Supplementary Information for:

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

Ji Hyun Bak, Min Hyeok Kim, Lei Liu, Changbong Hyeon

S1 Appendix

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 2**).

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 [1–5]. However, chromosome dynamics in each phase during the cell cycle is slow enough [6–9] that the system remains in local mechanical equilibrium over an extended time period, as captured by the stable patterns in the Hi-C data [10]. Although the notion of cell-to-cell variation in a population of cells is appreciated in the literature [7, 9, 11–16], fluorescence measurement still indicates that the spatial distances between pairs of chromatin segments can be well described by the gaussian distribution [17–20] (**S1 Fig**). This motivates us to model the chromosome structure using a gaussian polymer network whose configuration fluctuates around a mechanically equilibrated local basin of attraction [10, 21–24].

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 [25, 26], and the fluctuation dynamics of native proteins (gaussian network model, [27]). 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 in the name of generalized Rouse model [28].

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 (**S1 Fig**), and the agreement of the 3D structural properties inferred by modeling approaches [10, 23, 24] that share the same philosophy, with particular emphasis on our recent approach of *heterogeneous loop model* (HLM) [10, 29], suggest that Gaussian polymer networks 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 $p_{ij} = \int_0^{r_c} dr_{ij} P(r_{ij}) \sim |i-j|^{-3/2}$.

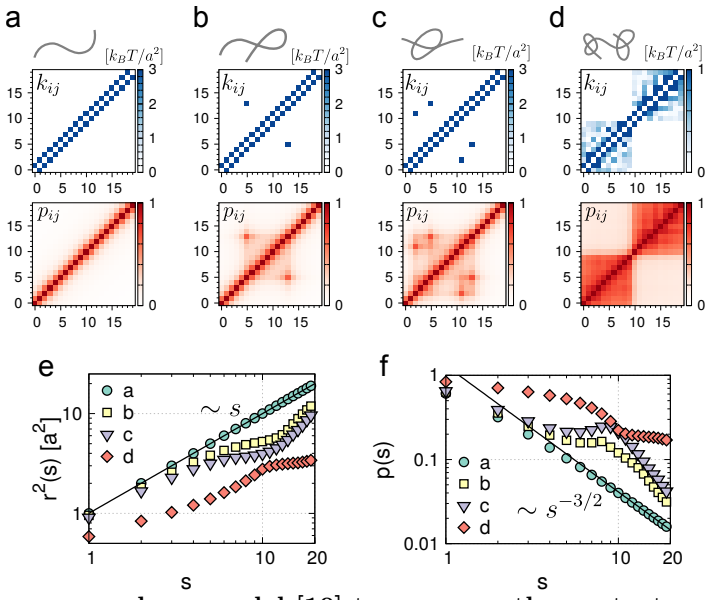


Fig A. Heterogenous loop model [10] to compare the contact probabilities of Gaussian polymer networks. (a-d) Four examples of polymer models composed of 20 monomers with different interaction strength matrix $[k_{ij}]$ (top row), and the corresponding contact probability matrices $[p_{ij}]$ (second row) 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.

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 [10] (**Fig A**). 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 $s \equiv |i - j|$ (**Fig A**), and the contact probability p_{ij} (or $p(s)$) is no longer described with a simple scaling relation (**Fig A**). 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 [10] 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.

References

1. Bruinsma R, Grosberg AY, Rabin Y, Zidovska A. Chromatin hydrodynamics. Biophys J. 2014;106(9):1871–1881.
2. Battle C, Broedersz CP, Fakhri N, Geyer VF, Howard J, Schmidt CF, et al. Broken detailed balance at mesoscopic scales in active biological systems. Science. 2016;352:604–607.
3. Hwang W, Hyeon C. Quantifying the Heat Dissipation from Molecular Motor's Transport Properties in Nonequilibrium Steady States. J Phys Chem Lett. 2017;8:250–256.
4. Hwang W, Hyeon C. Energetic Costs, Precision, and Transport Efficiency of Molecular Motors. J Phys Chem Lett. 2018;9:513–520.
5. Fang X, Kruse K, Lu T, Wang J. Nonequilibrium physics in biology. Rev Mod Phys. 2019;91:045004.
6. Rosa A, Everaers R. Structure and Dynamics of Interphase Chromosomes. PLoS Comput Biol. 2008;4:e1000153.
7. Kang H, Yoon YG, Thirumalai D, Hyeon C. Confinement-induced glassy dynamics in a model for chromosome organization. Phys Rev Lett. 2015;115:198102.
8. Liu L, Shi G, Thirumalai D, Hyeon C. Chain organization of human interphase chromosome determines the spatiotemporal dynamics of chromatin loci. PLoS Comp Biol. 2018;14(12):e1006617.
9. Shi G, Liu L, Hyeon C, Thirumalai D. Interphase Human Chromosome Exhibits Out of Equilibrium Glassy Dynamics. Nat Commun. 2018;9:3161.
10. Liu L, Kim MH, Hyeon C. Heterogeneous Loop Model to Infer 3D Chromosome Structures from Hi-C. Biophys J. 2019;117:613–625.
11. Shi G, Thirumalai D. Conformational heterogeneity in human interphase chromosome organization reconciles the FISH and Hi-C paradox. Nature Commun. 2019;10(1):1–10.
12. Bintu B, Mateo LJ, Su JH, Sinnott-Armstrong NA, Parker M, Kinrot S, et al. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. Science. 2018;362:eaau1783.
13. Nir G, Farabella I, Estrada CP, Ebeling CG, Beliveau BJ, Sasaki HM, et al. Walking along chromosomes with super-resolution imaging, contact maps, and integrative modeling. PLoS genetics. 2018;14(12):e1007872.
14. Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature. 2013;502(7469):59–64.
15. Stevens TJ, Lando D, Basu S, Atkinson LP, Cao Y, Lee SF, et al. 3D structures of individual mammalian genomes studied by single-cell Hi-C. Nature. 2017;544(7648):59–64.
16. Finn EH, Pegoraro G, Brandão HB, Valton AL, Oomen ME, Dekker J, et al. Extensive heterogeneity and intrinsic variation in spatial genome organization. Cell. 2019;176(6):1502–1515.

17. Wang S, Su JH, Beliveau BJ, Bintu B, Moffitt JR, Wu Ct, et al. Spatial organization of chromatin domains and compartments in single chromosomes. Science. 2016;353(6299):598–602.

