BMC Bioinformatics

METHODOLOGY ARTICLE

Open Access

Deciphering hierarchical organization of topologically associated domains through change-point testing

Haipeng Xing[1†], Yingru Wu[1†], Michael Q. Zhang[2] and Yong Chen[3*]

*Correspondence: chenyong@rowan.edu
†Haipeng Xing and Yingru Wu have contributed equally
[3] Department of Molecular and Cellular Biosciences, Rowan University, 201 Mullica Hill Rd, Glassboro, NJ 08028, USA
Full list of author information is available at the end of the article

Abstract

Background: The nucleus of eukaryotic cells spatially packages chromosomes into a hierarchical and distinct segregation that plays critical roles in maintaining transcription regulation. High-throughput methods of chromosome conformation capture, such as Hi-C, have revealed topologically associating domains (TADs) that are defined by biased chromatin interactions within them.

Results: We introduce a novel method, HiCKey, to decipher hierarchical TAD structures in Hi-C data and compare them across samples. We first derive a generalized likelihood-ratio (GLR) test for detecting change-points in an interaction matrix that follows a negative binomial distribution or general mixture distribution. We then employ several optimal search strategies to decipher hierarchical TADs with *p* values calculated by the GLR test. Large-scale validations of simulation data show that HiCKey has good precision in recalling known TADs and is robust against random collisions of chromatin interactions. By applying HiCKey to Hi-C data of seven human cell lines, we identified multiple layers of TAD organization among them, but the vast majority had no more than four layers. In particular, we found that TAD boundaries are significantly enriched in active chromosomal regions compared to repressed regions.

Conclusions: HiCKey is optimized for processing large matrices constructed from high-resolution Hi-C experiments. The method and theoretical result of the GLR test provide a general framework for significance testing of similar experimental chromatin interaction data that may not fully follow negative binomial distributions but rather more general mixture distributions.

Keywords: Hi-C data, Chromatin interaction, Hierarchical TADs, Change-points, Generalized likelihood-ratio test

Background

The eukaryotic genome is hierarchically organized in the nucleus, exhibiting well-maintained three-dimensional (3D) structures for its cellular functions. DNA and associated proteins constitute chromatin units, among which interactions are not random but precisely regulate transcription and replication during the cell cycle [1–3]. For example, the interactions between enhancers and their distal targeted genes are essential for

controlling gene expression strengths and tissue-specific expression patterns [4]. 3D chromosomal studies in prostate cancer, thalassemia, breast cancer and multiple myeloma have revealed that disordered interactions are closely related to gene dysregulation, contributing to the development of cancer and other genetic diseases [5–7]. Thus, estimating the 3D organization of chromosomes can provide important insight into not only the role of high-order chromatin compaction in gene regulation but also the way disordered chromatin interactions lead to diseases.

To systematically delineate chromatin interactions and 3D organization, novel experimental methods have been developed by employing high-throughput sequencing techniques. Chromosome conformation capture (3C) [8] and its high-throughput derivatives, such as ChIA-PET [9], HiChIP [10] and Hi-C [11, 12], have granted researchers comprehensive information on chromatin interactions and hierarchical chromosomal organizations, including active or repressive compartments (A/B compartments) [13], topologically associated domains (TADs) [11, 14, 15], CTCF protein-mediated loops [9] and enhancer-promoter interactions [4]. In general, a chromosome can be divided into active or repressed compartments (A/B compartments) corresponding to higher or lower gene expression levels [13]. The analysis of high-resolution Hi-C data has shown that chromosomes can be divided into functional units, called TADs, which are conserved across multiple human and mouse cell lines [11, 12, 16]. Furthermore, ChIA-PET data of CTCF, Cohesin and RNA PolII have revealed fine spatial structures of CTCF loops and enhancer-promoter interactions [4, 9]. Compared with ChIA-PET and other capture-based methods, Hi-C provides high-resolution unbiased signals of chromatin interactions [12].

TADs can be considered isolated structures that partition chromosomes into discrete functional regions and thus restrict regulatory activities within them [3, 11, 14]. To detect TAD structures from Hi-C data, many computational methods have been proposed by calculating the insulation scores or defining significance values of TAD boundaries. However, most of them are tools for detecting nonhierarchical TADs, such as Armatus [17], TopDom [18], HiCSeg [19], InsulationScore [20], Arrowhead [12] and DomainCaller [11]. Since TADs were shown to be hierarchically organized [11, 12], the estimated nonhierarchical TADs cannot fully describe the biological hierarchy in cell systems. As shown in Fig. 1a, a $\sim$ 3 Mb region on chr1 of the GM12878 cell line clearly exhibits four layers of TADs with different interaction strengths. To overcome the limitations of nonhierarchical TAD finders, another type of method, such as TADtree [21], IC-Finder [22], GMAP [23], Matryoshka [24] and 3DNetMod [25], has been proposed to find TADs and their nested sub-TAD organizations. Although these methods have given researchers new knowledge in understanding chromosomal organization, they still suffer from low precision or poor robustness against noise or high time consumption [26]. IC-Finder [22] employs a constrained hierarchical clustering strategy that iteratively groups objects into a hierarchy of clusters. Although it was robust against noise, it requires high sequencing depth [26]. Another method, GMAP [23], utilizes a Gaussian mixture model to iteratively identify TADs but is limited to two levels of TADs. TADtree [21] finds the best TAD hierarchy via a dynamic programming algorithm that was tested to be time consuming for large size Hi-C matrices [26]. 3DNetMod [25] uses network modularity theory to hierarchically cluster TADs; however, it is sensitive to multiple

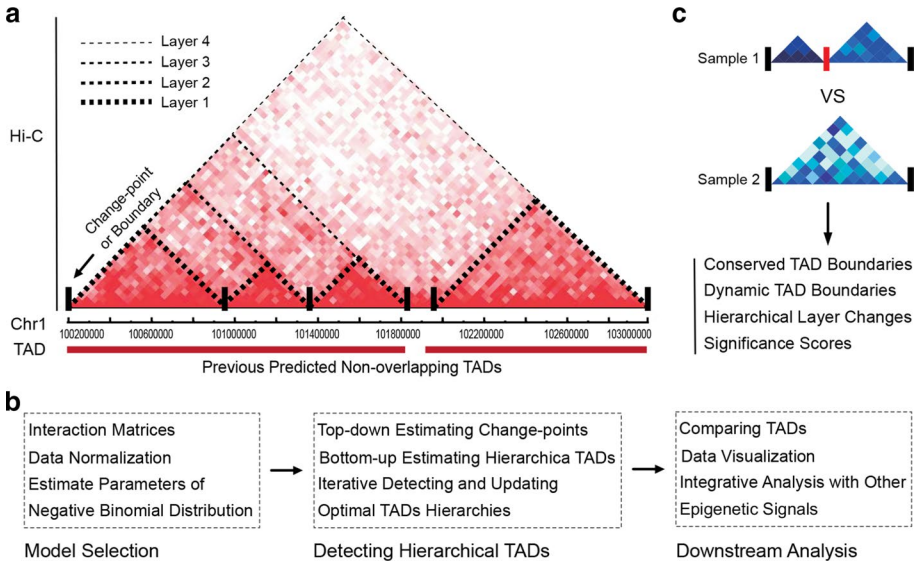

Fig. 1 HiCKey workflow. **a** Illustration of hierarchical organizations of chr1:100–103 Mb from in situ Hi-C data of GM12878. The potential four-layer organization is denoted by dotted lines. The non-overlapping TADs were obtained from a previous study [12]. **b** The workflow of HiCKey. HiCKey takes interaction matrices (or interaction list) as input, iteratively searches TADs, and outputs their hierarchies. It can also provide additional steps for TAD comparison and data visualization. **c** Comparing two sets of TADs structures

parameter settings and is less robust against experimental noise. One major challenge in detecting TAD structures is the experimental noise that mainly comes from random ligation of chromosomal segments during the cross-linking step and the "genomic distance effect" in Hi-C experiments, reducing the consistency and prevalence of higher-order structures [11, 12, 27]. Another obstacle in Hi-C data analysis is how interaction frequencies are distributed. Negative binomial (NB) distribution is the most widely used assumption, but it cannot fully capture the characteristics of chromatin interactions [12, 16] since confounding factors of Hi-C experiments may transform the interaction frequencies into more complicated distributions (e.g., a mixture of unknown discrete distributions). Thus, there are urgent requirements for new TAD detection methods to precisely estimate chromosome structure.

Understanding dynamic changes in TADs is also an important topic in Hi-C data analysis since disordered TADs are linked with cell-specific gene expression regulation or different developmental conditions. For example, Sauerwald and Kingsford et al. confirmed that conservation and dynamics of TAD boundaries were associated with distinct biological conditions or chromosomal variations by comparing a large number of Hi-C experiments of cell lines or tissues [28, 29]. Several methods have been proposed for detecting boundary changes in TADs, including HiCcompare [30], localTADSim [29], HOMER [31], HiCDB [32] and TADCompare [33]. The major strategy of these methods is to first detect TADs separately and then compare two sets of TAD boundaries. However, these methods usually require specific data types and lack statistical rigorousness. HOMER [31] only outputs different TAD regions by overlapping two sets of TADs but does not provide significance testing for boundary differences. Another method, TADCompare [33], has arisen as a potentially useful tool for comparing TAD boundaries.

This method proposes a new boundary score for differential boundary detection, time-course analysis of boundary changes, and consensus boundary calling but is limited to five types of boundary changes. LocalTADSim [29] requires using Armatus software or manually formatting their inputs as Armatus output. HiCDB [32] uses a new metric named relative local insulation that is similar to insulation score, but it is biased to top-ranked insulation scores.

