PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Identifying topologically associating domains using differential kernels

Luka Maisuradze[1], Megan C. King[2], Ivan V. Surovtsev[2], Simon G. J. Mochrie[3], Mark D. Shattuck[4], Corey S. O'Hern[3,5,6]*

1 Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut, United States of America, **2** Department of Cell Biology, Yale School of Medicine, New Haven, Connecticut, United States of America, **3** Department of Physics, Yale University, New Haven, Connecticut, United States of America, **4** Benjamin Levich Institute and Physics Department, The City College of New York, New York, New York, United States of America, **5** Department of Mechanical Engineering and Materials Science, Yale University, New Haven, Connecticut, United States of America, **6** Graduate Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, United States of America

* corey.ohern@yale.edu

Citation: Maisuradze L, King MC, Surovtsev IV, Mochrie SGJ, Shattuck MD, O'Hern CS (2024) Identifying topologically associating domains using differential kernels. PLoS Comput Biol 20(7): e1012221. https://doi.org/10.1371/journal.pcbi.1012221

Editor: Roland L. Dunbrack, Jr., Fox Chase Cancer Center, UNITED STATES OF AMERICA

Received: November 17, 2023

Accepted: June 3, 2024

Published: July 15, 2024

Data Availability Statement: All code used in the manuscript, including the KerTAD algorithm and analysis code is available at https://github.com/lmaisuradze/KerTAD/. All Hi-C maps used in the manuscript are either publicly available or made available at https://figshare.com/s/c92bd17f5bd0882fa3e0.

Funding: The authors acknowledge support from National Science Foundation, Grant No. 1830904 (to L.M., M.C.K., S.G.J.M., I.S, and C.S.O.) and National Science Foundation, Grant No. 2124558

Abstract

Chromatin is a polymer complex of DNA and proteins that regulates gene expression. The three-dimensional (3D) structure and organization of chromatin controls DNA transcription and replication. High-throughput chromatin conformation capture techniques generate Hi-C maps that can provide insight into the 3D structure of chromatin. Hi-C maps can be represented as a symmetric matrix $\mathcal{A}_{ij}$, where each element represents the average contact probability or number of contacts between chromatin loci *i* and *j*. Previous studies have detected topologically associating domains (TADs), or self-interacting regions in $\mathcal{A}_{ij}$ within which the contact probability is greater than that outside the region. Many algorithms have been developed to identify TADs within Hi-C maps. However, most TAD identification algorithms are unable to identify nested or overlapping TADs and for a given Hi-C map there is significant variation in the location and number of TADs identified by different methods. We develop a novel method to identify TADs, KerTAD, using a kernel-based technique from computer vision and image processing that is able to accurately identify nested and overlapping TADs. We benchmark this method against state-of-the-art TAD identification methods on both synthetic and experimental data sets. We find that the new method consistently has higher true positive rates (TPR) and lower false discovery rates (FDR) than all tested methods for both synthetic and manually annotated experimental Hi-C maps. The TPR for KerTAD is also largely insensitive to increasing noise and sparsity, in contrast to the other methods. We also find that KerTAD is consistent in the number and size of TADs identified across replicate experimental Hi-C maps for several organisms. Thus, KerTAD will improve automated TAD identification and enable researchers to better correlate changes in TADs to biological phenomena, such as enhancer-promoter interactions and disease states.

(to L.M. and C.S.O.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Chromatin, which encodes the genetic information for cells, must fold into the cell nucleus that is many times smaller in size. The folded 3D structure of chromatin in the nucleus enables gene expression and proper cell function. With the advent of advanced chromatin conformation capture techniques, we can identify topologically associating domains (TADs), which are regions of the genome that prefer to interact within themselves rather than with neighboring regions. Numerous methods have been developed to automatically detect TADs in Hi-C maps, however, they frequently disagree on the location and number of TADs. We develop a new algorithm, KerTAD, to identify TADs using techniques from image processing and computer vision. We find that our method is more accurate on both synthetic and manually-annotated experimental Hi-C maps than all tested methods. Our method also performs well in the presence of noise and sparsity, which are frequently encountered in experimental Hi-C maps. KerTAD will enable future studies to elucidate the role of TADs in gene regulation and disease formation.

Introduction

Chromatin is a polymer complex of DNA and proteins that forms chromosomes. Chromatin must undergo a highly organized compaction process to fit into the μm-sized nucleus. During this compaction process, chromatin forms hierarchical structures, such as loops, A/B compartments, and territories, across a range of length scales [1–4]. The spatial organization of chromatin is essential for many nuclear processes, such as DNA replication and transcription. For example, during transcription, enhancer and promoter DNA regions that are separated on the chromatin fiber must come into close proximity through the formation of loops to increase the transcription of target genes [1, 5]. Disruptions in chromatin loop formation can alter gene expression by preventing enhancer-promoter interactions [6, 7]. To better understand the structural organization of chromatin, chromosome conformation capture and proximity ligation derivative techniques (in particular Hi-C) have been developed to elucidate genome-wide spatial interactions and structures [8, 9]. Hi-C generates an interaction matrix, $\mathcal{A}_{ij}$, where each element represents the frequency with which two loci i and j on chromatin are close in space, averaged over a cell population [8]. Hi-C maps reveal significant interactions off the diagonal that are not expected for an extended polymer. In particular, Hi-C maps display topologically associating domains (TADs), or regions of increased self-interaction (with decreased interactions outside the region), typically presenting as a square of higher frequency centered on the diagonal [10, 11]. TADs often indicate the formation, elongation, and dissolution of loops. Loops enable enhancer-promoter interactions and TAD boundaries are frequently enriched for insulator proteins and transcription marks, which explains why enhancer-promoter interactions occur mostly within TADs [10, 12–16].

Several features of experimentally determined Hi-C maps, such as noise, sparsity, and low resolution, make TAD identification difficult. Further, TAD features are heterogeneous, e.g. while some TADs possess strong corner points and weak intensity in the interior of the TAD, others possess uniform intensity in the interior with weak borders. TADs are also often difficult to differentiate from the background power-law decay in the interaction frequency away from the diagonal that arises from expected distance-dependent polymer interactions [17]. The convention for TAD identification, or TAD calling, is to specify the starting and ending loci of each TAD in the interaction matrix $\mathcal{A}_{ij}$. However, TADs do not directly report on static

chromatin structure, instead they provide a statistical description of dynamic chromatin organization that is influenced by the experimental methods used to construct the Hi-C maps [12, 18, 19]. Currently, there is no ground-truth definition for TADs in Hi-C maps, and TAD definitions are scale- and resolution-dependent [12, 18, 20]. To illustrate this point, in Fig 1A and 1B, we show the same segment (from 9 to 13 Mb) of mouse chromosome 17 Hi-C map using both linear and logarithmic (base e) intensity scales, respectively. On the linear scale, TADs are not visible, whereas on the logarithmic scale, numerous overlapping and nested TADs appear. (See the Benchmarks subsection in the Materials and Methods for definitions of overlapping and nested TADs.) In Fig 1C we show the same segment of mouse chromosome

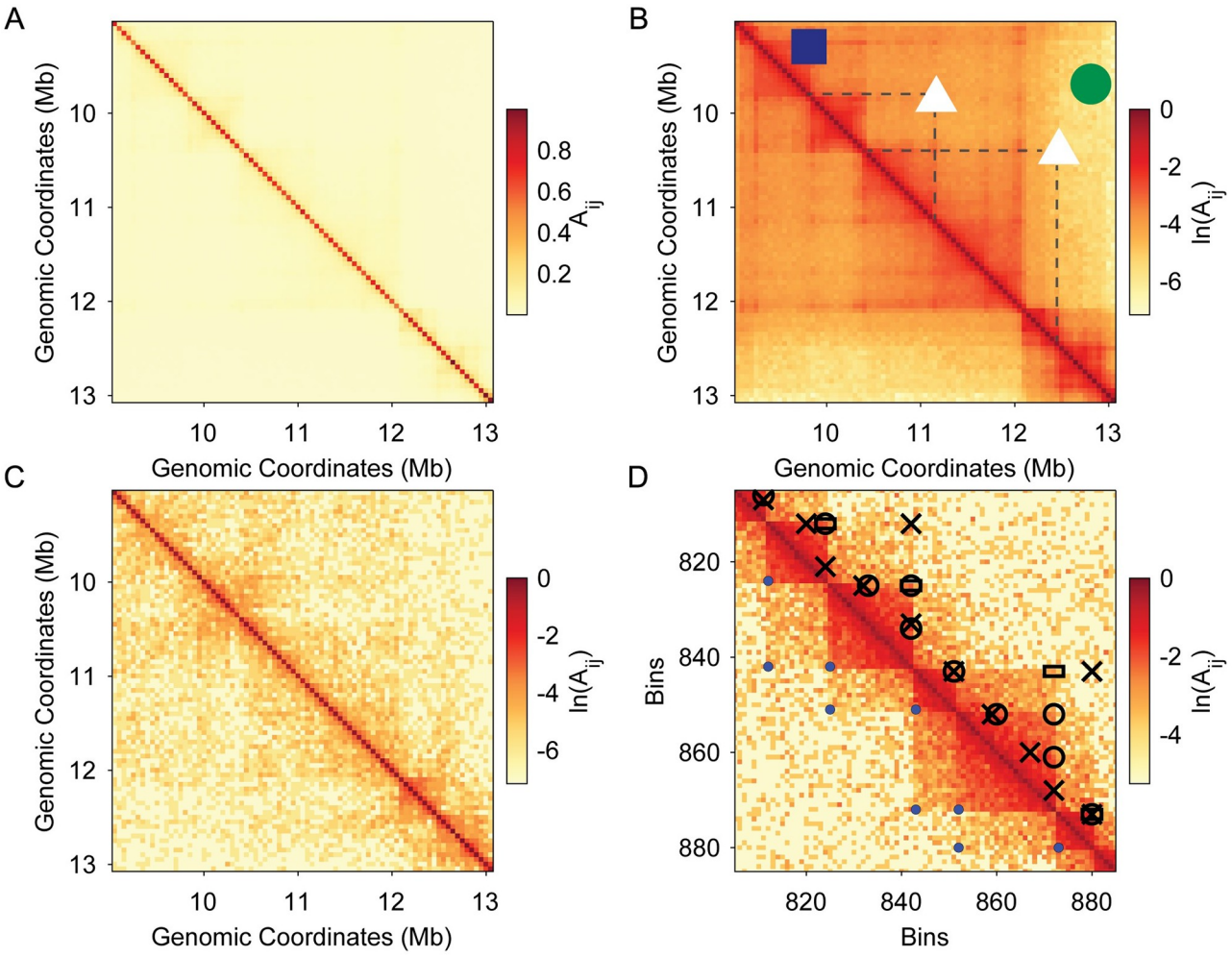

Fig 1. Challenges of TAD identification in Hi-C maps. A: The choice of scale and normalization of Hi-C maps impacts the visibility of TADs. Mouse Hi-C map of chromosome 17 (from 9 to 13 Mb) without preprocessing on a linear scale and normalized so that $0 \leq \mathcal{A}_{ij} \leq 1$ yields faint TADs. B: We show the same Hi-C map as in (A), but plotted on a natural logarithmic scale. The blue square indicates the corner of a clear TAD and the dotted lines in the upper triangular matrix denote its boundaries. The green circle shows a region of noise near a TAD boundary, and the white triangles (and associated dashed lines) indicate "borderline" TADs (structures for which it is unclear whether they would count as TADs) that were not visible in the left image. C: The same region of mouse chromosome 17, but from a different biological replicate with many more low intensity values off the diagonal. D: A synthetic Hi-C map generated from negative binomial distribution sampling with TADs identified (shown in the upper right triangle) using three state-of-the-art TAD calling algorithms: SpectralTAD (open circles), deDoc (crosses), and Armatus (open rectangles). Ground truth TADs (blue circles) are shown in the lower triangular matrix.

https://doi.org/10.1371/journal.pcbi.1012221.g001

17 on a logarithmic scale, but from a different biological replicate, showing a much sparser Hi-C map and replicate to replicate fluctuations.

Because there is currently no clear ground-truth definition of TADs in Hi-C maps, it is challenging to determine the accuracy of TAD calling algorithms on experimental data. However, TAD calling algorithms can be tested on synthetic data that mimics experimental Hi-C maps. The advantage of synthetic data is that it has a well-defined ground-truth and the noise and sparsity of the data can be tuned. To generate a possible ground truth for experimental Hi-C maps, a consensus manual annotation from multiple experts can be obtained. We can then benchmark TAD calling algorithms on their accuracy compared to the manually annotated experimental data [21].

