Multi-CD: Multi-scale discovery of Chromatin Domains

Min Hyeok Kim[1,*], **Ji Hyun Bak**[1,*], **Lei Liu**[1], **and Changbong Hyeon**[1,†]

[1]Korea Institute for Advanced Study, Seoul 02455, Korea; *These two authors contributed equally.; †Correspondence: hyeoncb@kias.re.kr

(Dated: June 21, 2019)

Abstract – **Identifying chromatin domains (CDs) from Hi-C data is currently a central problem in genome research. Here we present Multi-CD (https://github.com/multi-cd), a unified method to discover CDs at various genomic scales. Integrating approaches from polymer physics, financial market fluctuation analysis and Bayesian inference, Multi-CD identifies the CDs that best represent the global pattern of correlation manifested in Hi-C, and reveals the multi-scale structure of chromosome. At each scale, the CDs are consistent with the results of existing methods, as well as with biological data from independent sources. CD solutions compared across different scales and cell types allow us to quantify the hierarchy between four major families of CDs, and to glean the principles of chromatin organization: (i) Sub-TADs, TADs, and meta-TADs constitute a robust hierarchical structure. (ii) The assemblies of compartments and TAD-based domains are governed by distinct organizational principles. (iii) Sub-TADs are the common building blocks of chromosome architecture. CDs acquired from our interpretation of Hi-C data using Multi-CD not only provide new insights into chromatin organization, but also offer a quantitative account for its cell-type-dependence and function.**

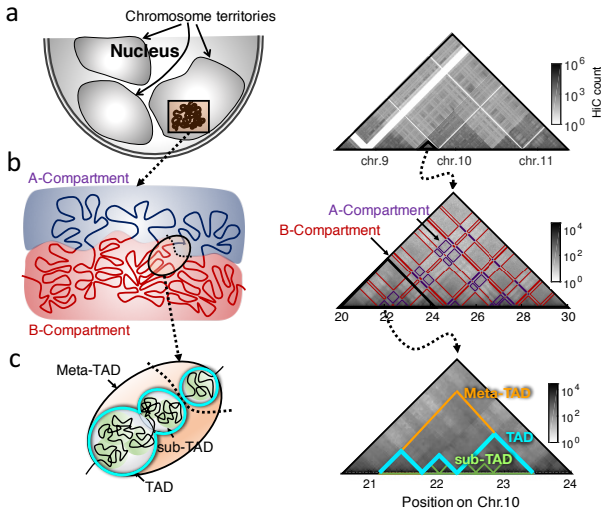

Figure 1: **The hierarchical organization of interphase chromosome and Hi-C maps at each scale.** **(a)** Chromosome territories in the cell nucleus, which are reflected as the higher intra-chromosomal counts in the Hi-C map. **(b)** Alternating blocks of active and inactive chromatins, segregated into A- and B-compartments, give rise to the checkerboard pattern on Hi-C. **(c)** Sub-megabase to megabase sized chromatin folds into TADs. According to recent studies, adjacent TADs are merged to meta-TAD [6], whereas individual TAD is further decomposed into sub-TADs [17, 28–31].

INTRODUCTION

Chromosome conformation capture (3C) and its derivatives, which are used to identify chromatin contacts through the proximity ligation techniques [1, 2], take center stage in studying the organization and function of chromosomes [3, 4]. It is clear from the genome-wide interaction profiles of Hi-C data that details of chromosome architecture not only vary with cell type but also with the transcription activity and the phase of cell cycle, underscoring the functional roles of chromosome structure in gene expression and regulation [5–13]. Since pathological states of chromatin are also manifested in Hi-C [14, 15], accurate characterization of chromatin domains (CDs) from Hi-C data is of utmost importance.

Before discussing our new method and algorithm, we give a brief overview of the current knowledge on scale-dependent organization of chromatin [6, 16–21]. Chromosomes packaged inside the nucleus are first segregated into their own territories (Fig. 1a) [18]. At the scale of $\gtrsim \mathcal{O}(10)$ Mb, alternating blocks of active and inactive chromatin are phase-separated into two megabase sized aggregates, called A- and B-compartments [16, 17, 20, 22] (Fig. 1b). Inter-chromosomal contact patterns based on high resolution Hi-C (5 kb [17], 25 kb [23]) have suggested more detailed classification of the compartments into at least six sub-compartments, A1, A2, B1, B2, B3 and B4, which were shown to be in good agreement with the finer details of epigenetic markers. Because of their well-conserved domain boundaries across cell/tissue types, topologically associated domains (TADs), detected at $\sim \mathcal{O}(10^{-1}) - \mathcal{O}(1)$ Mb [24–27], are deemed the basic functional unit of chromatin organization and gene regulation [18–21]. It was suggested that the proximal TADs in genomic neighborhood aggregate into a higher-order structural domain termed "meta-TAD" [6], although such proposal has not fully been accepted by the community. At smaller genomic scale,

each TAD is split into sub-structures called sub-TADs that display more localized contacts [17, 28–31] (Fig. 1c).

A number of algorithms make important contributions toward understanding the intra-chromosome architecture by identifying CDs from Hi-C data. However, CDs identified using different algorithms or parameters display significant variations, and currently there is no universally accepted definition of CD at each genomic scale. For example, the average size of a TAD varies from 100 kb to 2 Mb depending on the specific algorithm being used. Furthermore, still lacking is a unified algorithm to characterize CDs at multiple scales. Many of the existing algorithms, specialized at finding CDs at particular genomic scales, require that the Hi-C data be formatted at specific resolutions that match the scale of the target domains [16, 17, 24]. Although more recent methods, such as TADtree [32] and Armatus [33], considered hierarchical domain structures of chromosomes, most existing methods are motivated from the viewpoint of local pattern recognition analyses [16, 17, 24, 34], and do not take into account the underlying physical nature of chromosomes that chromosomes are a long polymer organized in 3 dimensions [6, 35–40].

Here, we interpret Hi-C data as the pairwise contact probability of a locally equilibrated polymer network whose inter-loci distances are restrained with varying strengths of interactions. Just like Hi-C map changes along cell cycle but is stably defined over a certain time interval of cell cycle phase, we assume that the strengths of inter-loci interactions are stably maintained over the time interval. Based on this model of the locally equilibrated model of polymer network, we derive the cross-correlation matrix from Hi-C and use it as the sole input for the algorithm of identifying CD at varying genomic scales (Multi-CD). The algorithm includes a tuning parameter which enables us to control the average domain size. We demonstrate the utility of Multi-CD by applying it to Hi-C data from various cell lines as well as to that of a particular cell line over multiple genomic scales. We also confirm that the CD structures identified at multiple genomic scales are consistent with the previous knowledge on chromatin organization, as well as with information from bio-markers. This study will show that amid the rapidly expanding volume of Hi-C data [10–12], Multi-CD holds good promise to more quantitative and principled determination of chromatin organization.

RESULTS

Overview of Multi-CD

The primary goal of this study is to extract information of CDs from Hi-C data at varying genomic scale of interest. First, we translate the Hi-C data into a cross-correlation matrix of polymer network, by noting that chromosomes are in essence a polymer network with pairwise interactions between the loci [35, 41, 42]. The three dimensional structure of chromosome is deemed in nonequilibrium steady state with "activity" where the energy source of ATP is internally injected to the system and is dissipated as heat. As long as the system remains in local mechanical equilibrium over an extended time scale, one can assume a stable structural ensemble manifested in Hi-C data. Using this assumption of local mechanical equilibrium, we model the dynamic fluctuation of polymer network configuration, representing chromosome structure, around its local equilibrium ensemble by using a sum of harmonic potentials, writing down the distance distribution between two loci i and j into gaussian:

$$P(r_{ij};\gamma_{ij}) = \frac{4}{\sqrt{\pi}}\gamma_{ij}^{3/2} r_{ij}^2 \exp(-\gamma_{ij} r_{ij}^2) \tag{1}$$

with $\gamma_{ij} = (\sigma_{ii} + \sigma_{jj} - 2\sigma_{ij})/2$, where $\sigma_{ij}\,(= \langle \delta\mathbf{r}_i \cdot \delta\mathbf{r}_j \rangle)$ is the positional covariance, determined by the topology of polymer network [42]. Indeed, distance distributions measured using fluorescence measurement between chromatin loci justifies this hypothesis (see Fig. S1). Importantly, our interpretation of chromosome conformation as a locally equilibrated, quasi-stable polymer network enables a one-to-one mapping of the contact probability $p_{ij} = \int_0^{r_c} dx P(x;\gamma_{ij})$ to the positional covariance σ_{ij}, and hence to the cross-correlation matrix, $(\mathbf{C})_{ij} = \sigma_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}}$ (see Methods). The cross-correlation matrix $\mathbf{C}$ normalizes the wide numerical range of the original Hi-C counts into the range between -1 and 1.

Clustering a correlation matrix into a finite number of correlated groups is a general problem discussed in diverse disciplines. Here, we adapted a formalism known as the "group model," developed for identifying the correlated groups of companies from empirical data of stock market price fluctuations [43–45]. Without ambiguity, the formalism can be applied to the clustering of correlated genomic loci in a chromosome. For a given correlation matrix $\mathbf{C}$, the group model finds the optimal solution of clustered loci groups (domains) that best explains the pattern manifested in $\mathbf{C}$. The domain solution for N loci can be written as a vector $\mathbf{s} = \{s_1, s_2, \ldots, s_N\}$, where s_i denotes the domain index for locus i. Technically, this procedure involves finding a vector $\mathbf{s}$ that maximizes the posterior distribution $p(\mathbf{s}|\mathbf{C})$ for a given correlation data $\mathbf{C}$; the optimal CD solution is found as $\mathbf{s}^* = \mathrm{argmax}_{\mathbf{s}}\, p(\mathbf{s}|\mathbf{C})$. Maximizing the posterior distribution in the form of $p(\mathbf{s}|\mathbf{C}) \propto e^{-\mathcal{H}(\mathbf{s}|\mathbf{C})/T}$ is equivalent to minimizing the cost function (or the effective Hamiltonian) $\mathcal{H}(\mathbf{s}|\mathbf{C})$. We consider the cost function of the form $\mathcal{H}(\mathbf{s}|\mathbf{C}) = \mathcal{E}(\mathbf{s}|\mathbf{C}) + \lambda\mathcal{K}(\mathbf{s})$, where $\mathcal{E}(\mathbf{s}|\mathbf{C})$ quantifies the goodness of clustering, and $\mathcal{K}(\mathbf{s})$ with $\lambda(\geq 0)$ promotes simpler CD solutions by penalizing the effective number of clusters (see Methods). This gives rise to a "tunable" group model that allows us to flexibly control the average size of domain solutions by changing the parameter λ. In light of the grand-canonical ensemble in statistical mechanics, T is the effective *temperature* of the system, and λ amounts to the *chemical potential*. Our "tunable group model," Multi-CD, applied to Hi-C data, is used to identify the four major CD families, namely, sub-TADs, TADs, meta-TADs, and compartments.

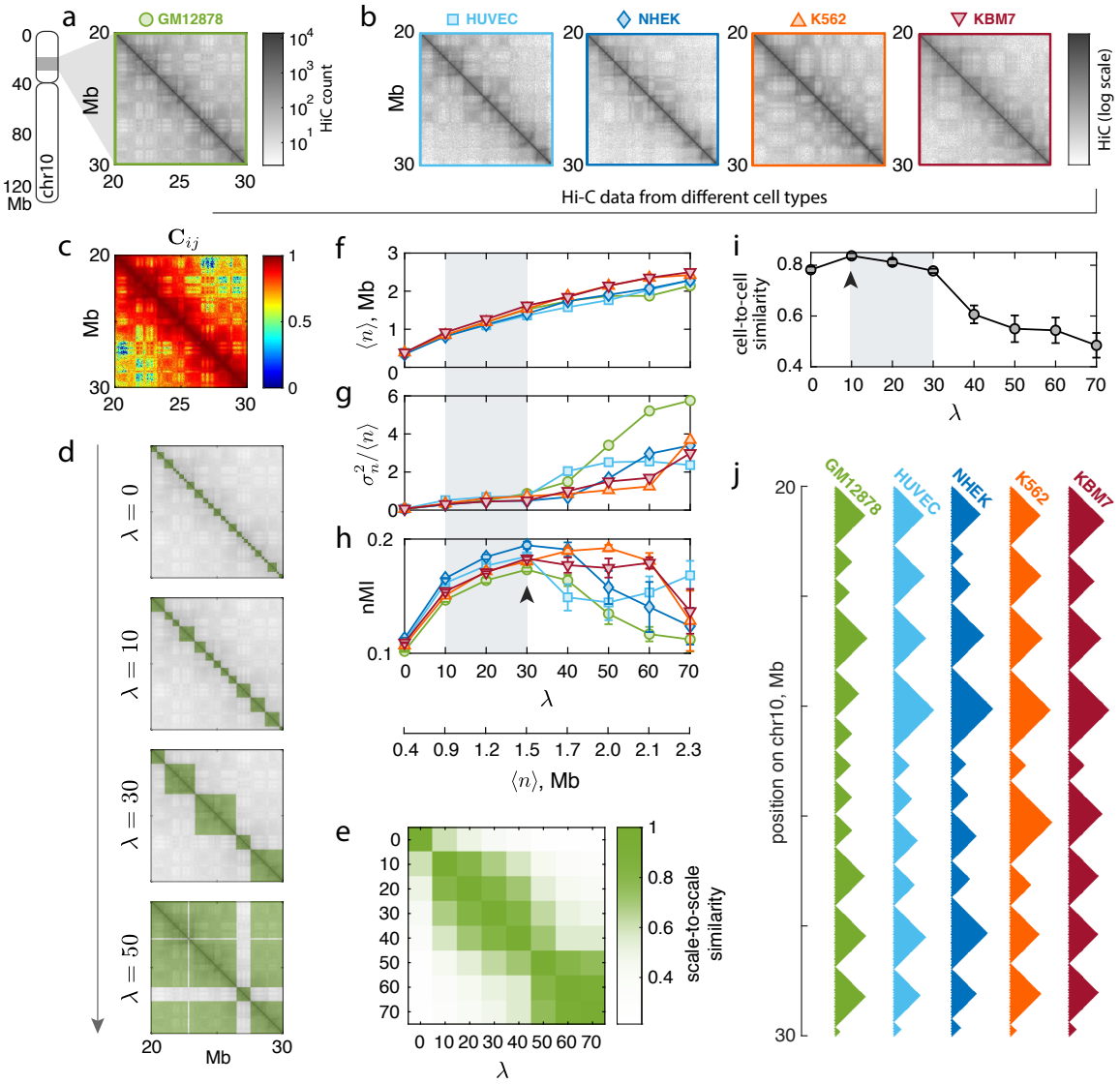

Figure 2: **Multi-scale domain solutions for various cell types, identified using Multi-CD. (a)** A subset of 50-kb resolution Hi-C data, covering a 10-Mb genomic region of chr10 in GM12878. **(b)** Similar subsets of Hi-C data from the same chromosome (chr10) in four other cell lines: HUVEC, NHEK, K562, and KBM7. **(c-e)** Applying Multi-CD to the Hi-C data in **a**, from GM12878. **(c)** The cross-correlation matrix $\mathbf{C}_{ij}$. **(d)** Domain solutions determined by Multi-CD at 4 different values of $\lambda = 0, 10, 30, 50$. Multi-CD captures less fragmented domains with increasing λ. **(e)** Similarity between domain solutions at different λ's, calculated in terms of Pearson correlation. The similarity matrix has its own modular structure, such that it is partitioned into two regions, $\lambda > 40$ and $\lambda < 40$. The boundary value $\lambda \approx 40$ corresponds to the average genomic size of $\langle n \rangle = 1.8$ Mb. **(f-h)** Statistics of the domain solutions, found from all five Hi-C data in **a-b**. As λ is varied, we plot **(f)** the average domain size, $\langle n \rangle$; **(g)** the index of dispersion in the domain size, $D(= \sigma_n^2/\langle n \rangle)$; **(h)** the normalized mutual information, nMI. **(i-j)** Comparison of domain solutions across cell types. **(i)** Average cell-to-cell similarity of domain solutions at fixed values of λ, in terms of Pearson correlations. **(j)** Domains obtained at $\lambda = 10$. See Fig. S3 for solutions at a smaller $\lambda = 0$ and a larger $\lambda = 40$.