18. Szabo Q, Jost D, Chang JM, Cattoni DI, Papadopoulos GL, Bonev B, et al. TADs are 3D structural units of higher-order chromosome organization in Drosophila. Science advances. 2018;4(2):eaar8082.

19. Finn EH, Pegoraro G, Shachar S, Misteli T. Comparative analysis of 2D and 3D distance measurements to study spatial genome organization. Methods. 2017;123:47–55.

20. Giorgetti L, Galupa R, Nora EP, Piolot T, Lam F, Dekker J, et al. Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. Cell. 2014;157(4):950–963.

21. Sauerwald N, Zhang S, Kingsford C, Bahar I. Chromosomal dynamics predicted by an elastic network model explains genome-wide accessibility and long-range couplings. Nucleic acids research. 2017;45(7):3663–3673.

22. Zhang S, Chen F, Bahar I. Differences in the intrinsic spatial dynamics of the chromatin contribute to cell differentiation. Nucleic Acids Research. 2020;48:1131–1145.

23. Le Treut G, Képès F, Orland H. A Polymer Model for the Quantitative Reconstruction of Chromosome Architecture from HiC and GAM Data. Biophys J. 2018;115(12):2286–2294.

24. Shinkai S, Nakagawa M, Sugawara T, Togashi Y, Ochiai H, Nakato R, et al. PHi-C: deciphering Hi-C data into polymer dynamics. NAR Genomics and Bioinfo. 2020;2(2):lqaa020.

25. Goldbart PM, Zippelius A. Amorphous solid state of vulcanized macromolecules: A variational approach. Phys Rev Lett. 1993;71:2256–2259.

26. Bryngelson JD, Thirumalai D. Internal Constraints Induce Localization in an Isolated Polymer Molecule. Phys Rev Lett. 1996;76:542–545.

27. Haliloglu T, Bahar I, Erman B. Gaussian Dynamics of Folded Proteins. Phys Rev Lett. 1997;79:3090–3093.

28. Hyeon C, Morrison G, Thirumalai D. Force-dependent hopping rates of RNA hairpins can be estimated from accurate measurement of the folding landscapes. Proc Natl Acad Sci U S A. 2008;105(28):9604–9609.

29. Liu L, Hyeon C. Revisiting the organization of Polycomb-repressed domains: 3D chromatin models from Hi-C compared with super-resolution imaging. Nucleic Acids Research. 2020; p. gkaa932.

Supplementary Information for:

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

Ji Hyun Bak, Min Hyeok Kim, Lei Liu, Changbong Hyeon

S2 Appendix

Derivation of the likelihood function

Here we derive the likelihood function, **Eq 14**.

Problem: We want to compute

$$p(\mathbf{x}|\mathbf{s},\mathbf{g}) = \left\langle \delta^N\left(\mathbf{x}-\mathbf{f}(\eta,\epsilon)\right)\right\rangle_{\eta,\epsilon} \tag{S2-1}$$

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

$$f_i(\eta,\epsilon)=\frac{\sqrt{g_{s_i}}\eta_i+\epsilon_i}{\sqrt{1+g_{s_i}}},\quad i=1,\cdots,N. \tag{S2-2}$$

Two lemmas will be useful. The Gaussian integral lemma:

$$\int_{\mathbb{R}^N} d\mathbf{z}\;\mathcal{N}(\mathbf{z}|\boldsymbol{\mu},M)\;e^{i\mathbf{a}^\top\mathbf{z}}=\exp\left(-\frac{1}{2}\mathbf{a}^\top M\mathbf{a}\right),\quad \mathbf{a}\in\mathbb{R}^N; \tag{S2-3}$$

and the Sherman-Morrison formula:

$$(A+\mathbf{u}\mathbf{v}^\top)^{-1}=A^{-1}-\frac{A^{-1}\mathbf{u}\mathbf{v}^\top A^{-1}}{1+\mathbf{v}^\top A^{-1}\mathbf{u}}. \tag{S2-4}$$

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 S2-1**, and evaluate the gaussian integrals using the lemma (**Eq S2-3**):

$$\begin{aligned} p(\mathbf{x}|\mathbf{s},\mathbf{g}) &= \int \frac{d\mathbf{k}}{(2\pi)^N}\, e^{i\mathbf{x}^\top\mathbf{k}} \int d\boldsymbol{\eta}\, \mathcal{N}(\boldsymbol{\eta})\, e^{-iA\boldsymbol{\eta}^\top\mathbf{k}} \int d\boldsymbol{\epsilon}\, \mathcal{N}(\boldsymbol{\epsilon})\, e^{-iB\boldsymbol{\epsilon}^\top\mathbf{k}} \\ &= \int \frac{d\mathbf{k}}{(2\pi)^N}\, \exp\left(i\mathbf{x}^\top\mathbf{k} - \frac{1}{2}(A\mathbf{k})^\top\Lambda(A\mathbf{k}) - \frac{1}{2}(B\mathbf{k})^\top\Sigma(B\mathbf{k})\right) \\ &= \int \frac{d\mathbf{k}}{(2\pi)^N}\, \exp\left(i\mathbf{x}^\top\mathbf{k} - \frac{1}{2}\mathbf{k}^\top Q\mathbf{k}\right), \end{aligned} \tag{S2-5}$$

where $Q \equiv (A\Lambda A + B\Sigma B)$. Recognizing that this is another (unnormalized) gaussian integral with covariance matrix Q^{-1}, we use the lemma (**Eq S2-3**) once again:

$$\begin{aligned} p(\mathbf{x}|\mathbf{s},\mathbf{g}) &= \sqrt{(2\pi)^N \det Q^{-1}} \int \frac{d\mathbf{k}}{(2\pi)^N}\, \mathcal{N}(\mathbf{k}|\mathbf{0},Q^{-1})\, e^{i\mathbf{x}^\top\mathbf{k}} \\ &= \exp\left(-\frac{1}{2}\mathbf{x}^\top Q^{-1}\mathbf{x} - \frac{1}{2}\log\det Q\right). \end{aligned} \tag{S2-6}$$

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

$$\log p(\mathbf{x}|\mathbf{s},\mathbf{g}) = -\frac{1}{2}\sum_{s=1}^{K}\left(\mathbf{x}_s^\top Q_s^{-1}\mathbf{x}_s + \log\det Q_s\right), \tag{S2-7}$$