Many algorithms have been developed to identify TADs using graph-theoretic, clustering, machine-learning, and image transform techniques [10, 22–29]. In Fig 1D we compare three state-of-the-art TAD calling algorithms on synthetic data generated by sampling from a negative binomial distribution (see the Complex Synthetic Hi-C Maps subsection in the Materials and Methods for further details) meant to mimic experimental mouse Hi-C maps. These TAD callers identify different numbers of TADs and in different locations, as expected from previous TAD identification algorithm comparison studies [21, 30–33]. Previous studies have found that on manually annotated GM12878 and hESC Hi-C maps at 50 kb resolution, current TAD calling algorithms rarely exceed a positive predictive value of 40% [21]. On synthetic data for overlapping and nested TADs, these methods mostly obtain a true positive rate (TPR) (see Metrics subsection in the Materials and methods) of $\lesssim 0.6$ [31, 33]. In addition, most current TAD-calling algorithms impose strong restrictions that limit their ability to call overlapping, nested, and gapped TADs. [21, 30–33].

In this article, we develop a novel TAD-calling algorithm, KerTAD, that applies gradient and other image operators on Hi-C maps to accentuate and extract their off-diagonal features. We show that KerTAD is more accurate than the current state-of-the-art methods as determined by previous studies [30–33] across three categories of Hi-C maps: synthetic maps generated via molecular dynamics simulations of block copolymers; synthetic maps with overlapping and nested TADs sampled from a binomial distribution of intensities; and manually annotated GM12878 maps at 50kb resolution. On all three datasets, KerTAD is the most accurate in terms of TPR while having a negligible false discovery rate (FDR). On synthetic data, our method has an average TPR of ≈ 0.98 and ≈ 0.99 on non-nested and nested maps, respectively, and a median TPR of ≈ 0.75 on manually annotated Hi-C maps. In addition, KerTAD is highly resistant to noise and sparsity, achieving a higher TPR at the highest level of noise tested than other methods with no noise. Because KerTAD outperforms every tested method on both manually annotated experimental and synthetic data, KerTAD is likely able to capture the underlying features in experimental Hi-C maps.

This article is organized as follows. In the Materials and methods section, we first describe the preprocessing of the input Hi-C maps and the generation of masks to identify key features of TADs in Hi-C maps. We also define the metrics for sensitivity and false discovery rate for comparing the predictions of KerTAD to ground truth for the synthetic and manually annotated Hi-C maps. We then define the techniques used for generating noise and sparsity in synthetic data. In the results section, we summarize the performance of KerTAD (as well as six other methods) in TAD identification on synthetic and manually annotated Hi-C maps. We also analyze replicate Hi-C maps across four organisms and compare the variation in number and mean size of TADs identified by three TAD identification algorithms. Finally, we discuss how the improved accuracy in TAD identification will enable more robust inferences between the identified TADs and chromatin organization.

Materials and methods

The description of the Materials and methods is organized into two sections. In the first section, we explain the new TAD identification algorithm, KerTAD, including the preprocessing steps and the application of masks to identify key features of TADs. In the second section, we discuss the implementation of six other state-of-the-art methods to identify TADs, metrics that we use to quantify the accuracy of the TAD identification methods, and techniques to generate sparse and noisy synthetic data. We describe the motivation and process of manually annotating experimental Hi-C maps, as well as the methods for comparing the accuracy of TAD identification methods on manually annotated experimental data. We finally describe in detail our analysis of the performance of several TAD identification algorithms on replicate non-annotated experimental Hi-C maps across several organisms.

KerTAD

KerTAD takes as input a symmetric $N \times N$ matrix, $\mathcal{A}_{ij}$, which gives the frequency of contacts between bins i and j and returns an $M \times 2$ matrix, where each row gives the corner location of one of the M TADs in $\mathcal{A}_{ij}$. The preprocessing step normalizes $\mathcal{A}_{ij}$ such that $\mathcal{A}_{ii} \geq \mathcal{A}_{ij}$ for all i, j and reduces fluctuations in $\mathcal{A}_{ij}$ while preserving edge features. The method then feeds the preprocessed Hi-C map into two separate pipelines, each of which generates a mask. One pipeline seeks to extract small-scale diffuse point features in the Hi-C map, while the other favors larger scale regions near corner points. The final TADs are given by the intersection of the two masks.

Preprocessing. There is no standard format or normalization scheme for Hi-C maps [34–40]. Because normalization is known to significantly affect TAD-calling performance [34], we first preprocess $\mathcal{A}_{ij}$ to satisfy the requirements below. First, we ensure that the diagonal elements of $\mathcal{A}_{ij}$ are the maxima in their respective rows, i.e. $\mathcal{A}_{ii} \geq A_{ij}$. If a given $\mathcal{A}_{ij} > \mathcal{A}_{ii}$, we then set $\mathcal{A}_{ii} = \mathcal{A}_{ij}$. This condition is reasonable in the sense that we should expect that local regions of chromatin interact with themselves more than any other region. We then locally row-normalize by re-setting $\mathcal{A}_{ij}$ to $(\mathcal{A}_{ij} - \sum_{j=1}^{N} \mathcal{A}_{ij}/N)/\sigma_i$, where σ_i is the standard deviation of the ith row of $\mathcal{A}_{ij}$. This normalization reduces global fluctuations and also perturbs the original $\mathcal{A}_{ij}$ less than other normalization schemes like requiring $\mathcal{A}_{ij}$ to be both row- *and* column-normalized [39]. We then filter $\mathcal{A}_{ij}$ with a Gaussian kernel with standard deviation $\sigma = 3\Gamma/2$ and filter size $2\lceil(2\sigma)\rceil + 1$, where ΓN^2 is the number of zero elements in $\mathcal{A}_{ij}$ and $\lceil\cdot\rceil$ is the ceiling function. This Gaussian filtering is performed since extremely sparse Hi-C maps can cause division by zero errors in the KerTAD masks. Additionally, normalization and applying a Gaussian kernel to $\mathcal{A}_{ij}$ reduces the total variation of $\mathcal{A}_{ij}$, which is defined as:

$$V(\mathcal{A}_{ij}) = \sum_{i=1}^{N}\sum_{j=1}^{N} |\Delta_y \mathcal{A}_{ij}| + |\Delta_x \mathcal{A}_{ij}|, \tag{1}$$

where $\Delta_y \mathcal{A}_{ij} = \mathcal{A}_{(i+1)j} - \mathcal{A}_{ij}$, $\Delta_x \mathcal{A}_{ij} = \mathcal{A}_{i(j+1)} - \mathcal{A}_{ij}$, and the outside bins of $\mathcal{A}_{ij}$ are given by $\mathcal{A}_{(N+1)j} = \mathcal{A}_{Nj}$, $\mathcal{A}_{i(N+1)} = \mathcal{A}_{iN}$, $\mathcal{A}_{i0} = \mathcal{A}_{i1}$, and $\mathcal{A}_{0j} = \mathcal{A}_{1j}$. While spatial variation is a hallmark of TADs, excessive total variation outside of TAD boundaries (such as speckle noise) can obscure the signal and make TAD identification challenging. It is important however to regulate standard smoothing techniques, like Gaussian blurring, since while they can reduce the total variation, they can also remove stark edge features that are essential for identifying TADs. Finally, $\mathcal{A}_{ij}$ is automatically segmented (if necessary) by finding outliers on the diagonal where

the ratio of zero elements to nonzero elements of the 5 elements around the diagonal (either to the left or to the right of the diagonal depending on the location of the diagonal) for each row is greater than 0.8. Further outliers are found using the Grubbs method if necessary and then all the adjacent non-outliers are segmented into separate maps to process [41].

Mask for corner point features. The mask for corner point features is designed to identify locations near the diagonal where there are strong changes in intensity, since these often indicate transitions between TADs, and then to generate a mask of possible corner point combinations in $\mathcal{A}_{ij}$. We first calculate the discrete partial derivative of $\mathcal{A}_{ij}$. We then feed the row vectors of the partial derivative map into a non-linear function that produces a similarity matrix. The similarity matrix is then filtered by applying a local maximum operator and global threshold, which identifies locations on the diagonal of $\mathcal{A}_{ij}$ where there are sharp local changes. We then use the identified locations on the diagonal to generate a binary mask of every TAD corner point combination, with each diagonal location representing one index of a possible TAD corner point. Differential operators in image processing are often represented as convolutions of an image with a kernel that is separable into at least one smoothing filter. Smoothing can reduce noise, but excessive smoothing removes edge features, making it difficult to determine TAD locations. Thus, we implement a low-order partial derivative map with no smoothing filter, $\Delta_y \mathcal{A}_{ij}$, with symmetric boundary conditions.

Next, we construct a list of row vectors $\{\vec{v}_1, \ldots, \vec{v}_N\}$, where $\vec{v}_i$ is the ith row of $\Delta_y \mathcal{A}_{ij}$. We then construct a similarity matrix, $\mathcal{S}_{ij}$,

$$\mathcal{S}_{ij} = (\max(\vec{v}_i) - \min(\vec{v}_i) + \max(\vec{v}_j) - \min(\vec{v}_j))||\vec{v}_i||_1||\vec{v}_j||_1, \tag{2}$$

and $\max(\vec{v}_i)$ and $\min(\vec{v}_i)$ return the maximum and minimum components of $\vec{v}_i$, respectively. Finally, we define the $N \times N$ binary mask of point features, $\mathcal{M}_{ij}$, as follows: for every i, j such that $i < j$, $\mathcal{M}_{ij} = 1$ if and only if $\mathcal{S}_{ii}$ and $\mathcal{S}_{jj}$ are both local maxima in their respective 3×3 local neighborhoods *and* $\mathcal{S}_{ii}, \mathcal{S}_{jj} \geq \Omega$, where Ω is the global threshold determined using the triangle algorithm [42] on $\mathcal{S}_{ij}$. Fig 2 illustrates the several intermediate steps and maps to transform an input Hi-C map, $\mathcal{A}_{ij}$, into $\mathcal{M}_{ij}$.

Mask for corner regions. While the previous mask captured *point* features of TADs spread throughout the Hi-C map, we also need a mask to identify the specific corner *regions* near the diagonal in $\mathcal{A}_{ij}$. As before, we calculate an image derivative, this time $\Delta_x \mathcal{A}_{ij}$, using periodic boundary conditions. For $i < j$, if $\Delta_x \mathcal{A}_{ij} > 0$ then $\Delta_x \mathcal{A}_{ij}$ is set to 0 and for $i > j$ if $\Delta_x \mathcal{A}_{ij} < 0$ then $\Delta_x \mathcal{A}_{ij}$ is set to 0. We then calculate

$$\mathcal{P}_{ij} = \sum_{k=1}^{N} (\Delta_x \mathcal{A}_{ik} \Delta_x \mathcal{A}_{kj}^T - \Delta_x \mathcal{A}_{ik}^T \Delta_x \mathcal{A}_{kj}). \tag{3}$$

$\mathcal{P}_{ij}$ has several important features. First, TAD corners and edges are maxima of $\mathcal{P}_{ij}$ in their local neighborhood as shown in Fig 3C. The diagonal elements of $\mathcal{P}_{ij}$ that correspond to TAD corner points (i.e. if $\mathcal{A}_{ij}$ is the corner point of a TAD, the corresponding points in $\mathcal{P}_{ij}$ are $\mathcal{P}_{ii}$ and $\mathcal{P}_{jj}$) are strongly negative minima in their neighborhood. Taking advantage of both of these facts, we construct the final binary mask $\mathcal{M}'_{ij}$:

$$\mathcal{M}'_{ij} = \begin{cases} 1 & \text{if } (-\mathcal{P}_{ij}(\mathcal{P}_{ii} + \mathcal{P}_{jj})) \geq \Omega \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where Ω is the threshold determined by the triangle method on the matrix, $\mathcal{C}_{ij} = -\mathcal{P}_{ij}(\mathcal{P}_{ii} + \mathcal{P}_{jj})$.

