

MSTD: an efficient method for detecting multi-scale topological domains from symmetric and asymmetric 3D genomic maps

Yusen Ye¹, Lin Gao^{1,*} and Shihua Zhang^{2,3,4,*}

¹School of Computer Science and Technology, Xidian University, Xi'an 710071, Shaanxi, China, ²NCMIS, CEMS, RCSDS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China, ³School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China and ⁴Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China

Received January 29, 2019; Revised March 06, 2019; Editorial Decision March 14, 2019; Accepted March 16, 2019

ABSTRACT

The chromosome conformation capture (3C) technique and its variants have been employed to reveal the existence of a hierarchy of structures in three-dimensional (3D) chromosomal architecture, including compartments, topologically associating domains (TADs), sub-TADs and chromatin loops. However, existing methods for domain detection were only designed based on symmetric Hi-C maps, ignoring long-range interaction structures between domains. To this end, we proposed a generic and efficient method to identify multi-scale topological domains (MSTD), including *cis*- and *trans*-interacting regions, from a variety of 3D genomic datasets. We first applied MSTD to detect promoter-anchored interaction domains (PADs) from promoter capture Hi-C datasets across 17 primary blood cell types. The boundaries of PADs are significantly enriched with one or the combination of multiple epigenetic factors. Moreover, PADs between functionally similar cell types are significantly conserved in terms of domain regions and expression states. Cell type-specific PADs involve in distinct cell type-specific activities and regulatory events by dynamic interactions within them. We also employed MSTD to define multi-scale domains from typical symmetric Hi-C datasets and illustrated its distinct superiority to the state-of-art methods in terms of accuracy, flexibility and efficiency.

INTRODUCTION

Folding of mammalian chromosomes into the nucleus has increasingly been recognized as an important factor in

gene regulation, cell fate decisions, and so on (1–3). However, how chromosomes fold into the nucleus is still obscure. The chromosome conformation capture (3C) technique and its variants such as Hi-C, ChIA-PET and capture Hi-C have been employed to uncover the chromatin loops and hierarchical chromatin structural domains in three-dimensional (3D) genome architecture. Specifically, recent studies have revealed direct physical interactions, such as long-range chromatin contacts between enhancer and target genes (4), actively co-regulated genes (5) and Polycomb-repressed genes (6). In addition, researchers have provided evidence for topologically associating domains (TADs) or sub-TADs (7–12). Furthermore, interactions between TADs at a variable distance result in active and inactive compartments, which are further subdivided into six different sub-compartments according to distinct patterns of histone modifications (4,13). Therefore, with the rapid accumulation of 3D genomic maps, developing efficient computational methods for detecting topological domains in chromosomal architecture is urgently needed.

Several methods have been developed to address these issues. For example, Dixon *et al.* adopted a hidden Markov model (HMM) method based on the directionality index from Hi-C maps to detect TADs (7). Lévy-Leduc *et al.* proposed a block-wise segmentation model for TAD detection and proved that maximization of the likelihood on the block boundaries is a one-dimensional (1D) segmentation problem (14). Similarly, Crane *et al.* developed an approach to transform the Hi-C contact matrix into an 1D insulation score vector for detecting topological structures (15). Shin *et al.* employed an efficient and deterministic method to systematically identify TADs with a set of statistical measures to evaluate their quality (16). However, these methods were only designed for detecting single-scale domains.

Recently, a few methods have been designed to explore the hierarchical organization of chromosomal architecture from symmetric Hi-C maps. For example, Filippova *et al.*

*To whom correspondence should be addressed. Tel: +86 10 82541360; Fax: +86 10 82541360; Email: zsh@amss.ac.cn
Correspondence should be addressed to Lin Gao. Email: lgao@mail.xidian.edu.cn

introduced a dynamic programming model to identify hierarchical domains by adjusting a single parameter (17). Weinreb *et al.* proposed a matrix decomposition model to infer a hierarchy of nested TADs based on ideal empirical distributions (18). Zhan *et al.* presented a method CaTCH to detect hierarchical trees of chromosomal domains given a certain degree of reciprocal insulation from Hi-C maps (19). However, these methods can only detect *cis*-interacting regions from Hi-C maps.

Very recent studies have revealed that both *cis*- and *trans*-interacting DNA regions exist in the cell nucleus (20). With the emerging higher resolution genome-wide interaction maps of chromatin such as promoter capture Hi-C maps (21) and ChIA-PET maps (22), many distal DNA fragments regulate their targets bypassing long chromatin regions have been observed. Therefore, how to develop a generic method to infer both *cis*- and *trans*-interacting regions from diverse types of 3D genomic data is still a grand challenge in computational biology.

To this end, we proposed a generic and efficient method to identify **multi-scale topological domains** (MSTD) from both asymmetric and symmetric 3D genomic datasets with a single adjustable parameter controlling domain scales. MSTD can detect long-range interacting domains such as those between promoters and regulatory elements at variable distances, which can not be addressed directly by existing methods. We detected promoter-anchored interacting domains (PADs) from promoter capture Hi-C datasets across 17 primary blood cell types. We observed that the boundaries of these PADs can be well specified by one or the combination of a few epigenetic factors, indicating that these factors should play a key role in the formation of genomic conformation. The analysis of the affinity relationship among cell types showed that PADs between functionally similar cell types have distinctly high conservation and consistent expression levels. Furthermore, dynamic PADs might perform specific cellular functions, while common ones take into account the underlying conditions of regular cell activities and participate in cell-specific regulatory events by dynamic interactions within each PAD. This suggests that PADs are important and basic units of genomic structure and function influencing gene regulation and cellular differentiation. We also employed MSTD to define multi-scale domains from symmetric Hi-C datasets with distinctly superior accuracy and efficiency compared to existing methods. Interestingly, the conservation of TADs between cells during continuous differentiation cycles is significantly higher than those between interval cycles. Last but not least, TADs are strongly correlated with epigenetic and transcriptional features.

MATERIALS AND METHODS

Materials

We first downloaded a comprehensive catalog of capture Hi-C datasets of 31 253 annotated promoters and 230 525 unique promoter-interacting regions (PIRs) in 17 human primary blood cell types (21). The datasets detect 698 187 high-confidence unique promoter interactions, of which 9.6% are promoter-promoter interactions and 90.4% are promoter-PIR interactions. The 17 cell types

are from different nodes of the hematopoietic tree, which can be roughly divided into two categories, including eight myeloid cell types (megakaryocytes (MK), erythroblasts (Ery), neutrophils (Neu), monocytes (Mon), endothelial precursors (Endp), macrophages M0 (Mac0), macrophages M1 (Mac1), macrophages M2 (Mac2)) and nine lymphoid cell types (naive B cells (nB), total B cells (tB), fetal thymus (FetT), Naive CD4⁺ T cells (nCD4), Total CD4⁺ T cells (tCD4), non-activated total CD4⁺ T cells (naCD4), activated total CD4⁺ T cells (aCD4), naive CD8⁺ T cells (nCD8), total CD8⁺ T cells (tCD8)). Interaction scores were computed for each fragment pairs between promoters and promoter interacting regions (PIRs). As described by (21), the scores were computed by the CHiCAGO pipeline for each cell types. We collected available histone modification ChIP-seq datasets, including H3K4me3, H3K4me1, H3K27ac, H3K36me3, H3K9me3, H3K27me3, H3K9ac and DNase-seq datasets for seven cell types (Neu, Mon, FetT, nCD4, tCD4, nCD8, tCD8) from the Roadmap Epigenetics project (http://egg2.wustl.edu/roadmap/web_portal/). In addition, we collected available RNA-seq datasets for 8 cell types (MK, Ery, Neu, Mon, Mac0, Mac1, Mac2, nCD4) from (21).