Discovery of chromatin domains at multiple scales

We applied Multi-CD to 50-kb resolution Hi-C of chromosome 10 from five different cell lines: GM12878, HUVEC, NHEK, K562, KBM7 (Fig. 2a-b). Given a Hi-C matrix, we first obtained the cross-correlation matrix $\mathbf{C}$ (Fig. 2c), and used Multi-CD to identify a set of CDs for each fixed value of λ (Fig. 2d). This resulted in a family of CD solutions at varying λ, with coarser CDs at larger values of λ. We note that the family of CD solutions for a given cell line are divided into two regimes. In the case of GM12878 (Fig. 2e), CD solutions can be partitioned into two groups below and above $\lambda \approx 40$. Solution families for other four cell lines are also similarly

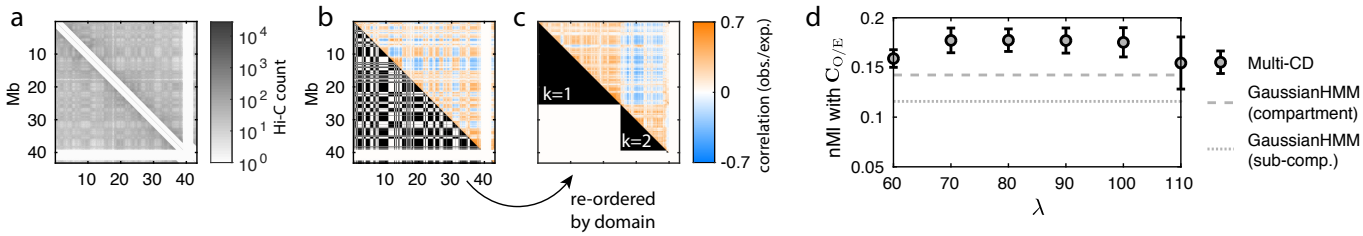

Figure 3: **Domain solutions for compartments.** **(a)** Input Hi-C data for compartment identification. The 2-Mb diagonal band was removed. **(b)** Demonstrated are the domain solution at $\lambda = 90$ obtained based on Hi-C data in **(a)** (lower triangle) and $\mathbf{C}_{\text{O/E}}$ (upper triangle). **(c)** The domain solution for compartments obtained using the original Hi-C data in **(b)** are re-ordered with the cluster (compartment) index. The two largest compartments ($k = 1, 2$), corresponding to B ($k = 1$) and A ($k = 2$) compartments, are depicted in the lower triangle. Clearly separated B- and A-compartments emerge from the correlation matrix $\mathbf{C}_{\text{O/E}}$ (upper triangle) when the rows and columns of $\mathbf{C}_{\text{O/E}}$ are also re-ordered in accordance with the domain solution (lower triangle). **(d)** nMI between domain solutions at varying λ and $\mathbf{C}_{\text{O/E}}$. The nMI of compartment structure with respect to $\mathbf{C}_{\text{O/E}}$ is maximized at $\lambda = 70 - 100$. The nMI values of sub-compartment (dashed line) and compartment (dotted line) from [17] are depicted for comparison.

divided (Fig. S2).

From the Multi-CD analysis for all five cell lines of chromosome 10, we find the following features which are also shared by chr 4, 11, and 19 (Fig. S4)

(i) In each of the cells, the average domain size $\langle n \rangle$ increased monotonically with λ (Fig. 2f).

(ii) There is a crossover point at $\lambda = \lambda_{cr}$ where the distribution of domain sizes suddenly changes. The variability of domain size, quantified in terms of the index of dispersion, $D\left(= \sigma_n^2/\langle n \rangle\right)$, is below 1 for small λ ($< \lambda_{cr}$), which means that the domain size is regular, but it exhibits transition at $\lambda_{cr} \approx 30 - 40$ ($\langle n \rangle_{cr} \approx 1.6$ Mb) for GM12878, HUVEC, and NHEK; and at $\lambda_{cr} \approx 60 - 70$ ($\langle n \rangle_{cr} \approx 2.2$ Mb) for K562 and KBM7 (Fig. 2g).

(iii) The goodness of CD solution quantified by the normalized mutual information (nMI, see Methods for its definition) against Hi-C data is maximized at $\lambda^* = 30$ in all the cell types, except for K562 ($\lambda^* = 50$) (Fig. 2h).

(iv) The extent of domain conservation is quantified in terms of the average cell-to-cell similarity over all the cell-type pairs, where the similarity is evaluated using the Pearson correlation (see Methods). We found strong cell-to-cell domain conservation in the range of $0 < \lambda \leq 30$, which corresponds to CD sizes $\langle n \rangle \lesssim 1.5$ Mb (Fig. 2i). The maximal extent of domain conservation across the cell lines is found at $\lambda = 10$ (Fig. 2i), at which the average domain size is $\langle n \rangle \approx 0.9$ Mb, which is close to the typical size of a TAD suggested in the previous studies [24, 25]. This finding is consistent with the widely accepted notion that TADs are the most well-conserved, common organizational and functional unit of chromosomes, across different cell types [21]. Thus, we identify the CDs found at $\lambda = 10$ as the TADs for human chromosome 10 (see Fig. 2j; also see Fig. S3 for domain solutions at $\lambda = 0$ and $\lambda = 40$).

Two families of chromatin domains

Fig. 2h shows that there is a special value of $\lambda^* \approx 30$ at which the CD solution best captures the pattern of Hi-C data. The family of CD solutions are also divided into two regimes at $\lambda \approx \lambda^*$.

TAD-based chromatin organization at $\lambda \leq \lambda^*$. What do the CDs at $\lambda^* = 30$ represent? First, the CDs at this scale have an average size of $\langle n \rangle^* \approx 1.5$ Mb (Fig. 2h), which is slightly greater than the size of TADs ($\langle n \rangle_{\text{TAD}} \approx 0.9$ Mb); Second, for GM12878 the CD solutions showing high similarity at $10 < \lambda < 40 \approx \lambda^*$ can be grouped together (Fig. 2e. See also the similarity matrices calculated for HUVEC and NHEK in Fig. S2). Based on these observations, we surmise that CDs at $\lambda^* \approx 30$ are associated with a higher-order structure of TADs, a "meta-TAD", which results from an aggregate consisting of multiple TADs in genomic neighborhood [6]. In contrast to the previous analysis which extended the range of meta-TAD to entire chromosome via hierarchical clustering analysis [6], the meta-TAD implicated from Multi-CD is confined in a finite range, so that it is well discerned from compartments and at the same time is more correlated with TADs (Fig. 2e). Notably, the pattern of CDs identified at $\lambda < \lambda^*$ is localized (see Fig. 2d, $\lambda = 0$, 10, 30). Our algorithm identifies the diagonal blocks of Hi-C data as the subsets of a hierarchically crumpled structure of chromatin chain [37, 46].

Compartment-like chromatin organization at $\lambda > \lambda^*$. The super-Mb sized domains are generally defined as the compartment in the chromosome organization [21]. In this scale, a direct application of Multi-CD to the cross-correlation matrix $\mathbf{C}$ (as in Fig. 2) is dominated by the strong local correlation from the loci pairs in genomic neighborhood. A simple and effective solution to capture the compartment-like structures is to exclude a narrow band along the diagonal of the Hi-C matrix (Fig. 3a; also see Methods). Then we can apply Multi-CD to identify two large compartments with alternating patterns (Fig. 3b). The results from Multi-CD success-

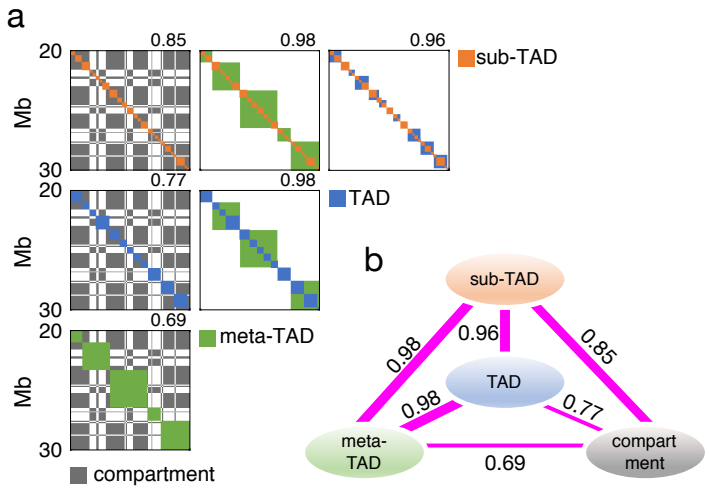

Figure 4: **Hierarchical organization of CD families (a)** Hierarchical structure of CDs are highlighted with the domain solutions for sub-TADs (red), TADs (green), meta-TADs (blue) and compartments (black). Each panel represents the superposition of two domain solution, and the hierarchy score h is provided above each panel. **(b)** A diagram of hierarchical relations between sub-TADs, TADs, meta-TADs and compartments based on the average hierarchy score calculated for chr10 of GM12878. Higher score for a pair of two different domains means that one domain is more nested to other domain.

fully capture the non-local correlations, as clearly seen with a reordering of indices (Fig. 3c). It is natural to associate a domain ($k = 1$) showing a greater number of contacts (Fig. S5) with the compartment B, which is usually more compact than compartment A; and $k = 2$ with compartment A. Further validation of the two compartments, by comparisons to epigenetic markers, will be presented below.

Conventionally, in order to identify chromosome compartments, the existing methods based on principal component analysis (PCA) use the Pearson correlation matrix of a low-resolution Hi-C data ($\mathbf{C}_{\mathrm{O/E}}$) [16] (see Methods), whose heatmap is typically featured with checkerboard pattern (upper triangular part of Fig. 3b). For the case of compartment, instead of making direct comparison with Hi-C data, we used the $\mathbf{C}_{\mathrm{O/E}}$ matrix to evaluate the goodness of compartment-like CD solutions, again calculating the normalized mutual information (nMI). We note that Multi-CD outperforms GaussianHMM [17], a widely accepted benchmark, in capturing the large-scale structures in Hi-C (Fig. 3d).

Hierarchical organization of chromatin domains

We examined the extent of hierarchical relationship between the four classes of CD solutions obtained at varying λ. From the diagram in which sub-TADs, TADs, meta-TADs and compartments are overlaid on top of each other (Fig. 4a), it is visually clear that sub-TADs or TADs almost never fail to be included inside the boundary of meta-TAD, whereas there are mismatches between the domain boundaries of meta-TADs and compartments. We evaluated the extent of overlap or domains-within-domains type of hierarchy between two domain solutions by means of the hierarchy scores (h) which quantifies the extent of inclusion of smaller domains into larger domains (see Methods).

Based on the hierarchy scores, calculated over the CD solutions from Hi-C data of GM12878 (Fig. 4b), we found the basic principles for chromatin organization: (i) The hierarchy scores between the pairs of TAD-related domains (sub-TADs, TADs, and meta-TADs) are all > 0.96, which is appreciably greater than that of any pair of TAD-related domains with compartments. (ii) The hierarchical links of TADs and meta-TADs with compartments are relatively weak. This implies that TADs or meta-TADs are not necessarily the components of compartments, which is also consistent with the recent reports that TADs and compartments are organized by different mechanisms [47, 48]. (iii) Although the hierarchy score between sub-TADs and compartments ($h = 0.85$) is not so large as those among the pairs of TAD-based domains, it is still greater than the hierarchy scores between TADs and compartments ($h = 0.77$) or between meta-TADs and compartments ($h = 0.69$). Thus, sub-TAD can be considered a good candidate for a common building block of the chromatin architecture.

Validation of domain solutions from Multi-CD

The CD solutions from Multi-CD are in good agreement with the previously proposed CDs, obtained from several different methods. Specifically, CDs correspond to the sub-TADs [17] in the prior-free solution at $\lambda = 0$, to the TADs [24, 49] at $\lambda \approx 10$, and to the compartments [17] at $\lambda \approx 90$ (see Fig. S6). When assessed in terms of nMI of acquired CD solutions against the input Hi-C data, Multi-CD outperforms other conventional methods (ArrowHead, DomainCaller, GaussianHMM) in identifying three distinct CD families (Fig. 5a).

In order to further validate the biological relevance of the CD solutions from Multi-CD, we compared them with several biomarkers that are known to be correlated with the spatial organization of the genome [50].

First, we calculated how much our domain boundaries obtained at $\lambda = 10$ are correlated with the CTCF signals which are known to capture TAD boundaries [24, 25] (Fig. 5b). Compared to the correlation (or extent of overlap quantified by $\chi(d)$. See Eq.21 and Fig. 5b) of CD solutions for $\lambda = 0$ with CTCF signals, the overlap at the domain boundary ($d \approx 0$) is stronger for solutions at $\lambda = 10$ and 20, which are in the parameter range where Multi-CD identifies TADs. We also observe that Multi-CD identifies TAD boundaries that are more sharply correlated with the CTCF binding sites than those identified by two popular methods, ArrowHead [17] and DomainCaller [24] (Fig. 5b). Specifically, when fitted to

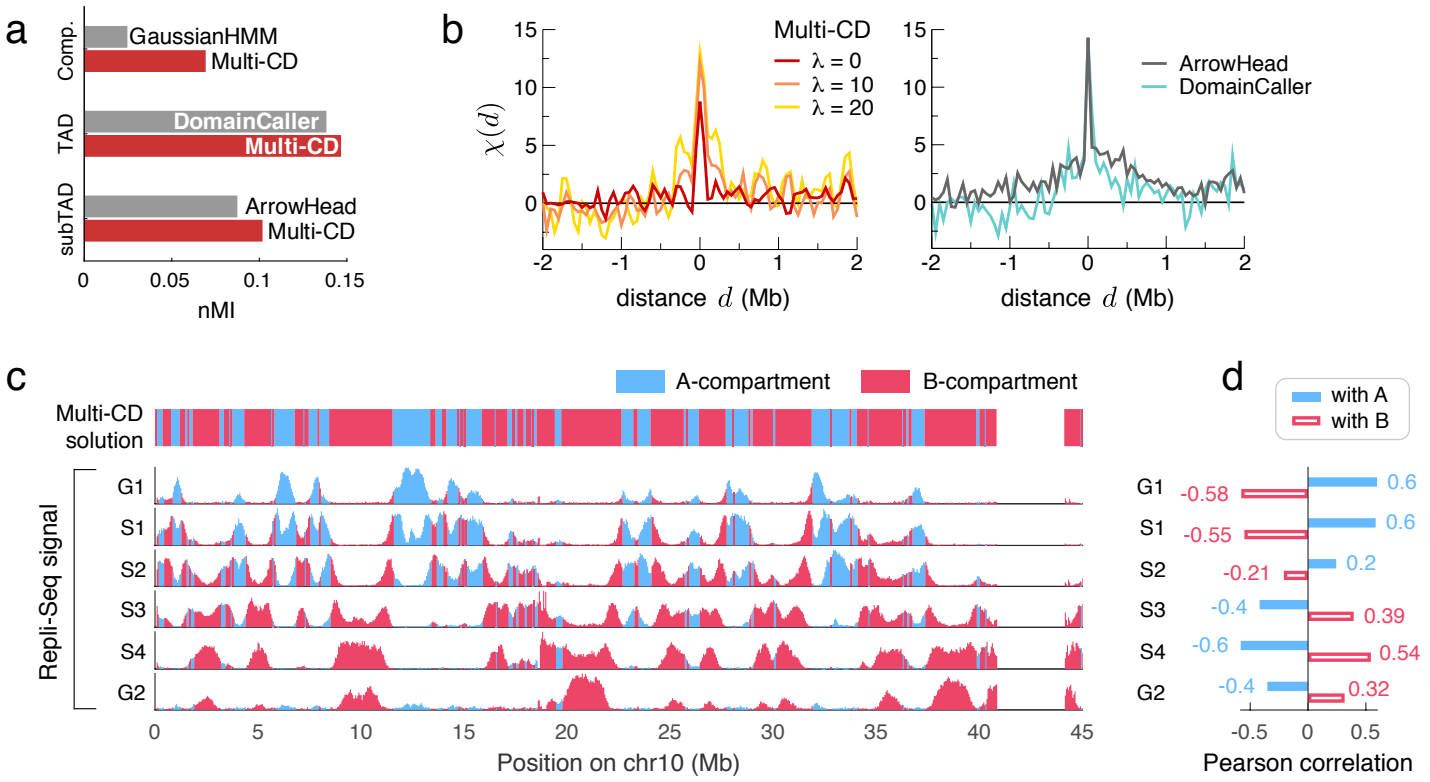

Figure 5: **Validation of CD solutions from Multi-CD. (a)** CD solutions (sub-TAD, TAD and compartment) assessed in terms of nMI against Hi-C data ($\log_{10}\mathbf{M}$): Multi-CD outperforms ArrowHead, DomainCaller and GaussianHMM at the corresponding scale. **(b)** The overlap function $\chi(d)$ calculated between CTCF enrichment and the domain boundaries obtained from different methods: $\chi(d)$ for Multi-CD at $\lambda = 0$, 10, and 20 (left), and $\chi(d)$ for ArrowHead and DomainCaller (right). **(c)** Genome-wide, locus-dependent replication signal. The genomic position of the two domains, B ($k = 1$) and A ($k = 2$) compartments obtained in Fig. 3, are shaded in red and blue, respectively. **(d)** Pearson correlation between the replication signal and the two compartments, compartment A (filled blue) and B (open red), respectively. The replication activities in the two compartments are anti-correlated. In the early phase of cell-cycle (G1, S1, S2) the replication of A-compartment is more active than B-compartment, but an opposite trend is observed in the later phase (S3, S4, G2).

exponential function, the correlation lengths are 34 kb ($\lambda = 0$), 143 kb ($\lambda = 10$), and 234 kb ($\lambda = 20$); whereas the correlation lengths obtained from ArrowHead and DomainCaller are $\gtrsim$ 900 kb (Fig. 5b).