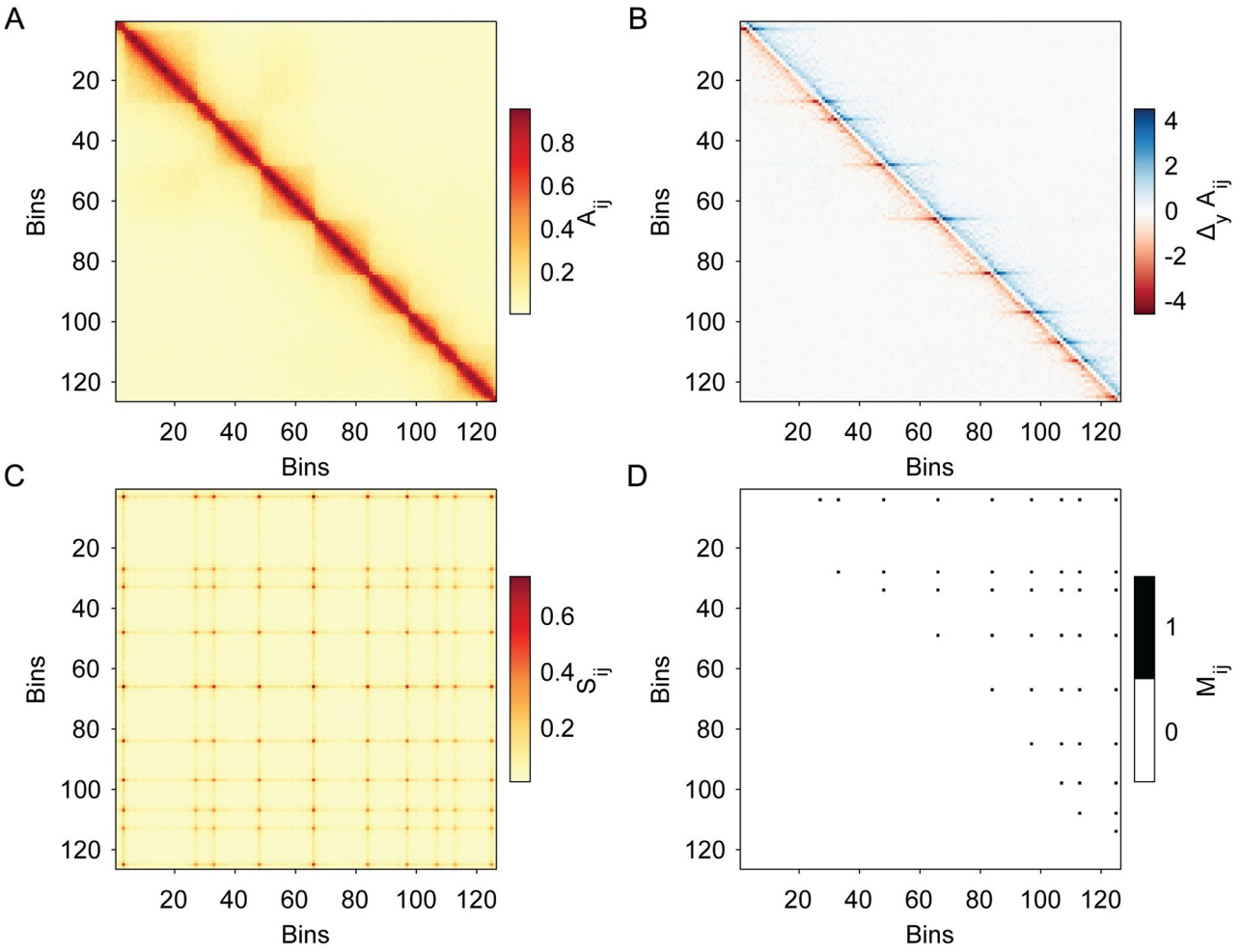

Fig 2. Illustration of the four steps in constructing the point feature binary mask. A: We start with a Hi-C map $\mathcal{A}_{ij}$ (with bins i and j labelled from 1 to 125). B: We first calculate the discrete partial derivative, $\Delta_y A_{ij}$. C: We then construct $\mathcal{S}_{ij}$ from a nonlinear function of the pairs of row vectors of $\Delta_y \mathcal{A}_{ij}$. D: The binary mask $\mathcal{M}_{ij}$ is obtained by combining a local maximum filter with binary thresholding of $\mathcal{S}_{ij}$. If $\mathcal{M}_{ij} = 1$ (black squares), $\mathcal{S}_{ii}$ and $\mathcal{S}_{jj}$ are both local maxima in their 3 × 3 windows and above the threshold set by the triangle method on $\mathcal{S}_{ij}$.

https://doi.org/10.1371/journal.pcbi.1012221.g002

Final mask and parameters. After constructing both masks, we take the element-wise product of $\mathcal{M}$ and $\mathcal{M}'$ to obtain the final binary mask, $\mathcal{B}_{ij} = \mathcal{M}_{ij}\mathcal{M}'_{ij}$. Each nonzero element of $\mathcal{B}$ represents a predicted TAD corner point. For the final output, KerTAD converts $\mathcal{B}$ to a 2 column list where each row represents the start and end index of a TAD corner point. We illustrate the full algorithm applied to chromosome 12 of a GM12878 cell line in Fig 4.

KerTAD does not require any user-provided parameters, taking only a Hi-C matrix as input. However, KerTAD has two optional hard-coded parameters that can be manually overridden for expert users to have more flexibility. First, the parameter κ is the maximum number of TADs that can be identified per row. Starting from the diagonal and moving outward for each row, every nth TAD where $n > \kappa$ is discarded. By default, $\kappa = 3$, since greater than 3 TADs per row is unlikely. Throughout the manuscript we use $\kappa = 3$ with the exception of simple Hi-C maps where we set $\kappa = 1$ (to mimic simple TAD callers). The other optional parameter is γ, which controls how many times the initial automatic segmentation of $\mathcal{A}_{ij}$ is broken

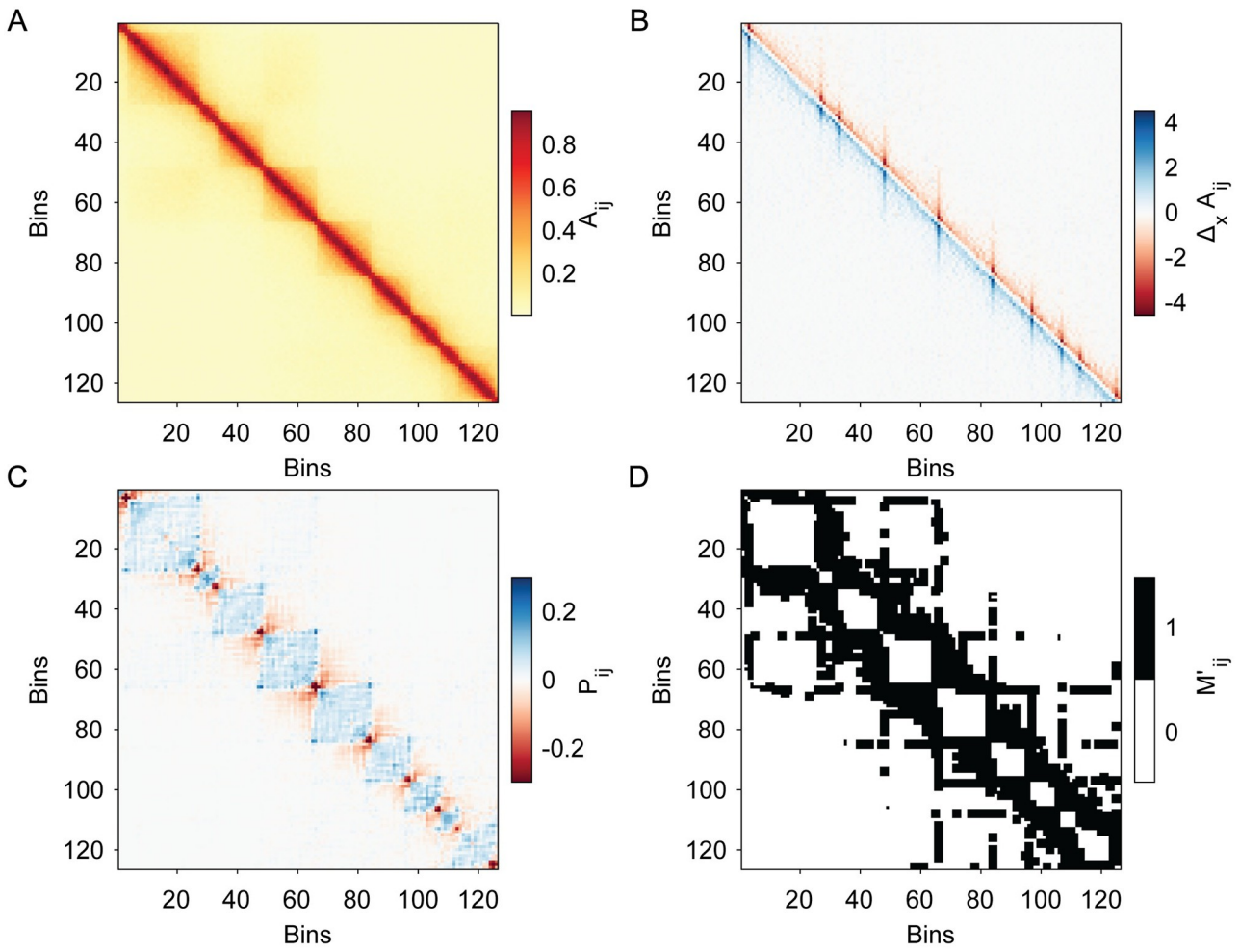

Fig 3. Illustration of the steps used to construct the mask $\mathcal{M}'_{ij}$ for identifying corner regions in Hi-C maps. A: We start with the same input Hi-C map $\mathcal{A}_{ij}$ as in Fig 2. B: We first calculate the discrete partial derivatives, $\Delta_x A_{ij}$. C: We then calculate $\mathcal{P}_{ij}$ from $\Delta_x A_{ij}$. D: We obtain the final binary mask $\mathcal{M}'_{ij}$ after applying a global threshold on $-\mathcal{P}_{ij}(\mathcal{P}_{ii} + \mathcal{P}_{jj})$.

https://doi.org/10.1371/journal.pcbi.1012221.g003

into smaller maps for additional segmentation. If a segmentation, $\mathcal{W}$, of $\mathcal{A}_{ij}$ spans from $\mathcal{A}_{xx}$ to $\mathcal{A}_{yy}$ (i.e. $\mathcal{W}_{ij} = \mathcal{A}_{(x+i-1)(y+j-1)}$), then if $\gamma = n$, $\mathcal{W}$ is divided into n further segments, $w_1, \ldots, w_n$, where w_i spans $\mathcal{W}_{(i+(i-1)\tau)(i+(i-1)\tau)}$ to $\mathcal{W}_{(i+i\tau)(i+i\tau)}$, $\tau = \lceil \frac{\gamma - x}{n} \rceil$, and $\lceil . \rceil$ is the ceiling function. By default, $\gamma = 2$. Further splitting Hi-C maps is useful for large and heterogeneous Hi-C maps where different regions have varying coverage and local intensity. Generally, TAD predictions scale with γ (i.e. increasing γ will increase the number of TADs predicted). γ can be tuned based on user preference in either direction. (Increasing γ will likely increase TPR, but it will also increase FDR.) For robust Hi-C maps, γ can be set to 1 (i.e. no further segmentations are calculated after the initial automatic segmentation). For this manuscript, we use $\gamma = 1$ for synthetic Hi-C maps and $\gamma = 2$ for all other Hi-C maps.

