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 multiple sequencing- and imaging-based datasets, we identify long-range chromosomal skeleton loop structures that span over 100 Mb, extending beyond the reach of several existing DNA loop models. Further epigenetic and spatial density analyses point to assembly formation 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, the normal human genome is composed of 46 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 expression[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 have 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 five 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. Reports also exist on the existence of stable, long-range genomic interactions (>10 Mb) in large genomes of species including human[4, 9-11], and mechanisms such as liquid-liquid phase separation mediated by Polycomb and HP1 are known to facilitate some such long-range contacts[12, 13]. However, there is a lack of systematic investigations on the existence of sequence-specific, conserved, and stable locus-locus interactions >50 Mb and the physical principles governing them. Here, we address this gap by characterizing spatial colocalization of pairs of distant loci on chromosomes of several mammalian and non-mammalian species. We demonstrate the conservation of their spatial proximity across diverse cell lines and elucidate the physical mechanisms that likely enforce this stable association.

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 chromosome 14 (chr14) in human MCF10A cells[14]. While most genomic pairs followed the expected *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 distal locus pairs with high contact frequencies were also observed in chr2 and chr14 of human IMR90 cells[15] (Extended Data Fig. 1b). Extending this analysis to 11 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 one subpopulation of chromatin. Notably, these long-range interactions extend beyond typical intrachromosomal contacts such as enhancer-promoter interactions, super-enhancers, and TADs[12, 13]. Given the entropic cost of maintaining substantial colocalization probability over such long genomic distances, we sought to explore their underlying mechanisms and potential functional significance.

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-range locus pairs often overshadow weaker signals from long-range ones. To address this challenge, 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 (Extended Data Fig. 1d&e, details in Methods). It then compares locus pairs with similar genomic distances and identifies outliers exhibiting abnormally high contact frequencies, defined as locus pairs whose signals exceed two times the interquartile range above the third quartile 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[14]. CLOD successfully identified a 150 kb loop (Extended Data Fig. 1f), previously validated as a promoter-enhancer interaction of *PRIP1* using a 3C assay[16]. We then applied CLOD to a Hi-C dataset from the HCT116 cell line[17] 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[18-20] (Extended Data Table 1). Across these datasets, we identified multiple colocalized locus pairs separated by 50 Mb or more 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). To assess whether the colocalization is restricted to human chromosomes, we examined colocalizations in long (>150 Mb) chromosomes of other species, including six mammals (*Mus musculus* (mouse) [21, 22], *Rattus norvegicus* (rat), *Macaca mulatta* (rhesus monkey), *Bos taurus* (cattle), *Ovis aries* (sheep), *Oryctolagus cuniculus* (rabbit)) and *Gallus gallus* (chicken) [22]. Consistently, we identified colocalization of a substantial number of genomic locus pairs spanning genomic distances exceeding 50 Mb (Extended Data Fig. 4, Extended Data Table 1), suggesting this spatial organization is broadly conserved.

To determine whether the CLOD-identified long-range colocalization was transient or persistent, we focused on a colocalized genomic pair on human chr14 separated by approximately 80 Mb (Fig. 1h) [23]. 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 &k, 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. 1g).

MERFISH chromatin tracing data reveal 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 can neither distinguish whether colocalized pairs originate from the same or different homologous chromosomes, nor determine whether the observed contact frequencies reflect widespread interactions among cells or rare subpopulations. To cross-validate the prevalence of long-range pair colocalization and assess their distribution at the single-chromatin level, we analyzed previously published IMR90 chromatin tracing data obtained using DNA MERFISH[24].

As a technical control, we first re-analyzed the 3D distances (d_1) of MERFISH-labeled neighboring genomic loci across chr21 with a genomic separation $l_1 = 50\,kb$. The median distance for region chr21:32.45-33.35 Mb (Extended Data Fig. 5a) was found to be 256 nm, consistent with the previous report[24]. The overall distance distribution of neighboring genomic loci across chr21 followed a log-normal distribution, with a median distance of 325 nm (Extended Data Fig. 5b, left).

Next, we calculated the spatial distances of all neighboring genomic locus pairs across 2,991 copies of chr2 that were fluorescently traced with $l_1 = 250\,kb$ (Fig 2a, green line). The distribution also exhibited a log-normal pattern, with a median distance of 462 nm and a first quartile distance d_r = 299 nm (Fig. 2b, Extended Data Fig. 5b, right). For subsequent analyses, we adopted d_r as an intrinsic distance reference, classifying a genomic locus pair as in spatial proximity if their measured spatial distance $d \leq d_r$. This criterion aligns with previous studies using MERFISH data at an approximately 30 kb resolution, which employed a cutoff distance of ~ 200 nm to define chromatin interactions[25], and 500 nm for the chr21 data with 50 kb resolution to identify genomic pairs in proximity[24].

Most previous studies have focused on short-range chromosome interactions, such as enhancer-promoter interactions, which occur within genomic distances of ≤ 200 kb, typically in the 20 – 50 kb range[26, 27]. Few studies have specifically focused on long-range chromatin interactions. Here, we calculated the spatial distances between pairs of loci with genomic separations of $l_2 \in (50, 100]$ Mb and $l_3 > 100$ Mb, respectively (Fig. 2a, purple and orange lines). Compared to d_1, the distributions of corresponding d_2 and d_3 shifted significantly toward larger values, with median distances of 1.86 μm and 2.08 μm, respectively (Fig. 2b). Interestingly, even at such large genomic separations, certain locus pairs exhibited spatial distances < d_r (Fig. 2b, arrow). In the following analyses, we focused on the genomic locus pairs with $l_3 > 100$ Mb and $d_3 < d_r$, denoting them as long-range colocalizations (LRCs).

As expected, individual chr2 copies show diverse conformations (see examples shown in Fig. 2c, Extended Data Fig. 5c&d, Extended Data Movies 2-4), consistent with previous studies[25]. Moreover, a negative correlation exists between the total number of LRCs of an individual chromosome copy and its conformation measured by convex hull volume (Fig. 2d), suggesting a

153 potential relationship between LRCs and chromosome conformation. Notably, in these
154 conformations, LRCs either form scattered small clusters or join into larger ones, along with
155 varying compactness of the chromatin structures. In addition, LRC counts per chr2 copy range
156 from fewer than 10 to over 10,000, with a peak at ~ 600 per copy (Extended Data Fig. 5e). These
157 LRCs can be formed due to transient contacts through thermal fluctuations, or via stable structures.
158 While transient LRCs should appear infrequently among the entire cell population, stable ones are
159 expected to recur in subpopulations of chr2 copies. To test this hypothesis, we analyzed the
160 occurrence frequency of 155,009 unique LRCs across 2,991 chr2 copies (Fig. 2e). We identified
161 7,366 LRCs with occurrence frequencies beyond the 95% population interval, and defined them
162 as stable LRCs (sLRCs) (Fig. 2e & f). Each sLRC appeared in at least 23 chr2 copies (~0.8% of
163 the dataset), with some found in over 10% of samples. That is, sLRCs exhibited a significantly
164 higher probability of occurring within 299 nm compared to other LRCs (Fig. 2g).

165 Some sLRC detected via MERFISH were also detected with CLOD-identified colocalized pairs
166 (> 100 Mb separation) from Hi-C (Fig. 2h). Note that Hi-C and MERFISH detect proximity
167 differently: Hi-C captures direct contact, whereas MERFISH assesses spatial distances, using a
168 299 nm threshold here. Technical limitations also contribute to discrepancies — Hi-C may miss
169 repetitive or low-accessibility regions, whereas its much larger sample size (~1 million cells) could
170 detect rare configurations potentially overlooked by MERFISH (with ~3,000 chr2 copies). Despite
171 these differences, the strong agreement between Hi-C and MERFISH data reveals that stable LRCs
172 are prevalent in chromatin structures.

