Long-range genomic loci stochastically assemble into combinatorial forms of chromosome skeleton

Jingyu Zhang[1,*], Siyuan Wang[2], Simon C Watkins[3], Jianhua Xing[1,4,5,*]

[1] Department of Computational and Systems Biology, University of Pittsburgh; Pittsburgh, PA 15232, USA.

[2] Department of Cell Biology, Yale School of Medicine; New Haven, CT 06510, USA.

[3] Department of Cell Biology, University of Pittsburgh; Pittsburgh, PA 15232, USA.

[4] UPMC-Hillman Cancer Center, University of Pittsburgh; Pittsburgh, PA, USA.

[5] Department of Physics and Astronomy, University of Pittsburgh; Pittsburgh, PA 15232, USA.

[*] To whom correspondence should be addressed. Email: xing1@pitt.edu, zhangjy4316@gmail.com

Abstract

One fundamental yet open question is how eukaryotic chromosomes fold into segregated territories, a process essential for gene transcription and cell fate. Through analyzing Hi-C and chromatin-tracing DNA-FISH data, we identify long-range chromo skeleton loop structures that span over 100 Mb, extending beyond the reach of many existing DNA loop models. Further spatial density analyses point to assembly formation also independent of major nuclear structures. Some long-range loops share a subset of genomic loci, which serve as nucleation centers and drive loop clustering. These complexes are highly stable, as shown by live-cell imaging with sequence-specific fluorescent labeling, and biophysical model analyses reveal a multivalent binding mechanism. Our findings suggest a redundant, distributed cluster mechanism that ensures robustness across cell types and against mutations, guiding both chromosome compaction and the formation of smaller-scale chromosomal structures.

Main Text

A central problem in structural biology and biological physics is how macromolecules fold into three-dimensional structures. An emerging "local-to-global" mechanism highlights the role of local cooperativity among peptide monomers in guiding the formation of a native protein structure from an expansive conformational space[1]. In contrast to typical polypeptides with 300 – 400 amino acids, normal human genome is composed of 48 long-chain polymers, ranging from ~ 40 to 250 Mb, with a total linear length extending up to two meters. Given that the 3D structures play a central role in regulating cell identity and gene expressions[2], an intriguing question is how these linear DNA molecules compact into segregated chromatin territories within a cell nucleus, typically only about 10 μm in diameter.

Chromatin, unlike a polypeptide, is often considered to lack a stable, ordered three-dimensional structure due to its significantly greater length. When modeled as a linear polymer chain without

structural supports such as protein bridges or long non-coding RNA networks, theoretical analyses predict that the average contact frequency (f) between two chromosomal fragments depends on their genomic separation (l) and follows a power-law decay, $f \propto l^{\alpha}$. Previous Hi-C measurements confirmed this relationship, with an exponent $\alpha \sim -1$ within the 1 – 5 Mb range (Fig. 1a, Extended Data Fig. 1a-c). This result aligns with polymer models describing chromatin as a crumpled polymer[3-5] or emphasizing the excluded volume effect[6].

However, deviations from the power-law behavior have been observed at both short and long genomic distances. At genomic distances ~100 – 200 kb, the deviations (Fig. 1a, Extended Data Fig. 1a-c) suggest the presence of stable chromatin loops, potentially formed through a loop-extrusion mechanism[7]. Beyond 5 Mb, Hi-C data reveal further deviations, where the *f-l* curves level off (Fig. 1a blue line, Extended Data Fig 1a&b). This phenomenon has been attributed to polymer models incorporating random-sized loop formation between chromatin segments[8]. Additional factors, such as A/B compartment segregation and nuclear structure tethering, also contribute to deviations from the crumpled polymer model at large genomic distances[4,9].

In sum, previous studies suggest that chromosomes form functionally important local structures such as topologically associated domains (TADs) and enhancer-promoter interactions, typically spanning within ~ 1 Mb. In contrary, in large genomes like the human genome, it is considered rare to have stable proximities between locus pairs separated by more than 10 Mb long-range genomic loci, though scattered instances of locus pairs interacting across distances exceeding 20 Mb have been reported[4,9,10]. In this study, we investigate the presence and conservation of spatial proximity between two well-separated loci on the same human chromosome across various cell lines and explore the underlying physical mechanisms that may explain such stable contacts.

Hi-C data reveals the prevalence of long-range genomic locus colocalization beyond 50 Mb

We first analyzed Hi-C data for chromosome 2 (chr2) and 14 (chr14) in MCF10A cells. While most genomic pairs followed the expected Hi-C *f-l* curve, certain pairs (e.g., red box in Fig. 1a & Extended Data Fig. 1a) showed significant deviations, with contact frequencies comparable to the median value of 1-Mb pairs despite their much larger genomic separation. Similar high-interaction locus pairs were also observed in chr2 and chr14 of IMR90 (Extended Data Fig. 1b). Extending this analysis to 13 additional human cell lines revealed that such deviations are prevalent among genomic locus pairs separated by even more than 50 Mb (Extended Data Fig. 2), suggesting their stable three-dimensional proximity in at least a subpopulation of chromatin. Notably, these long-range interactions extend beyond typical intrachromosomal contacts such as enhancer-promoter interactions, super-enhancer, and TAD[11,12]. Given the entropic cost of maintaining substantial colocalization probability over such long genomic distances, we sought to explore their underlying mechanisms and potential functional significances.

While the above statistical analyses identified long-range locus pairs with spatial proximity, detecting them quantitatively from noisy Hi-C data remains challenging, as strong signals from

short-ranged locus pairs often overshadow weaker signals from long-range ones. To address this, we developed a computational pipeline, Chromosome Long-range colocalization identifier through Outlier Detection (CLOD) (Fig. 1b, details in Methods). The pipeline operates on the premise that colocalized locus pairs increase the contact frequency of their neighboring loci along the genome. To enhance the signal-to-noise ratio, CLOD employs a sliding-window averaging approach, allowing for the distinction of colocalization signals from background noise. It then compares locus pairs with similar genomic distances and identifies outliers exhibiting abnormally high contact frequencies. Specifically, outliers are defined as locus pairs whose signals exceed two times the interquartile range above the third quantile of all pairs with similar genomic distances ($\frac{Signal-Q3}{Q3-Q1} > 2$).

To assess CLOD's performance in detecting experimentally verified colocalized locus pairs, we first analyzed an MCF7 Hi-C dataset[13]. CLOD successfully identified a 150 kb loop (Extended Data Fig. 1d), previously validated as a promoter-enhancer interaction of *PRIP1* using a 3C assay[14]. We then applied CLOD to a Hi-C dataset from the HCT116 cell line[15] and detected a colocalized genomic pair separated by 27 Mb (Fig. 1c). This locus pair was independently confirmed by fluorescence in situ hybridization (FISH) [10]. Collectively, these results demonstrate that the CLOD pipeline can effectively detect colocalized locus pairs across both short and long genomic distances.

We then applied CLOD to Hi-C datasets from a panel of four normal differentiated human cell lines, five cancer cell lines, and four embryonic cell lines (Extended Data Table 1). Across these datasets, we identified multiple colocalized locus pairs separated by 50 Mb or further in chr2 and chr14, separately (Fig. 1d&e and Extended Data Fig. 3). Notably, some of these long-range pairs were consistently detected across multiple cell lines of different types (Fig 1f&g).

Specifically, CLOD identified a colocalized genomic pair on chr14 separated by approximately 80 Mb (Fig. 1h). To determine whether this long-range colocalization was transient or persistent, we labeled the loci with two-color CRISPR-dCas9 guided fluorescent proteins in HEK293T cells (Fig. 1i, Methods, Extended Data Table 2). Live-cell imaging confirmed the sustained proximity and coordinated movement of green and red puncta, with their co-fluctuations persisting throughout the observed 10 min in some cases (Fig. 1j &i, and Extended Data Movie S1). Extended imaging further showed that the neighboring puncta remained in proximity during an imaging period of about one hour (Fig. 1l and Extended Data Fig. 1e).

MERFISH chromatin tracing data reveals prevalence and conservation of colocalized loci > 100 Mb in human chr2.

While bulk Hi-C provides an ensemble-averaged view of chromosome conformations, it lacks the resolution to capture individual chromosome structures. It cannot distinguish whether colocalized pairs originate from the same or different homologous chromosomes, nor can it determine whether observed contact frequencies reflect widespread interactions among cells or rare sub-populations. To cross-validate the prevalence of long-range pair colocalization and assess their distribution at

112 the single chromatin level, we analyzed previously published IMR90 chromatin tracing data
113 obtained using DNA MERFISH[16].