Based on the above observations, we propose a novel computational method, called HiCKey, to decipher the hierarchical organization of chromatin interactions in Hi-C data (Fig. 1b). We derived a generalized likelihood-ratio test (GLR) for calling TAD boundaries (change-points), which is a matrix-variant change-point testing method in the literature. HiCKey can be applied to different interaction strength distributions. This is important for statistical analysis of Hi-C data, which is composed of biological interactions, random missing interactions and random ligation noise. Furthermore, the p values of a change-point from different Hi-C matrices can be combined by Fisher's method, providing a measure of whether a boundary is conserved across different samples (Fig. 1c). We demonstrated the performance and robustness of HiCKey using substantial validations of simulation studies. By applying HiCKey to seven human cell lines, we identified not only multiple layers of TAD organization in each cell line but also TAD structures consisting of different gene expression or histone modification signals. We found that TAD boundaries are significantly enriched in active chromosomal regions, indicating that fine TAD architectures are employed for precise gene transcription control. These results show the advantages of HiCKey in detecting TADs and provide novel biological discoveries revealing the association of chromosomal organization and gene regulation.

Methods

Modelling TADs organization

Hi-C experiments generate a symmetric n by n matrix $\mathbf{X} = (x_{ij}) \in \mathbb{R}^{n\times n}$, where n is the number of bins and x_{ij} is the frequency of chromatin interactions between a pair of genomic loci i and j. Assume there are K change-points located at $1 = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_K < n$. These change-points divide all chromosomal bins into $K + 1$ non-overlapping TADs, as shown in Fig. 2a. For change-points τ_a, τ_b and τ_c, the TAD between τ_a and τ_b is $A_{a,b}$. The rectangle between $A_{a,b}$ and $A_{b,c}$ is $R_{a,b,c}$ (R). We aim to detect (1) all the change-points and (2) the hierarchical organization of these change-points.

Previous Hi-C studies have revealed that within-TAD interactions are much stronger than cross-TAD interactions [11, 12, 16], as shown in a Hi-C matrix of GM12878 (Fig. 1a). In general, the average interactions in neighbouring blocks, $A_{0,1}$ and $A_{1,2}$, can be different, but they are both stronger than the interactions in R (Fig. 2a). Additionally, the interaction strength decreases as the distance from the diagonal increases. These biological observations raise statistical insight that a change-point, τ_1, will lead to significant distribution differences among $A_{0,1}$, $A_{1,2}$ and R.

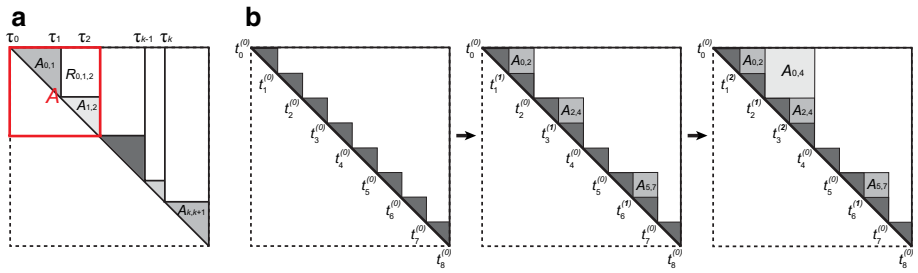

Fig. 2 Schematic plot of hierarchical TAD detection. **a** Block diagonal with K change-points and one change-point (red region). Sub-matrices are shown with a red region that has a single change-point. **b** The bottom-up procedure for detecting hierarchical TADs. At each iteration, the p values are iteratively calculated for all boundaries, and the layer labels of inner boundaries of two blocks are iteratively updated if they satisfy merge conditions

In theoretical statistics, this problem is known as change-point analysis. We note that various frequentist and Bayesian methods have been developed for multiple change-point analyses of uni- and multi-variate data over the last few decades. In particular, Bayesian methods assume that multiple change-points follow a stochastic process and solve the inference problem through Markov chain Monte Carlo (MCMC) simulations [34, 35]. Exact Bayesian inference approaches with efficient approximations were also developed for various problems [36–38]. The frequentist methods provide more technical tools, including dynamic programming algorithms to solve maximum likelihood estimation [39, 40], binary segmentation [41], information criteria model selection [42, 43], and penalized likelihood or cost function approaches [44, 45]. However, none of these methods can be directly applied to Hi-C data because they are matrix-variate. Simplifying them to multi-variate vectors and feeding them to existing change-point methods are not optimal. Furthermore, a general tool needs to be developed to address cases that are more complicated than some single distribution assumption (e.g., NB).

Generalized likelihood-ratio test for change-points

In Fig. 2a, the red block is a sub-matrix from τ_0 to τ_2 in matrix X. Its upper triangular part is denoted by A. Assume that all interaction reads are independent random variables from an NB family. If there exists a change-point, we then have three sets of block-wise constant parameters. Otherwise, all the parameters will be the same. Specifically,

$$x_{ij} \sim NB(\mu_k, r), \quad 1 \le i \le j \le n, \quad \mu_k = \begin{cases} \mu_1, & \text{if } (i,j) \in A_{0,1} \\ \mu_2, & \text{if } (i,j) \in A_{1,2} \\ \mu_0, & \text{if } (i,j) \in R \end{cases}$$

where μ_k is the mean of the NB distribution and r is a nuisance parameter with a positive value. In A, we consider the hypothesis test H_0: there is no change-point against H_1: there is one change-point at unknown position $\tau_1 = m (1 < m < n)$, such that $A = A_{0,1} \cup A_{1,2} \cup R$. Let S_A, S_{A_k} and S_R be the sums of x_{ij} in the corresponding regions.

$$S_A = \sum_{(i,j)\in A} x_{ij}, \quad S_R = \sum_{(i,j)\in R_{0,1,2}} x_{ij}, \quad S_{A_1} = \sum_{(i,j)\in A_{0,1}} x_{ij}, \quad S_{A_2} = \sum_{(i,j)\in A_{1,2}} x_{ij}$$

When the location of change-point $\tau_1 = m$ is known, the logarithm of the generalized likelihood-ratio test statistic is

$$\begin{aligned} GLR_{NB,m} = \sum_{k=1,2} \Big\{ & S_{A_k} \log\Big(\frac{S_{A_k}/|A_{k-1,k}|}{r + S_{A_k}/|A_{k-1,k}|}\Big) \\ & + r|A_{k-1,k}| \log\Big(\frac{r}{r + S_{A_k}/|A_{k-1,k}|}\Big)\Big\} \\ & + \Big(S_R \log\Big(\frac{S_R/|R|}{r + S_R/|R|}\Big) + r|R| \log\Big(\frac{r}{r + S_R/|R|}\Big)\Big) \\ & - \Big(S_A \log\Big(\frac{S_A/|A|}{r + S_A/|A|}\Big) + r|A| \log\Big(\frac{r}{r + S_A/|A|}\Big)\Big) \end{aligned}$$

where $|\cdot|$ is the cardinality of a set.

Theorem 1 *The GLR statistic, $GLR_{NB,m}$, is asymptotically equivalent to the following scan statistic if m/n holds constant as $n \to \infty$*

$$Z_m := \frac{1}{2\sigma_0^2} \left\{ \frac{\Big(S_{A_1} - \frac{|A_{0,1}|}{|A_{0,1}\cup R|} S_{A_1\cup R}\Big)^2}{|A_{0,1}|\Big(1 - \frac{|A_{0,1}|}{|A_{0,1}\cup R|}\Big)} + \frac{\Big(S_{A_1\cup R} - \frac{|A_{0,1}\cup R|}{|A|} S_A\Big)^2}{|A_{0,1}\cup R|\Big(1 - \frac{|A_{0,1}\cup R|}{|A|}\Big)} \right\} \tag{1}$$

where $S_{A_1\cup R} = S_{A_1} + S_R$, σ_0^2 is the variance under the null hypothesis, which can be estimated.

The first term on the right-hand side in (1) describes the difference between $A_{0,1}$ and R, while the second term describes the difference between $A_{0,1} \cup R$ and $A_{1,2}$. Note that if the change-point position is known, the partition ratio m/n of the matrix A is fixed. Hence, it is natural to assume that m/n holds constant as $n \to \infty$. It is easy to see that the asymptotic distribution of Z_m is chi-square. However, in practice, m is unknown, so we need to define a new test statistic and study its asymptotic properties.

Definition 1 We define the following test statistic:

$$\widetilde{Z} = \max_{\xi < m \le n-\xi} Z_m, \tag{2}$$

Z_m is calculated for each ($\xi < m \le n - \xi$), and we take the supremum, where ξ is the minimum size of a TAD.

The minimal TAD size is defined by default as the up integer of 100 kb/resolution. Here, 100 kb is estimated by the observed sub-TAD size in real biological datasets, since more than 95% of sub-TAD sizes are larger than 100 kb among multiple cell types of Rao's Hi-C data [12]. For example, for the interaction matrix with 40 k resolution, 3 (round 100/40 up to 3) is set as the minimal TAD size. Moreover, the p value threshold

and the minimal TAD size threshold are flexible in HiCKey software and can be changed by users for specific research goals.

As a result, we can eliminate the assumption on the original distribution. The above $\widetilde{Z}$ can be used to detect the change in means and is not limited to an NB distribution. Under the null hypothesis, we take the upper triangular part, A, as a discrete-time random-walk with a two-dimensional time index. By Donsker's invariance principle, we obtain the asymptotic distributions of Z_m and $\widetilde{Z}$ in the following theorem.