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 S2-7**, and show that the resulting expression corresponds to **Eq 14**. First, the quadratic term can be expanded by using the Sherman-Morrison formula (**Eq S2-4**):

$$Q_s^{-1} = (\beta_s^2 I_{n_s} + (\alpha_s\mathbf{1}_{n_s})(\alpha_s\mathbf{1}_{n_s})^\top)^{-1} = \frac{1}{\beta_s^2}\left(I - \frac{(\alpha_s^2/\beta_s^2)\mathbf{1}\mathbf{1}^\top}{1+(\alpha_s^2/\beta_s^2)\mathbf{1}^\top\mathbf{1}}\right). \tag{S2-8}$$

The quadratic form is

$$\mathbf{x}_s^\top Q_s^{-1}\mathbf{x}_s = (1+g_s)\left(n_s - \frac{g_s c_s}{1+g_s n_s}\right), \tag{S2-9}$$

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 $\mathbf{z}$,

$$\lambda_s\mathbf{z} = \alpha_s^s(\mathbf{1}^\top\mathbf{z})\mathbf{1} + \beta_s^2\mathbf{z}; \tag{S2-10}$$

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

$$\det(Q_s) = (\alpha_s^2 n_s + \beta_s^2) \cdot (\beta_s^2)^{n_s - 1} = \frac{1 + g_s n_s}{(1+g_s)^{n_s}}, \tag{S2-11}$$

and

$$\log\det(Q_s) = \log(1 + g_s n_s) - n_s \log(1 + g_s). \tag{S2-12}$$

Substitution of **Eq S2-9** and **Eq S2-12** into **Eq S2-7** yields **Eq 14**.

Supplementary Information for:

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

Ji Hyun Bak, Min Hyeok Kim, Lei Liu, Changbong Hyeon

S3 Appendix

The inference algorithm

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 $t_{\mathrm{tot}} \geq 5\tau^*$ samples in the CD solution space. The "relaxation time" τ^* is defined as the number of steps at which the autocorrelation function $R(\tau)$, drops significantly ($< 1/e$). The autocorrelation function is calculated as

$$R(\tau) = \frac{1}{\sigma^2} \langle (\mathcal{H}(\mathbf{s}_t|\mathbf{C}) - \mu)(\mathcal{H}(\mathbf{s}_{t+\tau}|\mathbf{C}) - \mu) \rangle_t, \tag{S3-1}$$

where $\mathbf{s}_t$ is the t-th sample in the chain, and μ and σ are the mean and standard deviation of $\mathcal{H}$. The average $\langle \cdot \rangle_t$ is taken over all pairs of samples with a delay of τ.

Simulated annealing The simulated annealing process is described below. Also see **Fig A** for an example of simulated annealing in our Multi-CD algorithm.

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 $c_{\mathrm{cool}} \in (0, 1)$, such that the next temperature is $T_{r+1} = c_{\mathrm{cool}} \cdot T_r$. We used $c_{\mathrm{cool}} = 0.95$ 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 $\mathcal{H}(\mathbf{s}^*|\mathbf{C})$.

Robustness of solutions over data subset choices The domain solutions reported by Multi-CD are robust over different choices as to which subsets of Hi-C data we solve from. We showed that Multi-CD is practically *locality-preserving*, in the following sense. Suppose that $S_1, S_2 \subset \{1, 2, \cdots, N\}$ be two subsets (specifically, consecutive intervals) of the genomic range, and both include the two genomic segments i, j. At a given λ, if the pair (i, j) belongs to the same domain according to a domain solution based on the subset of data $\mathbf{C}_{S_1}$, most of the times it also belongs to the same domain when solved for the other subset $\mathbf{C}_{S_2}$. Also see **Fig B** for an example from the real data.

Computational cost Repeated sampling in the simulated annealing is the computational bottleneck for the current method. Whereas our final choice of parameters for the simulated annealing was on the conservative side, to prioritize accurate solutions over speed, it is often useful to adjust the parameters to enable lighter runs, especially for pilot studies. For the MCMC sampling at each fixed temperature, the chain length (set adaptively by the stopping condition; we used $5\tau^*$ throughout this study) could be reduced, for example to $3\tau^*$. In general, one can trade off the number of independent simulated annealing runs (try a larger number of initial configurations), which is readily parallelized, for a shorter sampling per run. For the simulated annealing, the temperature schedule can be accelerated by adjusting the cooling rate c_{cool} (currently 0.95); a smaller c_{cool}, such as 0.9, results in a faster annealing.

In addition to adjusting the simulated annealing parameters listed above, one can also use smaller data subsets (with smaller subset size N) for faster test runs. More specifically, we used a smaller data subset and adjusted simulated annealing parameters (shorter chain length, accelerated cooling rate, etc.) to perform pilot runs to determine a rough range of λ values that is meaningful for the given data. Then we performed a more thorough run to obtain the actual results. The full set of parameters that we used for the main analysis, as well as for the shorter test runs, can be found in our public code repository.

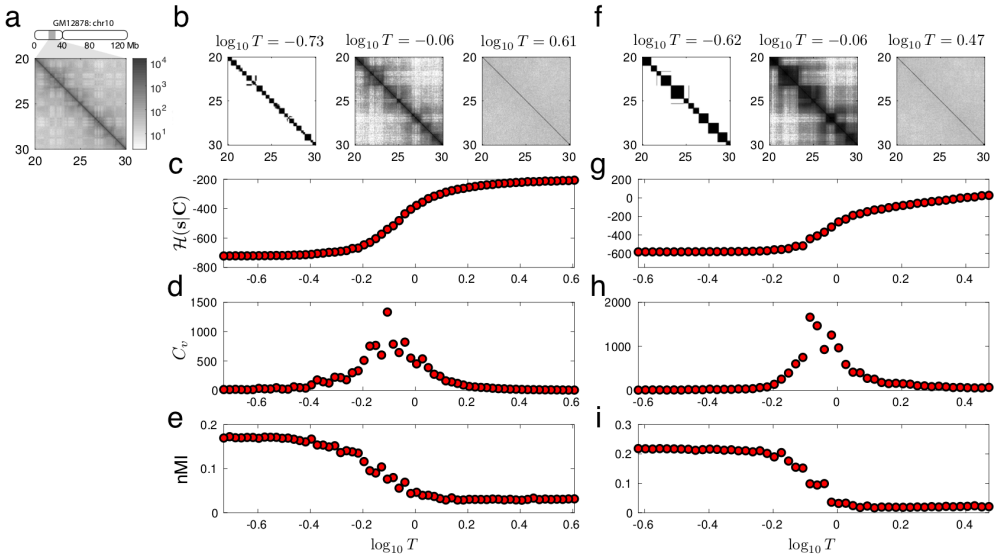


Fig A. Finding the best domain solution through simulated annealing. **(a)** A subset of Hi-C data, covering 10-Mb genomic region on chr10 of GM12878. **(b)** CD solutions, obtained from the Hi-C data in **(a)**, at three values of T for $\lambda = 0$. The CD solution at each T was constructed by $2,000$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