114 As a technical control, we first re-analyzed the 3D distances (d_1) of MERFISH-labeled
115 neighboring genomic loci across chr21 with a genomic separation $l_1 = 50\ kb$. The median
116 distance for region chr21:32.45 - 33.35 Mb (Extended Data Fig. 4a) was found to be 256 nm,
117 consistent with the previous report[16]. The overall distance distribution of neighboring genomic loci
118 across chr21 followed a log-normal distribution, with a median distance of 325 nm (Extended Data
119 Fig. 4b, left).

120 Next, we calculated the spatial distances of all neighboring genomic locus pairs across 2991
121 copies of chr2 that were fluorescently traced with $l_1 = 250\ kb$ (Fig 2a, green line). The
122 distribution also exhibited a log-normal pattern, with a median distance 462 nm and a first quantile
123 distance d_r = 299 nm (Fig. 2b, Extended Data Fig. 4b, right). For subsequent analyses, we adopted
124 d_r as an intrinsic distance reference, classifying a genomic locus pair as in spatial proximity if their
125 measured spatial distance $d \leq d_r$. This criterion aligns with previous studies using super-
126 resolution DNA FISH data at approximately 30 kb resolution, which employed a cutoff distance
127 of ~ 200 nm to define chromatin interactions[17], and 500 nm for the chr21 data with 50 kb resolution
128 to identify genomic pairs in proximity[18].

129 Most previous researches studied short-range chromosome interactions, such as enhancer-
130 promoter interactions, occur within genomic distances of ≤ 200 kb, typically in the 20 – 50 kb
131 range[19,20]. Few studies have specifically focused on long-range chromatin interactions. Here, we
132 calculated the spatial distances between pairs of loci with genomic separations of $l_2 \in (50, 100]$
133 Mb and $l_3 > 100$ Mb, respectively (Fig. 2a, purple and orange lines). Compared to d_1, the
134 distributions of corresponding d_2 and d_3 shifted significantly toward larger values, with median
135 distances of 1.86 μm and 2.08 μm, respectively (Fig. 2b). Interestingly, even at such large
136 genomic separations, certain locus pairs exhibited spatial distances < d_r (Fig. 2b, arrow). In the
137 following analyses, we focused on the genomic locus pairs with $l_3 > 100$ Mb and $d_3 < d_r$, denoting
138 them as long-range colocalization (LRC).

139 As expected, individual chr2 copies show diverse conformations (see examples shown in Fig.
140 2c, Extended Data Fig. 4c&d, Extended Data Movie 2-4), consistent with previous studies[17].
141 Moreover, a negative correlation exists between the total number of LRCs of an individual
142 chromosome copy and its conformation measured by convex hull volumes (Fig. 2d), suggesting a
143 potential relationship between LRC and chromosome conformation. Notably, in these
144 conformations, LRCs either form scattered small clusters or join into larger ones, along with
145 varying compactness of the chromatin structures. In addition, LRC counts per chr2 copy range
146 from fewer than 10 to over 10,000, with a peak at ~ 600 per copy (Extended Data Fig. 4e). These
147 LRCs can be formed due to transient contacts through thermal fluctuations, or via stable structures.
148 While the transient LRCs should appear infrequently among the entire cell population, stable ones
149 recur in subpopulations of chr2 copies. To test this, we analyzed the occurrence frequency of
150 155,009 unique LRCs across 2,991 chr2 copies (Fig. 2e). We identified 7,366 LRCs with

occurrence frequencies beyond the 95% population interval, defined them as stable LRCs (sLRCs) (Fig. 2e & f). Each sLRC appeared in at least 23 chr2 copies (~0.8% of the dataset), with some found in over 10% of samples. That is, sLRC exhibited a significantly higher probability of occurring within 299 nm compared to other LRC (Fig. 2g).

We then compared sLRC detected via DNA FISH with CLOD-identified colocalized pairs (> 100 Mb separation) from Hi-C. Most Hi-C outliers were also classified as sLRCs (Fig. 2h). Notably, Hi-C and DNA FISH detect proximity differently: Hi-C captures direct contact, whereas DNA FISH assesses spatial distances, using a 299 nm threshold here. Technical limitations also contribute to discrepancies — Hi-C may miss repetitive or low-accessibility regions, whereas its much larger sample size (~1 million cells vs. ~3,000 chr2 copies in DNA FISH) could detect rare configurations potentially overlooked by DNA FISH. Despite these differences, the strong agreement between Hi-C and DNA FISH reveals that LRCs are prevalent in chromatin structure, with a subset forming stable structural features.

A subset of genomic sites serves as nucleation centers (NCs) for clustering sLRCs.

Our structural analyses revealed that sLRC-associated genomic loci form intertwined networks with sLRCs acting as links (Fig. 2c, Extended Data Fig. 4c&d, Extended Data Movie S5-S8). Linkage maps from both Hi-C and DNA FISH data showed that certain genomic loci colocalize with multiple distant loci (Fig. 2f&h). These loci likely function as NCs, linking multiple sLRCs and forming the network (Fig. 3a). Plots of the number of sLRCs that a locus was associated with clearly showed that some specific locations across chr2, such as regions at ~28 – 40 Mb and the ends of chr2, spatially colocalized to significantly more well-separated cis-loci compared to other loci on the same chromosome, and thus can serve as NCs (Fig. 3b, left & Fig. 3c). Notably, locus 242 Mb (NC 242 Mb) exhibited the most sLRCs and was analyzed separately for clear visualization.

Further statistical analyses showed that 26% of sLRC-associated loci formed only one pair of sLRC, while 50% established fewer than five. However, a subset of loci participated in over 150 sLRCs (Fig. 3b, right and Fig. 3d). Even within individual chr2 copies, some loci can be paired with up to ~ 90 others distant loci to form sLRCs (Fig. 3d). A striking example was NC 242 Mb, which paired with loci across the chromosome. Two examples shown in Fig. 3e illustrate the extensiveness of sLRC partners at individual chr2 copy levels. At the population level, 89.8% of the 2,991 observed chromosomes had at least one sLRC partner at NC 242 Mb. Other NCs also showed frequent and diverse long-range interactions. For example, NC 34.25 Mb, which was found in 38% of chr2 copies, colocalized to majorly three distant regions and formed various sLRC in different chr2 (Fig. 3f). Extended Data Table 3 lists all loci associated with at least two sLRCs in chr2 copies.

To reveal the basic patterns of chromosome conformation composed by NCs, we modeled sLRCs as networks. As expected, the result exhibited that the distant regions on chr2 were clustered via sLRC, where NCs served as hubs in the net (two representative chr2 copies shown in Fig. 4g), and most of individual chr2 copy had less than 30 isolated clusters (Fig. 4h). Further

statistical analyses revealed that most NC-coordinated sLRCs clusters (59.2%) contained only 2 – 5 loci, while 5.4% had over 100 loci as nodes (Fig. 3i). Detailed analyses revealed distinct structural motifs (Fig. 3j): some NCs formed isolated single-NC clusters, while others coalesced into multi-NC hubs via direct colocalization or shared pairing loci, forming clusters with 100 – 780 nodes.

These findings suggested that NCs link multiple genomic regions, assembling hub-like domains that may be associated with chromosome conformation and territory segregation.

Neighboring LRCs cooperate to form stable complexes through multivalent binding

DNA FISH data revealed the widespread presence of stable LRC ($p \geq 0.8\%$) in chr2 copies, with some occurring in 5 – 10% of all copies (Fig. 2e). However, consider an isolated sLRC pair that stochastically transits between bound and unbound states, with association (α) and dissociation (γ) rate constants (Fig. 4a left). After dissociation, the loci diffuse apart due to thermal fluctuations, making re-encounters highly unlikely (α << 1). Achieving $p = \frac{\alpha}{\alpha+\gamma} \geq 0.8\%$ would require an exceptionally low γ, implying the need for strong and specific interactions, or alternative mechanisms stabilizing spatial proximity.

A potential clue lies in Fig. 2f, which shows some sLRC clustering together. Statistical analysis confirmed that most sLRC-associated loci have their nearest neighboring sLRC genomic loci within 500 kb (Extended Data Fig. 5a). We hypothesized that neighboring sLRCs stabilize each other via a multivalent binding mechanism — where an unbound sLRC pair remains spatially close due to adjacent bound sLRCs. This leads to an increased effective association rate constant $\alpha + \delta$, where δ represents the enhancement due to neighboring sLRCs, consequently the sLRC cluster becomes an effective two-conformation system with no or at least one bound sLRC, respectively (Fig. 4a, middle & right).