Next, we compared our compartment-like domains with the replication timing profile (GM12878 Repli-Seq data) [7, 51]. The large-scale domains from Multi-CD (at $\lambda = 90$) are in good agreement with the patterns of replication timing anticipated for the A/B compartments, which exhibits anti-correlated activation/repression along the replication cycle (Fig. 5c-d). Specifically, the replication signal in the Multi-CD-identified compartment A (blue shade in Fig. 5c) is active in the early phases (G1, S1, S2), whereas it is repressed (or deactivated) in the late phases (S3, S4, G2). An entirely opposite trend is observed for B-compartment (red shade in Fig. 5c): the replication activity in B compartment is repressed in the early phases (G1, S1, S2), but is activated in the late phases (S3, S4, G2). The Pearson correlation between the replication signals (Fig. 5d) quantitatively confirms the clear contrast between the replication timing of A/B compartments. Correlation between A/B compartments determined by Multi-CD and compartmentalized patterns of the histone modifications are also shown in Fig. S7, which validates the domain solutions of compartments identified by Multi-CD.

Chromatin organization and its link to gene expression

To demonstrate that chromatin organization is closely linked to gene expression, we overlaid the RNA-seq profiles on the 10 Mb range of the TAD solutions identified for the chr10 of five cell lines (GM12878, HUVEC, NHEK, K562, KBM7) (Fig. 2j). At around 26.8 Mb position of this chromosome, we found APBB1IP gene, that is transcriptionally active in GM12878 and KBM7 but not in HUVEC, NHEK and K562 (Fig. 6). We also consulted the Gene-

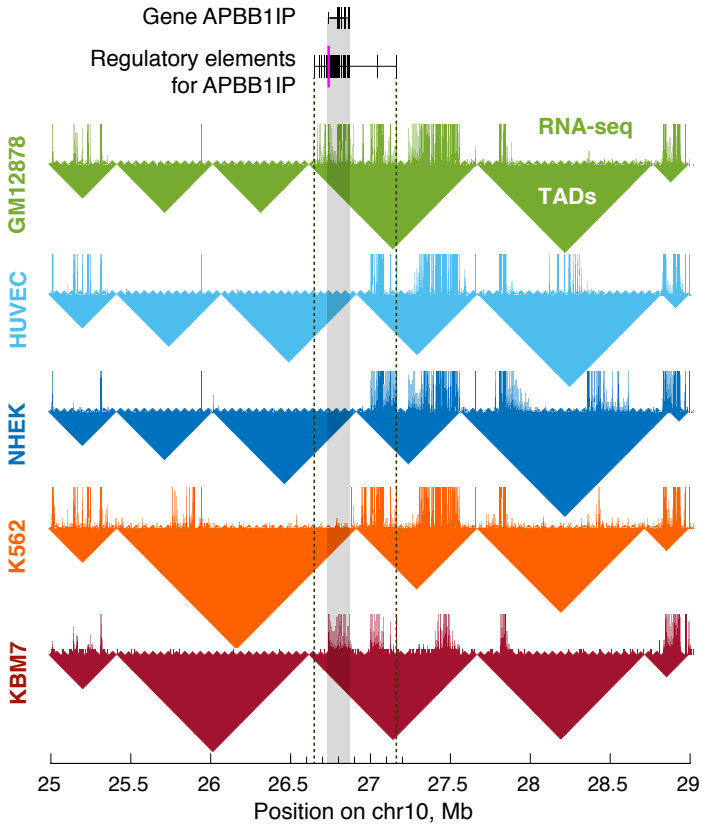

Figure 6: **Cell-line dependent TAD organization and its link to gene expression.** The RNA-seq signals from five different cell lines (colored lines) are shown on top of the TAD solutions obtained by Multi-CD (triangles with matching colors). At the top shown are the the position of a specific gene APBB1IP (top row), and the regulatory elements associated with this gene (second row), including the enhancers and the promoter (the position of promotor is marked with a magenta line). APBB1IP is transcriptionally active only in two cell lines, GM12878 and KBM7. In the two cell lines, the regulatory elements are fully enclosed in the same TAD.

Hancer database [52] to identify the regulatory elements for this gene (enhancers and promoters) in the interval between 26.65 and 27.15 Mb. Remarkably, our Multi-CD solutions show that the interval associated with the regulatory elements is fully enclosed in a single TAD in GM12878 and KBM7, whereas it is split into two different domains in the other three cell lines (Fig. 6). This suggests that for the gene expression it is critical to have all regulatory elements within the same TAD, which is consistent with the understanding that TADs are the functional boundaries for genetic interactions [5, 6, 18, 20].

DISCUSSION

What fundamentally differentiates Multi-CD from other approaches rests on the algorithm by which the pattern of CD is identified. In the conventional methods, local features of Hi-C data, such as CD boundaries or loops enriched with higher contact frequencies, are key for CD-identification, and it is usually required that the Hi-C data is prepared in a specific range of resolution that matches the scale of domain to be identified. In contrast, Multi-CD solves the problem of global pattern clustering as its basic algorithm for CD discovery. Therefore, Multi-CD can find CDs across a wide range of scales without resorting to a coarse-grained version of Hi-C data or to a particular bin size at each scale of interest. Moreover, as Multi-CD is based on a physical model of polymer network, the method can offer a physically relevant interpretation of Hi-C data.

Multi-CD uses a tuning parameter λ, which is tantamount to the "chemical potential" in statistical thermodynamics, to set the average domain size, giving rise to λ-dependent CD solution for a given Hi-C data. nMI comparing CD solutions with the 50-kb resolution Hi-C data is maximized at $\lambda \approx 30$, which corresponds to ~ 1.5 Mb in domain size (length) (Fig. 2f). Notably, 1.5 Mb, the average size of CD that we can best read off from the 50-kb resolution Hi-C data [17] used in this study, is also similar to the domain size detected by a recently proposed TAD detection algorithm called deDoC [53]. In essence, the concept of "graph structural entropy" used in deDoC is also based on global pattern recognition. The authors of deDoC, who developed deDoC as a TAD detection algorithm, have concluded that their $\sim$ 2-Mb-sized domain solution from their analyses on 40-kb data of Dixon *et al.* [24] was the best solution for TAD, based on their finding that deDoC identified domain solution displayed the lowest structural entropy in comparison with all the five other TAD detection algorithms they tested. Interestingly, we also found that the best domain solution from varying λ, assessed in terms of nMI with Hi-C heatmap, was when $\lambda \approx 30$, which corresponds to the genomic length of 1.5 Mb; however, we do not conclude CD solution at $\lambda = 30$ represents the solution for canonical TAD. Instead, we surmise the domain solution at $\lambda = 30$ is for meta-TAD, an aggregate of TADs in genomic neighborhood. As indicated by the domain solutions from Multi-CD at varying λ, the extent of domain conservation across different cell types are maximized at $\lambda = 10$ ($\langle n \rangle = 0.8$ Mb). To be consistent with the general notion that TADs are the functional unit of chromosome conserved across different cell types and species [24], CD solution obtained at $\lambda = 10$ is better interpreted as the solution for TAD.

We showed that the characteristics of CD solutions shared by the TAD-like domains do not precisely hold together in compartment-like domains. This finding is consistent with the recent insightful studies which report that compartments and TADs do not necessarily have a hierarchical relationship because they are formed by different mechanisms of motor-driven active loop extrusion and microphase separation [48, 54–56]. Notably, even when clear mismatches are present between the meta-TAD and compartment, the

sub-TADs are, in most of the cases, a part of the compartment (Fig. 4). This finding points to sub-TADs as the fundamental building blocks of the higher domain organization. In fact, the existence of sub-TADs is robust even when a higher resolution Hi-C data is analyzed. From a clustering analysis on 5-kb resolution HiC data, the boundaries of $\sim$ 300-kb-sized sub-TAD are clear and consistent with those obtained from 50-kb resolution Hi-C (see Fig. S8).

Although there are methods that report hierarchical CDs [32, 33], Multi-CD makes significant advances both algorithmically and conceptually. Multi-CD formulates the problem in terms of a state-space model of the high-dimensional domain solution vector, $\mathbf{s}$, instead of finding a set of intervals, so that it can detect non-local domains with better flexibility. Multi-CD also avoids the high false-negative rate that is typical of the previous method (e.g., TADtree [32]) that focuses on the nested domain structure (Fig. S9). Further, employing an appropriate prior to explore the solution space effectively, Multi-CD can avoid the problem encountered in Armatus [33] which skips detection of domains in some part of Hi-C data while its single scale parameter is varied (compare Fig. S9e and Fig. S9f).

In order to glean genome function from Hi-C data that vary with genomic state [10–13], a computationally efficient and accurate method to identify CD structures is of vital importance. In summary, we developed Multi-CD, a novel and versatile method for CD-identification. The method identifies multi-scale structures of chromatin domains by solving the global optimization problem. We find that the chromatin domains identified from Multi-CD are in excellent match with biological data such as CTCF binding sites and replication timing signal, supplementing the existing methods. Quantitative analyses of CD structures identified by this unified algorithm across multiple genomic scales and various cell types not only offer general physical insight into how chromatin is organized in the nucleus but also will be of practical use to decipher broad spectrum of Hi-C data obtained under various conditions.

ONLINE CONTENT

All methods and additional information are available in Methods and Supplementary Figures.

References

[1] Dekker, J., Rippe, K., Dekker, M., & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).

[2] Davies, J. O., Oudelaar, A. M., Higgs, D. R., & Hughes, J. R. How best to identify chromosomal interactions: a comparison of approaches. *Nat. Methods* **14**, 125 (2017).

[3] Misteli, T. Beyond the sequence: cellular organization of genome function. *Cell* **128**, 787–800 (2007).

[4] Bickmore, W. A & van Steensel, B. Genome architecture: domain organization of interphase chromosomes. *Cell* **152**, 1270–1284 (2013).

[5] Lanctôt, C., Cheutin, T., Cremer, M., Cavalli, G., & Cremer, T. Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nat. Rev. Genetics* **8**, 104 (2007).

[6] Fraser, J., Ferrai, C., Chiariello, A. M., Schueler, M., Rito, T., et al. Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Mol. Sys. Biol.* **11**, 852 (2015).

[7] Ryba, T., Hiratani, I., Lu, J., Itoh, M., Kulik, M., et al. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res.* **20**, 761–770 (2010).

[8] Hiratani, I., Ryba, T., Itoh, M., Yokochi, T., Schwaiger, M., et al. Global reorganization of replication domains during embryonic stem cell differentiation. *PLoS Biol.* **6**, e245 (2008).

[9] Cavalli, G & Misteli, T. Functional implications of genome topology. *Nat. Struct. Mol. Biol.* **20**, 290–299 (2013).

[10] Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., et al. Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331 (2015).

[11] Krijger, P. H. L & De Laat, W. Regulation of disease-associated gene expression in the 3D genome. *Nat. Rev. Mol. Cell Biol.* **17**, 771 (2016).

[12] Dixon, J. R., Xu, J., Dileep, V., Zhan, Y., Song, F., et al. Integrative detection and analysis of structural variation in cancer genomes. *Nat. Genetics* **50**, 1388 (2018).

[13] Shao, Y., Lu, N., Wu, Z., Cai, C., Wang, S., et al. Creating a functional single-chromosome yeast. *Nature* **560**, 331 (2018).

[14] Franke, M., Ibrahim, D. M., Andrey, G., Schwarzer, W., Heinrich, V., et al. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265–269 (2016).

[15] Baca, S. C., Prandi, D., Lawrence, M. S., Mosquera, J. M., Romanel, A., et al. Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666–677 (2013).

[16] Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).

[17] Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., et al. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665 – 1680 (2014).

[18] Cremer, T & Cremer, M. Chromosome territories. *Cold Spring Harbor Perspectives in Biology* **2** (2010).

[19] Dekker, J & Heard, E. Structural and functional diversity of topologically associating domains. *FEBS Letters* **589**, 2877–2884 (2015).

[20] Sexton, T & Cavalli, G. The role of chromosome domains in shaping the functional genome. *Cell* **160**, 1049 – 1059 (2015).

[21] Dixon, J. R., Gorkin, D. U., & Ren, B. Chromatin domains: the unit of chromosome organization. *Mol. Cell* **62**, 668–680 (2016).

[22] Wang, S., Su, J.-H., Beliveau, B. J., Bintu, B., Moffitt, J. R., et al. Spatial organization of chromatin domains and compartments in single chromosomes. *Science* **353**, 598–602 (2016).

[23] Rowley, M. J., Nichols, M. H., Lyu, X., Ando-Kuri, M., Rivera, I. S. M., et al. Evolutionarily conserved principles predict 3D chromatin organization. *Mol. Cell* **67**, 837–852 (2017).

[24] Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).

[25] Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., et al. Spatial partitioning of the regulatory landscape of the X-inactivation center. *Nature* **485**, 381 (2012).

[26] Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., et al. Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome. *Cell* **148**, 458 – 472 (2012).

[27] Pope, B. D., Ryba, T., Dileep, V., Yue, F., Wu, W., et al. Topologically associating domains are stable units of replication-timing regulation. *Nature* **515**, 402 (2014).

[28] Phillips-Cremins, J. E., Sauria, M. E., Sanyal, A., Gerasimova, T. I., Lajoie, B. R., et al. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153**, 1281–1295 (2013).

[29] Jin, F., Li, Y., Dixon, J. R., Selvaraj, S., Ye, Z., et al. A high-resolution map of three-dimensional chromatin interactome in human cells. *Nature* **503**, 290 (2013).

[30] Rocha, P. P., Raviram, R., Bonneau, R., & Skok, J. A. Breaking TADs: insights into hierarchical genome organization. *Epigenomics* **7**, 523–526 (2015).

[31] Wang, Q., Sun, Q., Czajkowsky, D. M., & Shao, Z. Sub-kb Hi-C in D. melanogaster reveals conserved characteristics of TADs between insect and mammalian cells. *Nature communications* **9**, 188 (2018).

[32] Weinreb, C & Raphael, B. J. Identification of hierarchical chromatin domains. *Bioinformatics* **32**, 1601–1609 (2016).

[33] Filippova, D., Patro, R., Duggal, G., & Kingsford, C. Identification of alternative topological domains in chromatin. *Algorithms for Molecular Biology* **9**, 14 (2014).

