lash, A. E. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**(1), 207–210 (2002).
42. Newman, M. E. J. & Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E* **69**(2), 026113 (2004).
43. Reichardt, J. & Bornholdt, S. Statistical mechanics of community detection. *Phys. Rev. E* **74**(1), 016110 (2006).

Acknowledgements

Author contributions

A.H. and L.L. devised the study. D.B. prepared the data. A.H. and D.B. performed the experiments and analysed the results. All authors wrote, edited, and accepted the manuscript in its final form.

Funding

Open access funding provided by Umea University.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-39522-7.

Correspondence and requests for materials should be addressed to L.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.