A unified framework for inferring the multi-scale organization of chromatin domains from Hi-C

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

[a]Korea Institute for Advanced Study, Seoul 02455, Korea; [b]Redwood Center for Theoretical Neuroscience, University of California, Berkeley, CA 94720, USA

This manuscript was compiled on April 6, 2020

Abstract: **Identifying chromatin domains (CDs) from high-throughput chromosome conformation capture (Hi-C) data is currently a central problem in genome research. Here we present a unified algorithm, Multi-CD, which infers CDs at various genomic scales by leveraging the information from Hi-C. By integrating a model of the chromosome from polymer physics, statistical physics-based clustering analysis, and Bayesian inference, Multi-CD identifies the CDs that best represent the global pattern of correlation manifested in Hi-C. The multi-scale intra-chromosomal structures compared across different cell types allow us 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 different organizational principles. (iii) Sub-TADs are the common building blocks of chromosome architecture. CDs obtained from Multi-CD applied to Hi-C data enable a quantitative and comparative analysis of chromosome organization in different cell types, providing glimpses into structure-function relationship in genome.**

polymer network | chromosome | clustering analysis | tunable group model | Bayesian inference

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 chromosome research (3, 4). Square-block and checkerboard patterns manifested in Hi-C data provide glimpses into the organization of chromatin chains inside cell nuclei. Despite cell-to-cell variations inherent to Hi-C data, which is in practice collected from a heterogeneous cell population, the cell-type specificity of chromosome architecture gleaned from Hi-C is still clear. Furthermore, the change of Hi-C pattern with the transcription activity and the phase of cell cycle underscores the functional roles of chromosome structure in gene regulation (5–14). Given that pathological states of chromatin are also manifested in Hi-C (15, 16), accurate characterization of chromatin domains (CDs) from Hi-C data is of great importance in advancing our quantitative understanding of the genome function.

Inside cell nuclei each chromosome made of $\sim \mathcal{O}(10^2)$ Mb DNA is segregated into its own territory (Fig. 1a) (17). 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 (18–21) (Fig. 1b). Topologically associated domains (TADs), detected at $\sim \mathcal{O}(10^{-1}) - \mathcal{O}(1)$ Mb (22–25), are considered the basic functional unit of chromatin organization and gene regulation because of their well-conserved domain boundaries across cell/tissue types (17, 20, 26, 27). It was suggested that the proximal TADs in genomic neighborhood aggregate into a higher-order structural domain termed "meta-TAD" (6). At smaller genomic scale, each TAD is further split into sub-TADs that display more localized contacts (19, 28–31) (Fig. 1c).

Currently there are many algorithms available to identify CDs from Hi-C data, making important contributions to understanding the intra-chromosome architecture. However, CDs identified using different algorithms or parameters display significant variations, and there is no generally accepted definition for the above-mentioned CD at each 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. In many of the existing algorithms specialized in finding CDs at particular genomic scales, Hi-C data should first be formatted at specific resolution (18, 19, 22). Although there are methods (32, 33) developed for identifying hierarchical domain structures of chromosomes, their algorithms rely on local pattern recognition analyses (18, 19, 22, 34), not implementing the physical viewpoint that chromosomes are a three dimensional object made of a long polymer (6, 14, 35–40).

Here, we interpret Hi-C data as pairwise contact probability matrix resulting from polymer networks whose inter-monomer distances are harmonically restrained. The cross-correlation matrix derived from Hi-C is used as the sole input for Multi-CD, the algorithm that we have developed to identify CD at varying genomic scales. Agreement of CD solutions from Multi-CD with the previous knowledge on chromatin organization as well as with information from bio-markers indicates the reliability of Multi-CD. A single algorithm-based solutions of CD allow us to assess the multi-scale structure of chromatin from a more unified perspective. We assert that amid the rapidly expanding volume of Hi-C data (10–12), Multi-CD holds good promise to quantitative and principled determination of chromatin organization.

Theory

Transforming Hi-C into a cross-correlation matrix using a polymer network model. A chromosome can be viewed as a polymer chain that is folded to a network structure characterized with multiple cross-links (35, 41, 42). Even in the interphase that displays less amount of activity than mitotic phase, continuous events of free energy consumption break the detailed balance condition, driving the chromosome out of equilibrium (43–47). However, chromosome dynamics in each

Author contributions: J.H.B., M.H.K., L.L., and C.H. designed and performed research, J.H.B., M.H.K., and L.L. analyzed data; J.H.B., M.H.K., and C.H. wrote the paper.

The authors declare no competing interest.

*J.H.B. and M.H.K. contributed equally to this work.

[1]To whom correspondence should be addressed. E-mail: hyeoncb@kias.re.kr

Data sharing: Code for the algorithm presented in this paper is available online (https://github.com/multi-cd).

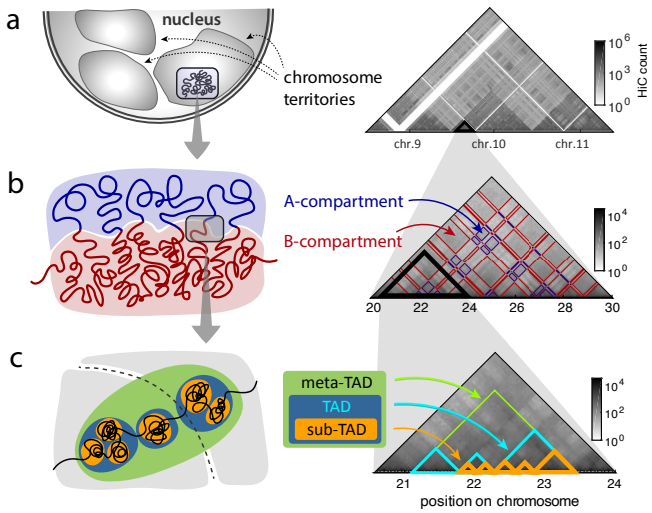

Fig. 1. The hierarchical organization of interphase chromosome and Hi-C map. **(a)** Chromosome territories in the cell nucleus, which are manifested as the higher intra-chromosomal counts in the Hi-C map. **(b)** Alternating blocks of active and inactive chromatins, segregated into A- and B-compartments, give rise to the checkerboard pattern on Hi-C. **(c)** Sub-megabase to megabase sized chromatin folds into TADs. Adjacent TADs are merged to meta-TAD (6), and individual TAD is further decomposed into sub-TADs (19, 28–31).

phase is slow (39, 40, 48, 49), such that the system remains in local mechanical equilibrium over an extended period of time, as captured by the stable patterns in the Hi-C data (14). Although a great amount of cell-to-cell variation is expected for a population of cells (39, 49, 50), fluorescence measurement indicates that the spatial distances between pairs of chromatin segments can be well described by the gaussian distribution (21, 51–53) (see Fig. S1). This motivates us to model the chromosome structure using a gaussian polymer network whose configuration fluctuates around its local mechanical equilibrium state (54, 55). See SI Appendix for more justifications for the use of the Gaussian polymer network model.

For a polymer chain whose long-range pairwise interactions are restrained via harmonic potentials with varying stiffness, the distance between each pair of segments i and j is written in the following form

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

with $\gamma_{ij}^{-1} = 2(\sigma_{ii}+\sigma_{jj}-2\sigma_{ij})$, where σ_{ij} is the positional covariance determined by the topology of polymer network (14, 42). The contact probability between the two segments, p_{ij}, is the probability that their pairwise distance r_{ij} is below a cutoff r_c. In other words, it is calculated using $p_{ij} = \int_0^{r_c} dr\, P(r;\gamma_{ij})$, where $P(r;\gamma_{ij})$ is the distance distribution in Eq. 1. Importantly, this model establishes a one-to-one mapping from the contact probability p_{ij} to the parameter γ_{ij}, which allows one to determine the covariance matrix $\{\sigma_{ij}\}$ and consequently the cross-correlation matrix, $(\mathbf{C})_{ij} = \sigma_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}}$ (see Materials and Methods for mathematical details). As a result, the following transformation from p_{ij} to $(\mathbf{C})_{ij}$ is conceived:

$$p_{ij} \longrightarrow \gamma_{ij} \longrightarrow \sigma_{ij} \longrightarrow (\mathbf{C})_{ij}. \qquad [2]$$

The resulting cross-correlation matrix $\mathbf{C}$ is used as the input data for our domain inference algorithm (see Fig. 2).

We propose a principled interpretation of Hi-C data as a *contact probability matrix*, which can be derived from a mathematically tractable yet physically meaningful model of gaussian polymer network. As a pre-processing method, our approach can replace the common but arbitrary use of nonlinear (most often logarithmic) scaling of the Hi-C data.

Modeling the correlations with the group model. Clustering a correlation matrix into a finite number of correlated groups is a general problem discussed in diverse disciplines. We adapted a statistical mechanical formalism known as the "group model," developed for identifying the correlated groups of companies from empirical data of stock market price fluctuations (56–58). Without ambiguity, the formalism can be applied to the clustering of correlated genomic segments in a chromosome.

Let us assume that each genomic segment $i \in \{1, 2, \cdots, N\}$ belongs to a chromatin domain s_i. Then the vector $\mathbf{s} = (s_1, s_2, \ldots, s_N)$ can be called the *domain solution* for the N segments. For example, a state $\mathbf{s} = (1, 1, 1, 2, 2, 3)$ describes a structure where the 6 genomic segments are clustered into 3 domains. Indexing of the domains is arbitrary. If there are K distinct domains in the solution, we can always index the domains such that $s_i \in \{1, 2, \cdots, K\}$.

We also assume that the cross-correlation matrix $\mathbf{C}$, captured by Hi-C, is essentially described by the correlation of a set of hidden variables $\{x_i\}$ where x_i represents the "genomic state" of the i-th chromatin segment. Without loss of generality we can only consider the case where x_i has zero mean and unit variance; or simply standardize as $(x_i - \langle x_i \rangle)/\sigma_{x_i} \to x_i$. Adapting the formalism in Refs. (56, 57), we assume that each x_i obeys the following stochastic equation

$$x_i = \sqrt{\frac{g_{s_i}}{1+g_{s_i}}}\eta_{s_i} + \frac{1}{\sqrt{1+g_{s_i}}}\epsilon_i \qquad [3]$$

where η_{s_i} and ϵ_i are two independent and identically distributed (i.i.d) random variables with $\eta_{s_i}, \epsilon_i \sim \mathcal{N}(0,1)$, that are linked to the domain (s_i) and the individual segment (i) respectively. The parameter $g_{s_i}(\geq 0)$ is associated with each domain s_i, such that a larger g_{s_i} indicates a stronger contribution from the domain-dependent variable η_{s_i}. The cross-correlation between two segments i and j is written as

$$\langle x_i x_j \rangle = \frac{g_{s_i}}{1+g_{s_i}}\delta_{s_i s_j} + \frac{1}{1+g_{s_i}}\delta_{ij}. \qquad [4]$$

In light of Eq. 4, the first term of Eq. 3 on the right hand side contributes to intra-CD correlation of the s_i-th CD with increasing g_{s_i}; the second term of Eq. 3 corresponds to a noise that randomizes the intra-domain correlation. Regarding $(\mathbf{C})_{ij}$ calculated from Hi-C as the correlation between the stochastic variables x_i and x_j, each representing the genomic state of segment (Eq. 4), namely,

$$(\mathbf{C})_{ij} \Leftrightarrow \langle x_i x_j \rangle, \qquad [5]$$

we use $\mathbf{C}$ as an input data for the clustering approach prescribed by the group model.

Our goal is to find the domain solution $\mathbf{s} = (s_1, s_2, \cdots, s_N)$ that best represents the pattern in the correlation matrix $\mathbf{C}$, with an appropriate set of strength parameters $\mathbf{g} = (g_1, g_2, \cdots, g_K)$, where K is the number of distinct domains in the solution. Using the group model described above, we can calculate the likelihood $p(\mathbf{C}|\mathbf{s}, \mathbf{g})$ that the observed correlation

matrix $\mathbf{C}$ was drawn from an underlying set of domains $\mathbf{s}$ with strengths $\mathbf{g}$ (see SI Appendix for derivation). The log-likelihood is written as a sum over all domains in the solution $\mathbf{s}$:

$$\log p(\mathbf{C}|\mathbf{s},\mathbf{g}) = -\frac{1}{2}\sum_{k=1}^{K}\left[(1+g_k)\left(n_k - \frac{g_k c_k}{1+g_k n_k}\right) - n_k \log(1+g_k) + \log(1+g_k n_k)\right], \quad [6]$$

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. The log-likelihood in Eq. 6 is maximized at $\hat{g}_k = (c_k - n_k)/(n_k^2 - c_k)$ for each k, allowing us to consider the reduced likelihood $p(\mathbf{C}|\mathbf{s}) \equiv \max_{\mathbf{g}} p(\mathbf{C}|\mathbf{s},\mathbf{g}) = p(\mathbf{C}|\mathbf{s},\hat{\mathbf{g}})$.

For convenience, we define an energy-like cost function $\mathcal{E}(\mathbf{s}|\mathbf{C})$ for a domain solution $\mathbf{s}$ given a correlation matrix $\mathbf{C}$, such that the problem of finding the maximum-likelihood solution $\mathbf{s}$ is equivalent to finding the $\mathbf{s}$ that minimizes $\mathcal{E}(\mathbf{s}|\mathbf{C})$. Specifically, it is useful to write the likelihood function as $p(\mathbf{C}|\mathbf{s}) \propto \exp(-\mathcal{E}(\mathbf{s}|\mathbf{C})/T)$ to resemble a Boltzmann distribution, with

$$\mathcal{E}(\mathbf{s}|\mathbf{C}) = \frac{1}{2}\sum_{k=1}^{K}\left[\log\frac{c_k}{n_k} + (n_k - 1)\log\frac{n_k^2 - c_k}{n_k^2 - n_k}\right], \quad [7]$$

where the new parameter $T > 0$ corresponds to the thermodynamic temperature. The best domain solution can be found by using the simulated annealing method where T is slowly decreased.

Inferring the domain solutions at multiple scales: a tunable group model. Besides evaluating how well a domain solution $\mathbf{s}$ explains the correlation pattern in the data, which is captured by the likelihood, we also want to impose an additional preference to more parsimonious solutions. This is done by introducing a prior distribution, $p(\mathbf{s})$. Here we use the form $p(\mathbf{s}) = \exp(-\lambda\mathcal{K}(\mathbf{s})/T)$, where $\mathcal{K}(\mathbf{s})$ is a function that gives a larger value for a more fragmented solution $\mathbf{s}$ (larger number of domains), and $\lambda\,(\geq 0)$ is a parameter that scales the overall strength of the prior. Specifically, $\mathcal{K}(\mathbf{s})$ is defined such that $\log\mathcal{K}(\mathbf{s})$ is the entropy of $\mathbf{s}$:

$$\mathcal{K}(\mathbf{s}) = \exp\left(-\sum_{k=1}^{K} p_k \log p_k\right), \quad p_k = n_k/N. \quad [8]$$

This quantity measures the effective number of domains; for example, $\mathcal{K}(\mathbf{s}) = K$ when the domain sizes are uniform.

Taking a Bayesian approach, we combine the likelihood and the prior, to construct the posterior distribution according to the Bayes rule: $p(\mathbf{s}|\mathbf{C}) \propto p(\mathbf{C}|\mathbf{s})\,p(\mathbf{s})$. The posterior is written in terms of an effective Hamiltonian $\mathcal{H}$:

$$p(\mathbf{s}|\mathbf{C}) \propto e^{-\mathcal{H}(\mathbf{s}|\mathbf{C})/T}, \quad \mathcal{H}(\mathbf{s}|\mathbf{C}) = \mathcal{E}(\mathbf{s}|\mathbf{C}) + \lambda\mathcal{K}(\mathbf{s}). \quad [9]$$

Finding the *maximum a posteriori* (MAP) estimate, $\mathbf{s}^* = \mathrm{argmax}_{\mathbf{s}}\, p(\mathbf{s}|\mathbf{C})$, is equivalent to solving a global minimization problem for $\mathcal{H}(\mathbf{s}|\mathbf{C})$. As the $\mathbf{s}$-space is expected to be high-dimensional and is likely characterized with multiple local minima, we use simulated annealing (59) to find the energy-minimizing $\mathbf{s}^*$ (Fig. S2) (see Materials and Methods for the details).