[34] Shin, H., Shi, Y., Dai, C., Tjong, H., Gong, K., et al. Topdom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res.* **44**, e70–e70 (2015).

[35] Mateos-Langerak, J., Bohn, M., de Leeuw, W., Giromus, O., Manders, E. M., et al. Spatially confined folding of chromatin in the interphase nucleus. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 3812–3817 (2009).

[36] Barbieri, M., Chotalia, M., Fraser, J., Lavitas, L.-M., Dostie, J., et al. Complexity of chromatin folding is captured by the strings and binders switch model. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 16173–16178 (2012).

[37] Halverson, J. D., Smrek, J., Kremer, K., & Grosberg, A. Y. From a melt of rings to chromosome territories: the role of topological constraints in genome folding. *Rep. Prog. Phys.* **77**, 022601 (2014).

[38] Brackley, C. A., Brown, J. M., Waithe, D., Babbs, C., Davies, J., et al. Predicting the three-dimensional folding of cis-regulatory regions in mammalian genomes using bioinformatic data and polymer models. *Genome Biol.* **17**, 59 (2016).

[39] Shi, G., Liu, L., Hyeon, C., & Thirumalai, D. Interphase Human Chromosome Exhibits Out of Equilibrium Glassy Dynamics. *Nat. Commun.* **9**, 3161 (2018).

[40] Liu, L., Shi, G., Thirumalai, D., & Hyeon, C. Chain organization of human interphase chromosome determines the spatiotemporal dynamics of chromatin loci. *PLoS Comp. Biol.* **14**, e1006617 (2018).

[41] Bryngelson, J. D & Thirumalai, D. Internal constraints induce localization in an isolated polymer molecule. *Phys. Rev. Lett.* **76**, 542–545 (1996).

[42] Bohn, M., Heermann, D. W., & van Driel, R. Random loop model for long polymers. *Phys. Rev. E.* **76**, 051805 (2007).

[43] Laloux, L., Cizeau, P., Bouchaud, J.-P., & Potters, M. Noise dressing of financial correlation matrices. *Phys. Rev. Lett.* **83**, 1467 (1999).

[44] Noh, J. D. Model for correlations in stock markets. *Phys. Rev. E.* **61**, 5981 (2000).

[45] Giada, L & Marsili, M. Data clustering and noise undressing of correlation matrices. *Phys. Rev. E.* **63**, 061101 (2001).

[46] Grosberg, A., Nechaev, S., & Shakhnovich, E. The role of topological constraints in the kinetics of collapse of macromolecules. *J. Phys.* **49**, 2095–2100 (1988).

[47] Nora, E. P., Goloborodko, A., Valton, A.-L., Gibcus, J. H., Uebersohn, A., et al. Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell* **169**, 930–944 (2017).

[48] Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., et al. Two independent modes of chromatin organization revealed by cohesin removal. *Nature* **551**, 51 (2017).

[49] Serra, F., Baù, D., Goodstadt, M., Castillo, D., Filion, G. J., et al. Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. *PLoS Comp. Biol.* **13**, 1–17 (2017).

[50] Misteli, T. Beyond the sequence: cellular organization of genome function. *Cell* **128**, 787–800 (2007).

[51] Hansen, R. S., Thomas, S., Sandstrom, R., Canfield, T. K., Thurman, R. E., et al. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 139–144 (2010).

[52] Fishilevich, S., Nudel, R., Rappaport, N., Hadar, R., Plaschkes, I., et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database* **2017**, 1–17 (2017).

[53] Li, A., Yin, X., Xu, B., Wang, D., Han, J., et al. Decoding topologically associating domains with ultra-low resolution Hi-C data by graph structural entropy. *Nature Commun* **9**, 3265 (2018).

[54] Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., et al. Formation of chromosomal domains by loop extrusion. *Cell Reports* **15**, 2038–2049 (2016).

[55] Gassler, J., Brandão, H. B., Imakaev, M., Flyamer, I. M., Ladstätter, S., et al. A mechanism of cohesin-dependent loop extrusion organizes zygotic genome architecture. *EMBO J.* **36**, 3600–3618 (2017).

[56] Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N., & Mirny, L. A. Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E6697–E6706 (2018).

[57] Szabo, Q., Jost, D., Chang, J.-M., Cattoni, D. I., Papadopoulos, G. L., et al. Tads are 3d structural units of higher-order chromosome organization in drosophila. *Science Advances* **4** (2018).

[58] Finn, E. H., Pegoraro, G., Shachar, S., & Misteli, T. Comparative analysis of 2d and 3d distance measurements to study spatial genome organization. *Methods* **123**, 47–55 (2017).

[59] Giorgetti, L., Galupa, R., Nora, E. P., Piolot, T., Lam, F., et al. Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell* **157**, 950–963 (2014).

ACKNOWLEDGEMENT

We thank Roger Oria Fernandez for the helpful feedback on the code. We thank the KIAS Center for Advanced Computation for providing computing resources. C.H. acknowledges the partial support from the National Research Foundation of Korea (NRF-2018R1A2B3001690).

AUTHOR CONTRIBUTIONS

M.H.K., J.H.B., L.L., and C.H. conceived the project and designed the algorithm. M.H.K., J.H.B. implemented the algorithm. M.H.K., J.H.B., and L.L. analyzed all the results. M.H.K., J.H.B., and C.H. wrote the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

METHODS

Data acquisition

Hi-C data. We applied Multi-CD on the 50-kb-resolution Hi-C data of chr10 from five different cell types (GM12878, HUVEC, NHEK, K562, and KBM7). The data were obtained through GEO data repository (GSE63525-*cell type*-primary) [17].

Biological markers. The domain solutions from Multi-CD were compared with known biological markers. We obtained these data mostly from the ENCODE project [60]. Specifically, we used the enrichment data of the transcriptional repressor CTCF measured in a Chip-Seq assay from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwTfbs/wgEncodeUwTfbsGm12878CtcfStdPkRep1.narrowPeak.gz. We binned the CTCF assay at 50-kb resolution (to match the Hi-C format). If there are multiple signal enrichments in a single bin, we took the average value. Because each CTCF signal has a finite width, there are occasional cases where a signal ranges across two bins; in those cases we evenly divided the signal strength into the two bins. The Repli-seq signals in the six phases G1, S1, S2, S3, S4, and G2 were obtained from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwRepliSeq/, and were averaged over 50-kb windows along the genome to construct the replication timing profiles. The 11 histone mark signals were obtained from http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/signal/jan2011/bigwig/, and undergone the same preprocess as Repli-seq. The RNA-seq data for the four cell lines GM12878, HUVEC, NHEK and K562 were also obtained from http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeCaltechRnaSeq/. RNA-seq for the cell line KBM7 were separately obtained from https://opendata.cemm.at/barlowlab/2015_Kornienko_et_al/hg19/AK_KBM7_2_WT_SN.F.bw. Information about the APBB1IP gene was obtained from the GeneCards database https://www.genecards.org/cgi-bin/carddisp.pl?gene=APBB1IP. The information about its regulatory elements was obtained specifically from the GeneHancer [52] database http://hgdownload.cse.ucsc.edu/gbdb/hg19/geneHancer/geneHancerInteractionsAll.hg19.bb, accessible through the UCSC Genome Browser Module.

Pre-processing of Hi-C data

Normalization and contact probability. We performed Knight-Ruiz (KR) normalization [61] on Hi-C data, using the normalization vectors provided by Ref. [17] along with the Hi-C matrices. The resulting matrix elements are in similar orders of magnitude as in the original Hi-C, unlike in a standard application of KR normalization where they are normalized to row-sums of 1. In order to use the KR-normalized Hi-C map, $\mathbf{M}$, as the contact probability matrix $\mathbf{P}$, defined element-wise as $(\mathbf{P})_{ij} = p_{ij}$, we first divided $\mathbf{M}$ by the mean value of the greatest matrix elements typically concentrated near the diagonal band of $\mathbf{M}$-matrix, i.e., $\mu = \frac{1}{N-1}\sum_{i=1}^{N-1}\mathbf{M}_{i,i+1}$, and next multiplied a constant value μ_c. In other words, we rescaled $\mathbf{M}_{ij}$ into $p_{ij} = (\mu_c/\mu)\mathbf{M}_{ij}$. We set $\mu_c = 0.9$ and regarded the elements of $\mathbf{P}$ greater than 1 as outliers and set them to 1, which effectively filters the unusually high contact signals from the actual data. For the contact probability for a pair of loci, we used p_{ij}, a rescaled and high-intensity-signal-filtered version of $\mathbf{M}_{ij}$.

Correlation matrix from Hi-C. A chromosome can be regarded as a polymer chain containing N monomers, each of which (i-th monomer or locus) corresponds to the i-th genomic segment and its spatial position is written as $\mathbf{r}_i$. Employing the idea of the random loop model (RLM) [42], which has been proposed for modeling chromosome conformation, we interpret that chromosome conformation is described with an ideal polymer network crosslinked at multiple sites, which is deemed a chromosome conformation in local mechanical equilibrium. In RLM, the position vector of the chromosome $\mathbf{r} = (\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_N)$ obeys gaussian distribution with zero mean $\langle \mathbf{r} \rangle = \mathbf{0}$ and covariance matrix Σ, the probability of relative distance $P(r_{ij})$ between i and j-th monomers in 3D is given as Eq. 1, or

$$P(r_{ij};\gamma_{ij}) = \frac{4}{\sqrt{\pi}}\gamma_{ij}^{3/2} r_{ij}^2 \exp(-\gamma_{ij} r_{ij}^2),$$

where $\gamma_{ij} = 1/2(\sigma_{ii} + \sigma_{jj} - 2\sigma_{ij})$ and $\sigma_{ij} \equiv (\Sigma)_{ij}$. Provided that the contact between two monomers is formed when their distance r_{ij} is within a certain cutoff distance r_c, the contact probability (p_{ij}) can be calculated as

$$p_{ij} = \int_0^{r_c} P(r_{ij};\gamma_{ij}) dr_{ij} = \mathrm{erf}(\gamma_{ij}^{1/2} r_c) - 2r_c\sqrt{\frac{\gamma_{ij}}{\pi}} e^{-\gamma_{ij} r_c^2}, \quad (2)$$

with $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x dt e^{-t^2}$. Note that p_{ij} is a monotonically increasing function of $\gamma_{ij} (\geq 0)$. Therefore, given covariance matrix Σ, we can explicitly calculate the contact probability p_{ij} through Eq.(2). In inverse problem, the covariance matrix Σ can be inferred from contact probability matrix $\mathbf{P}$. However, this inverse problem requires additional assumption about the variance of each monomer σ_{ii} from the definition of $\gamma_{ij} = 1/2(\sigma_{ii} + \sigma_{jj} - 2\sigma_{ij})$. We assume that all variances have identical value ($\sigma_{ii} = \sigma_{jj} = \sigma_c$), which generates the following normalized covariance matrix (i.e. correlation matrix, $\mathbf{C}$)

$$(\mathbf{C})_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}} = 1 - \frac{1}{2\sigma_c\gamma_{ij}}. \quad (3)$$

The parameter σ_c sets the overall intensity of $\mathbf{C}$. Here, we set this variable as the median of $1/2\gamma_{ij}$ to maintain the balance of $\mathbf{C}_{ij}$.

The primary goal of this study is to extract information of CDs from Hi-C correlation matrix ($\mathbf{C}$). In fact, a very similar problem has been posed for stock market price fluctuations [43, 44]. Adapting the formalism in References [43, 44], we assumed that x_i, which

stands for the "genomic state" (or "transcriptional state") of i-th locus, obeys the following stochastic equation in terms of the standardized variable ξ_i, i.e., $\xi_i = (x_i - \langle x_i \rangle)/\sigma_{x_i}$

$$\xi_i = y(\eta_{s_i}, \epsilon_i) = \sqrt{\frac{g_{s_i}}{1+g_{s_i}}}\eta_{s_i} + \frac{1}{\sqrt{1+g_{s_i}}}\epsilon_i \tag{4}$$

where s_i denotes the index of CD to which the i-th locus is clustered, and the parameter g_{s_i} ($-1/2 \leq g_{s_i} \leq \infty$) defines the strength of intra-CD correlation; η_{s_i} and ϵ_i are the independent and identically-distributed (i.i.d) random variables with zero mean and unit variance, $\eta_{s_i}, \epsilon_i \sim \mathcal{N}(0,1)$. From Eq.4, the cross-correlation between the loci i and j is written as

$$\langle \xi_i \xi_j \rangle = \frac{g_{s_i}}{1+g_{s_i}}\delta_{s_i s_j} + \frac{1}{1+g_{s_i}}\delta_{ij}. \tag{5}$$

Therefore, in light of Eq.5, the first term of Eq. 4 on the right hand side represents intra-CD variation of the s_i-th CD where intra-domain correlation increases with g_{s_i}; the second term of Eq. 4 corresponds to a noise that randomizes the intra-domain correlation dictated by the first term. By matching Eq. 3 with Eq. 5

$$(\mathbf{C})_{ij} := \langle \xi_i \xi_j \rangle. \tag{6}$$

one can use the cross-correlation matrix $\mathbf{C}$ from σ_{ij} as an input for Multi-CD.

Removal of the diagonal band for identifying compartments. The Hi-C matrix shows that the interaction strength is highly concentrated near the diagonal elements, which makes it difficult to identify the compartment characterized with the long-range interaction pattern. To circumvent this issue, the previous methods have either intentionally reduced the resolution of Hi-C data (usually to 1 Mb) [16] or used only inter-chromosomal interactions [17]. In this study, as a similar motivation we ignore the diagonal band of $\mathbf{C}$. Specifically, we set all elements in $\mathbf{C}$ to zero if its genomic distance ($|i-j|$) is smaller than l_c. We scanned l_c and found that the CD solutions most consistent with the compartment are obtained for $l_c^* = 40$ ($\simeq$ 2 Mb in 50-kb resolution). This value is almost identical to the crossover value of domain size n^* which sets the boundary between TAD-like and compartment-like domains.

The observed/expected (O/E) matrix and its Pearson correlation matrix. The O/E matrix was used to account for the genomic distance-dependent contact number due to random polymer interactions in chromosome [16]. Each pair (i,j) in O/E matrix is calculated by taking the count number M_{ij} (observed number) and dividing it by average contacts within the same genomic distance $d = |i-j|$ (expected number). Since the expected number could be noisy, one smooths it by increasing the window size (see refs [16, 17] for further details). In this study, we used the expected number obtained from [17]. The Pearson correlation matrix of the O/E ($\mathbf{C}_{\text{O/E}}$) represents the overall contact pattern through pairwise correlation coefficients between loci.

Identifying chromatin domains

Our goal is to find the chromatin domain (CD) solution $\mathbf{s} = \{s_1, s_2, \cdots, s_N\}$, as well as a set of parameters for the intra-CD correlation $G = \{g_1, g_2, \cdots, g_{K_{\max}}\}$, that best represent the pattern in the Hi-C data. Each $\mathbf{s}$ is a *state* in the CD solution space. For each genomic locus $i \in \{1, 2, \cdots, N\}$, the element of the CD solution $s_i = k$ indicates that the i-th locus belongs to the domain k ($= 1, 2, \cdots, K_{\max}$). It is always ensured that the elements of each $\mathbf{s}$ spans a set of consecutive integers from 1 to $K_{\max}$, where $K_{\max}$ is the number of distinct domains in the solution. For example, a state $\mathbf{s} = \{1,1,1,2,2,3\}$ describes a structure where the 6 loci are clustered into 3 domains as $\{\{s_1, s_2, s_3\}, \{s_4, s_5\}, \{s_6\}\}$ with the corresponding strength of three intra-domain correlation $G = \{g_1, g_2, g_3\}$.