Theorem 2 *Consider a Gaussian random field $G(s, t)$, with location indices s and t, defined on the upper triangular part of a unit square $B = \{(s,t)|0 \le s \le t \le 1\}$. Assuming that $m/n \to t \in (0,1)$ as $n \to \infty$, then the regions $\frac{A_{0,1}}{\sqrt{n^2/2}}$, $\frac{A_{1,2}}{\sqrt{n^2/2}}$ and $\frac{R}{\sqrt{n^2/2}}$ converge to regions $\widetilde{A}_1 = \{(\widetilde{s},\widetilde{t})|0 \le \widetilde{s} \le \widetilde{t} \le t\}$, $\widetilde{A}_2 = \{(\widetilde{s},\widetilde{t})|t \le \widetilde{s} \le \widetilde{t} \le 1\}$, and $\widetilde{R} = B - \widetilde{A}_1 - \widetilde{A}_2$, respectively (note that $\frac{A}{\sqrt{n^2/2}} \to B$). Correspondingly,*

$$Z_m \to g_t := \frac{(G_{\widetilde{A}_1} - \frac{t}{2-t} G_{\widetilde{A}_1 \cup \widetilde{R}})^2}{2t^2(1-t)/(2-t)} + \frac{(G_{\widetilde{A}_1 \cup \widetilde{R}} - t(2-t)G_{\widetilde{A}})^2}{t(1-t)^2(2-t)},$$

and if $\xi/n \to \delta > 0$,

$$\widetilde{Z} \to g_\delta := \max_{\delta < t < 1-\delta} g_t.$$

Details of the Gaussian random field construction and proof are included in Additional file 1. This asymptotic property is extremely helpful in high-resolution Hi-C data. Consider a TAD with a fixed chromosomal size (typically 1 Mb), where the higher the resolution is, the more reads TAD contains in the Hi-C matrix. In practice, Monte Carlo simulations can be used to obtain the asymptotic distribution from the above theorem. A histogram and a kernel density estimation are included in Additional file 1: Fig. S1. Since our GLR testing theoretically converges for different types of distributions, the parameters can be applied to different datasets.

Detecting hierarchical TADs

We propose an iterative algorithm to implement the GLR test in estimating hierarchical TADs. The first step is binary segmentation to identify all the change-points (TAD boundaries). In the Hi-C matrix, we first find one change-point that has the maximum Z_m in Eq. (2), resulting in two diagonal sub-matrices. Iteratively, one change-point is found for each sub-matrix, until the sub-matrix has a size smaller than the lower bound, 2ξ. In the second step, we use a pruning process to test each change-point in reverse order to which they are identified and to remove insignificant change-points. A p-value threshold, α_0, is needed for all the tests.

(binary segmentation)

while there is a diagonal block, A, with size $T \geq 2\xi$

  for i from $\xi + 1$ to $T - \xi + 1$ in A

    calculate Z_i

  find a change-point at $\arg\max Z_i$

  record the order we identify that change-point

(pruning)

for each change-point, τ_t, in reverse order to which they are identified

  take the sub-matrix from its left closest τ_{t-1} to right closest τ_{t+1} and

  calculate the GLR test statistic $\widetilde{Z} = Z_t$

  If p-value $> \alpha_0$

    eliminate the change-point

Although binary segmentation in a top-down strategy can provide hierarchical organization as divisive clustering or a decision tree, it may contain more false hierarchical structures. We use a bottom-up procedure to merge neighbouring blocks and update the layer labels of boundaries (please see Fig. 2b for a demo example). More specifically, we recalculate the p value for each potential boundary outputted in the top-down step by testing its flanking blocks with the attached rectangle sub-matrix. The p values of all boundaries are ranked in descending order, and their layer labels are initialized as zero. In descending order, two neighbouring blocks are parallelly merged into one if their inner boundary p value is larger than a threshold $\alpha_1 = 1e{-}5$. The layer label of the inner boundary of merged blocks increases by one. In each iteration, the boundary p values between merged blocks are recalculated. The iteration continues until no remaining blocks satisfy the merge conditions. Here, α_0 and α_1 are used to control the number of potential TAD boundaries and the number of hierarchical branches, respectively. In HiCKey, $\alpha_1 = 1e{-}5$ was used by default, which can well-delineate local hierarchical structures in real data analysis. These two parameters can be reset by users for different TAD detection goals. If α_1 increases to α_0, more blocks are considered as individual TADs (less hierarchical). In contrast, if α_1 is smaller, more blocks are grouped into hierarchical structures.

Recalculate the p-values for all boundaries by using two neighbouring blocks and their attached rectangle sub-matrix.

While there are p-values of boundaries larger than $\alpha_1 = 1e-5$.

for each boundary t_m in descending order

if its p-value $> \alpha_1$ and its two neighbouring blocks have not been merged at the current step, merge two neighbouring blocks as one and update the layer label of their inner boundary by one.

for each new boundary t_m

recalculate the GLR test and obtain the p-value

rank p-values in descending order.

Validating performance by simulation studies

Hi-C data have random interactions generated in random ligation of DNA segments [12, 16]. Polymer models show a decrease in the random interaction score as the distance between two loci increases. Because there is no true answer for TAD boundaries in real Hi-C datasets for validation, we first tested HiCKey on two simulation datasets that were originally created for assessing several computational methods by Forcato [46]. These datasets were simulated by a quasi-negative-binomial generator modified from Lun [47], with each y_{ij} specifically designed to approximate real Hi-C data well. Using the same datasets facilitates the comparison between HiCKey and other methods. The specific datasets we used are as follows:

- [(Sim1)] *Matrices without nested TADs*. It consists of 20 simulated Hi-C matrices with noise levels of 4%, 8%, 12% and 16%. These matrices contain no nested TAD structure, and each matrix has a size of approximately 4500 with 171 diagonal blocks.
- [(Sim2)] *Matrices with nested TADs*. It consists of 20 simulated Hi-C matrices with noise levels of 4%, 8%, 12% and 16%. These matrices differ from Sim1 in that they contain nested TADs. In particular, each matrix has a size of approximately 4500 and contains 910 diagonal blocks with three layers of hierarchical structure.

A previous study [47] reported that the noise level, which they refer to as the biological coefficient of variation, varies between 0 and 16%.

The performance was evaluated by four measures. First, the true positive rate (TPR) was defined as the number of detected true boundaries divided by the number of total true boundaries. Second, the false discovery rate (FDR) was defined as the number of falsely detected boundaries divided by the total number of detected boundaries. Third, the difference between the estimated and true number of TAD boundaries was defined as $\widehat{K} - K$. If there were several matrices in a simulation dataset, we calculated the average of all $\widehat{K} - K$ of the matrices. Fourth, to evaluate the consistency between true hierarchical TAD structures and HiCKey TADs, we calculated the Fowlkes–Mallows index (B_k) [48], where k is the hierarchical level.

Since there are three levels of hierarchical structures embedded in the Sim2 dataset, B_1, B_2 and B_3 were calculated for each hierarchical level from the bottom layer B_1 to the outer layer B_3. It was noted that the B_k index lies between 0 and 1. If two partitions were perfectly matched, then $B_k = 1$. For each noise level, we calculated the average score of B_k for 1000 random initializations. To obtain the Fowlkes–Mallows indices under the null hypothesis that the two clusterings are unrelated, we calculated control B_k between the true hierarchical structures and randomly relabelled HiCKey TADs (relabelling) by using the Fowlkes and Mallows formula [48].

Validating the robustness of HiCKey by simulation studies

To evaluate the robustness of HiCKey against different initial boundaries (change-points), we performed validations on simulated datasets Sim1 and Sim2 and real datasets of hESC and IMR90 cell lines. For each dataset, we first ran HiCKey ordinarily and recorded the number of detected change-points as well as their locations. Then, we ran HiCKey 1000 times with random selection of the first change-point in the whole matrix. We evaluated the result consistency of randomly starting and the ordinary run by using the criteria TPR and $\widehat{K} - K$.

To test the performance of HiCKey on datasets with different distributions, we generated simulation matrices whose entries followed Gaussian and NB distributions. We considered the following two scenarios for the Hi-C matrix of size 500×500:

- [(Sim3)] *Gaussian distribution.* Let $K = 31$ (change-point numbers), and their locations were uniformly drawn with the smallest block size that was larger than 4. We set the mean of each element, u_{ij}, as $u_{ij} = \mu_k \sim \text{Gamma}(4, 18)$ for $(i,j) \in A_{k-1,k}$, $k = 1, \ldots, 31$, and $\mu_0 = 0$ for $(i,j) \in A - \cup_{k=1}^{31} A_{k-1,k}$, where the numbers were estimated by real Hi-C data. The values of x_{ij} were generated by

$$x_{ij} \sim \begin{cases} N(u_{ij}, \sigma^2) & \text{for } (i,j) \in A_{k-1,k}, \\ \max\{N(0, \sigma^2), 0\} & \text{o.w.} \end{cases}$$

- [(Sim4)] *Poisson and negative binomial distributions.* Let $K = 31$. Change-point locations and element means μ_k were similarly generated as (Sim3). Furthermore, x_{ij} was generated by an NB model [47]

$$x_{ij} \sim \begin{cases} NB(\nu^{-1}, (1 + \nu\mu_k)^{-1}) & \text{for } (i,j) \in A_{k-1,k}, \\ 0.5\mathbf{1}_{\{0\}} + 0.5NB(\nu^{-1}, (1 + \nu \cdot \min_k\{\mu_k\})^{-1}) & \text{o.w.} \end{cases}$$

  Note that x_{ij} in the complementary region $A - \cup_{k=1}^{K+1} A_{k-1,k}$ follows a mixture of point mass and NB distribution.