173 To further validate these findings, we analyzed the locus-pair contact frequencies of IMR90 chr2
174 from a micro-C dataset[28] and compared them with MERFISH results. Compared to standard Hi-C,
175 the superior coverage per genomic bin of micro-C[29] provides sufficient signal-to-noise ratios for
176 direct analysis of raw read counts without using additional algorithms such as CLOD. When
177 considering micro-C data with contact counts >10 as colocalization pairs, 556 sLRCs that were
178 identified from MERFISH data are also present in micro-C data (Extended Data Fig. 5e, left). Even
179 when applying stringent filtering criteria (contact counts >20), we still identified a subset of LRCs
180 present in both micro-C and MERFISH (Extended Data Fig. 5e, right). Therefore, both
181 sequencing-based and imaging-based approaches consistently identified sLRCs in IMR90 chr2. In
182 addition, LRCs and sLRCs were also present in other long chromosomes of IMR90 cells (Extended
183 Data Fig. 5f).

184 To assess whether LRCs exist at the single-cell level in other species, we examined a DNA
185 seqFISH+ dataset of chr1 of mouse NMuMg cells (Fig. 2i) [30]. Applying the same analytical
186 framework used for MERFISH, we measured spatial distance distributions for genomic loci in the
187 seqFISH+ data separated by 25 kb, 250 kb, and over 100 Mb, respectively. Those with 25 kb have
188 the smallest median 3D distance at 253 nm. Notably, the first quartile of spatial distances for loci
189 250 kb apart was 281 nm in seqFISH+, closely matching the 299 nm threshold from MERFISH
190 data (Extended Data Fig. 5g). We therefore adopted a universal LRC definition across platforms:
191 loci with genomic separation >100 Mb and spatial proximity <299 nm. The seqFISH+ dataset

192 provided substantially lower genomic coverage per chromosome than MERFISH, with most long-
193 range loci pairs detected in only ~ 200 – 300 chr1 copies out of 617 or more copies (Extended Data
194 Fig. 5h). To identify robust colocalization, we defined stable LRCs (sLRCs) as those present in at
195 least 10, or ~5% of, individual chromosomes (Extended Data Fig. 5h&i). All these results further
196 demonstrate that colocalization of genomically distant loci is a universal phenomenon in long
197 chromosomes, irrespective of species origin or chromosome identity.

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

199 Our structural analyses revealed that sLRC-associated genomic loci form intertwined networks
200 with sLRCs acting as links (Fig. 2c&i, Extended Data Fig. 5c&d, Extended Data Movies S5-S8).
201 Linkage maps from Hi-C, micro-C, and MERFISH data showed that certain genomic loci
202 colocalize with multiple distant loci (Fig. 2f&h Extended Data Fig. 5e). These loci likely function
203 as NCs, linking multiple sLRCs and forming the network (Fig. 3a). Some specific locations across
204 chr2, such as regions at ~28 – 40 Mb and the ends of IMR90 chr2, spatially colocalized with
205 significantly more well-separated cis-loci compared to other loci on the same chromosome and
206 thus can serve as NCs (Fig. 3b, left & Fig. 3c). Notably, locus 242 Mb (NC 242 Mb) were
207 associated with the most sLRCs and was analyzed separately for clear visualization.

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

219 To reveal the basic patterns of chromosome conformation composed of NCs, we represented
220 sLRCs as networks. As expected, the results showed that distant regions on chr2 clustered via
221 sLRCs, where NCs served as hubs in the network (with two representative chr2 copies shown in
222 Fig. 3g), and most individual chr2 copies had fewer than 30 isolated clusters (Fig. 3h). Further
223 statistical analyses revealed that most NC-coordinated sLRCs clusters (59.2%) contained only 2 –
224 5 loci, while 5.4% had over 100 loci as nodes (Fig. 3i). Detailed analyses revealed distinct
225 structural motifs (Fig. 3j): some NCs formed isolated single-NC clusters, while others coalesced
226 into multi-NC hubs via direct colocalization or shared pairing loci, forming clusters with 100 –
227 780 nodes.

228 To confirm that NCs engage in multiple LRCs across diverse experimental platforms and cellular
229 contexts, we further analyzed the 3D genome conformation of other cell lines from human and

other species captured by other techniques, including K562 and GM12878 by Linking mRNA to Chromatin Architecture (LiMCA, Extended Data Fig. 6a) [31], HEK293T by single-cell split-pool recognition of interactions by tag extension (scSPRITE, Extended Data Fig. 6b) [32], human sperm Genome Architecture Mapping (GAM, Extended Data Fig. 6c) [33], IMR90 by micro-C (Extended Data Fig. 6d), and NMuMg by seqFISH+ (Fig. 2i). Hub-like loci appeared in all these cell types and across different 3D architecture detection techniques. Moreover, some NCs are conserved in different cell types. For example, loci ~32-35 Mb associated with multiple sLRCs were identified both in IMR90 by MERFISH and micro-C, and in K562 and GM12878 by LiMCA. Loci around the 3'-end of chr2 form another NC in IMR90 by MERFISH, human sperm by GAM, and HEK293T by scSPRITE.

These findings suggest 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 MERFISH data revealed the widespread presence of sLRCs ($p \geq 0.8\%$) in IMR90 chr2 copies, with some occurring in 5 – 10% of all copies (Fig. 2e). Notably, the genomic distributions of the two ends of sLRCs are not uniform but clustered along the genome (Fig. 2f). Statistical analysis confirmed that most sLRC-associated loci have their nearest neighboring sLRC genomic loci within 500 kb (Extended Data Fig. 7a). Here we defined the genomic proximity distance between two sLRC pairs a (loci a_1, a_2) and b (loci at b_1, b_2) as $l_{ab} = max\ ((|a_1 - b_1|, |a_1 - b_2|), (|a_2 - b_1|, |a_2 - b_2|))$ (Fig. 4b, Extended Data Fig. 7b). Therefore, we grouped sLRCs into metaLRCs, where each sLRC had at least one other sLRC within 250 kb at both ends (Fig 4c). This process grouped 72% of sLRCs into 736 metaLRCs, each containing 2 to over 40 sLRCs (Extended Data Fig. 7c).

We hypothesized that neighboring sLRCs stabilize each other, as revealed from the live-cell imaging studies (Fig. 1i–l). To examine this hypothesis, we analyzed two classes of minimal biophysical models with and without cooperativity between sLRCs within a metaLRC (Methods). In both models, colocalization of an sLRC pair is assumed to be held by some interactions, e.g., mediated by proteins or noncoding RNAs. Consequently, a pair exists in either a bound or unbound state, distinguished by whether the two loci colocalize. In the absence of other sLRC pairs, an isolated sLRC pair stochastically transits between the two states with association (α) and dissociation (γ) rate constants (Fig. 4a, left). For simplicity, assuming that the N sLRC pairs within a metaLRC have the same dynamics and transit independently, one can estimate from the data $\frac{\alpha}{\gamma} = \frac{p_1}{Np_0}$, where p_0 and p_1 are the frequencies of observing zero or one bound sLRC, respectively. The probability of observing $i > 1$ bound sLRC within the metaLRC can then be predicted without any adjustable parameter to be $p_i = \frac{(N-i+1)\alpha}{i\gamma} p_{i-1}$.

The cooperative model assumes that the transition dynamics of neighboring sLRCs are not independent due to a multivalent binding mechanism. In the presence of other bound sLRC pairs

267 within the metaLRC, the two loci of an unbound sLRC pair remain spatially close, leading to an
268 increased effective association rate constant. For simplicity, assuming that the effective
269 enhancement δ is the same in the presence of one or more neighboring bounded sLRC pairs, then
270 its value can be determined by fitting the measured distribution of bound sLRCs, p_i vs. i for i > 1.
271 Consequently the metaLRC as a whole would appear as an effective two-conformation system
272 with no or at least one bound sLRC, respectively (Fig. 4a, middle & right).

273 We tested the predictions of the two models against the MERFISH data. Figure 4d depicted one
274 representative metaLRC. Correlation analyses over MERFISH data confirmed the cooperativity
275 between sLRCs within the metaLRC (Fig. 4e). Note that some sLRCs have higher correlations
276 than others within the same metaLRC, indicating that the assumption of a single value of δ is only
277 crudely valid. The model without cooperativity significantly underestimated the probabilities of
278 p_i for i >1, while a single parameter $\delta > 0$ significantly improves the agreement between the
279 experimental data and theoretical predictions. Two additional examples shown in the Extended
280 Data Fig. 5d&e also support the cooperative model. The bimodal distribution in Extended Data
281 Fig. 5d reveals an even stronger cooperativity than the simple model we considered, suggesting
282 that the enhancement δ increases with the number of bounded sLRCs.