We also collected two sets of Hi-C maps. The first dataset includes mouse cortex, ES and human ES cell types binned at 40 kb resolution (7). For mouse cortex and ES cell types, we collected ChIP-seq datasets from the ENCODE project (<http://www.genome.ucsc.edu/ENCODE/>), including the architectural protein CTCF, seven histone modifications (H3K4me3, H3K9ac, H3K36me3, H3K4me1, H3K27ac, H3K27me3, H3K9me3), RNA Polymerase II, EP300, DNase-seq and RNA-seq data from (23). For human ES cell type, we obtained human housekeeping genes from (24) and TSS of protein coding genes of hg19 RefSeq in GENCODE database. The second dataset contains proliferating mouse embryonic stem cells (ESC), intermediate neuronal precursor cells (NPC) and post-mitotic neurons (neurons) Hi-C datasets through different differentiation periods, binned at 50Kb resolution from (25). We normalized these Hi-C maps as previously described by Yaffe and Tanay (26).

Definition of chromatin domains for Hi-C and promoter capture Hi-C datasets

The whole-genome Hi-C generates the chromatin interactions from all genomic regions to all genomic regions, while promoter capture Hi-C generates the chromatin interactions from promoters to a set of genomic regions. If the data are expressed in chromatin interaction matrixes, the rows and columns of the interactions matrixes from the whole-genome Hi-C are the same, while the rows and columns of the interactions matrixes from promoter capture Hi-C are different (the rows could be the promoter regions, and the columns could be any genomic regions).

The whole-genome Hi-C maps and the promoter capture Hi-C maps can be represented as symmetric and asymmetric data matrices respectively. Here, the 'asymmetric' 3D genomics maps mean the latter case. Specifically, a promoter capture Hi-C data matrix can be considered as an asymmetric submatrix sampled from a corresponding Hi-C data ma-

trix about all promoters. This results in a rectangle map in which the rows are promoters and the columns are promoter interacting regions respectively. Existing methods designed for TADs detection cannot be applied to such asymmetric maps. In contrast, MSTD can identify topological domains from both asymmetric and symmetric 3D genomic datasets.

Generally, Hi-C maps are summarized as a symmetric matrix A with bins of a fixed width (e.g. 40 kb). Considering the symmetry of Hi-C matrix, in which $A_{i,j}$ ($1 \leq i \leq n, 1 \leq j \leq n$) represents the interacting frequencies between bins i and j , we first define the diagonal blocks D_k of high intensity with relative strong interactions as the topologically associating domains (TADs) (Figure 1A),

$$D_k = \{(i, j) : b_{2k-1} \leq i \leq b_{2k}, b_{2k-1} \leq j \leq b_{2k}\},$$

$$(k = 1, \dots, K),$$

where b_{2k-1}, b_{2k} ($1 \leq b_{2k-1} < b_{2k} \leq n$) represent the true TAD boundaries of the k_{th} TAD and K is the number of TADs. We further define the non-diagonal blocks N_l as pairwise topologically associating domains (PTADs) (Figure 1B),

$$N_l = \{(i, j) : b_{2l-1}^1 \leq i \leq b_{2l}^1, b_{2l-1}^2 \leq j \leq b_{2l}^2\},$$

$$(l = 1, \dots, L),$$

where b_{2l-1}^1, b_{2l}^1 ($1 \leq b_{2l-1}^1 < b_{2l}^1 \leq n$) and b_{2l-1}^2, b_{2l}^2 ($1 \leq b_{2l-1}^2 < b_{2l}^2 \leq n$) denote the boundaries of two distinct chromatin regions forming the l_{th} PTAD and L is the number of PTADs. In this study, we denote promoter capture Hi-C map as an asymmetric rectangular contact matrix B , in which $B_{i,j}$ ($1 \leq i \leq n, 1 \leq j \leq m$) represents interaction frequencies between promoters and PIRs. As a special type of PTADs, we define the special non-diagonal blocks S_x as promoter-anchored interacting domains (PADs) from promoter capture Hi-C maps (Supplementary Figure S1A),

$$S_x = \{(i, j) : b_{2x-1}^1 \leq i \leq b_{2x}^1, b_{2x-1}^2 \leq j \leq b_{2x}^2\},$$

$$(x = 1, \dots, X),$$

where b_{2x-1}^1, b_{2x}^1 ($1 \leq b_{2x-1}^1 < b_{2x}^1 \leq n$) and b_{2x-1}^2, b_{2x}^2 ($1 \leq b_{2x-1}^2 < b_{2x}^2 \leq m$) represent respectively the boundaries of promoters and PIRs forming the x_{th} PAD and X is the number of PADs.

MSTD

MSTD was inspired by a fast density-based clustering method, which is designed for grouping data points (27). Here, we group the strong contact interactions, which form square or rectangle submatrices corresponding to domains defined as above (Figure 1). Given a symmetric or asymmetric matrix A , where $A_{i,j}$ ($1 \leq i \leq n, 1 \leq j \leq m$) denotes the interaction frequencies between bin i and j . Firstly, for each element (i, j) (a diagonal element (i.e., $i = j$) for a symmetric matrix in Figure 1C or a non-zero element for an asymmetric matrix in Figure 1E), MSTD computes two indexes: (1) local density $\rho_{ij}(w)$ defined as the average of the interaction frequencies with a cutoff radius window w (Supplementary Figure S1B, C and Supplementary Methods), and