$NB(\nu^{-1}, (1 + \nu\mu_k)^{-1})$ provides an NB distribution with mean μ_k and variance $\mu_k + \nu\mu_k^2$. The parameter $\sqrt{\nu}$, referred to as the *biological coefficient of variation* (BCV) [47], varies from 0 to 16%. Hence, we set $\sqrt{\nu} = 0, 0.05, 0.10, 0.15$ in (Sim4). Note that $\nu = 0$ corresponds to the case in which x_{ij} follows a Poisson distribution with mean μ_k. In (Sim3), σ^2 was specified as $\sigma^2 \approx 72 + 72^2 \cdot \nu$, where 72 was the mean of Gamma(4, 18).

Hi-C datasets for real data case studies

High-resolution in situ Hi-C data of seven cell lines produced by Rao [12] were downloaded from the Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/) with the accession number GSE63525. We applied HiCKey on 25 kb resolution Hi-C data for all seven cell lines, which included GM12878, HMECs, HUVECs, IMR90 cells, K562 cells, KBM7 cells, and NHEKs. In addition, we downloaded their predicted TADs using the Arrowhead method [12] and denoted them by Rao-TAD in the rest of the paper. We also downloaded early Hi-C data generated by Dixon [11] for two human cell lines, HESC and IMR90, whose resolutions are 40 kb. Histone modifications and TF binding peaks were obtained from ENCODE and the Roadmap epigenomics project using the WashU genome browser [49].

Comparing TADs across samples

Since HiCKey outputs p values of TAD boundaries, they can be extended to compare boundary differences across cell lines. For a boundary, m, we calculated its p values, $p_1(m)$ and $p_2(m)$, in two different samples of Hi-C matrices. Assuming two Hi-C experiments are independent, we used Fisher's method [50] to combine two p values into one test statistic $\chi_4^2 := -2\ln(p_1(m)) - 2\ln(p_2(m))$. Here, χ_4^2 follows a chi-squared distribution of 4 degrees of freedom, and its p value is denoted by p_f. The p value $p_f(m)$ will decrease if $p_1(m)$ and/or $p_2(m)$ decrease.

Memory and running time optimization

Recent in situ Hi-C experiments have generated datasets with a resolution as high as 1 kb, resulting in large matrices. Therefore, it is essential to optimize memory usage and running time. Some Hi-C data processing pipelines, such as HiC-Pro [51], place much emphasis on memory efficiency. Many existing TAD detection methods use a matrix as input; however, this is infeasible for high-resolution data. For example, storing a single Hi-C matrix of 5 kb resolution under double precision might require 18 G memory. We used several strategies for computing resource optimization.

First, high-resolution Hi-C matrices are sparse, as most elements are zero. Our program can read both matrix or list forms (Rao's data consist of non-zero elements and their indices [12]). HiCKey stores only non-zero elements in the upper triangular part of the Hi-C matrix. Second, the top-down binary segmentation is the most time-consuming step of HiCKey. To calculate all Z_m in $\mathbf{X} = (x_{ij}) \in \mathbb{R}^{n\times n}$, we first calculate the sums of every row and column. As m moves from 1 to n, only linear arithmetic operations are needed.

Second, for the best case in which each change-point is allocated in the middle of matrix A, we need at most $log_2 n$ iterations with operations $O(n^2)$. In the worst case, each iteration generates one sub-matrix as small as possible and the other as large as possible. This results in at most n/ξ iterations with operations $O(n^3)$ (see detailed calculations in Additional file 1). All the tests were conducted on a regular laptop with an Intel(R) Core(TM) i5-7200U CPU with a 2.50 GHz processor and 12 GB memory.

Results

Performance of HiCKey in detecting TADs

To understand the performance of HiCKey in detecting TADs, we first tested it on large-scale simulated Hi-C matrices [46], which included two types of data, Sim1 without nested TAD structures and Sim2 with nested TAD structures. For both Sim1 and Sim2, four noise levels of 4%, 8%, 12% and 16% were added to test the robustness of HiCKey against the random collision noise of interactions. Testing at different noise levels is critical since large numbers of random collision interactions are observed in HiC data [12, 16]. First, on dataset Sim1, HiCKey achieved a high TPR of 0.9988 under a 4% noise level (Fig. 3. Additional file 1: Table S1). As the noise level increased from 4 to 16% (four-fold change), the TPR decreased to 0.9459 (0.947-fold change). The fold change ratio of TPR and noise was 0.24 (0.947/4), indicating that the TPR of HiCKey was robust against noise changes. When the noise level increased from 4 to 8% and 12%, the FDR slightly increased from 0 to 0.0173 and 0.1176, respectively. Additionally, the FDR increased to 0.3618 at the 16% noise level. We also compared the number of TADs estimated by HiCKey with the true value. We found that HiCKey produced a very accurate average number of TADs at the 4% noise level ($\widehat{K} - K = -0.2$). As the noise level increased, the estimated number of TADs increased ($\widehat{K} - K$ as 1.8, 19.4 and 82.6 for noise levels 8%, 12% and 16%, respectively). We then tested HiCKey on dataset Sim2. At a 4% noise level,

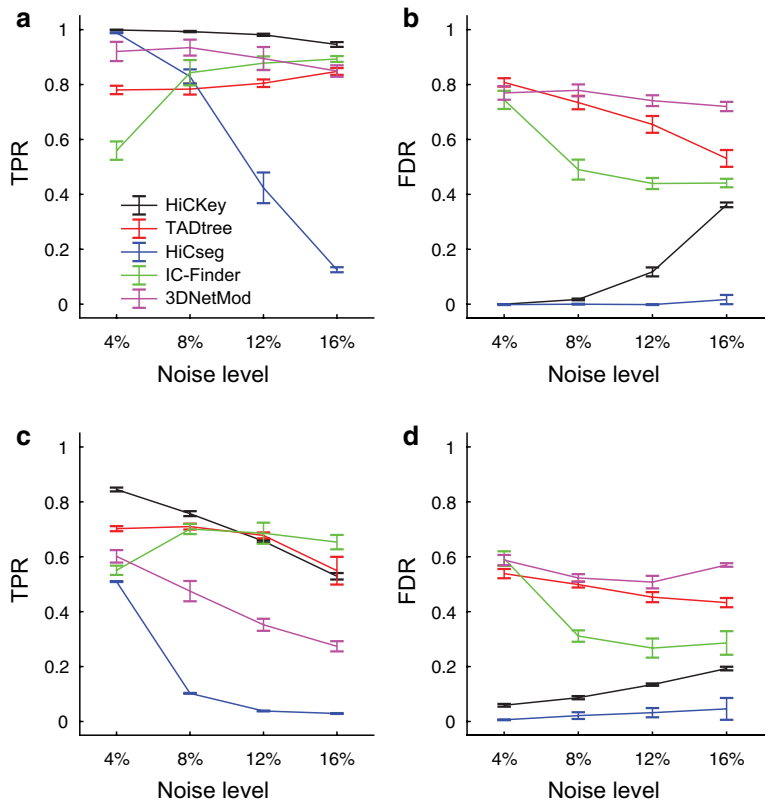

Fig. 3 The performance of four methods. **a** TPR of Sim1 dataset. **b** FDR of the Sim1 dataset. **c** TPR of the Sim2 dataset. **d** FDR of the Sim2 dataset

HiCKey achieved 0.845 TPR and 0.059 FDR (Additional file 1: Table S2). When the noise level increased to 8% and 12%, the TPR decreased to 0.757 and 0.6568, respectively. We noticed that TPR and noise level were linearly correlated ($R^2 = 0.9926$), demonstrating that HiCKey can remain stable with these noise changes. In summary, these validation results at different noise levels suggest that HiCKey is robust against random collision noise, especially for noise levels ranging from 0 to 12%.

We calculated the Fowlkes–Mallows index $B_k (k = 1, 2, 3)$ [48] to evaluate the consistency between true hierarchical TAD structures and HiCKey TADs, as there are three layers of hierarchical structures embedded in the Sim2 dataset. The B_k index lies between 0 and 1, where larger index scores indicate higher similarities among two compared hierarchical structures. Overall, we found that the B_k indices were larger than 0.8196 for all three hierarchical layers at the 4% noise level (Additional file 1: Table S3). When noise levels increased to 16%, they decreased slightly but maintained fair scores all larger than 0.6887. At four noise levels, the control Fowlkes–Mallows indices were no more than 0.01 after recalculating B_k between the true hierarchical structures and randomly relabelled HiCKey TADs (Additional file 1: Table S3). We also found that the B_2 and B_3 indices were more stable than B_1 when noise levels increased, suggesting that HiCKey is more robust against noise in detecting second and third levels of TAD structures.

We examined practical running time and memory usage on a real Hi-C matrix of chr1 (the largest among 23 chromosomes) in the GM12878 cell line [12]. Under resolutions of 50 kb, 25 kb, 10 kb and 5 kb, the running times of HiCKey were 23 s, 53 s, 106 s and 157 s, respectively. Additionally, the practical running time was approximately $O(n^2)$ (Additional file 1: Fig. S2a). Its memory usage was 170 Mb, 389 Mb, 768 Mb and 980 Mb, respectively, and was also approximately $O(n^2)$ (Additional file 1: Fig. S2b). Taken together, HiCKey requires reasonable computing resources in processing high-resolution Hi-C matrices.

Robustness against different distributions and initialization

In Hi-C data analysis, the read counts of interactions were usually assumed to follow an NB distribution [12, 52, 53] or Poisson distribution [54] or normalized data [30]. However, these distribution models cannot fully capture the characteristics of chromatin interactions in Hi-C experiments due to the divergent confounding factors observed in real biological systems, which results in a complicated mixed model [8, 27, 46]. In HiCKey, we derived a GLR test that can be broadly used for multiple distributions but is not limited to the NB distribution. To test the performance of HiCKey on different distributions, we simulated 1000 interaction matrices with normal and NB distributions (see details in the Methods section). At four different noise levels, we found that the TPRs were all larger than 0.99, while the FDRs were less than 0.0055 (Additional file 1: Table S4), suggesting that HiCKey is robust against different distributions.