To test this model, we first defined the genomic proximity between two pairs of sLRCs. Given pairs a (loci a_1, a_2) and b (loci at b_1, b_2), we defined their genomic proximity distance as $l_{ab} = \max\left(\min(|a_1 - b_1|, |a_1 - b_2|), \min(|a_2 - b_1|, |a_2 - b_2|)\right)$ (Fig 4b, Extended Data Fig. 5b). The sLRC proximity distribution exhibited three peaks with 28% of pairs within 12 Mb (Fig. 4b). To detect multivalent sLRCs, we used a more stringent threshold, grouping sLRCs into metaLRCs, where each sLRC had at least one other sLRC with 250 kb at both ends (Fig 4c). This process grouped 72% sLRCs into 736 metaLRCs, each containing 2 to over 40 sLRCs (Extended Data Fig. 5c).

The multivalent binding model predicts that within a metaLRC, the binding of one sLRC enhances the binding of others, leading to positively correlated binding states (Fig 4a, middle). Fig. 4d and Extended Data Fig. 5d&e depicted three representative metaLRCs. Correlation analyses over MERFISH data confirmed this expectation for three representative metaLRCs (Fig 4e, Extended Data Fig. 5d&e, middle). Note that some sLRCs have higher correlations than others within the same metaLRC.

To quantify the cooperative effect, we constructed a minimal model treating all sLRCs within a metaLRC equally (Methods). Assumes a constant increase in α by δ due to neighboring bound sLRCs (Fig. 4a, middle), we obtained $\frac{\alpha}{\gamma} = \frac{p_1}{Np_0}$, where N is the number of sLRCs in the metaLRC, p_0 and p_1, the frequency of observing zero or one bound sLRC, respectively. The effective δ was determined by fitting the distribution of bound sLRCs, p_i vs. i for i > 1, which significantly exceeded predictions from a non-cooperative model ($\delta = 0$) (Fig. 4f).

Analysis of individual chr2 copies further supported the model. The spatial distances between two tagged genomic distant regions with bound form were much shorter compared to those in free form chr2 copies (Fig. 4g, Extended Data Movies 9-12). When one or more sLRCs were bound, the two genomic regions within the metaLRC were held closer, though additional bound sLRCs (> 1) did not significantly alter spatial distance (Fig. 4h). The observed non-monotonic p_i v.s. i curves for certain representative metaLRCs (e.g., 33.75-35.5 Mb & 146-146.25 Mb; 119.25-123.75 Mb & 242 Mb) suggested even stronger cooperativity than the minimal model predicted (see Methods) (Extended Data Fig. 5d&e).

For all metaLRCs of chr2 with N > 2, α/γ values from FISH data were < 0.007 (Fig. 4i), indicating that associating two distant loci is entropically unfavorable. However, the corresponding $(\alpha + \delta)/\gamma$ values were up to 10 – 100× higher, making the probability of observing a bound metaLRC over 1 – 10% (Extended Data Fig. 5f). That is, pairs of distant loci can remain stably proximate within a chromosome via neighbor-assisted multivalent binding, significantly enhancing structural stability and long-range genomic interactions.

Structural analyses reveal multiple candidate mechanisms for sLRC formation.

Colocalization of distant cis-loci had been intermittently observed, with three general mechanisms proposed to explain LRC formation (Fig. 5a, I-III). (I) A/B compartment, or loose/compact region segregation may bring distant B compartment loci in transient proximity[21]. (II) Transcriptionally active loci, chromatin around which are typically loose and accessible, may colocalize via attaching to shared nuclear structures like speckles, or nucleoli[22]. (III) The nuclear membrane may recruit multiple loci into spatial proximity[23]. An alternative mechanism (IV) suggests stable sLRC formation independent of major nuclear structures or compartment segregation. While the first three have been proposed previously[24], sLRC assemblies formed through Mechanism IV were considered rare.

Examination of A/B compartment identities among sLRC-forming loci revealed no clear pattern or correlation between the two loci of each sLRC pair (Extended Data Fig. 6a-c), while the majority of sLRCs have two loose ends (Extended Data Fig. 6b, bottom), which is consistent with the observation from a previous study[18]. Note that the A/B compartment was calculated from bulk level data. That is, an ensemble of chromosomes may be composed with subpopulations with and without the presence of an sLRC under study, thus conclusions from such bulk-averaged data are not definite. Bulk averaging also prevented us from concluding whether sLRCs or the loci linked to sLRCs have specific epigenetic patterns (Extended Data Fig. 6 d&e).

Using MERFISH data, we quantified local geometric structures within 299 nm of a locus by defining regional density of short-ranged (< 10 Mb) neighbors (RD-SN) and regional density heterogeneity (RDH), which measures the uniformity of local spatial distributions (Fig. 5b). Compared RD-SN to A/B compartment identified from Hi-C data, an A compartment has a higher probability of having a loose neighborhood than a B compartment (Extended Data Fig. 6f). Then one expects that Mechanism I and III lead to high RD-SN for both loci associated with a sLRC, Mechanism II leads to low-low RD-SN, while the pattern for Mechanism IV is undetermined. Furthermore, a locus next to a nuclear structure (Mechanism II and III) due to the excluded volume of the latter. Therefore, a combined signature of mixed low/high RD-SN and low RDH only comes from Mechanism IV. Analyses of the sLRCs revealed broad distributions of these two quantities, implying possible contribution from diverse mechanisms on sLRC formation (Fig. 5c&d), as also reflected from examining representative structures (Fig. 5e).

Furthermore, we observed spatial colocalization of *cis*-loci that are well separated on the genome in other chromosomes with different lengths. Specifically, we analyzed the colocalization state of locus pairs in chromosome 21 further than 20 Mb, which is about half of the region (~36.3 Mb) that was examined with MERFISH. The distribution of most spatial distances between loci pairs longer > 20 Mb in chr21 range between 0 and 5 μm, and colocalization pairs of loci denote to those whose spatial distances are closer than 299 nm in individual chromosome structure as forementioned (Extended Data Fig. 6g). As expected, both linkage maps and 3D structure plots show the existence and variety of LRCs in individual chr21 (Fig. 5f). While most of 32,341 loci pairs longer than 20 Mb were only observed in less than 200 out of 7591 chromosomes, 1718 or 5% of them appeared in more than 325 chromosomes, which were identified as stable colocalization loci pairs here (Fig. 5g). Moreover, similar to chr2, linkage map shown the stable colocalization loci pairs reveals that the clustering of sLRC and loci served as nucleation centers also exist in chr21 (Fig. 5h).

Together, all analyses indicate the prevalence of long-range colocalization in chromosome and the epigenetic, positional, and spatial distribution results suggest an additional, previously underappreciated mechanism contributing to stable sLRC formation.

DISCUSSION

Our analyses of Hi-C and DNA FISH chromosome tracing data reveal widespread colocalization of genomic loci separated by distances far exceeding a typical TAD. Statistically, the probability of multiple genomic loci with $l \geq 100$ Mb being in physical proximity without stable structural mechanisms would be negligible. While the exact molecular mechanism remains unclear, some sLRCs may form through anchoring to nuclear structures such as nucleoli, speckles, or the nuclear envelope. However, the presence of many sLRCs and NCs away from these structures suggests additional nuclear structure-independent mechanisms, such as molecular assembly or condensates. Notably, in contrast to assumptions in random loop models[8], some large-sized loops appear sequence-specific, implying potential roles of non-coding DNA elements. For instance, the 242

304 Mb locus, enriched in repetitive sequences, may serve as a selective binding site for partnering
305 loci, either directly or through bridging molecular factors.

306 Besides their varied formation mechanism, these sLRC structures likely play diverse functional
307 roles, including transcriptional regulation and broader chromosome 3D organization. Proper DNA
308 folding into spatially segregated territories with minimal entanglement and knot formation is
309 essential for eukaryotic genome organization and gene regulation. This process must be robust
310 across all the cell types, including those with abnormal karyotypes, and must function despite cell
311 type-specific euchromatin/heterochromatin partitioning, folding stochasticity, and mutations.
312 From an engineering perspective, achieving this through a single centralized mechanism requiring
313 precise tuning would be challenging. Instead, our findings suggest a redundant, distributed
314 component mechanism that facilitates chromosome folding into compact structures (Fig. 5j). The
315 genome harbors an extensive repertoire of NCs and partnering loci separated by long genomic
316 distances, from which only a subset is available for a given cell type. At the individual chromosome
317 level, subsets of these loci stochastically form sLRCs and larger structures, contributing to the
318 observed heterogeneity of chromosome structures[25]. The structures provide a "divide-and-conquer"
319 mechanism to facilitate forming chromosome structures at smaller length scales, e.g., by forming
320 segregated globally compact chromosome configurations and close-end boundary conditions for
321 simultaneous crumpled folding at multiple regions (Extended Data Fig. 4d)[5], and enhanced contact
322 frequencies for forming loops at smaller scales.

323 It is important to note that our study employed stringent criteria to identify representative
324 examples rather than exhaustively cataloging all stable long-range colocalizations. While we
325 focused on loci separated by >100 Mb, similar structures exist at shorter genomic distances,
326 particularly in smaller chromosomes such as chr21 (~48 Mb).