Likelihood: goodness of clustering. To formulate the clustering problem, we first consider the likelihood of data, at a given CD solution $\mathbf{s}$ with strength parameters G: [45]

$$p(\text{data}|\mathbf{s}, G) = \left\langle \prod_{i=1}^{N} \delta\left(\xi_i - \sqrt{\frac{g_{s_i}}{1+g_{s_i}}}\eta_{s_i} - \frac{1}{\sqrt{1+g_{s_i}}}\epsilon_i\right)\right\rangle_{\boldsymbol{\eta},\boldsymbol{\epsilon}} \tag{7}$$

where $\langle \cdots \rangle_{\boldsymbol{\eta},\boldsymbol{\epsilon}}$ denotes an average over the gaussian noises for η_{s_i} and ϵ_i. The data dependence of the likelihood is written simply in terms of the correlation matrix $\mathbf{C}$ (Eq. 6). With standard calculations involving Gaussian integrals, the corresponding log-likelihood can be written as:

$$\log p(\mathbf{C}|\mathbf{s}, G) = -\frac{1}{2}\sum_{k=1}^{K_{\max}}\left[(1+g_k)\left(n_k - \frac{g_k c_k}{1+g_k n_k}\right) - n_k \log(1+g_k) + \log(1+g_k n_k)\right],$$

where $n_k = \sum_{i=1}^{N} \delta_{s_i,k}$ is the size of domain k, and $c_k = \sum_{i,j=1}^{N} \mathbf{C}_{ij}\delta_{s_i,k}\delta_{s_j,k}$ is the sum of all intra-domain correlation elements. Conveniently, the likelihood $p(\mathbf{C}|\mathbf{s}, G)$ is maximized at $g_k = (c_k - n_k)/(n_k^2 - c_k)$ for a given $\{\mathbf{s}, \mathbf{C}\}$. We define the cost function $\mathcal{E}(\mathbf{s}|\mathbf{C})$ as the negative log-likelihood at this likelihood-maximizing G:

$$\mathcal{E}(\mathbf{s}|\mathbf{C}) = \frac{1}{2}\sum_{k=1}^{K_{\max}}\left[\log\frac{c_k}{n_k} + (n_k - 1)\log\frac{n_k^2 - c_k}{n_k^2 - n_k}\right], \tag{8}$$

such that $\max_G p(\mathbf{C}|\mathbf{s}, G) = \exp(-\mathcal{E}(\mathbf{s}|\mathbf{C}))$. The cost function evaluates the (negative) goodness of clustering for a given CD solution $\mathbf{s}$.

Prior: preference to simpler solutions. Because of the structural hierarchy inherent to chromosome and the ensemble characteristic of the Hi-C measurement, it is still an issue to define CDs at a certain length scale of interest. In order to construct a unified formalism that can control the overall domain size in a CD solution, we

introduce a prior of the form $p(\mathbf{s}) = \exp(-\lambda\mathcal{K}(\mathbf{s}))$, where $\mathcal{K}(\mathbf{s})$ increases with the complexity of the solution $\mathbf{s}$. Specifically, we define $\mathcal{K}(\mathbf{s})$ as

$$\mathcal{K}(\mathbf{s}) = \exp\left(-\sum_{k=1}^{K_{\max}} \frac{n_k}{N} \log \frac{n_k}{N}\right), \tag{9}$$

such that it measures the effective number of CDs from the domain size distribution. For example, in the limit where all CDs are of the same size, $\mathcal{K}(\mathbf{s}) = K_{\max}$. This formulation is also equivalent to adding a regularizer to the cost function $\mathcal{E}$, such that the total cost function $\mathcal{H}$ becomes:

$$\mathcal{H}(\mathbf{s}|\mathbf{C}) = \mathcal{E}(\mathbf{s}|\mathbf{C}) + \lambda\mathcal{K}(\mathbf{s}), \tag{10}$$

where the parameter λ controls the relative weight of $\mathcal{K}(\mathbf{s})$ with respect to $\mathcal{E}(\mathbf{s}|\mathbf{C})$.

Posterior distribution. Then the posterior distribution is given by the following Bayes rule:

$$p(\mathbf{s}|\mathbf{C}) \propto p(\mathbf{C}|\mathbf{s})\, p(\mathbf{s}) = \exp(-\mathcal{H}(\mathbf{s}|\mathbf{C})). \tag{11}$$

We remark that this formulation is analogous to the grand canonical ensemble in statistical mechanics. The total cost function $\mathcal{H}(\mathbf{s}|\mathbf{C})$ can be thought of as the effective Hamiltonian of the system; $\mathcal{E}(\mathbf{s}|\mathbf{C})$ amounts to the energy of the system, and $\mathcal{K}(\mathbf{s})$ the effective number of particles (in this case CDs) with a chemical potential of λ. It is natural to introduce an effective temperature T, so that the probability of having a state $\mathbf{s}$ is given as

$$p(\mathbf{s}|\mathbf{C}) \propto \exp\left[-\frac{1}{T}\left(\mathcal{E}(\mathbf{s}|\mathbf{C}) + \lambda\mathcal{K}(\mathbf{s})\right)\right]. \tag{12}$$

A higher temperature T makes the distribution flatter; in other words, it *tempers* the distribution. This is useful for an efficient posterior inference through simulated annealing.

Metropolis-Hastings sampling. We use Markov chain Monte Carlo (MCMC) sampling to find the maximum value of the posterior distribution, or equivalently the minimum value of the total cost function $\mathcal{H}$. The standard Metropolis-Hastings (MH) routine was used, such that at each trial move from the current state $\mathbf{s}$ to the next state $\mathbf{s}'$, the move is accepted with a probability $\min(1, \alpha)$, where $\alpha(\mathbf{s}, \mathbf{s}') = \exp\left[-(\mathcal{H}(\mathbf{s}'|\mathbf{C}) - \mathcal{H}(\mathbf{s}|\mathbf{C}))/T\right]$. We used "single mutation" proposals, as described below.

To ensure that a steady state is reached, we continue the sampling until each chain collects $t_{\text{tot}} \geq 5\tau^*$ samples in the CD solution space. Here τ^* is the "relaxation time" defined as the number of steps it takes until the autocorrelation function $R(\tau)$, drops significantly: $\tau^* = \text{argmin}_\tau\, |R(\tau) - 1/e|$. The autocorrelation function is calculated from the value of the total cost function $\mathcal{H}$, as

$$R(\tau) = \frac{1}{\sigma^2}\, \langle(\mathcal{H}(\mathbf{s}_t|\mathbf{C}) - \mu)(\mathcal{H}(\mathbf{s}_{t+\tau}|\mathbf{C}) - \mu)\rangle_t, \tag{13}$$

where $\mathbf{s}_t$ is the t-th sample in the chain, and μ and σ are the mean and standard deviation of $\{\mathcal{H}(\mathbf{s}_t|\mathbf{C})\}$, respectively. The running time average in Eq.13, $\langle\cdots\rangle_t \left[= \frac{1}{t_{\text{tot}}-\tau}\int_0^{t_{\text{tot}}-\tau} dt(\cdots)\right]$, is taken over all the pairs of samples with the time gap of τ. We stop the sampling as soon as the total sampling time is five times longer than the relaxation time ($t_{\text{tot}} > 5 \cdot \tau^*(t_{\text{tot}})$), so that the state $\mathbf{s}_t$ (or $\mathcal{H}(\mathbf{s}_t|\mathbf{C})$) is practically in steady states.

Single mutation in the CD solution space. In the space of CD solutions, we define a single mutation as a move from a state $\mathbf{s}$ to another state $\mathbf{s}'$, such that the two CD solutions $(\mathbf{s}, \mathbf{s}')$ differ only by one genomic locus. In other words, it is a move with distance $d(\mathbf{s}, \mathbf{s}') = 1$, where the distance between the two CD states is defined as the number of loci with differing domain memberships, $d(\mathbf{s}, \mathbf{s}') = \sum_{i=1}^{N} \text{XOR}(s_i, s_i')$. More precisely, because a CD solution is invariant upon permutations of the domain indices, d is uniquely defined as the *minimal* number of mismatches over all possible domain index permutations. We define the set of all single-mutated neighbors around a state $\mathbf{s}$ as $S_1(\mathbf{s}) = \{\mathbf{s}' : d(\mathbf{s}, \mathbf{s}') = 1\}$.

Simulated annealing. We use the simulated annealing to explore the high-dimensional CD solution space which is also likely characterized with multiple local minima. We start from a finite temperature $T = T_0 > 0$ and slowly decrease it to $T \to 0$, letting the system relax toward the global minimum of the configurational landscape of $\mathcal{H}$ (Fig. S10). The simulated annealing process is described below.

Initialization. An initial configuration $\mathbf{s}^{(0)}$ is generated in two random steps. First, the total number of CDs, $K_{\max}$, is drawn randomly from the set of integers $\{1, \cdots, N\}$. Then, each genomic locus $i \in \{1, \cdots, N\}$ is allocated randomly into one of the CDs, $k \in \{1, 2, \cdots, K_{\max}\}$. The initial temperature T_0 is determined such that the acceptance probability for the "worst" move around $\mathbf{s}^{(0)}$ is 0.5. Specifically, it is given as $T_0 = \text{argmin}_T\, |\exp(-\Delta\mathcal{H}_1/T) - 0.5|$, where $\Delta\mathcal{H}_1 = \max_{\mathbf{s}\in S_1(\mathbf{s}_0)}\{\mathcal{H}(\mathbf{s}|\mathbf{C}) - \mathcal{H}(\mathbf{s}^{(0)}|\mathbf{C})\}$ is the energy difference to the least favorable move among the set of all single mutations.

Iteration. At each step r, the temperature is fixed at T_r. We sample the target distribution $p_r(\mathbf{s}|\mathbf{C}) \propto \exp(-\mathcal{H}(\mathbf{s}|\mathbf{C})/T_r)$, using the Metropolis-Hastings sampler described above. For the next step $r+1$, the temperature is lowered by a constant cooling factor $c_{\text{cool}} \in (0, 1)$, such that the next temperature is $T_{r+1} = c_{\text{cool}} \cdot T_r$. We used $c_{\text{cool}} = 0.95$ in this study.

Final solution. The annealing process is repeated until the temperature reaches a small (but finite) value T_f. We used $T_f = 0.03$. Then we quench the system to the closest local minimum by performing a "zero-temperature" sampling, in which a proposed move is always accepted if it lowers the cost function. This process is simply to remove any remaining fluctuation from the finite temperature, which is usually very small at this point. Because there is still no guarantee that the global minimum is found, we tried a batch of at least 10 different initial configurations and chose the final state

$\mathbf{s}^*$ that gives the minimal $\mathcal{H}(\mathbf{s}^*|\mathbf{C})$.

Analysis on subsets of Hi-C data. Our method allows the user to break down the Hi-C data into subsets, as long as the CDs are localized within the subsets (Fig. S11). This saves the algorithm from the large memory requirement of dealing with the entire intrachromosomal Hi-C (for example, Hi-C of chromosome 10 has 2711 bins in 50-kb resolution). For the analysis of the 50-kb resolution Hi-C data in this paper, we used subsets of the data that correspond to 40-Mb ranges along the genome, or 800 bins.

The overall schematic involving the algorithm of Multi-CD. A schematic diagram of Multi-CD is provided in Fig. S12.

Analysis and evaluation of domain solutions

Index of dispersion. The index of dispersion for the domain size distribution is defined as $D = \sigma_n^2/\langle n\rangle$, where $\langle n\rangle$ is the average size of a domain, and σ_n^2 is the variance. It measures how clustered or dispersed a given distribution is, compared to a normal distribution. If $D < 1$, it indicate that the domain sizes are all very similar. If $D > 1$, on the other hand, it means that the domain size distribution is over-dispersed and heterogeneous, which may be the case when there are a few large domains and many small ones.

Similarity of two distinct CD solutions using Pearson correlation. In order to measure the extent of similarity between two CD solutions $\mathbf{s}$ and $\mathbf{s}'$, we consider the Pearson correlation at the level of loci pairs. We start by constructing the binary matrices $\mathbf{B}$ and $\mathbf{B}'$ that represent the two CD solutions, where $(\mathbf{B})_{ij} = B_{ij} = \delta_{s_i s_j}$, such that the matrix element are all 1's within the same CD and 0 otherwise. Considering the lower triangular elements of $\mathbf{B}$, we can calculate the mean $\bar{b} = \frac{2}{N(N-1)}\sum_{i<j} B_{ij}$ and the variance $\sigma_B^2 = \frac{2}{N(N-1)}\sum_{i<j}(B_{ij}-\bar{b})^2$; similarly, $\bar{b}'$ and $\sigma_{B'}^2$ for $\mathbf{B}'$. The similarity between $\mathbf{B}$ and $\mathbf{B}'$ is quantified with the Pearson correlation

$$\rho = \frac{\mathrm{cov}(B, B')}{\sigma_B \sigma_{B'}}, \tag{14}$$

where the element-wise covariance is $\mathrm{cov}(B, B') = \frac{2}{N(N-1)}\sum_{i<j}(B_{ij}-\bar{b})(B'_{ij}-\bar{b}')$.

Normalized mutual information. We use the mutual information to evaluate how well a CD solution $\mathbf{s}$ captures the visible patterns in the pairwise correlation data. We consider the binary grouping matrix $(\mathbf{B})_{ij} = B_{ij} = \delta_{s_i,s_j}$ for the CD solution of interest, and compare it to the input data matrix $(\mathbf{A})_{ij} = A_{ij}$. In this study, either $\log_{10}\mathbf{M}$ or $\mathbf{C}_{\mathrm{O/E}}$ was used in the place of $\mathbf{A}$. Assuming that we can treat the matrix elements $a \in \mathbf{A}$ and $b \in \mathbf{B}$ as two random variables, we can construct the joint distribution $p(a, b)$ by binning and counting, as:

$$p(a, b) = \frac{2}{N(N-1)}\sum_{i<j}\delta_{A_{ij},a}\delta_{B_{ij},b}, \tag{15}$$

where the Kronecker delta for the continuous variable a should be understood in a discretized fashion. That is, $\delta_{A_{ij},a} = 1$ if $A_{ij} \in [a, a+\Delta a)$ and 0 otherwise, where we used $\Delta a = [\max\{A_{ij}\} - \min\{A_{ij}\}]/100$ to discretize the values into 100 bins. It is straightforward to obtain the marginal distributions as $p(a) = \sum_b p(a, b)$ and $p(b) = \sum_a p(a, b)$. We can use the standard definitions to calculate the marginal entropies, $H(a) = -\sum_a p(a)\log(p(a))$ and $H(b) = -\sum_b p(b)\log(p(b))$, as well as the mutual information

$$I(a; b) = \sum_a\sum_b p(a, b)\log\left[\frac{p(a, b)}{p(a)p(b)}\right]. \tag{16}$$

Note that the sum runs over the discretized values of a that are the endpoints of the bins used for counting, and over $\{0, 1\}$ for the binary variable b. Finally, the normalized mutual information (nMI) is defined as

$$\mathrm{nMI}(a; b) = \frac{I(a; b)}{\sqrt{H(a)\,H(b)}}. \tag{17}$$

Hierarchy score. We define the hierarchy score to quantify the hierarchical relationship between two CD solutions, $\mathbf{s}$ and $\mathbf{s}'$. We assume that we know the average domain sizes for the two solutions: we will say that $\mathbf{s}$ is a set of smaller CDs and $\mathbf{s}'$ a set of larger CDs. Then the perfect hierarchy condition can be defined as in the following statement: if two loci i, j belong to the same CD in the smaller-scale $\mathbf{s}$ ($s_i = s_j$), then they also belong to the same CD in the larger-scale $\mathbf{s}'$ ($s'_i = s'_j$). Here we extend this idea to evaluate the extent of overlap of $\mathbf{s}$ to $\mathbf{s}'$. To begin, for each domain $k \in \mathbf{s}$ in the smaller-scale CD solution, the best overlap of this domain on $\mathbf{s}'$ is quantified by the single-domain hierarchy score h_1:

$$h_1(k \to \mathbf{s}') = \max_{k' \in \mathbf{s}'}\left[\frac{\sum_i \delta_{s_i,k}\,\delta_{s'_i,k'}}{\sum_i \delta_{s_i,k}}\right] \tag{18}$$

We get a maximum score $h_1(k \to \mathbf{s}') = 1$ if there is a $k' \in \mathbf{s}'$ such that the smaller domain $k \in \mathbf{s}$ is completely included in the larger domain $k' \in \mathbf{s}'$. On the other hand, the worst score is obtained when the domains in the two CD solutions $\mathbf{s}$ and $\mathbf{s}'$ are completely uncorrelated, in which case h_1 only reflects the overlap "by chance". The chance level is written as $\bar{h}_1(\mathbf{s}') = \langle n\rangle_{\mathbf{s}'}/N$, where $\langle n\rangle_{\mathbf{s}'} = \langle\sum_i \delta_{s'_i,k'}\rangle_{k' \in \mathbf{s}'}$ is the average domain size of the larger solution $\mathbf{s}'$. This naturally defines a normalized score

$$\hat{h}_1(k \to \mathbf{s}') = \frac{h_1(k \to \mathbf{s}') - \bar{h}_1(\mathbf{s}')}{1 - \bar{h}_1(\mathbf{s}')}. \tag{19}$$

Consequently, the hierarchy score $h(\mathbf{s} \to \mathbf{s}')$ of the entire CD solution $\mathbf{s}$ on $\mathbf{s}'$ is calculated as a weighted sum of h_1 as

$$h(\mathbf{s} \to \mathbf{s}') = \sum_{k \in \mathbf{s}}\hat{h}_1(k \to \mathbf{s}')\frac{n_k}{N}, \tag{20}$$

where $n_k = \sum_i \delta_{s_i,k}$ is the size of domain k in $\mathbf{s}$. In this study, we are interested in $h(\mathbf{s}^{\lambda_1} \to \mathbf{s}^{\lambda_2})$ for CD solutions evaluated at two distinct values $\lambda_1 < \lambda_2$, knowing that the average domain sizes are $\langle n\rangle_{\mathbf{s}^{\lambda_1}} < \langle n\rangle_{\mathbf{s}^{\lambda_2}}$.