Our algorithm, Multi-CD, is a principled method for inferring the chromatin domain (CD) solutions at multiple scales.

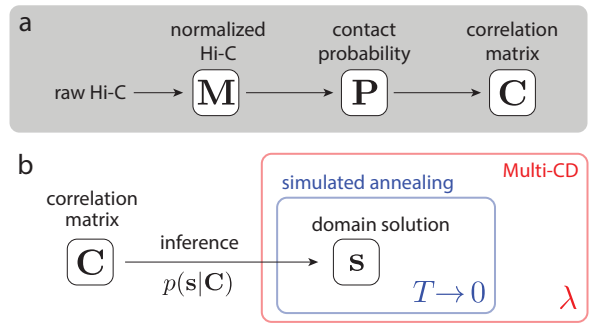

Fig. 2. Schematic of the Multi-CD algorithm. (a) Pre-processing: Hi-C data provide information of the correlation pattern. **(b)** Inference: for a given correlation matrix $\mathbf{C}$, the Multi-CD algorithm finds chromatin domain (CD) solutions $\mathbf{s}$ at multiple scales. The algorithm of Multi-CD is repeatedly applied to $\mathbf{C}$ to determine the best domain solutions at different values of λ (outer red box). At each λ, the best domain solution is found through a simulated annealing, in which the effective temperature T is gradually decreased (inner blue box). We use a MCMC sampling method to approximate the posterior distribution $p(\mathbf{s}|\mathbf{C}) \propto \exp(-\mathcal{H}(\mathbf{s}|\mathbf{C};\lambda)/T)$ and to find the domain solution $\mathbf{s}$ at each fixed temperature T. The boxes in red and blue represent the two levels of iterations varying λ and T, respectively, in the algorithm.

The group model becomes "tunable" at the level of inference through the parameter λ: as λ is increased, the resulting estimate $\mathbf{s}^*$ tends to have a fewer number of domains. In parallel to the statistical physics problem of a grand-canonical ensemble, T is the effective *temperature* of the system, and λ amounts to the negative *chemical potential*.

Results

Discovery of chromatin domains at multiple scales. Given a raw Hi-C matrix, one can use our Multi-CD algorithm for transforming it to a correlation matrix $\mathbf{C}$ (Fig. 2a), as well as for identifying a set of CDs for each fixed λ (Fig. 2b).

Here we applied Multi-CD on a sample subset of Hi-C data from a commonly used human lymphocyte cell line, GM12878, at 50-kb resolution. After the transformation of the raw Hi-C (Fig. 3a) into a correlation matrix (Fig. 3b), Multi-CD was employed to infer a family of CD solutions that vary with λ (Fig. 3c). We also applied Multi-CD to four other cell lines, HUVEC, NHEK, K562, and KBM7 (Fig. 3d), and analyzed the CD solutions for all five cell lines. All results shown in the main text consider chromosome 10; see Fig. S3 for similar results from three other chromosomes (chr4, 11, 19).

We observed several general features from the families of CD solutions in these cell lines:

(i) The average domain size $\langle n\rangle$ always increased monotonically with λ (Fig. 3e), as expected from our construction of the prior.

(ii) The domain sizes were relatively homogeneous in the small-domain regime (small λ), but became heterogeneous after a cross-over point (Fig. 3f). To quantify this, we defined the *index of dispersion* for the domain sizes, which is simply the variance-to-mean ratio $D = \sigma_n^2/\langle n\rangle$. If the domains were generated by randomly selecting the boundaries along the genome, the dispersion would be $D = 1$; a smaller $D < 1$ indicates that domain sizes are more homogeneous than random. A larger $D > 1$ means that the domain sizes are heterogeneous. The crossover points arise at $\lambda_{cr} \approx 30 - 40$ ($\langle n\rangle_{cr} \approx 1.6$ Mb) for GM12878, HUVEC, and NHEK; and $\lambda_{cr} \approx 60 - 70$ ($\langle n\rangle_{cr} \approx 2.2$ Mb) for K562 and KBM7. We

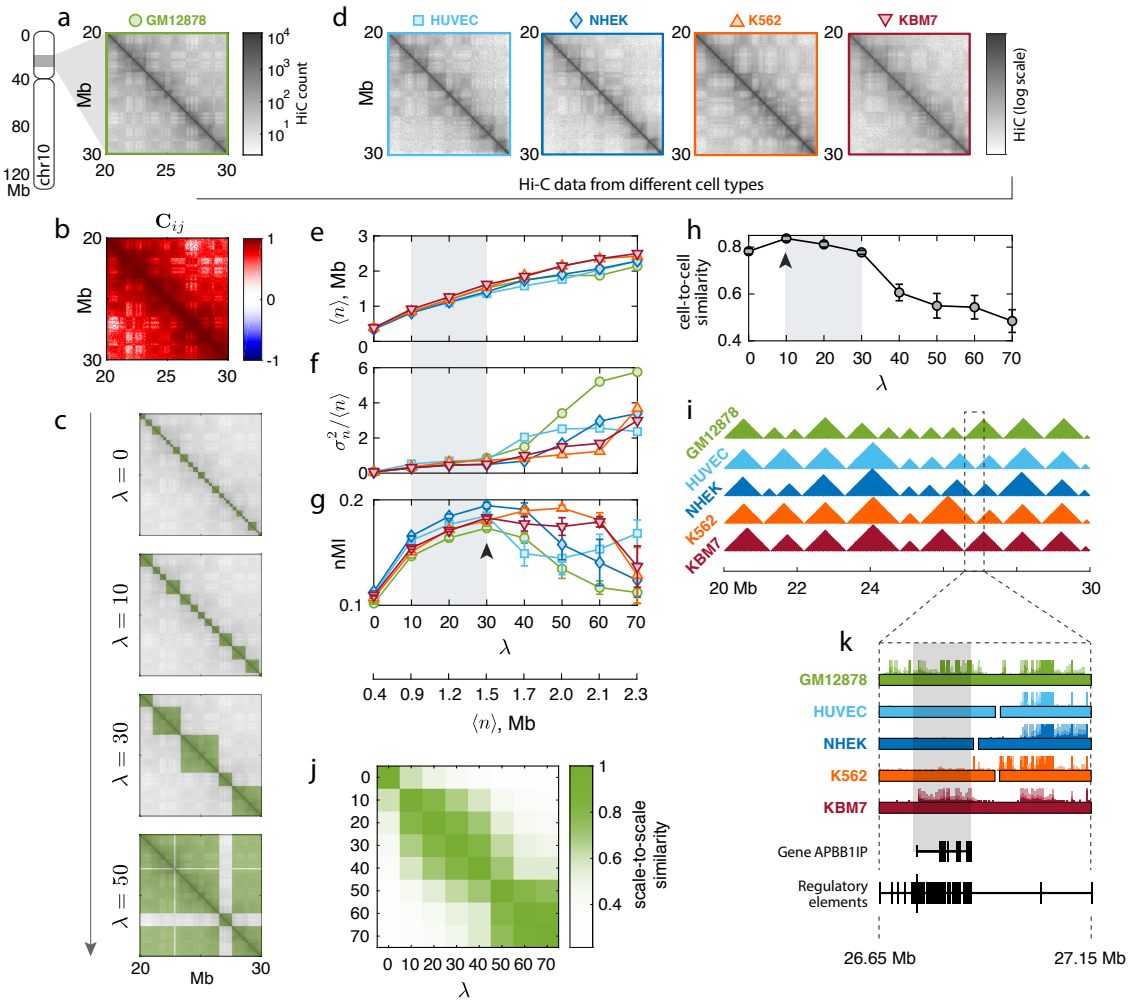

Fig. 3. Multi-scale chromatin domain solutions for various cell types. (a) A subset of 50-kb resolution Hi-C data, covering a 10-Mb genomic region of chr10 in GM12878. **(b)** The cross-correlation matrix $\mathbf{C}_{ij}$ for the corresponding subset. See Fig. S4 for a full-chromosome view. **(c)** Multi-CD applied to the correlation matrix in **b**. Domain solutions determined at 4 different values of $\lambda = 0, 10, 30, 50$. **(d)** Hi-C data from the same chromosome (chr10) in four other cell lines: HUVEC, NHEK, K562, and KBM7. Same subset as in **a**. **(e-g)** Characteristics of the domain solutions determined for all five cell lines in **a** and **d**: **(e)** the average domain size, $\langle n \rangle$; **(f)** the index of dispersion in the domain size, $D(= \sigma_n^2/\langle n \rangle)$; **(g)** the normalized mutual information, nMI. **(h-i)** Comparison of domain solutions across cell types. **(h)** Average cell-to-cell similarity of domain solutions, in terms of Pearson correlations, at varying λ. **(i)** Domain solutions obtained at $\lambda = 10$ for 5 different cell types. See Fig. S5 for solutions at $\lambda = 0$ and $\lambda = 40$. **(j)** Similarity between domain solutions at different λ's, shown for GM12878. See Fig. S6 for corresponding results for the other four cell lines. **(k)** RNA-seq signals from the five cell lines (colored hairy lines), on top of the TAD solutions (filled boxes), in a genomic interval that contains the regulatory elements associated with a gene APBB1IP. APBB1IP is transcriptionally active only in two cell lines, GM12878 and KBM7, where the regulatory elements are fully enclosed in the same TAD. See Fig. S7 for a larger figure.

observed that the onset of heterogeneity was related to the appearance of non-local domains (Fig. 3c).

(iii) We quantified the goodness of each CD solution by comparing its corresponding binary matrix against the Hi-C data in terms of the normalized mutual information (nMI; see Materials and Methods for details). There is a scale, λ^*, at which the diagonal block pattern manifested in Hi-C data is most accurately captured. In Fig. 3g, the best solution was found at $\lambda^* \approx 30$ for GM12878, HUVEC, and NHEK; the λ^*'s were identified at larger values for K562 and KBM7. As an interesting side note, K562 and KBM7 belong to immortalized leukemia cell lines, whereas the other three cell types are normal cells; the different statistical property of Hi-C patterns manifested in λ^* may hint at a link between the pathological state and a coarser organization of the chromosome.

(iv) The CD solutions inferred by Multi-CD, especially the families of local CDs, appeared to be *conserved* across different cell types (Fig. 3h-i). We quantified the extent of domain conservation, in terms of the Pearson correlation (see Materials and Methods), averaged over all pairs of different cell types. Domain conservation was strong for smaller domains at $\lambda \leq 30$ ($\langle n \rangle \lesssim 1.5$ Mb), with the strongest conservation at $\lambda = 10$ (Fig. 3h). The CDs at $\lambda = 10$ are shown in Fig. 3i for five different cell types.

(v) Finally, we quantified the similarity between pairs of CD solutions obtained at different scales, again using the similarity measure based on Pearson correlation. In the case of GM12878, the family of CD solutions is divided into two regimes; the smaller-scale CD solutions from a range of $10 \leq \lambda \leq 40$ are correlated among themselves, and the larger-scale CD solutions from $\lambda > 40$ as well. CD solutions below and above $\lambda \approx 40$ are not correlated with each other (Fig. 3j; also see Fig. S6 for the other cell lines).

The division boundary in (v) is found at a λ value in the sim-

ilar range with the best-clustering scale λ^*, and the crossover λ_{cr} from local/homogeneous to non-local/heterogeneous CDs (compare Fig. S6 to Fig. 3f-g). Hereafter we will refer to the two regimes as the family of *local* CDs with homogeneous size distribution ($\lambda \lesssim \lambda^*$), and the family of *non-local* CDs with heterogenous size distribution ($\lambda \gtrsim \lambda^*$).

TAD-like organizations in the family of local CDs. We identified at least three important scales in the family of local CDs. First of all, there was a scale at which domain conservation was maximized across different cells ($\lambda = 10$). This observation 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 (27, 60). Thus, for the example from human chromosome 10, we identify the CDs found at this scale $\lambda = 10$ as the TADs. The average domain size at $\lambda = 10$ was $\langle n \rangle \approx 0.9$ Mb, which agrees with the typical size of TADs as suggested by previous studies (22, 23).

The goodness of clustering, on the other hand, was maximized at a larger scale, $\lambda \approx 30$ ($\langle n \rangle \approx 1.5$ Mb) for GM12878. The CDs at this scale turned out to be aggregates of multiple TADs in the genomic neighborhood, from visual inspection (see Fig. 3c), or as quantified in terms of our nestedness score (see Materials and Methods). We therefore identify these CDs as the "meta-TADs", a higher-order structure of TADs, adopting the term of Ref. (6). In contrast to a previous analysis that extended the range of meta-TADs to the entire chromosome (6), we use the term meta-TAD exclusively for the larger-scale local CDs, distinguishing them from the non-local structures (i.e., compartments, discussed below). We note, however, that the terminologies of TADs and the meta-TADs are still not definitive – a recently proposed algorithm based on structural entropy minimization (61) found that the "best" solutions were found at ~ 2 Mb domains, which is consistent with our findings, although these domains were called the TADs in Ref. (61).

Finally, a trivial but special scale is $\lambda = 0$, where no additional preference for coarser CDs is imposed. The CDs at this scale are supposed to best explain the local correlation pattern that is reflected in the strong Hi-C signals near the diagonal. These smaller CDs are almost completely nested in the TADs and the meta-TADs; we can therefore call them the sub-TADs. We also confirmed that the sub-TAD solutions were not limited by the resolution of the Hi-C data; sub-TADs were robustly reproduced from a finer, 5-kb Hi-C (Fig. S8).

The first three panels in Fig. 3c shows three representative TAD-like CD solutions at $\lambda \leq \lambda^*$: sub-TADs ($\lambda = 0$; smallest CDs), TADs ($\lambda = 10$, strongest domain conservation), and meta-TADs ($\lambda = \lambda^* = 30$, largest nMI). The nested structure is reminiscent of the hierarchically crumpled structure of chromatin chains (37, 62).

Chromatin organization and its link to gene expression. The CD solutions from Multi-CD can shed important insights into the link between chromatin organization and gene expression. To demonstrate this, we overlaid the RNA-seq profiles on the TAD solutions, identified for the corresponding subset of the chr10 of five cell lines (GM12878, HUVEC, NHEK, K562, KBM7) (Fig. 3k; also see Fig. S7). At around 26.8 Mb position of this chromosome, we found a gene APBB1IP, which is transcriptionally active in GM12878 and KBM7 but not in HUVEC, NHEK and K562. Consulting the GeneHancer

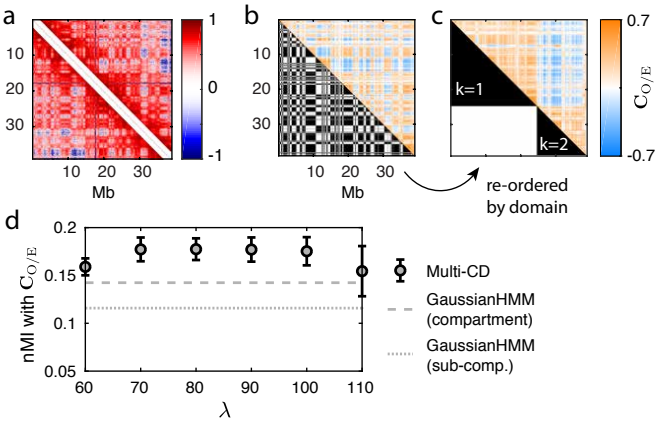

Fig. 4. Domain solutions for compartments. **(a)** Input correlation data for compartment identification. The 2-Mb diagonal band was removed. **(b)** Lower triangle: CDs obtained at $\lambda = 90$, based on the diagonal-band-removed data, which we identify as the compartments. Upper triangle: the $\mathbf{C}_{\mathrm{O/E}}$ matrix shown for comparison. **(c)** Same pair of data, after re-ordering to collect the two largest CDs in our solution (lower triangle), with $k = 1, 2$ as the B- and A-compartments respectively. The $\mathbf{C}_{\mathrm{O/E}}$ is simultaneously reordered to show a clear separation of correlation patterns (upper triangle). **(d)** nMI between CDs at varying λ and $\mathbf{C}_{\mathrm{O/E}}$, showing a plateau in the range $70 \leq \lambda \leq 100$. CDs inferred by Multi-CD show consistently higher nMI, compared to sub-compartments (dashed line) and compartments (dotted line) from a previous method (19).

database (63), we identified the regulatory elements for this gene (enhancers and promoters) in the interval between 26.65 and 27.15 Mb. Notably, our Multi-CD solutions show that the interval associated with the regulatory elements is fully enclosed in the same TAD in GM12878 and KBM7, whereas it is split into different TADs in the other three cell lines (Fig. 3k). The observation suggests that, for a gene to be expressed, it is critical that all regulatory elements are within the same TAD; this is consistent with the understanding that TADs define the functional boundaries for genetic interactions (5, 6, 17, 20).