327 A key unsolved question is how sLRC loci locate and associate with each other. One plausible
328 mechanism is that, during post-mitotic DNA decondensation, some genomically distant loci may
329 be in spatial proximity and associate stochastically, leading to subsequent formation of additional
330 sLRCs and stabilization of the assembled structures.

331 In summary, our analyses of chromosome tracing and Hi-C data uncover prevalent, conserved
332 large-sized loops and their clustering into stable structures. These findings raise important
333 questions regarding the molecular mechanisms, temporal dynamics, and functional significances
334 of these structures on chromosome configurations, gene transcriptional activity, and cell type
335 regulation.

336

337

References

1 Dill, K. A. & MacCallum, J. L. The Protein-Folding Problem, 50 Years On. *Science* **338**, 1042, doi:10.1126/science.1219021 (2012).

2 Bolzer, A. *et al.* Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol.* **3**, e157, doi:10.1371/journal.pbio.0030157 (2005).

3 Lieberman-Aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **326**, 289 (2009).

4 Wang, S. *et al.* Spatial organization of chromatin domains and compartments in single chromosomes. *Science* **353**, 598 (2016).

5 Grosberg, A., Rabin, Y., Havlin, S. & Neer, A. Crumpled Globule Model of the Three-Dimensional Structure of DNA. *Europhysics Letters* **23**, 373, doi:10.1209/0295-5075/23/5/012 (1993).

6 Gürsoy, G., Xu, Y., Kenter, A. L. & Liang, J. Spatial confinement is a major determinant of the folding landscape of human chromosomes. *Nucleic Acids Research* **42**, 8223-8230, doi:10.1093/nar/gku462 (2014).

7 Polovnikov, K. E. *et al.* Crumpled Polymer with Loops Recapitulates Key Features of Chromosome Organization. *Physical Review X* **13**, 041029, doi:10.1103/PhysRevX.13.041029 (2023).

8 Mateos-Langerak, J. *et al.* Spatially confined folding of chromatin in the interphase nucleus. *Proceedings of the National Academy of Sciences* **106**, 3812-3817, doi:10.1073/pnas.0809501106 (2009).

9 Quinodoz, S. A. *et al.* Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell* **174**, 744-757.e724, doi:10.1016/j.cell.2018.05.024 (2018).

10 Jäger, R. *et al.* Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nature communications* **6**, 6178 (2015).

11 Pombo, A. & Dillon, N. Three-dimensional genome architecture: players and mechanisms. *Nature Reviews Molecular Cell Biology* **16**, 245-257, doi:10.1038/nrm3965 (2015).

12 Whyte, Warren A. *et al.* Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. *Cell* **153**, 307-319, doi:http://dx.doi.org/10.1016/j.cell.2013.03.035 (2013).

13 Barutcu, A. R. *et al.* Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells. *Genome Biology* **16**, 214, doi:10.1186/s13059-015-0768-0 (2015).

14 Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290-294 (2013).

15 Rao, S. S. P. *et al.* Cohesin Loss Eliminates All Loop Domains. *Cell* **171**, 305-320.e324, doi:https://doi.org/10.1016/j.cell.2017.09.026 (2017).

16 Su, J.-H., Zheng, P., Kinrot, S. S., Bintu, B. & Zhuang, X. Genome-scale imaging of the 3D organization and transcriptional activity of chromatin. *Cell* **182**, 1641-1659. e1626 (2020).

17 Bintu, B. *et al.* Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science* **362**, eaau1783 (2018).

18 Su, J.-H., Zheng, P., Kinrot, S. S., Bintu, B. & Zhuang, X. Genome-Scale Imaging of the 3D Organization and Transcriptional Activity of Chromatin. *Cell* **182**, 1641-1659.e1626, doi:https://doi.org/10.1016/j.cell.2020.07.032 (2020).
19 Chepelev, I., Wei, G., Wangsa, D., Tang, Q. & Zhao, K. Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. *Cell research* **22**, 490-503 (2012).
20 Furlong, E. E. & Levine, M. Developmental enhancers and chromosome topology. *Science* **361**, 1341-1345 (2018).
21 Gibson, B. A. *et al.* Organization of Chromatin by Intrinsic and Regulated Phase Separation. *Cell* **179**, 470-484.e421, doi:https://doi.org/10.1016/j.cell.2019.08.037 (2019).
22 Kim, J., Venkata, N. C., Hernandez Gonzalez, G. A., Khanna, N. & Belmont, A. S. Gene expression amplification by nuclear speckle association. *Journal of Cell Biology* **219**, e201904046, doi:10.1083/jcb.201904046 (2019).
23 van Steensel, B. & Belmont, A. S. Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression. *Cell* **169**, 780-791, doi:10.1016/j.cell.2017.04.022 (2017).
24 Dekker, J. & Mirny, L. A. The chromosome folding problem and how cells solve it. *Cell* **187**, 6424-6450, doi:10.1016/j.cell.2024.10.026 (2024).
25 Finn, E. H. *et al.* Extensive Heterogeneity and Intrinsic Variation in Spatial Genome Organization. *Cell* **176**, 1502-1515.e1510, doi:https://doi.org/10.1016/j.cell.2019.01.020 (2019).

Abbreviations and glossary:

LRC: Long-range colocalization for genomic locus pairs > 100 Mb and within 299 nm in spatial distance.

sLRC: Stable long-range colocalization for LRC pairs with occurrence frequency > 4% in the IMR90 chr2 MERFISH dataset.

metaLRC: a group of sLRCs, in which for each sLRC there was at least one other sLRC with 250 kb for both ends.

NC: Nucleation center referring to locus having multiple sLRC partner loci.

RD-SN: Regional density of short-ranged neighbors (< 10 Mb) within a 299 nm range of a tagged locus.

RDH: Reginal density heterogeneity.

Acknowledgments: We thank Tom Mistelli, Harinder Singh, Jing Chen, Jian Liu for helpful discussions.

Author contributions:

Conceptualization: JZ, JX

Methodology: JZ, SW, SCW, JX

Investigation: JZ, JX

Visualization: JZ, JX

Funding acquisition: JX

Project administration: JX

Supervision: JX

Writing – original draft: JX

Writing – review & editing: JZ, SW, SCW, JX

Competing interests: Authors declare that they have no competing interests.