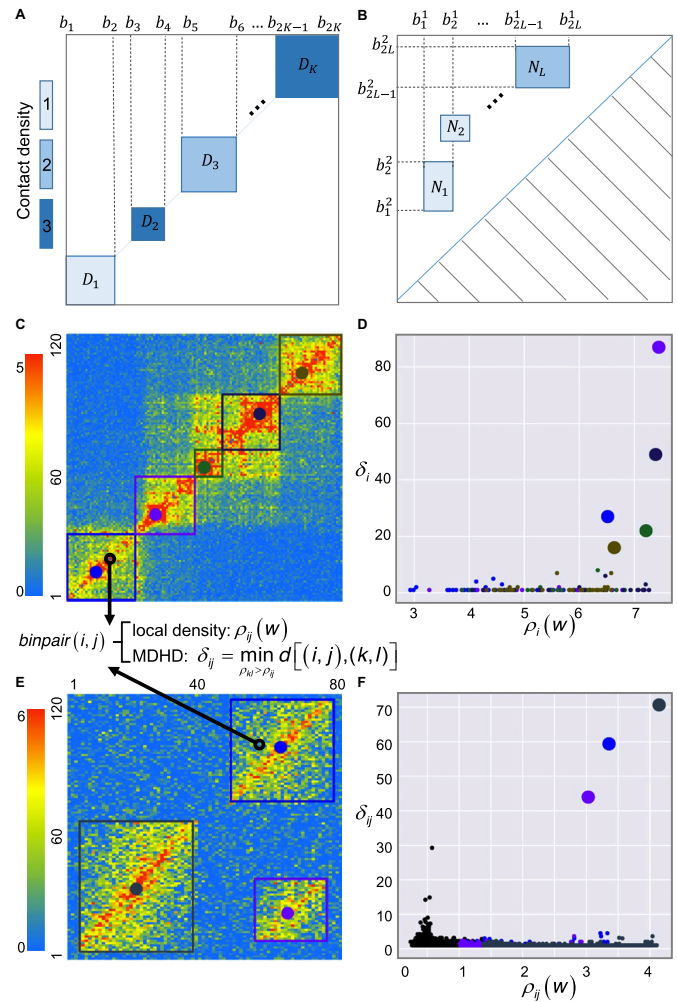


Figure 1. Illustration of MSTD for detecting TADs and PTADs in a symmetric and asymmetric matrix respectively. (A, B) Examples of TADs $\{D_1, D_2, \dots, D_K\}$ and PTADs $\{N_1, N_2, \dots, N_L\}$. The color depth represents the density of domains. (C) MSTD computes two indexes including local density and minimum distance of higher density (MDHD) for each diagonal element k in the symmetric matrix. The value of heatmap represents normalized interacting scores. (D) The decision graph of clustering results for (C). Clustering centers are marked by big dots with different colors. Starting with the centers, the remaining elements are assigned to the same cluster as its nearest neighbor element of higher density layer by layer. The clustering boundaries are defined to be the outermost elements of the same cluster in the two directions. (E) MSTD computes two indexes similar to (C) for each non-zero element k in the asymmetric matrix. (F) The decision graph of clustering results for (E). For the asymmetric matrix, MSTD employs all non-zero elements and defines the clustering boundaries in four directions. Noise elements were colored by black.

(2) δ_{ij} is computing by minimum distance between the element (i, j) and any other element with a higher local density (MDHD),

$$\delta_{ij} = \min_{\rho_{kl} > \rho_{ij}} d[(i, j), (k, l)], \quad (\delta_{ij} \leq Max),$$

where $d[(i, j), (k, l)]$ is the distance between the element (i, j) and element (k, l) . For the efficiency of the algorithm, We search for a nearest neighbor element of higher density of the element (i, j) in the range of Max (100 is used as the default value), which is not sensitive to the results (Supple-

mentary Materials). Note that local density ρ_{ij} of the clustering center element is locally or globally maximum, and its δ_{ij} are much larger than those of its neighbors. Thus, MSTD determined clustering centers as elements for which the value δ_{ij} is anomalously large, and the scale of domain is control by an adjustable parameter δ_T ($\delta_{ij} > \delta_T$). Next, the remaining elements are assigned to the same cluster as its nearest neighbor element of higher density layer by layer. Meanwhile, individual outlier elements with high δ and low ρ (Figure 1D) and the elements associated with these outlier elements will not be assigned to any cluster. Finally, the boundaries for each cluster are defined as the outermost elements of the same cluster in all directions. If the elements whose local density is significantly lower than their centers' ($\delta_{ij} < 0.2 \times \delta_{center}$), this element is removed from this cluster.

In summary, MSTD can identify multi-scale topological domains from both asymmetric and symmetric datasets with a single adjustable parameter controlling domain scales. As the adjustable parameter increases, the number of clusters decreases and the size increases. This process furthermore controls the scale of domains, which together form the hierarchical chromatin structure.

Definition of high (low) co-expressed PADs

We first defined high (low) expression genes for a given cell type if their expression level is higher (lower) than their corresponding median of gene expression level across all available cell types. We defined a PAD to be *high-* (*low-*) co-expressed or not by evaluating the difference between the number of high (low) expression genes in the genome region of this PAD, and the average number of high (low) expression genes in cyclically permuted genomes with the same number of genes, which is weighted by its standard deviation (P -value < 0.05). Every PAD obtains a z-score value to describe the level of co-expression.

Parameter selection for MSTD and other methods

In order to compare MSTD with three popular methods including DI (7), TopDom (16) and HicSeg (14) for identifying TADs, we next applied these four methods onto Hi-C datasets of mouse embryonic stem cells (ESC), intermediate neuronal precursor cells (NPC) and post-mitotic neurons (Neurons) through different differentiation periods (ESC-NPC-Neurons), binned at 50 Kb resolution (25). Since DI has been generally accepted to detect TADs in a number of existing studies, here we regard the size range of its TADs (from 700 to 1200 kb) as the standard scale. In order to get a similar domain scale for a fair comparison, we set the adjustable parameters window = 12 for TopDom, and $\delta_T = 7$ for MSTD, respectively. For HicSeg, we set the maximum number of boundary points (Maxk = chromosome length/850Kb) and data distribution (distrib = 'G') respectively. Based on the above parameter settings, all four methods identify TADs with nearly identical domain scales (Supplementary Table S1).

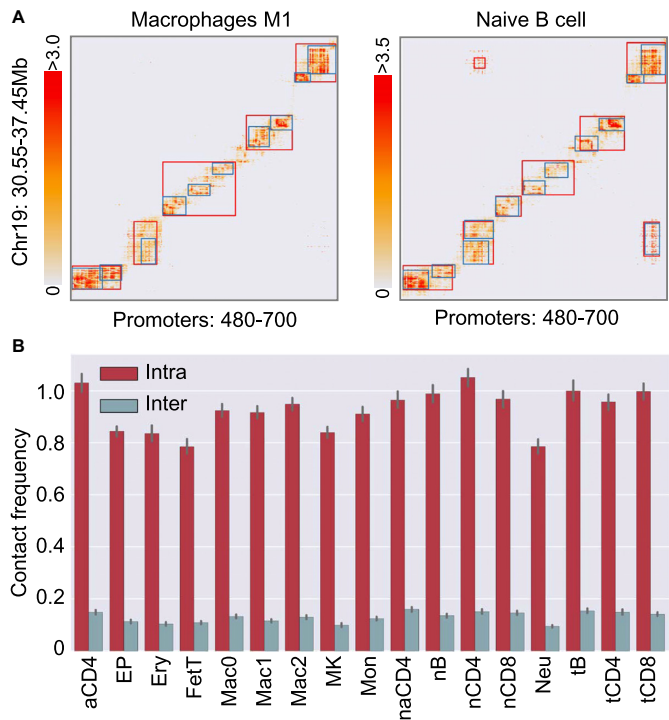


Figure 2. Illustration of PADs detected from promoter capture Hi-C datasets. (A) Examples of PADs detected by MSTD with $\delta_T = 100$ (blue) and $\delta_T = 200$ (red). The value of heatmaps represents CHiCAGO score. (B) Comparison of the average contact frequency of *intra*-PADs and *inter*-PADs in 17 human primary blood cell types with $\delta_T = 100$.

The statistical measurements for TADs quality

We compared the four methods using two quality measurements: the average Pearson's correlation coefficient (PCC) of chromosome-wide contact profile of m bins within the TAD D_k , which is defined as $PCC(D_k) = \frac{2}{n(n-1)} \sum_{i=1}^m \sum_{j=i+1}^m corr(D_k^{(i)}, D_k^{(j)})$, and the differ-

ence of average interaction frequency between intra-domain and the corresponding inter-domain (DIFF), which is defined as $DIFF(D_k) = Intra(D_k) - Inter(D_k, D_{k+1})$, where $Intra(D_k)$ denotes the average of interaction frequency between bins within the same TAD D_k and $Inter(D_k, D_{k+1})$ denotes the average of interaction frequency between bins in TAD D_k and bins in adjacent TAD D_{k+1} (16).