Compartments as the best domain solution that coexists with TAD-like domains. The super-Mb sized domains are generally defined as the compartments in the chromosome organization (27). Compartments are characterized by the checkerboard pattern in off-diagonal part of correlation matrix, being highly non-local CDs. Our formulation of the group model has the flexibility for dealing with the non-local CDs naturally. However, a naïve application of Multi-CD by increasing λ did not identify the compartments; some non-local CDs were found (Fig. 3c), but they do not correspond to alternating patterns characteristic to compartments.

We hypothesize that compartments correspond to a secondary CD solution, that coexists with a best solution that was already identified. Assuming a statistical independence between the two solutions and the additivity of cross-correlations (extension of Eq. 3), the inference of the secondary solution is reduced to a standard Multi-CD applied to a modified input data, which is essentially the result of taking out the best CD solution from the original correlation matrix $\mathbf{C}$ (see SI Appendix). Here we consider a simplified version of this problem, and remove from $\mathbf{C}$ a diagonal band of width 2 Mb, similar to the size of meta-TADs (Fig. 4a). Applied to the modified Hi-C, Multi-CD successfully captures the non-local correlations, and identifies two large compartments with alternating patterns (Fig. 4b). The correspondence is clearer when the

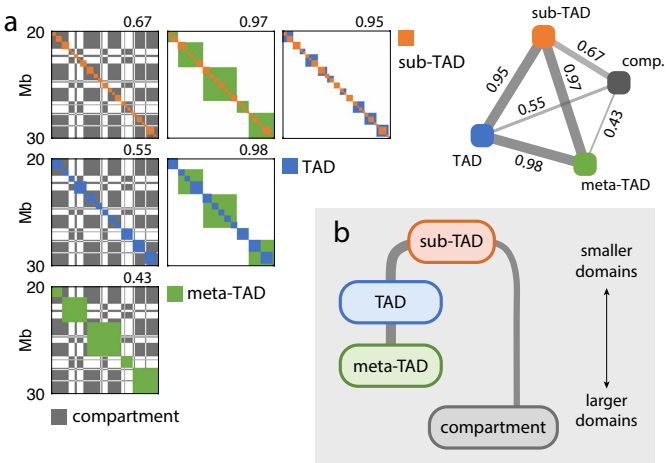

Fig. 5. Hierarchical organization of CD families (a) Hierarchical structure of CDs are highlighted with the domain solutions for sub-TADs (red), TADs (green), meta-TADs (blue) and compartments (black). Shown for chr10 of GM12878. Each square panel overlays a pair of CD solutions; number above the panel reports the nestedness score. Inset: a reprint of the nestedness scores in a tetrahedral visualization with the four representative CD solutions. **(b)** A schematic diagram of inferred hierarchical relations between sub-TADs, TADs, meta-TADs and compartments, based on our calculation of nestedness scores.

indices of segments are re-ordered (Fig. 4c). Because the larger CD ($k = 1$) shows a greater number of contacts (Fig. S9), it can be associated with the B-compartment, which is usually more compact; $k = 2$ is associated with the A-compartment. Further validation of the two compartments will be presented below, through comparisons with epigenetic markers.

To compare the goodness of our compartments with existing methods, we calculated the nMI against the $\mathbf{C}_{\mathrm{O/E}}$ matrix, the conventional form for compartment identification (see Materials and Methods) (18). We find that Multi-CD outperforms GaussianHMM (19), a widely accepted benchmark, in capturing the large-scale structures in Hi-C (Fig. 4d).

Multi-scale, hierarchical organization of chromatin domains. Now that we identified four classes of CD solutions, namely sub-TADs, TADs, meta-TADs and compartments, we examined their hierarchical relationships. Note that these CDs were obtained independently at the respective λ values, not through a hierarchical merging. Sub-TADs or TADs are almost always nested inside a meta-TAD, and TADs inside a meta-TAD, whereas there are mismatches between the TAD-like domains and the compartments. We quantified this relationship in terms of a nestedness score h, such that $h = 0$ indicates the chance level and $h = 1$ a perfect nestedness (see Materials and Methods), along with a visual comparison of each pair of CD solutions (Fig. 5a). This analysis confirms that the hierarchy between any pair of TAD-like domains (sub-TADs, TADs, and meta-TADs) is appreciably strong. On the other hand, the hierarchical links between the TAD-like domains and compartments are much weaker, which is again consistent with the recent reports that TADs and compartments are organized by different mechanisms (64, 65).

Although the nestedness score between sub-TADs and compartments (nestedness score $h = 0.67$) is not so large as those among the pairs of TAD-based domains, it is still greater than those between TADs and compartments ($h = 0.55$) or between meta-TADs and compartments ($h = 0.43$). Thus, sub-TAD can be considered a common building block of the chromatin architecture (see Fig. 5b).

Validation of domain solutions from Multi-CD. The CD solutions from Multi-CD are in good agreement with the results of several existing methods. Specifically, our CDs correspond to the previously proposed sub-TADs (19) at $\lambda = 0$, to the TADs (22) at $\lambda \approx 10$, and to the compartments (19) at $\lambda \approx 90$ (see Fig. S10). When assessed in terms of the nMI, Multi-CD outperforms the corresponding alternatives (ArrowHead (19), DomainCaller (22), GaussianHMM (19) for sub-TADs, TADs, meta-TADs) at the respective scales (Fig. 6a).

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 (66). All results shown here are for chr10 of GM12878.

First, we calculated how much the boundaries of our sub-TAD and TAD solutions are correlated with the CTCF signals, which are known to be linked to TAD boundaries (22, 23) (Fig. 6b). We quantified this in terms of a correlation function, $\chi(d)$, where d is the genomic distance between a domain boundary and each CTCF signal (see Materials and Methods). The correlation function shows a strong enrichment of CTCF signals at domain boundaries (high peak of correlation at $d \approx 0$), as well as precision (fast decay of correlation as d increases). Multi-CD performs similarly to ArrowHead and DomainCaller in terms of the enrichment at the boundary, and does better in terms of the precision (Fig. 6b). Specifically, when fitted to exponential decays, the correlation lengths are 34 kb ($\lambda = 0$) and 143 kb ($\lambda = 10$) for Multi-CD, compared to $\gtrsim$ 900 kb for the two previous methods (Fig. 6b).

Next, we compared our compartment solutions (CDs at $\lambda = 90$, shown in Fig. 4b) with the replication timing profiles (Repli-Seq), which are known to correlate differently with the A- and B- compartments (7, 67). Our inferred compartments exhibit the anticipated patterns of replication timing (Fig. 6c); the A-compartment shows an activation of replication signals in the early-phases (G1, S1, S2) and a repression in the later phases (S3, S4, G2), whereas the B-compartment shows an opposite trend. There is a clear anti-correlation between the replication patterns in the two compartments along the replication cycle (Fig. 6c), as quantified in terms of the Pearson correlation (Fig. 6d). Comparison to other epigenetic markers, such as the pattern of histone modifications, further confirms the association of our CD solutions with the A/B-compartments (Fig. S11).

Discussion

Multi-CD has many essential advantages that will make it a useful tool for the study of chromatin organization. As a computational algorithm, Multi-CD includes two core steps: the pre-processing of raw Hi-C data into a correlation matrix, and the inference of chromatin domain (CD) solutions from the correlation matrix. The pre-processing is based on a model of gaussian polymer network, allowing a physically justifiable interpretation of the Hi-C data.

The domain identification problem is formulated by combining the group model with a Bayesian inference for the CD solution. The formulation of Multi-CD that optimizes the recognition of global pattern appeared in Hi-C naturally deals

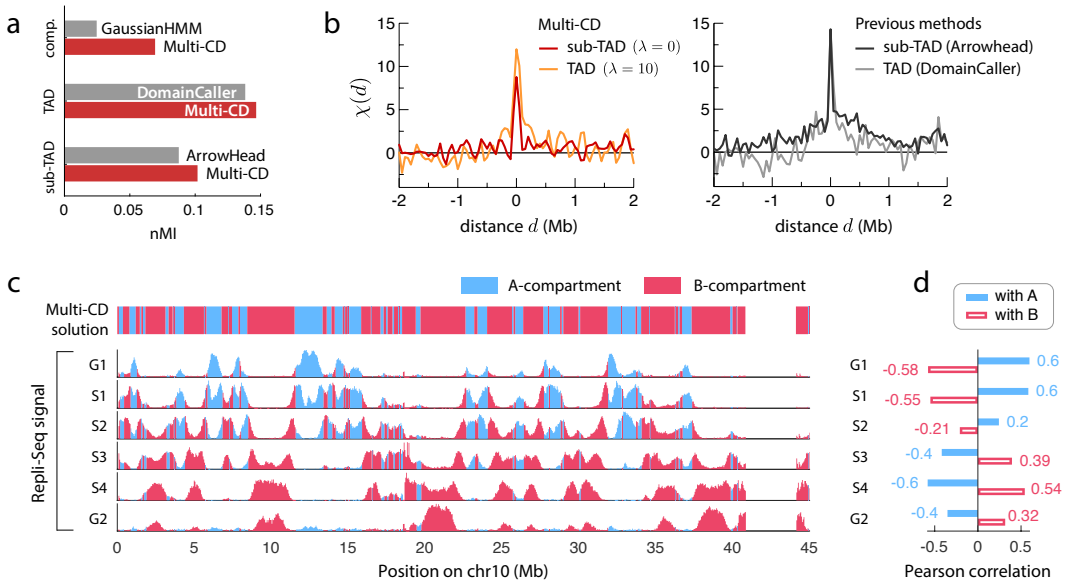

Fig. 6. Validation of CD solutions from Multi-CD. (a) In terms of the normalized mutual information between the CD solutions and the input data, Multi-CD outperforms ArrowHead, DomainCaller and GaussianHMM at the corresponding scales (sub-TAD, TAD and compartment). **(b)** The correlation function $\chi(d)$ between CTCF signals and the domain boundaries. Shown for sub-TADs and TADs, obtained from Multi-CD (left); from ArrowHead and DomainCaller (right). **(c)** Genome-wide, locus-dependent replication signal. Top panel shows the A- (blue) and B- (red) compartments inferred by Multi-CD. Bottom panels show the replication signals in six different phases in the cell cycle, shaded in matching colors for the two compartments. **(d)** Pearson correlation between the replication signals and the two compartments A (filled blue) and B (open red).

with non-local CDs, which differentiates Multi-CD from previous methods that focus on local features in Hi-C, such as CD boundaries or loops enriched with higher contact frequencies. Moreover, Multi-CD can find CDs across a wide range of scales without having to adjust or down-sample the Hi-C data to match the scale of CDs to be identified, which is an important improvement over many existing methods.

An important feature of Multi-CD, as emphasized in the name, is that it provides a unified framework to identify CDs at multiple scales, where the scales of the CDs are tuned by a single parameter λ. The resulting *family* of CD solutions allow quantitative comparisons between CD solutions at different scales. The analysis revealed special scales at which the CD solutions are particularly interesting: sub-TADs ($\lambda = 0$), TADs ($\lambda = 10$, where domain conservation was strongest), and meta-TADs ($\lambda = 30$, where the correlation pattern was best captured). At larger scales, we found that compartments ($\lambda = 90$) emerge as a secondary solution that can be inferred after removing the local signals in the correlation matrix that correspond to the TAD-like solutions. We confirmed that Multi-CD successfully reproduces, or even outperforms, the existing methods to identify CDs at the specific scales. Importantly, Multi-CD achieves this performance through a single unified algorithm, which not only identifies the specific CD solutions accurately, but also allows a comparative analysis of the multi-scale family of solutions.

In particular, we characterized the hierarchical organization of the chromatin by quantifying the similarity and the nestedness between CD solutions at two different scales. We showed that the characteristics of CD solutions shared by the local, TAD-like domains do not precisely hold together in the non-local, compartment-like domains. This finding is consistent with the recent studies which report that compartments and TADs are formed by different mechanisms of motor-driven active loop extrusion and microphase separation, and that they do not necessarily have a hierarchical relationship (65, 68–70). Meanwhile, the sub-TADs are nested in each of the other three solutions, including the compartments (Fig. 5), indicating that sub-TADs are the fundamental building blocks of the higher-order CD organization. In fact, the existence of sub-TADs is robust when a finer-resolution Hi-C is considered. Applying Multi-CD on Hi-C data at 5-kb resolution, we clearly recover the sub-TADs that are consistent with the sub-TADs obtained from the 50-kb Hi-C (see Fig. S8).

While there are methods that report hierarchical CDs (32, 33), Multi-CD makes significant advances both algorithmically and conceptually. Multi-CD can detect non-local domains with better flexibility instead of finding a set of intervals. 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. S12). 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 (Fig. S12).

Multi-CD is a method of great flexibility that can be readily applied to analyze any dataset that exhibits pairwise correlation patterns. However, two cautionary remarks are in place for more careful interpretation of the results. (i) In general, the relevant values of λ depend on the resolution of the input Hi-C dataset, as well as on the cell type. While λ is a useful parameter that allows comparative analysis, its specific value does not carry any biological significance. Although we referred to a specific CD solution by the corresponding value of λ in the current analysis (Fig. 3), the lesson should *not* be that TADs, for instance, always correspond to the particular value of λ; instead, TADs should be identified as the most conserved CD solutions across cell types after scanning a range of λ's.

(ii) Multi-CD is agnostic about whether the collected data is homogeneous or heterogeneous. Application of Multi-CD to single-cell Hi-C data, and the subsequent interpretation of the result, would be straightforward; however, if the input Hi-C data were an outcome of a mixture of heterogeneous subpopulations, the solution from Multi-CD would correspond to their superposition. This is a fundamental issue inherent to any Hi-C data analysis method. Despite the presence of cell-to-cell variations, the population-averaged pattern manifest in Hi-C carries a rich set of information that is specific to the cell type. The need for interpretable inference methods that can extract valuable insights into the spatial organization of the genome, including ours, is still high.

To recapitulate, in order to glean genome function from Hi-C data that varies with the genomic state (10–13), a computationally accurate method to identify CD structures is of vital importance. Multi-CD is a physically principled method that identifies multi-scale structures of chromatin domains by solving the global optimization problem. We find the chromatin domains identified from Multi-CD in excellent match with biological data such as CTCF binding sites and replication timing signal. Quantitative analyses of CD structures identified across multiple genomic scales and various cell types offer general physical insight into chromatin organization inside cell nuclei.

Materials and Methods

Interpretation of Hi-C data.