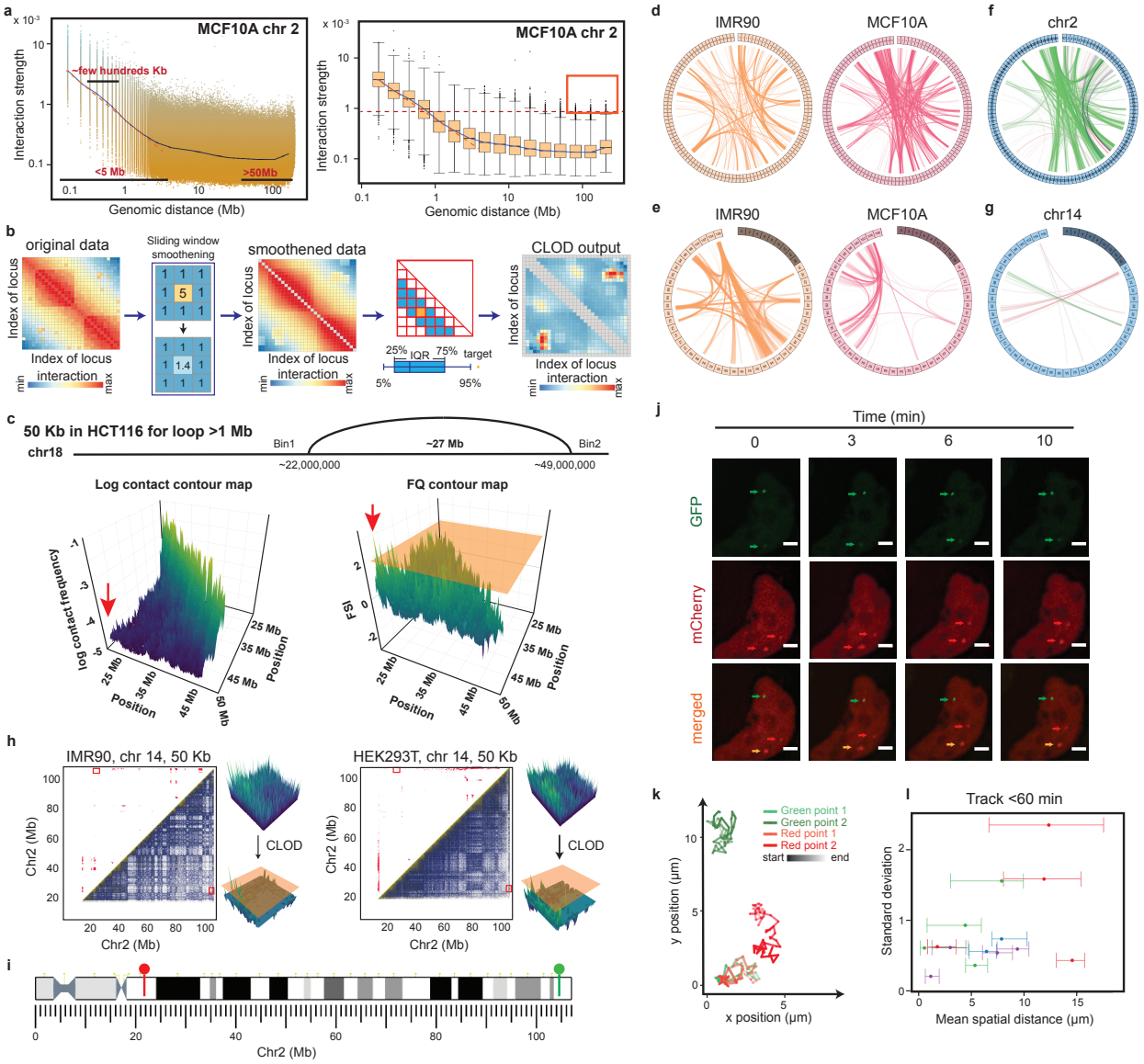

Fig. 1| Hi-C data reveals prevalent existence of colocalized pairs of long-range genomic loci. **a**, Scatter plot (left) and boxplot (right) showing the decay trend (blue line) of contact frequency with increasing genomic distance. Red boxed points represent locus pairs > 50 Mb apart with contact frequency exceeding the third quartile of those observed for ~1 Mb pairs. **b**, Schematic of the CLOD pipeline. **c**, Experimental validation of colocalized pairs predicted by CLOD using Hi-C data from HCT116 cells at 50 kb resolution. **d-e**, Representative linkage maps of colocalized locus pairs > 50 Mb on chr2 and chr14 in two different cell lines. **f-g**, Linkage maps of colocalized pairs > 50 Mb apart, shared across at least two of the following cell types: four embryonic cell lines (green), four normal differentiated cell lines (pink), and four cancer cell lines (blue). **h**, CLOD-based identification of colocalized pairs on chr14 from Hi-C data of IMR90 and HEK293T

444 cells at 50 kb resolution. **i**, CRISPR-dCas9 two-color labeling sites on chr14. **j**, Live-cell imaging
445 snapshots of a HEK293T cell labeled with RFP and GFP. Scale bar: 2 μm. **k**, 10-min time-resolved
446 trajectories of blue and red puncta in the labeled cell from panel j. **l**, The distance fluctuations
447 between blue and red puncta pairs in four cells that were tracked for 60 min. The left and right bars
448 indicate the maximum and minimum distances between a pair of puncta.

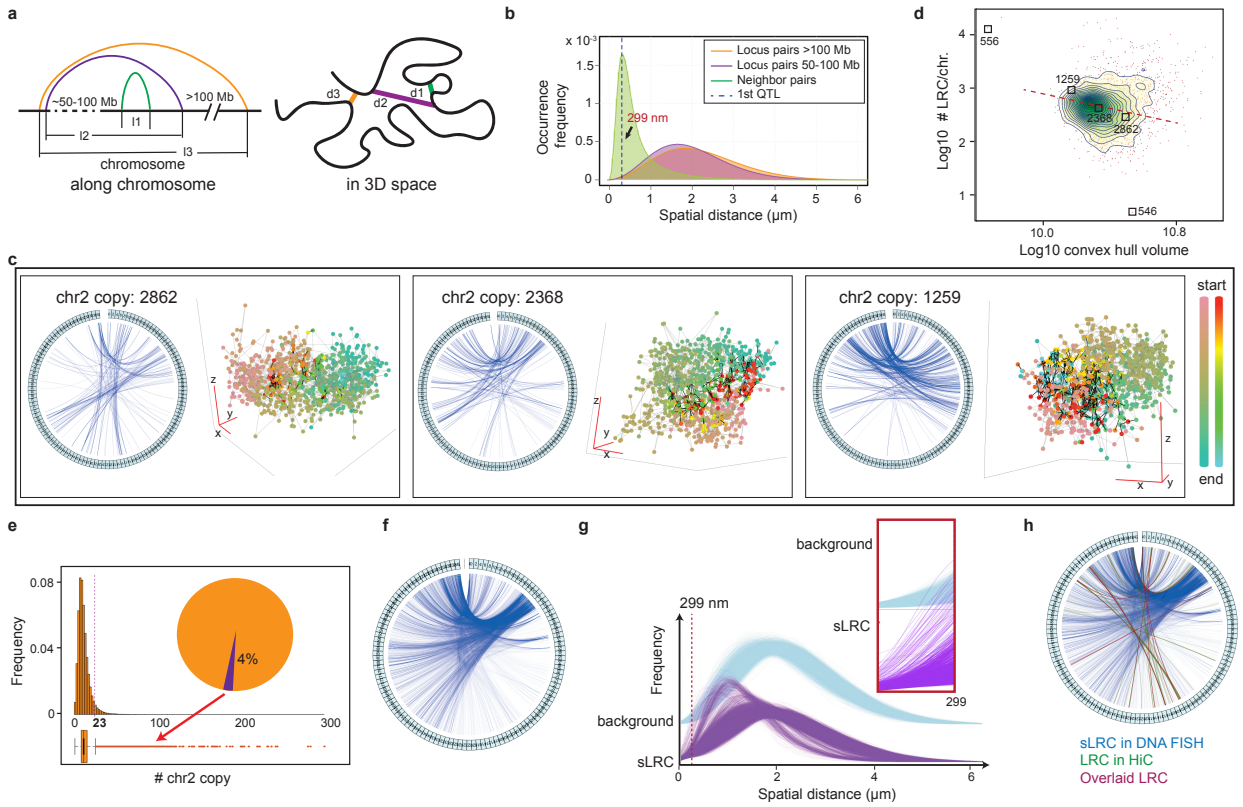

Fig. 2|MERFISH chromatin tracing reveals widespread long-range colocalization (LRC). a, Schematic illustration of genomic pairs and their genomic versus spatial distances. **b**, Density curves of spatial distance distributions for loci pairs with genomic separations of 250 kb, 50-100 Mb, and >100 Mb. The dashed line represents the first quantile of neighboring locus pairs. **c**, LRC linkage maps and 3D structures of 3 representative chromosomes. **d**, Scatter plot showing the relationship between chromosome compaction, quantified by the convex hull volume, and the number of LRCs. **e**, Histogram (top) and boxplot (bottom) showing the distribution of chr2 copy counts (n = 2991) containing LRCs. LRCs present in >23 copies (red dashed line) are classified as stable LRCs (sLRCs). **f**, Identification of chr2 sLRCs from DNA FISH data. **g**, Spatial distance distributions of locus pairs with genomic separation >100 Mb. Purple lines (n = 6443): sLRCs; blue lines (n = 6443): randomly selected non-sLRCs. The red dashed line represents the first quantile of spatial distances between neighbor loci on chr2 (299 nm). **h**, Comparison of colocalized chr2 locus pairs (>100 Mb apart) identified in IMR90 cells using DNA FISH (blue lines) and Hi-C (green lines) data. Overlapping links are highlighted in brown.