To test whether HiCKey is sensitive to the initial choice of change-point allocation, we constructed new validations by randomly selecting the initial location. We performed validations on simulation datasets Sim1 (Additional file 1: Table S5) and Sim2 (Additional file 1: Table S6) and real datasets of hESCs (Additional file 1: Table S7) and IMR90 cell lines (Additional file 1: Table S8). Regarding these validations, TPR rates were all

larger than 90%, while the means of $\hat{K} - K$ were very small, indicating that HiCKey is robust against the initial boundary selection.

Comparisons with other methods

We compared HiCKey with four popular methods, HiCSeg [19], TADtree [21], IC-Finder [22] and 3DNetMod [25], on simulation datasets Sim1 and Sim2. For Sim1, HiCKey achieved not only higher TPRs at four noise levels but also slow declines (Fig. 3a and Additional file 1: Table S1). The FDR of HiCKey at 16% noise was smaller than those of IC-Finder, TADtree and 3DNetMod (Fig. 3b). We found that HiCSeg tended to retrieve large TADs, resulting in fewer detected TAD boundaries. For example, HiC-Seg's $\widehat{K} - K$ were -2.2, -29.2, -99.2 and -150.2 at noise levels of 4%, 8%, 12% and 16%, respectively, explaining why their FDRs were always lower but TPRs decreased sharply. TADtree and IC-Finder outputted more TADs than the true value at the 4% noise level ($\widehat{K} - K = 397.6$ and 207.4, respectively), but the number of falsely identified TADs decreased with increasing noise levels. 3DNetMod tended to output more TADs than the true numbers at four noise levels and thus had higher FDR rates. For Sim2, HiCKey achieved the highest TPRs at noise levels of 4% and 8% but slightly dropped below TADtree and IC-Finder at noise levels of 12% and 16% (Fig. 3c and Additional file 1: Table S2). However, TADtree, 3DNetMod and IC-Finder suffered from much higher FDRs (Fig. 3d). Taken together, HiCKey achieved good performance, especially for lower noise levels.

We specifically investigated two regions with hierarchical TADs in dataset Sim2 that were used for comparative analysis in a previous study [46] (Fig. 4). The results showed that HiCKey, TADtree and 3DNetMod can detect single and hierarchical TADs for both cases, while IC-Finder and HiCSeg can only output bottom single TADs. At E1–E6, we found that HiCKey can correctly detect not only all the TAD boundaries but also their hierarchical organization, while TADtree and 3DNetMod made a few false predictions (E1, E3, E6) or missed outputting bottom TADs (E2, E4, E9). Region E7, consisting of complicated hierarchical structures, was a challenge for HiCKey and TADtree. Although HiCKey correctly detected all the boundaries in E7, it missed the outer layer. TADtree missed several boundaries and wrongly merged a neighbouring TAD into the block. 3DNetMod showed good hierarchical details at E7 and reported a large hierarchical TAD at E6, which seems to have no clear interaction blocks. Overall, these genome-wide analyses and detailed examples showed that HiCKey has good performance in detecting TAD boundaries and their hierarchical organization.

Hierarchical architecture of chromosomal organization

We applied HiCKey to in situ Hi-C data of seven cell lines. HiCKey successfully outputted the boundary positions, p values and hierarchical levels. Following Weinreb and Raphael's method [21], we defined the root TAD as order 1. If a root TAD had sub-level TADs, the two sub-TADs were of order 2. Similarly, sub-sub-TADs were of order 3, and so on. This is the same as how we defined the hierarchical order of TADs in the Methods section. In total, we detected 8586, 8200, 8903, 9043, 8801, 6246 and 7726 TADs in the GM12878, hMEC, HUVEC, IMR90, K562, KBM7 and NHEK cell lines, respectively (Additional file 2). In each cell line, we compared our results with

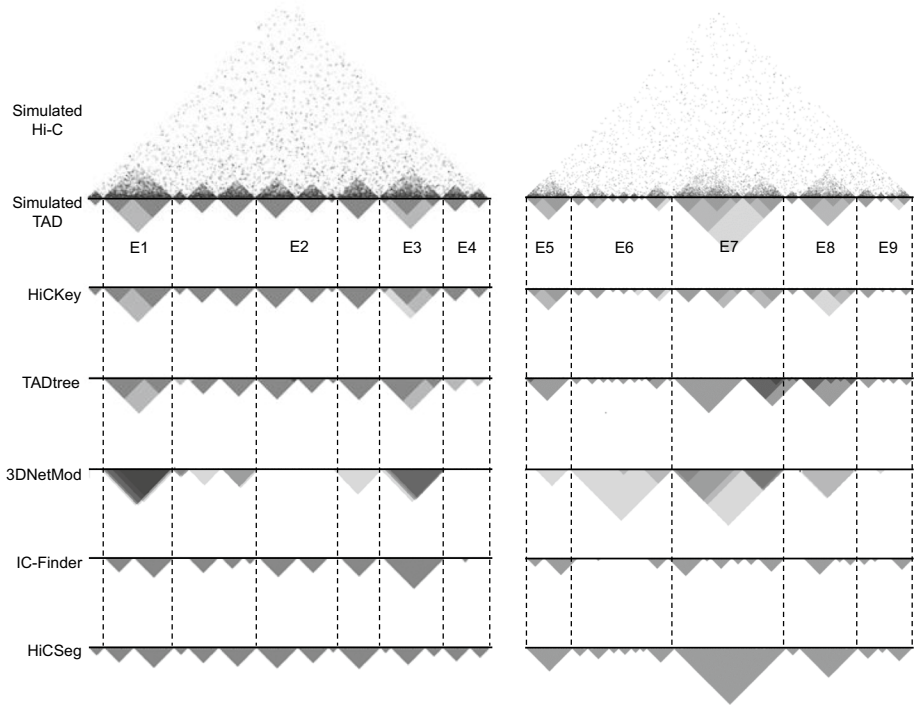

Fig. 4 True and estimated hierarchical structures of two samples in the Sim2 dataset. **a** Simulated case1, chr: 75600000–82720000. **b** Simulated case2, chr: 127000000–137000000

Rao-TAD. First, we found that their allocations of TADs in the 23 chromosomes were similar (p-values > 0.23 in all seven cells, two sides chi-square test). Second, we considered that a Rao-TAD boundary was matched if there was a HiCKey boundary located within its 2-bin (50 kb) distance. The proportions of matches between Rao-TADs and HiCKey TADs were 48.90%, 65.56%, 58.71%, 57.98%, 52.22%, 51.14% and 57.90% for GM12878, hMECs, HUVECs, IMR90, K562, KBM7 and NHEKs, respectively. Furthermore, we used the hypergeometric test (all the chromosomal bins were taken as population, boundaries of Rao-TAD as successes, HiCKey boundaries as samples, and the matched ones as sampled successes) to measure how significantly HiCKey TAD boundaries matched with Rao-TAD boundaries. On seven cell lines, the p values were all calculated as less than 1.0e−10, indicating that they were significantly matched.

Multiple levels of TADs were detected in each cell line. For example, Fig. 5 demonstrates our estimation of hierarchical TADs and Rao-TAD in a local region of chr3: 10700000–11300000 of GM12878. Overall, HiCKey TAD estimations exhibited more hierarchical layers than Rao-TADs. There was no Rao-TAD within a large sub-region (chromosome 3:10700000–11300000); however, clear blocks were observed from the Hi-C interaction heatmap. Extending the analysis to the genome-wide level, we found that although the highest order of TADs can reach 7, most of them exhibited an order 1 or 2 (97.86% $\pm$ 0.71%, Additional file 1: Table S9), suggesting that hierarchical TADs are enriched in certain chromosomal regions.

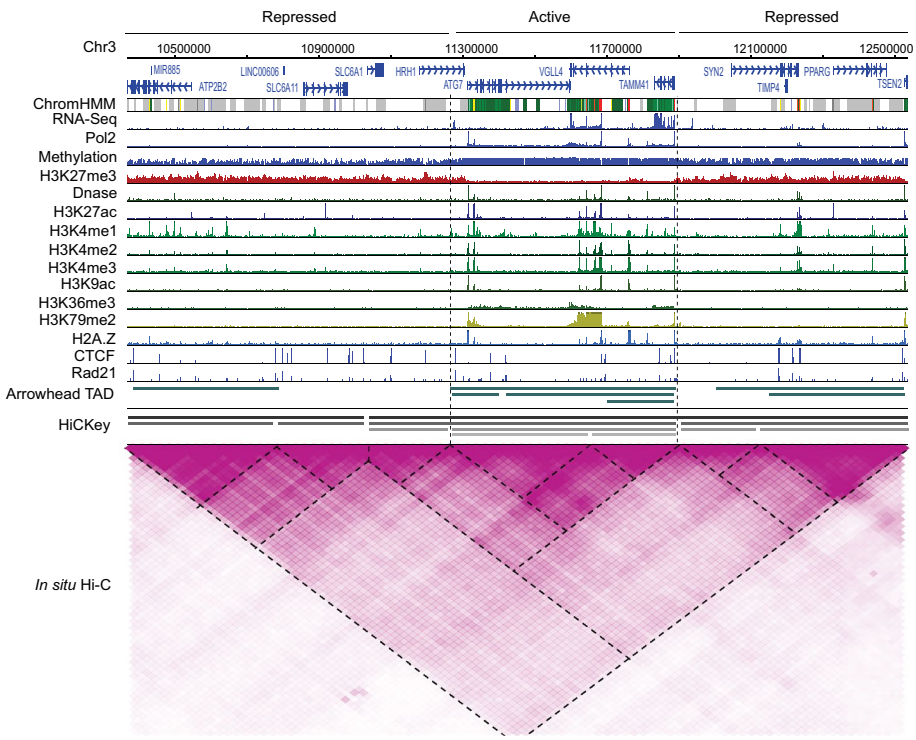

Fig. 5 Integrative analysis of TAD and histone modifications. A 2 Mb chr3 region of the GM12878 cell line shows one active region surrounded by two repressive regions. All signals are normalized by the 90-quantile value within this region. The tracks were generated by the WashU genome browser http://epigenomegateway.wustl.edu/legacy/

Hierarchical organizations are enriched in active regions compared with repressive regions

To test whether biased hierarchical organizations at different chromosomal regions are related to different biological insights, we performed an integrative analysis of TAD structures and epigenetic markers. Previous association analysis revealed that neighbouring TADs usually have different histone modification patterns [55] and that TAD boundaries are primarily associated with CTCF and Rad21 binding peaks [11, 12]. Here, we downloaded several histone signals and protein binding peaks for GM12878. We examined a $\sim$ 2 Mb region (chr3: 10700000–11300000, Fig. 5) that was partitioned into three parts, an active region flanked by two repressed regions. The active region was exhibited by biological signals, such as high RNA-seq signals of genes, Pol2 binding peaks, and multiple histone modifications (e.g., H3K4me3 and H3K27ac). We observed that the active region contained more hierarchical TADs than the repressed regions (Fig. 5, HiCKey track).