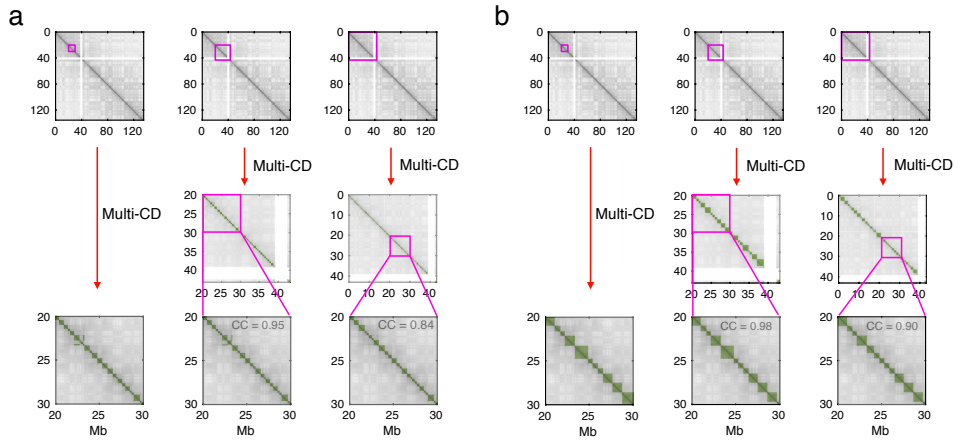
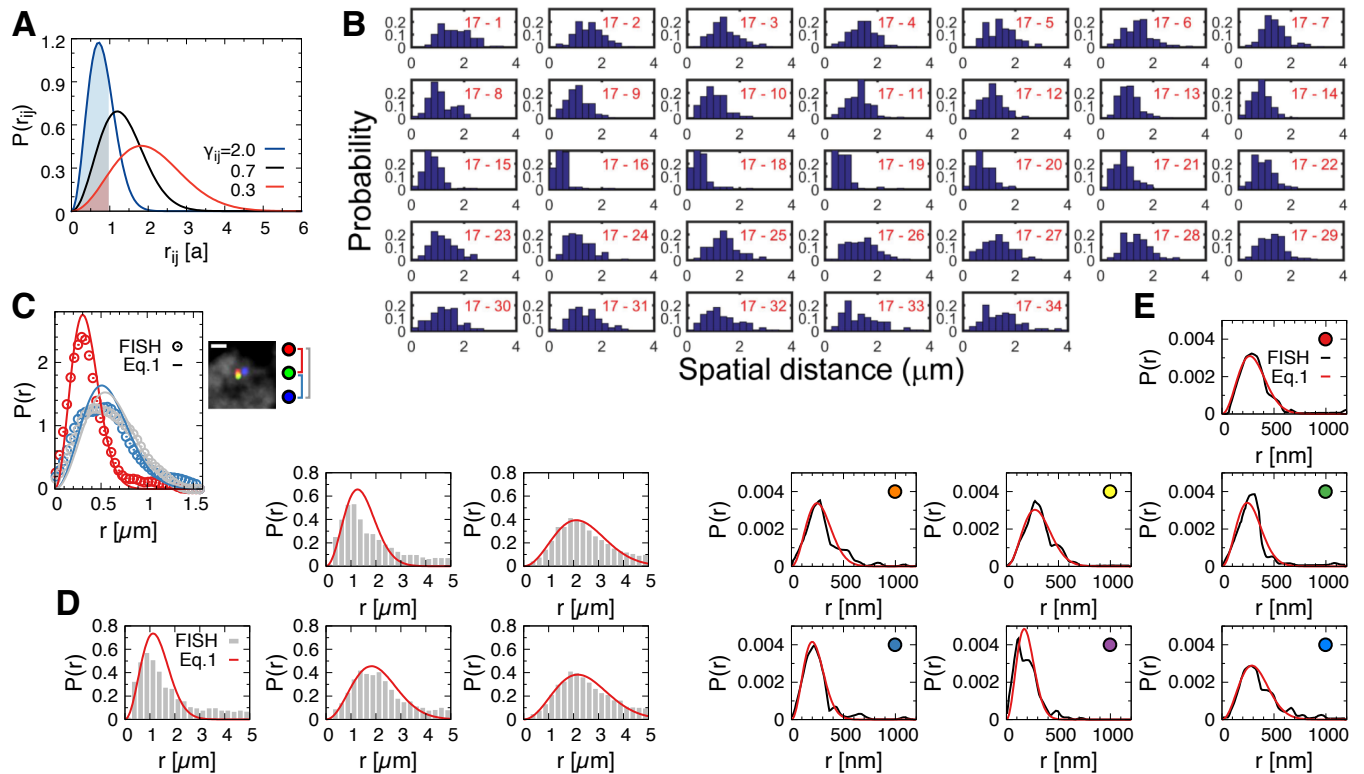
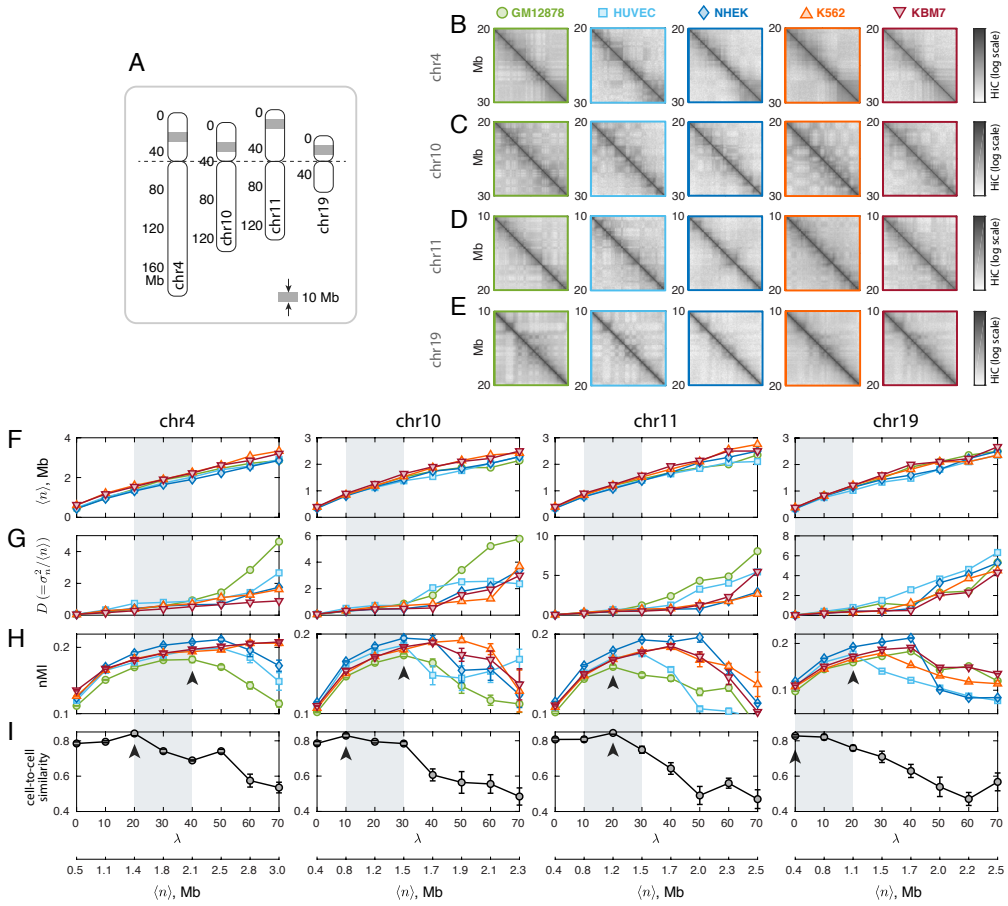
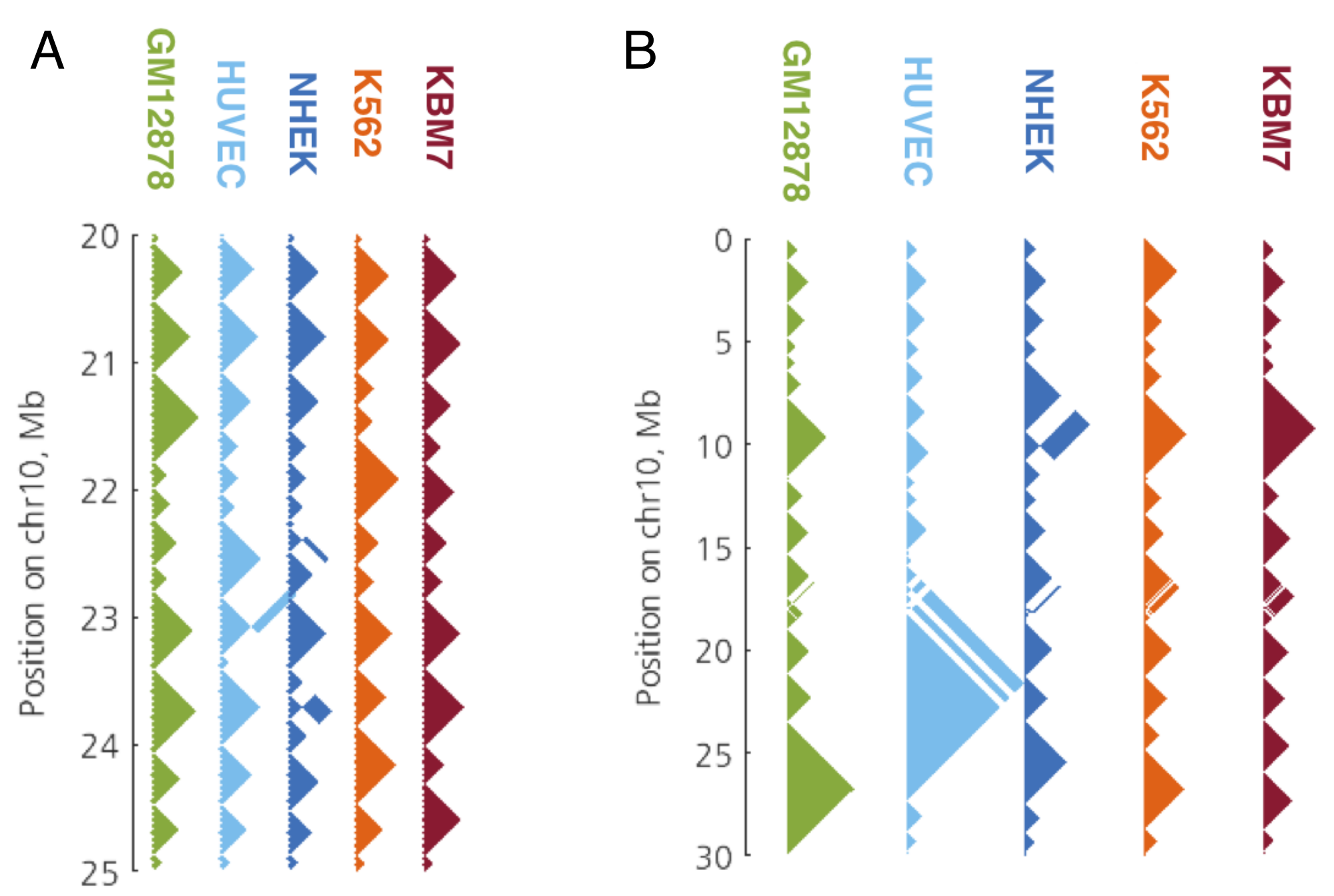
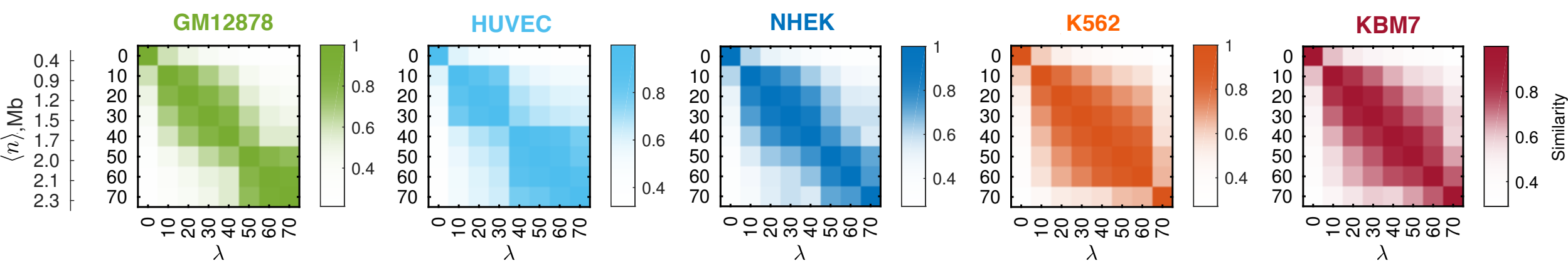