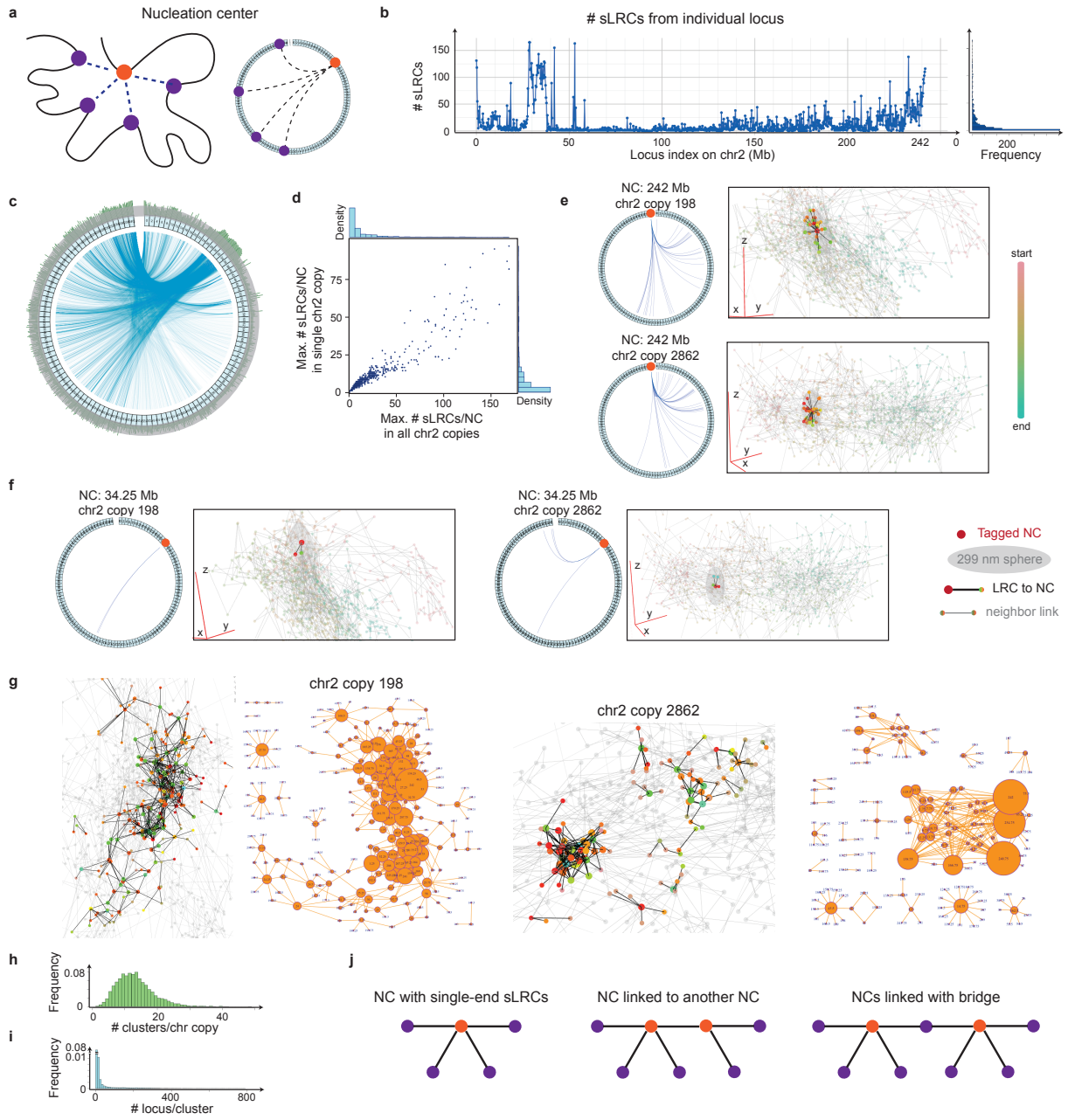

Fig. 3|Nucleation centers (NC) coordinate clustering of sLRCs. a, Left: Schematic illustration of a nucleation center (NC) (orange) forming spatial proximity with multiple partner LRC loci (purple). Right: Ring plot depicting the same interaction network. **b**, Left: Line plot showing the number of sLRCs originating from a given locus, based on 2,991 chr2 copies. Right: Histogram illustrating the distribution of loci (x-axis) connecting to a given number of sLRC (y-axis). **c**, sLRC linkage map with corresponding numbers of potential sLRCs (green bars outside the ring) originating from specific loci in the chr2. **d**, Scatter plot comparing the total number of potential sLRC from targeted loci across the analyzed chr2 population to the maximum number of sLRCs from a single targeted locus in an individual chr2 copy. Top and right panels: Corresponding distributions of numbers shown in the scatter plot. **e-f**, Linkage maps and 3D structures

475 highlighting two representative NCs (NC 242 Mb and NC 34.25 Mb) on chr2, along with their
476 associated sLRCs. **g**, 3D structural representations and network maps showing connections among
477 NC loci in two representative chr2 copies. **h**, Histogram displaying the number of locus clusters
478 — defined as sets of loci connected to at least one other locus — per individual chr2 copy. **i**,
479 Histogram depicting the number of loci per locus-cluster, as identified in panel h. **j**, Diagrams
480 illustrating single NC elements and various types of connections between NCs.

481

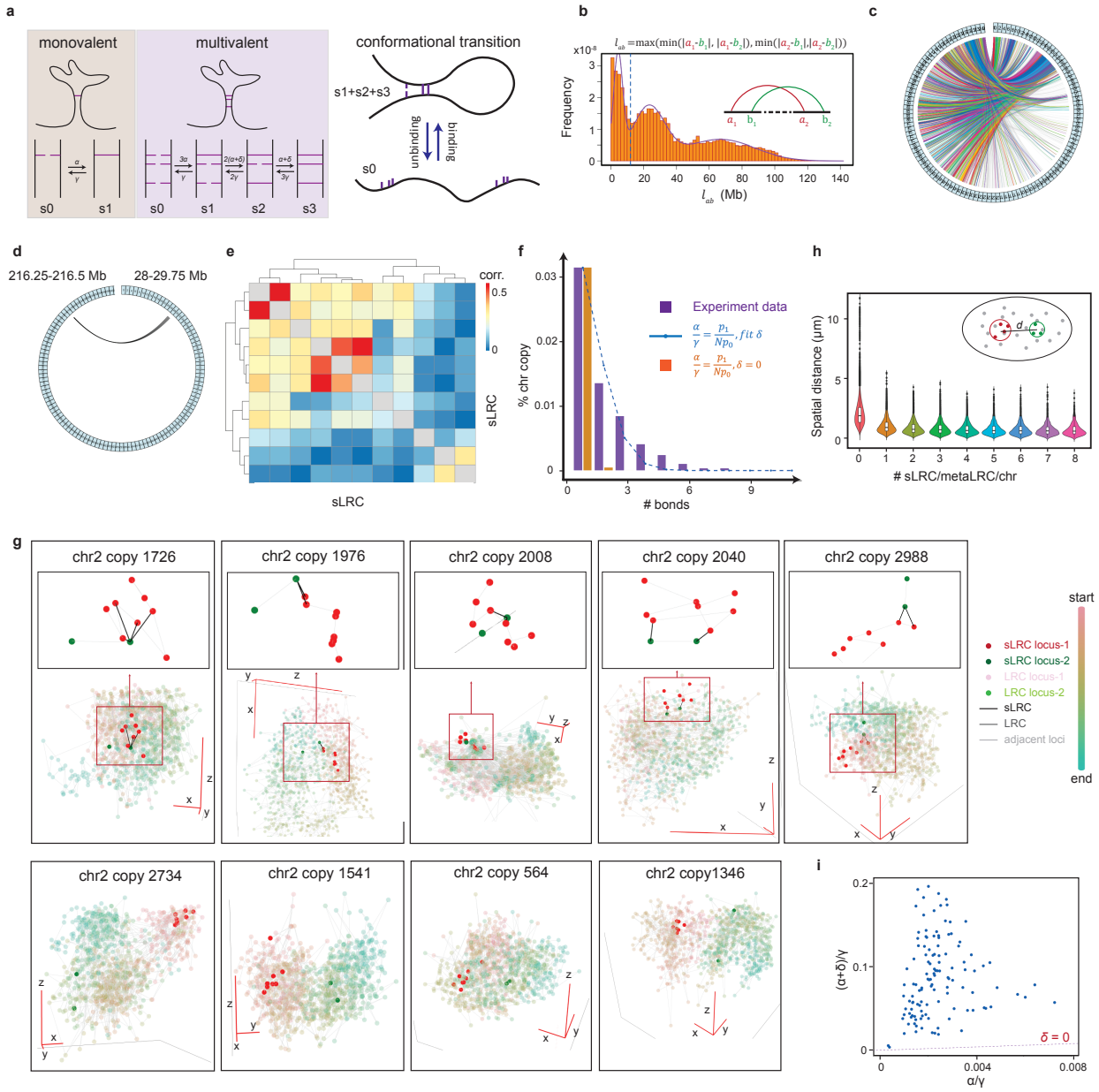

Fig. 4|DNA FISH data analysis reveals that multivalent binding stabilizes sLRC clusters. a, Left: Schematic illustration of monovalent binding vs. multivalent binding models. α: association rate constant of a single pair in the absence of other bound sLRCs; γ: dissociation rate constant of a single pair; δ: association rate enhancement due to neighboring bound sLRCs. Right: A multivalent site is coarse-grained into a two-state system — either free or bound (with one or more bound sLRCs) two-state system. **b**, Frequency distribution of the genomic separation between two LRCs. **c**, sLRC and metaLRC linkage map. The outer circle represents chr2, and each line denotes a sLRC. Gray lines: isolated sLRC; colored lines (except gray): sLRCs belonging to a metaLRC. **d**, A representative metaLRC. **e**, Heatmap showing correlation between sLRCs within the metaLRC from panel d. **f**, Comparison of the probabilities of observing various numbers of bound

493 sLRCs within the metaLRC (purple bar) vs. predictions based on an independent sLRC binding
494 model (orange bar), and a multivalent binding model (blue dashed line). p_0 is not shown. **g**,
495 Representative 3D structures of chr2, highlighting the metaLRC from panel d. Top: Structures
496 with at least one bound sLRC within the metaLRC. Bottom: Structures with no bound sLRC within
497 the metaLRC. **h**, Violin plot showing the relationship between 3D distances (between centers of
498 minimal-sized bounding spheres of the two genomic regions forming the metaLRC and the number
499 of bound sLRCs per individual chr2 copy. **i**, Multivalent binding model parameters for metaLRCs,
500 obtained through fitting DNA FISH data.