Benchmarks

When determining the accuracy of TAD identification methods, we first categorize the Hi-C maps into two types: synthetic and experimental Hi-C maps. For synthetic Hi-C maps, we also

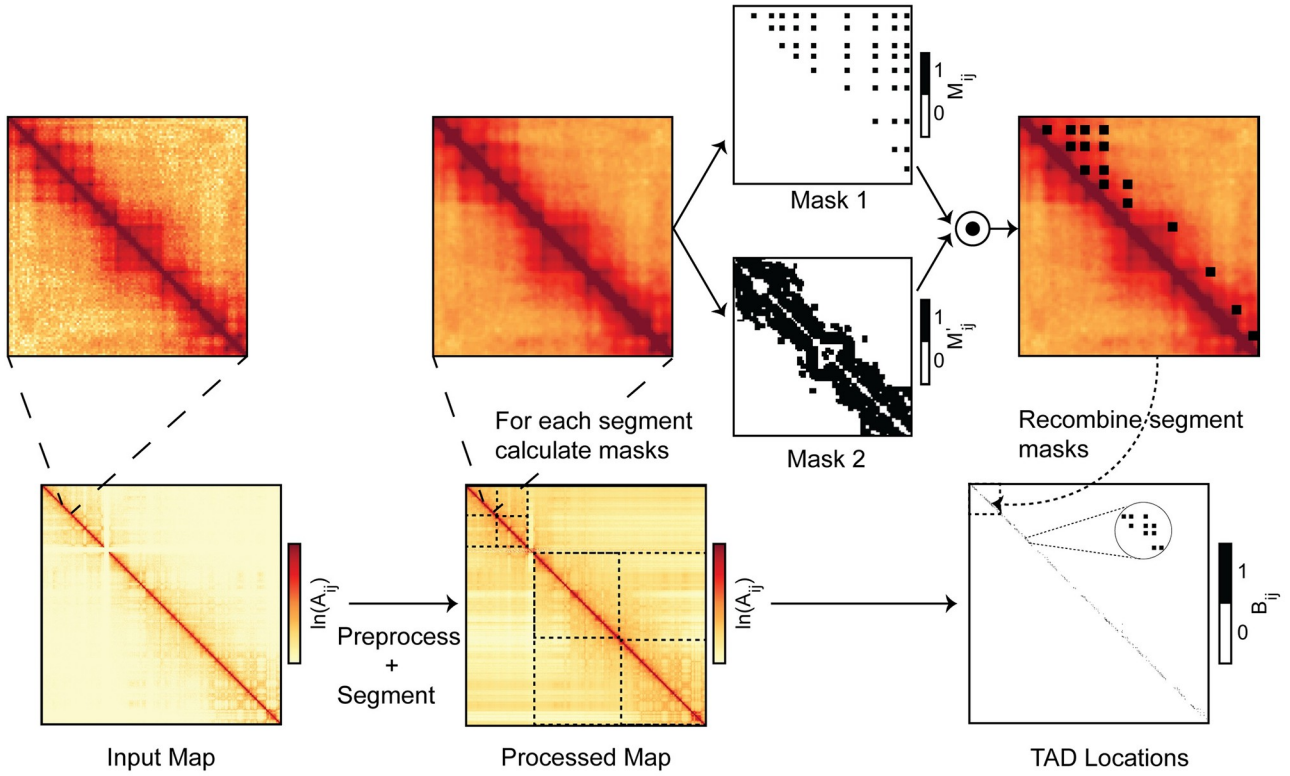

Fig 4. Graphical representation of KerTAD algorithm. The progression through the KerTAD algorithm is depicted above: the initial input Hi-C map undergoes preprocessing, followed by segmentation of the preprocessed map, application of masks 1 and 2 (see Figs 2 and 3) for each segment, calculating their element-wise product (shown overlaid with the input segment), followed by recombination of each binary output for each segment into the final TAD predictions.

https://doi.org/10.1371/journal.pcbi.1012221.g004

distinguish between "simple" and "complex" Hi-C maps. For simple Hi-C maps, each element on the diagonal of $\mathcal{A}_{ij}$ must belong to one and only one TAD. This condition implies that i) $\mathcal{A}_{ij}$ has no nested or overlapping TADs and ii) $\mathcal{A}_{ij}$ has no gaps between TADs. Thus, in a simple Hi-C map, if a TAD is identified over a set of diagonal elements, e.g. from A_{ii} to A_{jj}, there are no other TADs within that set and the next TAD must start at $A_{(j+1)(j+1)}$. Complex Hi-C maps are defined as any Hi-C map $\mathcal{A}_{ij}$ that is not simple, i.e. $\mathcal{A}_{ij}$ has either nested, overlapping, or gapped TADs. A nested TAD is a TAD with its corner point located at $\mathcal{A}_{ij}$ (where $j > i$) while there exists another TAD corner at $\mathcal{A}_{kl}$ (where $l > k$), where $k \leq i$ and $j \leq l$. An overlapping TAD has a corner at $\mathcal{A}_{ij}$ (where $j > i$) and another TAD corner at $\mathcal{A}_{kl}$ (where $l > k$), where $k < i$ and $i < l < j$ or $i < k < j$ and $j < l$. A Hi-C map possesses a gapped TAD if there exists an element on the diagonal, $\mathcal{A}_{ii}$, that does not belong to any TAD. In S1 Fig, we show a graphical illustration and examples from the GM12878 chromosome 6 Hi-C map of nested, overlapping, and gapped TADs.

We analyze the performance of TAD identification algorithms on simple and complex synthetic Hi-C maps separately. Many TAD identification algorithms assume that the input Hi-C maps are simple. This additional information provides constraints on the locations of TADs, which can lead to enhanced accuracy for these algorithms. However, the additional constraints do not improve TAD prediction in manually annotated experimental Hi-C maps, as most experimental Hi-C maps are not simple. In previous work comparing the performance of

TAD identification algorithms, the top performers on simple and complex synthetic maps were different [31, 33]. In the Results section, we show that KerTAD is highly accurate in identifying TADs in both simple and complex Hi-C maps, while not presupposing that a given Hi-C map is simple or complex.

Simple synthetic Hi-C maps. To compare the performance of different TAD identification algorithms for simple, synthetic Hi-C maps, we consider 100 Hi-C maps generated by molecular dynamics (MD) simulations of block copolymers from previous studies [43]. In these MD simulations, chromatin is modeled as a bead-spring polymer with non-bonded, purely repulsive interactions to prevent bead overlaps, non-specific short-ranged attractive interactions between bead pairs to induce compaction, and specific short-ranged attractive interactions between bead pairs to mimic TADs that occur in specific epigenomic profiles.

From previous studies [21, 30–33] we select the top performing TAD identification algorithms for simple, synthetic maps. Namely, we compare KerTAD with TopDom [27], HICSeg [28], and CHDF [29]. We perform TAD identification on the set of 100 simple, synthetic Hi-C maps discussed above. (Note that TopDom, HICSeg, and CHDF do not identify nested or overlapping TADs.) For TopDom we count the "domain" predictions and set the window size to 5 as done in previous work [31, 33] for the same synthetic Hi-C maps. Again following previous work [21, 30, 31, 33], we set the max TAD size parameter for CHDF to 50 and for HICSeg we use the "G" distribution. When comparing TAD predictions from KerTAD to those for the other algorithms on the simple, synthetic Hi-C maps, we impose a further restriction on our identified TADs. Since KerTAD can identify nested and overlapping TADs, it has more chances to identify correct TADs compared to methods that are unable to call nested and overlapping TADs. Thus, we set $\kappa = 1$, considering only the innermost TAD corners with the smallest distance from the diagonal.

Complex synthetic Hi-C maps. For generating complex, synthetic Hi-C maps, we use a variation of a previously developed procedure [30, 44] that mimics mouse embryonic stem cells by sampling from a negative binomial distribution of Bernoulli trials, where successful trials represent contacts between chromatin loci. The distribution is characterized by a location-dependent variance $\sigma_{ij}^2 = \mu_{ij} + r\mu_{ij}^2$ (with dispersion factor $r = 0.01$) and mean $\mu_{ij} = \langle \mathcal{A}_{ij} \rangle$. The location-dependent mean is defined by

$$\mu_{ij} = K_d \delta_{ij} + \theta_{ij} K_t (i - j + 1)^c + \mathcal{N}_{noise}, \tag{5}$$

where δ_{ij} is the Kronecker-delta, K_d gives $\langle \mathcal{A}_{ii} \rangle$, K_t and c are parameters that control the power-law decay of $\langle \mathcal{A}_{ij} \rangle$ away from the diagonal. ($K_d = 35$, $K_t = 28$, and $c = -0.69$ were selected to match $\langle \mathcal{A}_{ij} \rangle$ in chromosome five in IMR90 replicate B.) $\theta_{ij} = 1$ when $\mathcal{A}_{ij}$ is inside of a TAD (excluding diagonal elements) and 0 otherwise [17, 30]. TAD boundary lengths are selected randomly from a uniform distribution with widths from 5 to 20 bins (where each bin represents 40 kb). We then remove randomly selected TADs from this list and fill in the gaps with larger overlapping and nested TADs. The deletion process involves randomly selecting 25% of TADs in the lower layer, removing them, and then adding a new TAD block that spans the length of the two TADs between any deleted TADs, thus creating nested and overlapping TADs. More specifically, if two TADs around a deleted TAD have corner points located at (x_1, y_1) and (x_2, y_2), the new TAD will have a corner point located at (x_1, y_2). $\mathcal{N}_{noise}$ is a random variable that mimics weak and non-specific ligation events by sampling (with replacement) a fraction of randomly selected elements of $\mathcal{A}_{ij}$ and adding a constant, K_{noise}. (We set $K_{\text{noise}} = 5$.) The likelihood that an element of $\mathcal{A}_{ij}$ receives a noise impulse scales with $(i - j + 1)^c$.

We generate 100 complex, synthetic Hi-C maps using this protocol with $\mathcal{N}_{noise} = 0$, where each Hi-C map has on average 150 TADs. From previous studies [21, 31–33] we select the top

performing TAD callers on similar datasets of complex, synthetic Hi-C maps. We compare KerTAD with deDoc [24], Armatus [22], and SpectralTAD [25]. As before, we follow the default or recommended parameters for each algorithm. For Armatus we set $g = 0.05$ and $s = 0.05$ [30], for SpectralTAD we use levels = 2, and for deDoc we use both the dedoc(M) and dedoc(E) predictions, removing duplicates. The accuracy of TAD identification was determined for these three methods, along with KerTAD, for each complex, synthetic Hi-C map.

Noise and sparsity. To test the robustness of the TAD identification algorithms, we compare TAD predictions for two sets of new complex, synthetic Hi-C maps with varying levels of added noise and sparsity. In the first set, we generate 10 complex Hi-C maps with $\mathcal{N}_{noise} = 0$ (as previously described) and for each, construct an additional 20 Hi-C maps, with varying levels of noise (totalling 210 Hi-C maps). Because many TAD identification algorithms only accept integer counts, we do not use additive Gaussian noise. Instead, we randomly sample $\mathcal{A}_{ij}$ (with replacement) and add a constant additive impulse, $K_{\text{noise}} = 5$, as described previously for $\mathcal{N}_{noise}$. The noise is parameterized by χ, which represents the number of added impulses divided by the number of elements of $\mathcal{A}_{ij}$. To generate the noisy maps, we increase χ in increments of 0.05 starting from 0 to 1. For the second set, we perform the same procedure but instead add sparsity to $\mathcal{A}_{ij}$ by setting random elements of $\mathcal{A}_{ij}$ equal to 0. Sparsity is parameterized by ξ, which is the fraction of elements of $\mathcal{A}_{ij}$ that are set to zero compared to the total number of elements. We generate 200 sparse maps by increasing ξ in increments of 0.05 starting from 0 to 0.95 ($\xi = 1$ would mean a map of only 0s). For KerTAD, we turn off outlier detection for highly sparse Hi-C maps to avoid runtime errors.

Experimental maps. To obtain ground truth for experimental Hi-C maps, we follow the previous manual annotations performed on Hi-C maps for the GM12878 cell line at 50 kb resolution for the 40–45 Mb regions of 10 different chromosomes (chromosomes 2, 3, 4, 5, 6, 7, 12, 18, 20, and 22) [21]. In the original annotations, "any identifiable TAD structure" was annotated and the positive predictive value (PPV) of the identified TADs was calculated for seven TAD identification algorithms [21]. However, calculating PPV does not penalize TAD callers that miss "obvious" TADs and even TPR may be inappropriate for gauging TAD prediction accuracy if the annotations are too lenient. In addition, likely due to differences in the pipeline or visualization, we found that many of the original annotations were displaced or pointed at no features or structures. Thus, using the original annotations as a guide, we keep the most "obvious" TADs and then calculate TPR to capture the accuracy of the TAD identification methods. Because the annotations are not meant to be exhaustive, we do not calculate FDR. Because the experimental Hi-C maps are complex, we use deDoc, Armatus, and SpectralTAD, as well as KerTAD, to identify TADs in the manually annotated GM12878 Hi-C maps. For the input maps to each TAD caller, we used the cutout sections of the genome except for Armatus, which returned no TADs with the smaller map (a previously described bug) and for which we used the full intrachromosomal map as input.