Correlation between CTCF signal and domain boundaries. The validity of domain boundaries, determined from various CD-identification methods including Multi-CD, is assessed in terms of their correlation with the CTCF signal. We write $\phi_{\text{CTCF}}(i)$ to indicate the CTCF signal at locus i. We also define a binary variable $\psi_{\text{DB}}(i)$ that indicates the boundaries of a CD solution $\mathbf{s}$, such that $\psi_{\text{DB}}(i)$ is 1 if the i-th locus is precisely in the domain boundary, $\psi_{\text{DB}}(i) = 0$, otherwise ($\psi_{\text{DB}}(i) = (1 - \delta_{s_i,s_{i-1}}\delta_{s_i,s_{i+1}})$). We evaluated a distance-dependent, normalized overlap function $\chi(d)$, defined as

$$\chi(d) = \frac{\langle \delta\phi_{\text{CTCF}}(i+d)\psi_{\text{DB}}(i)\rangle_i}{\langle \psi_{\text{DB}}\rangle} \approx \frac{\sum_{i=1}^{N-d} \phi_{\text{CTCF}}(i+d)\psi_{\text{DB}}(i)}{\sum_{i=1}^{N} \psi_{\text{DB}}(i)} - \frac{1}{N}\sum_{i=1}^{N} \phi_{\text{CTCF}}(i), \quad (21)$$

where $\delta\phi_{\text{CTCF}} = \phi_{\text{CTCF}} - \langle \phi_{\text{CTCF}}\rangle$ and the approximation sign is used because of $\frac{N}{N-d} \approx 1$ for $N \gg d$. If the domain boundaries determined from a CD-identification method is correlated with TAD-capturing CTCF signal, a sharply peaked and large amplitude overlap function ($\chi(d)$) is expected at $d = 0$.

Correlation between epigenetic marks and compartments. We calculate the correlation of our compartment solutions with the epigenetic marks. Given a compartment solution $\mathbf{s}$ with two large domains A and B, we consider two binary vectors $\mathbf{q}^{(A)}$ and $\mathbf{q}^{(B)}$, such that $q_i^{(A)} = 2(\delta_{s_i,A} - \frac{1}{2})$ and $q_i^{(B)} = 2(\delta_{s_i,B} - \frac{1}{2})$. In other words, we set $q_i^{(A)} = +1$ if the i-th locus belongs to compartment A, and $q^{(A)} = -1$ if not. Let $\mathbf{h}$ be a set of epigenetic marks measured across the genome, such that h_i indicates the value at the i-th locus. Then we calculate the Pearson correlations, $c_A = (\mathbf{h}\cdot\mathbf{q}^{(A)})/|\mathbf{h}||\mathbf{q}^{(A)}|$ and $c_B = (\mathbf{h}\cdot\mathbf{q}^{(B)})/|\mathbf{h}||\mathbf{q}^{(B)}|$, between the epigenetic mark $\mathbf{h}$ and the two compartments, respectively.

Code availability

The Matlab software package and associated documentation are available online (https://github.com/multi-cd).

References for Methods

[60] Consortium, E. P et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57 (2012).

[61] Knight, P. A & Ruiz, D. A fast algorithm for matrix balancing. *IMA J. Numer. Anal.* **33**, 1029–1047 (2013).

Supplementary Figures

Fig S1. Distance distributions of loci pairs are described by Gaussian.

Fig S2. Scale-to-scale similarity of domain solutions for different cell lines.

Fig S3. Examples of Multi-CD domain solutions at different scales.

Fig S4. Chromatin domain solutions for chromosomes 4, 10, 11 and 19.

Fig S5. Intra-domain contact profiles for the two compartment solutions.

Fig S6. Comparison of domain solutions from Multi-CD and other methods at specific scales.

Fig S7. Comparison of histone marks and compartments.

Fig S8. Identification of sub-TAD boundaries at 5-kb resolution.

Fig S9. Comparison to existing algorithms for identifying domains at multiple scales.

Fig S10. Finding the best domain solution through simulated annealing.

Fig S11. Robustness of clustering solutions over different subsets of Hi-C data.

Fig S12. Schematic of the Multi-CD algorithm.

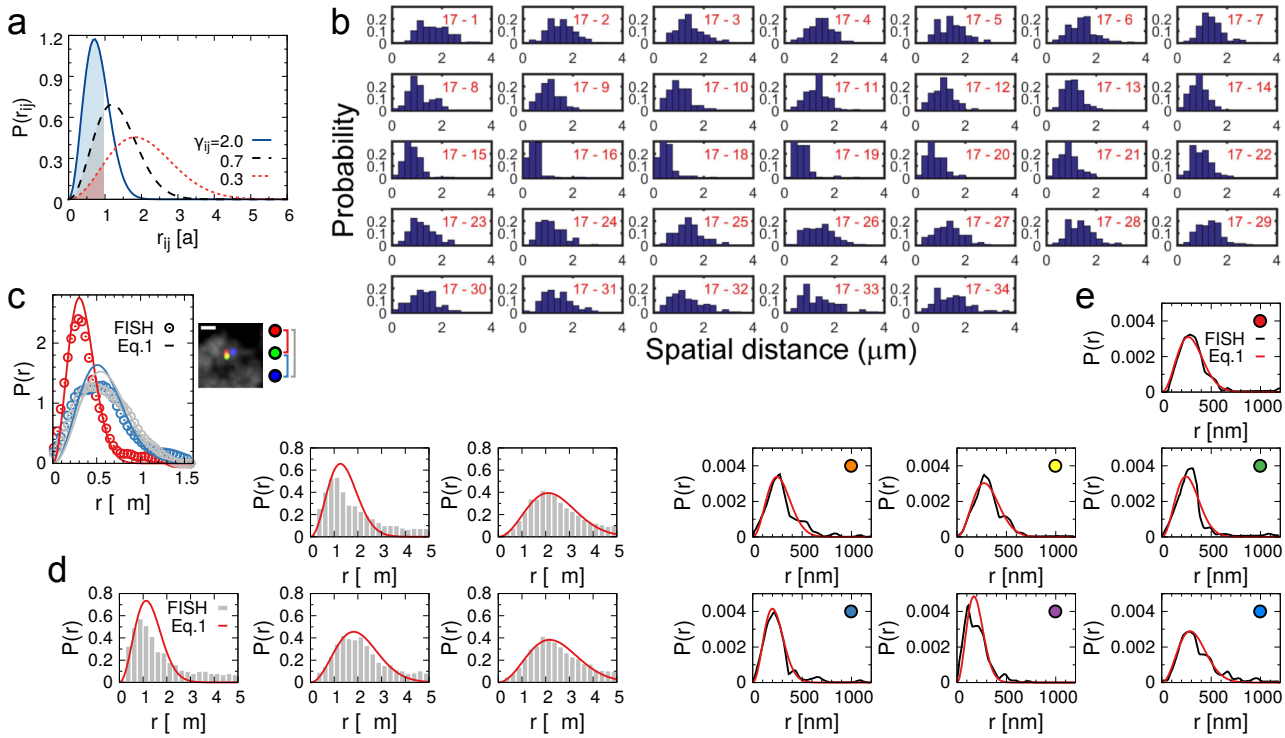

Figure S1: **Distance distributions of loci pairs are described by Gaussian. (a)** Gaussian probability distribution plotting $P(r_{ij})$ with different values of γ_{ij} (Eq. 1). The shaded area in different colors represents the corresponding values of contact probabilities, (Eq. 2 at $r_c = 1$) **(b)** Distance distributions between one TAD (TAD17) and other TADs on Chr21 in human IMR90 cells measured with FISH. This figure was adapted from Fig.S3 in [22]. **(c)** Distance distributions between three FISH probes on the X chromosome of male *Drosophila* embryos. The experimental data were digitized from Fig.3B in [57]. Their best fits to Eq. 1 are plotted with solid lines. **(d)** Distance distributions between five pairs of FISH probes on chr1 in fibroblast cells. The experimental data (histograms) were digitized from Fig.4B in [58]. The fits using Eq. 1 are plotted with solid lines. **(e)** Distance distributions between seven pairs of FISH probes in the Tsix/Xist region on the X chromosome of mouse ESC. The experimental data (black lines) were digitized from Fig.2F in [59], and their corresponding fits are shown in red.

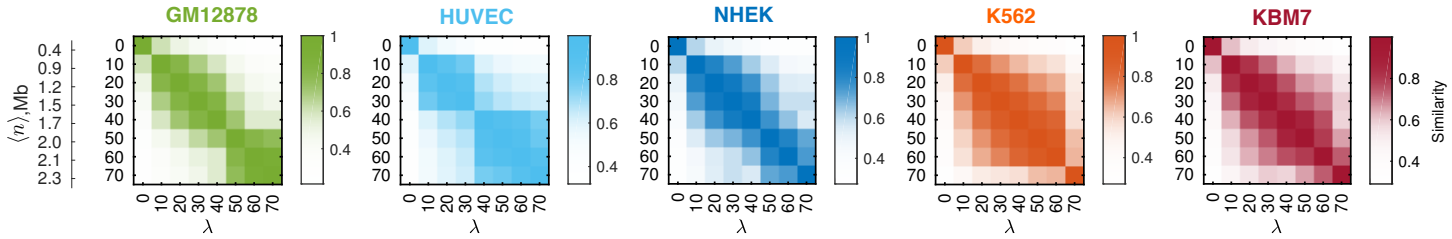

Figure S2: **Scale-to-scale similarity of domain solutions for different cell lines.** We calculate the similarity between domain solutions at different λ in terms of Pearson correlation. The calculation was performed for chromosome 10 from five different cell lines.

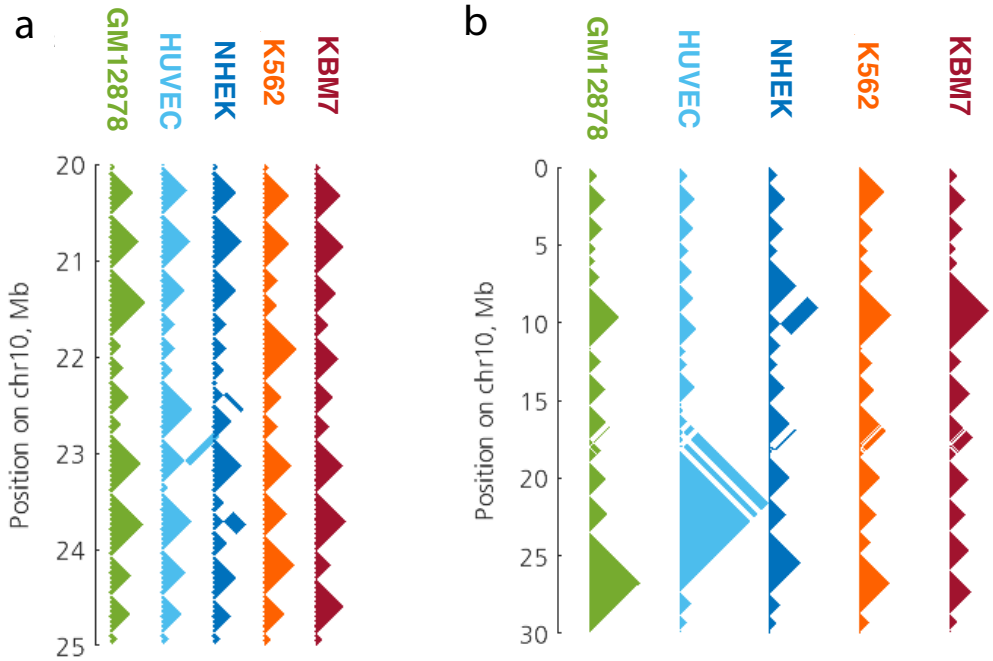

Figure S3: **Examples of Multi-CD domain solutions at different scales.** Shown are the domain solutions obtained from Multi-CD for the five different cell lines (GM12878, HUVEC, NHEK, K562, KBM7), at (a) $\lambda = 0$ and (b) $\lambda = 40$.