283 Analysis of individual chr2 copies further supported the cooperative model. The spatial
284 distances between two tagged genomically distant regions in bound form were much shorter
285 compared to those in free form chr2 copies (Fig. 4g, Extended Data Movies 9–12). When one or
286 more sLRCs were bound, the two genomic regions within the metaLRC were held closer, while
287 additional bound sLRCs (> 1) did not significantly alter spatial distance (Fig. 4h). The observed
288 non-monotonic p_i vs. i curves for certain representative metaLRCs (e.g., 33.75–35.5 Mb & 146–
289 146.25 Mb; 119.25–123.75 Mb & 242 Mb) suggested even stronger cooperativity than the minimal
290 model predicted (see Methods) (Extended Data Fig. 7d&e).

291 Next, we examined all metaLRCs of chr2 with $N > 2$. The α/γ values directly obtained from the
292 FISH data were < 0.007 (Fig. 4i), indicating that associating two distant loci is entropically
293 unfavorable. However, the corresponding $(\alpha + \delta)/\gamma$ values were up to $10 - 100\times$ higher, making
294 the probability of observing a bound metaLRC over 1 – 10% (Extended Data Fig. 7f). That is,
295 pairs of distant loci can remain stably proximate within a chromosome via neighbor-assisted
296 multivalent binding, significantly enhancing structural stability and long-range genomic
297 interactions.

298 **Structural analyses reveal multiple candidate mechanisms for sLRC formation.**

299 Colocalization of distant cis-loci has been intermittently observed. Three general mechanisms have
300 been proposed to explain LRC formation (Fig. 5a, I–III): (I) A/B compartment, or loose/compact
301 region segregation, may bring distant B compartment loci into transient proximity[34] or relatively
302 stable colocalization via polycomb structures[11]; (II) Transcriptionally active loci, around which
303 chromatin is typically loose and accessible, may colocalize via attaching to shared nuclear
304 structures like speckles or nucleoli[35]; (III) The nuclear membrane may recruit multiple loci into

305 spatial proximity[36]. Alternatively, one may conceive mechanism (IV) as stable sLRC formation
306 independent of major nuclear structures or compartment segregation. While the first three have
307 been proposed previously[37], sLRC assemblies formed through Mechanism IV were considered
308 rare.

309 sLRC-forming loci are not significantly enriched in protein-coding gene-rich or gene-rare
310 regions (Extended Data Fig. 8a, top) nor in lncRNA-coding genes (Extended Data Fig. 8a, bottom).
311 Examination of A/B compartment identities among sLRC-forming loci revealed no clear pattern
312 or correlation between the two loci of each sLRC pair (Extended Data Fig. 8b–d), while the
313 majority of sLRCs have two ends at the chomatin loose region (Extended Data Fig. 8c, bottom),
314 which is consistent with the observation from a previous study[24]. Note that the A/B compartment
315 was calculated from bulk-level data. That is, an ensemble of chromosomes may be composed of
316 subpopulations with and without the presence of an sLRC under study; thus, results using such
317 bulk-averaged data are not conclusive. Bulk averaging also prevented us from concluding whether
318 sLRCs or the loci linked to sLRCs have specific epigenetic patterns (Extended Data Fig. 8e&f).

319 To connect the sLRC and associated loci to epigenetic marks and nuclear structures at the single-
320 chromosome level, we investigated the seqFISH+ dataset[30]. This dataset with single chromosome
321 resolution combines spatial coordinate information of each location on the mouse genome with its
322 epigenetic modification marks, such as H3K4me2, H3K27ac, H3K9me3, H3K27me3, etc., as well
323 as marks for nuclear structures, such as Polycomb (EZH2), speckles (SF3A66), and lamina-
324 associated domains (Lamin B1). To be consistent with the resolution of MERFISH, we grouped
325 the DNA seqFISH+ data to a 250 kb resolution, calculated the median mark level, and used the 3rd
326 quantile as the gate to identify whether a region has a high level of the targeted marker (Extended
327 Data Fig. 9a).

328 The epigenetic profiling reveals subsets of sLRCs exhibiting epigenomic features consistent
329 with the three identified formation mechanisms (Mechanisms I–III, Fig. 5a). For instance, among
330 the 323 detected copies of the 3.25 and 125.15 Mb locus pair on NMuMg chr1, 15 colocalized
331 copies exhibited elevated H3K9me3 and uniformly depleted euchromatin marks (H3K4me2 and
332 H3K27ac) relative to non-colocalized copies (Extended Data Fig. 9b-i). Another pair (15.25 and
333 151.25 Mb) displayed high H3K27me3 enrichment along with the Polycomb marker EZH2
334 (Extended Data Fig. 9b–ii). Both examples are consistent with compressed encounters driven by
335 A/B compartment segregation (Mechanism I).

336 A colocalized pair in Extended Data Fig. 9c instead showed high levels of active chromatin
337 markers (H3K4me2 and H3K27ac), enrichment of the nuclear speckle marker SF3A66, and low
338 H3K9me3 and H3K27me3, consistent with nuclear speckle–associated assembly (Mechanism II).
339 Extended Data Fig. 9d illustrates an example potentially following Mechanism III, with high
340 Lamin B1 enrichment indicating association with lamina-associated domains.

341 Notably, some sLRCs on mouse chr1 could not be assigned to these categories: the regional
342 epigenetic mark levels of the two ends of an sLRC show no strong correlation (Extended Data Fig.
343 9e–i); or all epigenetic mark levels of both ends are relatively low or show no clear pattern

344 (Extended Data Fig. 9e–ii). These observations suggest additional formation mechanisms,
345 collectively classified as Mechanism IV.

346 The MERFISH data contain no epigenetic information at single-chromosome level. Instead, we
347 examined the local density of a locus, noting that the two ends of an sLRC formed through
348 Mechanism I-III are expected to share similar local packing structures. Specifically, we quantified
349 local geometric structures within 299 nm of a locus by defining regional density of short-range (<
350 10 Mb) neighbors (RD-SN) and regional density heterogeneity (RDH), which measures the
351 uniformity of local spatial distributions (Fig. 5b). Comparing RD-SN to A/B compartments
352 identified from Hi-C data, an A compartment has a higher probability of having a loose
353 neighborhood than a B compartment (Extended Data Fig. 8g). Then one expects that Mechanism
354 I and III lead to high RD-SN for both loci associated with an sLRC, Mechanism II leads to low-
355 low RD-SN, while the pattern for Mechanism IV is undetermined. Furthermore, a locus next to a
356 nuclear structure (Mechanism II and III) is expected to exhibit high RDH due to the excluded
357 volume of the latter. Therefore, a combined RD-SN and RDH analysis can identify sLRCs not
358 convincingly explainable by Mechanisms I-III, e.g., those with both RD-SN and RDH low on both
359 loci. Indeed, sLRCs show combinatorial patterns of these two quantities, implying existence of
360 diverse mechanisms to sLRC formation (Fig. 5c&d), as also reflected from examining
361 representative structures (Fig. 5e).

362 Furthermore, we observed spatial colocalization of *cis*-loci that are well separated on the genome
363 in chromosomes shorter than 150 Mb. Specifically, we analyzed the colocalization state of locus
364 pairs in chromosome 21 of IMR90 cells at 20 Mb, which is about half of the region (~36.3 Mb)
365 labeled with MERFISH. The distribution of spatial distances between most locus pairs longer than
366 20 Mb in chr21 ranged between 0 and 5 μm, and colocalized pairs of loci denote those whose
367 spatial distances are closer than 299 nm in individual chromosome structures as aforementioned
368 (Extended Data Fig. 8h). As expected, both linkage maps and 3D structure plots show the existence
369 and variety of LRCs in individual chr21 (Fig. 5f). While most of 32,341 loci pairs longer than 20
370 Mb were only observed in fewer than 200 out of 7,591 chr21 copies, 1,718 or 5% of them appeared
371 in more than 325 copies, which were identified as stable colocalization locus pairs here (Fig. 5g).
372 Moreover, similar to chr2, a linkage map showing the stable colocalization loci pairs reveals that
373 clustering of sLRCs and loci serving as nucleation centers also exists in chr21 (Fig. 5h).