Normalization and contact probability. Here we describe how the Hi-C data can be interpreted as a set of contact probabilities for pairs of genomic segments, p_{ij}. Typically, a Hi-C matrix have widely varying row-sums; for example, the net count of the i-th segment in the experiment is much larger than the net count of the j-th segment. To marginalize out this site-wise variation and only focus on the differential strengths of pairwise interactions, the raw Hi-C matrix $\mathbf{M}_{\rm raw}$ is normalized to have uniform row and column sums. This is achieved using the Knight-Ruiz (KR) algorithm (71), which finds a vector $\mathbf{v} = (v_1, \cdots, v_N)$ for calculating $(\mathbf{M})_{ij} = v_i v_j (\mathbf{M}_{\rm raw})_{ij}$, such that each row (column) in $\mathbf{M}$ sums to 1.

We assume that the normalized Hi-C signal is proportional to the contact probability: $(\mathbf{M})_{ij} \propto p_{ij}$. Note that p_{ij} is the probability that the two segments i and j are within a contact distance, and the rows of the contact probability matrix $(\mathbf{P})_{ij} = p_{ij}$ is not required to sum to 1. Because the proportionality constant is unknown *a priori*, however, we have a free parameter to choose. We do this by fixing the average nearest-neighbor contact probability, $\bar{p}_1 = \langle p_{i,i+1} \rangle$. We expect the $\bar{p}_1$ to be relatively close to 1, assuming that nearest-neighbor contact is likely; but not exactly 1, because there are variations among the nearest-neighbor Hi-C signal. In this work we chose $\bar{p}_1 = 0.9$. The resulting contact probability matrix $\mathbf{P}$ is given as $(\mathbf{P})_{ij} = \min(1, \hat{p}_{ij})$, with $\hat{p}_{ij} = (\bar{p}_1/\mu)(\mathbf{M})_{ij}$, where $\mu = \langle \mathbf{M}_{i,i+1} \rangle$ is the Hi-C signals averaged over the nearest-neighbors. At $\bar{p}_1 = 0.9$, in our case, the fraction of over-saturated elements ($\hat{p}_{ij} > 1$) was sufficiently small.

Building correlation matrix from Hi-C. A chromosome can be regarded as a polymer chain containing N monomers, each of which corresponds to the i-th genomic segment and its spatial position is written as $\mathbf{r}_i$. Adapting the random loop model (RLM) (42), we interpret chromosome conformation as described by an ideal polymer network, with cross-links between spatially close segments. In RLM, we describe the spatial positions of the polymer segments using a gaussian distribution with zero mean and and a covariance matrix Σ, with elements $(\Sigma)_{ij} = \sigma_{ij} = \langle \delta \mathbf{r}_i \cdot \delta \mathbf{r}_j \rangle$. It follows that the distance $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ between two monomers i and j can be described in the form of a weighted gaussian function (Eq. 1) where the variance $(2\gamma_{ij})^{-1}$ is associated with the covariance matrix elements as $\gamma_{ij}^{-1} = 2(\sigma_{ii} + \sigma_{jj} - 2\sigma_{ij})$. The contact probabilities can be calculated from the distribution of pairwise distances (Eq. 1), by saying that two segments i and j are in contact when their distance r_{ij} is below a cutoff, r_c. In other words, we write

$$p_{ij} = \int_0^{r_c} P(r_{ij}; \gamma_{ij}) dr_{ij} = \mathrm{erf}(\gamma_{ij}^{1/2} r_c) - 2r_c \sqrt{\frac{\gamma_{ij}}{\pi}} e^{-\gamma_{ij} r_c^2}, \quad [10]$$

where $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x dt e^{-t^2}$. The value of γ_{ij} is uniquely determined for each p_{ij}; once we have the γ_{ij}'s, we can reconstruct the covariance matrix $\{\sigma_{ij}\}$. Because the diagonals are underdetermined in Hi-C (self-contacts are not reported), we assume a uniform variance $\sigma_{ii} = \sigma_{jj} = \sigma_c$ along the diagonal. Note that although the value of γ_{ij} depends on the choice of r_c, its effect is only to scale the γ_{ij}'s as $\gamma_{ij} \to r_c^2 \gamma_{ij}$, and consequently the σ_{ij}'s.

Finally, we normalize the covariance matrix to build the correlation matrix $\mathbf{C}$:

$$(\mathbf{C})_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}} = \frac{\sigma_{ij}}{\sigma_c} = 1 - \frac{1}{4\sigma_c \gamma_{ij}}. \quad [11]$$

It appears that σ_c sets the overall intensity of $\mathbf{C}$. Here, we chose the value of σ_c as the median of $1/4\gamma_{ij}$, i.e., $\sigma_c = \mathrm{median}(1/4\gamma_{ij})$. This cancels out the scaling effect of r_c in σ_{ij}, so that the choice of r_c does not affect the ultimate construction of the correlation matrix $\mathbf{C}$.

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 (18). 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 out by increasing the window size (see refs (18, 19) for further details). In this study, we used the expected number obtained from (19). The Pearson correlation matrix of the O/E ($\mathbf{C}_{\rm O/E}$) represents the overall contact pattern through pairwise correlation coefficients between segments.

Determination of optimal CD solutions.

Metropolis-Hastings sampling. Markov chain Monte Carlo (MCMC) sampling was employed to find the minimum value of the total cost function $\mathcal{H}$. 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 $\alpha(\mathbf{s}, \mathbf{s}') = \exp\left[-(\mathcal{H}(\mathbf{s}'|\mathbf{C}) - \mathcal{H}(\mathbf{s}|\mathbf{C}))/T\right]$.

In sampling the space of CD solutions, a move from a state $\mathbf{s}$ to another state $\mathbf{s}'$ is defined such that the two CD solutions $(\mathbf{s}, \mathbf{s}')$ differ only by one genomic segment. More precisely, because a CD solution is invariant upon permutations of the domain indices, the distance between $\mathbf{s}$ and $\mathbf{s}'$ is uniquely defined as the *minimal* number of mismatches over all possible domain index permutations.

To ensure that the sampling is properly conducted, we continue the sampling until each chain collects $t_{\rm tot} \geq 5\tau^*$ 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: $\tau^* = \mathrm{argmin}_\tau |R(\tau) - 1/e|$. The autocorrelation function is calculated from the value of the total cost function $\mathcal{H}$, as

$$R(\tau) = \frac{1}{\sigma^2} \langle (\mathcal{H}(\mathbf{s}_t|\mathbf{C}) - \mu)(\mathcal{H}(\mathbf{s}_{t+\tau}|\mathbf{C}) - \mu) \rangle_t, \quad [12]$$

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. In Eq.12, the running time average is taken over all the pairs of samples with the time gap of τ.

We note that sampling is the computational bottleneck for our method; our stop condition (at $5\tau^*$) was chosen conservatively for accurate solutions. In practice, the sampling time can be reduced at the cost of an increased batch size (number of different initial configurations, as described below in simulated annealing), which is significantly cheaper if parallel computing is used.

Simulated annealing. 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, is drawn randomly from the set of integers $\{1, \cdots, N\}$. Then, each genomic segment $i \in \{1, \cdots, N\}$ is allocated randomly into one of the CDs, $k \in \{1, 2, \cdots, K\}$. The initial temperature T_0 is determined such that the acceptance probability for the "worst" move around $\mathbf{s}^{(0)}$ is 0.5.

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 is repeated until the temperature reaches T_f. We used $T_f = 0.03$. Then we quench the system to the closest local minimum by performing gradient descent. 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. S13). This saves the algorithm from the large memory requirement of dealing with the entire intra-chromosomal 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.

Analysis and evaluation of domain solutions.

Similarity of two distinct CD solutions using Pearson correlation. To measure the extent of similarity between two CD solutions $\mathbf{s}$ and $\mathbf{s}'$, we evaluate the Pearson correlation. The binary matrices $\mathbf{B}$ and $\mathbf{B}'$ that represent the two CD solutions, are defined such that the matrix element are all 1's within the same CD and 0 otherwise. i.e., $(\mathbf{B})_{ij} = B_{ij} = \delta_{s_i s_j}$. The similarity between $\mathbf{B}$ and $\mathbf{B}'$ is quantified using the Pearson correlation

$$\rho = \langle \delta B \delta B' \rangle \Big/ \sqrt{\langle (\delta B)^2 \rangle \, \langle (\delta B')^2 \rangle}, \qquad [13]$$

where $\langle \delta B \delta B' \rangle = \langle (B_{ij} - \bar{B})(B'_{ij} - \bar{B}') \rangle_{i \neq j}$, and $\langle (\delta B)^2 \rangle = \langle (B_{ij} - \bar{B})^2 \rangle_{i \neq j}$. The average $\langle \cdot \rangle_{i \neq j}$ runs over all distinct pairs.

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 $\mathbf{C}_{\text{O/E}}$ was used for $\mathbf{A}$. Treating the matrix elements $a \in \mathbf{A}$ and $b \in \mathbf{B}$ as two random variables, we construct the joint distribution

$$p(A, B) = \langle \delta_{A_{ij}, a} \delta_{B_{ij}, b} \rangle_{i \neq j} \qquad [14]$$

where $\langle \cdot \rangle_{i \neq j}$ is an average over all distinct pairs. The Kronecker delta for the continuous variable a is defined in a discretized fashion: that is, $\delta_{A_{ij}, a} = 1$ if $A_{ij} \in [a, a + \Delta a)$ and 0 otherwise, where $\Delta a (= [\max\{A_{ij}\} - \min\{A_{ij}\}]/100)$ is used for discretization into 100 bins. Then we can calculate the mutual information,

$$I(A; B) = \sum_{a \in \mathbf{A}} \sum_{b \in \mathbf{B}} p(a, b) \log \left[p(a, b) / p(a) p(b) \right], \qquad [15]$$

and the normalized mutual information (nMI),

$$\text{nMI}(A; B) = I(A; B) \Big/ \sqrt{H(A) \cdot H(B)}, \qquad [16]$$

where $H(X) = -\sum_{x \in \mathbf{X}} p(x) \log p(x)$ is the marginal entropy.

Nestedness of CD solutions. Here we define a measure to quantify the nestedness between two CD solutions, $\mathbf{s}$ (assumed to have smaller domains on average) and $\mathbf{s}'$ (larger domains). The idea is the following: $\mathbf{s}$ is perfectly nested in $\mathbf{s}'$ if, whenever two sites belong to a same domain in $\mathbf{s}$, they also belong to a same domain in $\mathbf{s}'$. For each domain $k \in \mathbf{s}$, we consider the best overlap of this domain k on the other solution $\mathbf{s}'$:

$$h_1(k \to \mathbf{s}') = \max_{k' \in \mathbf{s}'} v_{k,k'} / n_k \qquad [17]$$

where $v_{k,k'}$ is the number of overlapping sites between two domains $k \in \mathbf{s}$ and $k' \in \mathbf{s}'$, and n_k is the size of domain k. The highest score $h_1(k \to \mathbf{s}') = 1$ is obtained when domain k is fully included in one of the domains in $\mathbf{s}'$. The null hypothesis corresponds to where the domains in $\mathbf{s}$ and $\mathbf{s}'$ are completely uncorrelated, in which case h_1 only reflects the overlap "by chance". The chance level $\bar{h}_1(k \to \mathbf{s}')$ is calculated by making n_k random draws from $\mathbf{s}'$; we averaged over 100 independent trials. We normalize the score as

$$\hat{h}_1(k \to \mathbf{s}') = \frac{h_1 - \bar{h}_1}{1 - \bar{h}_1}, \qquad [18]$$

such that $\hat{h}_1(k \to \mathbf{s}') = 0$ indicates the chance level, and $\hat{h}_1(k \to \mathbf{s}') = 1$ means a perfect nestedness. Finally, we define the nestedness score $h(\mathbf{s} \to \mathbf{s}')$ for the entire CD solution as a weighted average:

$$h(\mathbf{s} \to \mathbf{s}') = \sum_{k \in \mathbf{s}} \hat{h}_1(k \to \mathbf{s}') \cdot n_k / N. \qquad [19]$$

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. Suppose that the CTCF signal at genomic segment i is given as $\phi_{\text{CTCF}}(i)$. Then, we can consider an overlap function between $\phi_{\text{CTCF}}(i)$ and a CD-boundary indicating function $\psi_{\text{DB}}(i)$, where $\psi_{\text{DB}}(i) = 1$ if the i-th segment is precisely at the domain boundary; $\psi_{\text{DB}}(i) = 0$, otherwise. We evaluated a distance-dependent, normalized overlap function $\chi(d)$, defined as

$$\chi(d) = \frac{\langle \delta\phi_{\text{CTCF}}(i + d) \psi_{\text{DB}}(i) \rangle_i}{\langle \psi_{\text{DB}} \rangle}, \qquad [20]$$

where $\delta\phi_{\text{CTCF}} = \phi_{\text{CTCF}} - \langle \phi_{\text{CTCF}} \rangle$. If the domain boundaries determined from Multi-CD is well 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 $\mathbf{q}^{(B)}$, where $q_i^{(A)} = +1$ if the i-th segment belongs to compartment A, and $q^{(A)} = -1$ otherwise. For a set of epigenetic marks measured across the genome is represented with $\mathbf{h}$, where its component h_i denotes the value at the i-th genomic segment, the correlation between the solutions of compartment A and B and the epigenetic marks can be evaluated using the Pearson correlations as:

$$c_A = (\mathbf{q}^{(A)} \cdot \mathbf{h}) / |\mathbf{q}^{(A)}||\mathbf{h}|, \quad c_B = (\mathbf{q}^{(B)} \cdot \mathbf{h}) / |\mathbf{q}^{(B)}||\mathbf{h}|. \qquad [21]$$

Data availability. All data used in the paper were obtained from publicly available repositories. See SI Appendix.

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

ACKNOWLEDGMENTS. We thank Roger Oria Fernandez for critical comment on our script deposited in GitHub. This work was supported in part by a KIAS Individual Grant at Korea Institute for Advanced Study (No. CG035003 to C.H.). We thank the Center for Advanced Computation in KIAS for providing computing resources.

1. J. O. Davies, A. M. Oudelaar, D. R. Higgs & J. R. Hughes, How best to identify chromosomal interactions: a comparison of approaches. *Nat. Methods*, **14**, 125, (2017).
2. J. Dekker, K. Rippe, M. Dekker & N. Kleckner, Capturing chromosome conformation. *Science*, **295**, 1306–1311, (2002).
3. T. Misteli, Beyond the sequence: cellular organization of genome function. *Cell*, **128**, 787–800, (2007).
4. W. A. Bickmore & B. van Steensel, Genome architecture: domain organization of interphase chromosomes. *Cell*, **152**, 1270–1284, (2013).
5. C. Lanctôt, T. Cheutin, M. Cremer, G. Cavalli & T. Cremer, Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. *Nat. Rev. Genetics*, **8**, 104, (2007).
6. J. Fraser, C. Ferrai, A. M. Chiariello, M. Schueler, T. Rito, G. Laudanno, M. Barbieri, B. L. Moore, D. C. Kraemer, S. Aitken et al., Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Mol. Sys. Biol.*, **11**, 852, (2015).
7. T. Ryba, I. Hiratani, J. Lu, M. Itoh, M. Kulik, J. Zhang, T. C. Schulz, A. J. Robins, S. Dalton & D. M. Gilbert, Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res.*, **20**, 761–770, (2010).