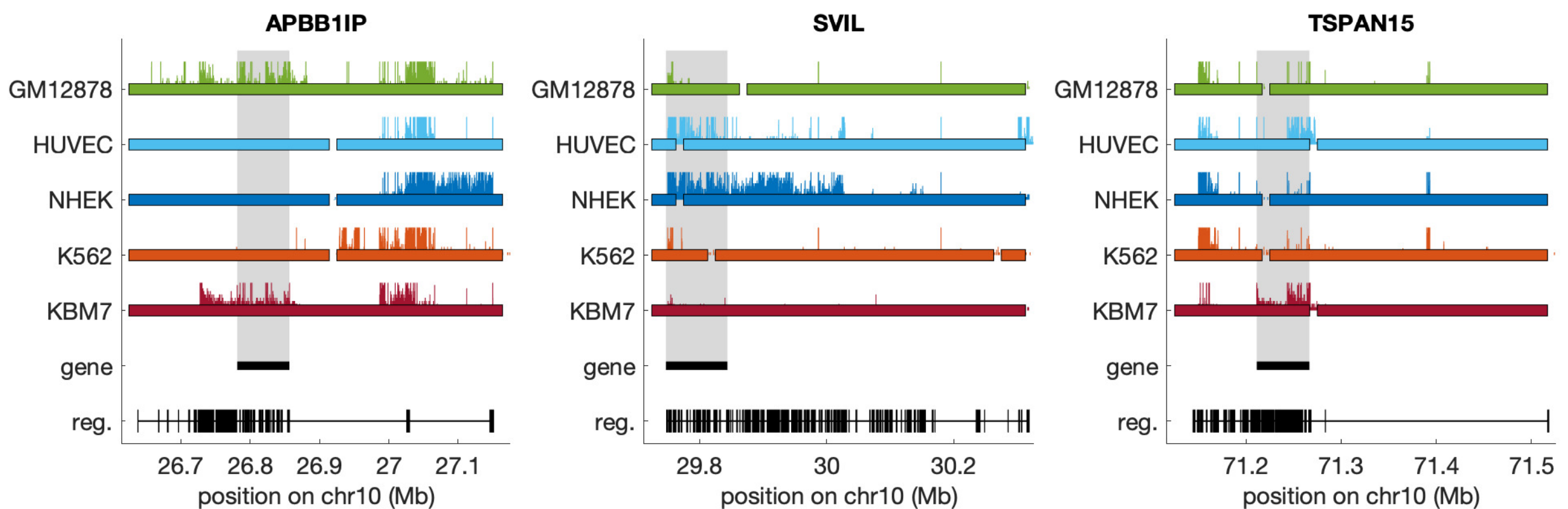
Fig B. Robustness of clustering solutions over different subsets of Hi-C data. The Hi-C data demarcated by the purple squares on the top panels are the input data used for Multi-CD analysis. The three panels from left to right on the bottom are the domain solutions from 10-Mb, 20-Mb, and 40-Mb Hi-C inputs. **(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.