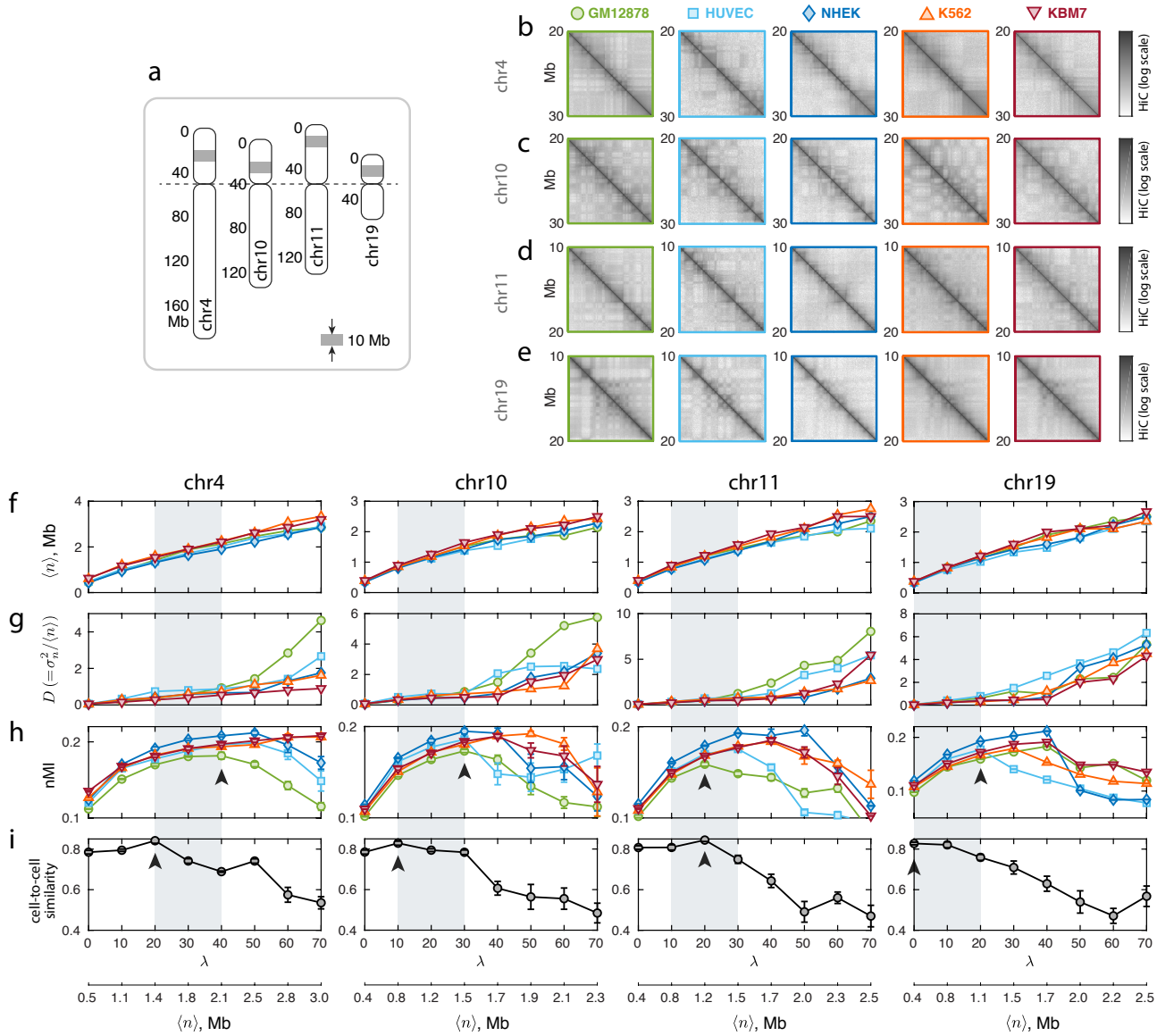

Figure S4: **Chromatin domain solutions for chromosomes 4, 10, 11 and 19**. **(a)** Relative sizes of chromosomes considered, aligned at the centromeres. The gray shade in each chromosome indicates the 10-Mb interval for which we show the Hi-C data in the next panels. **(b-e)** Hi-C data for the corresponding 10-Mb genomic intervals of **(b)** chr4, **(c)** chr10, **(d)** chr11, and **(e)** chr19, for the five different cell lines respectively. All the panels for chr10 are reprints of Fig. 2 in the main text. **(f-i)** Statistics of the domain solutions for chr4, chr10, chr11, and chr19. The five cell lines are color coded as indicated at the top of **(b)**. **(f)** Mean domain size $\langle n \rangle$ as a function of λ. **(g)** The index of dispersion $D(= \sigma_n^2/\langle n \rangle)$ of domain sizes. **(h)** The goodness of domain solutions, measured in terms of the normalized mutual information with respect to Hi-C data ($\log_{10} \mathrm{M}$). **(i)** The similarity of domain solutions across the five different cell types, measured by the Pearson correlation between binarized contact matrices. For each chromosome, arrows indicate the likely TAD scale (highest cell-to-cell similarity) and the likely meta-TAD scale (where the nMI is high and the index of dispersion D starts to diverge).

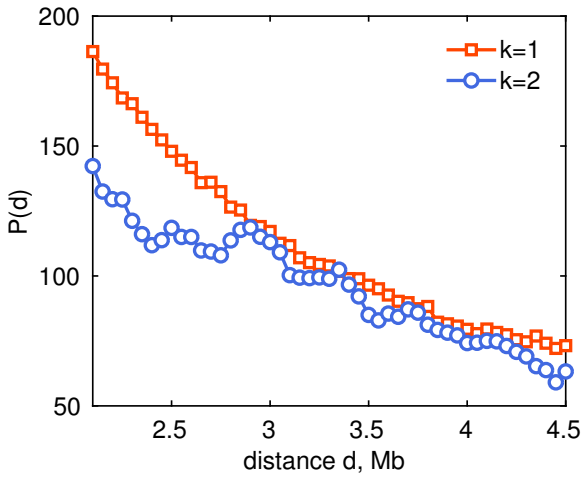

Figure S5: **Intra-domain contact profiles for the two compartment solutions.** We plot the genomic distance-dependent contact number for the two large domains $k = 1$ and $k = 2$, from the solution at $\lambda = 90$ for chromosome 10 in GM12878. At short genomic distance, the domain solution of $k = 1$ is characterized with a greater number of contacts than $k = 2$, which suggests that $k = 1$ domain is locally more compact. We therefore associate the first domain $k = 1$ to the B-compartment, and the second domain $k = 2$ to the A-compartment.

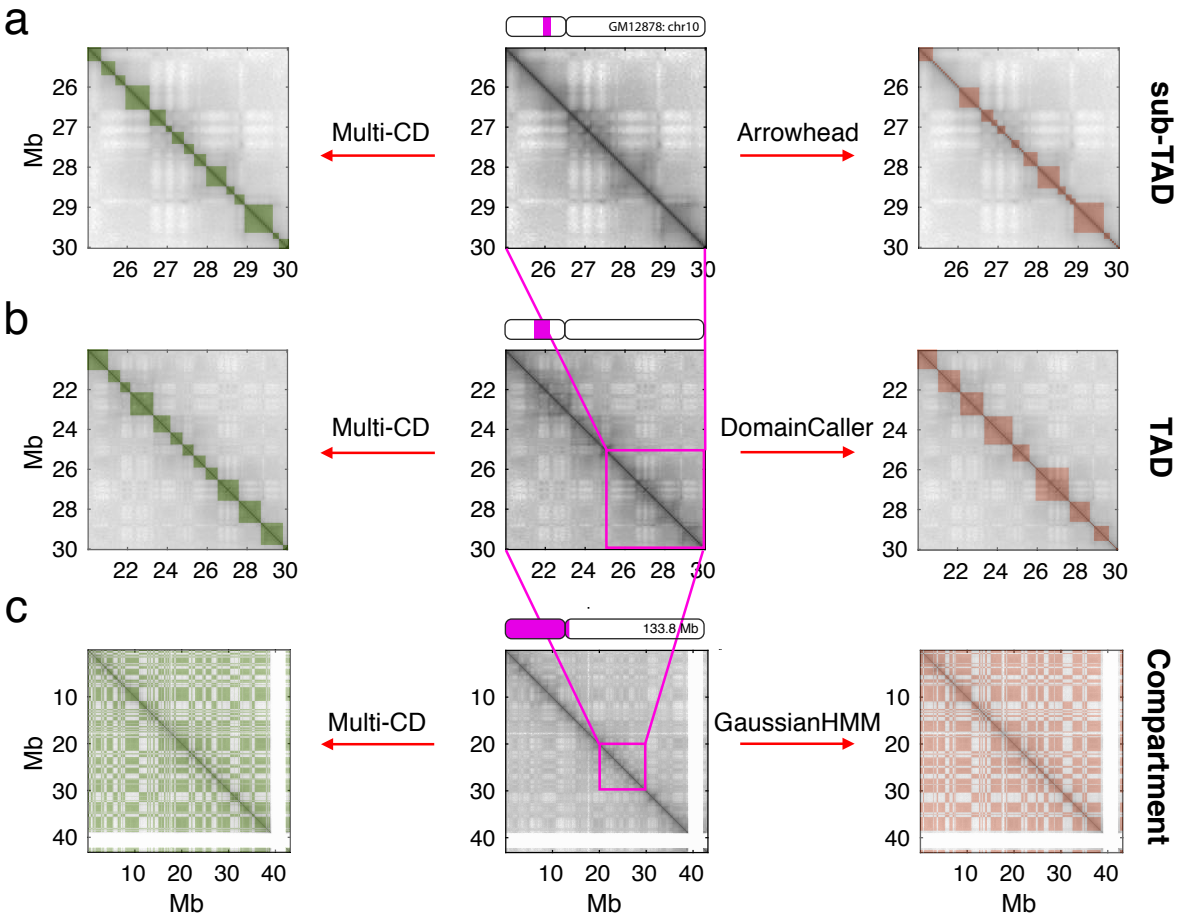

Figure S6: **Comparison of domain solutions from Multi-CD and other methods at specific scales.** Comparison between domain solutions obtained by three popular algorithms (ArrowHead, DomainCaller, GaussianHMM) (right column) and those by Multi-CD (left column), applied to 50-kb resolution Hi-C data. Three subsets from the same Hi-C data ($\log_{10} \mathbf{M}$), with different magnification (5, 10, and 40 Mb from top to bottom), are given in the middle column. ArrowHead algorithm [17] was used for identifying the domain structures of sub-TADs, DomainCaller [24] for TADs, and Gaussian Hidden Markov Model (GaussianHMM) [17] for compartments. Multi-CD use $\lambda = 0$, 10, 90, as the parameter values for identifying sub-TADs, TADs, and compartments, respectively.

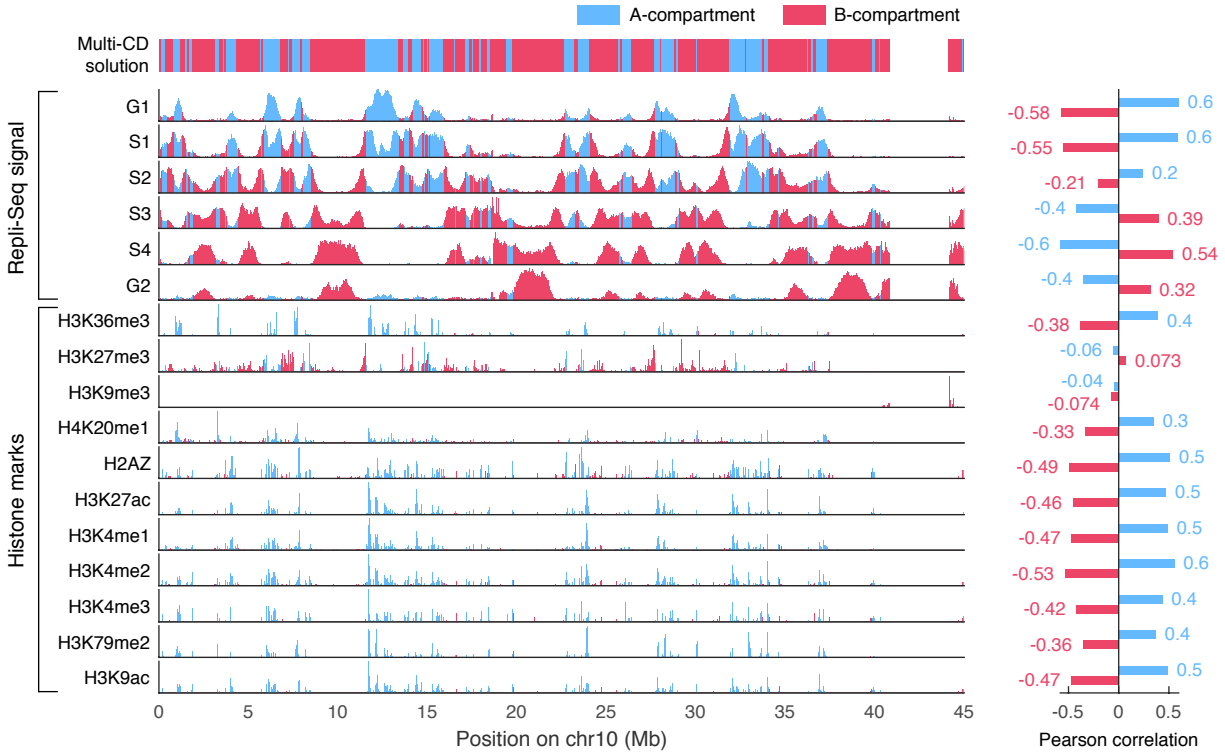

Figure S7: **Comparison of histone marks and compartments.** Extension of Fig. 5c-d which make comparison between the CD solutions for A/B compartments by Multi-CD and epigenetic marks. The upper part with Repli-Seq signals is a reprint from the main text figure. The lower part shows histone marks on the corresponding genomic range. Majority of the histone marks are correlated with the A-compartment. The values of Pearson correlation between Repli-Seq signal or histone marks and A/B compartment are given on the right.

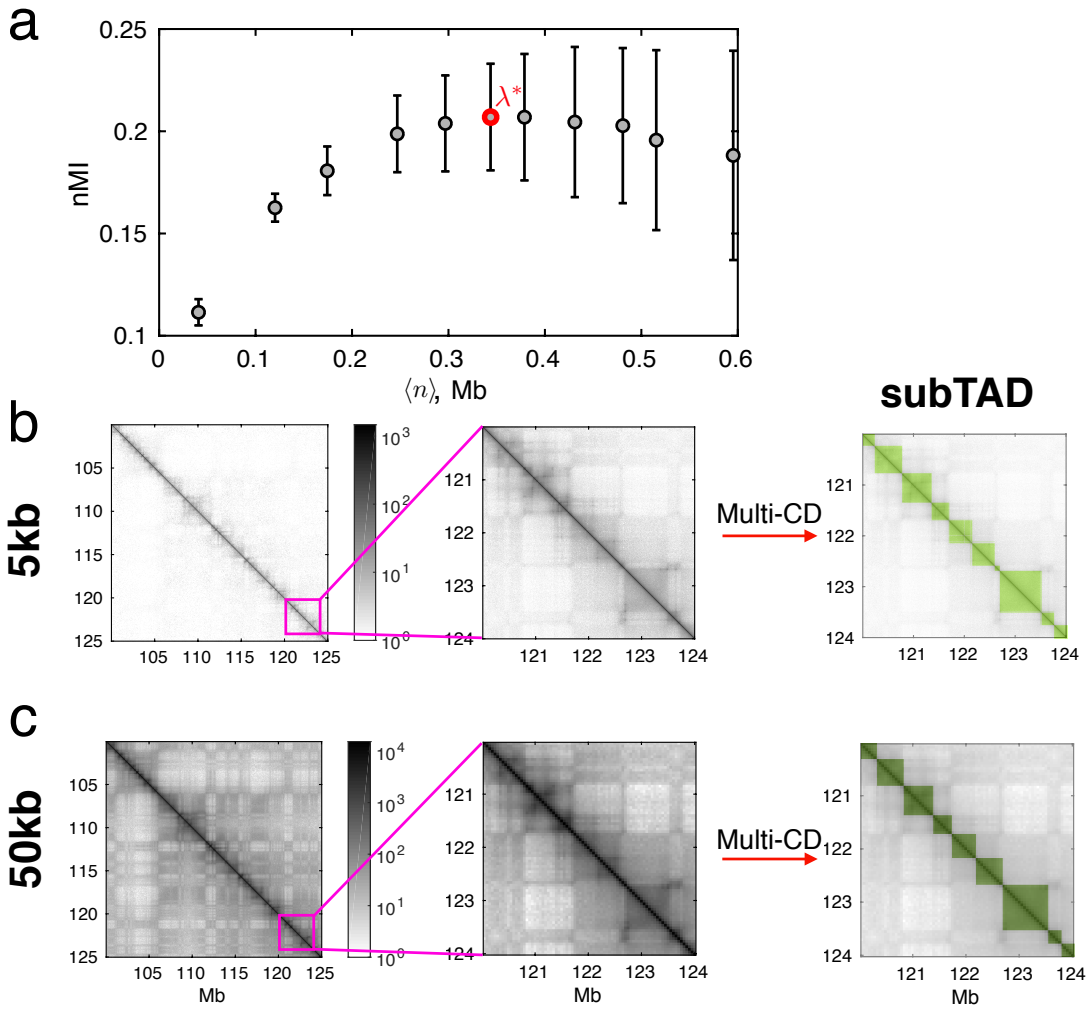

Figure S8: **Identification of sub-TAD boundaries at 5-kb resolution.** **(a)** The optimum cluster size, best describing 5-kb resolution Hi-C map in terms of nMI, is determined at $\langle n \rangle = 0.35$ Mb, which is consistent with the sub-TAD size determined from 50-kb resolution Hi-C at $\lambda = 0$. **(b-c)** Comparison between Multi-CD solutions at different resolutions of the input Hi-C data, that point to the robustness of sub-TAD boundaries regardless of Hi-C resolution. **(b)** The best CD solution (corresponding to $\lambda = \lambda^*$ in panel **(a)**) for the 5-kb resolution Hi-C data in the 120-124 Mb region of the genome. **(c)** Solution for the same genomic interval from 50-kb Hi-C, determined at $\lambda = 0$. The two CD solutions are effectively identical, which supports our interpretation of sub-TAD as the unit of hierarchical chromosome organization.