8. I. Hiratani, T. Ryba, M. Itoh, T. Yokochi, M. Schwaiger, C.-W. Chang, Y. Lyou, T. M. Townes, D. Schübeler & D. M. Gilbert, Global reorganization of replication domains during embryonic stem cell differentiation. *PLoS Biol.*, **6**, e245, (2008).
9. G. Cavalli & T. Misteli, Functional implications of genome topology. *Nat. Struct. Mol. Biol.*, **20**, 290–299, (2013).
10. J. R. Dixon, I. Jung, S. Selvaraj, Y. Shen, J. E. Antosiewicz-Bourget, A. Y. Lee, Z. Ye, A. Kim, N. Rajagopal, W. Xie, Y. Diao, J. Liang, H. Zhao, V. V. Lobanenkov, J. R. Ecker, J. A. Thomson & B. Ren, Chromatin architecture reorganization during stem cell differentiation. *Nature*, **518**, 331, (2015).
11. P. H. L. Krijger & W. De Laat, Regulation of disease-associated gene expression in the 3D genome. *Nat. Rev. Mol. Cell Biol.*, **17**, 771, (2016).
12. J. R. Dixon, J. Xu, V. Dileep, Y. Zhan, F. Song, V. T. Le, G. G. Yardımcı, A. Chakraborty, D. V. Bann, Y. Wang, R. Clark, L. Zhang, H. Yang, T. Liu, S. Iyyanki, L. An, C. Pool, T. Sasaki, J. C. Rivera-Mulia, H. Ozadam, B. R. Lajoie, R. Kaul, M. Buckley, K. Lee, M. Diegel, D. Pezic, C. Ernst, S. Hadjur, D. T. Odom, J. A. Stamatoyannopoulos, J. R. Broach, R. C. Hardison, F. Ay, W. S. Noble, J. Dekker, D. M. Gilbert & F. Yue, Integrative detection and analysis of structural variation in cancer genomes. *Nat. Genetics*, **50**, 1388, (2018).
13. Y. Shao, N. Lu, Z. Wu, C. Cai, S. Wang, L.-L. Zhang, F. Zhou, S. Xiao, L. Liu, X. Zeng, H. Zheng, C. Yang, Z. Zhao, G. Zhao, J.-Q. Zhou, X. Xue & Z. Qin, Creating a functional single-chromosome yeast. *Nature*, **560**, 331, (2018).
14. L. Liu, M. H. Kim & C. Hyeon, Heterogeneous Loop Model to Infer 3D Chromosome Structures from Hi-C. *Biophys. J.*, **117**, 613–625, (2019).
15. M. Franke, D. M. Ibrahim, G. Andrey, W. Schwarzer, V. Heinrich, R. Schöpflin, K. Kraft, R. Kempfer, I. Jerković, W. L. Chan, M. Spielmann, B. Timmermann, L. Wittler, I. Kurth, P. Cambiaso, O. Zuffardi, G. Houge, L. Lambie, F. Brancati, A. Pombo, M. Vingron, F. Spitz & S. Mundlos, Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature*, **538**, 265–269, (2016).
16. S. C. Baca, D. Prandi, M. S. Lawrence, J. M. Mosquera, A. Romanel, Y. Drier, K. Park, N. Kitabayashi, T. Y. MacDonald, M. Ghandi et al., Punctuated evolution of prostate cancer genomes. *Cell*, **153**, 666–677, (2013).
17. T. Cremer & M. Cremer, Chromosome territories. *Cold Spring Harbor Perspectives in Biology*, **2**, (2010).
18. E. Lieberman-Aiden, N. L. Van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner et al., Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293, (2009).
19. S. S. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander & E. L. Aiden, A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, **159**, 1665 – 1680, (2014).
20. T. Sexton & G. Cavalli, The role of chromosome domains in shaping the functional genome. *Cell*, **160**, 1049 – 1059, (2015).
21. S. Wang, J.-H. Su, B. J. Beliveau, B. Bintu, J. R. Moffitt, C.-t. Wu & X. Zhuang, Spatial organization of chromatin domains and compartments in single chromosomes. p., aaf8084, (2016).
22. J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu & B. Ren, Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380, (2012).
23. E. P. Nora, B. R. Lajoie, E. G. Schulz, L. Giorgetti, I. Okamoto, N. Servant, T. Piolot, N. L. van Berkum, J. Meisig & J. Sedat, Spatial partitioning of the regulatory landscape of the X-inactivation center. *Nature*, **485**, 381, (2012).
24. T. Sexton, E. Yaffe, E. Kenigsberg, F. Bantignies, B. Leblanc, M. Hoichman, H. Parrinello, A. Tanay & G. Cavalli, Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome. *Cell*, **148**, 458 – 472, (2012).
25. B. D. Pope, T. Ryba, V. Dileep, F. Yue, W. Wu, O. Denas, D. L. Vera, Y. Wang, R. S. Hansen, T. K. Canfield et al., Topologically associating domains are stable units of replication-timing regulation. *Nature*, **515**, 402, (2014).
26. J. Dekker & E. Heard, Structural and functional diversity of topologically associating domains. *FEBS Letters*, **589**, 2877–2884, (2015).
27. J. R. Dixon, D. U. Gorkin & B. Ren, Chromatin domains: the unit of chromosome organization. *Mol. Cell*, **62**, 668–680, (2016).
28. J. E. Phillips-Cremins, M. E. Sauria, A. Sanyal, T. I. Gerasimova, B. R. Lajoie, J. S. Bell, C.-T. Ong, T. A. Hookway, C. Guo, Y. Sun, M. J. Bland, W. Wagstaff, S. Dalton, T. C. McDevitt, R. Sen, J. Dekker, J. Taylor & V. Corces, Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*, **153**, 1281–1295, (2013).
29. F. Jin, Y. Li, J. R. Dixon, S. Selvaraj, Z. Ye, A. Y. Lee, C.-A. Yen, A. D. Schmitt, C. Espinoza & B. Ren, A high-resolution map of three-dimensional chromatin interactome in human cells. *Nature*, **503**, 290, (2013).
30. P. P. Rocha, R. Raviram, R. Bonneau & J. A. Skok, Breaking TADs: insights into hierarchical genome organization. *Epigenomics*, **7**, 523–526, (2015).
31. Q. Wang, Q. Sun, D. M. Czajkowsky & Z. Shao, Sub-kb Hi-C in D. melanogaster reveals conserved characteristics of TADs between insect and mammalian cells. *Nature communications*, **9**, 188, (2018).
32. C. Weinreb & B. J. Raphael, Identification of hierarchical chromatin domains. *Bioinformatics*, **32**, 1601–1609, (2016).
33. D. Filippova, R. Patro, G. Duggal & C. Kingsford, Identification of alternative topological domains in chromatin. *Algorithms for Molecular Biology*, **9**, 14, (2014).
34. H. Shin, Y. Shi, C. Dai, H. Tjong, K. Gong, F. Alber & X. J. Zhou, Topdom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res.*, **44**, e70–e70, (2015).
35. J. Mateos-Langerak, M. Bohn, W. de Leeuw, O. Giromus, E. M. Manders, P. J. Verschure, M. H. Indemans, H. J. Gierman, D. W. Heermann, R. Van Driel & S. Goetze, Spatially confined folding of chromatin in the interphase nucleus. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 3812–3817, (2009).
36. M. Barbieri, M. Chotalia, J. Fraser, L.-M. Lavitas, J. Dostie, A. Pombo & M. Nicodemi, 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. J. D. Halverson, J. Smrek, K. Kremer & A. Y. Grosberg, From a melt of rings to chromosome territories: the role of topological constraints in genome folding. *Rep. Prog. Phys.*, **77**, 022601, (2014).
38. C. A. Brackley, J. M. Brown, D. Waithe, C. Babbs, J. Davies, J. R. Hughes, V. J. Buckle & D. Marenduzzo, Predicting the three-dimensional folding of cis-regulatory regions in mammalian genomes using bioinformatic data and polymer models. *Genome Biol.*, **17**, 59, (2016).
39. G. Shi, L. Liu, C. Hyeon & D. Thirumalai, Interphase Human Chromosome Exhibits Out of Equilibrium Glassy Dynamics. *Nat. Commun.*, **9**, 3161, (2018).
40. L. Liu, G. Shi, D. Thirumalai & C. Hyeon, Chain organization of human interphase chromosome determines the spatiotemporal dynamics of chromatin loci. *PLoS Comp. Biol.*, **14**, e1006617, (2018).
41. J. D. Bryngelson & D. Thirumalai, Internal constraints induce localization in an isolated polymer molecule. *Phys. Rev. Lett.*, **76**(3), 542–545, (1996).
42. M. Bohn, D. W. Heermann & R. van Driel, Random loop model for long polymers. *Phys. Rev. E.*, **76**, 051805, (2007).
43. R. Bruinsma, A. Y. Grosberg, Y. Rabin & A. Zidovska, Chromatin hydrodynamics. *Biophys. J.*, **106**, 1871–1881, (2014).
44. C. Battle, C. P. Broedersz, N. Fakhri, V. F. Geyer, J. Howard, C. F. Schmidt & F. C. MacKintosh, Broken detailed balance at mesoscopic scales in active biological systems. *Science*, **352**, 604–607, (2016).
45. W. Hwang & C. Hyeon, Quantifying the heat dissipation from molecular motor's transport properties in nonequilibrium steady states. *J. Phys. Chem. Lett.*, **8**, 250–256, (2017).
46. W. Hwang & C. Hyeon, Energetic costs, precision, and transport efficiency of molecular motors. *J. Phys. Chem. Lett.*, **9**, 513–520, (2018).
47. X. Fang, K. Kruse, T. Lu & J. Wang, Nonequilibrium physics in biology. *Rev. Mod. Phys.*, **91**, 045004, (2019).
48. A. Rosa & R. Everaers, Structure and dynamics of interphase chromosomes. *PLoS Comput. Biol.*, **4**, e1000153, (2008).
49. H. Kang, Y.-G. Yoon, D. Thirumalai & C. Hyeon, Confinement-induced glassy dynamics in a model for chromosome organization. *Phys. Rev. Lett.*, **115**, 198102, (2015).
50. G. Shi & D. Thirumalai, Conformational heterogeneity in human interphase chromosome organization reconciles the FISH and Hi-C paradox. *Nature Commun.*, **10**, 1–10, (2019).
51. Q. Szabo, D. Jost, J.-M. Chang, D. I. Cattoni, G. L. Papadopoulos, B. Bonev, T. Sexton, J. Gurgo, C. Jacquier, M. Nollmann, F. Bantignies & G. Cavalli, TADs are 3D structural units of higher-order chromosome organization in Drosophila. *Science Advances*, **4**, (2018).
52. E. H. Finn, G. Pegoraro, S. Shachar & T. Misteli, Comparative analysis of 2d and 3d distance measurements to study spatial genome organization. *Methods*, **123**, 47–55, (2017).
53. L. Giorgetti, R. Galupa, E. P. Nora, T. Piolot, F. Lam, J. Dekker, G. Tiana & E. Heard, Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell*, **157**, 950–963, (2014).
54. N. Sauerwald, S. Zhang, C. Kingsford & I. Bahar, Chromosomal dynamics predicted by an elastic network model explains genome-wide accessibility and long-range couplings. *Nucleic acids research*, **45**, 3663–3673, (2017).
55. S. Zhang, F. Chen & I. Bahar, Differences in the intrinsic spatial dynamics of the chromatin contribute to cell differentiation. *Nucleic Acids Research*, **x**, gkz1102, (2019).
56. L. Laloux, P. Cizeau, J.-P. Bouchaud & M. Potters, Noise dressing of financial correlation matrices. *Phys. Rev. Lett.*, **83**, 1467, (1999).
57. J. D. Noh, Model for correlations in stock markets. *Phys. Rev. E.*, **61**, 5981, (2000).
58. L. Giada & M. Marsili, Data clustering and noise undressing of correlation matrices. *Phys. Rev. E.*, **63**, 061101, (2001).
59. S. Kirkpatrick, D. Gelatt & M. P. Vecchi, Optimization by simulated annealing. *Science*, **220**, 671–680, (1983).
60. M. Vietri Rudan, C. Barrington, S. Henderson, C. Ernst, D. Odom, A. Tanay & S. Hadjur, Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Reports*, **10**, 1297 – 1309, (2015).
61. A. Li, X. Yin, B. Xu, D. Wang, J. Han, Y. Wei, Y. Deng, Y. Xiong & Z. Zhang, Decoding topologically associating domains with ultra-low resolution Hi-C data by graph structural entropy. *Nature Commun*, **9**, 3265, (2018).
62. A. Grosberg, S. Nechaev & E. Shakhnovich, The role of topological constraints in the kinetics of collapse of macromolecules. *J. Phys.*, **49**, 2095–2100, (1988).
63. S. Fishilevich, R. Nudel, N. Rappaport, R. Hadar, I. Plaschkes, T. Iny Stein, N. Rosen, A. Kohn, M. Twik, M. Safran, D. Lancet & D. Cohen, GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database*, **2017**, 1–17, (2017).
64. E. P. Nora, A. Goloborodko, A.-L. Valton, J. H. Gibcus, A. Uebersohn, N. Abdennur, J. Dekker, L. A. Mirny & B. G. Bruneau, Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell*, **169**, 930–944, (2017).
65. W. Schwarzer, N. Abdennur, A. Goloborodko, A. Pekowska, G. Fudenberg, Y. Loe-Mie, N. A. Fonseca, W. Huber, C. H. Haering, L. Mirny et al., Two independent modes of chromatin organization revealed by cohesin removal. *Nature*, **551**, 51, (2017).
66. T. Misteli, Beyond the sequence: cellular organization of genome function. *Cell*, **128**, 787–800, (2007).
67. R. S. Hansen, S. Thomas, R. Sandstrom, T. K. Canfield, R. E. Thurman, M. Weaver, M. O. Dorschner, S. M. Gartler & J. A. Stamatoyannopoulos, Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl. Acad. Sci. U. S. A.*, **107**, 139–144, (2010).
68. G. Fudenberg, M. Imakaev, C. Lu, A. Goloborodko, N. Abdennur & L. A. Mirny, Formation of chromosomal domains by loop extrusion. *Cell Reports*, **15**, 2038–2049, (2016).
69. J. Gassler, H. B. Brandão, M. Imakaev, I. M. Flyamer, S. Ladstätter, W. A. Bickmore, J.-M. Peters, L. A. Mirny & K. Tachibana, A mechanism of cohesin-dependent loop extrusion organizes zygotic genome architecture. *EMBO J.*, **36**, 3600–3618, (2017).
70. J. Nuebler, G. Fudenberg, M. Imakaev, N. Abdennur & L. A. Mirny, Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proc. Natl. Acad. Sci. U. S. A.*, **115**, E6697–E6706, (2018).

71. P. A. Knight & D. Ruiz, A fast algorithm for matrix balancing. *IMA J. Numer. Anal.*, **33**, 1029–1047, (2013).
72. E. P. Consortium et al., An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57, (2012).
73. P. M. Goldbart & A. Zippelius, Amorphous solid state of vulcanized macromolecules: A variational approach. *Phys. Rev. Lett.*, **71**(14), 2256–2259, (1993).
74. J. D. Bryngelson & D. Thirumalai, Internal constraints induce localization in an isolated polymer molecule. *Phys. Rev. Lett.*, **76**(3), 542–545, (1996).
75. T. Haliloglu, I. Bahar & B. Erman, Gaussian dynamics of folded proteins. *Phys. Rev. Lett.*, **79**(16), 3090–3093, (1997).
76. C. Hyeon, G. Morrison & D. Thirumalai, Force-dependent hopping rates of RNA hairpins can be estimated from accurate measurement of the folding landscapes. *Proceedings of the National Academy of Sciences*, **105**, 9604–9609, (2008).

Supporting Information Appendix

A. Data acquisition.

Hi-C data. Hi-C data were obtained through GEO data repository (GSE63525-*celltype*-primary) (19), where *celltype* is replaced by one of the five different cell types that were considered in our analysis (GM12878, HUVEC, NHEK, K562, and KBM7).

Biological markers. The domain solutions from Multi-CD were compared with known biological markers. We obtained these data mostly from the ENCODE project (72). 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 (63) database http://hgdownload.cse.ucsc.edu/gbdb/hg19/geneHancer/geneHancerInteractionsAll.hg19.bb, accessible through the UCSC Genome Browser Module.