For experimental Hi-C maps without manual annotations, we evaluate in situ Hi-C maps for four organisms: fruit fly S2 cells [45] (4DN accession code: 4DNESFOADERB), zebrafish embryos [46] (4DN accession code: 4DNESV5PGOUC), mouse CH12.LX cells [17] (4DN accession code: 4DNESK95HVFB), and human HCT-116 cells [47] (4DN accession code: 4DNES3QAGOZZ). All Hi-C maps were obtained from the 4DN data portal [48] and the `.pairs` files for each biological and technical replicate were converted to `.cool` files and then intrachromosomal Hi-C maps at 50 kb resolution were extracted using Cooler [49]. For zebrafish Hi-C maps, we analyzed three biological replicates with one technical replicate for each biological replicate. For fruit fly Hi-C maps, we also analyzed three biological replicates with one technical replicate each. For mouse Hi-C maps, we used three biological replicates

with 11, 2, and 2 technical replicates. For human Hi-C maps, we analyzed six biological replicates with 3, 4, 2, 3, 2, and 2 technical replicates. For each Hi-C map, we perform TAD identification using KerTAD and the top performers in TPR for the simple and complex Hi-C map categories: TopDom and deDoc. For TopDom we used a window size of 10 following the recommendation for 50kb resolution from previous work [21]. Because TopDom generated an error message for chromosome Y of biological replicate 2 for fruit fly, we do not include that Hi-C map in our analysis for TopDom. We calculate the total number of identified TADs by summing the number of predicted TADs for each intrachromosomal map for each replicate. We also calculate the mean size of the identified TADs for each intrachromosomal map. We characterize the distribution of the number of TADs and mean sizes of TADs over replicates for each organism by calculating the median, maximum, and minimum values.

Accuracy metrics. We apply each TAD identification algorithm to each synthetic or manually annotated experimental Hi-C map and compare the lists of identified TADs to ground truth. For a predicted TAD corner point located at $\mathcal{A}_{ij}$, we denote it as a "true positive" if and only if there is a ground truth TAD with the same corner point coordinates. We calculate two metrics for each synthetic and experimental Hi-C map for every algorithm: TPR $= p/\mathcal{G}$ and FDR $= (\mathcal{T} - p)/\mathcal{T}$, where p is the number of true positives, $\mathcal{G}$ is the total number of ground truth TADs, and $\mathcal{T}$ is the total number of TADs predicted. In manually annotated experimental Hi-C maps, since the TAD corners are often difficult to define, a "true positive" is counted as long as the ground truth coordinate is one of the coordinates in the 3×3 square centered around the predicted TAD corner point.

Results

In this section, we compare the performance of KerTAD against current state-of-the-art TAD identification methods using two metrics: the ability to reliably identify ground truth TADs (TPR) and the ability to avoid predicting incorrect TADs (FDR). We compare the accuracy of seven different methods on two sets of synthetic Hi-C maps: a set of simple Hi-C maps obtained from MD simulations of block copolymers and a set of complex Hi-C maps generated by sampling a negative binomial distribution. We also calculate TPR and FDR for the same TAD identification algorithms on manually annotated Hi-C maps from the GM12878 cell line. Finally, we calculate the number and size of TADs obtained using each algorithm on in-situ experimental Hi-C maps for four organisms: mouse, human, fruit fly, and zebrafish.

On the 100 simple, synthetic Hi-C maps, our method gives the highest median TPR ≈ 0.99 and the lowest median FDR ≈ 0.02 of all surveyed methods (Fig 5A). The next best performing algorithm, TopDom, had a comparable median TPR ≈ 0.94 and median FDR ≈ 0.03, but TopDom yields a significantly larger variance with a minimum TPR ≈ 0.65 compared to ≈ 0.88 for our method. In Fig 5A, we also show that the other TAD identification algorithms, CHDF and HiCSeg, performed poorly on the simple, synthetic Hi-C maps with a median TPR < 0.6 and median FDR > 0.2. (Note that the median FDR ≈ 0.7 for CHDF was larger than its median TPR ≈ 0.5.) In previous work, [31, 33] CHDF was reported to perform very well on this synthetic dataset (hence why it was selected for comparison), scoring a mean TPR ≈ 0.965 and FDR ≈ 0.381. Even so, KerTAD still outperforms CHDF in both TPR and FDR. In fact, KerTAD scores a higher mean and minimum TPR than all 27 surveyed TAD callers in previous work [31, 33]. Furthermore, when running our method on simple, synthetic Hi-C maps, we did not allow it to call nested or overlapping TADs. Without this restriction, the median TPR was even greater than 0.99, while maintaining small median FDR.

For the 100 complex, synthetic Hi-C maps, the differences in the median TPR between the new method and the other tested algorithms are more pronounced, as shown in Fig 5B. The

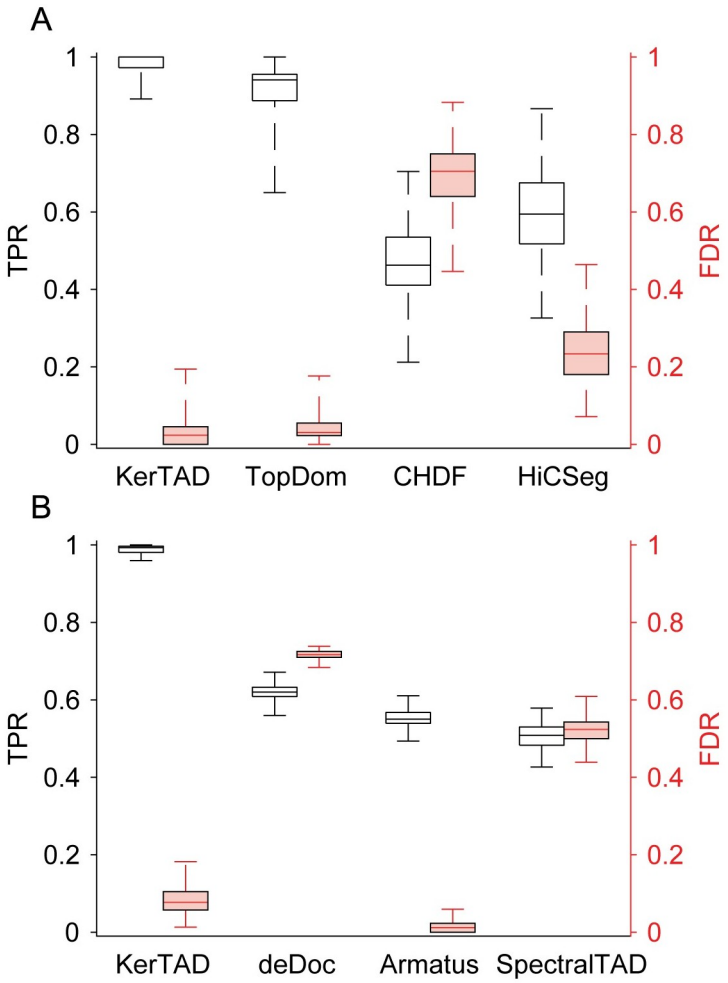

Fig 5. TPR and FDR on simple and complex synthetic Hi-C maps. A: Box plots of TPR (black; left axis) and FDR (red; right axis) calculated by comparing the ground truth TADs from 100 simple, synthetic Hi-C maps generated by MD simulations of block copolymers [43] and TADs predicted by KerTAD, TopDom, CHDF, and HiCSeg. The box edges represent the 25th and 75th percentiles in TPR/FDR, and the central line in each box indicates the median. The error bars represent the maximum and minimum TPR or FDR. B: Box plots for TPR (black) and FDR (red) for 100 complex, synthetic Hi-C maps that mimic mouse embryonic stem cells by sampling from a negative binomial distribution [30, 31]. We show the TPR and FDR for KerTAD, deDoc, Armatus, and SpectralTAD.

https://doi.org/10.1371/journal.pcbi.1012221.g005

new method obtains a median TPR ≈ 0.98, while the next best TAD identification method, deDoc, on complex synthetic maps only had a median TPR ≈ 0.65. The remaining algorithms, Armatus and SpectralTAD, were roughly comparable in TPR performance with deDoc. For FDR, Armatus performed the best (median 0.01) followed by KerTAD (median 0.08). DeDoc and SpectralTAD had significantly higher FDRs with both greater than 0.45.

We also studied the impact of impulse noise on the calculations of TPR and FDR on complex, synthetic Hi-C maps. We find that our method is highly resistant to noise. In Fig 6A, we show that the mean TPR decays slowly with increasing χ, i.e. the mean TPR > 0.8 across all tested values of χ. In contrast, none of the other tested algorithms achieve a mean TPR of 0.70 or greater at any χ. In the regime $0.4 \leq \chi \leq 0.6$, FDR for DeDoc decreases compared to the results for less noisy Hi-C maps. This regime is likely a result of deDoc showing a unique sharp drop in the number of TAD predictions with increasing noise, taking a "safer" approach

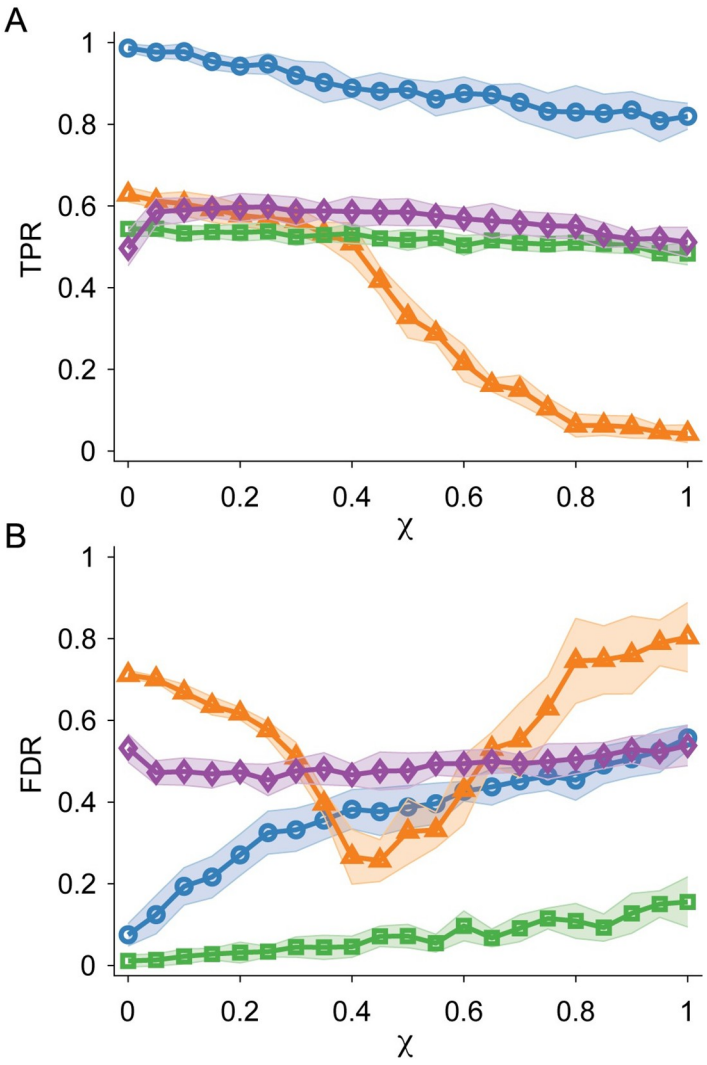

Fig 6. TPR and FDR for TAD prediction in Hi-C maps with added noise. A: TPR averaged over 200 complex synthetic Hi-C maps plotted versus the impulse noise fraction χ. We calculate TPR by comparing the ground truth of the synthetic Hi-C maps with the predicted TADs for KerTAD (blue circles), deDoc (orange triangles), Armatus (green squares), and SpectralTAD (purple diamonds). Shaded regions denote plus and minus one standard deviation about the mean given by the symbol. B: FDR plotted versus χ for the same data in A.

https://doi.org/10.1371/journal.pcbi.1012221.g006

to TAD identification with noisier Hi-C maps (i.e. for $\chi = 0.6$, deDoc makes an average of only 56 TAD predictions, whereas methods like SpectralTAD make on average over 210 TAD predictions). While this technique results in a lower FDR, it also results in fewer identified TADs (because of fewer predictions) and thus a lower TPR. In fact, deDoc drops much more rapidly in TPR over the same regime compared to other methods.