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

377

378 ## DISCUSSION

379 Our analyses of Hi-C and DNA FISH chromosome tracing data reveal widespread colocalization
380 of genomic loci separated by distances far exceeding a typical TAD. It is important to note that
381 our study employed stringent criteria to identify representative examples rather than exhaustively

382 cataloging all stable long-range colocalizations. While we focused on loci separated by >100 Mb,
383 similar structures exist at shorter genomic distances, particularly in smaller chromosomes such as
384 chr21 (~48 Mb).

385 Statistically, the probability of multiple genomic loci with $l \geq 100$ Mb being in physical
386 proximity without stable structural mechanisms would be negligible. Some sLRCs may form
387 through known mechanisms such as compartment segregation or anchoring to nuclear structures
388 such as nucleoli, speckles, or the nuclear envelope. However, many sLRCs and NCs are buried in
389 the nuclear interior, show no signature of compartment segregation, and are away from large
390 nuclear structures, suggesting additional nuclear structure-independent mechanisms. generically
391 assigned to Mechanism IV.

392 The molecular nature of the interactions holding the assemblies needs further investigation. An
393 assembly can be liquid-, gel-, or solid-like, similar to what is observed in other contexts. The
394 existence of NCs suggests that they may play roles in LRC formation different from their
395 partnering loci. These NC sites can either be nucleation centers for liquid droplet formation or act
396 as scaffolds for solid assembly mediated by elements such as non-coding RNAs or bridging
397 proteins. For instance, the human chr2 242 Mb locus is enriched in repetitive sequences, which
398 may serve as selective binding sites for partnering loci, either directly or through other bridging
399 factors. Another 32–35 Mb locus on human chr2 is also rich in repetitive sequences and non-
400 coding RNAs such as LIN1[38]. Multi-way chromosome assemblies have been identified in different
401 contexts, such as enhancer-promoter hubs[39], superenhancers, and the multi-chromosome enhancer
402 clusters controlling monoallelic olfactory receptor expression in olfactory sensory neurons[40-42].
403 The assemblies of sLRCs and metaLRCs may form under similar molecular mechanisms. One
404 known example is that Polycomb repressive complex 2 uses RNA binding to coordinate long-
405 range chromosome loop formation and H3K27me3 spreading[42]. Falk et al. argue that attractions
406 between heterochromatic regions and lamina-chromatin interactions, i.e., Mechanisms I and III,
407 are critical to recapitulate compartmented chromatin structures[43]. Analyses in this work reveal
408 existence of long-range colocalization structures showing no clear features of lamina-chromatin
409 interactions and heterochromatin, and their molecular mechanism(s) of formation need further
410 investigation.

411 A key unsolved question is how sLRC loci locate and associate with each other. One plausible
412 mechanism is that, during post-mitotic DNA decondensation, some genomically distant loci may
413 be in spatial proximity and associate stochastically, leading to subsequent formation of additional
414 sLRCs and stabilization of the assembled structures. Liquid-phase droplet fusion can also drive
415 two regions into proximity to form an sLRC[44].

416 Besides their varied formation mechanisms, these sLRC structures likely play diverse functional
417 roles, including transcriptional regulation and broader chromosome 3D organization. Proper DNA
418 folding into spatially segregated territories with minimal entanglement and knot formation is
419 essential for eukaryotic genome organization and gene regulation. This process must be robust
420 across all cell types, including those with abnormal karyotypes, and must function despite cell

421 type-specific euchromatin/heterochromatin partitioning, folding stochasticity, and mutations.
422 From an engineering perspective, achieving this through a single centralized mechanism requiring
423 precise tuning would be challenging. Instead, our findings suggest a redundant, distributed
424 component mechanism that facilitates chromosome folding into compact structures (Fig. 5i). The
425 genome harbors an extensive repertoire of NCs and partnering loci separated by long genomic
426 distances, from which only a subset is available for a given cell type. At the individual chromosome
427 level, subsets of these loci stochastically form sLRCs and larger structures, contributing to the
428 observed heterogeneity of chromosome structures[45]. These structures provide a "divide-and-
429 conquer" mechanism to facilitate forming chromosome structures at smaller length scales, e.g., by
430 forming segregated globally compact chromosome configurations, and closed-end boundary
431 conditions for simultaneous crumpled folding at multiple regions (Extended Data Fig. 4d) [5], and
432 enhanced contact frequencies for forming loops at smaller scales.

433 In summary, our analyses of chromosome tracing and Hi-C data uncover prevalent, conserved
434 large-sized loops and their clustering into stable structures. These findings raise important
435 questions regarding the molecular mechanisms, temporal dynamics, and functional significance of
436 these structures on chromosome configurations, gene transcriptional activity, and cell type
437 regulation.

438

439

References

1. Dill, K.A. & MacCallum, J.L. The protein-folding problem, 50 years on. *science* **338**, 1042-1046 (2012).
2. Bolzer, A. *et al.* Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS biology* **3**, e157 (2005).
3. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science* **326**, 289-293 (2009).
4. Wang, S. *et al.* Spatial organization of chromatin domains and compartments in single chromosomes. *Biophysical Journal* **112**, 217a (2017).
5. Grosberg, A., Rabin, Y., Havlin, S. & Neer, A. Crumpled globule model of the three-dimensional structure of DNA. *Europhysics letters* **23**, 373 (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 (2014).
7. Polovnikov, K.E. *et al.* Crumpled polymer with loops recapitulates key features of chromosome organization. *Physical Review X* **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 (2009).
9. Quinodoz, S.A. *et al.* Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. *Cell* **174**, 744-757. e724 (2018).
10. Jäger, R. *et al.* Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nature communications* **6**, 6178 (2015).
11. Kraft, K. *et al.* Polycomb-mediated genome architecture enables long-range spreading of H3K27 methylation. *Proceedings of the National Academy of Sciences* **119**, e2201883119 (2022).
12. Pombo, A. & Dillon, N. Three-dimensional genome architecture: players and mechanisms. *Nature reviews Molecular cell biology* **16**, 245-257 (2015).
13. Whyte, W.A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307-319 (2013).
14. 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 (2015).
15. Rao, S.S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680 (2014).
16. Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290-294 (2013).
17. Rao, S.S. *et al.* Cohesin loss eliminates all loop domains. *Cell* **171**, 305-320. e324 (2017).
18. Dixon, J.R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331-336 (2015).
19. 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 (2019).
20. Ren, B. *et al.* High-resolution Hi-C maps highlight multiscale 3D epigenome reprogramming during pancreatic cancer metastasis. *Journal of Hematology & Oncology* **14**, 120 (2021).
21. Liu, M. *et al.* Tracing the evolution of single-cell cancer 3D genomes: an atlas for cancer gene discovery. *bioRxiv*, 2023.2007. 2023.550157 (2024).