B. Justifications for the Gaussian polymer network for modeling chromosomes.

Here we provide additional justifications for the use of harmonic potentials in the effective Hamiltonian, and consequently, a gaussian distribution for pairwise distances (Eq. 1).

The concept of an effective Hamiltonian consisting of harmonic potential terms is not new; it has been widely employed to study a variety of systems, including the phase transition of vulcanized macromolecules with increasing numbers of crosslinks (73, 74), and the fluctuation dynamics of native proteins (gaussian network model, (75)). Furthermore, a slightly modified, but essentially identical, form of Hamiltonian was used to study the dynamics of folding/unfolding transitions of a single RNA molecule under external force (generalized Rouse model, (76)).

Whereas the success of the gaussian polymer network model does not necessarily guarantee its extension to the modeling of chromosomes, our use of a gaussian distribution for the pairwise distance between two segments in the polymer is empirically justified. The Gaussian-like pairwise distance distributions reported by fluorescence measurements of the chromosome (Fig. S1), and the agreement of the 3D structural properties inferred by a modeling approach that shares the same philosophy (*heterogeneous loop model* or HLM, (14)), suggest that a harmonic spring network provides a reasonable approximation of the energy landscape for the mixture of those subpopulations.

As a side note, it is worth highlighting the versatility of the Gaussian polymer network model in representing the complex topology of chromosome conformation. For the conventional Rouse chain whose monomers along the backbone are constrained by an energy hamiltonian $H = (k/2)\sum_i^{N-1}(r_{i+1} - r_i)^2$ with a uniform spring constant k, it is straightforward to show that $\langle r_{ij}^2 \rangle \sim |i-j|$. Furthermore, if two monomers are in close proximity to form a contact ($r_{ij} < r_c$), then one can obtain the contact probability between monomers i and j in the chain backbone as $p_{ij} = \int_0^{r_c} dr_{ij} P(r_{ij}) \sim |i-j|^{-3/2}$. However, adding just a few non-nearest-neighbor interactions to the Rouse model makes the results highly nontrivial. To illustrate this, we explicitly compared the contact probability map of a linear Gaussian chain (Rouse chain), and those of Gaussian polymer network models with varying numbers of non-nearest-neighbor interactions, which were calculated from the HLM-generated structural ensemble (14) (see Fig. S14). The statistical behavior of Gaussian polymer network model differs from that of the linear "Gaussian" chain. The mean square distance $\langle r_{ij}^2 \rangle$ no longer scales linearly with the separation $s \equiv |i-j|$ (see Fig. S14e), and the contact probability p_{ij} (or $p(s)$) is no longer described with a simple scaling relation (Fig. S14f). The simple modification to the Rouse model, resulting in the Gaussian polymer network model, allows one to explore many different issues of chromosomes. In fact, our recent work based on HLM (14) demonstrated several case studies, substantiating the various experimental measurements on chromosome conformation by solving the inverse-problem of inferring chromosome structures from Hi-C data.

C. Derivation of the likelihood function.

Here we show how to derive the likelihood function, Eq. 6.

Problem. We want to compute

$$p(\mathbf{x}|\mathbf{s},\mathbf{g}) = \left\langle \delta^N \left(\mathbf{x} - \mathbf{f}(\eta,\epsilon)\right) \right\rangle_{\eta,\epsilon} \qquad \text{[S1]}$$

with the following assumptions:

- $\mathbf{x} \in \mathbb{R}^N$ is a sequence of normalized and uncorrelated observations, with zero mean $\langle \mathbf{x} \rangle = \mathbf{0}_N$ and unit covariance $\mathrm{Cov}(\mathbf{x}) = I_N$.
- $\mathbf{s} = (s_1, \cdots, s_N)$ is a clustering map that assigns each site $i \in \{1, \cdots, N\}$ to a cluster index $s_i \in \{1, \cdots, K\}$. Without loss of generality, we can assume that $s_i \le s_j$ whenever $i < j$ (ordered indexing).
- $\eta \sim \mathcal{N}(\mathbf{0}_N, \Lambda)$ and $\epsilon \sim \mathcal{N}(\mathbf{0}_N, \Sigma)$ are i.i.d. gaussian random variables, where Λ and Σ are $N \times N$ covariance matrices. The cluster-dependent covariance is a block diagonal matrix $\Lambda = [\Lambda_s] = [\mathbf{1}_{n_s}\mathbf{1}_{n_s}^\top]$, defined element-wise as $(\Lambda)_{ij} = \delta_{s_i,s_j}$. The site-wise variation is assumed to be uncorrelated, with a unit covariance matrix $\Sigma = I_N$, or $(\Sigma)_{ij} = \delta_{ij}$.
- The clustering strength $\mathbf{g} = (g_1, \cdots, g_K)$ parameterizes the target function $\mathbf{f}$, defined element-wise as

$$f_i(\eta,\epsilon) = \frac{\sqrt{g_{s_i}}\eta_i + \epsilon_i}{\sqrt{1+g_{s_i}}}, \quad i = 1, \cdots, N. \qquad \text{[S2]}$$

Lemma: gaussian integral.

$$\int_{\mathbb{R}^N} d\mathbf{z}\, \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, M)\, e^{i\mathbf{a}^\top \mathbf{z}} = \exp\left(-\frac{1}{2}\mathbf{a}^\top M \mathbf{a}\right), \quad \mathbf{a} \in \mathbb{R}^N. \qquad \text{[S3]}$$

Lemma: Sherman-Morrison formula.

$$(A + \mathbf{u}\mathbf{v}^\top)^{-1} = A^{-1} - \frac{A^{-1}\mathbf{u}\mathbf{v}^\top A^{-1}}{1 + \mathbf{v}^\top A^{-1}\mathbf{u}}. \qquad \text{[S4]}$$

Solution. Let us abbreviate the coefficients as $\alpha_s \equiv \sqrt{g_s/(1+g_s)}$ and $\beta_s \equiv 1/\sqrt{1+g_s}$, such that $f_i = \alpha_{s_i}\eta_i + \beta_{s_i}\epsilon_i$. Further define $A \equiv \mathrm{diag}(\alpha_{s_i})$ and $B \equiv \mathrm{diag}(\beta_{s_i})$, to write $\mathbf{f} = A\boldsymbol{\eta} + B\boldsymbol{\epsilon}$. Taking

the inverse Fourier transform of the Dirac delta function, we can write

$$\delta^N(\mathbf{x}-\mathbf{f}) = \int \frac{d\mathbf{k}}{(2\pi)^N} e^{i(\mathbf{x}-\mathbf{f})^\top \mathbf{k}} = \int \frac{d\mathbf{k}}{(2\pi)^N} e^{i(\mathbf{x}-A\boldsymbol{\eta}-B\boldsymbol{\epsilon})^\top \mathbf{k}},$$

where $\int = \int_{\mathbb{R}^N}$ unless otherwise specified. Now we can rewrite Eq. S1, and evaluate the gaussian integrals using the lemma (Eq. S3):

$$\begin{aligned} & p(\mathbf{x}|\mathbf{s},\mathbf{g}) \\ &= \int \frac{d\mathbf{k}}{(2\pi)^N}\ e^{i\mathbf{x}^\top\mathbf{k}} \int d\boldsymbol{\eta}\ \mathcal{N}(\boldsymbol{\eta})\ e^{-iA\boldsymbol{\eta}^\top\mathbf{k}} \int d\boldsymbol{\epsilon}\ \mathcal{N}(\boldsymbol{\epsilon})\ e^{-iB\boldsymbol{\epsilon}^\top\mathbf{k}} \\ &= \int \frac{d\mathbf{k}}{(2\pi)^N}\ \exp\left(i\mathbf{x}^\top\mathbf{k} - \frac{1}{2}(A\mathbf{k})^\top\Lambda(A\mathbf{k}) - \frac{1}{2}(B\mathbf{k})^\top\Sigma(B\mathbf{k})\right) \\ &= \int \frac{d\mathbf{k}}{(2\pi)^N}\ \exp\left(i\mathbf{x}^\top\mathbf{k} - \frac{1}{2}\mathbf{k}^\top Q\mathbf{k}\right), \end{aligned} \tag{S5}$$

where $Q \equiv (A\Lambda A + B\Sigma B)$. Recognizing that this is another (unnormalized) gaussian integral with covariance matrix Q^{-1}, use the lemma (Eq. S3) once again:

$$\begin{aligned} p(\mathbf{x}|\mathbf{s},\mathbf{g}) &= \sqrt{(2\pi)^N \det Q^{-1}} \cdot \frac{1}{(2\pi)^N} \int d\mathbf{k}\ \mathcal{N}(\mathbf{k}|\mathbf{0},Q^{-1})\ e^{i\mathbf{x}^\top\mathbf{k}} \\ &= \exp\left(-\frac{1}{2}\mathbf{x}^\top Q^{-1}\mathbf{x} - \log\det Q\right). \end{aligned} \tag{S6}$$

With uncorrelated $\boldsymbol{\epsilon}$, both Q and Q^{-1} are block diagonal matrices, the exponent is completely separable by clusters:

$$\log p(\mathbf{x}|\mathbf{s},\mathbf{g}) = \frac{1}{2}\sum_{s=1}^{K}\left(\mathbf{x}_s^\top Q_s^{-1}\mathbf{x}_s + \log\det Q_s\right), \tag{S7}$$

where $\mathbf{x}_s$ is the corresponding n_s-dimensional subset of $\mathbf{x}$, and $Q_s = A_s\Lambda_s A_s + B_s\Sigma_s B_s = \alpha_s^2\Lambda_s + \beta_s^2\Sigma_s$, is the $n_s \times n_s$ block matrix corresponding to cluster index s; element-wise, $(Q_s)_{ij} = \alpha_s^2 + \beta_s^2\delta_{ij}$.

We now simplify the two terms in the summand of Eq. S7, and show that the resulting expression corresponds to Eq. 6. First, the quadratic term can be expanded by using the Sherman-Morrison lemma Eq. S4:

$$Q_s^{-1} = (\beta_s^2 I_{n_s} + (\alpha_s\mathbf{1}_{n_s})(\alpha_s\mathbf{1}_{n_s})^\top)^{-1} \tag{S8}$$

$$= \frac{1}{\beta_s^2}\left(I - \frac{(\alpha_s^2/\beta_s^2)\mathbf{1}\mathbf{1}^\top}{1+(\alpha_s^2/\beta_s^2)\mathbf{1}^\top\mathbf{1}}\right). \tag{S9}$$

The quadratic form is

$$\mathbf{x}_s^\top Q_s^{-1}\mathbf{x}_s = (1+g_s)\left(n_s - \frac{g_s c_s}{1+g_s n_s}\right), \tag{S10}$$

where $\mathbf{x}_s^\top\mathbf{x}_s = \sum_{i=1}^N (x_i)^2\delta_{s_i,s} \approx \langle x_i^2\rangle \sum_{i=1}^N \delta_{s_i,s} = n_s$, and $\mathbf{x}_s^\top(\mathbf{1}\mathbf{1}^\top)\mathbf{x}_s = \sum_{i,j=1}^N x_i x_j \delta_{s_i,s}\delta_{s_j,s} \equiv c_s$.

Second, the log-determinant term can be calculated by considering the eigenvalues of the matrix Q_s. Solving for $Q_s\mathbf{z} = \lambda_s\mathbf{z}$ for an arbitrary n_s-dimensional vector $\mathbf{z}$,

$$\lambda_s\mathbf{z} = \alpha_s^s(\mathbf{1}^\top\mathbf{z})\mathbf{1} + \beta_s^2\mathbf{z}; \tag{S11}$$

there are two types of solutions. The first possibility is to have the eigenvector $\mathbf{z} \propto \mathbf{1}$, in which case $\lambda_{s,1} = \alpha_s^2 n_s + \beta_s^2 = (1+g_s n_s)/(1+g_s)$. The other possibility is to have $(\lambda_s - \beta_s^2)\mathbf{z}$ vanish, where $\lambda_{s,2} = \cdots = \lambda_{s,n_s} = \beta_s^2 = 1/(1+g_s)$; the degenerate eigenvectors span the remaining (n_s-1)-dimensional subspace. Therefore

$$\det(Q_s) = (\alpha_s^2 n_s + \beta_s^2)\cdot(\beta_s^2)^{n_s-1} = \frac{1+g_s n_s}{(1+g_s)^{n_s}}, \tag{S12}$$

and

$$\log\det(Q_s) = \log(1+g_s n_s) - n_s\log(1+g_s). \tag{S13}$$

Substituting Eq. S10 and Eq. S13 into Eq. S7, we have derived Eq. 6 as shown the main text.

D. Modification of the correlation matrix for the secondary domain solution. Here we show how the correlation matrix can be modified to solve for a secondary domain solution (for example compartments), by taking out the contribution from the primary domain solution (for example the meta-TADs), which is supposed to be known already.

We extend the group model to consider a bivariate grouping, $i \mapsto (s_i, u_i)$, where each genomic locus i simultaneously belongs to a primary group s_i and a secondary group u_i, presumably at different scales. Generalizing Eq. 3, we assume a linear model

$$x_i = \frac{1}{\sqrt{1+g_{s_i}+h_{u_i}}}\left(\epsilon_i + \sqrt{g_{s_i}}\,\eta_{s_i} + \sqrt{h_{u_i}}\,\xi_{u_i}\right),$$

where ξ_{u_i} and h_{u_i} are respectively the random variable and the grouping strength parameter that correspond to the secondary group u_i. If we further assume that s_i and u_i are statistically independent, the pairwise correlation between two loci i and j can be written as

$$\langle x_i x_j\rangle = \frac{1}{1+g_{s_i}+h_{u_i}}\left(\delta_{ij} + g_{s_i}\delta_{s_i,s_j} + h_{u_i}\delta_{u_i,u_j}\right);$$

the contributions from different groups are additive.

We can do a straightforward algebra to rearrange the above model as

$$\frac{1+g_{s_i}+h_{u_i}}{1+h_{u_i}}\left(\langle x_i x_j\rangle - \frac{g_{s_i}\,\delta_{s_i,s_j}}{1+g_{s_i}+h_{u_i}}\right) = \frac{\delta_{ij}+h_{u_i}\delta_{u_i,u_j}}{1+h_{u_i}},$$

such that the right-hand side becomes the single-group model. The left-hand side of this expression is a normalized residual of the correlation, which we will call C_{ij}^{res}. If we have already inferred the primary group s_i and the corresponding strength $\tilde{g}$ without considering the secondary group, the correspondence to this two-group model (due to normalization) is given as $g = (1+h)\tilde{g}$. Substituting this, and replacing the model correlation $\langle x_i x_j\rangle$ with the data C_{ij}, the left-hand side of the previous expression is simplified to

$$C_{ij}^{\text{res}} = (1+\tilde{g}_{s_i})\,C_{ij} - \tilde{g}_{s_i}\delta_{s_i,s_j}. \tag{S14}$$

C^{res} is written in terms of the original data C and the primary solution s (and $\tilde{g}$) only; it is independent of the unknown secondary solution u (and h). Now u is the solution of a modified single-group problem, using this residual correlation C^{res} as the input data.