In addition, we investigated the effect of sparsity on the ability of TAD identification algorithms to predict TAD locations. To incorporate sparsity, we modify complex synthetic maps by randomly selecting elements in $\mathcal{A}_{ij}$ and replacing them with 0. In Fig 7A, we show that our method achieves a higher mean TPR at almost every ξ than all other tested TAD identification algorithms. We find that the mean TPR for KerTAD is significantly higher for the majority of ξ values tested; for example, our method achieves a higher mean TPR at $\xi = 0.5$ than the second

best algorithm, deDoc, at $\xi = 0$. The mean FDR for our method also grows more slowly compared to the other tested algorithms, only passing a mean FDR of 0.5 at large sparsity, $\xi > 0.6$. (See Fig 7B.) Notably, while KerTAD shows the largest TPR and smallest FDR, deDoc possesses the slowest rate of change in FDR below $\xi < 0.7$. This behavior is likely explained by the fact that the approach for identifying TADs by deDoc, which involves partitioning the graph generated by the Hi-C matrix based on minimal structural entropy, is relatively resistant to local sparsity. With higher sparsity, the deDoc TAD predictions reduce in size (with the average TAD size approaching 1 bin with higher ξ) and deDoc begins to predict the same one bin TADs for distinct maps. SpectralTAD generated error messages for large values of ξ and returned no predicted TADs (thus, for these maps we set TPR = 0 and FDR = 1). As a result, the error bars for SpectralTAD for these values of ξ (roughly between 0.3 and 0.6) are very large, since they include 0s for TPR and 1s for FDR.

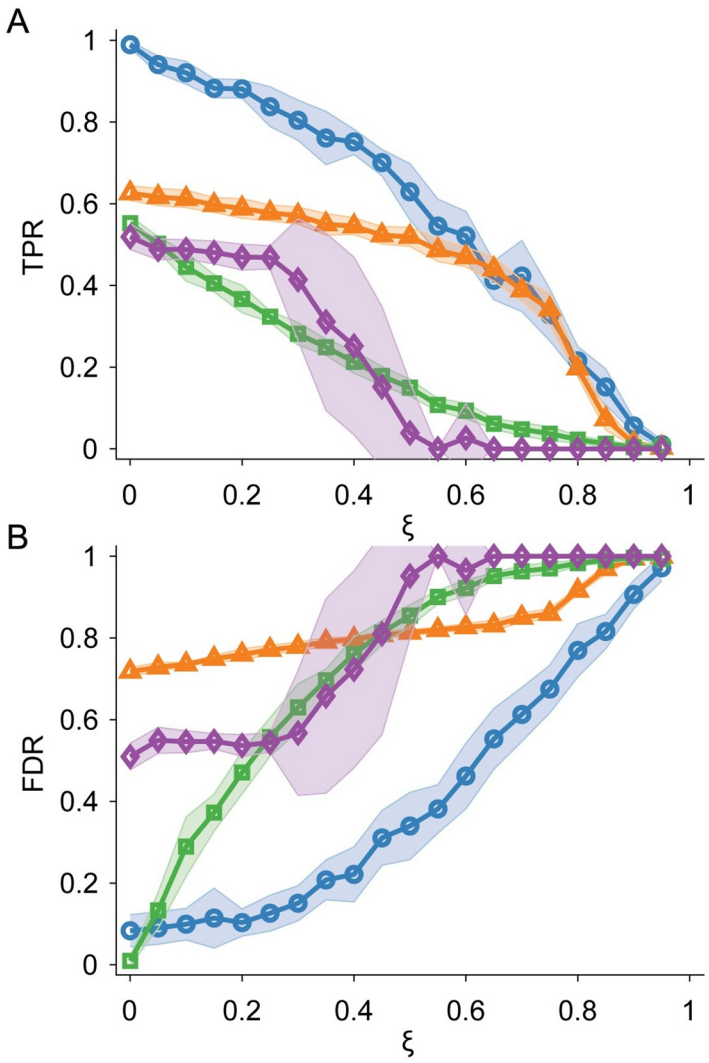

Fig 7. TPR and FDR for TAD prediction in Hi-C maps with added sparsity. A: TPR averaged over 200 complex synthetic Hi-C maps plotted versus the sparsity fraction ξ. We calculate TPR by comparing the ground truth of the synthetic Hi-C maps with the predicted TADs for the KerTAD (blue circles), deDoc (orange triangles), Armatus (green squares), and SpectralTAD (purple diamonds). Shaded regions denote plus and minus one standard deviation about the mean given by the symbol. B: FDR plotted versus ξ for the same data in A.

https://doi.org/10.1371/journal.pcbi.1012221.g007

In addition to assessing the performance of TAD identification algorithms on synthetic Hi-C maps, we also determined their performance on manually annotated Hi-C maps from the GM12878 cell line. We calculated TPR and FDR averaged over 10 chromosomes (chromosomes 2, 3, 4, 5, 6, 7, 12, 18, 20, and 22) by treating the manual annotations as the ground truth. We show in Fig 8 that our new method achieves a median TPR of nearly 0.80, while the next best performer, deDoc, obtains a median TPR of only $\sim$ 0.4. When using the original annotations, we also found that KerTAD outperformed the other techniques by a factor of ≈ 2 (KerTAD had a TPR of 0.4 while the next best, deDoc, had a TPR of 0.2). However, we were unable to precisely match the maps the original annotations used, with many annotated TADs pointing to no visible structure and hence the original annotation TPRs are likely not very meaningful. We do not include TopDom in the manually annotated comparisons as TopDom does not call nested or overlapping TADs. We find that TopDom achieves a median TPR of about 0.2 on the manually annotated Hi-C maps, which is surprisingly better than Armatus and near the performance of SpectralTAD, despite TopDom being unable to call nested and overlapping TADs.

Our new TAD identification method achieves a higher TPR and lower FDR on both simple and complex synthetic Hi-C maps, as well as on manually annotated experimental Hi-C maps. Additionally, our new method achieves and maintains the highest TPR in Hi-C maps with added noise and sparsity. Based on these results, we suggest that our method will have the highest accuracy of TAD identification on non-annotated experimental Hi-C maps. We compare the TAD predictions for the top-performing algorithms on synthetic Hi-C maps, manually annotated experimental Hi-C maps, as well as on non-annotated experimental Hi-C maps for four organisms: zebrafish, fruit fly, mouse, and human. In Fig 9A, we find that deDoc, TopDom, and our method predict different median total numbers of TADs (over the intrachromosomal Hi-C maps for all technical and biological replicates). For example, deDoc gives a median of 3370 TADs for zebrafish, while TopDom predicts roughly a factor of three fewer TADs. For zebrafish and fruit fly, we find that the fluctuations in the number of predicted

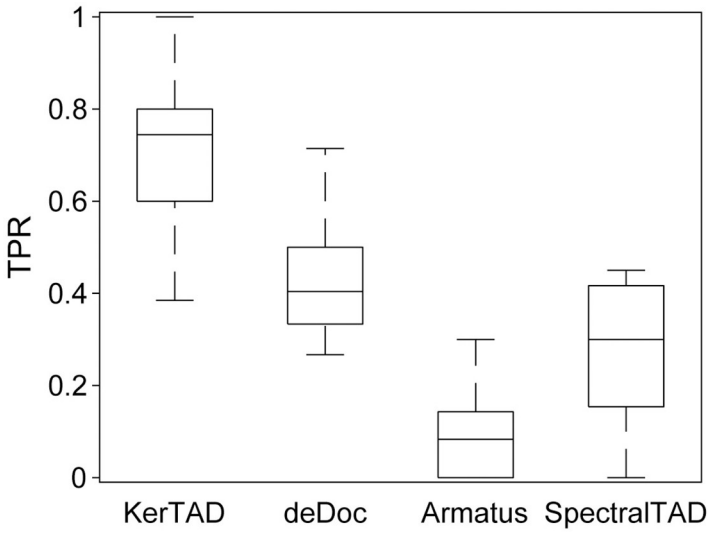

Fig 8. TPR on manually annotated Hi-C maps. A: Box plots of TPR calculated by comparing the ground truth TADs from ten manually annotated GM12878 Hi-C maps to those predicted by KerTAD, deDoc, Armatus, and SpectralTAD. The box edges represent the 25th and 75th percentiles in TPR and the central line in each box indicates the median. The error bars represent the maximum and minimum values of TPR

https://doi.org/10.1371/journal.pcbi.1012221.g008

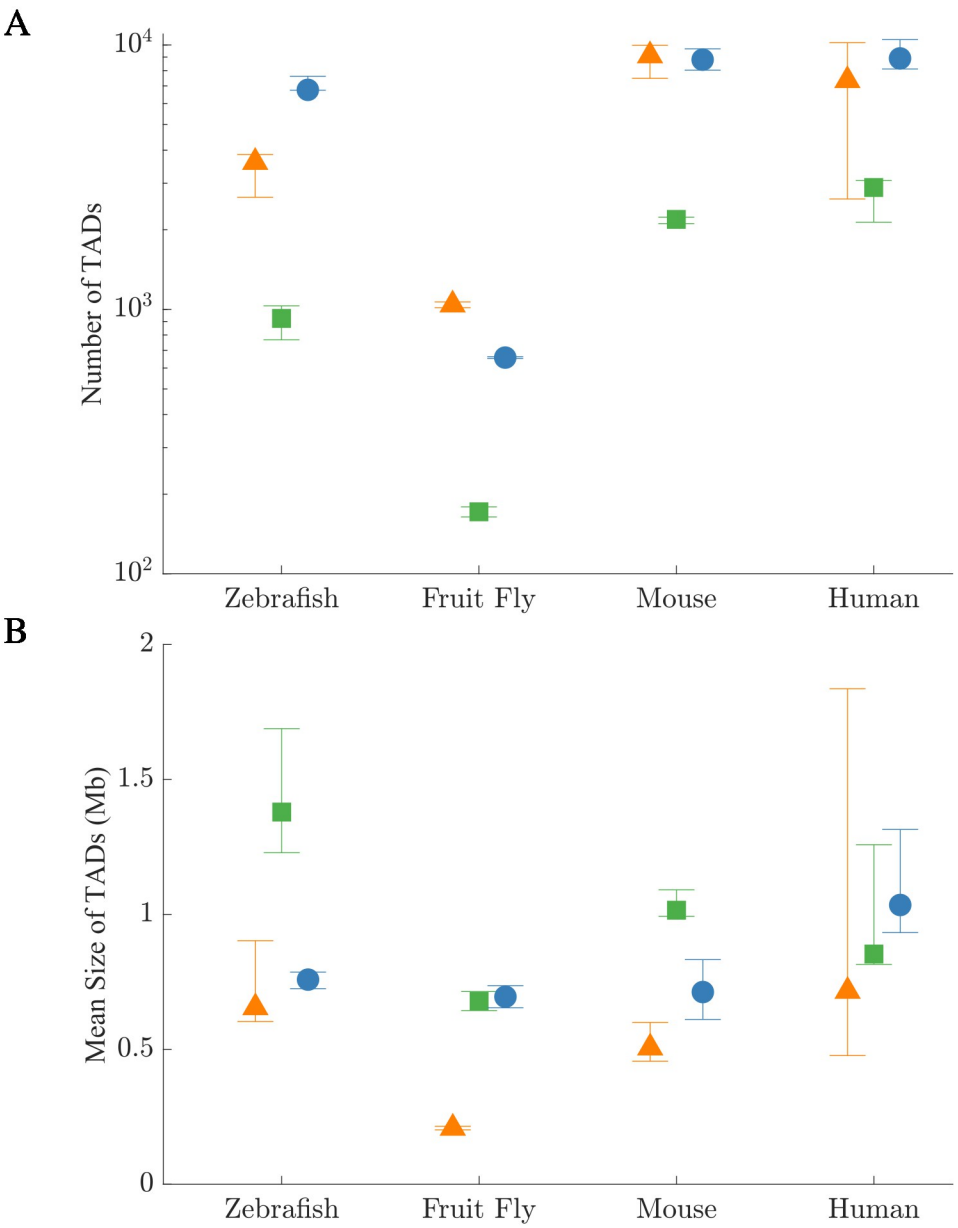

Fig 9. Number and size of TADs predicted across Hi-C map replicates from four organisms. A: The number of TADs predicted from whole-genome in situ Hi-C data for technical and biological replicates of four organisms (fruit fly, human, mouse, and zebrafish) using the KerTAD (blue circles), deDoc (orange triangles), and TopDom (green squares). The symbols indicate the median number of TADs and the error bars indicate the maximum and minimum values over replicates. B: Average size of the TADs in Mb for the same organisms, set of replicates, and TAD identification algorithms in A.