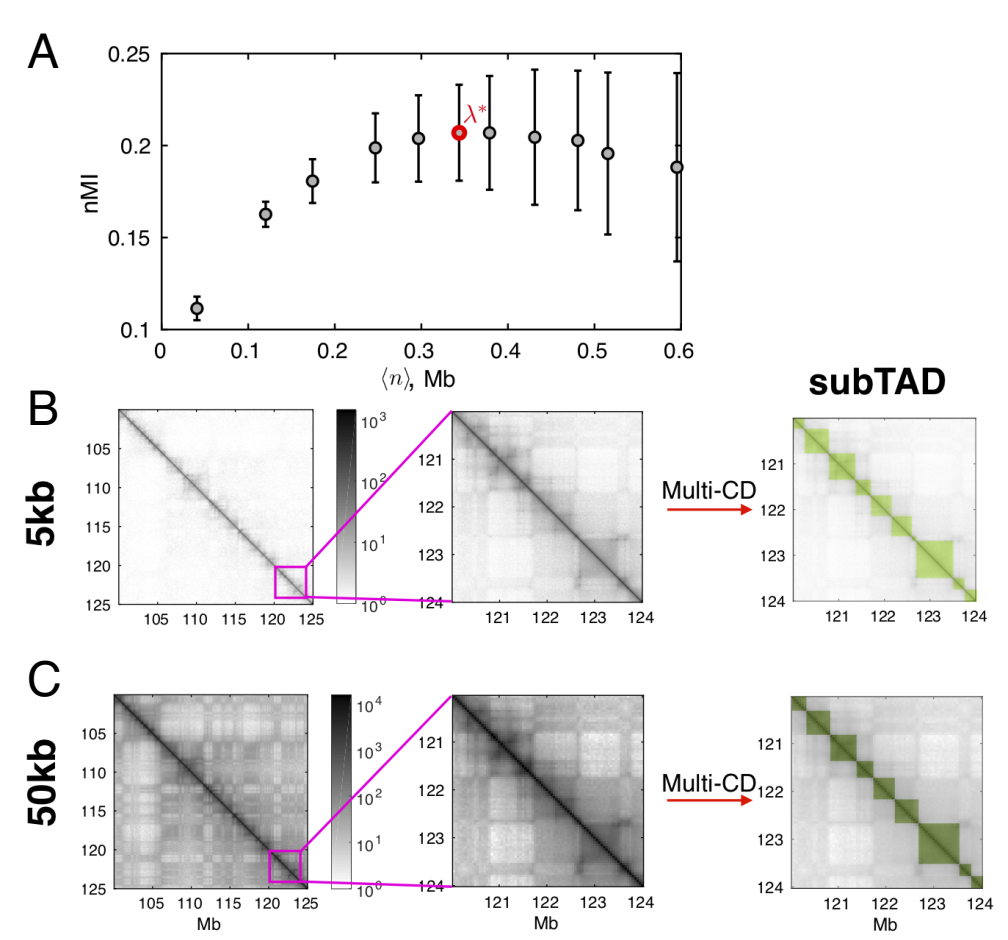

A
P(r_ij)
r_ij [a]
γ_ij=2.0
0.7
0.3
B
Probability
Spatial distance (μm)
17 - 1
17 - 2
17 - 3
17 - 4
17 - 5
17 - 6
17 - 7
17 - 8
17 - 9
17 - 10
17 - 11
17 - 12
17 - 13
17 - 14
17 - 15
17 - 16
17 - 18
17 - 19
17 - 20
17 - 21
17 - 22
17 - 23
17 - 24
17 - 25
17 - 26
17 - 27
17 - 28
17 - 29
17 - 30
17 - 31
17 - 32
17 - 33
17 - 34
C
P(r)
FISH
Eq.1
r [μm]
D
E
r [nm]