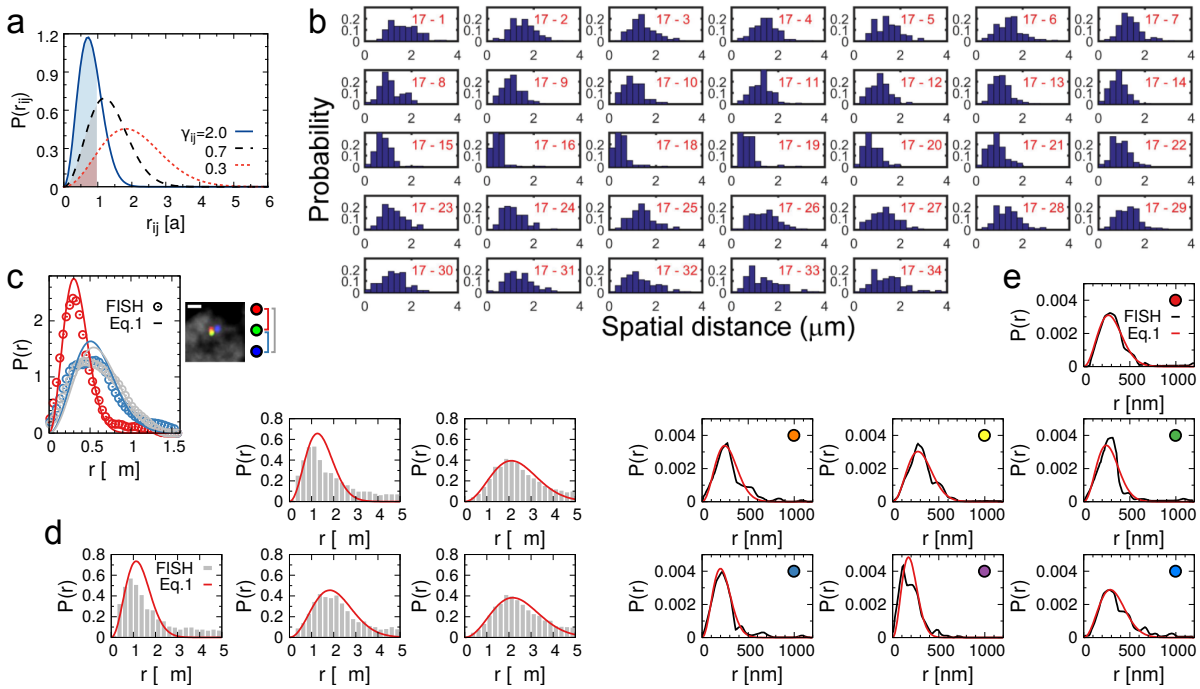

Fig. S1. Distance distributions of segment pairs are described by Gaussian. (a) Gaussian probability distribution plotting $P(r_{ij})$ with different values of γ_{ij} (Eq. 1). The shaded area in different colors represents the corresponding values of contact probabilities, (Eq. 10 at $r_c = 1$) **(b)** 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 (21). **(c)** Distance distributions between three FISH probes on the X chromosome of male *Drosophila* embryos. The experimental data were digitized from Fig.3B in (51). 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 (52). 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 (53), and their corresponding fits are shown in red.

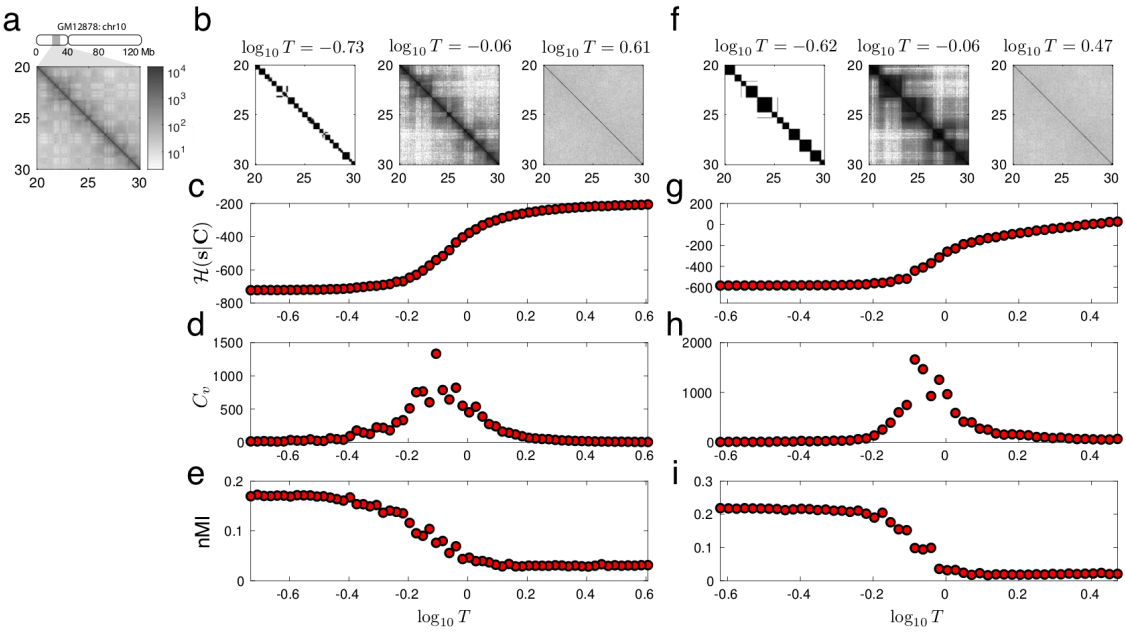

Fig. S2. Finding the best domain solution through simulated annealing. (a) A subset of Hi-C data, covering 10-Mb genomic region on chr10 of GM12878. **(b)** CD solutions, obtained from the Hi-C data in **(a)**, at three values of T for $\lambda = 0$. The CD solution at each T was constructed by $2,000$ 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 $\lambda = 10$.

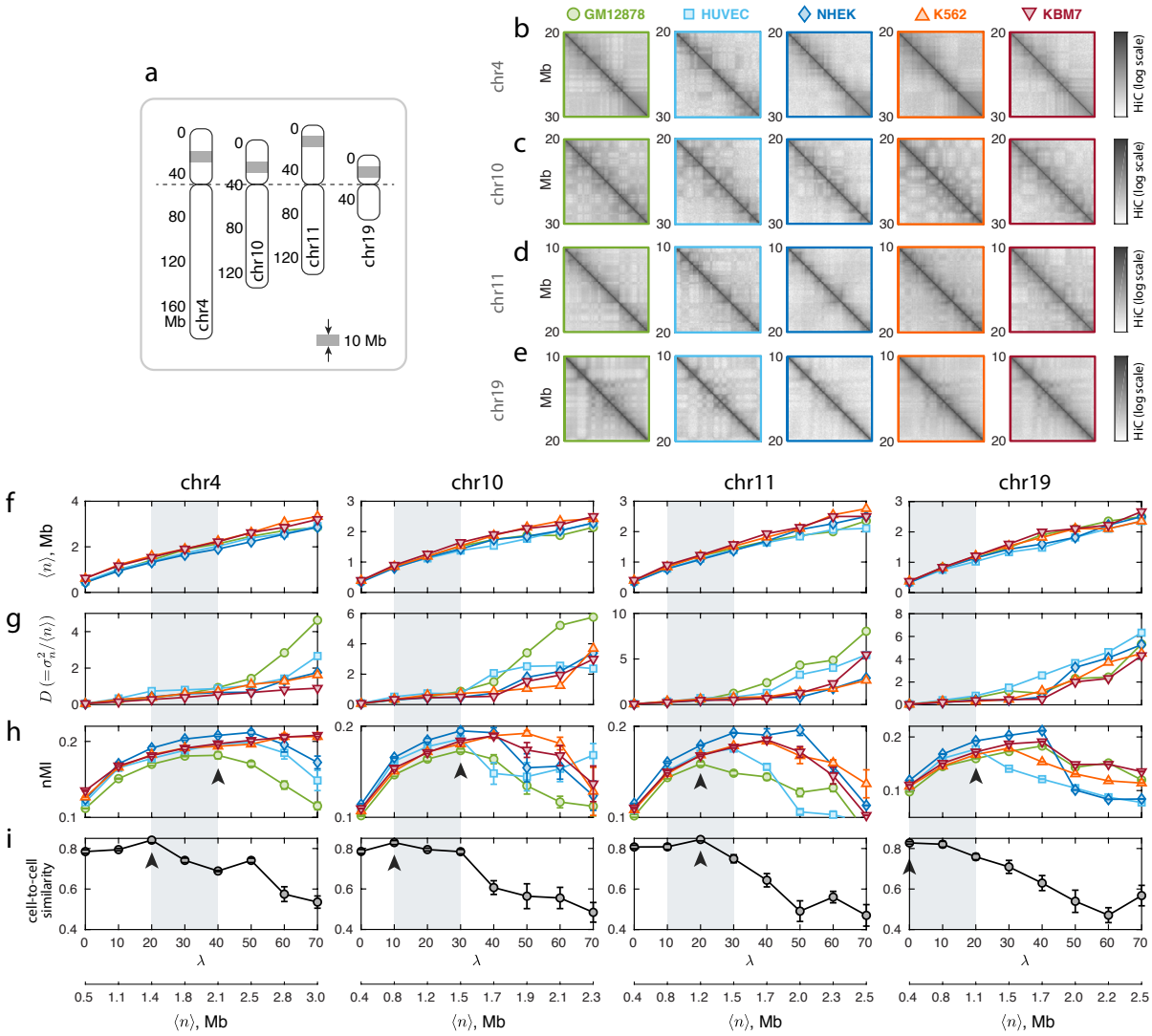

Fig. S3. Chromatin domain solutions for chromosomes 4, 10, 11 and 19. **(a)** Relative sizes of chromosomes considered, aligned at the centromeres. The gray shade in each chromosome indicates the 10-Mb interval for which we show the Hi-C data in the next panels. **(b-e)** Hi-C data for the corresponding 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. 3 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 ($\log_{10}$ 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).

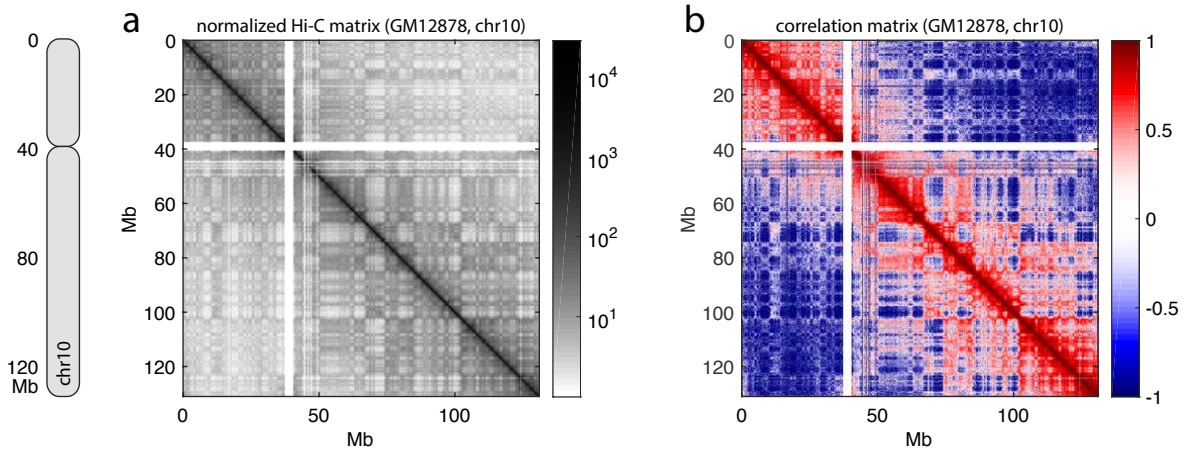

Fig. S4. A full-chromosome view of the input data, for the chromosome 10 of GM12878. **(a)** the normalized Hi-C matrix $\mathbf{M}$, and **(b)** the correlation matrix $\mathbf{C}$.

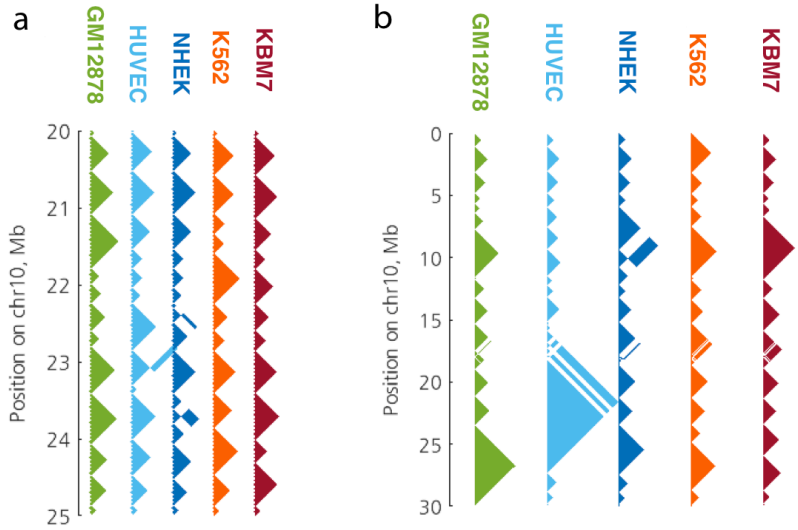

Fig. S5. Examples of Multi-CD domain solutions at different scales. Shown are the domain solutions obtained from Multi-CD for the five different cell lines (GM12878, HUVEC, NHEK, K562, KBM7), at (a) $\lambda = 0$ and (b) $\lambda = 40$.

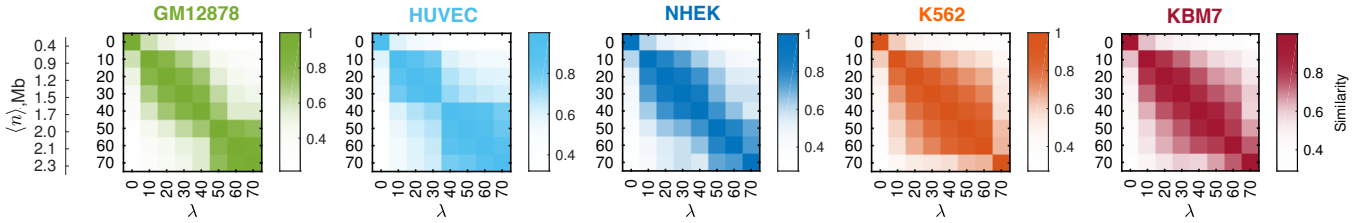

Fig. S6. Scale-to-scale similarity of domain solutions for different cell lines. We calculate the similarity between domain solutions at different λ in terms of Pearson correlation. The calculation was performed for chromosome 10 from five different cell lines.

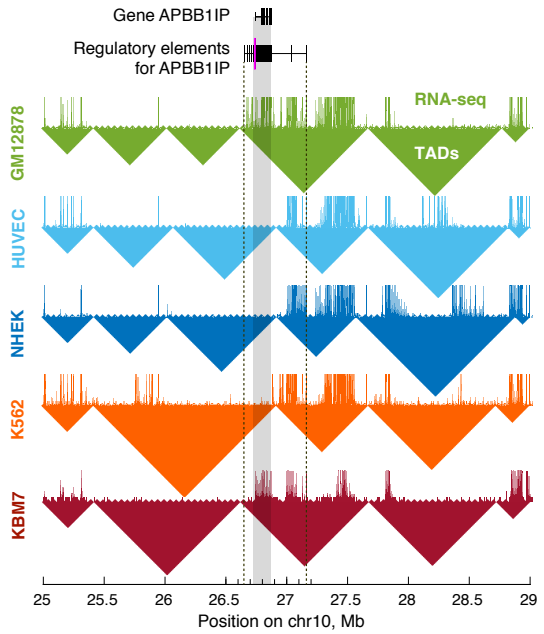

Fig. S7. Cell-line dependent TAD organization and its link to gene expression. The RNA-seq signals from five different cell lines (colored hairy lines) are shown on top of the TAD solutions obtained by Multi-CD (triangles with matching colors). At the top shown are the the position of a specific gene 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.