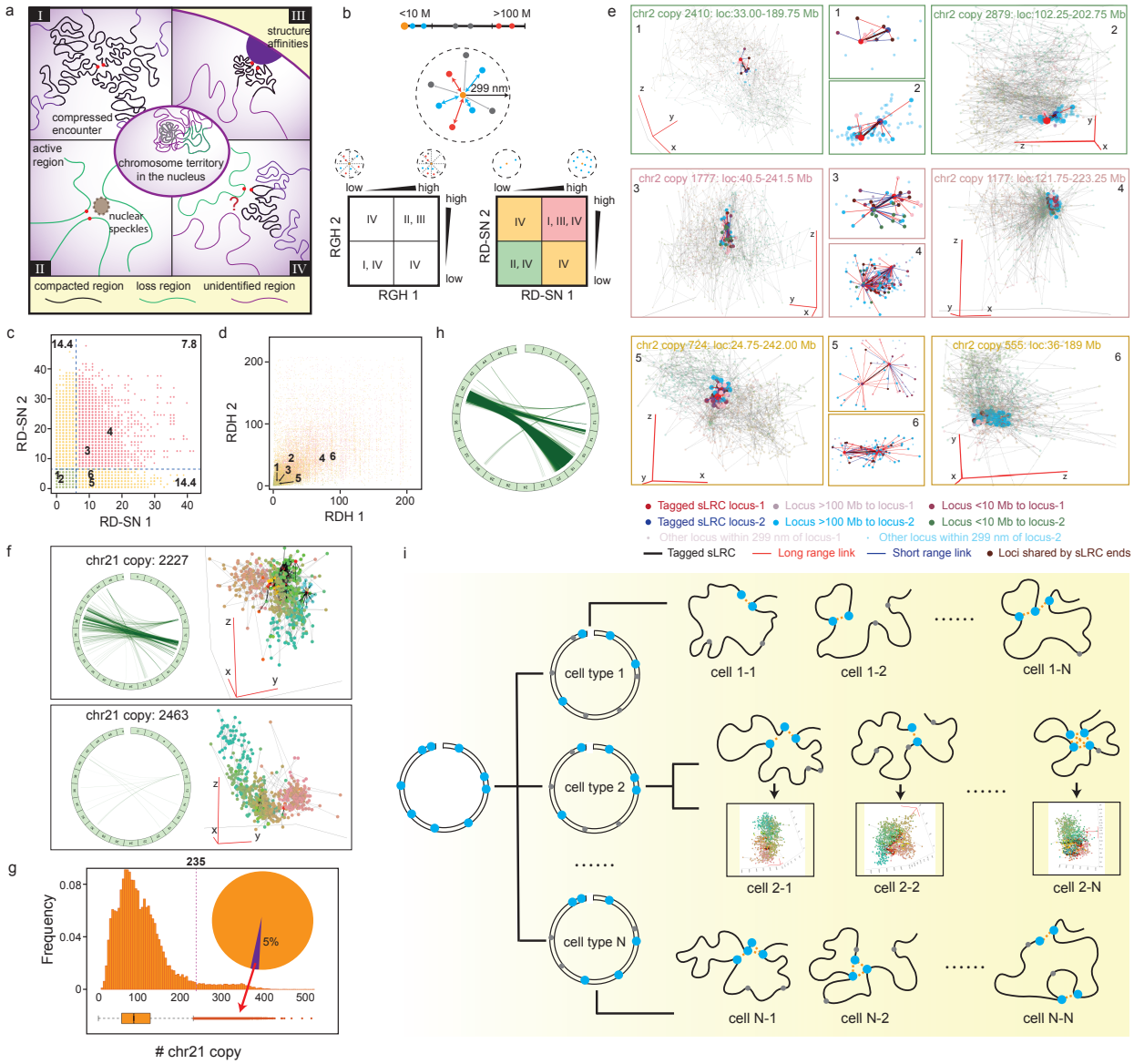

Fig. 5|Possible mechanisms of sLRC formation and functional roles on chromosome folding. **a**, Schematic of four possible mechanisms for sLRC formation. I & II: A/B compartment segregation leads to compressed encounters of genomically distant loci in the B and A compartments, respectively. III: Colocalization of transcriptionally inactive, genomically distant loci interacting with common nuclear membrane structures. IV: Specific interactions between genomically distant loci form an assembly without the involvement of large nuclear structures. **b**, Definition of regional density of short-ranged neighbors (RD-SN) and regional density heterogeneity (RDH) to characterize structural properties of sLRCs formed via mechanisms in panel a. **c-d**, Scatter plots of RD-SN and RDH around both ends of individual sLRCs in each chr2 copy. Each dot represents one sLRC in a single chr2 copy. **e**, 3D structures of six representative chr2 copies highlighted in panel c&d. **f**, Representative MERFISH structures of chr21. Black lines represent LRCs. **g**, Histogram (top) and boxplot (bottom) showing the distribution of chr21 copy

514 counts (n = 7591) containing LRCs. LRCs present in >325 copies (red dash line) are classified as
515 stable LRCs (sLRCs). **h**, Identification of chr2 sLRCs from DNA FISH data. **i**, Schematic
516 illustration of the proposed excessively redundant distributed component mechanism for
517 chromosome folding.

518

519

520

521

522

523 **Methods**

524 Hi-C data analysis

525 Raw Hi-C data were downloaded from GEO, as shown in Extended Data Table 1, and were aligned
526 to the human genome *hg38* using the Burrows-Wheeler Alignment tool[26]. Hi-C contacts were then
527 identified and filtered with Pairtools[27]. The contact matrices used for the CLOD algorithm were
528 created and balanced with Cooler[28].

529 Chromosome Long-range colocalization identifier through Outlier Detection (CLOD) pipeline

530 To establish a minimalistic and robust pipeline to detect long-range colocalization between cis-
531 loci based on Hi-C results, we smoothed the noisy Hi-C data and then highlighted the robust off-
532 diagonal signals representing chromatin contacts between distal genomic regions. In detail, first
533 we used the balanced Hi-C results from Cooler and converted the original reads to the average of
534 the 7x7 window to smooth the random noisy signals. Then to give prominence to the signal at off-
535 diagonal regions of contact matrix, we grouped all pairs of loci with the same genomic distances
536 (l) and those with $l \pm resolution$ and calculated the quantiles. The final output of the CLOD is
537 the fold change of the interquartile range above the third quantile of all pairs with similar genomic
538 distances ($\frac{Signal-Q3}{Q3-Q1}$).

539 MERFISH data

540 The Euclidean coordinates of chromosome fragments were directly downloaded from Zenodo
541 repository (records:3928890) [18].

542 CRISPR-dCas9 two-color labeling

543 The plasmid cloning method used for CRISPR-dcas9 fluorescent labeling was adapted from a
544 previously established protocol[29]. The sgRNAs used in this study are listed in Extended Data Table
545 2.

546 Cell culture

547 HEK293T cells (#ATCC-CRL-3216) were cultured in DMEM (Invitrogen #11995073)
548 supplemented with 10% FBS (Gbico #10437028), 100 units/ml penicillin, and 100 µg/ml
549 streptomycin (Invitrogen). Cells were maintained at 37°C in 5% CO2, and used within 15–30
550 passages for experiments.

551 For transfection, HEK293T cells were seeded onto 35 mm glass-bottom dishes one day before the
552 experiments. In the next day, 750 ng of donor template and 750 ng of sgRNA-Cas9 plasmid were
553 co-transfected using FuGENE HD Transfection Reagent (Promega #E2311). Cells were incubated
554 for 48 hours before imaging[29].

555 Time-lapse imaging

556 For imaging presentation (Fig. 1j), cells were visualized using a Nikon Ti microscope equipped
557 with a 100x oil objective. Excitation wavelengths were 470 nm for GFP and 555 nm for RFP. Cells
558 were imaged every 10 seconds for 10 minutes with an exposure time of 50 ms.

559 For trajectory tracking (Fig. 1l), cells were maintained in a Tokai Hit Microscope Stage Top
560 Incubator and imaged using a Nikon Ti microscope with FITC and TRITC channels (excitation
561 wavelength is 470 nm and 555 nm for GFP and RFP, respectively; 60x oil objective). Cells were
562 imaged every minute for 60 minutes with an exposure time of 100 ms[30].