https://doi.org/10.1371/journal.pcbi.1012221.g009

TADs (given by the difference in the maximum and minimum values) over replicates for each TAD identification algorithm is smaller than the range in the median predictions between algorithms. Among the TAD identification methods tested, TopDom and our method have comparable variations in the number of TADs among replicates, while deDoc showed larger variations, especially for the human Hi-C maps. In Fig 9B, we show the predictions of the mean size of TADs identified by each algorithm. For the mouse and fruit fly Hi-C maps, we

find small variations among the methods on the mean size of TADs, while for zebrafish and human Hi-C maps there are large differences in the TAD sizes. For human Hi-C maps, our method and TopDom predict similar mean sizes for TADs (0.8–1.2 Mb), while deDoc shows large fluctuations in the sizes of TADs among replicates. (Note that the fluctuations in the TAD sizes over replicates obtained from our method and TopDom are comparable.). While KerTAD shows the smallest variation in mean sizes for human Hi-C maps, the range is significant ($\approx$ 0.4 Mb). This variation occurs partly because there are more human Hi-C maps in the dataset analyzed (384 total intrachromosomal Hi-C maps) compared to the other organisms (315 for mouse, 75 for zebrafish, and 14 for fruit fly) as well as the fact that there is more variation in coverage and sparsity across the human Hi-C maps than the other organisms (for instance, the mean element-wise range across across chromosome 1 for all human Hi-C maps in the dataset is $\approx$ 1.02 contacts per bin averaged for all maps whereas for mouse Hi-C maps it is $\approx$ 0.13 contacts per bin). In Fig 10A, 10B and 10C we show Hi-C maps with superimposed TAD predictions for different TAD identification algorithms. In Fig 10D, we show a non-annotated human lymphoblastoid Hi-C map with superimposed TAD predictions from KerTAD, deDoc, and TopDom. While there are some TADs for which all methods agree, we find large variability in the locations and number of predicted TADs.

Discussion

In this article, we developed a novel algorithm, KerTAD, to identify TADs in Hi-C maps. Most previous TAD calling algorithms assume simple Hi-C maps, i.e. each diagonal element of $\mathcal{A}_{ij}$ must belong to one and only one TAD. For simple Hi-C maps, when a TAD is identified at element i and j, the next TAD must have a starting index of $j + 1$ and there can be no additional TADs between i and j. In contrast, our method does not assume that Hi-C maps are simple and can accurately identify nested, overlapping, and gapped TADs. Among the few algorithms that can identify TADs in complex Hi-C maps, which is necessary for accurate TAD identification in experimental Hi-C maps, there is a large discrepancy in the number and size of TADs called, even among replicate Hi-C maps from the same experiment. Here, we present a novel algorithm that consistently outperforms other TAD identification algorithms on synthetic and manually annotated Hi-C maps, while being robust to noise and sparsity.

KerTAD uses two kernel-based techniques that detect complementary features of Hi-C maps. The method focuses on regions of Hi-C maps near the diagonal where there are large changes in intensity and strong corner points. We show that KerTAD outperforms six state-of-the-art TAD identification algorithms on both synthetic and manually annotated experimental Hi-C maps. In particular, we calculate the TPR and FDR by comparing the results for the predicted TADs for each algorithm to ground truth for the synthetic and experimental manually annotated Hi-C maps. We also test the performance of the TAD identification algorithms on complex, synthetic Hi-C maps with increasing levels of impulse noise and sparsity. For all of the Hi-C maps with ground truth that we tested (i.e. simple and complex synthetic, noisy and sparse, and manually annotated, experimental), our method has the highest TPR and negligible FDR.

We also find that our method has low variance in the median number and size of TADs across replicates for the experimental Hi-C maps without ground truth. In previous work [21, 31] that evaluated TAD identification algorithms, algorithms that can identify nested and overlapping TADs predict more TADs and possess higher variance in the number of identified TADs over replicates. This result is consistent with the fact that simple TAD identification algorithms can only call at most N TADs for a Hi-C map with $N \times N$ elements, whereas algorithms for complex Hi-C maps can identify at most N^2 TADs. Our results also show that

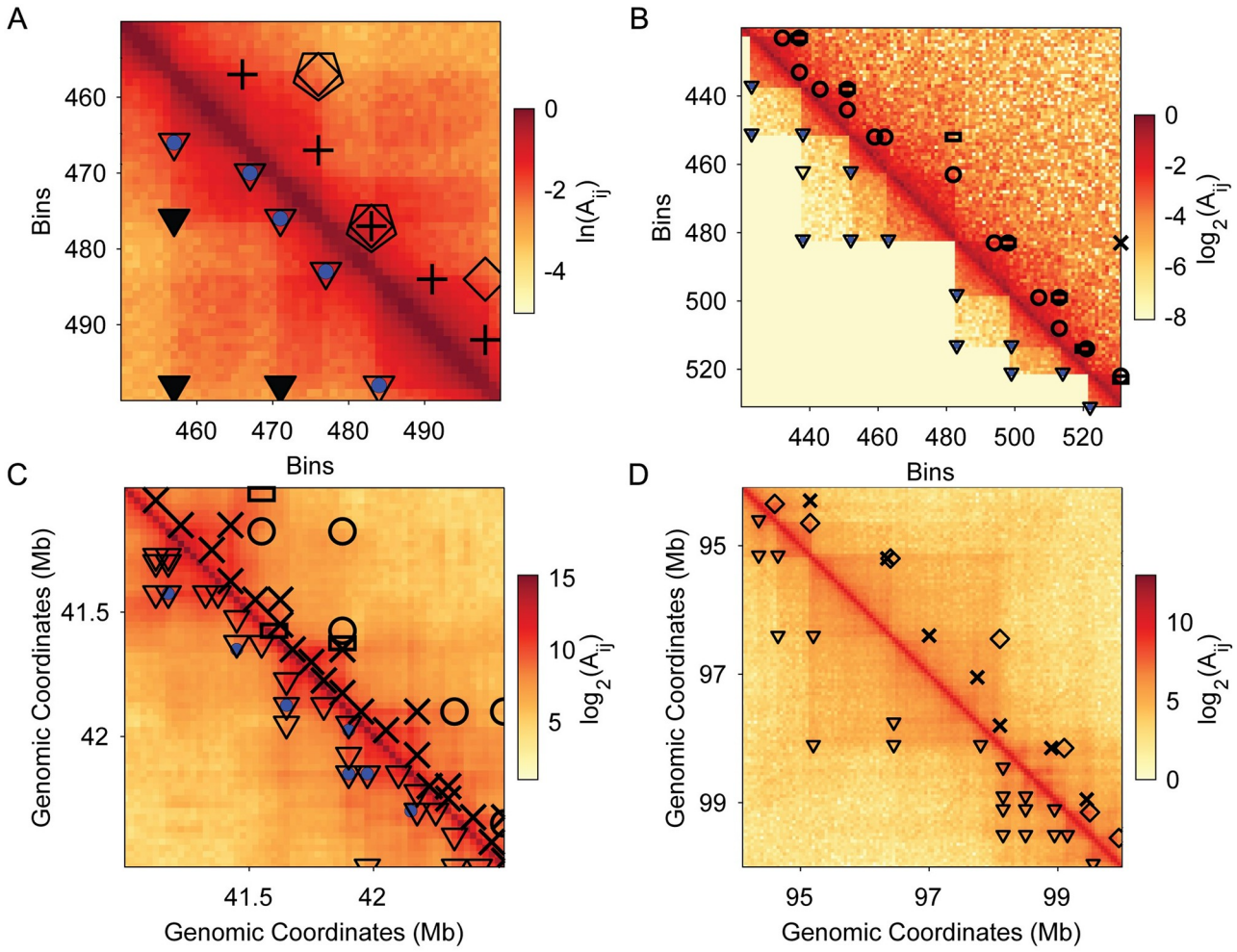

Fig 10. Demonstration of TAD predictions across four different types of Hi-C maps. A: Simple synthetic Hi-C map (on ln scale) with TAD predictions from the four tested TAD identification algorithms. We show the predicted TADs for KerTAD (open triangles) and ground truth (blue circle) in the lower triangular matrix. The predicted TADs for HiCSeg (open pentagons), TopDom (open diamonds), and CHDF (crosses) are shown in the upper triangular matrix. Gray triangles are examples of TADs that KerTAD identifies if the restriction of one TAD per row for simple maps is removed. B: Complex synthetic Hi-C map (on $\log_2$ scale) with added noise, $\chi = 1$. The upper triangular matrix shows the TAD predictions for deDoc (crosses), Armatus (open rectangles), and SpectralTAD (open circles). The lower triangular matrix shows the same Hi-C map with no added noise, $\chi = 0$, ground truth (blue circles), and the predictions of KerTAD on the noisy Hi-C map (we show the KerTAD predictions in the lower triangular matrix for better visibility). C: Manually annotated GM12878 chromosome 18 Hi-C map at 50kb resolution. Predictions from the same TAD identification algorithms in B are shown. D: TAD predictions on non-annotated chromosome 4 of human lymphoblastoid Hi-C map.

https://doi.org/10.1371/journal.pcbi.1012221.g010

algorithms for complex Hi-C maps identify more TADs than those for simple Hi-C maps, e.g. deDoc identifies significantly more TADs and with higher variance among replicates than TopDom. However, unlike deDoc, our method, which can identify TADs in complex Hi-C maps, shows significantly lower variation among replicates, with maximum and minimum values for the numbers and sizes of TADs comparable to those for TopDom. The fact that our method generates results for the numbers and sizes of TADs with small variations among replicates suggests that our method identifies the most important features of Hi-C maps that are insensitive to noise and sparsity.

While KerTAD outperforms other current TAD identification algorithms on synthetic Hi-C maps, it can be improved. For Hi-C maps where there are high-intensity regions

compared to the local neighborhood, we find that despite TVR reducing the variation, our method still tends to identify TADs in the regions of high intensity, rather than in regions of low intensity. Since TADs are usually defined *locally*, using global techniques that threshold across the whole Hi-C map will invariably suffer from this problem. Unfortunately, this results in a well-known dilemma: if one does not normalize weaker intensity regions, the algorithm will miss TADs, but normalizing weak intensity regions will bring out noise causing false positive TADs. This can be controlled to some degree by separating large maps into smaller ones (i.e. setting $\gamma = 1$), but risks "cutting off" TAD boundaries. In future work, we will develop new techniques to reduce noise, while maintaining the ability to identify TADs in weak intensity regions.

Because our method possesses the highest accuracy on synthetic and manually annotated experimental Hi-C maps, we hypothesize that our method will be accurate in capturing the true number and size of TADs in experimental Hi-C maps. However, it is worth reiterating that there is currently no ground truth definition of TADs in experimental Hi-C maps, which means that TPR and FDR on synthetic and manually annotated data, while useful, are only proxies for the accuracy of TAD identification algorithms on experimental Hi-C maps. Previous research groups [21, 30, 31, 33] have benchmarked their TAD identification algorithms using different metrics. For example, several studies have searched for correlations between predicted TAD boundaries and CTCF enrichment as a measure of TAD identification accuracy. However, this benchmark may not be related to benchmarks that rely on visual identification of TADs in experimental Hi-C maps.

Currently, there can be large variations in the experimentally determined Hi-C maps from one experiment to the next. As chromatin conformation capture experiments continue to improve, it will be possible to determine well-defined, relatively noise-free, and experimentally reproducible Hi-C maps. It is also important to understand how Hi-C maps depend on the phase of the cell cycle, cell type, cell-to-cell fluctuations, and tissue type in each organism. After such experimental studies are carried out and well-defined Hi-C maps are obtained, computational studies can be carried out to determine in an unsupervised way the important features that distinguish one Hi-C map from another. After identifying these key features, further studies can be carried out to understand the spatiotemporal dynamics of chromatin that give rise to each of the key features in Hi-C maps.