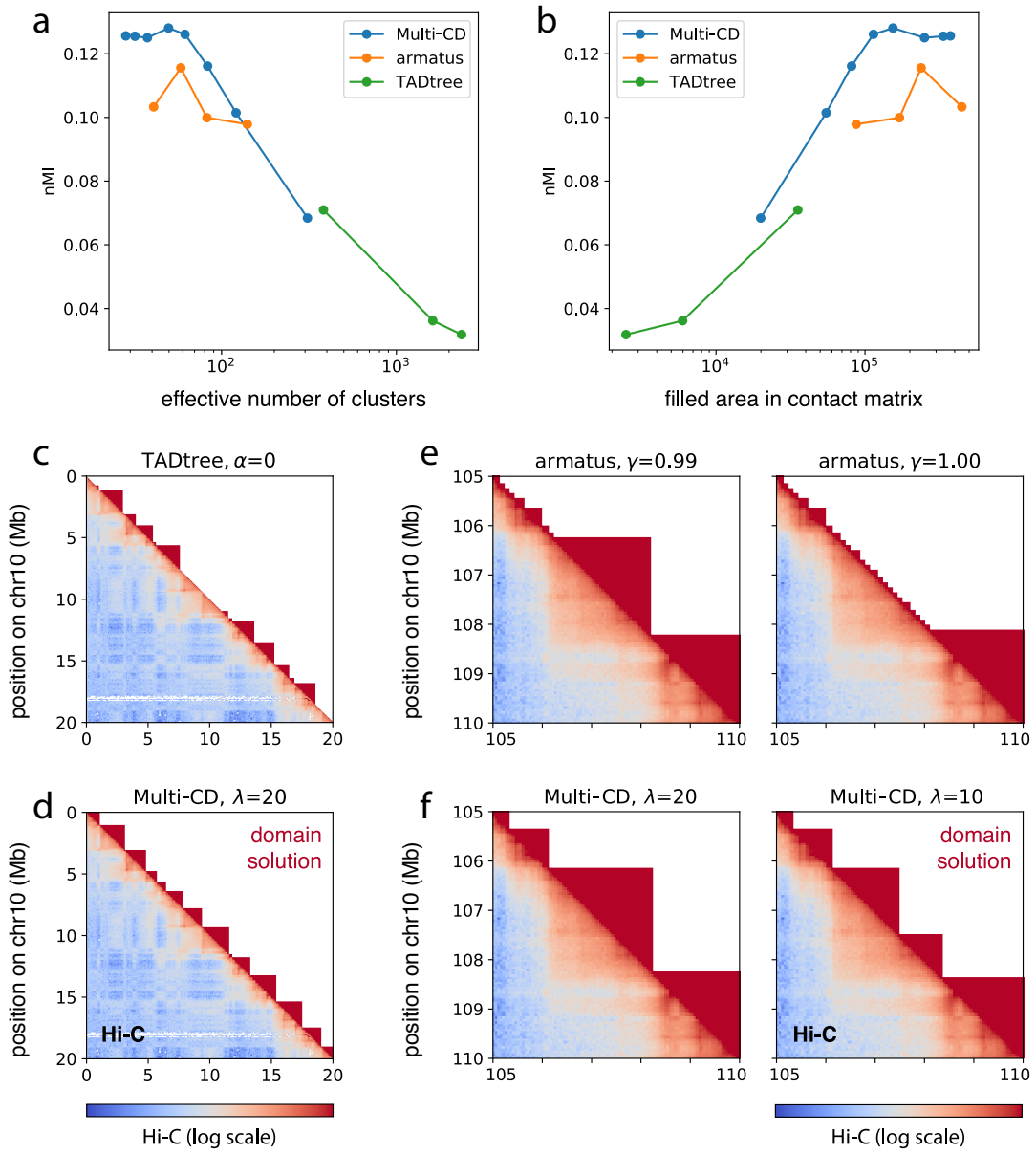

Figure S9: **Comparison to existing algorithms for identifying domains at multiple scales.** **(a,b)** Normalized mutual information between domain solutions at multiple scales, from Multi-CD, Armatus [33] and TADtree [32] respectively, and the log10 of KR-normalized Hi-C matrix for chr10 of the cell line GM12878. The scale of a domain solution $\mathbf{s}$ is measured in two ways, in terms of **(a)** the effective number of clusters, $\mathcal{K}(\mathbf{s}) = \exp\left(-\sum_{k=1}^{K_{\max}}(n_k/N)\log(n_k/N)\right)$, where $n_k = \sum_{i=1}^{N}\delta_{s_i,k}$ is the domain size; and **(b)** the total area of 1's in the corresponding binary contact matrix, $(\text{area}) = \sum_{i,j=1}^{N} B_{ij}$ where $B_{ij} = \delta_{s_i,s_j}$. All domain solutions from TADtree and Armatus were obtained using the respective default parameter settings. **(c-f)** Visual comparison of domains found by **(c,d)** TADtree and Multi-CD, and **(e,f)** Armatus and Multi-CD, at matching scales in terms of the average domain size. Domain solutions are shown in the upper triangle, colored by red (intra-domain) and white (extra-domain) for effective visualization. The lower triangle plots the corresponding subset of the Hi-C data (KR-normalized and in log10). Refer to the original papers [32, 33] for the definitions of the respective control parameters α (TADtree) and γ (Armatus).

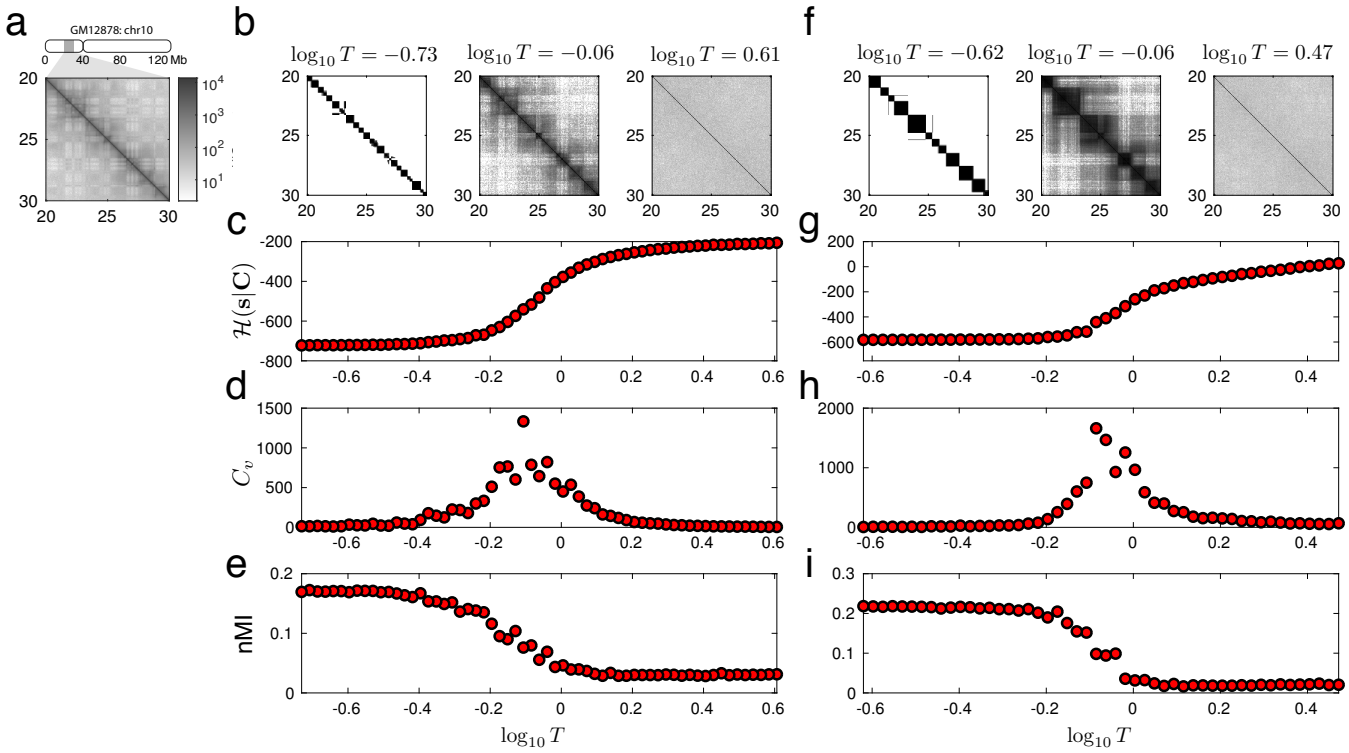

Figure S10: **Finding the best domain solution through simulated annealing.** **(a)** A subset of Hi-C data, covering 10-Mb genomic region on chr10 of GM12878. **(b)** CD solutions, obtained from the Hi-C data in **(a)**, at three values of T for $\lambda = 0$. The CD solution at each T was constructed by $2,000$ sample trajectories being equilibrated. **(c-e)** We plot three quantities over varying T, where the simulated annealing from high to low T (right to left in figure) was used as a sampling protocol. **(c)** The effective energy hamiltonian $\mathcal{H}(\mathbf{s}|\mathbf{C})$. **(d)** The heat capacity $C_v = \langle \delta \mathcal{H}^2 \rangle / T^2$. **(e)** The normalized mutual information (nMI) between the domain solution and Hi-C matrix ($\log_{10} \mathbf{M}$). **(f-i)** Same analyses repeated for $\lambda = 10$.

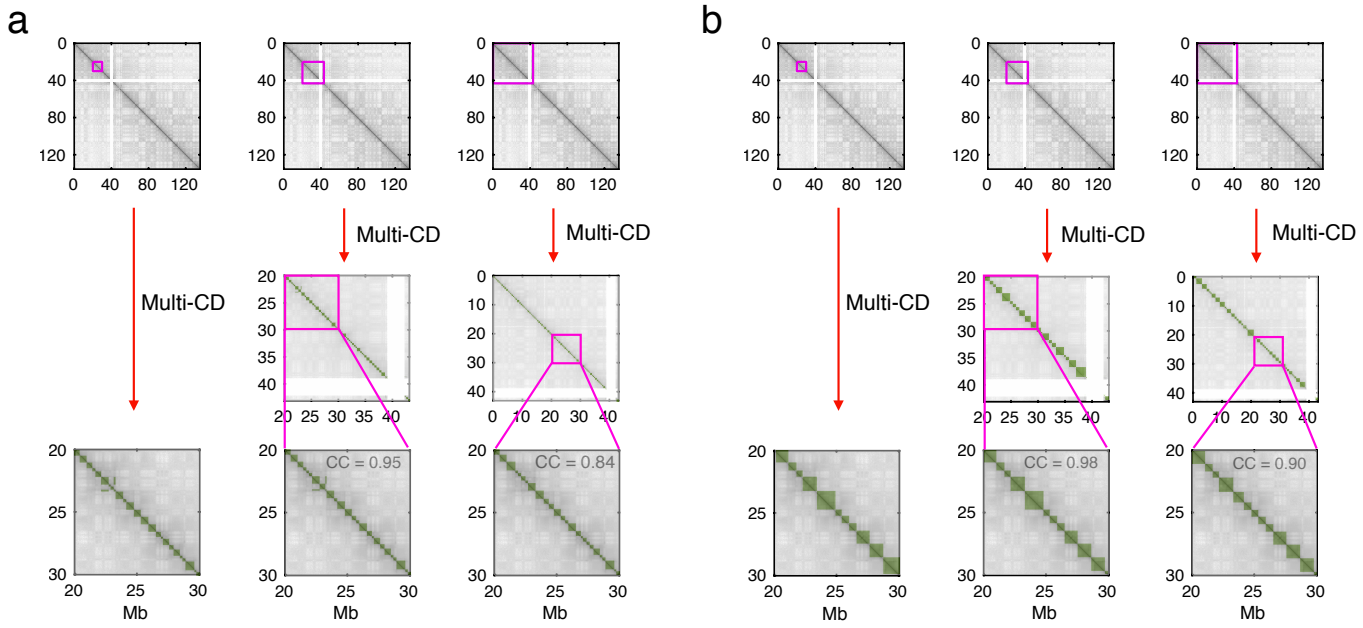

Figure S11: **Robustness of clustering solutions over different subsets of Hi-C data.** Here we compare domain solutions from Hi-C inputs of different size. Multi-CD is confirmed to be *locality-preserving*. That is, the sets of domain solutions determined from Hi-C inputs with different sizes remain almost identical to each other. The Hi-C data demarcated by the purple squares on the top panels are the input data used for Multi-CD analysis. The three panels from left to right on the bottom are the domain solutions from 10-Mb, 20-Mb, and 40-Mb Hi-C inputs. **(a)** For $\lambda = 0$, the correlation coefficients of 20-Mb Hi-C and 40-Mb Hi-C generated domain solutions with respect to the 10-Mb Hi-C generated one is 0.95 and 0.84, respectively. **(b)** Same calculations were carried out for λ=10.

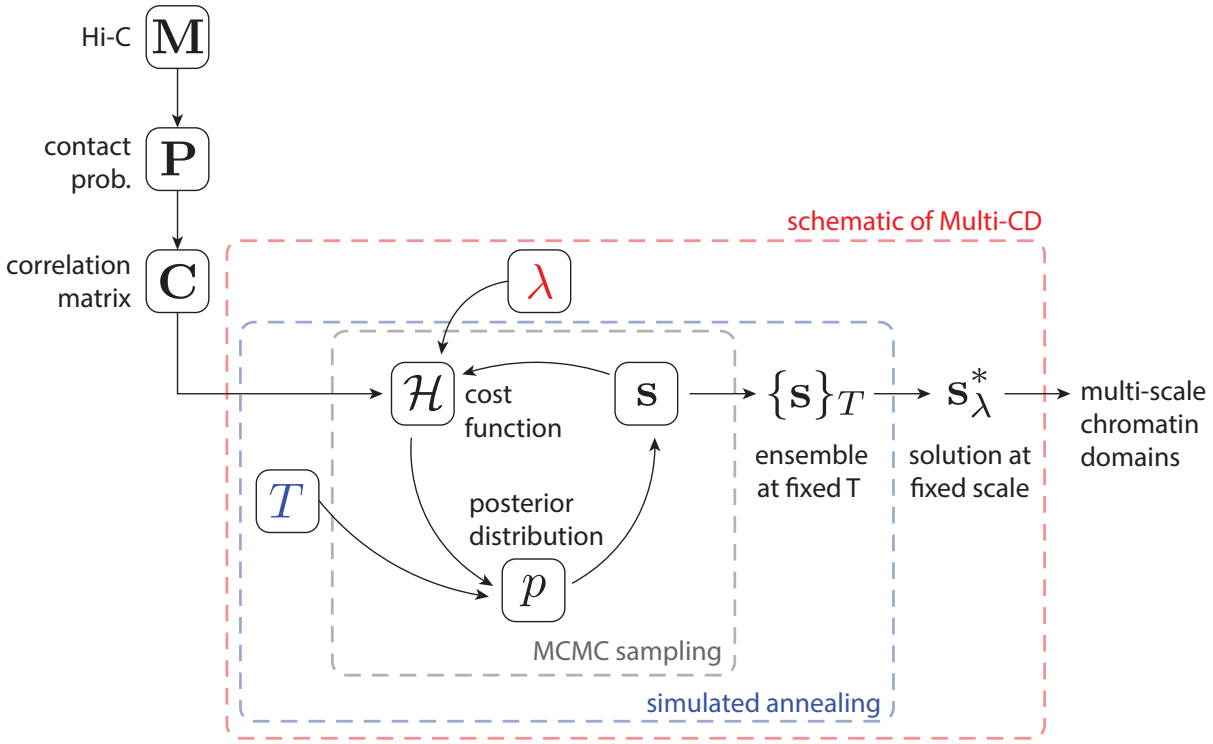

Figure S12: **Schematic of the Multi-CD algorithm.** The diagram illustrates the three levels of iterations in Multi-CD. Ultimately, Multi-CD identifies a family of chromatin domains at multiple scales as the control parameter λ is varied. At each fixed λ, the best domain solution is found through a simulated annealing, in which the effective temperature T is gradually decreased. At each fixed T, the tempered posterior distribution is approximated by the Markov chain Monte Carlo method, which samples multiple domain solutions, $\mathbf{s}$, according to the posterior distribution $p(\mathbf{s}|\mathbf{C}) \propto \exp(-\mathcal{H}(\mathbf{s}|\mathbf{C};\lambda)/T)$.