RESULTS

Identifying PADs from asymmetric promoter capture Hi-C datasets

Firstly, we applied MSTD to identify multi-scale PADs from promoter capture Hi-C maps across 17 human primary blood cell types (Figure 2A and Supplementary Methods) (21). MSTD can not only find *cis*-interacting PADs within the continuous genome, but also identify *trans*-interacting PADs between distal regulatory elements and target promoters (Figure 2A). The results below are based

on parameter $\delta_T = 100$ for detecting PADs, unless noted otherwise.

MSTD costs only about 7.3 minutes for each chromosome (Supplementary Table S2). Also, the size of the most of PADs detected by MSTD is in the range between 200Kb and 2Mb (Supplementary Figure S2A), as suggested by TADs from symmetric Hi-C maps in previous studies (8,15). Moreover, we noted that the same chromosome between functionally similar cell types obtains very similar number of PADs (Supplementary Figure S2B), and average interaction frequency of *intra*-PADs is significantly higher than that of *inter*-PADs across 17 human primary blood cell types (Figure 2B), indicating that PADs relate to cell activities distinctly.

Epigenetic feature is a good predictor of PADs

We only defined the boundaries and centers of PADs along promoter interacting regions (PIRs) axis not promoters' axis (Materials and Methods) and found that epigenetic features of PADs show strong difference on these PIRs' regions compared with random ones (Figure 3, Supplementary Figures S3–S5). Specifically, for monocytes cell type, the promoter mark H3K4me3 is strongly enriched in the boundaries of PADs (even more than twice that of *intra*-PADs) and the transcribed region mark (H3K36me3) is enriched in the regions slightly deviating from boundary regions, indicating that the boundary elements of PADs are significantly associated with promoters (Figure 3A). Meanwhile, the enhancer marks (H3K4me1, H3K27ac) are also enriched around topological boundaries, which might be related with chromatin loops in the boundary regions (Figure 3A). The transcribed region mark (H3K36me3) and the enhancer marks (H3K4me1, H3K27ac) demonstrate significant changes between *intra*-PADs and *inter*-PADs, suggesting PADs play key roles in active gene regulation and expression (Figure 3A). Meanwhile, the repressive mark (H3K27me3) is enriched in topological boundaries, which can be associated with the formation of repressive domains. Interestingly, the signals of H3K9me3 in *inter*-PADs are slightly higher than those in the *intra*-PADs, and both of which are significantly lower than random ones, revealing that PADs can segregate the spread of the heterochromatin (Figure 3A). Thus, multiple epigenetic features can identify different functional PAD boundaries. We also observed that multiple epigenetic features (except the heterochromatin mark) are enriched around the regions with high interaction intensity (PAD centers) (Figure 3A). Furthermore, we guess that one or the combination of several epigenetic features contributes to the formation of PADs, which can be associated with different functional units (Figure 3B). For example, H3K27ac tends to combine H3K4me3 or H3K4me1 to mark active boundaries of PADs, while H3K27me3 tends to mark repressive boundaries alone (Figure 3B). Therefore, histone modifications and DNase signal seem to be a good predictor of PADs and different functional units may have different strategies to specify chromatin domains.

PADs can well capture the hematopoietic tree structure

The distribution of PADs among different cell types should reflect the affinity relationship of the 17 blood cell types.

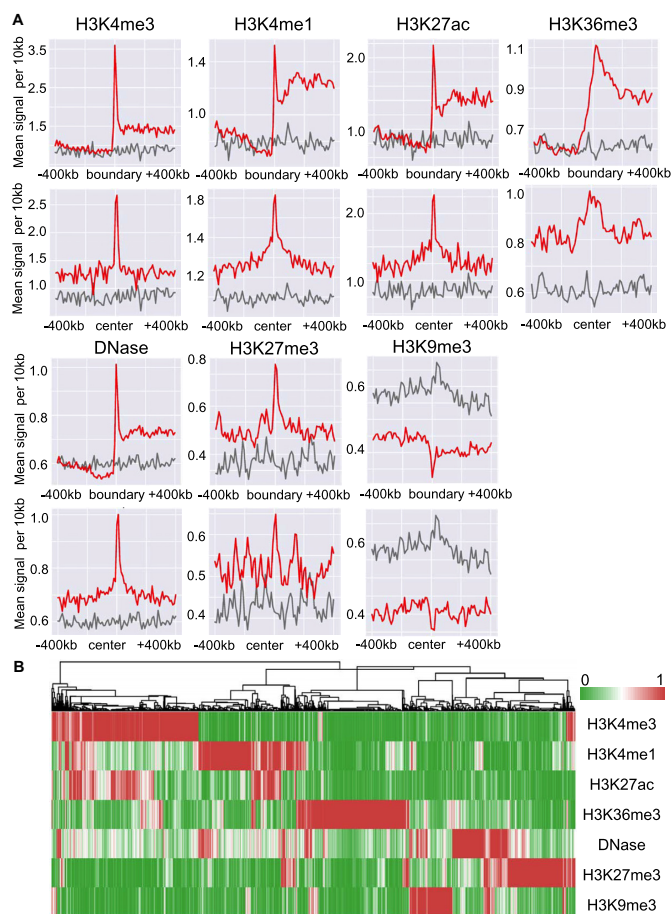


Figure 3. PADs are relating to epigenetic factors. (A) Enrichment analysis of seven epigenetic factors around boundaries and centers of PADs along PIRs axis not promoters' axis in the human monocytes cell type. The observed and random signals are marked with red and gray colors, respectively. (B) Hierarchical clustering of boundaries of PADs according to their epigenetic factors. The value of heatmap represents normalized score of columns of different epigenetic signals (min-max normalization to [0,1]).

We regarded that the PADs are conserved between different cell types if their boundaries are adjacent (Supplementary Methods) and noted that 40–70% of PADs are conserved between the same type of cells, yet only 30–50% of PADs are conserved between the different type of cells, indicating the conservation and dynamics of chromatin structure during cell development and differentiation (Supplementary Figure S6A). Furthermore, we derived five distinct classes based on the overlapping ratio of their PADs, which is highly consistent with the hematopoietic tree (28), suggesting dynamic chromatin structures contribute to gene cellular functions (Figure 4A). The expression level of genes within these PADs distinguishes them into high (low) co-expressed ones and divided them into four categories of cell types consistent with hematopoietic tree, indicating these PADs play important roles in gene expression and regulation (Figures 4B, C and Materials and Methods) (28). Interestingly, the gene sets tend to be very diverse across cell types within high and low co-expressed PADs respectively, indicating their underlying dynamical characteristics (Figure 4D). The above results reveal that the gradual change