A
chr4
chr10
chr11
chr19
10 Mb
B
C
D
E
GM12878
HUVEC
NHEK
K562
KBM7
HiC (log scale)
chr4
chr10
chr11
chr19
F
G
H
I
⟨n⟩, Mb
D (=σ²n/⟨n⟩)
nMI
cell-to-cell similarity
λ
⟨n⟩, Mb


A
GM12878
HUVEC
NHEK
K562
KBM7
Position on chr10, Mb
20
21
22
23
24
25
B
GM12878
HUVEC
NHEK
K562
KBM7
Position on chr10, Mb
0
5
10
15
20
25
30


GM12878
HUVEC
NHEK
K562
KBM7
⟨n⟩,Mb
0.4
0.9
1.2
1.5
1.7
2.0
2.1
2.3
0
10
20
30
40
50
60
70
λ
1
0.8
0.6
0.4
Similarity

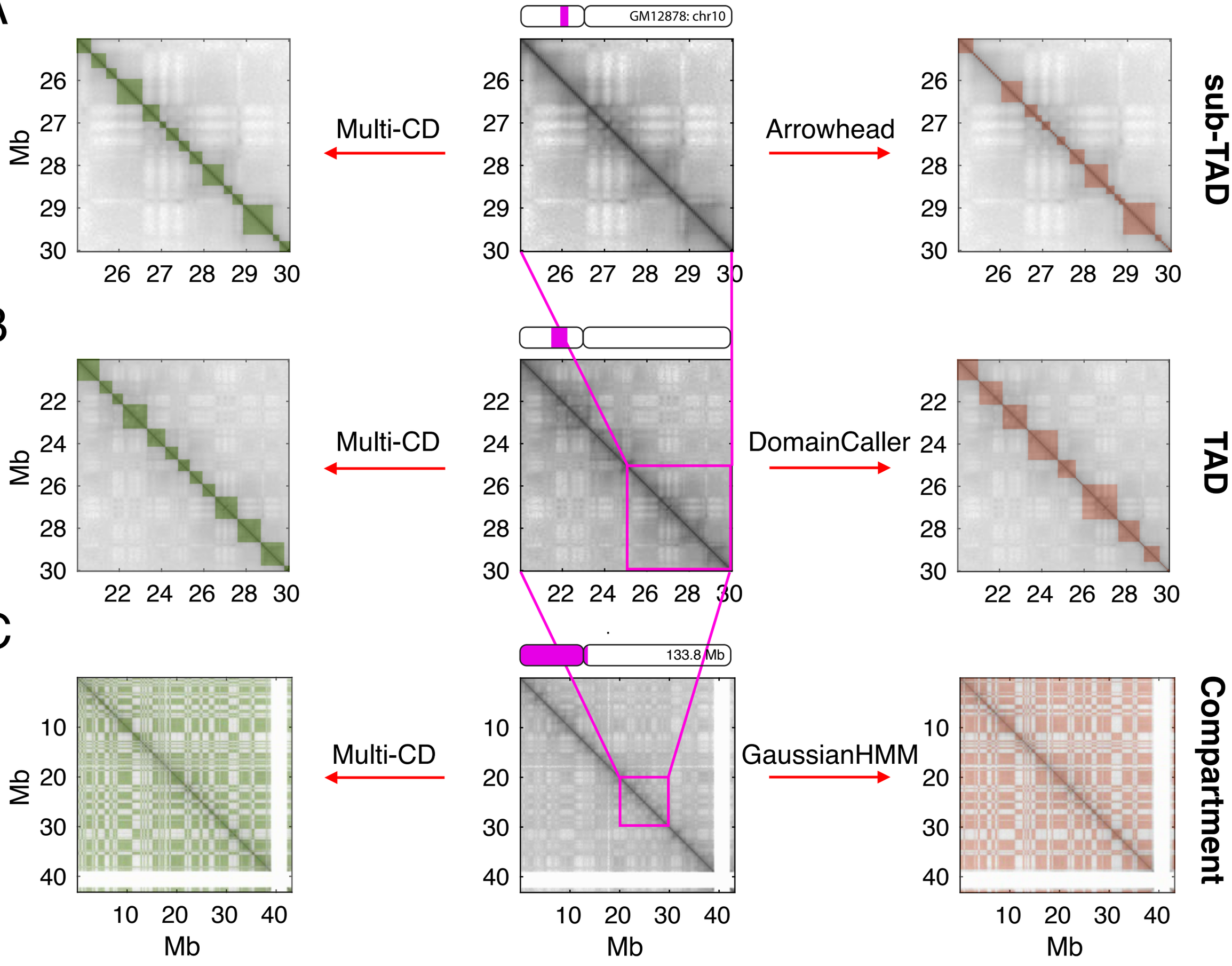

APBB1IP
SVIL
TSPAN15
GM12878
HUVEC
NHEK
K562
KBM7
gene
reg.
26.7
26.8
26.9
27
27.1
29.8
30
30.2
71.2
71.3
71.4
71.5
position on chr10 (Mb)
position on chr10 (Mb)
position on chr10 (Mb)