We next examined genome-wide TAD boundary enrichment in active chromosomal regions. To search genome-wide, we used the active/repressive annotations of chromosomal regions for six cell lines [55] and compared the numbers and layers of boundaries between them. First, we confirmed that among all six cell lines, TAD boundaries were enriched in active regions compared with repressive regions (p-value < 0.01, one-sided Fisher exact test, Additional file 1: Table S10). For instance, among the estimated 8200 active/repressive TAD boundaries in the IMR90 cell line, 5494 were located in active

regions (64,220 bins of 40 kb), while only 2706 were in repressive regions (46,400 bins of 40 kb). Next, we checked the layer annotation of boundaries. By comparing the layer number distributions, we found that they were significantly different, as the average layer number in active regions was larger than that in repressive regions (p-value < 0.01, K–S test). Thus, these genome-wide analysis results demonstrated that active chromosomal regions usually contain more TAD boundaries and richer TAD structures, indicating that active regions may employ more precise spatial organizations to regulate gene expression.

Detecting conserved and dynamic TAD boundaries between cells

Conserved and dynamic boundaries, as a result of cell-specific gene expression organization/regulation, different developmental conditions, or chromosomal variants in diseases, can be compared by Hi-C data of two different samples [7, 16]. Here, we examined the 8801 TAD boundaries of the K562 myelogenous leukaemia cell line and 9043 TAD boundaries of the IMR90 normal human fibroblast cell line detected by HiCKey. First, we found that 7286 boundaries were co-localized within a 2-bin distance. These high proportions of matched boundaries in K562 (82.79%) and IMR90 (80.57%) cells are consistent with early observations that TADs are conserved among mammalian cells [11, 12, 16]. Second, among the co-localized boundaries, 7280 of them have Fisher's combined p values $p_f < 0.01$, indicating that they are conserved in both cell lines. We also detected 1621 K562 TAD boundaries changed in IMR90, and 1763 IMR90 TAD boundaries changed in the K562 cell line, providing potential candidates for TAD boundaries that may be involved in cell-specific regulation.

To demonstrate dynamic changes in TADs and their potential biological functions related to gene transcription, we investigated a large chromosomal region ($\sim$ 4.2 Mb, chr2:38900000–43100000) of the K562 and IMR90 cell lines as representative. In this region, three large TADs were estimated in both K562 and IMR90, but a TAD (chr2:40000000–42050000) showed novel hierarchical sub-TAD structures in IMR90 but not in K562 (Fig. 6). Moreover, two sub-level TADs (chr2:40000000–41000000 and chr2:41000000–42050000) were observed in IMR90 with clear Hi-C interaction patterns. In contrast, their interactions were uniform in K562 cells. Furthermore, two smaller blocks (chr2:40000000–40700000 and chr2:40700000–41000000) were detected in the sub-TAD (chr2:40000000–41000000) of IMR90. The smaller block (chr2:40000000–40700000) included the protein-coding gene SLC8A1 and the lncRNA gene SLC8A1-AS1. SLC8A1 (solute carrier family 8 member A1) is a protein-coding gene linked to multiple diseases, such as long Qt syndrome 9, cardiac diseases and aromatase deficiency [56–58]. We found that this small block was active in IMR90 but not in K562 by several signals. First, SLC8A1 was highly expressed in IMR90 cells but not in K562 cells. This was consistent with the novel binding signals of Pol2, H2A. Z, CTCF and Rd21 in IMR90. In addition, novel histone modifications were observed in IMR90 but not in K562. Another small block (chr2:40700000–41000000) was at the 5' region of the SLC8A1 gene, but no genes were included. It also contained novel signals of CTCF and Rad21, suggesting that its generation may be mediated by CTCF and Rad21 to regulate SLC8A1 expression. Overall, these results demonstrate that HiCKey can detect not only TAD hierarchies but also their difference across samples. The hierarchical TAD

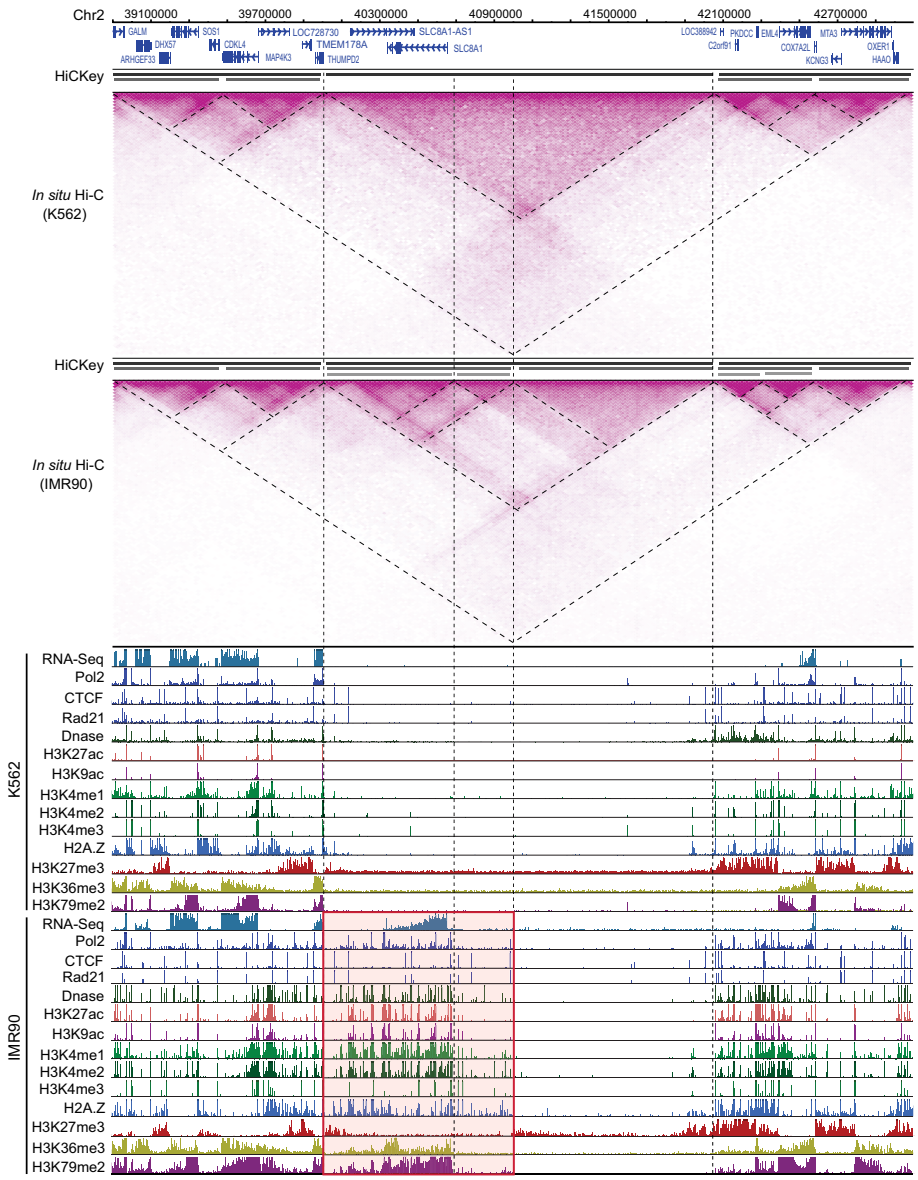

Fig. 6 Different TAD structures of chr2:38900000–43100000 in K562 and IMR90. A chr2:40000000–41000000 region showed hierarchical TADs in IMR90 but not in K562. ChIP-seq signals are normalized by the 90-quantile value within this region. The tracks were generated by the WashU genome browser http://epigenomegateway.wustl.edu/legacy/

structures in IMR90 further support our findings that active regions of chromosomes usually contain more and richer TAD structures for regulating gene expression.

Discussion

The identification of TADs and their hierarchical structures is extremely important in the study of chromatin interactions. We developed a novel GLR test to detect change-points in Hi-C matrices and studied its asymptotic properties. Based on the GLR test, we introduced HiCKey to decipher the hierarchical structure of TADs. The performance of HiCKey is endorsed by extensive simulation and real data analysis. The retrieved

hierarchical TADs are consistent with diverse biological signals, including histone modification and ChIP-seq data. We further found much more detailed TAD structures in active chromosomal regions. Comparative analysis of TADs across various cell lines revealed that different TAD organizations harbour disease-related genes, providing insights into how disordered interactions are linked to different cancer types.