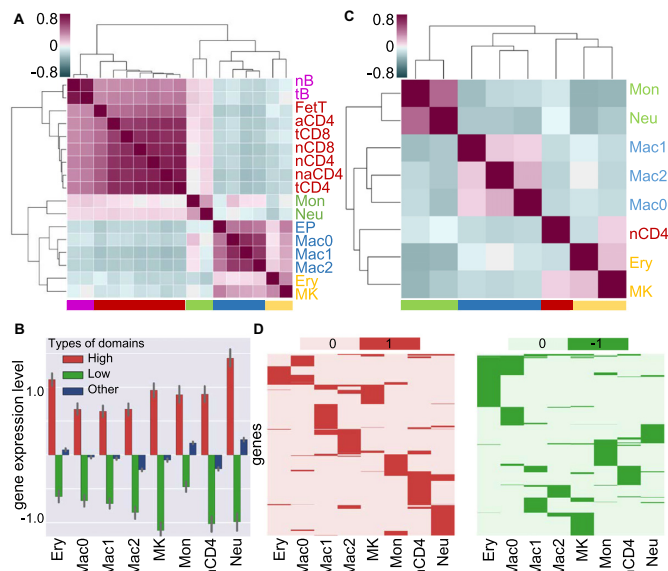


Figure 4. PADs can well capture the hematopoietic tree structure with distinct expression features. (A) Hierarchical clustering of the 17 cell types in terms of overlapping ratio of their PADs. Five clusters are marked by different colors. The value of heatmap represents Pearson's correlation coefficient about the overlap ratio of PADs. (B) The expression level of genes within high (low) expression level and other domains. (C) Hierarchical clustering of the eight cell types in terms of the correlation of z -score of PADs expression levels. The value of heatmap represents Pearson's correlation coefficient based on the expression level of PADs. (D) The genes of high (low) expression domains are marked by red and green respectively across eight cell types.

of the chromosomal structures (e.g. PADs) regulate the expression level of specific genes, which in turn promote the occurrence of cell development and differentiation. Therefore, it is important to distinguish these cell-specific PADs compared with stable ones for understanding the formation mechanism and cellular functions of PADs.

Next, we identified 693 lymphoid-specific, 426 myeloid-specific and 3808 common PADs according to the occurrence features among 17 human primary blood cell types to explore their cellular functions (Supplementary Methods). We further obtained 308 lymphoid-specific, 195 myeloid-specific and 1380 common PADs, whose two interacting chromatin regions containing more than one marker genes (Supplementary Methods and Supplementary Table S3). Among those PADs, MSTD pioneered the discovery of 180 trans-interacting PADs without any overlap between their two interacting chromatin regions (Figure 5, Supplementary Figures S7–S14 and Supplementary Tables S4 and S5).

For example, we found two interesting *trans*-interacting PADs within the same region of genome—one myeloid-specific PAD and a common PAD shared by 16 cell types (except aCD4) (Figure 5 and Supplementary Figures S7 and S8). For the myeloid-specific PAD, we can see that active TSS and transcription states occupies the two interacting chromatin regions of this PAD based on the ChromHMM annotations and active markers including H3K4me3, H3K27ac and DNase are significantly enriched in them too, which may suggest the chromatin loops within this PAD bring active promoters into the same nuclear

space to initiate transcription (29). Moreover, E2F, p53 motifs and E2F, E2F1 motifs are enriched in the upstream and downstream regions of this PAD, which contain genes related to innate immune response of myeloid tissues (upstream: HIST1H2BC, downstream: HIST1H2BJ and HIST1H2BK) (Supplementary Tables S4 and S5) (30). This suggests that the chromatin loops among this myeloid-specific PAD play a special role in the cooperation between E2F1 and p53 to specifically activate innate immune response genes in myeloid tissues, which could promote the occurrence of apoptosis (31).

For the lymphoid and myeloid common PAD, the two interacting chromatin regions contain several genes from histone families (H1, H2A, H2B, H3 and H4), which regulates circulating iron and mediates the regulation of lymphocyte and leukocyte (such as activation, cell-cell adhesion, immune response, antigen processing and presentation) (Figure 5 and Supplementary Table S4) (30). Furthermore, we identified 529 common and 40 dynamic chromatin loops between lymphoid and myeloid tissues from high-density chromatin loops of this PAD (P -value < 0.01) (Supplementary Methods). Surprisingly, we found all of dynamic chromatin loops occur in lymphoid tissues and these dynamics interacting regions contains BTN3A2 gene involved in the adaptive immune response of lymphoid tissues (Supplementary Figure S6B and Supplementary Table S4) (30). Thus, this PAD might participate in the underlying functions of regular activities of blood cells while enforce the adaptive immune response in lymphoid tissue. It can be seen that the upstream region of the common PAD contains that of the myeloid-specific PAD, and compared with the common PAD, the myeloid-specific PAD performs specific cellular functions by dynamic interactions with the different remote-region (Figure 5).

Another example is a lymphoid-specific PAD, which is characterized by a handful of active marks (Supplementary Methods and Supplementary Figure S9). Both chromatin regions of this PAD contain more than one motifs associated with lymphocyte (Supplementary Table S5). Specifically, GATA3 motif is a transcriptional activator, which binds to the enhancer of the T-cell receptor alpha and delta genes and Nur77 motif participate in modulating apoptosis in developing thymocytes (32). Both GATA3 and Nur77 motifs are enriched in the upstream regions of this PAD, while TFE3 and SPIB motifs are enriched in its downstream regions. Previous studies have showed that TFE3 motif regulates T-cell-dependent antibody in aCD4 cell type and thymus-dependent humoral immunity, and SPIB motif is a lymphoid-specific enhancer (32). In addition, Dontje *et al.* revealed that Delta-like1-induced Notch1 signaling pathway directs T cells and plasmacytoid dendritic cells decision by controlling the levels of GATA3 and SPIB (33). More interestingly, within the genome region of the PAD, RC3H2 gene participates in T cell activation and differentiation involving in immune response and PTGS1 gene is related to blood pressure and circulation (Supplementary Table S4) (30). Thus, this PAD could perform lymphoid-specific functions by T cell specific marker gene expression, which are regulated by GATA3, SPIB and other transcription factors corporately (e.g. TEF3 and Nur77) (Supplementary Methods).