22. Li, D. *et al.* Comparative 3D genome architecture in vertebrates. *BMC biology* **20**, 99 (2022).
23. Zuin, J. *et al.* Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proceedings of the National Academy of Sciences* **111**, 996-1001 (2014).
24. 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).
25. Bintu, B. *et al.* Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science* **362**, eaau1783 (2018).
26. 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).
27. Furlong, E.E. & Levine, M. Developmental enhancers and chromosome topology. *Science* **361**, 1341-1345 (2018).
28. Palikyras, S. *et al.* Rapid and synchronous chemical induction of replicative‐like senescence via a small molecule inhibitor. *Aging Cell* **23**, e14083 (2024).
29. Krietenstein, N. *et al.* Ultrastructural details of mammalian chromosome architecture. *Molecular cell* **78**, 554-565. e557 (2020).
30. Takei, Y. *et al.* Spatial multi-omics reveals cell-type-specific nuclear compartments. *Nature* **641**, 1037-1047 (2025).
31. Wu, H. *et al.* Simultaneous single-cell three-dimensional genome and gene expression profiling uncovers dynamic enhancer connectivity underlying olfactory receptor choice. *Nature Methods* **21**, 974-982 (2024).
32. Arrastia, M.V. *et al.* A single-cell method to map higher-order 3D genome organization in thousands of individual cells reveals structural heterogeneity in mouse ES cells. *bioRxiv*, 2020.2008.2011.242081 (2020).
33. Irastorza-Azcarate, I. *et al.* Extensive folding variability between homologous chromosomes in mammalian cells. *Molecular Systems Biology* **21**, 735-775 (2025).
34. Gibson, B.A. *et al.* Organization of chromatin by intrinsic and regulated phase separation. *Cell* **179**, 470-484. e421 (2019).
35. 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 (2019).
36. Van Steensel, B. & Belmont, A.S. Lamina-associated domains: links with chromosome architecture, heterochromatin, and gene repression. *Cell* **169**, 780-791 (2017).
37. Dekker, J. & Mirny, L.A. The chromosome folding problem and how cells solve it. *Cell* **187**, 6424-6450 (2024).
38. Percharde, M. *et al.* A LINE1-Nucleolin Partnership Regulates Early Development and ESC Identity. *Cell* **174**, 391-405.e319 (2018).
39. Lando, D. *et al.* Enhancer-promoter interactions are reconfigured through the formation of long-range multiway hubs as mouse ES cells exit pluripotency. *Molecular Cell* **84**, 1406-1421.e1408 (2024).
40. 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* **113**, E2889-E2898 (2016).

532 41. Pourmorady, A. & Lomvardas, S. Olfactory receptor choice: a case study for gene
533 regulation in a multi-enhancer system. *Current Opinion in Genetics & Development* **72**,
534 101-109 (2022).
535 42. Kraft, K. *et al.* Polycomb-mediated genome architecture enables long-range spreading of
536 H3K27 methylation. *Proceedings of the National Academy of Sciences* **119**,
537 e2201883119 (2022).
538 43. Falk, M. *et al.* Heterochromatin drives compartmentalization of inverted and
539 conventional nuclei. *Nature* **570**, 395-399 (2019).
540 44. Strom, A.R. *et al.* Phase separation drives heterochromatin domain formation. *Nature*
541 **547**, 241-245 (2017).
542 45. Finn, E.H. *et al.* Extensive heterogeneity and intrinsic variation in spatial genome
543 organization. *Cell* **176**, 1502-1515. e1510 (2019).
544
545

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, or the top 5% in IMR90 chr1, chr3 and chr4 MERFISH dataset, or ⩾15 chromosomes (~7% in detected chromosomes) in NmuMg chr1 DNA seqFISH+.

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: Regional density heterogeneity.

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

Funding:

National Institute of General Medical Sciences R01GM148525 (JX)

The Charles E. Kaufman Fund of the Pittsburgh Foundation KA2018-98550 (JX)

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.

Figure captions

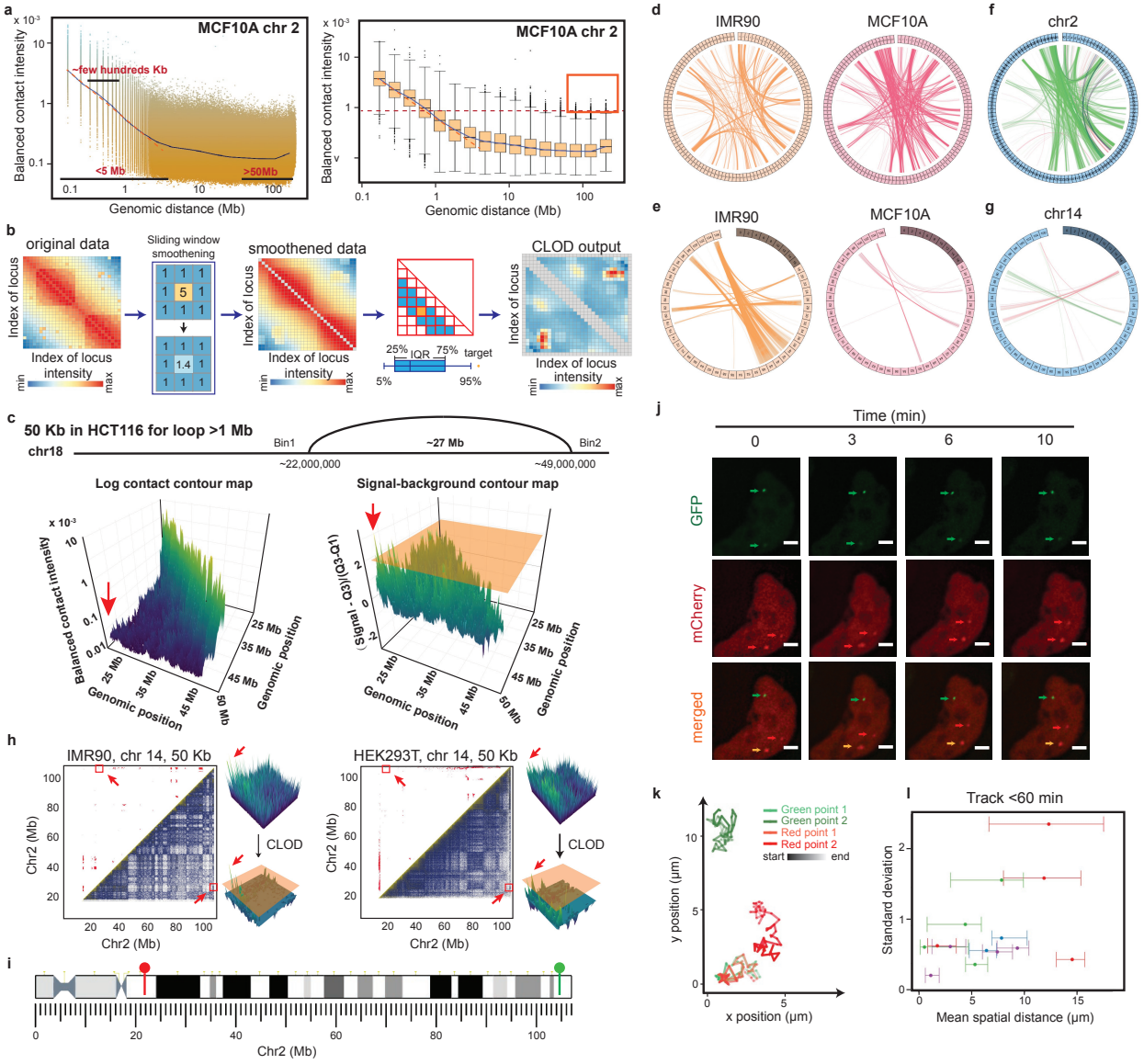

Fig. 1 Hi-C data reveals prevalent existence of colocalized pairs of long-range genomic loci. **a**, Scatter plot (left) and boxplot (right) showing the decay trend (blue line) of balanced contact intensity with increasing genomic distance. Red boxed points represent locus pairs > 50 Mb apart with contact intensity 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. Left: Raw contact map between the two segments. Right: $\frac{Signal-Q3}{Q3-Q1}$ contour map of the same region as the left one. **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

588 cell types: four embryonic cell lines (green), four normal differentiated cell lines (pink), and four
589 cancer cell lines (blue). **h**, CLOD-based identification of colocalized pairs on chr14 from Hi-C
590 data of IMR90 and HEK293T cells at 50 kb resolution. **i**, CRISPR-dCas9 two-color labeling sites
591 on chr14. **j**, Live-cell imaging snapshots of a HEK293T cell labeled with RFP and GFP. Scale bar:
592 2 μm. **k**, 10-min time-resolved trajectories of blue and red puncta in the labeled cell from panel j.
593 **l**, Distance fluctuations between blue and red puncta pairs in four cells that were tracked for 60
594 min. The left and right bars indicate the maximum and minimum distances between a pair of puncta.