563 Imaging processing

564 Due to the low number of cells, cell segmentation and puncta recognition were performed by
565 CellProfiler[31] and corrected manually.

566 Derivation of the minimal multivalent binding model

567 Consider a metaLRC that contains N pairs of sLRCs. Given the observed abnormally high contact
568 frequency of a sLRC pair as compared to non-sLRC pairs separated with similar genomic distance,
569 it is likely that the two loci are held together through some direct interactions or mediated by other
570 molecular components. Thus, we can assume that an sLRC pair i exists in one of the two possible
571 configurations, $\sigma_i = 1$ if the pair is bound (with spatial distance $d_r <= 299$ nm), and 0 otherwise.
572 We assume this metaLRC is genomically distant from other metaLRCs and sLRCs, so for a good
573 approximation there is no need to explicitly consider the influence of the latter on the association-
574 dissociation dynamics of sLRCs in this tagged metaLRC.

575 Suppose that without influence from other sLRCs within the tagged metaLRC, an sLRC pair has
576 a bare association rate α and a dissociation rate γ. Then at kinetic equilibrium, the probability in
577 the bound state is $\frac{\alpha}{\alpha+\gamma}$, and from the FISH data we expect that $\alpha \ll \gamma$. Here for simplicity we
578 assume that each sLRC pair has the same association-dissociation kinetic parameters. Let's focus
579 on one pair i. When there is one or more other bound sLRC pairs within the metaLRC, after
580 dissociation the two loci of pair i are constrained by the nearby bound sLRCs and cannot diffuse
581 far away from each other before rebind. Note that the genomic distance between these two loci of
582 a sLRC ≥ 100 Mb, and by definition the genomic distance between two sLRCs i and j within a
583 metaLRC $l_{ij} << 100$ Mb. Consequently, at the presence of other bound sLRC pair(s), the two loci
584 of a dissociated sLRC have an increased probability to confront each other and rebind, so the
585 effective association constant changes to $\alpha + \delta$, and we expect that $\delta \gg \alpha$. For simplicity, given
586 $l_{ij} << 100$ Mb let us assume that each sLRC pair has the same effect on any other one, and the
587 effect is not additive but δ remains the same value at the presence of one or more bound sLRC
588 pairs. Given the genomic distance between two loci belonging to two different sLRCs, we further
589 assume that there is no direct interaction between two sLRC pairs to simplify the analyses. That
590 is, we assume that the dissociation constant of a sLRC pair, γ, is not affected by the presence of
591 other bound sLRC pairs.

592 With the above model, one can write down a set of master equation for finding that the metaLRC
593 exists with 0, 1,…, N pairs of bound sLRCs,

$\frac{dp_0}{dt} = -Nap_0 + \gamma p_1,$

$$\frac{dp_1}{dt} = -\big((N-1)(a+\delta)+\gamma\big)p_1 + 2\gamma p_2 + Nap_0,$$

$\frac{dp_2}{dt} = -\big((N-2)(a+\delta)+2\gamma\big)p_2 + 3\gamma p_3 + (N-1)(a+\delta)p_1,$

...

$\frac{dp_{N-1}}{dt} = -\big((a+\delta)+(N-1)\gamma\big)p_{N-1} + N\gamma p_N + 2(a+\delta)p_{N-2}$,

$\frac{dp_N}{dt} = -N\gamma p_N + (a+\delta)p_{N-1}.$

Then at steady state,

$p_1 = \frac{Na}{\gamma}p_0,$

$p_2 = \frac{(N-1)(a+\delta)}{2\gamma}p_1,$

$p_3 = \frac{(N-2)(a+\delta)}{3\gamma}p_2,$

...

$p_N = \frac{a+\delta}{N\gamma}p_{N-1}.$

Define the probability of observing presence of one or more pairs as

$p = 1 - p_0 = \sum_{i=1,\dots,N} p_i.$

One has,

$$p = \left(\frac{N!\,(a+\delta)^{N-1}}{N!\,\gamma^{N-1}} + \cdots + \frac{(N-1)N(a+\delta)}{2\gamma} + N\right)\frac{a}{\gamma}p_0$$

$$= \left(\frac{N!\,(a+\delta)^{N}}{N!\,\gamma^{N}} + \cdots + \frac{(N-1)N(a+\delta)^2}{2\gamma^2} + N\frac{(a+\delta)}{\gamma}\right)\frac{a}{a+\delta}p_0$$

$$= \left(\left(1+\frac{a+\delta}{\gamma}\right)^N - 1\right)\frac{a}{a+\delta}p_0 = \frac{\left(\left(1+\frac{a+\delta}{\gamma}\right)^N - 1\right)\frac{a}{a+\delta}}{\left(\left(1+\frac{a+\delta}{\gamma}\right)^N - 1\right)\frac{a}{a+\delta}+1}.$$

In the limit $\alpha \ll \delta \ll \gamma$, the above expression can be further simplified as,

$$p \approx \frac{N\left(1+\frac{(N-1)}{2}\frac{a+\delta}{\gamma}\right)\frac{a}{\gamma}}{N\left(1+\frac{(N-1)}{2}\frac{a+\delta}{\gamma}\right)\frac{a}{\gamma}+1} \approx \frac{N\left(1+\frac{(N-1)}{2}\frac{\delta}{\gamma}\right)\frac{a}{\gamma}}{N\left(1+\frac{(N-1)}{2}\frac{\delta}{\gamma}\right)\frac{a}{\gamma}+1}.$$

Several mechanisms can further increase the cooperativity beyond what described by a constant δ value in the above model. Presence of additional bound sLRC pair may further constrain the diffusion of the two loci of a dissociated sLRC, leading to further increase of the value of δ. That is, the value of δ may increase with the number of bound sLRC pairs. The bound sLRC pairs may physically contact and stabilize each other with an increased γ value. Indeed, we observed non-monotonic histograms of p_i for some metaLRCs (Extended Data Fig. 5d&e). With a thermodynamic equilibrium model, one can introduce some binding free energies to describe the cooperativity as in allosteric effect and a previous cooperative enhancer binding model[32]. For simplicity in this work we restricted the analyses to the minimal model discussed above.

Note that all the loci belonging to a metaLRC can be clustered as two clumped genomic regions that are genomically distant (i.e., ~ 100 Mb or further). Subsequent analyses further coarse-grained the binding state of a metaLRC as $s = 1$ if $(\sum_{i=1}^{N} \sigma_i) \geq 1$and 0 otherwise.

References

26 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

27 Open2C *et al.* Pairtools: From sequencing data to chromosome contacts. *PLOS Computational Biology* **20**, e1012164, doi:10.1371/journal.pcbi.1012164 (2024).

28 Abdennur, N. & Mirny, L. A. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics* **36**, 311-316 (2020).

29 Chen, Y.-J. *et al.* Rapid, modular, and cost-effective generation of donor DNA constructs for CRISPR-based gene knock-in. *Biology Methods and Protocols* **5**, bpaa006, doi:10.1093/biomethods/bpaa006 (2020).

30 Wang, W., Poe, D., Yang, Y., Hyatt, T. & Xing, J. Epithelial-to-mesenchymal transition proceeds through directional destabilization of multidimensional attractor. *eLife* **11**, e74866, doi:10.7554/eLife.74866 (2022).

31 Stirling, D. R. *et al.* CellProfiler 4: improvements in speed, utility and usability. *BMC bioinformatics* **22**, 1-11 (2021).

32 Tian, X. J., Zhang, H., Sannerud, J. & Xing, J. Achieving diverse and monoallelic olfactory receptor selection through dual-objective optimization design. *Proceedings of the National Academy of Sciences of the United States of America* **113**, E2889-2898, doi:10.1073/pnas.1601722113 (2016).

33 Dixon, J. R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331-336, doi:10.1038/nature14222 (2015).

34 Rhie, S. K. *et al.* A high-resolution 3D epigenomic map reveals insights into the creation of the prostate cancer transcriptome. *Nature Communications* **10**, 4154, doi:10.1038/s41467-019-12079-8 (2019).

35 Ren, B. *et al.* High-resolution Hi-C maps highlight multiscale 3D epigenome reprogramming during pancreatic cancer metastasis. *Journal of Hematology & Oncology* **14**, 120, doi:10.1186/s13045-021-01131-0 (2021).

36 Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell.* **159**, 1665-1680, doi:10.1016/j.cell.2014.11.021 (2014).

658 37 Zuin, J. *et al.* Cohesin and CTCF differentially affect chromatin architecture and gene
659 expression in human cells. *Proceedings of the National Academy of Sciences* **111**, 996-
660 1001, doi:10.1073/pnas.1317788111 (2014).