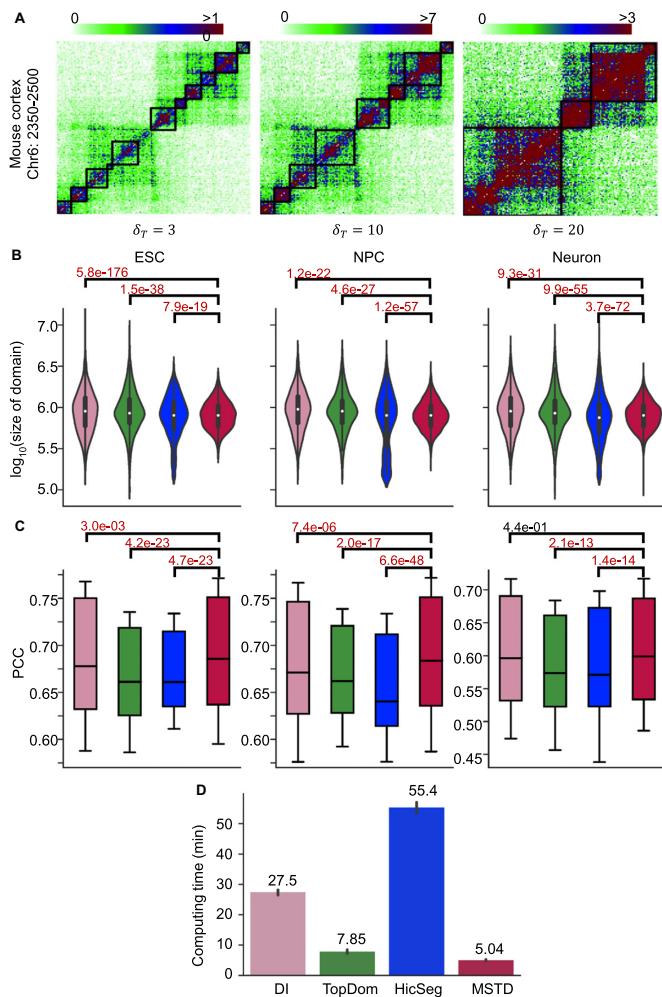


Figure 6. Illustration of multi-scale domains determined by MSTD from Hi-C datasets. (A) Multi-scale topological domain examples detected by MSTD with different MDHD thresholds (i.e. $\delta_T = 3, 10, 20$) of chr19: 2350–2500 bins of Hi-C maps in the mouse cortex cell (40 kb resolution). The value of heatmap represents normalized interacting scores. (B) The size distribution of TADs detected by DI, TopDom, HicSeg, MSTD in three mouse cells (P -value $< 7.9 \times 10^{-19}$, F -test). (C) Quality evaluation of TADs detected by the four methods in three cells using PCC measurements (P -value $< 3.0 \times 10^{-3}$ except one case, T -test). The dark horizontal lines in boxes represent mean. The box surrounding each mean indicates the middle part of the data, which is the range from the 15th to 85th percentile. (D) Computing time of the four methods for domain detection across mouse ESC, NPC and Neurons cells.

incides well with compacted and elongated domain conformations in biology (27,34).

MSTD is friendly to users in parameter selections with very short computing time. Specifically, users can set an adjustable parameter δ_T to control the domain scale (Supplementary Figure S15 and Supplementary Methods). Meanwhile, MSTD takes ~ 5.04 min, while TopDom takes ~ 7.85 min, DI takes ~ 27.5 min and HicSeg spends 55.4 min approximately for detecting TADs of all chromosomes of three cell lines (mouse ESC, NPC and Neurons) in same computer environment (inter core 3.4 GHz and 24G RAM) (Figure 6D).

TADs show distinct biological conservation and specificity among ESC, NPC and Neurons

Previous studies have shown that TADs are conserved between cell types (7,35). We observed the average of common boundary ratio (CBR) is about 50%–80% and the average of TADs overlap ratio (TOR) is approximately 35%–70% during cell differentiation, which is similar with previous studies (Dixon *et al.*, 2012) (Supplementary Methods and Supplementary Figure S19). Interestingly, the conservation between cells during continuous differentiation cycle is significantly higher than those between the interval cycles. It will be useful to further investigate the result and extend it to specific functional and evolutionary events (25).

TADs are strongly correlated with epigenetic patterns and expressed features

The conservation of TADs during cell differentiation promotes us to explore the mechanisms of TADs formation. Recent studies showed that a common property of CTCF and highly active housekeeping genes may create insulating force of TAD boundaries (24,35). Intriguingly, architectural protein CTCF is indeed highly enriched at TAD boundaries in mouse ES and cortex cells (Figure 7A and Supplementary Figure S20), and housekeeping exons and genes tend to be located in the boundary region of human ES cell (Figure 7B). Furthermore, H3K4me3, H3K9ac and RNA Polymerase II are significantly enriched and H3K36me3 is slightly enriched in the boundary regions of TADs for mouse ES and cortex cells, indicating the formation of TADs is indeed promoter-related (35). Meanwhile, we found that transcription start sites (TSS) and expressed genes (FPKM >3) are also enriched around TAD boundaries (Figures 7A and B). Those observations reveal that factors associated active promoters and gene bodies can contribute to the formation of TADs boundaries (Figure 7A and Supplementary Figure S20). These evidences can help to explain the formation model of TADs, such as, loop extrusion model (35,36). Moreover, enhancer-marks (H3K4me1, H3K27ac and EP300) show a bimodal distribution around TAD boundaries (Supplementary Figure S20). These peaks are located at 100–200 kb away from the TADs boundaries, which may be associated with domains of gene co-regulation patterns (35). H3K4me1 has a different bimodal patterns between mouse ES and cortex cells, which may be attributed to H3K4me1 marks sequences with region of DNA methylation loss in human mesenchymal stem cells (Supplementary Figure S20) (37). Interestingly, open chromatin mark (DNase signal) shows a significant enrichment in TAD boundaries of mouse ESC cell and a three peaks distribution around TADs boundaries of cortex cell (Figure 7A and Supplementary Figure S20). In addition, inactive chromosome mark (H3K27me3) shows a slight peak in TADs boundaries for mouse ESC cell but not for cortex cell, which may be due to the responsibility of H3K27me3 for the repression of genes involved in cellular development and differentiation (Figure 7A and Supplementary Figure S20) (38). Finally, H3K9me3 is the mark of heterochromatin and shows no enrichment in both mouse ES and cortex cells (Figure 7A and Supplementary Figure S20).