595

596

597

598

599

600

601

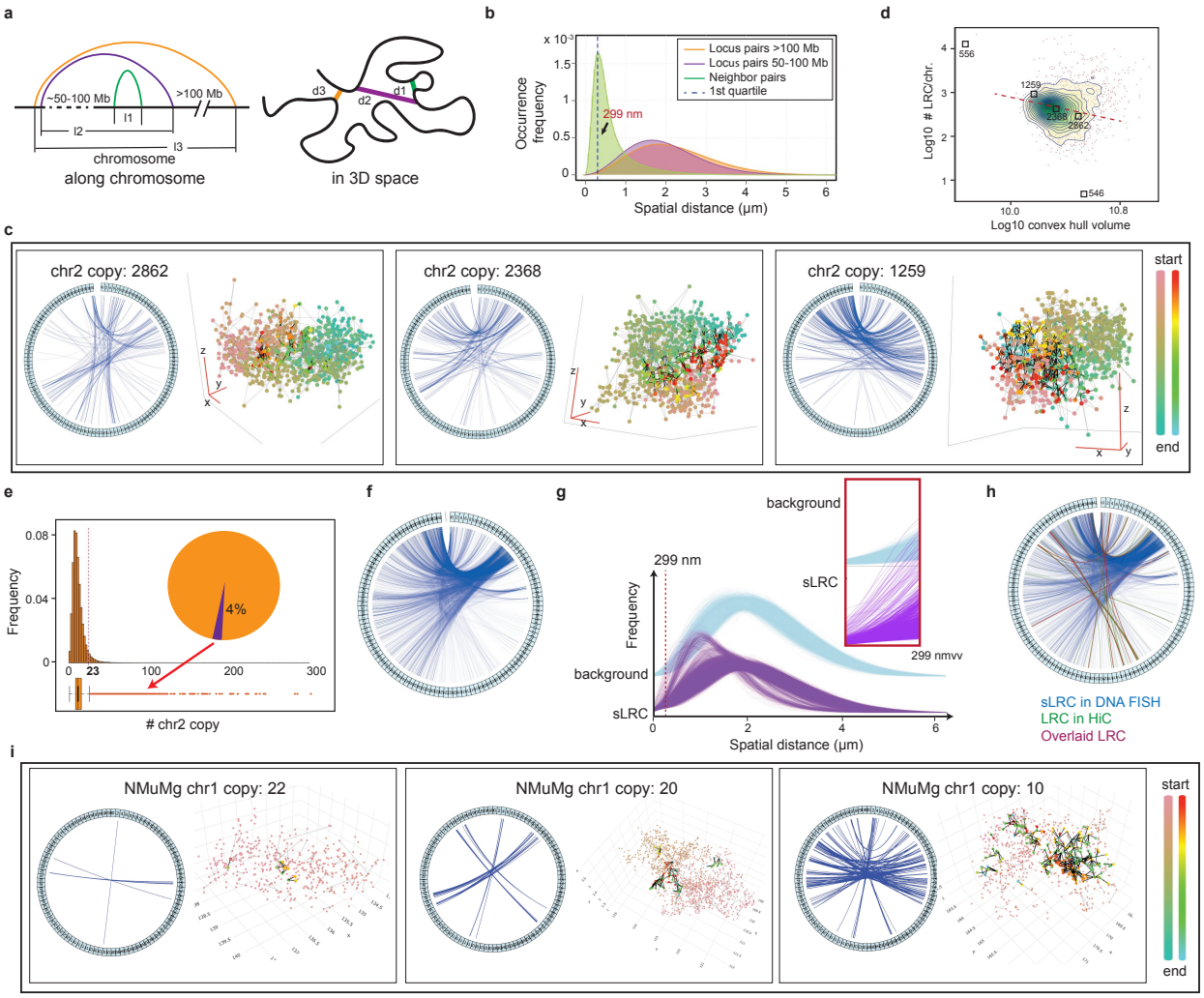

Fig. 2 MERFISH chromatin tracing reveals widespread long-range colocalization (LRC). a, Schematic illustration of genomic pairs and their genomic versus spatial distances. **b**, Density curves of spatial distance distributions for loci pairs with genomic separations of 250 kb, 50-100 Mb, and >100 Mb. 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). The pie diagram shows the percentage of LRCs classified as 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 MERFISH (blue lines) and Hi-C (green lines) data. Overlapping links were highlighted in brown.

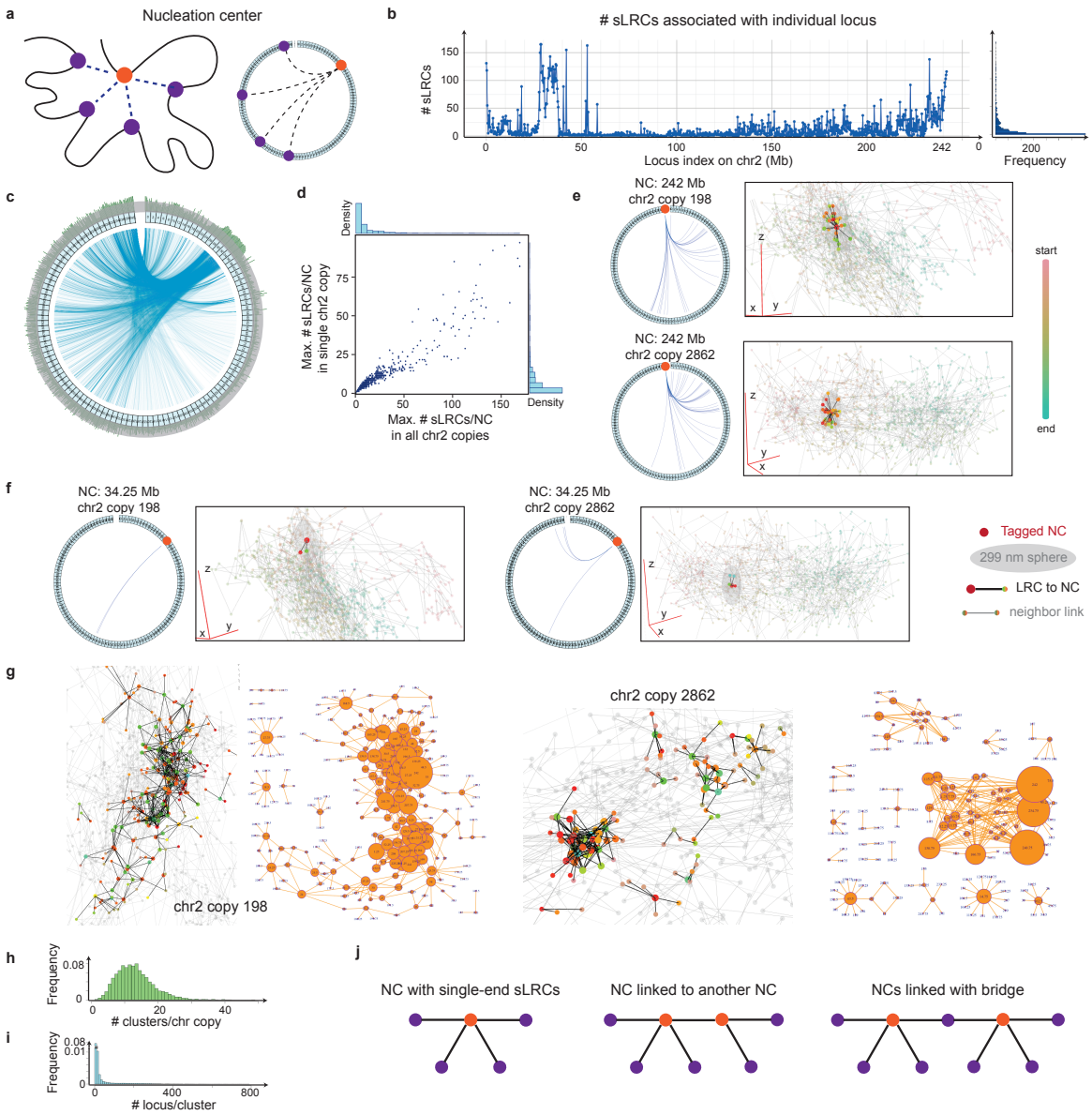

Fig. 3 Nucleation centers (NC) coordinate clustering of sLRCs. a, Left: Schematic illustration of a nucleation center (NC) (orange) forming spatial proximity with multiple partner LRC loci (purple). Right: Ring plot depicting the same interaction network. **b**, Left: Line plot showing the number of 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 sLRCs 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 highlighting two representative NCs (NC 242 Mb and NC 34.25 Mb) on chr2, along with their

associated sLRCs. **g**, 3D structural representations and network maps showing connections among NC loci in two representative chr2 copies. **h**, Histogram displaying the number of locus clusters – defined as sets of loci connected to at least one other locus – per individual chr2 copy. **i**, Histogram depicting the number of loci per locus-cluster, as identified in panel h. **j**, Diagrams illustrating single NC elements and various types of connections between NCs.