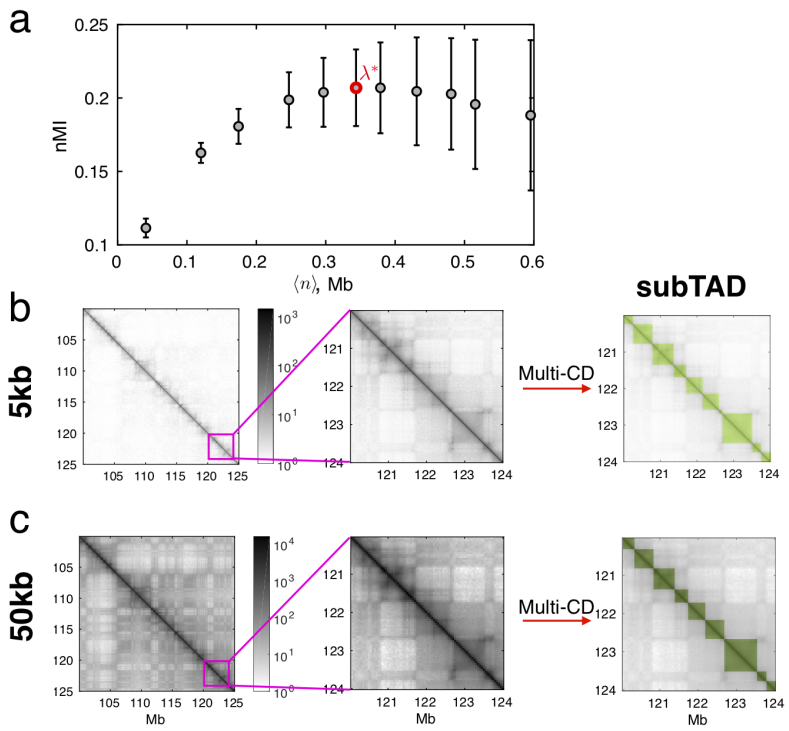

Fig. S8. Identification of sub-TAD boundaries at 5-kb resolution. **(a)** The optimum cluster size, best describing 5-kb resolution Hi-C map in terms of nMI, is determined at $\langle n \rangle = 0.35$ Mb, which is consistent with the sub-TAD size determined from 50-kb resolution Hi-C at $\lambda = 0$. **(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.

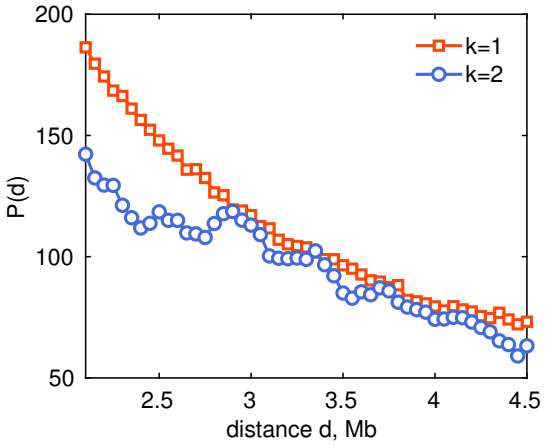

Fig. S9. Intra-domain contact profiles for the two compartment solutions. We plot the genomic distance-dependent contact number for the two large domains $k = 1$ and $k = 2$, from the solution at $\lambda = 90$ for chromosome 10 in GM12878. At short genomic distance, the domain solution of $k = 1$ is 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.

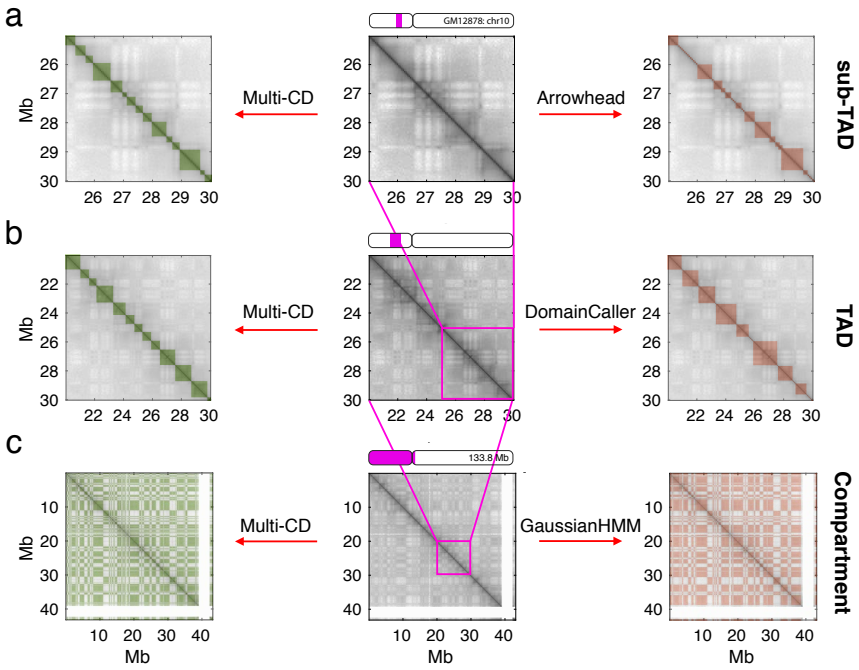

Fig. S10. Comparison of domain solutions from Multi-CD and other methods at specific scales. Comparison between domain solutions obtained by three popular algorithms (ArrowHead, DomainCaller, GaussianHMM) (right column) and those by Multi-CD (left column), applied to 50-kb resolution Hi-C data. Three 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 (19) was used for identifying the domain structures of sub-TADs, DomainCaller (22) for TADs, and Gaussian Hidden Markov Model (GaussianHMM) (19) for compartments. Multi-CD use $\lambda = 0$, 10, 90, as the parameter values for identifying sub-TADs, TADs, and compartments, respectively.

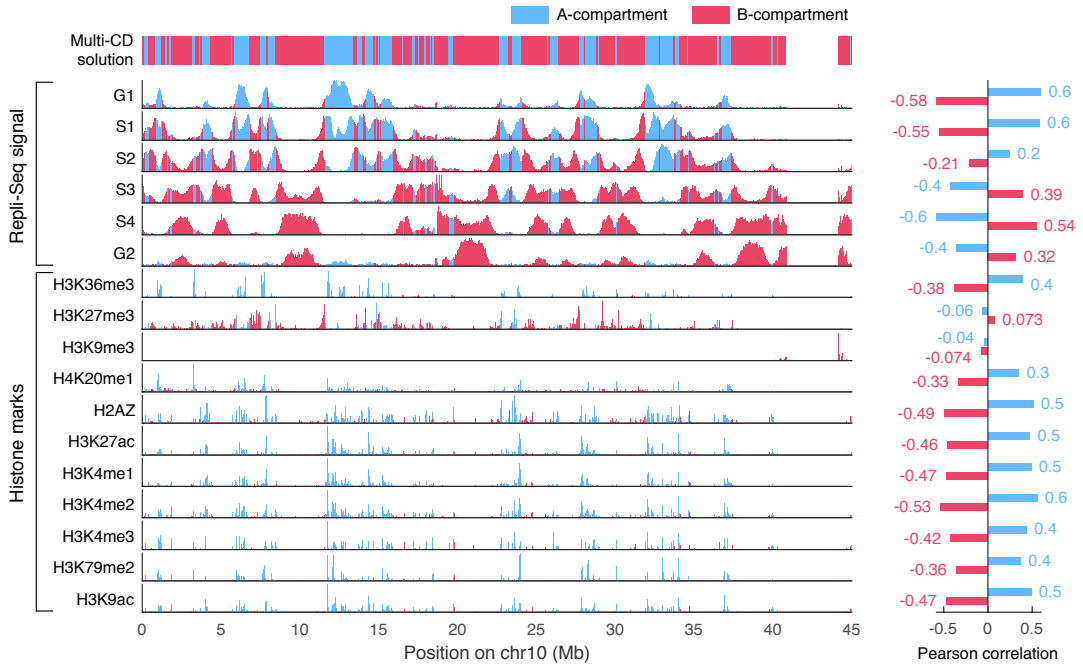

Fig. S11. Comparison of histone marks and compartments. Extension of Fig. 6c-d which make comparison between the CD solutions for A/B-compartments by Multi-CD and epigenetic marks. The upper part with Repli-Seq signals is a reprint from the main text figure. The lower part shows histone marks on the 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.

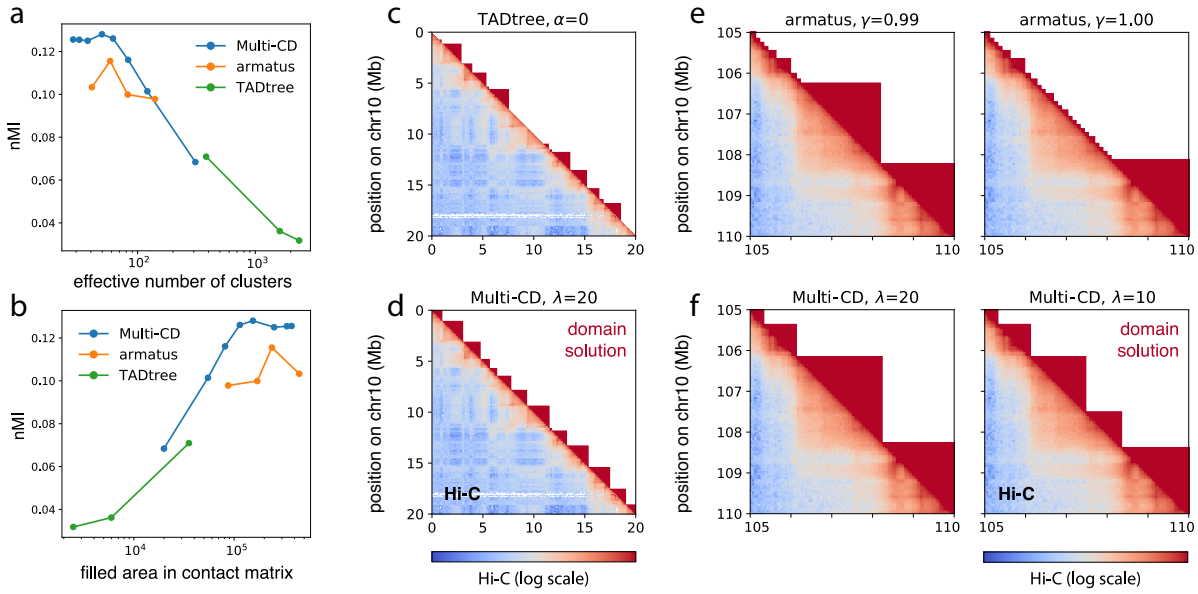

Fig. S12. Comparison to existing algorithms for identifying domains at multiple scales. (a,b) Normalized mutual information between domain solutions at multiple scales, from Multi-CD, Armatus (33) and TADtree (32) respectively, and the log10 of KR-normalized Hi-C matrix for chr10 of the cell line GM12878. The scale of a domain solution $\mathbf{s}$ is measured in two ways, in terms of **(a)** the effective number of clusters, $\mathcal{K}(\mathbf{s}) = \exp\left(-\sum_{k=1}^{K}(n_k/N)\log(n_k/N)\right)$, 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).

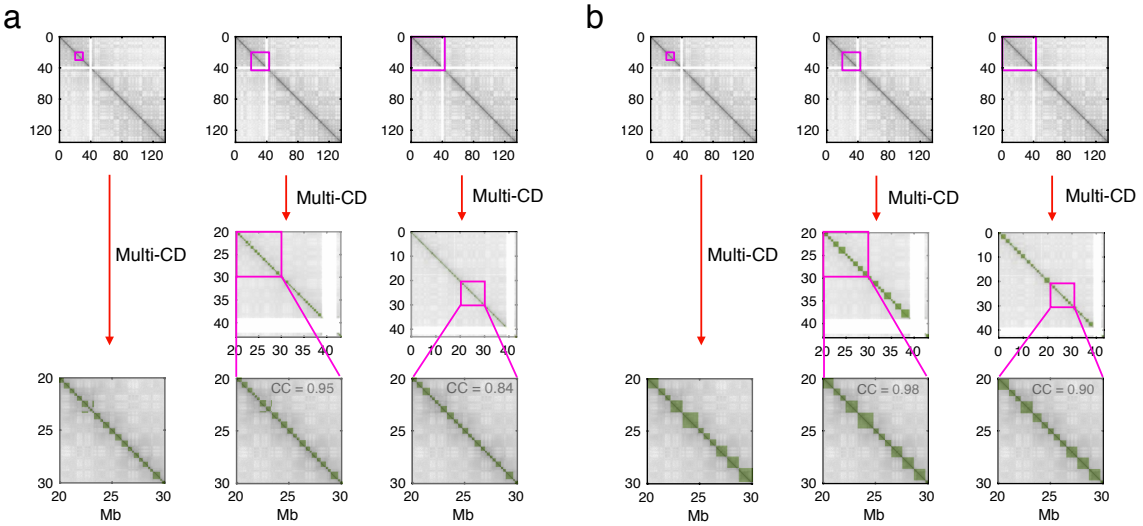

Fig. S13. Robustness of clustering solutions over different subsets of Hi-C data. Here we compare domain solutions from Hi-C inputs of different size. Multi-CD is confirmed to be *locality-preserving*. That is, the sets of domain solutions determined from Hi-C inputs with different sizes remain almost 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.

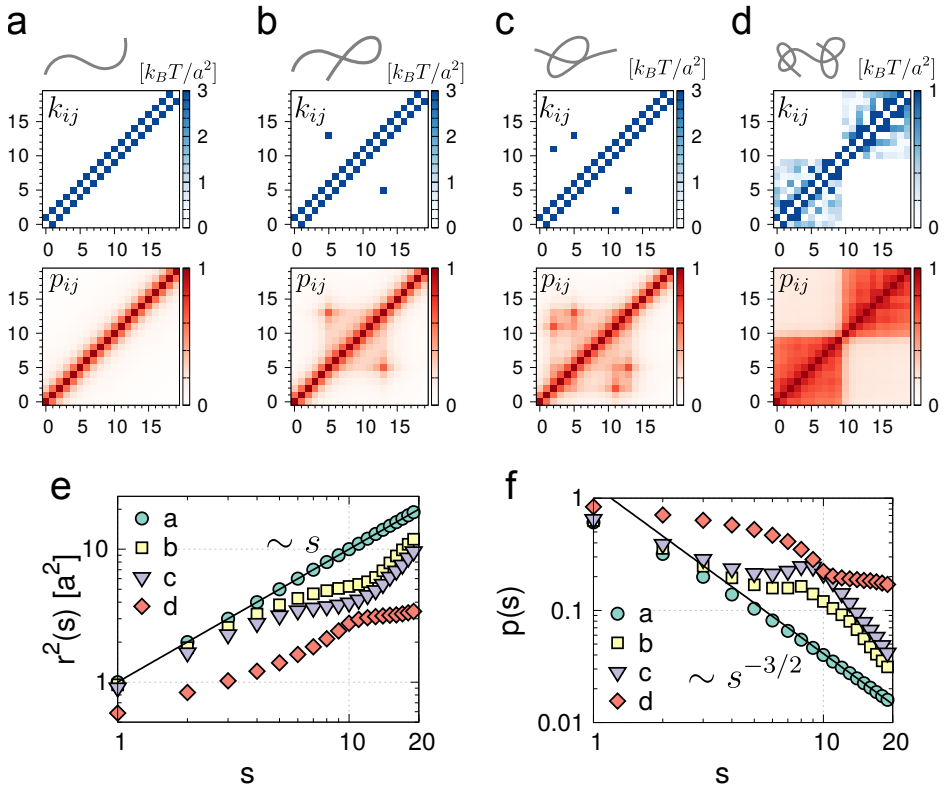

Fig. S14. Heterogenous loop model (14) to compare the contact probabilities of Gaussian polymer networks. (a-d) Four examples of polymer models composed of 20 monomers with different interaction strength matrix $[k_{ij}]$ (top row), and the corresponding contact probability matrices $[p_{ij}]$ (second row) calculated with $r_c = 1$. **(e)** the mean square distance and **(f)** the contact probability $p(s)$ are calculated as a function of the genomic distance, s, for the four different models (a-d). Scaling results in (e) and (f) show that even the Gaussian polymer network model can produce rich multi-scale structure with domains.