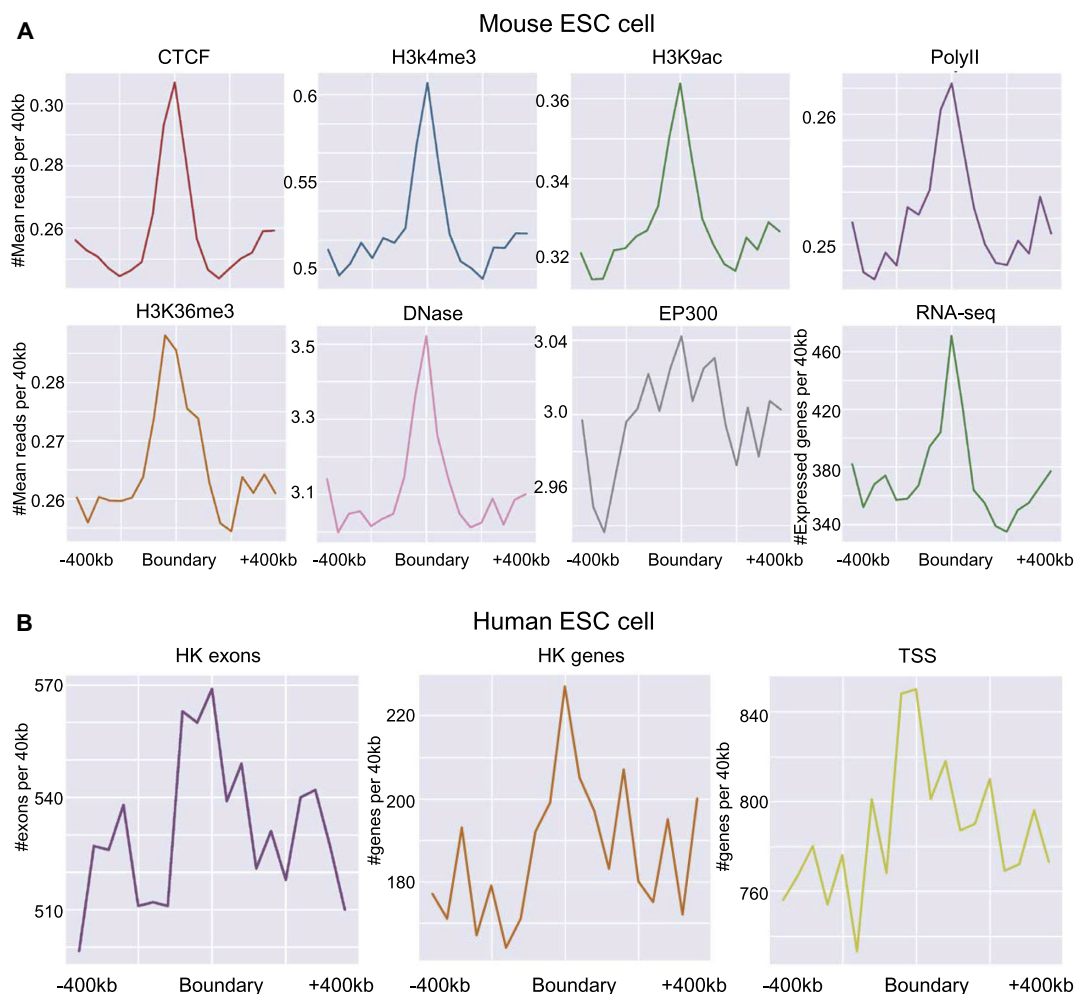


Figure 7. TADs are strongly related to epigenetic patterns and expression features. **(A)** The enrichment analysis of architectural protein CTCF, histone modifications, RNA Polymerase II, DNase-seq, EP300 elements and expressed genes around the boundaries of TADs identified by MSTD in mouse ES cell. **(B)** The enrichment analysis of housekeeping (HK) exons, HK genes and TSS of protein coding genes of hg19 RefSeq in GENCODE database in the boundaries of TADs identified by MSTD in human ES cell.

The case study of PTADs

We also applied MSTD to identify PTADs on 5M away from the diagonal of Hi-C maps of mouse cortex and ES cell line at 40 kb resolution (7). We found 43 and 4 PATDs on chromosome 19 of these two cell lines respectively (Supplementary Figure S21A and B). For examples, we found one PTAD is specific in mouse cortex cell line (Supplementary Figure S21C). The two interacting chromatin regions of the PTAD contains genes (CPEB3 and DNMBP) involving in neural system, respectively. Previous studies have shown that CPEB3 activity and CPEB3-dependent protein synthesis can facilitate hippocampal plasticity and hippocampal-dependent memory storage (39) and DNMBP regulates actin cytoskeleton and synaptic vesicle pools, whose expression is lower in neuropathologically-confirmed Alzheimer's brains (40). We also found that the two interacting chromatin regions of the PTAD contains multiple genes (Cyp2c44, Cyp26c1 and Cyp26a1) in CYP family. The CYP enzymes are tissue- and cell type-specific expression in the brains and these

isozymes can metabolize a vast array of compounds including centrally acting drugs, neurotoxins, neurotransmitters, and neurosteroids (41). This suggests that PTADs could play an important role in neural system of mouse. Another example is a common PTAD in the two mouse cell lines (Supplementary Figure S21D and E). We found enhancer marks (H3K27ac and H3K4me1) are enriched in two interacting chromatin regions of the PTAD. These enhancer regions may constitute 'super enhancer', which regulates functional genes with the PTAD. Further investigation shows that one end of the PTAD contains multiple genes (Add3, Dusp5 and Adra2a) that regulate vascular process in circulatory system (29). This suggests that the PTAD may be an important and basic unit of genomic structure and function in mouse cells.

DISCUSSION

The chromosome conformation capture (3C) and its variants produce billions of read pairs that are used to draw genome-wide chromatin contact maps for multiple tissues

and species. The accumulation of these data poses unprecedented challenges to computational biologists. Different tools have been developed to detect domains from symmetric Hi-C maps (42). However, these methods pay no attention to interaction structures between domains (*trans*-domains) directly. MSTD begin to explore a variety of chromatin structures, including multi-scale TADs, PTADs and PADs from diverse chromatin contact maps, especially promoter capture Hi-C maps. Exploring a variety of chromatin structures will help to better understand the formation of spatial chromatin architecture and distinguish cellular function units. Meanwhile, 3D genomic maps of higher resolution will contribute to the presentation of landscape in the nucleus clearly and accurately.

We believe that the identification of multi-scale *cis*- and *trans*-domains will provide insights into local chromatin structure and promote the construction of 3D genomic model. Meanwhile, the analysis combining a variety of functional factors gradually explain cellular functions and biological processes of these structural units. With the development of new technologies, more diverse chromatin data will be generated for solving unknown biological problems. For examples, single cell Hi-C maps provide an opportunity to study the conservation and heterogeneity from cell to cell, revealing more granular cellular functions and biological processes in cell cycle (43,44). Moreover, what the difference is between PADs and TADs will be a very interesting question. We hope to make some deep exploration about their connection in the near future when both Hi-C and promoter capture Hi-C maps are available from the same cell lines or tissues. MSTD is a very flexible and powerful framework, which is expected to be applied to such new data in near future.

DATA AVAILABILITY

MSTD (version 0.0.2) is free, open source software under MIT License (OSI-compliant), is implemented in python3.6, and is freely available at <https://github.com/zhanglabtools/MSTD> or <http://page.amss.ac.cn/shihua.zhang/software.html>.

All datasets in this manuscript are public datasets. Their available link addresses are placed on Supplementary information.

SUPPLEMENTARY DATA

[Supplementary Data](#) are available at NAR Online.

ACKNOWLEDGEMENTS

Yusen Ye would like to thank the support of the Academy of Mathematics and Systems Science at CAS during his visit.

FUNDING

National Natural Science Foundation of China [61873198, 61532014, 61432010, 61672407 to L.G.; 11661141019, 61621003, 61422309, 61379092 to S.Z.]; Strategic Priority Research Program of the Chinese Academy of Sciences (CAS) [XDB13040600]; Key Research Program of

the Chinese Academy of Sciences [KFZD-SW-219]; National Key Research and Development Program of China [2017YFC0908405]; CAS Frontier Science Research Key Project for Top Young Scientist [QYZDB-SSW-SYS008]. Funding for open access charge: National Natural Science Foundation of China.

Conflict of interest statement. None declared.

REFERENCES

- Bickmore, W.A. and van Steensel, B. (2013) Genome architecture: domain organization of interphase chromosomes. *Cell*, **152**, 1270–1284.
- Sexton, T. and Cavalli, G. (2015) The role of chromosome domains in shaping the functional genome. *Cell*, **160**, 1049–1059.
- Pombo, A. and Dillon, N. (2015) Three-dimensional genome architecture: players and mechanisms. *Nat. Rev. Mol. Cell Biol.*, **16**, 245–257.
- Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D. and Lander, E.S. (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.
- Schoenfelder, S., Sexton, T., Chakalova, L., Cope, N.F., Horton, A., Andrews, S., Kurukuti, S., Mitchell, J.A., Umlauf, D. and Dimitrova, D.S. (2010) Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat. Genet.*, **42**, 53–61.
- Bantignies, F., Roure, V., Comet, I., Leblanc, B., Schuettengruber, B., Bonnet, J., Tixier, V., Mas, A. and Cavalli, G. (2011) Polycomb-dependent regulatory contacts between distant Hox loci in *Drosophila*. *Cell*, **144**, 214–226.
- Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S. and Ren, B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.
- Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J. and Sedat, J. (2012) Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, **485**, 381–385.
- Phillips-Cremins, J.E., Sauria, M.E., Sanyal, A., Gerasimova, T.I., Lajoie, B.R., Bell, J.S., Ong, C.-T., Hookway, T.A., Guo, C. and Sun, Y. (2013) Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*, **153**, 1281–1295.
- Dali, R. and Blanchette, M. (2017) A critical assessment of topologically associating domain prediction tools. *Nucleic Acids Res.*, **45**, 2994–3005.
- Norton, H.K., Emerson, D.J., Huang, H., Kim, J., Titus, K.R., Gu, S., Bassett, D.S. and Phillips-Cremins, J.E. (2018) Detecting hierarchical genome folding with network modularity. *Nat. Methods*, **15**, 119.
- Malik, L. and Patro, R. (2018) Rich chromatin structure prediction from Hi-C data. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, doi:10.1109/TCBB.2018.2851200.
- Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragozy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
- Lévy-Leduc, C., Delattre, M., Mary-Huard, T. and Robin, S. (2014) Two-dimensional segmentation for analyzing Hi-C data. *Bioinformatics*, **30**, i386–i392.
- Crane, E., Bian, Q., McCord, R.P., Lajoie, B.R., Wheeler, B.S., Ralston, E.J., Uzawa, S., Dekker, J. and Meyer, B.J. (2015) Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature*, **523**, 240–244.
- Shin, H., Shi, Y., Dai, C., Tjong, H., Gong, K., Alber, F. and Zhou, X.J. (2016) TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res.*, **44**, e70.
- Filippova, D., Patro, R., Duggal, G. and Kingsford, C. (2014) Identification of alternative topological domains in chromatin. *Algorithms Mol. Biol.*, **9**, 14.