Considerations for different distributions of chromatin interactions

To explain chromatin interaction strength decay, polymer models propose that the average pairwise contact probability decreases and asymptotically follows a power law with a given contour distance [59–61]. NB is another popular model used not only in the simulation of Hi-C data but also in other types of interaction data, such as ChIA-PET, CAPTURE-seq and HiChIP. Although such methods can approximately explain the interacting decay along the off-diagonal of the Hi-C matrix, they are challenged by several confounding factors in real biological systems. First, most promoter-enhancer interactions are usually regulated by divergent TF proteins or structural proteins, such as Pol2, GATA1/2, CTCF and Rad21. These deterministic factors have been evolutionarily fixed in different cell lines, thereby reducing the randomness of chromatin interactions that is captured in Hi-C experiments. In fact, most promoter-enhancer interactions are observed over short distances (tens of kb) from gene promoters to their neighbouring enhancers. Second, different experimental factors, such as the cross-linking of chromatin, chromatin digestion, and streptavidin pull-down of biotinylated ligations, may also affect the signal/noise ratio of Hi-C data [62]. Existing distribution models, for example, the NB distribution, cannot fully capture the characteristics of chromatin interactions in Hi-C experiments [60, 63]. In HiCKey, we derived a GLR test that can be broadly used for multiple distributions and is not limited to NB. This could potentially be very useful for complex Hi-C or ChIP-PET or HiChIP experiments in which interaction strengths may follow some unknown distribution.

Phase transition, loop extrusion and chromosomal hierarchies

Although Hi-C data provide a landscape of interacting strengths among chromosomes, mechanistic explanations of how chromatin interactions are dynamically formatted and regulated are lacking. At large scales (e.g., A/B compartments and TADs), it seems that interactions can be self-organized by levels of epigenetic modifications that are formalized as the phase transition in the neighbourhood of typical physiological conditions [61, 64, 65]. To further consolidate this phase-transition model, it will be highly desirable to manipulate in vivo histone-tail modifications for comparative Hi-C analysis. At smaller scales and active chromosomal regions, the loop extrusion model is proposed to link the CTCF and cohesin proteins into the formation of local looping structures [66–68]. The loop extrusion model suggests that structural maintenance of chromosome proteins (cohesin or condensin) progressively extrudes chromatin until it is blocked by CTCF bound to properly oriented site pairs [69–71]. In our comparative studies of K562 and IMR90 cell lines, we found that CTCF and Rad21 binding sites were remarkably changed within the newly established TADs (Fig. 6), indicating that CTCF-loop domains and enhancer-promoter interactions may be established via an extrusion process involving cohesin and CTCF. In summary, these observations and studies suggest that the

phase-transition model and loop extrusion model take place at different scales and chromatin states.

Integrating other Omics data

With the hierarchical organization of chromatin interactions available, we can deeply investigate the biological functions or principles of how histone modifications are coordinately used, as well as how gene expression is dynamically regulated. Since histone modifications, TF binding sites, and gene expression have been collected for hundreds of cell lines in ENCODE and Roadmap Epigenomics projects, we integrated them for locus-specific and genome-wide analysis. As shown in Fig. 5, we demonstrated that TADs estimated by HiCKey are consistent with histone modifications, including H3K27ac, H3Kme1/2/3 and many others. Gene expression from RNA-seq and TF binding peaks (Pol2, CTCF and H2A. Z) also confirmed the active and repressive compartment of chromosomal regions. By using multiple omics signals, we further observed different signal patterns in a comparative analysis of K562 and IMR90 (Fig. 6). We found that the newly expressed genes at active regions in IMR90 were occupied by multiple active histone signals. The new binding peaks of CTCF and Rad21 suggest that these structural proteins may be involved in the local chromosomal conformation for SLC8A1 gene expression in IMR90. Both examples indicate that integrative analysis of multiple omics data and hierarchical organizations is a promising method to fully understand chromosomal compartments and functions.

Conclusions

In this work, we presented an efficient method, HiCKey, for detecting and comparing hierarchical TAD structures in Hi-C datasets. We especially derived a GLR test that worked for general distributions. The theoretical results of the GLR test can be used in similar experimental data (such as HiChIP, ChIA-PET and Drop-seq), whose signal may not fully follow the NB distribution but more general mixture distributions. HiCKey was evaluated by using large simulation data and real Hi-C data of mammalian cell lines. First, large-scale validations on simulation data (with or without nested Hi-C structures) show that HiCKey has good precision in recalling known TADs and is robust against random collision noise of chromatin interactions. Second, HiCKey was successfully applied to in situ Hi-C data of seven human cell lines, and its predictions are supported by diverse epigenetic markers and exhibit novel biological discoveries. We concordantly identified multiple layers of TAD organization among these cell lines. In particular, TAD boundaries were found to be significantly enriched in active chromosomal regions compared to repressed regions. HiCKey was manipulated by C++ language for high operation speed. It accepts multiple input formats of the Hi-C matrix and is optimized for processing large matrices constructed from high-resolution Hi-C experiments. With more Hi-C and similar experimental datasets available, we believe our method and theoretical framework will highly inspire computational biologists to design novel pipelines by using the GLR test to elucidate the hierarchical organization of locus-specific chromatin interactions in mammalian genomes or other types of deep sequencing data analysis.

Abbreviations
TAD: Topologically associating domain; GLR: Generalized likelihood ratio; NB: Negative binomial; MCMC: Markov chain Monte Carlo; TPR: True positive rate; FDR: False discovery rate; GEO: Gene expression omnibus; ENCODE: Encyclopaedia of DNA elements.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-021-04113-8.

Additional file 1. Including the proof of Theorems 1 and 2, calculation of time complexity, Figs. S1 and S2, and Tables S1–S10.

Additional file 2. Including the predicted TAD boundaries of nine cell lines.

Acknowledgements
We would like to thank Catherine Tang and Nakoa Kristen Webber for critical reading. We would like to thank Dr. Mattia Forcato for providing detailed information helping with the comparison of HiCKey with other methods. We thank the four anonymous reviewers for their insightful suggestions.

Authors' contributions
HX, YC and MQZ initiated the concept and supervised the study. HX, YW and YC designed the methodology. YW and YC performed the data analysis. YW implemented the software. YC, YW and HX drafted and reviewed the paper. All authors have read and approved the final manuscript.

Funding
Publication costs are funded by Rowan University Startup Grant, 2019, (PI, Yong Chen). Moreover, research reported in this project was partially supported by the NIH Grant R01MH109616, the Cecil H. and Ida Green Endowment, the SKR&DPC Grant (2017YFA0505503) (PI, Michael Q. Zhang), and NSF DMS-1612501 (PI, Haipeng Xing). The funders had no roles in the design of the study and collection, analysis, and interpretation of data or in writing the manuscript.

Availability of data and materials
HiCKey is coded in C++. HiCKey is open source available in the GitHub repository (https://github.com/YingruWuGit/HiCKey).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, 100 Nicolls Rd, Stony Brook, NY 11794, USA. [2] Center for System Biology, University of Texas at Dallas, 800 W Campbell Rd, Richardson, TX 75080, USA. [3] Department of Molecular and Cellular Biosciences, Rowan University, 201 Mullica Hill Rd, Glassboro, NJ 08028, USA.

Received: 25 August 2020 Accepted: 30 March 2021
Published online: 10 April 2021

References

1. Cavalli G, Misteli T. Functional implications of genome topology. Nat Struct Mol Biol. 2013;03(20):290–9.
2. Gibcus J, Dekker J. The hierarchy of the 3D genome. Mol Cell. 2013;03(49):773–82.
3. Bonev B, Cavalli G. Organization and function of the 3D genome. Nat Rev Genet. 2016;10(17):661–78.
4. Liu X, Chen Y, Zhang Y, Liu Y, Liu N, Botten G, et al. Multiplexed capture of spatial configuration and temporal dynamics of locus-specific 3D chromatin by biotinylated dCas9. Genome Biol. 2020;12:21.
5. Ramanand SG, Chen Y, Yuan J, Daescu K, Lambros M, Houlahan KE, et al. The landscape of RNA polymerase II associated chromatin interactions in prostate cancer. J Clin Invest. 2020;130:4.
6. Liu X, Zhang Y, Chen Y, Li M, Zhou F, Li K, et al. In situ capture of chromatin interactions by biotinylated dCas9. Cell. 2017;08(170):1028–43.
7. Spielmann M, Lupiáñez D, Mundlos S. Structural variation in the 3D genome. Nat Rev Genet. 2018;04:19.
8. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. Science. 2002;295(5558):1306–11.
9. Tang YY, Holzel B, Posner M. The neuroscience of mindfulness meditation. Nat Rev Neurosci. 2015;03:16.
10. Mumbach MR, Rubin AJ, Flynn RA, Dai C, Khavari PA, Greenleaf WJ, et al. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nat Methods. 2016;07(13):919–22.