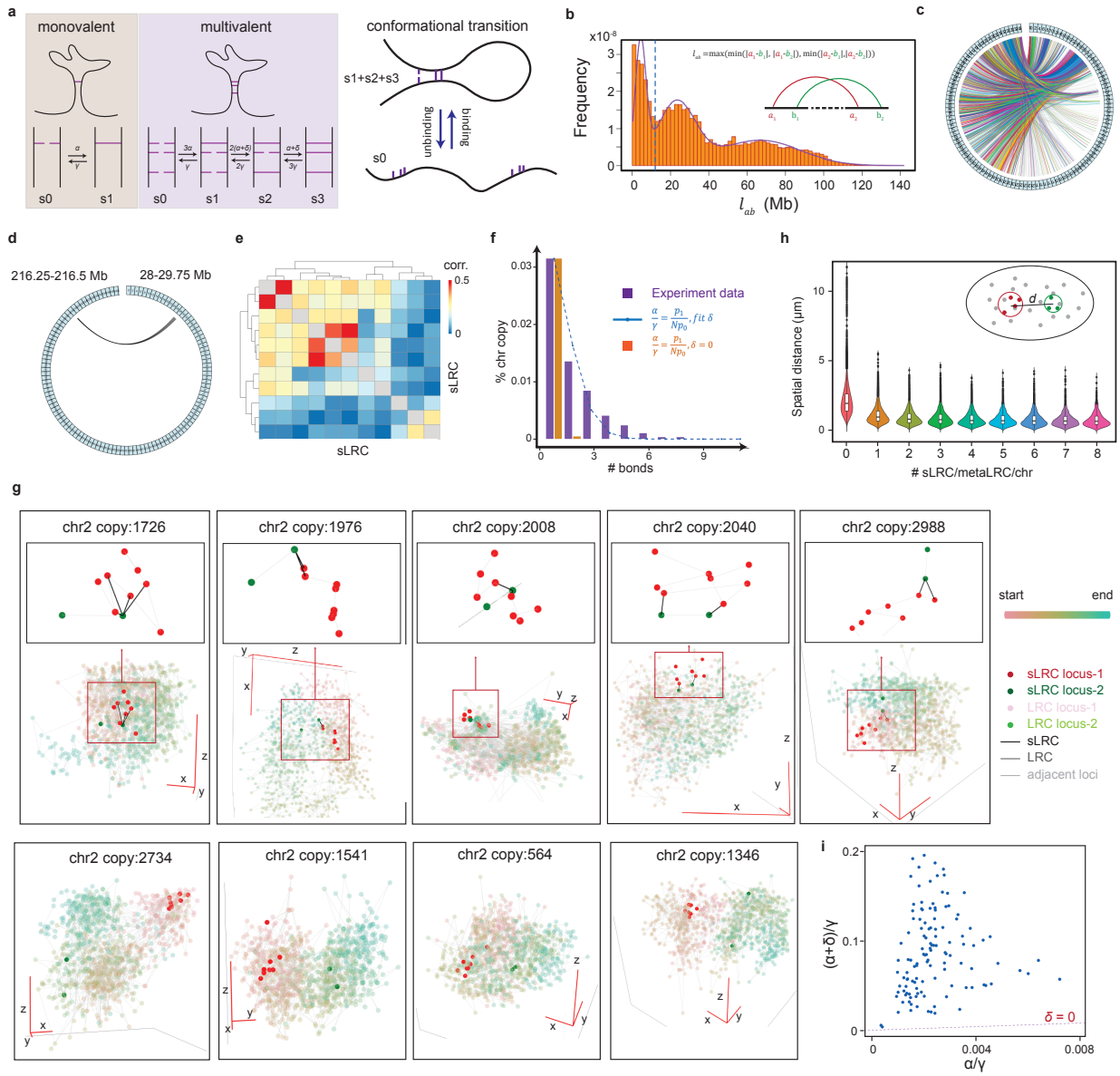

Fig. 4 DNA FISH data analysis reveals that multivalent binding stabilizes sLRC clusters. a, Left: Schematic illustration of monovalent binding vs. multivalent binding models. α: association rate constant of a single pair in the absence of other bound sLRCs; γ: dissociation rate constant of a single pair; 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 sLRCs within the metaLRC (purple bar) vs. predictions based on an independent sLRC binding

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

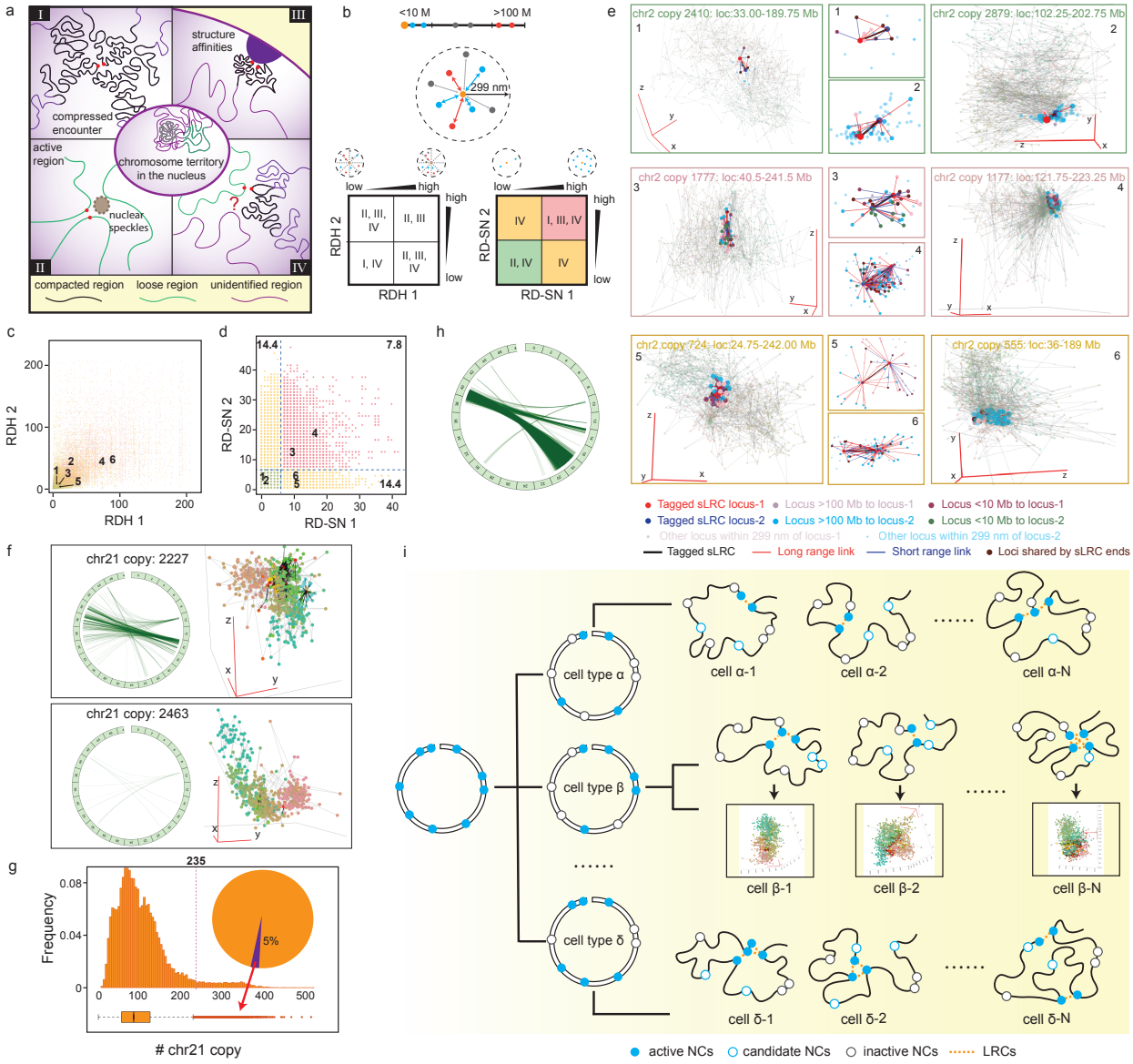

Fig. 5 Possible mechanisms of sLRC formation and functional roles on chromosome folding. **a**, Schematic of four possible mechanisms for sLRC formation. I & II: A/B compartment segregation leads to compressed encounters of genomically distant loci in the B and A compartments, respectively. III: Colocalization 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-range 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