18. Weinreb, C. and Raphael, B.J. (2016) Identification of hierarchical chromatin domains. *Bioinformatics*, **32**, 1601–1609.
19. Zhan, Y., Mariani, L., Barozzi, I., Schulz, E.G., Bluthgen, N., Stadler, M., Tiana, G. and Giorgetti, L. (2017) 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.*, **27**, 479–490.
20. Fraser, J., Rousseau, M., Shenker, S., Ferraiuolo, M.A., Hayashizaki, Y., Blanchette, M. and Dostie, J. (2009) Chromatin conformation signatures of cellular differentiation. *Genome Biol.*, **10**, R37.
21. Javierre, B.M., Burren, O.S., Wilder, S.P., Kreuzhuber, R., Hill, S.M., Sewitz, S., Cairns, J., Wingett, S.W., Várnai, C. and Thiecke, M.J. (2016) Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell*, **167**, 1369–1384.
22. Zhang, Y., Wong, C.-H., Birnbaum, R.Y., Li, G., Favaro, R., Ngan, C.Y., Lim, J., Tai, E., Poh, H.M. and Wong, E. (2013) Chromatin connectivity maps reveal dynamic promoter–enhancer long-range associations. *Nature*, **504**, 306.
23. Shen, Y., Yue, F., McClery, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L. and Lobanenkov, V.V. (2012) A map of the cis-regulatory sequences in the mouse genome. *Nature*, **488**, 116.
24. Eisenberg, E. and Levanon, E.Y. (2013) Human housekeeping genes, revisited. *Trends Genet.*, **29**, 569–574.
25. Fraser, J., Ferrai, C., Chiariello, A.M., Schueler, M., Rito, T., Laudanno, G., Barbieri, M., Moore, B.L., Kraemer, D.C. and Aitken, S. (2015) Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Mol. Syst. Biol.*, **11**, 852.
26. Yaffe, E. and Tanay, A. (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.*, **43**, 1059.
27. Rodriguez, A. and Laio, A. (2014) Clustering by fast search and find of density peaks. *Science*, **344**, 1492–1496.
28. Görgens, A., Radtke, S., Möllmann, M., Cross, M., Dürig, J., Horn, P.A. and Giebel, B. (2013) Revision of the human hematopoietic tree: granulocyte subtypes derive from distinct hematopoietic lineages. *Cell Rep.*, **3**, 1539–1552.
29. Sutherland, H. and Bickmore, W.A. (2009) Transcription factories: gene expression in unions? *Nat. Rev. Genet.*, **10**, 457.
30. Consortium, G.O. (2015) Gene ontology consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.
31. Polager, S. and Ginsberg, D. (2009) p53 and E2f: partners in life and death. *Nat. Rev. Cancer*, **9**, 738–748.
32. Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R. and Magrane, M. (2004) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.
33. Dontje, W., Schotte, R., Cupedo, T., Nagasawa, M., Scheeren, F., Gimeno, R., Spits, H. and Blom, B. (2006) Delta-like1-induced Notch1 signaling regulates the human plasmacytoid dendritic cell versus T-cell lineage decision through control of GATA-3 and Spi-B. *Blood*, **107**, 2446–2452.
34. de Wit, E. (2017) Capturing heterogeneity: single-cell structures of the 3D genome. *Nat. Struct. Mol. Biol.*, **24**, 437–438.
35. Dixon, J.R., Gorkin, D.U. and Ren, B. (2016) Chromatin domains: the unit of chromosome organization. *Mol. Cell*, **62**, 668–680.
36. Sanborn, A.L., Rao, S.S., Huang, S.-C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A. and Li, J. (2015) Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, E6456–E6465.
37. Fernández, A.F., Bayón, G.F., Urduño, R.G., Torano, E.G., García, M.G., Carella, A., Petrus-Reurer, S., Ferrero, C., Martínez-Camblor, P. and Cubillo, I. (2015) H3K4me1 marks DNA regions hypomethylated during aging in human stem and differentiated cells. *Genome Res.*, **25**, 27–40.
38. Kuzmichev, A., Nishioka, K., Erdjument-Bromage, H., Tempst, P. and Reinberg, D. (2002) Histone methyltransferase activity associated with a human multiprotein complex containing the Enhancer of Zeste protein. *Genes. Dev.*, **16**, 2893–2905.
39. Pavlopoulos, E., Trifilieff, P., Chevaleyre, V., Fioriti, L., Zairis, S., Pagano, A., Malleret, G. and Kandel, E.R.J.C. (2011) Neuralized1 activates CPEB3: a function for nonproteolytic ubiquitin in synaptic plasticity and memory storage. *Cell*, **147**, 1369–1383.
40. Casoli, T., Di Stefano, G., Fattoretti, P., Giorgetti, B., Baliani, M., Lattanzio, F., Aicardi, G. and Platano, D. (2012) Dynamin binding protein gene expression and memory performance in aged rats. *Neurobiol. Aging*, **33**, 618.
41. Miksys, S. and Tyndale, R.F.J.N. (2009) Brain drug-metabolizing cytochrome P450 enzymes are active in vivo, demonstrated by mechanism-based enzyme inhibition. *Neuropsychopharmacology*, **34**, 634.
42. Forcato, M., Nicoletti, C., Pal, K., Livi, C.M., Ferrari, F. and Bicciato, S. (2017) Comparison of computational methods for Hi-C data analysis. *Nat. Methods*, **14**, 679.
43. Stevens, T.J., Lando, D., Basu, S., Atkinson, L.P., Cao, Y., Lee, S.F., Leeb, M., Wohlfahrt, K.J., Boucher, W. and O’Shaughnessy-Kirwan, A. (2017) 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature*, **544**, 59–64.
44. Ramani, V., Deng, X., Qiu, R., Gunderson, K.L., Steemers, F.J., Disteche, C.M., Noble, W.S., Duan, Z. and Shendure, J. (2017) Massively multiplex single-cell Hi-C. *Nat. Methods*, **14**, 263–266.