11. Dixon J, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in Mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485:376–80.
12. Rao SSP, Huntley MH, Durand N, Stamenova EK, Bochkov I, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2015;159(7):1665–80.
13. Naumova N, Dekker J. Integrating one-dimensional and three-dimensional maps of genomes. J Cell Sci. 2010;06(123):1979–88.
14. Crane E, Bian Q, McCord R, Lajoie B, Wheeler B, Ralston E, et al. Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature. 2015;06:523.
15. Nora E, Lajoie B, Schulz E, Giorgetti L, Okamoto I, Servant N, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature. 2012;04(485):381–5.
16. Dixon J, Gorkin DU, Ren B. Chromatin domains: the unit of chromosome organization. Mol Cell. 2016;62(5):668–80.
17. Filippova D, Patro R, Duggal G, Kingsford C. Identification of alternative topological domains in chromatin. Algorithms Mol Biol AMB. 2014;9:14.
18. Shin H, Shi Y, Dai C, Tjong H, Gong K, Alber F, et al. TopDom: an efficient and deterministic method for identifying topological domains in genomes. Nucl Acids Res. 2015;44:gkv1505.
19. Lévy-Leduc C, Delattre M, Mary-Huard T, Robin S. Two-dimensional segmentation for analyzing Hi-C data. Bioinformatics. 2014;30(17):i386–92.
20. Gong Y, Lazaris C, Sakellaropoulos T, Lozano A, Kambadur P, Ntziachristos P, et al. Stratification of TAD boundaries identified in reproducible Hi-C contact matrices reveals preferential insulation of super-enhancers by strong boundaries. Nat Commun. 2017;11:141481.
21. Weinreb C, Raphael B. Identification of hierarchical chromatin domains. Bioinformatics. 2015;32(11):1601–9.
22. Haddad N, Vaillant C, Jost D. IC-finder: inferring robustly the hierarchical organization of chromatin folding. Nucl Acids Res. 2017;45(10):e81.
23. Yu W, He B, Tan K. Identifying topologically associating domains and subdomains by Gaussian Mixture model And Proportion test. Nat Commun. 2017;12:8.
24. Malik L, Patro R. Rich chromatin structure prediction from Hi-C data. IEEE/ACM Trans Comput Biol Bioinf. 2018;PP:1.
25. Norton H, Emerson D, Huang H, Kim J, Titus K, Gu S, et al. Detecting hierarchical genome folding with network modularity. Nat Methods. 2018;02:15.
26. Zufferey M, Tavernari D, Oricchio E, Ciriello G. Comparison of computational methods for the identification of topologically associating domains. Genome Biol. 2018;19(217):1–18.
27. Yardimci G, Ozadam H, Sauria M, Ursu O, Yan KK, Yang T, et al. Measuring the reproducibility and quality of Hi-C data. Genome Biol. 2019;03:20.
28. Sauerwald N, Kingsford C. Quantifying the similarity of topological domains across normal and cancer human cell types. Bioinformatics. 2018;07(34):i475–83.
29. Sauerwald N, Singhal A, Kingsford C. Analysis of the structural variability of topologically associated domains as revealed by Hi-C. NAR Genom Bioinform. 2020;03:2.
30. Stansfield J, Cresswell K, Vladimirov V, Dozmorov M. HiCcompare: An R-package for joint normalization and comparison of HI-C datasets. BMC Bioinformatics. 2018;12:19.
31. Heinz S, Texari L, Hayes MGB, Urbanowski M, Chang MW, Givarkes N, et al. Transcription elongation can affect genome 3D Structure. Cell. 2018;174(6):1522–36.
32. Chen F, Li G, Zhang M, Chen Y. HiCDB: a sensitive and robust method for detecting contact domain boundaries. Nucl Acids Res. 2018;09:46.
33. Cresswell K, Dozmorov M. TADCompare: an R package for differential and temporal analysis of topologically associated domains. Front Genet. 2020;03:11.
34. Liu J, Lawrence C. Bayesian inference on biopolymer models. Bioinformatics. 1999;15:38–52.
35. Wang J, Zivot E. A Bayesian time series model of multiple structural changes in level, trend, and variance. J Bus Econ Stat. 2000;18(3):374–86.
36. Lai TL, Xing H. Stochastic change-point ARX-GARCH models and their applications to econometric time series. Stat Sin. 2013;23(4):1573–94.
37. Xing H, Mo Y, Liao W, Zhang M. Genomewide localization of protein-DNA binding and histone modification by BCP with ChIP-seq data. PLoS Comput Biol. 2012;8(7):e1002613.
38. Xing H, Sun N, Chen Y. Credit rating dynamics in the presence of unknown structural breaks. J Bank Finance. 2012;36(1):78–89.
39. Bai J, Perron P. Estimating and testing linear models with multiple structural changes. Econometrica. 1998;66(1):47–78.
40. Perron P, Qu Z. Estimating and testing multiple structural changes in multivariate regressions. Econometrica. 2007;02(75):459–502.
41. Matteson DS, James NA. A nonparametric approach for multiple change point analysis of multivariate data. J Am Stat Assoc. 2014;109(505):334–45.
42. Shen J, Zhang N. Change-point model on nonhomogeneous Poisson processes with application in copy number profiling by next-generation DNA sequencing. Ann Appl Stat. 2012;6:476–96.
43. Zhang N, Siegmund D. A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. Biometrics. 2007;63:22–32.
44. Lavielle M. Using penalized contrasts for the change-point problem. Sig Process. 2005;08(85):1501–10.
45. Harchaoui Z, Lévy-Leduc C. Multiple change-point estimation with a total variation penalty. J Am Stat Assoc. 2010;105(492):1480–93.
46. Forcato M, Nicoletti C, Pal K, Livi C, Ferrari F, Bicciato S. Comparison of computational methods for Hi-C data analysis. Nat Methods. 2017;14:679–85.
47. Lun A, Smyth G. DiffHic: a bioconductor package to detect differential genomic interactions in Hi-C data. BMC Bioinform. 2015;16:258.

48. Fowlkes EB, Mallows CL. A method for comparing two hierarchical clusterings. J Am Stat Assoc. 1983;78(383):553–69.
49. Li D, Hsu S, Purushotham D, Sears RL, Wang T. WashU epigenome browser update 2019. Nucl Acides Res. 2019;07(47):158–65.
50. Fisher RA. Questions and answers #14. Am Stat. 1948;2(5):30–1.
51. Servant N, Varoquaux N, Lajoie B, Viara E, Chen CJ, Vert JP, et al. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. 2015;12:16.
52. Carty M, Zamparo L, Sahin M, González A, Pelossof R, Elemento O, et al. An integrated model for detecting significant chromatin interactions from high-resolution Hi-C data. Nat Commun. 2017;17(8):1–10.
53. Cook KB, Hristov BH, Roch KGL, Vert JP, Noble WS. Measuring significant changes in chromatin conformation with ACCOST. Nucl Acids Res. 2020;48(5):2303–11.
54. Djekidel MN, Chen Y, Zhang MQ. FIND: differential chromatin interactions detection using a spatial Poisson process. Genome Res. 2018;28(3):412–22.
55. Chen Y, Wang Y, Xuan Z, Chen M, Zhang M. De novo deciphering three-dimensional chromatin interaction and topological domains by wavelet transformation of epigenetic profiles. Nucl Acids Res. 2016;44:gkw225.
56. Kepp K, Org E, Sõber S, Kelgo P, Viigimaa M, Veldre G, et al. Hypervariable intronic region in NCX1 is enriched in short insertion–deletion polymorphisms and showed association with cardiovascular traits. BMC Med Genet. 2010;01(11):15.
57. Kennedy R, Ovsyannikova I, Haralambieva I, Lambert N, Pankratz V, Poland G. Genome-wide SNP associations with rubella-specific cytokine responses in measles-mumps-rubella vaccine recipients. Immunogenetics. 2014;05:66.
58. Roberts D, Matsuda T, Bose R. Molecular and functional characterization of the human platelet Na^{+}/Ca^{2+} exchangers. Br J Pharmacol. 2011;07(165):922–36.
59. Esposito A, Annunziatella C, Bianco S, Chiariello A, Fiorillo L, Nicodemi M. Models of polymer physics for the architecture of the cell nucleus. Wiley Interdiscip Rev Syst Biol Med. 2018;12(11):e1444.
60. Nicodemi M, Pombo A. Models of chromosome structure. Curr Opin Cell Biol. 2014;05(28C):90–5.
61. Haddad N, Jost D, Vaillant C. Perspectives: using polymer modeling to understand the formation and function of nuclear compartments. Chromosome Res. 2017;01:25.
62. Belton JM, McCord R, Gibcus J, Naumova N, Zhan Y, Dekker J. Hi-C: a comprehensive technique to capture the conformation of genomes. Methods (San Diego). 2012;05:58.
63. Rottenberg H, Covian R, Trumpower B. Membrane potential greatly enhances superoxide generation by the cytochrome bc1 complex reconstituted into phospholipid vesicles. J Biol Chem. 2009;06(284):19203–10.
64. Lesage A, Dahirel V, Victor JM, Barbi M. Polymer coil-globule phase transition is a universal folding principle of Drosophila epigenetic domains. Epigenet Chromatin. 2019;12:12.
65. Khanna N, Zhang Y, Lucas J, Dudko O, Murre C. Chromosome dynamics near the sol–gel phase transition dictate the timing of remote genomic interactions. Nat Commun. 2019;12:10.
66. Nuebler J, Fudenberg G, Imakaev M, Abdennur N, Mirny L. Chromatin organization by an interplay of loop extrusion and compartmental segregation. Proc Natl Acad Sci. 2018;07(115):201717730.
67. Rowley M, Corces V. Organizational principles of 3D genome architecture. Nat Rev Genet. 2018;10:19.
68. Schwarzer W, Abdennur N, Goloborodko A, Pekowska A, Fudenberg G, Loe-Mie Y, et al. Two independent modes of chromatin organization revealed by Cohesin removal. Nature. 2017;551:09.
69. Sanborn A, Rao S, Huang SC, Durand N, Huntley M, Bochkov I, et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc Natl Acad Sci USA. 2015;10:112.
70. Alipour E, Marko J. Self-organization of domain structures by DNA-loop-extruding enzymes. Nucl Acids Res. 2012;10:40.
71. Fudenberg G, Imakaev M, Lu C, Goloborodko A, Abdennur N, Mirny L. Formation of chromosomal domains by loop extrusion. Cell Rep. 2016;05:15.

Publisher's Note