676 counts (n = 7,591) containing LRCs. LRCs present in >325 copies (red dash line) are classified as
677 stable LRCs (sLRCs). **h**, Identification of chr2 sLRCs from DNA FISH data. **i**, Schematic
678 illustration of the proposed excessively redundant distributed component mechanism for
679 chromosome folding.

680

681

682

683

684

Methods

Hi-C data analysis

Raw Hi-C data of human were downloaded from GEO, as shown in Extended Data Table 1, and were aligned to the human genome *hg38* using the Burrows-Wheeler Alignment tool[46]. Hi-C contacts were then identified and filtered with Pairtools[47]. The contact matrices of balanced contact intensity used for the CLOD algorithm were created with Cooler[48]. Hi-C counts of other species were downloaded from GEO directly (Extended Data Table 1).

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

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

To test the impact of Hi-C resolution and window size on CLOD outlier detection and seek an appropriate combination, we analyzed three Hi-C resolutions (50 kb, 100 kb, and 250 kb) and five sliding window sizes (3×3, 5×5, 7×7, 9×9, and 11×11). The results showed that with different resolutions, the number of CLOD identified outliers varied from a few thousands to close to a hundred of thousands, while the ratios of CLOD identified outliers to total loci pairs ranged from 0.004 to 0.009 (Extended Data Fig. 1d). Since the amount of CLOD outliers obtained from 250 kb resolution was much less than those from resolution 50 kb and 100 kb, we compared the sliding window size using the latter two resolutions. With 50 kb resolution, the identified CLOD outliers were sensitive to window size, with 551 locus pairs consistently present in all sliding window sizes. CLOD outliers identified from 100 kb Hi-C resolution were more robust with diverse window sizes (Extended Data Fig. 1e). Balancing between processing time and identification efficiency, Hi-C long-range colocalization was identified using the 100 kb resolution with a 7×7 window, unless otherwise stated.

MERFISH data

The Euclidean coordinates of chromosome fragments were directly downloaded from Zenodo repository (records:3928890) .

DNA seqFISH+ data

The Euclidean coordinates of chromosome fragments and the level of epigenetic marks were directly downloaded from Zenodo repository (records:7693825) .

721 CRISPR-dCas9 two-color labeling

722 The plasmid cloning method used for CRISPR-dcas9 fluorescent labeling was adapted from a
723 previously established protocol[49]. The sgRNAs used in this study are listed in Extended Data Table
724 2.

725 Cell culture

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

730 For transfection, HEK293T cells were seeded onto 35 mm glass-bottom dishes one day before the
731 experiments. In the next day, 750 ng of donor template and 750 ng of sgRNA-Cas9 plasmid were
732 co-transfected using FuGENE HD Transfection Reagent (Promega #E2311). Cells were incubated
733 for 48 hours before imaging[49].

734 Time-lapse imaging

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

738 For trajectory tracking (Fig. 1l), cells were maintained in a Tokai Hit Microscope Stage Top
739 Incubator and imaged using a Nikon Ti microscope with FITC and TRITC channels (excitation
740 wavelength is 470 nm and 555 nm for GFP and RFP, respectively; 60x oil objective). Cells were
741 imaged every minute for 60 minutes with an exposure time of 100 ms[50].

742 Imaging processing

743 Due to the low number of cells, cell segmentation and puncta recognition were performed by
744 CellProfiler[51] and corrected manually.

745 Derivation of the minimal multivalent binding model

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

754 Suppose that without influence from other sLRCs within the tagged metaLRC, an sLRC pair has
755 a bare association rate α and a dissociation rate γ. Then at kinetic equilibrium, the probability in
756 the bound state is $\frac{\alpha}{\alpha+\gamma}$, and from the FISH data we expect that $\alpha \ll \gamma$. Here for simplicity we

757 assume that each sLRC pair has the same association-dissociation kinetic parameters. Let's focus
758 on one pair *i*. When there is one or more other bound sLRC pairs within the metaLRC, after
759 dissociation the two loci of pair *i* are constrained by the nearby bound sLRCs and cannot diffuse
760 far away from each other before rebind. Note that the genomic distance between these two loci of
761 a sLRC ≥ 100 Mb, and by definition the genomic distance between two sLRCs *i* and *j* within a
762 metaLRC l_{ij} << 100 Mb. Consequently, at the presence of other bound sLRC pair(s), the two loci
763 of a dissociated sLRC have an increased probability to confront each other and rebind, so the
764 effective association constant changes to $\alpha + \delta$, and we expect that $\delta \gg \alpha$. For simplicity, given
765 l_{ij} << 100 Mb let us assume that each sLRC pair has the same effect on any other one, and the
766 effect is not additive but δ remains the same value at the presence of one or more bound sLRC
767 pairs. Given the genomic distance between two loci belonging to two different sLRCs, we further
768 assume that there is no direct interaction between two sLRC pairs to simplify the analyses. That
769 is, we assume that the dissociation constant of a sLRC pair, γ, is not affected by the presence of
770 other bound sLRC pairs.

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

773 $\frac{dp_0}{dt} = -Nap_0 + \gamma p_1,$

774 $\frac{dp_1}{dt} = -\big((N-1)(a+\delta) + \gamma\big)p_1 + 2\gamma p_2 + Nap_0,$

775 $\frac{dp_2}{dt} = -\big((N-2)(a+\delta) + 2\gamma\big)p_2 + 3\gamma p_3 + (N-1)(a+\delta)p_1,$

776 ...

777 $\frac{dp_{N-1}}{dt} = -\big((a+\delta) + (N-1)\gamma\big)p_{N-1} + N\gamma p_N + 2(a+\delta)p_{N-2},$

778 $\frac{dp_N}{dt} = -N\gamma p_N + (a+\delta)p_{N-1}.$

779 Then at steady state,

780 $p_1 = \frac{Na}{\gamma}p_0,$

781 $p_2 = \frac{(N-1)(a+\delta)}{2\gamma}p_1,$

782 $p_3 = \frac{(N-2)(a+\delta)}{3\gamma}p_2,$

783 ...

784 $p_N = \frac{a+\delta}{N\gamma}p_{N-1}.$

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

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

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. 7d&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[40]. 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 $(\sum_{i=1}^{N} \quad \sigma_i) \geq 1$and 0 otherwise.

Reference

46. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics* **25**, 1754-1760 (2009).
47. Open2C *et al.* Pairtools: from sequencing data to chromosome contacts. *PLOS Computational Biology* **20**, e1012164 (2024).
48. Abdennur, N. & Mirny, L.A. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics* **36**, 311-316 (2020).
49. 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 (2020).

50. Wang, W., Poe, D., Yang, Y., Hyatt, T. & Xing, J. Epithelial-to-mesenchymal transition proceeds through directional destabilization of multidimensional attractor. *Elife* **11**, e74866 (2022).
51. Stirling, D.R. *et al.* CellProfiler 4: improvements in speed, utility and usability. *BMC bioinformatics* **22**, 433 (2021).