Supporting information

S1 Fig. Graphical description and examples of different types of TADs. We illustrate graphically (from left to right) nested, overlapping, and gapped TADs in the top row. In the bottom row, below each type of TAD, we show an example of that particular type of TAD (outlined using dotted lines) in an experimental Hi-C map of chromosome 6 from the human GM12878 cell line.
(TIF)

Author Contributions

Conceptualization: Luka Maisuradze, Mark D. Shattuck, Corey S. O'Hern.

Data curation: Luka Maisuradze.

Formal analysis: Luka Maisuradze.

Funding acquisition: Megan C. King, Simon G. J. Mochrie, Corey S. O'Hern.

Investigation: Luka Maisuradze.

Methodology: Luka Maisuradze.

Software: Luka Maisuradze.

Supervision: Mark D. Shattuck, Corey S. O'Hern.

Validation: Luka Maisuradze.

Visualization: Luka Maisuradze.

Writing – original draft: Luka Maisuradze.

Writing – review & editing: Luka Maisuradze, Megan C. King, Ivan V. Surovtsev, Simon G. J. Mochrie, Mark D. Shattuck, Corey S. O'Hern.

References

1. Boney B, Cavalli G. Organization and function of the 3D genome. Nat Rev Genet. 2016; 17: 661–678. https://doi.org/10.1038/nrg.2016.112
2. Bickmore WA, van Steensel B. Genome architecture: Domain organization of interphase chromosomes. Cell. 2013; 152(6):1270–84. https://doi.org/10.1016/j.cell.2013.02.001 PMID: 23498936
3. Dekker J, Marti-Renom MA, Mirny LA. Exploring the three-dimensional organization of genomes: Interpreting chromatin interaction data. Nat Rev Genet. 2013; 14: 390–403. https://doi.org/10.1038/nrg3454 PMID: 23657480
4. Yu M, Ren B. The three-dimensional organization of mammalian genomes. Annu Rev Cell Dev Biol. 2017; 33: 265–289. https://doi.org/10.1146/annurev-cellbio-100616-060531 PMID: 28783961
5. Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell. 2012; 148(1-2):84–98. https://doi.org/10.1016/j.cell.2011.12.014 PMID: 22265404
6. Therizols P, Illingworth RS, Courilleau C, Boyle S, Wood AJ, Bickmore WA. Chromatin decondensation is sufficient to alter nuclear organization in embryonic stem cells. Science. 2014; 346(6214):1238–42. https://doi.org/10.1126/science.1259587 PMID: 25477464
7. Lupiáñez DG, Spielmann M, Mundlos S. Breaking TADs: How alterations of chromatin domains result in disease. Trends Genet. 2016; 32: 225–237. https://doi.org/10.1016/j.tig.2016.01.003 PMID: 26862051
8. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009; 326(5950):289–93. https://doi.org/10.1126/science.1181369 PMID: 19815776
9. Hughes JR, Roberts N, McGowan S, Hay D, Giannoulatou E, Lynch M, et al. Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. Nat Genetic. 2014; 46(2):205–12. https://doi.org/10.1038/ng.2871
10. Dixon JR, 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: 376–380. https://doi.org/10.1038/nature11082 PMID: 22495300
11. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, et al. Three-dimensional folding and functional organization principles of the Drosophila genome. Cell. 2012; 148(3):458–472. https://doi.org/10.1016/j.cell.2012.01.010 PMID: 22265598
12. Sikorska N, Sexton T. Defining functionally relevant spatial chromatin domains: It is a TAD complicated. J Mol Biol. 2020; 432(3):7. https://doi.org/10.1016/j.jmb.2019.12.006 PMID: 31863747
13. Pope BD, Ryba T, Dileep V, Yue F, Wu W, Denas O. Topologically associating domains are stable units of replication-timing regulation. Nature. 2014; 515: 402–405. https://doi.org/10.1038/nature13986 PMID: 25409831
14. Dily FL, Baù D, Pohl A, Vicent GP, Serra F, Soronellas D. Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. Genes Dev. 2014; 28(19): 2151–2162. https://doi.org/10.1101/gad.241422.114 PMID: 25274727
15. Dekker J, Heard E. Structural and functional diversity of topologically associating domains. FEBS Lett. 2015; 589:2877–2884. https://doi.org/10.1016/j.febslet.2015.08.044 PMID: 26348399
16. Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, et al. Chromatin architecture reorganization during stem cell differentiation. Nature. 2015; 518: 331–336. https://doi.org/10.1038/nature14222 PMID: 25693564

17. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014; 159(7): 1665–1680. https://doi.org/10.1016/j.cell.2014.11.021 PMID: 25497547

18. de Wit E. TADs as the caller calls them. J Mol Biol. 2020; 432(3): 638–642. https://doi.org/10.1016/j.jmb.2019.09.026

19. Hansen AS, Cattoglio C, Darzacq X, Tjian R. Recent evidence that TADs and chromatin loops are dynamic structures. Nucleus. 2018; 9(1):20–32. https://doi.org/10.1080/19491034.2017.1389365 PMID: 29077530

20. Chang L, Ghosh S, Noordermeer D. TADs and their borders: Free movement or building a wall?. J Mol Biol. 2020; 432(3): 643–652. https://doi.org/10.1016/j.jmb.2019.11.025 PMID: 31887284

21. Dali R, Blanchette M. A critical assessment of topologically associating domain tools. Nucleic Acids Res. 2017; 45(6):2994–3005. https://doi.org/10.1093/nar/gkx145 PMID: 28334773

22. Filippova D, Patro R, Duggal G, Kingsford C. Identification of alternative topological domains in chromatin. Algorithms Mol Biol. 2014; 9:14. https://doi.org/10.1186/1748-7188-9-14 PMID: 24868242

23. Zhan Y, Mariani L, Barozzi I, Schulz EG, Blüthgen N, Stadler M, et al. Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes. Genome Res. 2017; 27(3): 479–490. https://doi.org/10.1101/gr.212803.116 PMID: 28057745

24. Li A, Yin X, Xu B, Wang D, Han J, Wei Yi, et al. Decoding topologically associating domains with ultra-low resolution Hi-C data by graph structural entropy. Nat Commun. 2018; 9:3265. https://doi.org/10.1038/s41467-018-05691-7 PMID: 30111883

25. Cresswell KG, Stansfield JC, Dozmorov MG. SpectralTAD: An R package for defining a hierarchy of topologically associated domains using spectral clustering. BMC Bionformatics. 2020; 21:319. https://doi.org/10.1186/s12859-020-03652-w

26. Serra F, Baù D, Goodstadt M, Castillo D, Filion GJ, Marti-Renom MA. Automatic analysis and 3D-modelling of Hi-C data using TADBit reveals structural features of the fly chromatin colors. PLoS Comput Biol. 2017; 13(7):1005665. https://doi.org/10.1371/journal.pcbi.1005665 PMID: 28723903

27. Shin H, Shi Y, Dai C, Tjong H, Gong K, Alber F. TopDom: An efficient and deterministic method for identifying topological domains in genomes. Nucleic Acid Res. 2016; 44:1505. https://doi.org/10.1093/nar/gkv1505 PMID: 26704975

28. Lévy-Leduc C, Delattre M, Mary-Huard T, Robin S. Two-dimensional segmentation for analyzing Hi-C data. Bioinformatics. 2014; 1; 30(17):i386–92. https://doi.org/10.1093/bioinformatics/btu443 PMID: 25161224

29. Wang Y, Li Y, Gao J, Zhang MQ. A novel method to identify topological domains using Hi-C data. Quant Biol. 2015; 3: 81–89. https://doi.org/10.1007/s40484-015-0047-9

30. Forcato M, Nicoletti C, Pal K, Livi CM, Ferrari F, Bicciato S. Comparison of computational methods for Hi-C data analysis. Nat Methods. 2017; 14: 679–685. https://doi.org/10.1038/nmeth.4325 PMID: 28604721

31. Liu K, Li H, Li Y, Wang J, Wang J. A comparison of topologically associating domain callers based on Hi-C data. IEEE/ACM Trans Comput Biol Bioinform. 2023; 20(1): 15–29. PMID: 35104223

32. Zufferey M, Tavernari D, Oricchio E, Ciriello G. Comparison of computational methods for the identification of topologically associating domains. Genome Biology. 2018; 19:217. https://doi.org/10.1186/s13059-018-1596-9 PMID: 30526631

33. Sefer E. A comparison of topologically associating domain callers over mammals at high resolutions. BMC Bioinformatics. 2022; 23:127. https://doi.org/10.1186/s12859-022-04674-2 PMID: 35413815

34. Lyu H, Liu E, Wu Z. Comparison of normalization methods for Hi-C data. Biotechniques. 2020; 68(2): 56–64. https://doi.org/10.2144/btn-2019-0105 PMID: 31588782

35. Hu M, Deng K, Selvaraj S, Qin ZH, Ren B, Liu JS. HiCNorm: Removing biases in Hi-C data via Poisson regression. Bioinformatics. 2012; 28(23): 3131–3133. https://doi.org/10.1093/bioinformatics/bts570 PMID: 23023982

36. Schmitt AD, Hu M, Ren B. Genome-wide mapping and analysis of chromosome architecture. Nat Rev Mol. Cell Biol. 2016; 17(12): 743–755. https://doi.org/10.1038/nrm.2016.104 PMID: 27580841

37. Cournac A, Marie-Nelly H, Marbouty M, Koszul R, Mozziconacci J. Normalization of a chromosomal contact map. BMC Genomics. 2012; 13(1): 436. https://doi.org/10.1186/1471-2164-13-436 PMID: 22935139

38. Imakaev M, Fudenberg G, Mccord RP, Naumova N, Goloborodko A, Lajoie BR, et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods. 2012; 9(10): 999–1003. https://doi.org/10.1038/nmeth.2148 PMID: 22941365

39. Knight PA, Ruiz D. A fast algorithm for matrix balancing. IMA J. Numer. Anal. 2013; 33(3): 1029–1047. https://doi.org/10.1093/imanum/drs019

40. Shavit Y, Lio P. Combining a wavelet change point and the Bayes factor for analysing chromosomal interaction data. Mol Biosyst. 2014; 10(6): 1576–1585. https://doi.org/10.1039/C4MB00142G PMID: 24710657

41. Grubbs F. Sample Criteria for Testing Outlying Observations. Ann. Math. Stat. 1950; 21(1): 27–58. https://doi.org/10.1214/aoms/1177729885

42. Zack GW, Rogers WE, Latt SA. Automatic measurement of sister chromatid exchange frequency. J Histochem Cytochem. 1977; 25(7):741–53. https://doi.org/10.1177/25.7.70454 PMID: 70454

43. Haddad N, Vaillant C, Jost D. IC-Finder: Inferring robustly the hierarchical organization of chromatin folding. Nucleic Acids Res. 2017; 45(10): e81. https://doi.org/10.1093/nar/gkx036 PMID: 28130423

44. Lun ATL, Smyth GK. diffHic: A Bioconductor package to detect differential genomic interactions in Hi-C data. BMC Bioinformatics. 2015; 16: 258. https://doi.org/10.1186/s12859-015-0683-0 PMID: 26283514

45. Ray J, Munn PR, Vihervaara A, Lewis JJ, Ozer A, Danko CG, et al. Chromatin conformation remains stable upon extensive transcriptional changes driven by heat shock. PNAS. 2019; 116(39): 19431–19439. https://doi.org/10.1073/pnas.1901244116 PMID: 31506350

46. Wike CL, Guo Y, Tan M, Nakamura R, Shaw DK, Díaz N, et al. Chromatin architecture transitions from zebrafish sperm through early embryogenesis. Genome Res. 2021; 31(6): 981–994. https://doi.org/10.1101/gr.269860.120 PMID: 34006569

47. Rao SSP, Huang SC, St Hilaire BG, Engreitz JM, Perez EM, Kieffer-Kwon KR, et al. Cohesin loss eliminates all loop domains. Cell. 2017; 171(2): 305–320. https://doi.org/10.1016/j.cell.2017.09.026 PMID: 28985562

48. Dekker J, Belmont AS, Guttman M, Leshyk VO, Lis JT, Lomvardas S, et al. 4D Nucleome Network. The 4D nucleome project. Nature. 2017; 549(7671):219–226. https://doi.org/10.1038/nature23884 PMID: 28905911

49. Abdennur N, Mirny LA. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. Bioinformatics. 2020; 36(1): 311–316. https://doi.org/10.1093/bioinformatics/btz540 PMID: 31290943