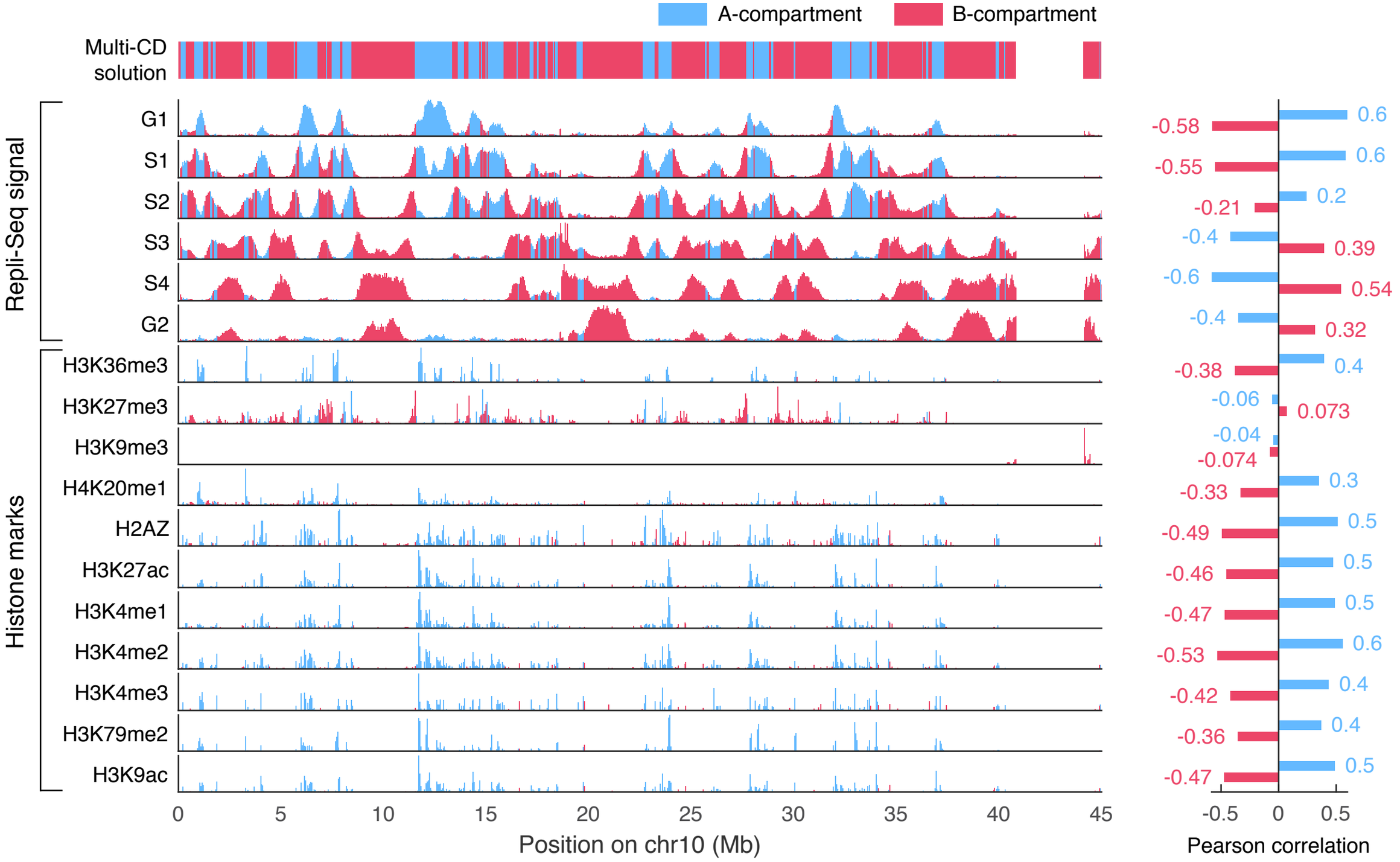

A
nMI
0.25
0.2
0.15
0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
$\langle n \rangle$, Mb
λ^*
B
5kb
subTAD
Multi-CD
C
50kb
Multi-CD
Mb

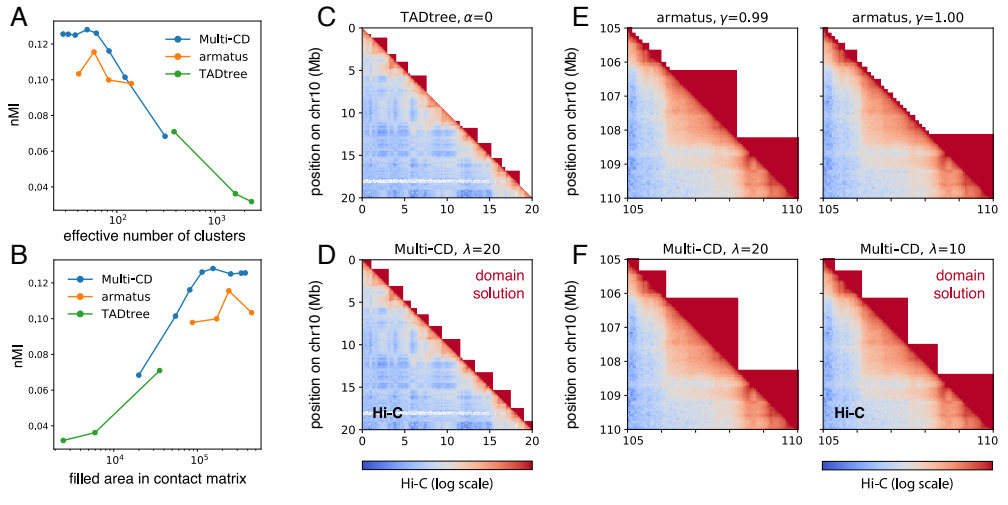
A

B

C


GM12878: chr10
Multi-CD
Arrowhead
sub-TAD
Multi-CD
DomainCaller
TAD
133.8 Mb
Multi-CD
GaussianHMM
Compartment
Mb


A-compartment
B-compartment
Multi-CD solution
Repli-Seq signal
G1
S1
S2
S3
S4
G2
Histone marks
H3K36me3
H3K27me3
H3K9me3
H4K20me1
H2AZ
H3K27ac
H3K4me1
H3K4me2
H3K4me3
H3K79me2
H3K9ac
-0.58
0.6
-0.55
0.6
-0.21
0.2
-0.4
0.39
-0.6
0.54
-0.4
0.32
-0.38
0.4
-0.06
0.073
-0.04
-0.074
-0.33
0.3
-0.49
0.5
-0.46
0.5
-0.47
0.5
-0.53
0.6
-0.42
0.4
-0.36
0.4
-0.47
0.5
0
5
10
15
20
25
30
35
40
45
Position on chr10 (Mb)
-0.5
0
0.5
Pearson correlation


A
Multi-CD
armatus
TADtree
nMI
effective number of clusters
B
filled area in contact matrix
C
TADtree, α=0
position on chr10 (Mb)
D
Multi-CD, λ=20
domain solution
Hi-C
Hi-C (log scale)
E
armatus, γ=0.99
armatus, γ=1.00
F
Multi-CD, λ=20
Multi-CD, λ=10